Proceedings
20th International Parallel and Distributed Processing Symposium

IPDPS 2006
Abstracts and CD-ROM

IEEE Catalog Number: 06TH8860
ISBN: 1-4244-0054-6
ISSN: 1530-2075

Proceedings
20th International Parallel and Distributed Processing Symposium

April 25–29, 2006
Rhodes Island, Greece

Sponsored by
IEEE Computer Society Technical Committee on Parallel Processing

In Cooperation with
IEEE Computer Society Technical Committee on Computer Architecture (TCCA)
IEEE Computer Society Technical Committee on Distributed Processing (TCDP)
ACM SIGARCH

Summary of Contents

Detailed Table of Contents

International Parallel and Distributed Processing Symposium
IPDPS 2006

Message from the General Co-Chairs


Welcome to the 20th International Parallel and Distributed Processing Symposium (IPDPS 2006), in the beautiful Island of Rhodes, Greece. It is a great honor for us to serve the international scientific community by organizing this major event of parallel and distributed computing, bringing together researchers, scientists, and students, from academia, research laboratories and industry. The Research Academic Computer Technology Institute (CTI) is proud to host this year's IPDPS.

These proceedings present the outstanding technical program of IPDPS 2006. We thank Program Chair Arnold L. Rosenberg, University of Massachusetts Amherst, USA, for all of his hard and excellent work in assembling this high quality program, in cooperation with his able Program Vice-Chairs: for "Algorithms" - Mikhail J. Atallah, Purdue University, USA; for "Applications" - David A. Bader, Georgia Institute of Technology, USA; for "Architectures" - Allan Gottlieb, New York University, USA; and for "Software" - Laxmikant Kale, University of Illinois, Urbana-Champaign, USA. We also thank the Program Committee members and external reviewers, listed elsewhere in these proceedings, who assisted the paper review and selection process.

We acknowledge the efforts of the authors of papers submitted to IPDPS 2006, without whom this conference series would not be possible. Due to the very competitive selection this year, many strong papers could not be included in the final program. We hope that authors will continue to view IPDPS as the premiere event for the dissemination of their novel, high quality research in parallel and distributed computing.

For the special speakers, we want to add our thanks to those expressed by the Program Chair in his message. In particular, we thank the four keynotes speakers: William Dally, Manish Gupta, Yves Robert, and Horst Simon.

We are grateful to the members of the IPDPS 2006 organizing committee. General Vice-Chair, Sotiris Nikoletseas, University of Patras and CTI, Greece, envisioned holding IPDPS 2006 in Greece, prepared and presented the corresponding proposal and co-cordinated several local organization aspects. General Vice-Chair Chip Weems, University of Massachusetts at Amherst, USA, has assisted with many aspects of the conference by offering his valuable experience, including provision of guidance to the Workshop Chairs.

A collection of 19 workshops planned to collocate with IPDPS this year have been co-ordinated through the efforts of Workshop Chair Alan Sussman, University of Maryland, USA, and Workshops Vice-Chair Yuanyuan Yang, State University of New York, Stony Brook, USA. We acknowledge the hard work of the organizers of each workshop, listed elsewhere in these proceedings. These workshops offer an opportunity to explore a great variety of special topics related to parallel and distributed computing, and are an important part of the IPDPS events.

There will be interesting commercial exhibits and talks by IPDPS 2006 industrial sponsors, through the efforts of the Commercial Presentations and Exhibits Chair John K. Antonio, University of Oklahoma, USA, and the Co-Chair Theodoros Komninos, Computer Technology Institute, Greece. The Proceedings Chair Shoukat Ali, University of Missouri-Rolla, USA, has done a great job in preparing the program/abstract book and the conference proceedings, including handling the camera-ready manuscripts for the conference and the associated workshops. We also appreciate the dissemination efforts of the Publicity Coordinators Bo Hong, Drexel University, USA (Americas), Cho-Li Wang, University of Hong Kong, China (Asia/Pacific RIM), and Ioannis Chatzigiannakis, Computer Technology Institute, Greece (Europe/Africa). We thank Anna Brown (www.mediagirl.com) for maintaining the IPDPS web page.

Sally Jelinek, Bill Pitts, and Susamma Barua are three perennial members of the IPDPS organizing committees. Myriad important administrative details are handled by them with great skill and experience. Our job as General Co-

Chairs would have been unbearably difficult without them. Sally Jelinek, Electronic Design Associates, Inc., USA, is the Production Chair, and manages the logistical arrangements, oversees publicity, is responsible for our web site, produces the call for papers, pulls together the front matter for the proceedings, helps construct the program and abstract book distributed at the conference etc., etc., etc. Local Arrangements Co-Chair Susamma Barua, California State University, Fullerton, USA, works on the conference all year managing the detailed arrangements for a variety of local aspects such as the hotel contract, meeting room space, audio-visual equipment, registration, wireless internet, and signage. The Finance Chair, Bill Pitts, Toshiba America Information Systems, Inc., USA, prepares an accurate and detailed conference budget, tracks all income and expenses, and produces the final fiscal report at the highest level of financial integrity. We are very grateful to Sally, Bill, and Susamma for shielding us from these time-consuming and important chores that in most conferences the General Co-Chairs would have to handle. This year two senior staff members of the Computer Technology Institute, Greece, Rozina Efstathiadou (head of the administrative and financial department) and Lena Gourdoupi (public relations and conference organization), contributed significantly to various, important local organization aspects, with hard work and professionalism.

Finally, we thank the IPDPS Steering Committee Co-Chairs, Viktor K. Prasanna, University of Southern California, USA, and George Westrom, Future Scientists & Engineers of America, USA, for their dedication and leadership of IPDPS. It is to their credit that this meeting has become the premiere international conference for parallel and distributed computing. In addition, we thank the rest of the Steering Committee, listed elsewhere in these proceedings, for their guidance and contributions to this meeting.

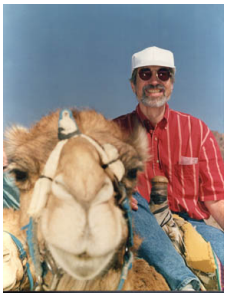
IPDPS is sponsored by the IEEE Computer Society, and held in cooperation with the ACM. Furthermore, IPDPS is the flagship activity of the IEEE Computer Society's Technical Committee on Parallel Processing (TCPP). If you are not already a member, please consider joining TCPP. For more information about TCPP, please see www.computer.org /tab/tclist/tcpp.htm.

As mentioned earlier, the Research Academic Computer Technology Institute (CTI) is hosting IPDPS 2006. CTI was founded in 1985, with head offices in Patras, Greece. It is a non-profit, financially and administratively independent institution, under the supervision of the Ministry of Education and Religious Affairs. Its workforce averages at about 260 persons, including experienced researchers, University faculty members, computer engineers and technicians, scientific staff specialized in other fields, post-graduate students and administrative staff. CTI establishes collaborative networks with a host of academic and research institutes, companies and institutions at an international, European and national level. At the same time, at the scientific level, CTI's researchers develop scientific collaborations in issues concerning basic and applied research with scientists and research centres belonging to the global scientific community. CTI's chief priority is research and development, focusing on certain areas of strategic importance, of both basic and applied nature. The research endeavours of the institute are stimulated by the framework and the objectives for research policy set by the European Union, in conjunction with the technological needs of the country. Please learn more about the activities at CTI from its web site: www.cti.gr. We thank CTI for hosting this year's IPDPS meeting.

We hope you find these proceedings very informative and useful. We encourage you to look also at past and future IPDPS proceedings.

IPDPS 2006 General Co-Chairs
Paul Spirakis, University of Patras and CTI, Greece
H. J. Siegel, Colorado State University, USA

Greetings from the Program Chair



This was a special year in the life of the *International Parallel and Distributed Processing Symposium (IPDPS)*. We received a record number of submissions, 531 — almost 100 more than the previous record and close to 200 more than in a recent "typical" year. Of these submissions, 16 were ruled "out of scope;" all others received the careful scrutiny of at least three reviewers. Based on these reviews, 125 papers were accepted for presentation at the conference. On behalf of the entire Organizing Committee, it is my pleasure to congratulate the authors of these 125 papers, while thanking the authors of all 531 submissions for submitting their research to *IPDPS* this year. The overall quality of the submissions raised the bar for the entire conference.

Special congratulations are in order for the authors of the four Best-Paper Award winning submissions:

- to Yaacov Fernandess and Dahlia Malkhi, for "On Collaborative Content Distribution Using Multi-Message Gossip" (Algorithms Track);
- to Anantharaman Kalyanaraman, Scott Emrich, Patrick Schnable, and Srinivas Aluru for "Assembling Genomes on Large-Scale Parallel Computers" (Applications Track);
- to Resit Sendag, Ayse Yilmazer, Joshua Yi, and Augustus Uht for "Quantifying and Reducing the Effects of Wrong-Path Memory References in Cache-Coherent Multiprocessor Systems" (Architecture Track);
- to Thomas Hart, Paul McKenney, and Angela Demke Brown for "Making Lockless Synchronization Fast: Performance Implications of Memory Reclamation" (Software Track).

These papers appear in a dedicated plenary session.

It is both an honor and a pleasure to thank our four Keynote speakers, William Dally, Manish Gupta, Yves Robert, and Horst Simon, for their important contributions to the intellectual excitement at *IPDPS'06*.

There is no space in this message for individually acknowledging all those who have made this Conference happen, so let me address them in groups. I have already thanked all of the authors who, by submitting their research, supplied the raw material for our excellent program. It is a real pleasure to thank the multitude of reviewers — both the official ones on the Program Committee and the unofficial ones whose help was enlisted for specific papers — who provided the perspectives and insights that allowed us to select the 125 accepted papers from among the 531 submissions. Their names appear elsewhere in the Proceedings.

The debt we all owe to the four Vice Chairs of the Program Committee, Mike Atallah (Algorithms Track), David Bader (Applications Track), Allan Gottlieb (Architecture Track), and Laxmikant (Sanjay) Kale (Software Track), transcends the expressive power of words. Faced with an unprecedented workload, these folks did an unbelievable job of coordinating and coaxing, encouraging and exhorting, that culminated in a 17-hour marathon Program Committee meeting during which the Program was selected. It was both an honor and a privilege to work with these diligent, dedicated professionals.

I extend warm thanks to H.J. Siegel and Paul Spirakis, our General Chairs, and to Viktor Prasanna, Co-Chair of our Steering Committee, for wise counsel on a broad range of issues—and for inviting me to serve as this year's program chair.

Last, but certainly not least, it is an unmitigated pleasure to thank the many people who made our work on the Program Committee easier in so many ways. I want to single out Sally Westrom of the *IPDPS* Organizing Committee, for guidance and help on myriad issues, and Xiaotao Wu of Columbia, for innumerable hints on how to navigate the EDAS System.

I wish you all a very productive conference, and a memorable stay on the beautiful Isle of Rhodes.

IPDPS 2006 Program Chair
Arnold L. Rosenberg
University of Massachusetts, Amherst

Message from the General Vice-Chair, Local Organization

The International Parallel and Distributed Processing Symposium (IPDPS) has become a major event for high quality research in several important areas of parallel and distributed computing (including algorithms, applications, architectures and software) and diverse topics related to emerging technologies and important special aspects. It is a great honor for me to co-organize such an event!

When two years ago, during IPDPS 2004 in Santa Fe, our proposal for holding IPDPS 2006 in Greece got accepted, I promised to do my best for contributing to its successful organization. Holding such an important and complex event, especially outside the American continent, is a difficult challenge and a great responsibility. Its successful organization involves several aspects and innumerable details that can be effectively handled only through the collaborative efforts of several people; I like to thank all people involved in organizing IPDPS 2006 for a fruitful cooperation.

I wish to especially thank Viktor Prasanna for his trust and guidance; Paul Spirakis for accepting to lead this effort and for coordinating local organization; Jose Rolim for his encouragement to hold such an event; the CTI people for an efficient cooperation.

I sincerely hope you will enjoy IPDPS 2006 in Rodos, the beautiful "island of roses," with its unique physical scenery, its long history and diverse civilization!

IPDPS 2006 General Vice-Chair, Local Organization
Sotiris Nikoletseas
University of Patras and CTI, Greece

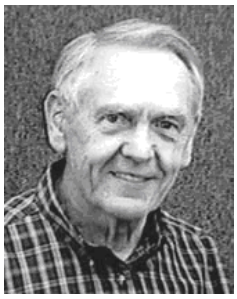
Message from the General Vice-Chair

In addition to the main conference, IPDPS has developed a wide range of additional elements in its annual program. These include 20 workshops, a commercial track with exhibits, presentations, and product tutorials, a general tutorial that is open to all participants, and a set of Birds- of-a-Feather sessions. Together, these elements greatly enrich the overall IPDPS experience for attendees. They also extend the program into a full five days and several evenings, and they expand the size of the IPDPS community to make it a larger and more vibrant intellectual gathering.

It takes many hands to organize these diverse activities, and ensure that they run smoothly. I would like to offer my thanks and gratitude to the following people for all of their hard work. The many workshops are made possible by Alan Sussman and Yuanyuan Yang, together with a host of organizers for each of the workshops. The commercial track owes its success to the efforts of John Antonio and Theodoros Komninos. Shoukat Ali, our proceedings chair, has discovered that the majority of his work is in handling the hundreds of workshop papers! Production chair Sally Jelinek works tirelessly to assemble all of the pieces of our program into advertising and both electronic and printed programs. The local arrangements co- chairs, Susamma Barua, Rozina Efstathiadou, and Lena Gourdoupi, take care of scheduling the many rooms, A/V equipment, and signs that are needed for all of our different events. And Bill Pitts is the dedicated finance chair who crunches the numbers that keep the whole endeavor afloat. As you encounter these folks around the conference, please take a moment to thank them. Volunteers are largely fueled by praise and the satisfaction of knowing that their work is appreciated. Let's all make sure that our excellent team of volunteers are aware of the gratitude that we feel for what they have done to make IPDPS the superb event that it is!

IPDPS 2006 General Vice-Chair
Charles Weems
University of Massachusetts, Amherst

Message from the Steering Co-Chairs

Welcome to Rhodes and to IPDPS 2006. As you may know, the meeting is held outside the US every third year, reflecting the true composition of the community it serves.

We are grateful to the host general chair, Paul Spirakis, for the several years of proposals and preparation that it took to bring us to Greece, and we thank his co-chair, H.J. Siegel, for bringing his template of experience to the tasks of conducting this event. This joint effort started with the organization of a superb group of volunteers whose accomplishments are well documented in previous messages.

The meeting has grown both with respect to the number of submitted papers as well as actual meeting participation. Organizing such an event at a distant location, in this case an island, is not an easy task. And this year presented a number of new challenges, which we can now report have been met with grace and excellence.

To start, we would like to acknowledge the contributions of Sotiris Nikoletseas and his local team in meeting the requirements of our conference while ensuring that IEEE Computer Society guidelines for organizing an event outside the US were strictly adhered to. This was most ably facilitated by the efforts of Bill Pitts, our finance chair, who implemented and oversaw an entirely new system for conference registration. Together Sotiris and Bill have sorted out the many contract and finance issues to support the success of this event, and we wish to give them our special thanks.

Arnold Rosenberg, our program chair, had the overwhelming task of handling over five-hundred submissions. Arny's efforts to ensure quality in the midst of what might have become chaos is greatly appreciated. In addition, we are thankful to him for his numerous suggestions to further improve the handling of submissions. We believe we have one of the strongest technical programs ever.

Although the proceedings continue as a publication of the IEEE CS Press, this year we handled production of the proceedings CD and book of abstracts through an alternate vendor. We are indebted to Shoukat Ali, our proceedings chair, whose inestimable effort allowed us to make this change and to once again bring together the regular symposium proceedings as well as those for the workshops in a timely manner.

As noted, we are planning for IPDPS 2009 and 2012 to be held outside the US, and we encourage you to contact us or any of the steering committee members with suggestions for locations and volunteers to lead such efforts.

We hope you enjoy the meeting, the antiquities, and the islands.

IPDPS 2006 Steering Co-Chairs
Viktor Prasanna, University of Southern California
George Westrom, Future Scientists & Engineers of America

IPDPS 2006 Organization

General Co-Chairs

Paul Spirakis, CTI & University of Patras, Greece
H.J. Siegel, Colorado State University, USA

General Vice-Chairs

Sotiris Nikoletseas, CTI & University of Patras, Greece
Charles Weems, University of Massachusetts Amherst, USA

Program Chair

Arnold L. Rosenberg, University of Massachusetts Amherst, USA

Steering Co-Chairs

Viktor K. Prasanna, University of Southern California
George Westrom, Future Scientists & Engineers of America

Steering Committee

David A. Bader, Georgia Institute of Technology, USA
K. Mani Chandy, California Institute of Technology, USA
Jean-Luc Gaudiot, University of California, Irvine, USA
Ali R. Hurson, Pennsylvania State University, USA
Joseph JaJa, University of Maryland, USA
F. Tom Leighton, MIT, USA
Sotiris E. Nikoletseas, CTI & University of Patras, Greece
Dhabaleswar K. Panda, Ohio State University, USA
Timothy Pinkston, University of Southern California, USA
José D.P. Rolim, University of Geneva, Switzerland
Arnold L. Rosenberg, University of Massachusetts Amherst, USA
Sartaj Sahni, University of Florida, USA
Behrooz Shirazi, Washington State University, USA
H. J. Siegel, Colorado State University, USA
Paul Spirakis, CTI & University of Patras, Greece
Hal Sudborough, University of Texas at Dallas, USA
Charles Weems, University of Massachusetts Amherst, USA

Workshops Committee

Chair: Alan Sussman, University of Maryland, USA
Vice-Chair: Yuanyuan Yang, State University of New York, Stony Brook, USA

Commercial Presentations & Exhibits Co-Chairs

John K. Antonio, University of Oklahoma, USA
Theodoros Komninos, Computer Technology Institute, Greece

Proceedings Chair

Shoukat Ali, University of Missouri-Rolla, USA

Finance Chair

Bill Pitts, Toshiba America Information Systems, Inc., USA

Local Arrangements Co-Chairs

Susamma Barua, California State University, Fullerton, USA
Rozina Efstathiadou, Computer Technology Institute, Greece
Lena Gourdoupi, Computer Technology Institute, Greece

Production Chair

Sally Jelinek, Electronic Design Associates, Inc., USA

Publicity Coordinators

Americas
Bo Hong, Drexel University, USA
Asia/Pacific Rim
Cho-Li Wang, University of Hong Kong, China
Europe/Africa
Ioannis Chatzigiannakis, Computer Technology Institute, Greece

Program Chair

Arnold L. Rosenberg, University of Massachusetts Amherst, USA

Program Vice-Chairs

Algorithms

Mikhail J. Atallah, Purdue University, USA

Applications

David A. Bader, Georgia Institute of Technology, USA

Architectures

Allan Gottlieb, New York University, USA

Software

Laxmikant Kale, University of Illinois, Urbana-Champaign, USA

Program Committee

Gagan AGRAWAL (Ohio State) USA
Ishfaq AHMAD (U. Texas/Arlington) USA
Shoukat ALI (U. Missouri/Rolla) USA
Nancy AMATO (Texas A&M) USA
Hagit ATTIYA (The Technion) Israel
Eduard AYGUADÉ (U. Politècnica de Catalunya) Spain
Olivier BEAUMONT (U. Bordeaux) France
Pete BECKMAN (Argonne National Lab.) USA
Michael BENDER (SUNY/Stony Brook) USA
Azzedine BOUKERCHE (U. Ottawa) Canada
David BROOKS (Harvard) USA
Doug BURGER (U. Texas/Austin) USA
Rajkumar BUYYA (U. Melbourne) Australia
Franck CAPPELLO (INRIA) France
Nikos CHRISOCHOIDES (Coll. of William & Mary) USA
Guojing CONG (IBM TJ Watson Research Center) USA
Toni CORTES (U. Politècnica de Catalunya) Spain
Karen DEVINE (Sandia National Lab.) USA
Guy EVEN (Tel-Aviv U.) Israel
Rainer FELDMANN (U. Paderborn) Germany
Wu-chun FENG (Los Alamos National Lab.) USA
John FEO (Cray Inc.) USA
Alfredo FERRO (U. Catania) Italy
Pierre FRAIGNIAUD (U. Paris-Sud) France
Mark A. FRANKLIN (Washington U.) USA
Efstratios GALLOPOULOS (U. Patras) Greece
Guang GAO (U. Delaware) USA
John GAROFALAKIS (U. Patras & CTI) Greece
James R. GOODMAN (U. Auckland) NZ
Ananth GRAMA (Purdue) USA
Manish GUPTA (IBM) USA
Attila GURSOY (Koc U.) Turkey
Kei HIRAKI (U. Tokyo) Japan
Hiroshi IMAI (U. Tokyo) Japan
Mary Jane IRWIN (Penn State) USA
Emmanuel JEANNOT (U. H. Poincaré Nancy-l & INRIA) France
Christos KAKLAMANIS (U. Patras) Greece
Vijay KARAMCHETI (NYU) USA
George KARYPIS (U. Minnesota) USA
Bradley C. KUSZMAUL (MIT) USA
Konrad LAI (Intel Labs) USA
Hyunyoung LEE (U. Denver) USA
Kai LI (Princeton) USA
Keqin LI (SUNY/New Paltz) USA
David LOWENTHAL (U. Georgia) USA
Paul LU (U. Alberta) Canada
Andrew LUMSDAINE (Indiana U.) USA
Xiaosong MA (North Carolina State & ORNL) USA
Grzegorz MALEWICZ (U. Alabama) USA
Allen D. MALONY (U. Oregon) USA
Bernard MANS (Macquarie Univ.) Australia
Satoshi MATSUOKA (Tokyo Inst. of Technology) Japan

Tim MATTSON (Intel) USA
Sally A. MCKEE (Cornell) USA
Celso MENDES (U. Illinois) USA
Ulrich MEYER (Max Planck Inst.) Germany
Russ MILLER (SUNY/Buffalo) USA
Michael O'BOYLE (U. Edinburgh) UK
Kunle OLUKOTUN (Stanford) USA
Marios PAPAEFTHYMIOU (U. Michigan) USA
Manish PARASHAR (Rutgers) USA
Kunsoo PARK (Seoul National U.) Korea
Yale N. PATT (U. Texas/Austin) USA
Fabrizio PETRINI (Pacific Northwest National Lab.) USA
Cynthia A. PHILLIPS (Sandia National Lab.) USA
Beth PLALE (Indiana U.) USA
C. Greg PLAXTON (U. Texas/Austin) USA
Sushil K. PRASAD (Georgia State) USA
Geppino PUCCI (U. Padova) Italy
Patrice QUINTON (U. Rennes) France
Sanguthevar RAJASEKARAN (U. Connecticut) USA
Partha RANGANATHAN (HP Labs) USA
Sanjay RANKA (U. Florida) USA
Lawrence RAUCHWERGER (Texas A&M) USA
Alexander REINEFELD (Zuse Inst.) Germany
Adi ROSÉN (The Technion) Israel
Larry RUDOLPH (MIT) USA
P. SADAYAPPAN (Ohio State) USA
Subhash SAINI (NASA Ames) USA
Pascal SAINRAT (U. Paul Sabatier) France
Christian SCHEIDELER (Johns Hopkins) USA
Christian SCHINDELHAUER (U. Paderborn) Germany
Karsten SCHWAN (Georgia Inst. Technology) USA
Peter SLOOT (U. Amsterdam) The Netherlands
Per STENSTROM (U. Chalmers) Sweden
Quentin F. STOUT (U. Michigan) USA
Michela TAUFER (U. Texas/El Paso) USA
Patricia J. TELLER (U. Texas/El Paso) USA
Denis TRYSTRAM (U. Grenoble) France
Dean TULLSEN (U. California/San Diego) USA
Mateo VALERO (U. Politècnica de Catalunya) Spain
Jeffrey VETTER (Oak Ridge National Lab.) USA
Uzi VISHKIN (U. Maryland) USA
Charles WEEMS (U. Massachusetts) USA
Uri WEISER (Intel) Israel
Li XIAO (Michigan State) USA
Zhiwei XU (Chinese Academy of Sciences) China
Yuanyuan YANG (SUNY/Stony Brook) USA
Albert Y. ZOMAYA (U. Sydney) Australia

IPDPS 2006 Technical Program

April 25-29, 2006 – Rhodes Island, Greece

Tuesday, April 25, 2006

Workshops 1-11

Wednesday, April 26, 2006

Keynote Speaker
Manish Gupta, IBM T.J. Watson Research Center
Massively Parallel Systems: Ready or Not, Here They Come

Rising power dissipation in microprocessor chips is driving computer architects towards a variety of solutions, all of which require exploiting greater degrees of parallelism. Hence, even though Moore's Law is alive, relying primarily on frequency scaling is no longer a viable path for meeting the growing computational needs of applications. There are several hurdles to exploiting greater levels of parallelism, such as, programming complexity, communication bottlenecks, interference from operating system services, and system management costs. We describe our experiences with the IBM Blue Gene project on pushing the limits of scalability in all aspects of system design. We will present our successes as well as outstanding challenges in programming and managing massively parallel systems. We will argue that we need another revolution in software to help achieve scientific breakthroughs and truly deliver on the promise offered by the next generation of high performance computing systems.

Plenary Session: Best Papers

Session 1: Scheduling

Session 2: P2P and Grid Computing, 1

Session 3: Memory Systems and Caches

Session 4: Consistency In Grids

Session 5: Hashing

Session 6: Parallel and Distributed Algorithms

Session 7: P2P and Grid Computing, 2

Session 8: Processor Designs

Thursday, April 27, 2006

Keynote Speaker
Yves Robert, LIP Laboratory - CNRS/ENS Lyon
Static Scheduling for Large-Scale Platforms: Can One Hope for Efficiency?

We discuss the potential and limitations of static scheduling techniques for heterogeneous clusters, grids, and large-scale decentralized platforms. We start with platform/application models and review several trade-offs between "tractability" and "accuracy". The traditional scheduling objective, namely, predicting and achieving optimal execution time (or, makespan), must be abandoned. However, we show how to approach this objective by using the power of divisible, steady-state, and flow-based approaches. For very large-scale platforms, a centralized scheduling mechanism is not realistic — but how can one even hope for decentralized yet provably efficient schedulers? We present sophisticated algorithmic approaches to this goal, illustrated by simple, yet significant applications, such as the problems of scheduling independent tasks and of implementing collective communications (e.g., broadcast, multicast, etc.).

Session 9: Load Balancing

Session 10: Computational Science: Biology, Chemistry, and Physics

Session 11: Performance Evaluation and Models

Session 12: Input/Output

Session 13: Scheduling, 2

Session 14: Data-Intensive Applications

Session 15: Energy Considerations

Session 16: Compilers and Optimization

Session 17: Memory Sharing

Session 18: Communication and Coordination

Session 19: Fault and Failure Tolerance

Session 20: MPI

Banquet

Invited Speaker: Bill Dally, Stanford University

Challenges and Opportunities for Parallel Computing

The next several years promise to be a golden age for parallel computing research. Parallel computing is becoming mainstream with even desktop computers having multiple processors. Key research problems stand in the way of efficiently using this emerging mainstream capability. Exploiting locality is a central problem. Bandwidth is the critical resource that dominates cost and performance of modern computing systems - arithmetic is almost free - and the cost of bandwidth increases greatly with distance. Exploiting locality requires an architecture that exposes data location at all levels – so it can be optimized by the programmer and compiler. Efficiently mapping to such an exposed-communication architecture requires new programming languages and compilation techniques. This talk will discuss the challenges and opportunities in parallel computing research with particular attention to the challenge of locality. Examples will be drawn from the Imagine and Merrimac stream processor projects.

Friday, April 28, 2006

Keynote Speaker

Horst D. Simon, Lawrence Berkeley National Laboratory

Progress in Supercomputing: The Top Three Breakthroughs of the Last 20 Years and the Top Three Challenges for the Next 20 Years

• As a community we have almost forgotten, what supercomputing was like twenty years ago in 1985. The state of the art system then was a 2 Gflop/s peak Cray-2, with at that time phenomenal 2 GBytes memory. It was the era of custom built vector mainframes, where anything beyond 100 Mflop/s sustained was considered excellent performance. The software environment was Fortran with vectorizing compilers (at best), and a proprietary operating system. There was hand tuning only, no tools, no visualization, and dumb terminals with remote batch. If one was lucky and had an account, remote access via 9600 baud was state-of-the-art. Usually a single code developer developed and coded everything from scratch.

• What a long way have we come in the last twenty years! Teraflop/s level performance on inexpensive, highly parallel commodity clusters, open source software, community codes, grid access via 10 Gbit/s, powerful visualization systems, and a productive development environment on a desktop system that is more powerful than the Cray-2 from 20 years ago – these are the characteristics of high performance computing in 2005.

• Of course a significant contribution to this progress is due to the continued increase of computing power following Moore's law. But what I want to argue here is, that progress was not just simply quantitative alone. We did not just get more of the same at a cheaper price. There were several powerful ideas and concepts that were shaped in the last 20 years, that made supercomputing the vibrant field that it is today. As an active researcher in the field for the last 25 years, I

will offer my subjective opinion, what were the real top breakthrough ideas that led to qualitative change and significant progress in our field.

• Retrospection leads to extrapolation: in the last part of the lecture, I will envision, what supercomputing will be like 20 years from now in the year 2025. Can we expect similar performance increases? How will supercomputing change qualitatively? And what are the top challenges that we will have to overcome to reach that vision of supercomputing in 2025?

Session 21: Routing

Session 22: Image Processing And Visualization

Session 23: Reconfigurable And Multiple-Width Systems

Session 24: Programming Abstractions

Session 25: Resource Allocation

Session 26: Partitioning And Refinement

Session 27: Collective Communication

Session 28: Distributed Coordination

Session 29: Symbolic Computing Applications

Session 30: Multithreading

Session 31: Runtime Optimizations

Session 32: Distributed Systems

Saturday, April 29, 2006

Workshops 12-19

IPDPS 2006 Reviewers

Tarek Abdelzaher
Manuel Acacio
Matthew Adiletta
Sanjeev Agarwal
Adnan Agbaria
Marco Morales Aguirre
Sayaka Akioka
Sadaf Alam
Enrique Alba
George Almasi
Jussara Almeida
Raed AlShaikh
Christoph Ambuehl
Patrick Amestoy
David Amzallag
Artur Andrzejak
Theodore A. Antonakopoulos
Gabriel Antoniu
Edoardo Aprá
Eric Aubanel
Enzo Auletta
Cevdet Aykanat
Pavan Balaji
Janaka Balasooriya
Roberto Baldoni
Kim Baldridge
Sudha Balla
Richard Barrett
Timothy Barth
Reuven Bar-Yehuda
Surender Baswana
Olivier Baudon
Andy Bavier
Micah Beck
Robert Belleman
Anne Benoit
Brahim Bensaou
Robert Bergeron
Vandy Berten
Alberto Bertoldo
Karan Bhatia
Mauro Bianco
Bryan Biegel
Gianfranco Bilardi
Vittorio Bilo
Joke Blom
Matt Blumrich
Vincent Boudet
Anu Bourgeois
Raouf Boutaba
Aurelien Bouteiller
Pascal Bouvry
Greg Bronevetsky
Marian Bubak
Marc Bui
Martin Burtscher
Kirk Cameron
Yves Caniou
Jiannong Cao
Ioannis Caragiannis
Eddy Caron
Francisco Cazorla
Brad Chamberlain
Roger Chamberlain
Sumir Chandra
Johnny Chang
Nicholas Chang
Sherry Chang
Claude Chaudet
Daniel Chavarria
Guilin Chen
Songqing Chen
John Chessa
Derek Chiou
Byung Kyu Choi
Anthony Chronopoulos
Yvonne Chu
Marcelo Cintra
Andrea Clementi
Johanne Cohen
Salvador Coll
Pham CongDuc
John Conner
Daniel Corlette
Patrick Crowley
Amitava Datta
Jaime Davila
Bronis R. de Supinski
Alexandre Denis
Jeremie Detrey
Akshaye Dhawan
Ned Dimitrov
Xiaoning Ding
Tai Do
Stefan Dobrev
Isaac Dooley
Gabor Dozsa
Nathalie Drach
Saurabh Drolia
Jose Duato
Michel Dubois
Philippe Duchon
Rocky Dunlap
Hans Eberle
Stefan Edelkamp
Rudolf Eigenmann
Benny Eitan
Magnus Ekman
Avshalom Elyada
Roee Engelberg
Luiz Angelo Estefanel
Lionel Eyraud
Gilles Fedak
Agustin Fernandez
Juan Fernandez Fernandez
Stefka Fidanova
Leslie Fife
Renato Figueiredo
Jose Flich
Jose Fortes
Vincent Freeh
Eric Freudenthal
Ron Gabor
Evghenii Gaburov
Clemente Galdi
Li Gao
Maria Garzaran
Georgi Gaydadjiev
Dominique Geniet
Amol Ghoting
Ran Ginosar
Filippo Gioachin
Andy Glew
Marc Gonzalez
Bernard Goossens
Kartik Gopalan
Rich Graham
Paul Gratz
Abdou Guermouche
Dimitrios Gunopulos
Lei Guo
Indranil Gupta
Jens Gustedt
Harshvardhan
Nicholas Harvey
David Hay
Danny Hendler
Thomas Herault
Amir Herzberg
Michael Hind

Mikael Hoegqvist
Alfons Hoekstra
Walter Hoffmann
Edward Hook
Chao Huang
Chun-Hsi Huang
Kevin Huck
Fabrice Huet
Jaehyuk Huh
Herbert Hum
Marty Humphrey
Luke Hunter
Felix Hupfeld
Sanjeliwala Huzefa
Adriana Iamnitchi
Jang-uk In
Dan Ionescu
David Irwin
Samir Jafar
Roozbeh Jafari
Sarangapani Jagannathan
G. John Janakiraman
Klaus Jansen
Song Jiang
Haoqiang Jin
Jose Joao
Mahmut Kandemir
Drona Kandhai
Panagiotis Kanellopoulos
Jaeyeon Kang
Kalapriya Kannan
Martin Karsten
Ronen Kat
Michael Kaufmann
Dan Keith
Andre Kerstens
Rajkumar Kettimuthu
Samee Khan
Changkyu Kim
Hyesoon Kim
Sangwook Kim
Iluju Kiringa
Gregory Koenig
Derrick Kondo
Charalampos Konstantopoulos
Ibrahim Korpeoglu
Annamaria Kovacs
Mehmet Koyuturk
Sriram Krishnamoorthy
Manojkumar Krishnan
Valeria Krzhizhanovskaya
Fabian Kuhn
Rakesh Kumar
Sameer Kumar
Yu-Kwong Kwok
Bjorn Landfeldt
Anders Landin
Isabelle Guérin Lassous
Edward Lee
Jack Y.B. Lee
Arnaud Legrand
Pierre Lemarinier
Andrew Lewis
Katarzyna (Kasia) Leyk
Li Li
Xiang-Yang Li
Yongfang Liang
Wei-keng Liao
Jyh-Ming Lien
Huaiyu Liu
Josep Llosa
Oleg Lodygensky
Charng-da Lu
Dong Lu
Jizhu Lu
Flaminia Luccio
Chris Lumb
Daniel Lynch
Olav Lysne
Nateri Madavan
Sanjay Kumar Madria
Muthucumaru Maheswaran
Konrad Malkowski
Joseph Manke
Loris Marchal
Andres Marquez
Angeles Martinez
Xavier Martorell
Christopher Marty
Mark Mathis
Domagoj Matijevic
Jean Mayo
Piyush Mehrotra
Dominique Mery
Brian Miller
Mark Miller
Geyong Min
Francois Modave
Justin Moore
Tomer Morad
Christine Morin
Alan Morris
Ioannis Mourtos
Steve Muir
Onur Mutlu
Ramdas Nagarajan
Farid Nait-abdesselem
Koji Nakano
Akihiro Nakao
Lata Naranayan
Sivaramakrishnan Narayanan
Satish Narayanasamy
Aroon Nataraj
Nicolas Navet
Vincent Neri
Dimitrios Nikolopoulos
Ronny Nitzsche
Marc Olano
Baudon Olivier
Ozcan Otzurk
Mohamed Ould-Khaoua
Oznur Ozkasap
Ozcan Ozturk
Bruce Palmer
Zhelong Pan
Jairo Panetta
Evi Papaioannou
Jehan-Francois Paris
Jin Park
Roger Pearce
Li-Shiuan Peh
David Peleg
Jack Perdue
Christian Perez
Erez Perlman
Pino Persiano
Frédéric Pétrot
Srdjan Petrovic
Seth Pettie
Congduc Pham
Andrea Pietracaprina
Stefan Plantikow
Arno Puder
Aaron Quigley
Martin Quinson
Moinuddin Qureshi
Rashedur Rahman
Karthick Rajamani
Sergio Rajsbaum
Murali Krishna Ramanathan
Govindarajan Ramaswamy
Lakshmish Ramaswamy
Nagarajan Ranganathan
Nitya Ranganathan
Subba Rao
Dror Rawitz
Shansi Ren
Yves Robert
Christine Rochange
Kenneth Roche
Sam Rodriguez
Thomas Roeblitz

Ronny Ronen
Philip Roth
Atanas Rountev
Alain Roy
Alexander Russell
Krzysztof Rzadca
Kunihiko Sadakane
Siti-Salwa Said
Suleyman Sair
Jafar Samir
Huzefa Sangeliwala
Karu Sankaralingam
Karthikeyan Sankaralingam
Jose-Renato Santos
Hamid Sarbazi-Azad
Erik Saule
Yucel Saygin
Michael Scarpa
Elad Schiller
Florian Schintke
Cristina Schmidt
Thorsten Schuett
Martin Schulz
Frank Seinstra
Selvakennedy Selvadurai
Franciszek Seredynski
Simha Sethumadhavan
André Seznec
Gad Shaeffer
Moni Shahar
Li Shang
Puneet Sharma
German Shegalov
Sameer Shende
Weisong Shi
Premkishore Shivakumar
H.J. Siegel
Gabriel Silberman
Anand Sivasubramaniam
Tor Skeie
Berend Smit
Aaron Smith
Jay Smith
Mingjun Song
Daniel Sorin
Carlos Sosa
Matthew Sottile
Neelam Soundarajan
Wojtek Spankowski
Wyatt Spear
Jeffrey Squyres
Nigamanth Sridhar
Santhosh Srinath
Ken Steele
Ivan Stojmenovic
Tom Stricker
Craig Stunkel
Aater Suleman
Yu Sun
Xian-He Sun
Raj Sunderraman
Dam Sunwoo
Alan Sussman
Frederic Suter
Peter Sweeney
Peter Szwed
Bosiljka Tadic
Yoav Talgam
Vanish Talwar
Xinyu Tang
Chunqiang Tang
Lydia Tapia
Eyad Taqieddin
Keita Teranishi
Vishal Thapar
Shawna Thomas
Vinod Tipparaju
Alfredo Tirado-Ramos
Mitul Tiwari
Sebastien Tixeuil
Olga Tkachyshyn
Corinne Touati
Meir Tsadik
Francis Tseng
Eric Tune
Pavel Tvrdik
Frank Vahid
Dick van-Albada
Aimé Vargas-Estrada
Sudharshan Vazhkudai
Alex Veidenbaum
Aline Viana
Anastasios Viglas
Frédéric Vivien
Frédéric Voisin
Frederic Wagner
Cong (James) Wang
Guiling Wang
Limin Wang
Xiaohui Wei
Jennifer Welch
Gordon Wilfong
Erling Wold
Rich Wolski
Tim Woodall
Jie Wu
Dawen Xie
Yuan Xie
Wanxia Xie
Qin Xin
Mi Yan
Wai-Gen Yee
Adi Yoaz
Hao Yu
Ting Yu
Xin Yuan
Maciej Zawodniok
Cheng Xiang Zhai
Lixin Zhang
Dayi Zhou
Jaroslaw Zola

Plenary Session

BEST PAPERS

On Collaborative Content Distribution using Multi-Message Gossip

Coby Fernandess[1] and Dahlia Malkhi[2]

[1] School of Engineering and Computer Science
The Hebrew University of Jerusalem
Jerusalem, Israel, Israel
fery@cs.huji.ac.il

[2] Microsoft Research
Silicon Valley Campus, CA, USA
dalia@microsoft.com

We study epidemic schemes in the context of collaborative data delivery. In this context, multiple chunks of data reside at different nodes, and the challenge is to simultaneously deliver all chunks to all nodes. Here we explore the inter-operation between the gossip of multiple, simultaneous message-chunks. In this setting, interacting nodes must select which chunk, among many, to exchange in every communication round.

We provide an efficient solution that possesses the inherent robustness and scalability of gossip. Our approach maintains the simplicity of gossip, and has low message, connections and computation overhead. Because our approach differs from solutions proposed by network coding, we are able to provide insight into the tradeoffs and analysis of the problem of collaborative content distribution. We formally analyze the performance of the algorithm, demonstrating its efficiency with high probability.

Assembling Genomes on Large-Scale Parallel Computers

Anantharaman Kalyanaraman[1], Scott J. Emrich[1], Patrick S. Schnable[2] and Srinivas Aluru[1]

[1] Department of Electrical and Computer Engineering
Iowa State University
Ames, IA, USA
{ananthk, semrich, aluru}@iastate.edu

[2] Departments of Agronomy, and Genetics, Development and Cell Biology
Iowa State University
Ames, IA, USA
schnable@iastate.edu

Assembly of large complex genomes from tens of millions of short genomic fragments is computationally demanding requiring hundreds of gigabytes of memory and tens of thousands of CPU hours. New gene-enrichment sequencing strategies are expected to further exacerbate this situation. In this paper, we present a massively parallel genome assembly framework. The unique features of our approach include space-efficient and on-demand algorithms that consume only linear space, and heuristic strategies that reduce the number of expensive pairwise sequence alignments while maintaining assembly quality. As part of the ongoing national efforts in maize genome sequencing, we applied our assembly framework to the largest available maize genomic data. We report the partitioning of more than 1.6 million fragments of over 1.25 billion nucleotides total size into genomic islands in 2 hours on 1,024 processors of an IBM BlueGene/L supercomputer.

Quantifying and Reducing the Effects of Wrong-Path Memory References in Cache-Coherent Multiprocessor Systems

Resit Sendag[1], Ayse Yilmazer[1], Joshua J. Yi[2] and Augustus K. Uht[1]

[1]*Electrical and Computer Engineering*
University of Rhode Island
Kingston, RI, USA
{*sendag, yilmazer, uht*}*@ele.uri.edu*

[2]*Networking and Computing Systems Group*
Freescale Semiconductor, Inc.
Austin, TX, USA
joshua.yi@freescale.com

High-performance multiprocessor systems built around out-of-order processors with aggressive branch predictors execute many memory references that turn out to be on a mispredicted branch path. Previous work that focused on uniprocessors showed that these wrong-path memory references may pollute the caches by bringing in data that are not needed on the correct execution path and by evicting useful data or instructions. Additionally, they may also increase the amount of cache and memory traffic. On the positive side, however, they may have a prefetching effect for memory references on the correct path. While computer architects have thoroughly studied the impact of wrong-path effects in uniprocessor systems, there is no previous work on its effects in multiprocessor systems. In this paper, we explore the effects of wrong-path memory references on the memory system behavior of shared-memory multiprocessor (SMP) systems for both broadcast and directory-based cache coherence. Our results show that these wrong-path memory references can increase the amount of cache-to-cache transfers by 32%, invalidations by 8% and 20% for broadcast and directory-based SMPs, respectively, and the number of writebacks by up to 67% for both systems. In addition to the extra coherence traffic, wrong-path memory references also increase the number of cache line state transitions by 21% and 32% for broadcast and directory-based SMPs, respectively. In order to reduce the performance impact of these wrong-path memory references, we introduce two simple mechanisms filtering wrong-path blocks that are not likely-to-be-used and wrong-path aware cache replacement that yield speedups of up to 37%.

Making Lockless Synchronization Fast: Performance Implications of Memory Reclamation

Thomas E. Hart[1], Paul E. Mckenney[2] and Angela Demke Brown[1]

[1]*Department of Computer Science*
University of Toronto
Toronto, Ontario, Canada
{*tomhart, demke*}*@cs.toronto.edu*

[2]*Linux Technology Center*
IBM Beaverton
Beaverton, Oregon, USA
paulmck@us.ibm.com

Achieving high performance for concurrent applications on modern multiprocessors remains challenging. Many programmers avoid locking to improve performance, while others replace locks with non-blocking synchronization to protect against deadlock, priority inversion, and convoying. In both cases, dynamic data structures that avoid locking, require a *memory reclamation scheme* that reclaims nodes once they are no longer in use.

The performance of existing memory reclamation schemes has not been thoroughly evaluated. We conduct the first fair and comprehensive comparison of three recent schemes—*quiescent-state-based reclamation*, *epoch-based reclamation*, and *hazard-pointer-based reclamation*—using a flexible microbenchmark. Our results show that there is no globally optimal scheme. When evaluating lockless synchronization, programmers and algorithm designers should thus carefully consider the data structure, the workload, and the execution environment, each of which can dramatically affect memory reclamation performance.

Session 1

SCHEDULING

Centralized Versus Distributed Schedulers for Multiple bag-of-task applications

Olivier Beaumont[1], Larry Carter[2], Jeanne Ferrante[2], Arnaud Legrand[3], Loris Marchal[4] and Yves Robert[4]

[1]*Laboratoire LaBRI*
CNRS-INRIA
Bordeaux, France
olivier.beaumont@labri.fr

[2]*Dept. of Computer Science and Engineering*
University of California
San Diego, CA, USA
{*carter, ferrante*}*@cs.ucsd.edu*

[3]*Laboratoire ID-IMAG*
CNRS-IN
Grenoble, France
arnaud.legrand@imag.fr

[4]*Laboratoire LIP*
CNRS-INRIA, École Normale Supérieure de Lyon
Lyon, France
{*loris.marchal, yves.robert*}*@ens-lyon.fr*

Multiple applications that execute concurrently on heterogeneous platforms compete for CPU and network resources. In this paper we consider the problem of scheduling applications to ensure fair and efficient execution on a distributed network of processors. We limit our study to the case where communication is restricted to a tree embedded in the network, and the applications consist of a large number of independent tasks that originate at the trees root. The tasks of a given application all have the same computation and communication requirements, but these requirements can vary for different applications. Each application is given a weight that quantifies its relative value. The goal of scheduling is to maximize throughput while executing tasks from each application in the same ratio as their weights.

We can find the optimal asymptotic rates by solving a linear program that expresses all necessary problem constraints, and we show how to construct a periodic schedule. For single-level trees, the solution is characterized by processing tasks with larger communication-to-computation ratios at children with larger bandwidths. For multi-level trees, this approach requires global knowledge of all application and platform parameters. For large-scale platforms, such global coordination by a centralized scheduler may be unrealistic. Thus, we also investigate decentralized schedulers that use only local information at each participating resource. We assess their performance via simulation, and compare to a centralized solution obtained via linear programming. The best of our decentralized heuristics achieves the same performance on about two-thirds of our test cases, but is far worse in a few cases. While our results are based on simplistic assumptions and do not explore all parameters (such as buffer size), they provide insight into the important question of fairly and optimally co-scheduling heterogeneous applications on heterogeneous grids.

A Strategy proof Mechanism for Scheduling Divisible Loads in Tree Networks

Thomas E. Carroll and Daniel Grosu

Department of Computer Science
Wayne State University
Detroit, MI, USA
{*tec, dgrosu*}*@cs.wayne.edu*

The underlying assumption of Divisible Load Scheduling is that the processors composing the network are obedient, *i.e.*, they do not "cheat" the algorithm. This assumption is unrealistic if the processors are owned by autonomous, self-interested organizations that have no *a priori* motivation for cooperation and they will manipulate the algorithm if it is beneficial to do so. In this paper we propose the strategy proof mechanism DLS-TL for scheduling divisible loads in tree networks. Our proposal augments Divisible Load Theory (DLT) with incentives such that it is beneficial for processors to report their true processing capacity and compute their assignments at full processing capacity. Additionally, incentives are provided for processors to report algorithm deviants. Deviants are penalized which abates the processors' willingness to deviate.

Real-Time Task Mapping and Scheduling for Collaborative In-Network Processing in DVS-Enabled Wireless Sensor Networks

Yuan Tian[1], Jarupan Boangoat[2], Eylem Ekici[3] and Fusun Ozguner[4]

[1]*Department of Electrical and Computer Engineering*
The Ohio State University
Columbus, Ohio, USA
tiany@ece.osu.edu

[2]*Department of Electrical and Computer Engineering*
The Ohio State University
Columbus, Ohio, USA
bea@ece.osu.edu

[3]*Department of Electrical and Computer Engineering*
The Ohio State University
Columbus, Ohio, USA
ekici@ece.osu.edu

[4]*Department of Electrical and Computer Engineering*
The Ohio State University
Columbus, Ohio, USA
ozguner@ece.osu.edu

With the increasing importance of energy consumption considerations and new requirements of emerging applications, in-network processing of information gains recognition as a viable solution for Wireless Sensor Networks (WSNs). The required processing capability can be achieved through locally collaborative information processing among sensors. Task mapping and scheduling plays an important role in efficient collaborative information processing. Although task mapping and scheduling in wired networks of processors has been well studied in the past, its counterpart for WSNs remains largely unexplored. In this paper, a task mapping and scheduling solution for real-time applications in WSNs, Real-time Task Mapping and Scheduling (RT-MapS), is presented. RT-MapS incorporates wireless channel modeling, Hyper-DAG extension, concurrent task mapping, communication and computation scheduling, and Dynamic Voltage Scaling (DVS) methods. Simulation results show significant performance improvements compared with existing mechanisms in terms of providing deadline guarantee with minimum energy consumption.

Flexible Tardiness Bounds for Sporadic Real-Time Task Systems on Multiprocessors

Umamaheswari Devi and James H. Anderson

Department of Computer Science
The University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599-3175
{uma, anderson}@cs.unc.edu

The earliest-deadline-first (EDF) scheduling of a sporadic real-time task system on a multiprocessor may require that the total utilization of the task system, U_{sum}, not exceed $(m+1)/2$ on m processors if every deadline needs to be met. In recent work, we considered the alleviation of this under-utilization for task systems that can tolerate deadline misses by bounded amounts (i.e., bounded tardiness). We showed that if $U_{sum} \leq m$ and tasks are not pinned to processors, then the tardiness of each task is bounded under both preemptive and non-preemptive EDF. However, the tardiness bounds derived are applicable to every task in the task system, i.e., any task may incur maximum tardiness. In this paper, we consider supporting tasks whose tolerances to tardiness are less than that known to be possible under EDF. We propose a new scheduling policy, called EDF-hl, that is a variant of EDF, and show that under EDF-hl, any tardiness, including zero tardiness, can be ensured for a limited number of *privileged* tasks, and that bounded tardiness can be guaranteed to the remaining tasks if their utilizations are restricted. EDFhl reduces to EDF in the absence of privileged tasks. The tardiness bound that we derive is a function of U_{sum}, in addition to individual task parameters. Hence, tardiness for all tasks can be lowered by lowering U_{sum}. A simulation-based evaluation of the tardiness bounds that are possible is provided.

Session 2

P2P and GRID COMPUTING, 1

Ad-hoc Distributed Spatial Joins on Mobile Devices

Panos Kalnis[1], Nikos Mamoulis[2], Spiridon Bakiras[3] and Xiaochen Li[1]

[1]Department of Computer Science
National University of Singapore
Singapore, Singapore
kalnis@comp.nus.edu.sg, g0202290@nus.edu.sg

[2]Department of Computer Science
The University of Hong Kong
Pokfulam, Hong Kong
nikos@cs.hku.hk

[3]Department of Computer Science
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
sbakiras@cs.ust.hk

PDAs, cellular phones and other mobile devices are now capable of supporting complex data manipulation operations. Here, we focus on ad-hoc spatial joins of datasets residing in multiple non-cooperative servers. Assuming that there is no mediator available, the spatial joins must be evaluated on the mobile device. Contrary to common applications that consider the cost at the server side, our main issue is the minimization of the transferred data, while meeting the resource constraints of the device. We show that existing methods, based on partitioning and pruning, are inadequate in many realistic situations. Then, we present novel algorithms that estimate the data distribution before deciding the physical operator independently for each partition. Our experiments with a prototype implementation on a WiFi-enabled PDA, suggest that the proposed methods outperform the competitors in terms of efficiency and applicability.

WaveGrid: a Scalable Fast-turnaround Heterogeneous Peer-based Desktop Grid System

Dayi Zhou and Virginia Lo

CIS Department
University of Oregon
Eugene, OR, USA
{dayizhou, lo}@cs.uoregon.edu

We propose a novel heterogeneous scalable desktop grid system, WaveGrid, which uses a peer-to-peer architecture and can satisfy the needs of applications with fast-turnaround requirements.

The challenges for fast-turnaround scheduling in a large heterogeneous peer-based desktop grid system include how to quickly discover available hosts with low message overhead; how to achieve high utilization of the available cycles in this opportunistic scheduling environment; and how to adapt to the heterogeneous environment for efficient scheduling. WaveGrid answers these challenges by letting hosts self-organize into a timezone-aware overlay network, which supports straightforward, quick resource discovery. Scheduling methods in WaveGrid take heterogeneity into account in selecting scheduling and migration targets. WaveGrid then rides the wave of available cycles by migrating jobs to hosts located in idle night-time zones around the globe.

We evaluate WaveGrid using a heterogeneous host CPU power profile based on empirical data collected from the global computing project BOINC. The simulation results show that WaveGrid performs consistently well with fast turnaround time and low migration overhead. It performs much better than other systems with respect to turnaround, stability and minimal impact on hosts.

Trust Overlay Networks for Global Reputation Aggregation in P2P Grid Computing

Runfang Zhou and Kai Hwang

Computer Science
University of Southern California
Los Angeles, California, USA
{*rzhou, kaihwang*}*@usc.edu*

This paper presents a new approach to trusted Grid computing in a Peer-to-Peer (P2P) setting. Trust and security are essential to establish lasting working relationships among the peers. A P2P reputation system collects peer trust scores and aggregates them to yield a global reputation. We use a new trust overlay network (TON) to model the trust relationships among the peers. After analyzing the eBay transaction trace data, we discover a power-law distribution in user feedbacks. We develop a new reputation system, PowerTrust, to leverage power-law feedback characteristics.

The PowerTrust system is built with locality-preserving hash functions and a lookahead random walk strategy. Dynamic system reconfiguration is enabled by the use of power nodes with well-established reputations. Through P2P simulation experiments on distributed file sharing and Grid parameter-sweeping applications (PSA), we demonstrate the PowerTrust advantages in fast reputation convergence and accurate ranking of peer reputations. We report performance results with enhanced query success rate and shortened job makespan in scalable P2P Grid applications.

An Adaptive Stabilization Framework for Distributed Hash Tables

Gabriel Ghinita[1] and Yong Meng Teo[2]

[1]*Department of Computer Science*
National University of Singapore
3 Science Drive 2, Singapore 117543
ghinitag@comp.nus.edu.sg

[2]*Department of Computer Science*
National University of Singapore
3 Science Drive 2, Singapore 117543
teoym@comp.nus.edu.sg

Distributed Hash Tables (DHT) algorithms obtain good lookup performance bounds by using deterministic rules to organize peer nodes into an overlay network. To preserve the invariants of the overlay network, DHTs use stabilization procedures that reorganize the topology graph when participating nodes join or fail. Most DHTs use periodic stabilization, in which peers perform stabilization at fixed intervals of time, disregarding the rate of change in overlay topology; this may lead to poor performance and large stabilization-induced communication overhead. We propose a novel adaptive stabilization framework that takes into consideration the continuous evolution in network conditions. Each peer collects statistical data about the network and dynamically adjusts its stabilization rate based on the analysis of the data. The objective of our scheme is to maintain nominal network performance and to minimize the communication overhead of stabilization.

Session 3

MEMORY SYSTEMS and CACHES

Enhancing L2 Organization for CMPs with a Center Cell

Chun Liu, Anand Sivasubramaniam, Mahmut Kandemir and Mary Jane Irwin

Computer Science and Engineering
Pennsylvania State Univerity
University Park, PA, USA
{*chliu, anand, kandemir, mji*}*@cse.psu.edu*

Chip multiprocessors (CMPs) are becoming a popular way of exploiting ever-increasing number of on-chip transistors. At the same time, the location of data on the chip can play a critical role in the performance of these CMPs because of the growing on-chip storage capacities and the relative cost of wire delays. It is important to locate the data at the right place at the right time in the on-chip cache hierarchy. This paper presents a novel L2 cache organization for CMPs with these goals in mind.

We first study the data sharing characteristics of a wide spectrum of multi-threaded applications and show that, while there are a considerable number of L2 accesses to shared data, the volume of this data is relatively low. Consequently, it is important to keep this shared data fairly close to all processor cores for both performance and power reasons. Motivated by this observation, we propose a small Center Cell cache residing in the middle of the processor cores which provides fast access to its contents. We demonstrate that this cache organization can considerably lower the number of block migrations between the L2 portions that are closer to each core, thus providing better performance and power.

Improving Cache Locality for Thread-Level Speculation

Stanley L. C. Fung and J. Gregory Steffan

Electrical and Computer Engineering
University of Toronto
Toronto, ON, Canada
{*sfung, steffan*}*@eecg.toronto.edu*

With the advent of chip-multiprocessors (CMPs), *Thread-Level Speculation* (TLS) remains a promising technique for exploiting this highly multithreaded hardware to improve the performance of an individual program. However, with such speculatively-parallel execution the cache locality once enjoyed by the original uniprocessor execution is significantly disrupted: for TLS execution on a four-processor CMP, we find that the data-cache miss rates are nearly four-times those of the uniprocessor case, even though TLS execution utilizes four private data caches (i.e., four-fold greater cache capacity).

We break down the TLS cache locality problem into instruction and data cache, execution stages, and parallel access patterns, and propose methods to improve cache locality in each of these areas. We find that for parallel regions across 13 SPECint applications our simple and low-cost techniques reduce data-cache misses by 38%, improve performance by 12.8%, and significantly improve scalability—further enhancing the feasibility of TLS as a way to capitalize on future CMPs.

On the Effectiveness of Speculative and Selective Memory Fences

Oliver Trachsel, Christoph Von Praun and Thomas R. Gross

Department of Computer Science
ETH Zurich
Zurich, Switzerland
{*trachsel, trg*}*@inf.ethz.ch, praun@acm.org*

Memory fences inhibit the reordering of memory accesses in modern microprocessors; fences are useful to implement synchronization and strong shared memory semantics in multi-threaded programs. A naive implementation of memory fences can result in a significant performance penalty for processors with deep pipelines supporting multiple concurrent memory accesses.

The paper compares three techniques to reduce the impact of memory fences: (1) Read-speculation allows reads that follow a fence to be issued while the fence is being processed; (2) Write-ahead additionally allows writes following a fence to proceed early; (3) Selective fences distinguish between memory accesses to thread-local and shared memory and enforce ordering only among accesses to shared memory.

We evaluate and compare the effectiveness of these techniques with a simulator derived from the Pentium 4 architecture. We report data for a storage model that uses memory fences to enforce the memory semantics at monitor boundaries.

Exploiting Locality: A Flexible DSM Approach

Håkan Zeffer, Zoran Radovic and Erik Hagersten

Department of Information Technology
Uppsala University
Uppsala, Sweden
{*hakan.zeffer, zoran.radovic, erik.hagersten*}*@it.uu.se*

No single coherence strategy suits all applications well. Many promising adaptive protocols and coherence predictors, capable of dynamically modifying the coherence strategy, have been suggested over the years.

While most dynamic detection schemes rely on plentiful of dedicated hardware, the customization technique suggested in this paper requires no extra hardware support for its per-application coherence strategy. Instead, each application is profiled using a low-overhead profiling tool. The appropriate coherence flag setting, suggested by the profiling, is specified when the application is launched.

We have compared the performance of a hardware DSM (Sun WildFire) to a software DSM built with identical interconnect hardware and coherence strategy. With no support for flexibility, the software DSM runs on average 45 percent slower than the hardware DSM on the 12 studied applications, while the flexibility can get the software DSM within 11 percent. Our all-software system outperforms the hardware DSM on four applications.

Session 4

CONSISTENCY in GRIDS

On Consistency Maintenance In Service Discovery

Vasughi Sundramoorthy, Pieter Hartel and Hans Scholten

Department of Computer Science
University of Twente
Enschede, The Netherlands
{*vasughi.sundramoorthy, pieter.hartel, hans.scholten*}*@utwente.nl*

Communication and node failures degrade the ability of a service discovery protocol to ensure Users receive the correct service information when the service changes. We propose that service discovery protocols employ a set of recovery techniques to recover from failures and regain consistency. We use simulations to show that the type of recovery technique a protocol uses significantly impacts the performance. We benchmark the performance of our own service discovery protocol, FRODO against the performance of first generation service discovery protocols, Jini and UPnP during increasing communication and node failures. The results show that FRODO has the best overall consistency maintenance performance.

Evaluation of UDDI as a Provider of Resource Discovery Services for OGSA-based Grids

Edward Benson, Glenn Wasson and Marty Humphrey

Computer Science Department
University of Virginia
Charlottesville, VA, USA
{*mah2h, wasson*}*@virginia.edu, humprey@cs.virginia.edu*

Grid computing involves networks of heterogeneous resources working in collaboration to solve problems that cannot be addressed by the resources of any one organization. A pervasive problem for Grid users is how best to discover the resources they need given dynamic Grid environments. UDDI, the Universal Description, Discovery and Integration framework, is an OASIS standard for publishing and querying discovery information for Web services, which to date, has received surprisingly little analysis as a discovery mechanism for Web service-based Grids, e.g. those based on the Open Grid Services Architecture (OGSA). This work identifies issues that must be addressed in order to make UDDI meet the requirements of OGSA discovery. We examine the performance implications of these issues using a freely available implementation of UDDI version 2. Based on our experimental results, we conclude that UDDI can be used for OGSA discovery, but the cost may be prohibitive for large Grids.

Monitoring Remotely Executing Shared Memory Programs in Software DSMs

Long Fei, Xing Fang, Y. Charlie Hu and Samuel P. Midkiff

School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN, USA
{*lfei, xfang, ychu, smidkiff*}*@purdue.edu*

Peer-to-Peer (P2P) cycle sharing over the Internet has become increasingly popular as a way to share idle cycles. A fundamental problem faced by P2P cycle sharing systems is how to incrementally monitor and verify, with low overhead, the execution of jobs submitted to a remote untrusted hosting machine, or cluster of machines. In this paper, we present the design and implementation of GridCop DSM, a novel incremental execution monitoring and verification scheme for software distributed shared memory (SDSM) programs running on remote clusters. Our scheme maximally leverages the shared memory abstraction provided by the SDSM system by extending the shared memory abstraction to the monitoring process by replicating one of the processes running on the host cluster to verify intermediate results at runtime. Our GridCop DSM employs two monitoring schemes: (i) a full-scale monitoring scheme that completely replicates the computation of a process running on the cluster, and (ii) a decoy monitoring scheme that deceives the host cluster into believing that full-scale monitoring is being performed without it ever actually being done, thereby incurring negligible overhead. Experiments show that the combined use of full-scale and decoy monitoring ensures faithful execution with low performance impact, even over a wide area network.

A Segment-Based DSM Supporting Large Shared Object Space

Benny Wang-leung Cheung and Cho-li Wang

Department of Computer Science
The University of Hong Kong
Hong Kong, Hong Kong
{*wlcheung, clwang*}*@cs.hku.hk*

This paper introduces a software DSM that can extend its shared object space exceeding 4GB in a 32-bit commodity cluster environment. This is achieved through the dynamic memory mapping mechanism, with local hard disks as backing store. We introduce the new concept of segments with intelligent splitting to reduce network traffic, false sharing as well as adapt better to the shared memory access patterns. A priority-based swapping algorithm is designed to reduce disk accesses for efficient dynamic memory mapping, and maximize the use of disk space as shared object space. A new queue-based scheme is also devised for efficient and simple management of memory blocks. The proposed solutions were implemented in LOTS V.2, and it can outperform its previous version when running small applications, while the maximum shared object space is increased to one-third of the total free disk space available among all the nodes.

Session 5

HASHING

D1HT: A Distributed One Hop Hash Table

Luiz Rodolpho Monnerat[1,2] and Claudio Luis De Amorim[2]

[1] *TI/TI-E&P/STEP*
PETROBRAS
Rio de Janeiro, RJ, Brazil
monnerat@cos.ufrj.br

[2] *COPPE - Computer and Systems Engineering*
Federal University of Rio de Janeiro
Rio de Janeiro, RJ, Brazil
amorim@cos.ufrj.br

Distributed Hash Tables (DHTs) have been used in a variety of applications, but most DHTs so far have opted to solve lookups with multiple hops, which sacrifices performance in order to keep little routing information and minimize maintenance traffic. In this paper, we introduce D1HT, a novel single hop DHT that is able to maximize performance with reasonable maintenance traffic overhead even for huge and dynamic peer-to-peer (P2P) systems. We formally define the algorithm we propose to detect and notify any membership change in the system, prove its correctness and performance properties, and present a Quarantine-like mechanism to reduce the overhead caused by volatile peers. Our analyses show that D1HT has reasonable maintenance bandwidth requirements even for very large systems, while presenting at least twice less bandwidth overhead than previous single hop DHT.

Hash-based Proximity Clustering for Load Balancing in Heterogeneous DHT Networks

Haiying Shen and Cheng-zhong Xu

Department of Electrical and Computer Engineering
Wayne State University
Detroit, MI, USA
{shy, czxu}@wayne.edu

DHT networks based on consistent hashing functions have an inherent load uneven distribution problem. The objective of DHT load balancing is to balance the workload of the network nodes in proportion to their capacity so as to eliminate traffic bottleneck. It is challenging because of the dynamism nature of DHT networks and time-varying load characteristics.

In this paper, we present a hash-based proximity clustering approach for load balancing in heterogeneity DHTs. In the approach, DHT nodes are classified as regular nodes and supernodes according to their computing and networking capacities. Regular nodes are grouped and associated with supernodes via consistent hashing of their physical proximity information on the Internet. The supernodes form a self-organized and churn resilient auxiliary network for load balancing. The hierarchical structure facilitates the design and implementation of a locality-aware randomized load balancing algorithm. The algorithm introduces a factor of randomness in the load balancing processes in a range of neighborhood so as to deal with both the proximity and dynamism. Simulation results show the superiority of the approach, in comparison with a number of other DHT load balancing algorithms. The approach performs no worse than existing proximity-aware algorithms and exhibits strong resilience to the effect of churn. It also greatly reduces the overhead of resilient randomized load balancing algorithms due to the use of proximity information.

DiST: Fully Decentralized Indexing for Querying Distributed Multidimensional Datasets

Beomseok Nam and Alan Sussman

UMIACS and Computer Science
University of Maryland
College Park, MD, 20742
{*bsnam, als*}*@cs.umd.edu*

Grid computing and Peer-to-peer (P2P) systems are emerging as new paradigms for managing large scale distributed resources across wide area networks. While Grid computing focuses on managing heterogeneous resources and relies on centralized managers for resource and data discovery, P2P systems target scalable, decentralized methods for publishing and searching for data. In large distributed systems, a centralized resource manager is a potential performance bottleneck and decentralization can help avoid this bottleneck, as is done in P2P systems. However, the query functionality provided by most existing P2P systems is very rudimentary, and is not directly applicable to Grid resource management. In this paper, we propose a fully decentralized multidimensional indexing structure, called *DiST*, that operates in a fully distributed environment with no centralized control. In DiST, each data server only acquires information about data on other servers from executing and routing queries. We describe the DiST algorithms for maintaining the decentralized network of data servers, including adding and deleting servers, the query routing algorithm, and failure recovery algorithms. We also evaluate the performance of the decentralized scheme against a more structured hierarchical indexing scheme that we have previously shown to perform well in distributed Grid environments.

Session 6

PARALLEL and DISTRIBUTED ALGORITHMS

Distributed Coloring in $\tilde{O}(\sqrt{\log n})$ Bit Rounds

Kishore Kothapalli[1], Melih Onus[2], Christian Scheideler[3] and Christian Schindelhauer[4]

[1]*Department of Computer Science*
Johns Hopkins University
Baltimore, Maryland, U.S.A.
kishore@cs.jhu.edu

[2]*Department of Computer Science*
Arizona State University
Tempe, Arizona, U.S.A.
Melih.Onus@asu.edu

[3]*Computer Science Department*
Technische Universität München
Garching, Germany
scheideler@in.tum.de

[4]*Computer Science Department*
University of Paderborn
Paderborn, Germany
schindel@upb.de

We consider the well-known vertex coloring problem: given a graph G, find a coloring of the vertices so that no two neighbors in G have the same color. Distributed algorithms that find a $(\Delta + 1)$-coloring in a logarithmic number of communication rounds, with high probability (w.h.p), are known since more than a decade. But what if the edges have orientations, i.e., the endpoints of an edge agree on its orientation? Interestingly, for the cycle in which all edges have the same orientation, we show that a simple randomized algorithm can achieve a 3-coloring with only $O(\sqrt{\log n})$ rounds of bit transmissions w.h.p. This result is tight because we also show that the bit complexity of coloring an oriented cycle is $\Omega(\sqrt{\log n})$, w.h.p., no matter how many colors are allowed. The 3-coloring algorithm can be easily extended to provide a $(\Delta + 1)$-coloring for oriented graphs of maximum degree Δ in $O(\sqrt{\log n})$ rounds of bit transmissions, w.h.p., if Δ is a constant, and the graph does not contain an oriented cycle of length less than $\sqrt{\log n}$. Using more complex algorithms, we show how to obtain an $O(\Delta)$-coloring for arbitrary oriented graphs with maximum degree Δ, and with no oriented cycles of length at most $\sqrt{\log n}$, using essentially $O(\log \Delta + \sqrt{\log n})$ rounds of bit transmissions.

Distributed Algorithm for a Color Assignment on Asynchronous Rings

Gianluca De Marco[1], Mauro Leoncini[2,3] and Manuela Montangero[2,3]

[1]*Dipartimento di Informatica e Applicazioni*
Università di Salerno
Baronissi (SA), Italy
demarco@dia.unisa.it

[2]*Dipartimento di Ingegneria dell'Informazione*
Università di Modena e Reggio Emilia
Modena, Italy
leoncini@acm.org, montangero.manuela@unimo.it

[3]*Istituto di Informatica e Telematica - CNR*
Pisa, Italy

We study a version of the β-assignment problem (introduced by G. J. Chang and P. H. Ho in 1998) on asynchronous rings: consider a set of items and a set of m colors, where each item is associated to one color. Consider also n computational agents connected by an asynchronous ring. Each agent holds a subset of the items, where initially different agents might hold items associated to the same color. We analyze the problem of distributively assigning colors to agents in such a way that(a) each color is assigned to one agent and (b) the number of different colors assigned to each agent is minimum. Since any color assignment requires that the items be distributed according to it (*e.g.* all items of the same color are to be held by only one agent), we define the cost of a color assignment as the amount of items that need to be moved, given an initial allocation. We first show that any distributed algorithm for this problem on the ring requires a communication complexity of $\Omega(n \cdot m)$ and then we exhibit a polynomial time distributed algorithm with message complexity matching the bound, that determines a color assignment with cost at most $(2 + \epsilon)$ times the optimal cost, for any $0 < \epsilon < 1$.

On the Packing of Selfish Items

Vittorio Bilò

Department of Mathematics
University of Lecce
Lecce, Italy
vittorio.bilo@unile.it

In the non cooperative version of the classical Minimum Bin Packing problem, an item is charged a cost according to the percentage of the used bin space it requires. We study the game induced by the selfish behavior of the items which are interested in being packed in one of the bins so as to minimize their cost. We prove that such a game always converges to a pure Nash equilibrium starting from any initial packing of the items, estimate the number of steps needed to reach one such equilibrium, prove the hardness of computing good equilibria and give an upper and a lower bound for the price of anarchy of the game. Then, we consider a multidimensional extension of the problem in which each item can require to be packed in more than just one bin. Unfortunately, we show that in such a case the induced game may not admit a pure Nash equilibrium even under particular restrictions. The study of these games finds applications in the analysis of the bandwidth cost sharing problem in non cooperative networks.

GPU-ABiSort: Optimal Parallel Sorting on Stream Architectures

Alexander Greß[1] and Gabriel Zachmann[2]

[1]*Institute of Computer Science II*
Rhein. Friedr.-Wilh.-Universität Bonn
Bonn, Germany
gress@cs.uni-bonn.de

[2]*Institute of Computer Science*
Clausthal University of Technology
Clausthal, Germany
zach@in.tu-clausthal.de

In this paper, we present a novel approach for parallel sorting on stream processing architectures. It is based on adaptive bitonic sorting. For sorting n values utilizing p stream processor units, this approach achieves the optimal time complexity $O((n \log n)/p)$.

While this makes our approach competitive with common sequential sorting algorithms not only from a theoretical viewpoint, it is also very fast from a practical viewpoint. This is achieved by using efficient linear stream memory accesses (and by combining the optimal time approach with algorithms optimized for small input sequences).

We present an implementation on modern programmable graphics hardware (GPUs). On recent GPUs, our optimal parallel sorting approach has shown to be remarkably faster than sequential sorting on the CPU, and it is also faster than previous non-optimal sorting approaches on the GPU for sufficiently large input sequences. Because of the excellent scalability of our algorithm with the number of stream processor units p (up to $n/\log^2 n$ or even $n/\log n$ units, depending on the stream architecture), our approach profits heavily from the trend of increasing number of fragment processor units on GPUs, so that we can expect further speed improvement with upcoming GPU generations.

Session 7

P2P and GRID COMPUTING, 2

An Authentication Protocol in Web-computing

Siman Wong

Dept. of Mathematics & Statistics
University of Massachusetts
Amherst, MA, USA
siman@math.umass.edu

A web-computing system (WCS) allows a host with limited resources to perform CPU intensive tasks by outsourcing the computations to external clients. But not every client is trusted, and redundancy in task assignment and auditing of results are needed to ensure the integrity of the results. This raises the question as to the efficiency and reliability of the system as measured against a given unit of the host's auditing time or cost. In this paper we propose a WCS with low overhead and has favorable error rate compared to a majority-voting scheme with similar efficiency. We can reduce the error rate by re-authenticating the results without having to resubmit any jobs, and we have an auditing strategy that in many cases is probabilistically better than random sampling.

A Design of Overlay Anonymous Multicast Protocol

Li Xiao[1], Xiaomei Liu[1], Wenjun Gu[2], Dong Xuan[2] and Yunhao Liu[3]

[1]*Department of Computer Science and Engineering*
Michigan State University
East Lansing, MI, USA
{lxiao, liuxiaom}@cse.msu.edu

[2]*Department of Computer Science and Engineering*
Ohio State University
Columbus, OH, USA
{gu, xuan}@cse.ohio-state.edu

[3]*Department of Computer Science*
Hong Kong University of Science and Technology
Kowloon, Hong Kong
liu@cs.ust.hk

Multicast services are demanded by a variety of applications. Many applications require anonymity during their communication. However, there has been very little work on anonymous multicasting and such services are not available yet. Since there are fundamental differences between multicast and unicast, the solutions proposed for anonymity in unicast communications cannot be directly applied to multicast applications. In this paper we define the anonymous multicast system, and propose a mutual anonymous multicast (MAM) protocol including the design of a unicast mutual anonymity protocol and construction and optimization of an anonymous multicast tree. MAM is self organizing and completely distributed. We define the attack model in an anonymous multicast system and analyze the anonymity degree. We also evaluate the performance of MAM by simulations.

IP over P2P: Enabling Self-configuring Virtual IP Networks for Grid Computing

Arijit Ganguly, Abhishek Agrawal, P. Oscar Boykin and Renato Figueiredo

Advanced Computing and Information Systems Lab
University of Florida
Gainesville, Florida, USA
{*aganguly, aagrawal, boykin, renato*}*@acis.ufl.edu*

Peer-to-peer (P2P) networks have mostly focused on task oriented networking, where networks are constructed for single applications, i.e. file-sharing, DNS caching, etc. In this work, we introduce IPOP, a system for creating virtual IP networks on top of a P2P overlay. IPOP enables seamless access to Grid resources spanning multiple domains by aggregating them into a virtual IP network that is completely isolated from the physical network. The virtual IP network provided by IPOP supports deployment of existing IP-based protocols over a robust, self-configuring P2P overlay. We present implementation details as well as experimental measurement results taken from LAN, WAN, and Planet-Lab tests.

Efficient Client-to-Server Assignments for Distributed Virtual Environments

Duong Nguyen Binh Ta and Suiping Zhou

School of Computer Engineering
Nanyang Technological University
Singapore, Singapore 639798
{*pa0236892b, asspzhou*}*@ntu.edu.sg*

Distributed Virtual Environments (DVEs) are distributed systems that allow multiple geographically distributed clients (users) to interact simultaneously in a computer-generated, shared virtual world. Applications of DVEs can be seen in many areas nowadays, such as online games, military simulations, collaborative designs, etc. To support large-scale DVEs with real-time interactions among thousands or more distributed clients, a geographically distributed server architecture (GDSA) is generally needed, and the virtual world can be partitioned into many distinct zones to distribute the load among the servers. Due to the geographic distributions of clients and servers in such architectures, it is essential to efficiently assign the participating clients to servers to enhance users' experience in interacting within the DVE. This problem is termed the client assignment problem. In this paper, we propose a two-phase approach, consisting of an initial assignment phase and a refined assignment phase to address this problem. Both phases are shown to be NP-hard, and several heuristic assignment algorithms are then devised based on this two-phase approach. Via extensive simulation studies with realistic settings, we evaluate these algorithms in terms of their performances in enhancing interactivity of the DVE.

Session 8

PROCESSOR DESIGNS

Exploiting Dataflow to Extract Java Instruction Level Parallelism on a Tag-based Multi-Issue Semi In-Order (TMSI) Processor

Hai-chen Wang and Chung-kwong Yuen

Dept. of Computer Science, School of Computing
National University of Singapore
Singapore, Singapore
{*wanghaci, yuenck*}*@comp.nus.edu.sg*

To design a Java processor with traditional modern processor architecture, the Instruction Level Parallelism (ILP) is not readily exploitable due to stack operands dependencies. This paper presents a dataflow-based instruction tagging scheme. With instruction tagging, the independent bytecode instruction groups with stack dependences are identified. Because there is no stack dependence among the different bytecode instruction groups, they can be executed in parallel. With the instruction tagging scheme, we propose a tag-based multi-issue semi-in-order (TMSI) Java processor. The processor takes advantage of instruction-tagging and stack-folding to generate the tagged register-based instructions. When the tagged instructions are ready, they are bundled out-of-order depending on data availability to form VLIW-like instruction words and issued in-order. To achieve high performance, a VLIW engine is employed. We have done the experiments in our TMSI simulation environment using SPECjvm98 and Linpack workload. The results indicate that the proposed processor has the good performance gain.

SAMIE-LSQ: Set-Associative Multiple-Instruction Entry Load/Store Queue

Jaume Abella[1,2] and Antonio González[1,2]

[1]*Intel Barcelona Research Center*
Intel & Universitat Politècnica de Catalunya
Barcelona, Spain
{*jaumex.abella, antonio.gonzalez*}*@intel.com*

[2]*Departament Arquitectura de Computadors*
Universitat Politècnica de Catalunya
Barcelona, Spain

The load/store queue (LSQ) is one of the most complex parts of contemporary processors. Its latency is critical for the processor performance and it is usually one of the processor hotspots.

This paper presents a highly banked, set-associative, multiple-instruction entry LSQ (SAMIE-LSQ) that achieves high performance with small energy requirements. Our approach relies on the fact that many in-flight memory instructions access the same cache lines. The SAMIE-LSQ groups those instructions accessing the same cache line in the same entry. This arrangement has a number of advantages. First, it significantly reduces the address comparison activity needed for memory disambiguation since there are less addresses to be compared. It also reduces the activity in the data TLB, the cache tag and cache data arrays by caching the cache line location and address translation in the corresponding SAMIE-LSQ entry. Hence, instructions in the same entry can reuse the translation, avoid the tag check and obtain the data accessing only the right cache way. Besides, the delay of the proposed scheme is lower than that required by a conventional LSQ.

We show that the SAMIE-LSQ saves 82% dynamic energy for the load/store queue, 42% for the L1 data cache and 73% for the data TLB, with a negligible impact on performance (0.6%).

Compiler Assisted Dynamic Management of Registers for Network Processors

Ryan Collins, Fernando Alegre, Xiaotong Zhuang and Santosh Pande

College of Computing
Georgia Institute of Technology
Atlanta, Georgia, USA
{*rcollins, fernando, xt2000, santosh*}*@cc.gatech.edu*

Modern network processors support high levels of parallelism in packet processing by supporting multiple threads that execute on a micro-engine. Threads switch context upon encountering long latency memory accesses and this way the parallelism and memory access can be overlapped. Context switches in the typical network processor architectures such as the IXP are designed to be very fast. However, the low overhead is partly achieved by leaving register management to programs, with minimal support from the hardware. The complexity of the multi-engine, multi-threaded environment makes manual register management a daunting task, which is better left to a compiler. However, a purely static analysis is unable to achieve full utilization of the register file due to conservative estimates of liveness. A register that is live across a context switch point must be considered live for the duration of all other threads, and so it must be assumed to be unavailable to other threads. In addition, aliasing further reduces the effectiveness of static analysis. The net effect is a large number of idle cycles that are still present after static optimization.

We propose a dynamic solution that requires minimal software and hardware support. On the software side, we take a pre-allocated binary file and annotate the potential context switch instructions with information about the dead registers. On the hardware side, we try to rename the transfer registers and addresses to dead general purpose registers and update the usage of registers. We then replace the long-latency memory instructions with fast move instructions in the architecture using the dynamic context. The results show up to 51% reduction in idle cycles and up to 14% increase in the throughput for hand coded applications on Intel IXP 1200 network processor.

Session 9

LOAD BALANCING

A New Analytical Method for Parallel, Diffusion-type Load Balancing

Petra Berenbrink[1], Tom Friedetzky[2] and Zengjian Hu[3]

[1] *School of Computing Science*
Simon Fraser University
Burnaby, BC, Canada
petra@cs.sfu.ca

[2] *Department of Computer Science*
Durham University
Durham, England, UK
tom.friedetzky@dur.ac.uk

[3] *School of Computing Science*
Simon Fraser University
Burnaby, BC, Canada
zhu@cs.sfu.ca

We propose a new proof technique which can be used to analyze many parallel load balancing algorithms. The technique is designed to handle concurrent load balancing actions, which are often the main obstacle in the analysis. We demonstrate the usefulness of the approach by analyzing various natural diffusion-type protocols. Our results are similar to, or better than, previously existing ones, while our proofs are much easier.

The key idea is to first sequentialize the original, concurrent load transfers, analyze this new, sequential system, and then to bound the gap between both.

Load Balancing in the Presence of Random Node Failure and Recovery

Sagar Dhakal[1], Majeed M. Hayat[1], Jorge E. Pezoa[1], Chaouki T. Abdallah[1], J. Doug Birdwell[2] and John Chiasson[2]

[1] *Dept. of Electrical and Computer Engineering*
University of New Mexico
Albuquerque, NM 87131-0001, USA
{dhakal, hayat, jpezoa, chaouki}@ece.unm.edu

[2] *Dept. of Electrical and Computer Engineering*
University of Tennessee
Knoxville, TN 37996-2100, USA
{birdwell, chiasson}@utk.edu

In many distributed computing systems that are prone to either induced or spontaneous node failures, the number of available computing resources is dynamically changing in a random fashion. A load-balancing (LB) policy for such systems should therefore be robust, in terms of workload re-allocation and effectiveness in task completion, with respect to the random absence and re-emergence of nodes as well as random delays in the transfer of workloads among nodes. In this paper two LB policies for such computing environments are presented: The first policy takes an initial LB action to preemptively counteract the consequences of random failure and recovery of nodes. The second policy compensates for the occurrence of node failure dynamically by transferring loads only at the actual failure instants. A probabilistic model, based on the concept of regenerative processes, is presented to assess the overall performance of the system under these policies. Optimal performance of both policies is evaluated using analytical, experimental and simulation-based results. The interplay between node-failure/recovery rates and the mean load-transfer delay are highlighted.

Dynamic Structured Partitioning for Parallel Scientific Applications with Pointwise Varying Workloads

Sumir Chandra[1], Manish Parashar[1] and Jaideep Ray[2]

[1]*ECE Dept./CAIP Center*
Rutgers University
Piscataway, NJ, USA
{*sumir, parashar*}*@caip.rutgers.edu*

[2]*Advanced Software R&D*
Sandia National Laboratories
Livermore, CA, USA
jairay@ca.sandia.gov

Parallel implementations of scientific applications involving the simulation of reactive flow on structured grids are challenging, since the underlying phenomena include transport processes with uniform computational loads as well as reactive processes having pointwise varying workloads. As a result, traditional parallelization approaches that assume homogeneous loads are not suitable for these simulations. This paper presents "*Dispatch*", a dynamic structured partitioning strategy that has been applied to parallel uniform and adaptive formulations of simulations with computational heterogeneity. *Dispatch* maintains the computational weights associated with pointwise processes in a distributed manner, computes the local workloads and partitioning thresholds, and performs in-situ locality-preserving load balancing. The experimental evaluation of *Dispatch* using an illustrative 2-D reactive-diffusion kernel demonstrates improvement in load distribution and overall application performance.

Accelerating Shape Optimizing Load Balancing for Parallel FEM Simulations by Algebraic Multigrid

Henning Meyerhenke, Burkhard Monien and Stefan Schamberger

Fakultät für Elektrotechnik, Informatik und Mathematik
Universität Paderborn
Paderborn, Germany
{*henningm, bm, schaum*}*@uni-paderborn.de*

We propose a load balancing heuristic for parallel adaptive finite element method (FEM) simulations. In contrast to most existing approaches, the heuristic focuses on good partition shapes rather than on minimizing the classical edge-cut metric. By applying Algebraic Multigrid (AMG), we are able to speed up the two most time consuming calculations of the approach while maintaining its large amount of natural parallelism.

Session 10

COMPUTATIONAL SCIENCE: BIOLOGY, CHEMISTRY, and PHYSICS

Parallelization and Performance Characterization of Protein 3D Structure Prediction of Rosetta

Wenlong Li[1], Tao Wang[1], Eric Li[1], David Baker[2], Li Jin[1], Steven Ge[1], Yurong Chen[1] and Yimin Zhang[1]

[1]*Intel China Research Center*
Intel Corporation
Beijing, China
{*wenlong.li, tao.wang, eric.q.li, steve.ge, yurong.chen, yimin.zhang*}*@intel.com, li.jin@inel.com*

[2]*Department of Biochemistry*
University of Washington
Washington, USA
dabaker@u.washington.edu

The prediction of protein 3D structure has become a hot research area in the post-genome era, through which people can understand a proteins function in health and disease, explore ways to control its actions and assist drug design. Many protein structure prediction approaches have been proposed in past decades. Among them, Rosetta is one of the best systems. However, the huge time complexity of Rosetta, e.g. a few days to predict a protein, limits its wide use in practice.To accelerate the prediction of protein 3D structure in Rosetta, this paper presents three different approaches, i.e., non-interactive, periodic interactive and asynchronous dynamic interactive scheme, to parallelize Rosetta. The asynchronous interactive scheme, with the adaptation of dynamic solution interaction, outperforms the other two, delivering much faster convergence speed and better solution quality. Detailed measurements and performance analysis also indicate that parallel Rosetta with asynchronous dynamic interactive scheme scales well.

Grid solutions for biological and physical cross-site simulations on the TeraGrid

S. Dong[1], N.t. Karonis[2,3] and G.e. Karniadakis[1]

[1]*Division of Applied Mathematics*
Brown University
Providence, RI, USA
{*sdong, gk*}*@dam.brown.edu*

[2]*Department of Computer Science*
Northern Illinois University
DeKalb, IL, USA
karonis@niu.edu

[3]*Mathematics and Computer Science Division*
Argonne National Lab
Argonne, IL, USA

Computational grids and grid middleware offer unprecedented computational power and storage capacity, and thus, have opened the possibility of solving problems that were previously not possible on even the largest single computational resources. These opportunities notwithstanding, the development of grid applications that run efficiently remains a challenge due to the inherent heterogeneity of networks and system architectures inherent in such environments. We present grid solutions to two grand challenge problems in computational mechanics. To study the scalability of our solutions we implemented both as MPI applications and ran them on the TeraGrid using NEKTAR and MPICH-G2. We present the results of our study which demonstrate near linear scalability in both applications when run across multiple TeraGrid sites and at a scale of hundreds or processors.

Achieving Strong Scaling with NAMD on Blue Gene/L

Sameer Kumar[1], Chao Huang[2], Gheorghe Almasi[3] and Laxmikant V. Kale[4]

[1] *T. J. Watson Research Center*
Yorktown Heights, NY, 10598
sameerk@us.ibm.com

[2] *Department of Computer Science*
University of Illinois
Urbana, IL, 61801
chuang10@uiuc.edu

[3] *T. J. Watson Research Center*
Yorktown Heights, NY, 10598
gheorghe@us.ibm.com

[4] *Department of Computer Science*
University of Illinois
Urbana, IL, 61801
kale@uiuc.edu

NAMD is a scalable molecular dynamics application, which has demonstrated its performance on several parallel computer architectures. Strong scaling is necessary for molecular dynamics as problem size is fixed, and a large number of iterations need to be executed to understand interesting biological phenomenon. The Blue Gene/L machine is a massive source of compute power. It consists of tens of thousands of embedded Power PC 440 processors. In this paper, we present several techniques to scale NAMD to 8192 processors of Blue Gene/L. These include topology specific optimizations, new messaging protocols, load-balancing, and overlap of computation and communication. We were able to achieve 1.2 TF of peak performance for cutoff simulations and 0.99 TF with PME.

Parallel ICA Methods for EEG Neuroimaging

Dan B. Keith, Christian C. Hoge, Robert M. Frank and Allen D. Malony

Neuroinformatics Center
University of Oregon
Eugene, OR, USA
{dkeith, hoge, rmfrank, malony}@cs.uoregon.edu

HiPerSAT, a C++ library and tools, processes EEG data sets with ICA (Independent Component Analysis) methods. *HiPerSAT* uses **BLAS**, **LAPACK**, **MPI** and **OpenMP** to achieve a high performance solution that exploits parallel hardware. ICA is a class of methods for analyzing a large set of data samples and extracting independent components that explain the observed data. ICA is used in EEG research for data cleaning and separation of spatiotemporal patterns that may reflect different underlying neural processes. We present two ICA implementations (FastICA and Infomax) that exploit parallelism to provide an EEG component decomposition solution of higher performance and data capacity than current MATLAB-based implementations. Experimental results and the methodology used to obtain them are presented. Integrating HiPerSAT with **EEGLAB** is described, as well as future plans for this research.

Session 11

PERFORMANCE EVALUATION and MODELS

Early Evaluation of the Cray XT3

Jeffrey S. Vetter, Sadaf R. Alam, Thomas H. Dunigan, Jr., Mark R. Fahey, Philip C. Roth and Patrick H. Worley

Oak Ridge National Laboratory
Oak Ridge, TN, 37831
vetter@computer.org, {alamsr, dunigan, faheymr, rothpc, worleyph}@ornl.gov

Oak Ridge National Laboratory recently received delivery of a 5,294 processor Cray XT3. The XT3 is Crays third-generation massively parallel processing system. The system builds on a single processor nodebuilt around the AMD Opteronand uses a custom chipcalled SeaStarto provide interprocessor communication. In addition, the system uses a lightweight operating system on the compute nodes. This paper describes our initial experiences with the system, including micro-benchmark, kernel, and application benchmark results. In particular, we provide performance results for strategic Department of Energy applications areas including climate and fusion. We demonstrate experiments on the installed system, scaling applications up to 4,096 processors.

A Study of the On-Chip Interconnection Network for the IBM Cyclops64 Multi-Core Architecture

Ying Ping Zhang, Taikyeong Jeong, Fei Chen, Haiping Wu, Ronny Nitzsche and Guang R. Gao

Department of Electrical and Computer Engineering
University of Delaware
Newark, DE, USA
{yzhang, ttjeong, fchen, hwu, ggao}@capsl.udel.edu, Ronny.Nitzsche@s2000.tu-chemnitz.de

The designs of high-performance processor architectures are moving toward the integration of a large number of multiple processing cores on a single chip. The IBM Cyclops-64 (C64) is a petaflop supercomputer built on multi-core system-on-a-chip technology. Each C64 chip employs a multistage pipelined crossbar switch as its on-chip interconnection network to provide high bandwidth and low latency communication between the 160 thread processing cores, the on-chip SRAM memory banks, and other components.

In this paper, we present a study of the architecture and performance of the C64 on-chip interconnection network through simulation. Our experimental results provide observations on the network behavior: (1) Dedicated channels can be created between any output port to input port of the C64 crossbar with latency as low as 7 cycles. The C64 crossbar has the potential reach the full hardware bandwidth, and exhibit a non-blocking behavior; (2) The C64 crossbar is a stable network; (3) The network logic design appears to provide a reasonable opportunity for sharing the channel bandwidth between traffic in either direction; (4) A simple circular neighbor arbitration scheme can achieve competitive performance level comparing to the complex segmented LRU (Least Recently Used) matrix arbitration scheme without losing the fairness. (5) Application-driven benchmarks provide comparable results to synthetic workloads.

A Performance Model for Fine-Grain Accesses in UPC

Zhang Zhang and Steven R. Seidel

Department of Computer Science
Michigan Technological University
Houghton, Michigan, U.S.
{zhazhang, steve}@mtu.edu

UPC's implicit communication and fine-grain programming style make application performance modeling a challenging task. The correspondence between remote references and communication events depends on the internals of the compiler and runtime system. This correspondence is often hidden from application developers. Aggressive optimizations allowed by the relaxed memory consistency model further blur this correspondence by transforming code structure. A modeling approach based on UPC platform benchmarking and code analysis is proposed. This approach abstracts a UPC platform according to its potential to apply a few common optimizations, then divides remote references in the application code into groups, based on a dependence analysis, that are amenable to each optimization. Each group is associated with a cost, obtained via benchmarking each potential optimization. The aggregated cost of these groups is the predicted cost of the application. Three simple UPC applications modeled using this approach usually yielded performance predictions within 15 percent of actual running times.

Analytical Performance Modelling of Adaptive Wormhole Routing in the Star Interconnection Network

Abbas Eslami Kiasari[1,2], Hamid Sarbazi-azad[1,2] and Mohamed Ould-khaoua[3]

[1]*IPM School of Computer Science*
Tehran, Iran
{kiasari, azad}@ipm.ir

[2]*Dept. of Computer Engineering*
Sharif University of Technology
Tehran, Iran

[3]*Dept. of Computing Science*
University of Glasgow
Glasgow, UK
mohamed@dcs.gla.ac.uk

The star graph was introduced as an attractive alternative to the well-known hypercube and its properties have been well studied in the past. Most of these studies have focused on topological properties and algorithmic aspects of this network. Although several analytical models have been proposed in the literature for different interconnection networks, none of them have dealt with star graphs. This paper proposes the first analytical model to predict message latency in wormhole-switched star interconnection networks with fully adaptive routing. The analysis focuses on a fully adaptive routing algorithm which has shown to be the most effective for star graphs. The results obtained from simulation experiments confirm that the proposed model exhibits a good accuracy under different operating conditions.

Session 12

INPUT/OUTPUT

Bitmap Indexes for Large Scientific Data Sets: A Case Study

Rishi Rakesh Sinha, Soumyadeb Mitra and Marianne Winslett

Department of Computer Science
University of Illinois, Urbana-Champaign
Urbana, IL, USA
{*rsinha, mitra1*}*@uiuc.edu, winslett@cs.uiuc.edu*

The data used by today's scientific applications are often very high in dimensionality and staggering in size. These characteristics necessitate the use of a good multidimensional indexing strategy to provide efficient access to the data. Researchers have previously proposed the use of bitmap indexes for high-dimension scientific data as a way of overcoming the drawbacks of traditional multidimensional indexes such as R-trees and KD-trees, which are bulky and whose performance does not scale well as the number of dimensions increases. However, the techniques proposed in previous work on bitmap indexes are not sufficient to address all problems that arise in practice. In experiments with real datasets, we experienced problems with index size and query performance. To overcome these shortcomings, we propose the use of adaptive, multilevel, multi-resolution bitmap indexes, and evaluate their performance in two scientific domains. Our preliminary experiments with a parallel query processor and index creator also show that it is very easy to parallelize a bitmap index.

MPI-IO/L: Efficient Remote I/O for MPI-IO via Logistical Networking

Jonghyun Lee[1], Robert Ross[1], Scott Atchley[2], Micah Beck[3] and Rajeev Thakur[1]

[1]*Mathematics and Computer Science Division*
Argonne National Laboratory
Argonne, IL, USA
{*jlee, rross, thakur*}*@mcs.anl.gov*

[2]*Myricom, Inc.*
Oak Ridge, TN, USA
atchley@myri.com

[3]*Department of Computer Science*
University of Tennessee
Knoxville, TN, USA
mbeck@cs.utk.edu

Scientific applications often need to access remotely located files, but many remote I/O systems lack standard APIs that allow efficient and direct access from application codes. This work presents MPI-IO/L, a remote I/O facility for MPI-IO using Logistical Networking. This combination not only provides high-performance and direct remote I/O using the standard parallel I/O interface but also offers convenient management and sharing of remote files. We show the performance trade-offs with various remote I/O approaches implemented in the system, which can help scientists identify preferable I/O options for their own applications. We also discuss how Logistical Networking could be improved to work better with parallel I/O systems such as ROMIO.

Evaluating I/O Characteristics and Methods for Storing Structured Scientific Data

Avery Ching[1], Alok Choudhary[1], Wei-keng Liao[1], Lee Ward[2] and Neil Pundit[2]

[1]*Department of EECS*
Northwestern University
Evanston, IL, USA
{*aching, choudhar, wkliao*}*@ece.northwestern.edu*

[2]*Scalable Computer Systems Department*
Sandia National Laboratories
Albuquerque, NM, USA
{*lee, pundit*}*@sandia.gov*

Many large-scale scientific simulations generate large, structured multi-dimensional datasets. Data is stored at various intervals on high performance I/O storage systems for checkpointing, post-processing, and visualization. Data storage is very I/O intensive and can dominate the overall running time of an application, depending on the characteristics of the I/O access pattern. Our NCIO benchmark determines how I/O characteristics greatly affect performance (up to 2 orders of magnitude) and provides scientific application developers with guidelines for improvement. In this paper, we examine the impact of various I/O parameters and methods when using the MPI-IO interface to store structured scientific data in an optimized parallel file system.

Dual-Layered File Cache On cc-NUMA System

Zhou Yingchao, Meng Dan and Ma Jie

Institute of Computing Technology
Chinese Academy of Sciences
Beijing, P.R.China
zhou.yingchao@gmail.com, {*md, majie*}*@ncic.ac.cn*

CC-NUMA is a widely adopted and deployed architecture of high performance computers. These machines are attractive for their transparent access to local and remote memory. However, the prohibitive latency gap between local and remote access deteriorates applications performance seriously due to memory access stalls. File system cache, especially, being shared by all processes, inevitably triggers many remote accesses. To address this problem, we suggest and implement a mechanism that uses local memory to cache remote file cache, of which the main purpose is to improve data locality. Using realistic workload on a two-node cc-NUMA machine, we show that the cost of such a mechanism is as low as 0.5%, the performance can be increased 14.3% at most, and the local hit ratio can be improved as much as 40%.

Session 13

SCHEDULING, 2

Dynamic Multi Phase Scheduling for Heterogeneous Clusters

Florina Monica Ciorba[1], Theodore Andronikos[1], Ioannis Riakiotakis[1], Anthony T. Chronopoulos[2] and George Papakonstantinou[1]

[1]*Electrical and Computer Engineering*
National Technical University of Athens
Athens, Attica, Greece
{*cflorina, tedandro, iriak, papakon*}*@cslab.ece.ntua.gr*

[2]*Department of Computer Science*
University of Texas at San Antonio
San Antonio, Texas, USA
atc@cs.utsa.edu

Distributed computing systems are a viable and less expensive alternative to parallel computers. However, concurrent programming methods in distributed systems have not been studied as extensively as for parallel computers. Some of the main research issues are how to deal with scheduling and load balancing of such a system, which may consist of heterogeneous computers. In the past, a variety of dynamic scheduling schemes suitable for parallel loops (with independent iterations) on heterogeneous computer clusters have been obtained and studied. However, no study of dynamic schemes for loops with iteration dependencies has been reported so far. In this work we study the problem of scheduling loops with iteration dependencies for heterogeneous (dedicated and non-dedicated) clusters. The presence of iteration dependencies incurs an extra degree of difficulty and makes the development of such schemes quite a challenge. We extend three well known dynamic schemes (CSS, TSS and DTSS) by introducing synchronization points at certain intervals so that processors compute in pipelined fashion. Our scheme is called dynamic multi-phase scheduling ($DMPS$) and we apply it to loops with iteration dependencies. We implemented our new scheme on a network of heterogeneous computers and studied its performance. Through extensive testing on two real-life applications (the heat equation and the Floyd-Steinberg algorithm), we show that the proposed method is efficient for parallelizing nested loops with dependencies on heterogeneous systems.

Using Virtual Grids to Simplify Application Scheduling

Richard Huang[1], Henri Casanova[2] and Andrew A. Chien[1]

[1]*Computer Science & Engineering and Center for Network Systems*
University of California, San Diego
La Jolla, California, USA
{*ryhuang, achien*}*@csag.ucsd.edu*

[2]*Information and Computer Sciences Department*
University of Hawai'i at Manoa
Honolulu, Hawaii, USA
henric@hawaii.edu

Users and developers of grid applications have access to increasing numbers of resources. While more resources generally mean higher capabilities for an application, they also raise the issue of application scheduling scalability. First, even polynomial time scheduling heuristics may take a prohibitively long time to compute a schedule. Second, and perhaps more critical, it may not be possible to gather all the resource information needed by a scheduling algorithm in a scalable manner. Our application focus is scientific workflows, which can be represented as Directed Acyclic Graphs (DAGs). Our claim is that, in future resource-rich environments, simple scheduling algorithms may be sufficient to achieve good workflow performances. We introduce a scalable scheduling approach that uses a resource abstraction called a virtual grid (VG). Our simulations of a range of typical DAG structures and resources demonstrate that a simple greedy scheduling heuristic combined with the virtual grid abstraction is as effective and more scalable than more complex heuristic DAG scheduling algorithms on large-scale platforms.

Enhancing Downlink Performance in Wireless Networks by Simultaneous Multiple Packet Transmission

Zhenghao Zhang and Yuanyuan Yang

Electrical and Computer Engineering
State University of New York at Stony Brook
Stony Brook, New York, USA
{*zhhzhang, yang*}*@ece.sunysb.edu*

In this paper we consider using simultaneous Multiple Packet Transmission (MPT) to improve the downlink performance of wireless networks. With MPT, the sender can send two compatible packets simultaneously to two distinct receivers and can double the throughput in the ideal case. We formalize the problem of finding a schedule to send out buffered packets in minimum time as finding a maximum matching problem in a graph. Since maximum matching algorithms are relatively complex and may not meet the timing requirements of real time applications, we give a fast approximation algorithm that is capable of finding a matching at least 3/4 of the size of a maximum matching in $O(|E|)$ time where $|E|$ is the number of edges in the graph. We also give analytical bounds for maximum allowable arrival rate which measures the speedup of the downlink after enhanced with MPT and our results show that the maximum arrival rate increases significantly even with a very small compatibility probability. We also use an approximate analytical model and simulations to study the average packet delay and our results show that packet delay can be greatly reduced even with a very small compatibility probability.

Instability in Parallel Job Scheduling Simulation: The Role of Workload Flurries

Dan Tsafrir and Dror G. Feitelson

School of Computer Science and Engineering
The Hebrew University
Jerusalem, Israel
{*dants, feit*}*@cs.huji.ac.il*

The performance of computer systems depends, among other things, on the workload. This motivates the use of real workloads (as recorded in activity logs) to drive simulations of new designs. Unfortunately, real workloads may contain various anomalies that contaminate the data. A previously unrecognized type of anomaly is workload flurries: rare surges of activity with a repetitive nature, caused by a single user, that dominate the workload for a relatively short period. We find that long workloads often include at least one such event. We show that in the context of parallel job scheduling these events can have a significant effect on performance evaluation results, e.g. a very small perturbation of the simulation conditions might lead to a large and disproportional change in the outcome. This instability is due to jobs in the flurry being effected in unison, a consequence of the flurry's repetitive nature. We therefore advocate that flurries be filtered out before the workload is used, in order to achieve stable and more reliable evaluation results (analogously to the removal of outliers in statistical analysis). At the same time, we note that more research is needed on the possible effects of flurries.

Session 14

DATA-INTENSIVE APPLICATIONS

Supporting Self-Adaptation in Streaming Data Mining Applications

Liang Chen and Gagan Agrawal

Department of Computer Science and Engineering
Ohio State University
Columbus, OHIO, U.S.A.
{*chenlia, agrawal*}*@cse.ohio-state.edu*

There are many application classes where the users are flexible with respect to the output quality. At the same time, there are other constraints, such as the need for real-time or interactive response, which are more crucial. This paper presents and evaluates a runtime algorithm for supporting adaptive execution for such applications. The particular domain we target is distributed data mining on streaming data. This work has been done in the context of a middleware system called GATES (Grid-based AdapTive Execution on Streams) that we have been developing.

The self-adaptation algorithm we present and evaluate in this paper has the following characteristics. First, it carefully evaluates the long-term load at each processing stage. It considers different possibilities for the load at a processing stage and its next stages, and decides if the value of an adaptation parameter needs to be modified, and if so, in which direction. To find the ideal new value of an adaptation parameter, it performs a binary search on the specified range of the parameter.

To evaluate the self-adaptation algorithm in our middleware, we have implemented two streaming data mining applications. The main observations from our experiments are as follows. First, our algorithm is able to quickly converge to stable values of the adaptation parameter, for different data arrival rates, and independent of the specified initial value. Second, in a dynamic environment, the algorithm is able to adapt the processing rapidly. Finally, in both static and dynamic environments, the algorithm clearly outperforms the algorithm described in our earlier work and an obvious alternative, which is based on linear-updates.

Distributed Antipole Clustering for Efficient Data Search and Management in Euclidean and Metric Spaces

Alfredo Ferro[1], Rosalba Giugno[1], Misael Mongiovì[1,2], Giuseppe Pigola[1] and Alfredo Pulvirenti[1]

[1]*Dept. of Mathematics and Computer Science*
University of Catania
Catania, Italy
{*ferro, giugno, pigola, apulvirenti*}*@dmi.unict.it,*
mongiovi@proteo.it

[2]*Research and Development Department*
Proteo S.p.A.
Catania, Italy

In this paper a simple and efficient distributed version of the recently introduced Antipole Clustering algorithm for general metric spaces is proposed. This combines ideas from the M-Tree, the Multi-Vantage Point structure and the FQ-Tree to create a new structure in the "bisector tree" class, called the Antipole Tree. Bisection is based on the proximity to an "Antipole" pair of elements generated by a suitable linear randomized tournament. The final winners (A,B) of such a tournament are far enough apart to approximate the diameter of the splitting set. A simple linear algorithm computing Antipoles in Euclidean spaces with exponentially small approximation ratio is proposed. The Antipole Tree Clustering has been shown to be very effective in important applications such as range and k-nearest neighbor searching, mobile objects clustering in centralized wireless networks with movable base stations and multiple alignment of biological sequences. In many of such applications an efficient distributed clustering algorithm is needed. In the proposed distributed versions of Antipole Clustering the amount of data passed from one node to another is either constant or proportional to the number of nodes in the network. The Distributed Antipole Tree is equipped with additional information in order to perform efficient range search and dynamic clusters management. This is achieved by adding to the randomized tournaments technique, methodologies taken from established systems such as BFR and BIRCH*. Experiments show the good performance of the proposed algorithms on both real and synthetic data.

Exploiting Programmable Network Interfaces for Parallel Query Execution in Workstation Clusters

Santhosh Kumar[1], M. J. Thazhuthaveetil[1,2] and R. Govindarajan[1,2]

[1]*Supercomputer Edn. and Res. Centre*
Indian Institute of Science
Bangalore, Karnataka, India
gvsk@hpc.serc.iisc.ernet.in, {mjt, govind}@serc.iisc.ernet.in

[2]*Supercomputer Edn. and Res. Centre*
Dept. of Computer Science and Automation
Indian Institute of Science
Bangalore, Karnataka, India

Workstation clusters equipped with high performance interconnect having programmable network processors facilitate interesting opportunities to enhance the performance of parallel application run on them. In this paper, we propose schemes where certain application level processing in parallel database query execution is performed on the network processor. We evaluate the performance of TPC-H queries executing on a high end cluster where all tuple processing is done on the host processor using a timed Petri net model, and find that tuple processing costs on the host processor dominate the execution time. These results are validated using a small cluster.

We therefore propose 4 schemes where certain tuple processing activity is offloaded to the network processor. The first 2 schemes offload the tuple splitting activity – computation to identify the node on which to process the tuples, resulting in an execution time speedup of 1.09 relative to the base scheme, but with I/O bus becoming the bottleneck resource. In the 3rd scheme in addition to offloading tuple processing activity, the disk and network interface are combined to avoid the I/O bus bottleneck, which results in speedups upto 1.16, but with high host processor utilization. Our 4th scheme where the network processor also performs a part of join operation along with the host processor, gives a speedup of 1.47% along with balanced system resource utilizations. Further we observe that the proposed schemes perform equally well even in a scaled architecture i.e., when the number of processors is increased from 2 to 64.

Design and Analysis of a Multi-dimensional Data Sampling Service for Large Scale Data Analysis Applications

Xi Zhang[1,2], Tahsin Kurc[1], Joel Saltz[1,2] and Srinivasan Parthasarathy[2]

[1]*Department of Biomedical Informatics*
The Ohio State University
Columbus, OH, USA
{xizhang, kurc, jsaltz}@bmi.osu.edu

[2]*Department of Computer Science and Engineering*
The Ohio State University
Columbus, OH, USA
srini@cse.ohio-state.edu

Sampling is a widely used technique to increase efficiency in database and data mining applications operating on large dataset. In this paper we present a scalable sampling implementation that supports efficient, multi-dimensional spatio-temporal sample generation on dynamic, large scale datasets stored on a storage cluster. The proposed algorithm leverages Hilbert space-filling curves in order to provide an approximate linear order of multidimensional data while maintaining spatial locality. This new implementation is then bootstrapped on top of our previous implementation, which efficiently samples large datasets along a single dimension (e.g., time), thereby realizing a service for spatio-temporal sampling. We evaluate the performance of our approach comparing it to the popular R-tree based technique. The experimental results show that our approach achieves up to an order of magnitude higher efficiency and scalability.

Session 15

ENERGY CONSIDERATIONS

Parallel Algorithms for Inductance Extraction of VLSI Circuits

Hemant Mahawar and Vivek Sarin

Department of Computer Science
Texas A&M University
College Station, Texas, USA
{*mahawarh, sarin*}*@cs.tamu.edu*

Inductance extraction involves estimating the mutual inductance in a VLSI circuit. Due to increasing clock speed and diminishing feature sizes of modern VLSI circuits, the effects of inductance are increasingly felt during the testing and verification stages. Hence, there is a need for fast and accurate inductance extraction software. A generalized approach for inductance extraction requires the solution of a dense complex symmetric linear system that models mutual inductive effects among circuit elements. Iterative methods are used to solve the system without explicit computation of the matrix itself. Fast hierarchical techniques are used to compute approximate matrix-vector products with the dense system matrix. This work presents an overview of a new parallel software package for inductance extraction of large VLSI circuits. The technique uses a combination of the solenoidal basis method and effective preconditioning schemes to solve the linear system. Fast Multipole Method (FMM) is used to compute approximate matrix-vector products with the inductance matrix. By formulating the preconditioner as a dense matrix similar to the coefficient matrix, we are able to use FMM for the preconditioning step as well. A two-tier parallelization scheme allows an efficient parallel implementation using both OpenMP and MPI directives simultaneously. The experiments conducted on various multiprocessor machines demonstrate the portability and parallel performance of the software.

Leakage-Aware Multiprocessor Scheduling for Low Power

Pepijn De Langen and Ben Juurlink

Computer Engineering Laboratory
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Delft, the Netherlands
{*pepijn, benj*}*@ce.et.tudelft.nl*

It is expected that (single chip) multiprocessors will increasingly be deployed to realize high-performance embedded systems. Because in current technologies the dynamic power consumption dominates the static power dissipation, an effective technique to reduce energy consumption is to employ as many processors as possible in order to finish the tasks as early as possible, and to use the remaining time before the deadline (the slack) to apply voltage scaling. We refer to this heuristic as Schedule and Stretch (S&S). However, since the static power consumption is expected to become more significant, this approach will no longer be efficient when leakage current is taken into account. In this paper, we first show for which combinations of leakage current, supply voltage, and clock frequency the static power consumption dominates the dynamic power dissipation. These results imply that, at a certain point, it is no longer advantageous from an energy perspective to employ as many processors as possible. Thereafter, a heuristic is presented to schedule the tasks on a number of processors that minimizes the total energy consumption. Experimental results obtained using a public task graph benchmark set show that our leakage-aware scheduling algorithm reduces the total energy consumption by up to 24% for tight deadlines (1.5x the critical path length) and by up to 67% for loose deadlines (8x the critical path length) compared to S&S.

A Dependable Infrastructure of the Electric Network for E-textiles

Nenggan Zheng[1], Zhaohui Wu[1], Man Lin[2] and Minde Zhao[1]

[1]*College of Computer Science*
Zhejiang University
Hangzhou, Zhejiang Province, P. R. China
{*zng, wzh, zmdd48*}*@zju.edu.cn*

[2]*Department of Computer Science*
St. Francis Xavier University
Antigonish, Nova Scotia, Canada
mlin@stfx.ca

Electronic textiles, known as computational fabrics, offer an emerging method for constructing wearable and large area applications. Because e-textiles are battery-driven and fault-prone systems, there is a need for developing a dependable infrastructure of the electric networks for e-textiles. In this paper, a new infrastructure of the power networks for e-textiles, Flexible Power Network (FPN), is presented. Instead of drawing power from a fixed battery as in the conventional electric networks, the power consuming nodes in a FPN can obtain power energy from one of the choices of batteries available with the help of the battery selectors. We also introduce the over current protectors into the battery nodes (BN) to protect the batteries from wasting the charge when short-circuit faults occur. The electric features of battery selectors and over current protectors, the two types of important electric devices used in FPNs, are illustrated in the paper. We have performed simulation experiments and the results show that our FPNs are more dependable than some common electric networks published before in the cases of short- and open-circuit faults.

Battery-Aware Router Scheduling in Wireless Mesh Networks

Chi Ma[1], Zhenghao Zhang[2] and Yuanyuan Yang[2]

[1]*Department of Computer Science*
State University of New York at Stony Brook
Stony Brook, New York, USA
mchi@cs.sunysb.edu

[2]*Department of Electrical and Computer Engineering*
State University of New York at Stony Brook
Stony Brook, New York, USA
{*zhhzhang, yang*}*@ece.sunysb.edu*

Wireless mesh networks recently emerge as a flexible, low-cost and multipurpose networking platform with wired infrastructure connected to the Internet. A critical issue in mesh networks is to maintain network activities for a long lifetime with high energy efficiency. As more and more outdoor applications require long-lasting, high energy efficient and continuously-working mesh networks with battery-powered mesh routers, it is important to optimize the performance of mesh networks from a battery-aware point of view. Recent study in battery technology reveals that discharging of a battery is nonlinear. Batteries tend to discharge more power than needed, and reimburse the over-discharged power later if they have sufficiently long recovery time. Intuitively, to optimize network performance, a mesh router should recover its battery periodically to prolong the lifetime. In this paper, we introduce a mathematical model on battery discharging duration and lifetime for wireless mesh networks. We also present a battery lifetime optimization scheduling algorithm (BLOS) to maximize the lifetime of battery-powered mesh routers. Based on the BLOS algorithm, we further consider the problem of using battery powered routers to monitor or cover a few hot spots in the network. We refer to this problem as the Spot Covering under BLOS Policy problem (SCBP). We prove that the SCBP problem is NP-hard and give an approximation algorithm called the Spanning Tree Scheduling (STS) to dynamically schedule mesh routers. The key idea of the STS algorithm is to construct a spanning tree according to the BLOS Policy in the mesh network. The time complexity of the STS algorithm is O(r) for a network with r mesh routers. Our simulation results show that the STS algorithm can greatly improve the lifetime, data throughput and power consumption efficiency of a wireless mesh network.

Session 16

COMPILERS and OPTIMIZATION

Optimizing Bandwidth Limited Problems Using One-Sided Communication and Overlap

Christian Bell[1,2], Dan Bonachea[1], Rajesh Nishtala[1] and Katherine Yelick[1,2]

[1]*Computer Science Division*
University of California, Berkeley
Berkeley, CA, USA
{*csbell, bonachea, rajeshn, yelick*}*@cs.berkeley.edu*

[2]*Computational Research Division*
Lawrence Berkeley National Laboratory
Berkeley, CA, USA

Partitioned Global Address Space languages like Unified Parallel C (UPC) are typically valued for their expressiveness, especially for computations with fine-grained random accesses. In this paper we show that the one-sided communication model used in these languages also has a significant performance advantage for bandwidth-limited applications. We demonstrate this benefit through communication microbenchmarks and a case-study that compares UPC and MPI implementations of the NAS Fourier Transform (FT) benchmark. Our optimizations rely on aggressively overlapping communication with computation but spreading communication events throughout the course of the local computation. This alleviates the potential communication bottleneck that occurs when the communication is packed into a single phase (e.g., the large all-to-all in a multidimensional FFT). Even though the new algorithms require more messages for the same total volume of data, the resulting overlap leads to speedups of over $1.75\times$ and $1.9\times$ for the two-sided and one-sided implementations, respectively, when compared to the default NAS Fortran/MPI release. Our best one-sided implementations show an average improvement of 15% over our best two-sided implementations. We attribute this difference to the lower software overhead of one-sided communication, which is partly fundamental to the semantic difference between one-sided and two-sided communication. Our UPC results use the Berkeley UPC compiler with the GASNet communication system, and demonstrate the portability and scalability of that language and implementation, with performance approaching 0.5 TFlop/s on the FT benchmark running on 512 processors.

Performance analysis of parallel programs via message-passing graph traversal

Matthew J. Sottile[1], Vaddadi P. Chandu[2] and David A. Bader[2]

[1]*Los Alamos National Laboratory*
Los Alamos, NM, USA
matt@lanl.gov

[2]*Georgia Institute of Technology*
Atlanta, GA, USA
{*vchandu, bader*}*@cc.gatech.edu*

The ability to understand the factors contributing to parallel program performance are vital for understanding the impact of machine parameters on the performance of specific applications. We propose a methodology for analyzing the performance characteristics of parallel programs based on message-passing traces of their execution on a set of processors. Using this methodology, we explore how perturbations in both single processor performance and the messaging layer impact the performance of the traced run. This analysis provides a quantitative description of the sensitivity of applications to a variety of performance parameters to better understand the range of systems upon which an application can be expected to perform well. These performance parameters include operating system interference and variability in message latencies within the interconnection network layer.

A Compiler-based Communication Analysis Approach for Multiprocessor Systems

Shuyi Shao[1], Alex K. Jones[2] and Rami Melhem[1]

[1]*Department of Computer Science*
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
{*syshao, mlehem*}*@cs.pitt.edu*

[2]*Department of Electrical and Computer Engineering*
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
akjones@ece.pitt.edu

In this paper we describe a compiler framework which can identify communication patterns for MPI-based parallel applications. This has the potential of providing significant performance benefits when connections can be established in the network prior to the actual communication operation. Our compiler uses a flexible and powerful communication pattern representation scheme that can capture the property of communication patterns and allows manipulations of these patterns. In this way, communication phases can be detected and logically separated within the application. Additionally, we extend the classification of static and dynamic communication patterns and operations to include persistent communications. Persistent communications appear dynamic, however, they remain unchanged for large segments of the application execution. Our compiler is capable of detecting both static and persistent communication patterns within an application. We show that for the NAS Parallel Benchmarks, 100% of the point-to-point communications can be classified as either static or persistent and, with the exception of IS, 100% of the collective were either static or persistent. By comparison to application trace data, the predicted LBMHD, CG and MG communication patterns have been verified.

A Code Motion Technique for Accelerating General-Purpose Computation on the GPU

Takatoshi Ikeda, Fumihiko Ino and Kenichi Hagihara

Graduate School of Information Science and Technology
Osaka University
Toyonaka, Osaka 560-8531, Japan
{*ikeda, ino, hagihara*}*@ist.osaka-u.ac.jp*

Recently, graphics processing units (GPUs) are providing increasingly higher performance with programmable internal processors, namely vertex processors (VPs) and fragment processors (FPs). Such newly added capabilities motivate us to perform general-purpose computation on GPUs (GPGPU) beyond graphics applications. Although VPs and FPs are connected in a pipeline, many GPGPU implementations utilize only FPs as a computational engine in the GPU. Therefore, such implementations may result in lower performance due to highly loaded FPs (as compared to VPs) being a performance bottleneck in the pipeline execution. The objective of our work is to improve the performance of GPGPU programs by eliminating this bottleneck. To achieve this, we present a code motion technique that is capable of reducing the FP workload by moving assembly instructions appropriately from the FP program to the VP program. We also present the definition of such movable instructions that do not change the I/O specification between the CPU and the GPU. The experimental results show that (1) our technique improves the performance of a Gaussian filter program with reducing execution time by approximately 40% and (2) it successfully reduces the FP workload in 10 out of 18 GPGPU programs.

Session 17

MEMORY SHARING

A Distributed Paging RAM Grid System for Wide-Area Memory Sharing

Rui Chu[1], Nong Xiao[1], Yongzhen Zhuang[2], Yunhao Liu[2] and Xicheng Lu[1]

[1]*National Key Laboratory for Parallel and Distributed Processing*
HuNan, China
{*rchu, nongxiao, xclu*}*@nudt.edu.cn*

[2]*Hong Kong University of Science and Technology*
Kowloon, Hong Kong
{*cszyz, liu*}*@cs.ust.hk*

Memory-intensive applications often suffer from the poor performance of disk swapping when memory is inadequate. Remote memory sharing schemes, which provide a remote memory that is faster than the local hard disk, are able to improve the performance of such applications. Due to the limitation of being applicable within single clusters only, however, most of the previous remote memory mechanisms, such as the network memory scheme, fail to be extendable into a large scale, distributed, heterogeneous, and dynamic environment. In this work, we propose a service-oriented grid memory sharing scheme, Distributed Paging RAM Grid (DPRG). We study the properties and criteria of large scale memory sharing, and then design major operations and optimizations to fit the usage of grid systems. We collect trace from our grid environment, and evaluate DPRG through comprehensive trace-driven simulations. Results show that DPRG significantly outperforms existing remote memory sharing schemes and supports grid computing applications effectively.

Detecting Phases in Parallel Applications on Shared Memory Architectures

Erez Perelman[1], Marzia Polito[2], Jean-yves Bouguet[2], John Sampson[1], Brad Calder[1] and Carole Dulong[2]

[1]*Computer Science and Engineering*
University of California, San Diego
La Jolla, CA, USA
{*eperelma, jsampson, calder*}*@cs.ucsd.edu*

[2]*Intel*
Palo Alto, CA, USA
{*marzia.polito, jean-yves.bouguet, carole.dulong*}*@intel.com*

Most programs are repetitive, where similar behavior can be seen at different execution times. Algorithms have been proposed that automatically group similar portions of a program's execution into phases, where samples of execution in the same phase have homogeneous behavior and similar resource requirements.

In this paper, we examine applying these phase analysis algorithms and how to adapt them to parallel applications running on shared memory processors. Our approach relies on a separate representation of each thread's activity. We first focus on showing its ability to identify similar intervals of execution across threads for a single run. We then show that it is effective at identifying similar behavior of a program when the number of threads is varied between runs. This can be used by developers to examine how different phases scale across different number of threads. Finally, we examine using the phase analysis to pick simulation points to guide multi-threaded simulation.

Coterminous Locality and Coterminous Group Data Prefetching on Chip-Multiprocessors

Xudong Shi[1], Zhen Yang[1], Jih-kwon Peir[1], Lu Peng[2], Yen-kuang Chen[3], Victor Lee[3] and Bob Liang[3]

[1]*Computer & Information Science & Engineering*
University of Florida
Gainesville, Florida, USA
{*xushi, zhyang, peir*}*@cise.ufl.edu*

[2]*Electrical & Computer Engineering*
Louisiana State University
Baton Rouge, Louisiana, USA
lpeng@lsu.edu

[3]*Architecture Research Lab*
Intel Corporation
Santa Clara, California, USA
{*yen-kuang.chen, victor.w.lee, bob.liang*}*@intel.com*

Due to shared cache contentions and interconnect delays, data prefetching is more critical in alleviating penalties from increasing memory latencies and demands on Chip-Multiprocessors (CMPs). Through deep analysis of SPEC2000 applications, we find that a part of the nearby data memory references often exhibit highly-repeated patterns with long, but equal block reuse distance. These references can form a coterminous group (CG). Coterminous locality is introduced as that when a member in a CG is referenced, the remaining members will likely be referenced in the near future. Based on the coterminous locality behavior, we implement a novel CG data prefetcher on CMPs. Performance evaluations show that the proposed prefetcher can accurately cover up to 40-50% of the total misses, and result in 50-60% of potential performance improvement for several selected workload mixes.

Session 18

COMMUNICATION and COORDINATION

Concurrent Counting is Harder than Queuing

Srikanta Tirthapura[1] and Costas Busch[2]

[1]Dept. of Elec. and Computer Engg
Iowa State University
Ames, IA, USA
snt@iastate.edu

[2]Computer Science Department
Rensselaer Polytechnic Inst.
Troy, NY, USA
buschc@cs.rpi.edu

In both distributed counting and queuing, processors in a distributed system issue operations which are organized into a total order. In counting, each processor receives the rank of its operation in the total order, where as in queuing, a processor gets back the identity of its predecessor in the total order. Coordination applications such as totally ordered multicast can be solved using either distributed counting or queuing, and it would be very useful to definitively know which of counting or queuing is a harder problem.

We conduct the first systematic study of the relative complexities of distributed counting and queuing in a concurrent setting. Our results show that concurrent counting is harder than concurrent queuing on a variety of processor interconnection topologies, including high diameter graphs such as the list and the mesh, and low diameter graphs such as the complete graph, perfect m-ary tree, and the hypercube. For all these topologies, we show that the concurrent delay complexity of a particular solution to queuing, the arrow protocol, is asymptotically smaller than a lower bound on the complexity of any solution to counting. As a consequence, we are able to definitively say that given a choice between applying counting or queuing to solve a distributed coordination problem, queuing is the better solution.

Relationships between communication models in networks using atomic registers

Lisa Higham[1] and Colette Johnen[2]

[1]Computer Science Department
University of Calgary
Alberta, Alberta, Canada
higham@cpsc.ugalgary.ca

[2]LRI-CNRS
Université Paris-Sud
Orsay, France
colette@lri.fr

A common way to model a distributed system is with a graph where nodes represent processors and there is an edge between two processors if and only if they can communicate directly. In shared-registers versions of this general description, neighbouring processorscommunicate by reading or writing shared registers, where each read or write is one atomic step. This paper defined two models of shared registers determined by selecting the register locations (processors or links). In the *atomic state* model each processor has a register; in the *atomic link* model, each communication link has a register. We determine under what conditions and with what robustness and/or failure-tolerance guarantees it is possible to transform a solution under the *atomic state* model into a solution under *atomic link* model. The fault-tolerant models considered in this paper are wait-freedom and self-stabilization. These questions are addressed by first establishing a framework for defining correct transformations, which may be useful for similar studies of the relationship between various models of distributed computation.

RAPID: An End-System Aware Protocol for Intelligent Data Transfer over Lambda Grids

Amitabha Banerjee[1], Wu-chun Feng[2], Biswanath Mukherjee[1] and Dipak Ghosal[1]

[1]*Computer Science*
University of California Davis
Davis, CA, USA
abanerjee@ucdavis.edu, {mukherje, ghosal}@cs.ucdavis.edu

[2]*Computer Science*
Virginia Tech
Blacksburgh, VA, USA
feng@cs.vt.edu

Next-generation e-Science applications will require the ability to transfer information at high data rates between distributed computing centers and data repositories. To support such applications, lambda grid networks have been built to provide large, on-demand bandwidth between end-points that are interconnected via optical circuit-switched lambdas. It is extremely important to develop an efficient transport protocol over such high-capacity, dedicated circuits.

Because lambdas provide dedicated bandwidth between endpoints, they obviate the need for network congestion control. Consequently, past research has demonstrated that rate-based transport protocols, such as RBUDP, are more effective than TCP in transferring data over lambdas. However, while lambdas eliminate congestion in the network, they ultimately push the congestion to the endpoints — congestion that current rate-based transport protocols are ill-suited to handle. In this paper we introduce a "Rate-Adaptive Protocol for Intelligent Delivery (RAPID)" of data that is lightweight and end-system performance-aware, so as to maximize end-to-end throughput while minimizing packet loss. Based on self monitoring of the dynamic task-priority at the receiving end-system, our protocol enables the receiver to proactively deliver feedback to the sender, so that the sender may adapt its sending rate to avoid congestion at the receiving end-system. This avoids large bursts of packet losses typically observed in current rate-based transport protocols. Over a 10-Gigabit link emulation of an optical circuit, RAPID reduces file-transfer time, and hence improves end-to-end throughput by as much as 25%.

Session 19

FAULT and FAILURE TOLERANCE

The Interleaved Authentication for Filtering False Reports in Multipath Routing based Sensor Networks

Youtao Zhang[1,3], Jun Yang[2] and Hai T Vu[3]

[1]Computer Science Department
University of Pittsburgh
Pittsburgh, PA, USA
zhangyt@cs.pitt.edu

[2]Computer Science and Engineering Department
University of California at Riverside
Riverside, CA, USA
junyang@cs.ucr.edu

[3]Computer Science Department
University of Texas at Dallas
Richardson, TX, USA
htv041000@utdallas.edu

In this paper, we consider filtering false reports in braided multipath routing sensor networks. While multipath routing provides better resilience to various faults in sensor networks, it has two problems regarding the authentication design. One is that, due to the large number of partially overlapped routing paths between the source and sink nodes, the authentication overhead could be very high if these paths are authenticated individually; the other is that false reports may escape the authentication check through the newly identified node association attack. In this paper we propose enhancements to solve both problems such that secure and efficient authentication can be achieved in multipath routing. The proposed scheme is (t+1)-resilient, i.e. it is secure with up to t compromised nodes. The upper bound number of hops that a false report may be forwarded in the network is O(t^2).

Necessary and Sufficient Conditions for 1-adaptivity

Joffroy Beauquier[1], Sylvie Delaet[2] and Sammy Haddad[2]

[1]PCRI,LRI (CNRS UMR 8623), INRIA Futurs
Paris XI
Orsay, France
jb@lri.fr

[2]LRI (CNRS UMR 8623)
Paris XI
Orsay, France
{delaet, haddad}@lri.fr

A 1-adaptive self-stabilizing system is a self-stabilizing system that can correct any memory corruption of a single process in one computation step. 1-adaptivity means that if in a legitimate state the memory of a single process is corrupted, then the next system transition will lead to a legitimate state and the system will recover a correct behavior. Thus 1-adaptive self-stabilizing algorithms guarantee the very strong property that a single fault is corrected immediately and consequently that it cannot be propagated. Our aim here is to study necessary and sufficient conditions to obtain that property in order to design such algorithms. In particular we show that this property can be obtained even under the distributed demon and that it can also be applied to probabilistic algorithms.

We provide two self-stabilizing 1-adaptive algorithms that demonstrate how the conditions we present here can be used to design and prove 1-adaptive algorithms.

A Proactive Fault-detection Mechanism in Large-scale Cluster Systems

Wu Linping[1,2], Meng Dan[1], Gao Wen[1] and Zhan Jianfeng[1]

[1]*Institute of Computing Technology*
Chinese Academy of Sciences
Beijing, China
{*wlp, md, gw, jfzhan*}*@ncic.ac.cn*

[2]*Graduate School of the Chinese Academy of Sciences*
Beijing, China

To improve the whole dependability of large-scale cluster systems, an online fault detection mechanism is proposed in this paper. This mechanism can detect the fault in time before node fails and enables the proactive fault management. The proposed mechanism is summarized as follows: First, the dynamic characteristics of cluster system running in normal activity are built using Time Series Analysis methods. Second, the fault detection process is implemented by comparing the current running state of cluster system with normal running model. The fault alarm decision is made immediately when the current running state deviates the normal running model. The experiment results show that this mechanism can detect the fault in cluster system in good time.

Algorithm-Based Checkpoint-Free Fault Tolerance for Parallel Matrix Computations on Volatile Resources

Zizhong Chen[1] and Jack Dongarra[1,2]

[1]*Department of Computer Science*
The University of Tennessee, Knoxville
Knoxville, TN, USA
{*zchen, dongarra*}*@cs.utk.edu*

[2]*Computer Science and Mathematics Division*
Oak Ridge National Laboratory
Oak Ridge, TN, USA

As the size of today's high performance computers increases from hundreds, to thousands, and even tens of thousands of processors, node failures in these computers are becoming frequent events. Although checkpoint/rollback-recovery is the typical technique to tolerate such failures, it often introduces a considerable overhead. Algorithm-based fault tolerance is a very cost-effective method to incorporate fault tolerance into matrix computations. However, previous algorithm-based fault tolerance methods for matrix computations are often derived using algorithms that are seldomly used in the practice of today's high performance matrix computations and have mostly focused on platforms where failed processors produce incorrect calculations.

To fill this gap, this paper extends the existing algorithm-based fault tolerance to the volatile computing platform where the failied processor stops working and applies it to scalable high performance matrix computations with two dimensional block cyclic data distribution. We show the practicality of this technique by applying it to the ScaLAPACK/PBLAS matrix-matrix multiplication kernel. Experimental results demonstrate that the proposed approach is able to survive process failures with a very low performance overhead.

Session 20

MPI

Collective Operations in NEC's High-performance MPI Libraries

Jesper Larsson Traff and Hubert Ritzdorf

C&C Research Laboratories, NEC Europe Ltd.
Sankt Augustin, Germany
{*traff, ritzdorf*}*@ccrl-nece.de*

We give an overview of the algorithms and implementations in the high-performance MPI libraries MPI/SX and MPI/ES of some of the most important collective operations of MPI (the *Message Passing Interface*). The infrastructure of MPI/SX makes it easy to incorporate new algorithms and algorithms for common special cases (e.g. a single SX node, or a single MPI process per SX node). Algorithms that are among the best known are employed, and special hardware features of the SX architecture and Internode Crossbar Switch (IXS) are exploited wherever possible. We discuss in more detail the implementation of `MPI_Barrier`, `MPI_Bcast`, the MPI reduction collectives, `MPI_Alltoall`, and the gather/scatter collectives.

Performance figures and comparisons to straightforward algorithms are given for a large SX-8 system, and for the *Earth Simulator*. The measurements show excellent absolute performance, and demonstrate the scalability of MPI/SX and MPI/ES to systems with large numbers of nodes.

Infiniband Scalability in Open MPI

Galen M. Shipman[1], Tim S. Woodall[1], Richard L. Graham[1], Arthur B. Maccabe[2] and Patrick G. Bridges[2]

[1]*Advanced Computing Laboratory*
Los Alamos National Laboratory
Los Alamos, NM, USA
{*gshipman, twoodall, rlgraham*}*@lanl.gov*

[2]*Department of Computer Science*
University of New Mexico
Albuquerque, NM, USA
{*maccabe, bridges*}*@cs.unm.edu*

Infiniband is becoming an important interconnect technology in high performance computing. Recent efforts in large scale Infiniband deployments are raising scalability questions in the HPC community. Open MPI, a new open source implementation of the MPI standard targeted for production computing, provides several mechanisms to enhance Infiniband scalability. Initial comparisons with MVAPICH, the most widely used Infiniband MPI implementation, show similar performance but with much better scalability characteristics. Specifically, small message latency is improved by up to 10% in medium/large jobs and memory usage per host is reduced by as much as 300%. In addition, Open MPI provides predictable latency that is close to optimal without sacrificing bandwidth performance.

Shared Receive Queue based Scalable MPI Design for InfiniBand Clusters

Sayantan Sur, Lei Chai, Hyun-wook Jin and Dhabaleswar K. Panda

Computer Science and Engineering
The Ohio State University
Columbus, Ohio, USA
{*surs, chail, jinhy, panda*}*@cse.ohio-state.edu*

Clusters of several thousand nodes interconnected with InfiniBand, an emerging high-performance interconnect, have already appeared in the Top 500 list. The next-generation InfiniBand clusters are expected to be even larger with tens-of-thousands of nodes. A high-performance scalable MPI design is crucial for MPI applications in order to exploit the massive potential for parallelism in these very large clusters. MVAPICH is a popular implementation of MPI over InfiniBand based on its reliable connection oriented model. The requirement of this model to make communication buffers available for each connection imposes a memory scalability problem. In order to mitigate this issue, the latest InfiniBand standard includes a new feature called Shared Receive Queue (SRQ) which allows sharing of communication buffers across multiple connections. In this paper, we propose a novel MPI design which efficiently utilizes SRQs and provides very good performance. Our analytical model reveals that our proposed designs will take only 1/10th the memory requirement as compared to the original design on a cluster sized at 16,000 nodes. Performance evaluation of our design on our 8-node cluster shows that our new design was able to provide the same performance as the existing design while requiring much lesser memory. In comparison to tuned existing designs our design showed a 20% and 5% improvement in execution time of NAS Benchmarks (Class A) LU and SP, respectively. The High Performance Linpack was able to execute a much larger problem size using our new design, whereas the existing design ran out of memory.

Executing MPI Programs on Virtual Machines in an Internet Sharing System

Zhelong Pan[1], Xiaojuan Ren[1], Rudolf Eigenmann[1] and Dongyan Xu[2]

[1]*School of Electrical and Computer Engineering*
Purdue University
West Lafayette, IN, USA
{*zpan, xren, eigenman*}*@purdue.edu*

[2]*Department of Computer Science*
Purdue University
West Lafayette, IN, USA
dxu@cs.purdue.edu

Internet sharing systems aim at federating and utilizing distributed computing resources across the Internet. This paper presents a user-level virtual machine (VM) approach to MPI program execution in an Internet sharing framework. In this approach, the resource consumer has its own operating system running on top of, and isolated from, the operating system of the resource provider. We propose an efficient socket virtualization technique to optimize VM network performance. Socket virtualization achieves the same network bandwidth as the physical network. In our LAN environment, it reduces the latency overhead from 172% (using existing TUN/TAP technique) to 35.6%. Performance results on MPI benchmarks show that our virtualization technique incurs small overhead compared with the physical host platform, while gaining in return a higher degree of guest isolation and customization. We also describe the key mechanisms that allow the employment of VMs in an existing Internet sharing system.

Adaptive Connection Management for Scalable MPI over InfiniBand

Weikuan Yu[1], Qi Gao[2] and Dhabaleswar K. Panda[3]

[1]Department of Computer Science and Engineering
The Ohio State University
Columbus, Ohio, USA
yuw@cse.ohio-state.edu

[2]Department of Computer Science and Engineering
The Ohio State University
Columbus, Ohio, USA
gaoq@cse.ohio-state.edu

[3]Department of Computer Science and Engineering
The Ohio State University
Columbus, Ohio, USA
panda@cse.ohio-state.edu

Supporting scalable and efficient parallel programs is a major challenge in parallel computing with the widespread adoption of large-scale computer clusters and supercomputers. One of the pronounced scalability challenges is the management of connections between parallel processes, especially over connection-oriented interconnects such as VIA and InfiniBand.

In this paper, we take on the challenge of designing efficient connection management for parallel programs over InfiniBand clusters. We propose adaptive connection management (ACM) to dynamically control the establishment of InfiniBand reliable connections (RC) based on the communication frequency between MPI processes. We have investigated two different ACM algorithms: an on-demand algorithm that starts with no InfiniBand RC connections; and a partial static algorithm with only $2 * logN$ number of InfiniBand RC connections initially. We have designed and implemented both ACM algorithms in MVAPICH to study their benefits. Two mechanisms have been exploited for the establishment of new RC connections: one using InfiniBand unreliable datagram and the other using InfiniBand connection management. For both mechanisms, MPI communication issues, such as progress rules, reliability and race conditions are handled to ensure efficient and light-weight connection management. Our experimental results indicate that ACM algorithms can benefit parallel programs in terms of the process initiation time, the number of active connections, and the resource usage. For parallel programs on a 16-node cluster, they can reduce the process initiation time by 15% and the initial memory usage by 18%.

Session 21

ROUTING

An Integrated Approach for Density Control and Routing in Wireless Sensor Networks

Isabela G. Siqueira[1], Carlos Maurício S. Figueiredo[1,2], Antonio Alfredo F. Loureiro[1], José Marcos Nogueira[1] and Linnyer Beatrys Ruiz[1]

[1]*Dept. of Computer Science*
Federal University of Minas Gerais
Belo Horizonte, MG, Brazil
{*isabela, mauricio, loureiro, jmarcos, linnyer*}*@dcc.ufmg.br*

[2]*Analysis, Research and Technological Innovation Center*
FUCAPI
Manaus, AM, Brazil

Wireless Sensor Networks (WSNs) are characterized by having scarce resources. The usual way of designing network functions is to consider them isolatedly, a strategy which may not guarantee the correct and efficient operation of WSNs. For this reason, in this paper we propose an integrated design of network functions. We take two important WSN functions — density control and routing — as an example and present two approaches to integrate them. In particular, we present two solutions, named RDC-Sync and RDC-Integrated, which integrate a geographical density control algorithm with tree routing. The simulations experiments performed prove that the integrated design improves the network performance, especially when density control and routing are fully integrated.

A Distributed Method for Dynamic Resolution of BGP Oscillations

Ahronovitz Ehoud, KÖnig Jean-claude and Saad Clément

LIRMM
Montpellier 2
Montpellier, France
{*aro, konig, saad*}*@lirmm.fr*

Autonomous Systems (AS) in the Internet use different protocols for internal and external routing. BGP is the only external protocol. It allows ASes to define their own routing policy independently. Many papers cited in reference deal with a divergence behavior due to this flexibility. In fact, when routing policies are not conflicting, BGP is self-stabilising, which means that whatever the network configuration, BGP converges to a stable solution. Unfortunately, as experienced on the Internet, AS routing policies may be uncoherent, thus generating oscillations. In this paper we propose a distributed dynamic method for detecting and solving oscillations of BGP. It respects private policy choices and requires only a few low level constraints in order to converge to a stable solution. Essentially, a router has to maintain only local path stateful information to detect instabilities. In this case, it generates and launches a token linked to a route. Each router makes the decision to forward or not the token according to local data and local policy. If the originating router receives back the token, then it marks the route as *barred*. Nevertheless, routes may furtherly be unmarked.

Segment-Based Routing: An Efficient Fault-Tolerant Routing Algorithm for Meshes and Tori

A. Mejia[1], J. Flich[1], J. Duato[1], Sven-arne Reinemo[2] and Tor Skeie[2]

[1]*Dpto de Informatica de Sistemas y Computadores*
Universidad Politecnica de Valencia
Valencia, Spain
{*andres, jflich, jduato*}*@gap.upv.es*

[2]*Simula Research Laboratory*
Simula Research Laboratory
Lysaker, Norway
{*svenar, tskeie*}*@simula.no*

Computers get faster every year, but the demand for computing resources seems to grow at an even faster rate. Depending on the problem domain, this demand for more power can be satisfied by either, massively parallel computers, or clusters of computers. Common for both approaches is the dependence on high performance interconnect networks such as Myrinet, Infiniband, or 10 Gigabit Ethernet. While high throughput and low latency are key features of interconnection networks, the issue of fault-tolerance is now becoming increasingly important. As the number of network components grows so does the probability for failure, thus it becomes important to also consider the fault-tolerance mechanism of interconnection networks. The main challenge then lies in combining performance and fault-tolerance, while still keeping cost and complexity low.

This paper proposes a new deterministic routing methodology for tori and meshes, which achieves high performance without the use of virtual channels. Furthermore, it is topology agnostic in nature, meaning it can handle any topology derived from any combination of faults when combined with static reconfiguration. The algorithm, referred to as Segment-based Routing (SR), works by partitioning a topology into subnets, and subnets into segments. This allows us to place bidirectional turn restrictions *locally* within a segment. As segments are independent, we gain the freedom to place turn restrictions within a segment independently from other segments. This results in a larger degree of freedom when placing turn restrictions compared to other routing strategies.

In this paper a way to compute segment-based routing tables is presented and applied to meshes and tori. Evaluation results show that SR increases performance by a factor of 1.8 over FX and up*/down* routing.

Network Uncertainty in Selfish Routing

Chryssis Georgiou, Theophanis Pavlides and Anna Philippou

Department of Computer Science
University of Cyprus
Nicosia, Cyprus
{*chryssis, phanosp, annap*}*@cs.ucy.ac.cy*

We study the problem of selfish routing in the presence of incomplete network information. Our model consists of a number of users who wish to route their traffic on a network of m parallel links with the objective of minimizing their latency. However, in doing so, they face the challenge of lack of precise information on the capacity of the network links. This uncertainty is modelled via a set of probability distributions over all the possibilities, one for each user. The resulting model is an amalgamation of the KP-model of [Koutsoupias and Papadimitriou, 1999] and the congestion games with user-specific functions of [Milchtaich, 1996].

We embark on a study of Nash equilibria and the price of anarchy in this new model. In particular, we propose polynomial-time algorithms for computing some special cases of pure Nash equilibria and we show that negative results of [Milchtaich, 1996], for the non-existence of pure Nash equilibria in the case of three users, do not apply to our model. Consequently, we propose an interesting open problem in this area, that of the existence of pure Nash equilibria in the general case of our model. Furthermore, we consider appropriate notions for the social cost and the price of anarchy and obtain upper bounds for the latter. With respect to fully mixed Nash equilibria, we propose a method to compute them and show that when they exist they are unique. Finally we prove that the fully mixed Nash equilibrium maximizes the social welfare.

Session 22

IMAGE PROCESSING and VISUALIZATION

MPEG-2 Decoding in a Stream Programming Language

Matthew Drake, Hank Hoffmann, Rodric Rabbah and Saman Amarasinghe

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, USA
{*madrake, hank, rabbah, saman*}*@mit.edu*

Image and video codecs are prevalent in multimedia devices, ranging from embedded systems, to desktop computers, to high-end servers such as HDTV editing consoles. It is not uncommon however that developers create and customize separate coder and decoder implementations for each of the architectures they target. This practice is time consuming and error prone, leading to code that is neither malleable nor portable. This paper describes an implementation of the MPEG-2 decoder using the StreamIt programming language. StreamIt is an architecture-independent stream language that aims to improve programmer productivity, while concomitantly exposing the inherent parallelism and communication topology of the application. The paper shows that MPEG is a good match for the streaming programming model and illustrates the malleability of the implementation using a simple modification to the decoder to support alternate color compression formats. StreamIt allows for modular application development, which increases code reuse, and reduces the complexity of the debugging process since stream components can be verified independently. This in turn leads to greater programmer productivity.

An Efficient and Scalable Parallel Algorithm for Out-of-Core Isosurface Extraction and Rendering

Qin Wang[1], Joseph Jaja[1] and Amitabh Varshney[2]

[1]*Dept. of Electrical and Computer Engineering*
University of Maryland
College Park, MD, USA
{*qinwang, joseph*}*@umiacs.umd.edu*

[2]*Dept. of Computer Science*
University of Maryland
College Park, MD, USA
varshney@cs.umd.edu

We consider the problem of isosurface extraction and rendering for large scale time varying data. Such datasets have been appearing at an increasing rate especially from physics-based simulations, and can range in size from hundreds of gigabytes to tens of terabytes. We develop a new simple indexing scheme, which makes use of the concepts of the interval tree and the span space data structures. The new scheme enables isosurface extraction and rendering in I/O optimal time, using more compact indexing structure and more effective bulk data movement than the previous schemes. Moreover, our indexing scheme can be easily extended to a multiprocessor environment in which each processor has access to its own local disk. The resulting parallel algorithm is provably efficient and scalable. That is, it achieves load balancing across the processors independent of the isovalue, with almost no overhead in the total amount of work relative to the sequential algorithm. We conduct a large number of experimental tests on the University of Maryland Visualization Cluster using the Richtmyer-Meshkov instability dataset, and obtain results that consistently validate the efficiency and the scalability of our algorithm.

Parallel Morphological Processing of Hyperspectral Image Data on Heterogeneous Networks of Computers

Antonio J. Plaza

Department of Computer Science
University of Extremadura
Avda. de la Universidad s/n, E-10071 Caceres, Spain
aplaza@unex.es

Recent advances in space and computer technologies are revolutionizing the way remotely sensed data is collected, managed and interpreted. The development of efficient techniques for transforming the massive amount of collected data into scientific understanding is critical for space-based Earth science and planetary exploration. Although most currently available parallel processing strategies for hyperspectral image analysis assume homogeneity in the computing platform, heterogeneous networks of computers represent a very promising cost-effective solution expected to play a major role in the design of high-performance computing platforms for many on-going and planned remote sensing missions. This paper explores techniques for mapping morphological hyperspectral analysis algorithms, characterized by their scalability and sub-pixel accuracy, onto heterogeneous parallel computers. Important aspects in algorithm design are illustrated by using both homogeneous and heterogeneous parallel computing facilities available at NASAs Goddard Space Flight Center and University of Maryland. Experiments reveal that heterogeneous networks of workstations represent a source of computational power that is both accessible and applicable in many remote sensing studies.

Acceleration of a Content-Based Image-Retrieval Application on the RDISK Cluster

Auguste Noumsi[1], Steven Derrien[2] and Patrice Quinton[3]

[1]*IRISA*
Université de Douala
Rennes, FRANCE
anoumsi@irisa.fr

[2]*IRISA*
Université de Rennes 1
Rennes, FRANCE
sderrien@irisa.fr

[3]*IRISA*
ENS Cachan
Rennes, FRANCE
quinton@irisa.fr

Because of the growing use of multimedia content over Internet, Content-Based Image Retrieval (CBIR) has recently received a lot of interest. While accurate search techniques based on local image descriptors exist, they suffer from very long execution time. We propose to accelerate CBIR on the RDISK machine, a cluster of FPGA-enhanced hard-drives, that follows the philosophy of smart-disks. Our platform combines coarse and fine grain parallelism thanks to the concurrent use of the cluster nodes and of a programmable logic device. The implementation of the CBIR application on this mixed hardware/software platform follows a strict methodology, that was validated on realistic data-set (image database of more than 30,000 images). This methodology allows us to adapt the original algorithm to suit a hardware implementation, and to select the values of some key design parameters to maximize global performance. Our preliminary results indicate that speed-ups between 120 and 200 could be obtained for a cluster of 32 nodes compared with a software implementation running on a standard desktop PC.

Session 23

RECONFIGURABLE and MULTIPLE-WIDTH SYSTEMS

Parallel FPGA-based All-Pairs Shortest-Paths in a Directed Graph

Uday Bondhugula[1], Ananth Devulapalli[2], Joseph Fernando[2], Pete Wyckoff[3] and P. Sadayappan[1]

[1]*Department of Computer Science and Engineering*
The Ohio State University
Columbus, OH, USA
{*bondhugu, saday*}*@cse.ohio-state.edu*

[2]*Ohio Supercomputer Center (Springfield)*
Springfield, OH, USA
{*ananth, fernando*}*@osc.edu*

[3]*Ohio Supercomputer Center*
Columbus, OH, USA
pw@osc.edu

With rapid advances in VLSI technology, Field Programmable Gate Arrays (FPGAs) are receiving the attention of the Parallel and High Performance Computing community. In this paper, we propose a highly parallel FPGA design for the Floyd-Warshall algorithm to solve the all-pairs shortest-paths problem in a directed graph. Our work is motivated by a computationally intensive bio-informatics application that employs this algorithm. The design we propose makes efficient and maximal utilization of the large amount of resources available on an FPGA to maximize parallelism in the presence of significant data dependences. Experimental results from a working FPGA implementation on the Cray XD1 show a speedup of 22 over execution on the XD1's processor.

Design flow for Optimizing Performance in Processor Systems with on-chip Coarse-Grain Reconfigurable Logic

Michalis D. Galanis, Gregory Dimitoulakos and Costas E. Goutis

VLSI Design Laboratory, ECE Department
University of Patras
Patras, Achaia, Greece
{*mgalanis, dhmhgre, goutis*}*@ee.upatras.gr*

A design flow for processor platforms with on-chip coarse-grain reconfigurable logic is presented. The reconfigurable logic is realized by a 2-Dimensional Array of Processing Elements. Performance is improved by accelerating critical software loops, called kernels, on the Reconfigurable Array. Basic steps of the design flow have been automated. A procedure for detecting critical loops in the input C code was developed, while a mapping technique for Coarse Grain Reconfigurable Arrays, based on software pipelining, was also devised. Analytical results derived from mapping five real-life DSP applications on eight different instances of a generic system architecture are presented. Large values of Instructions Per Cycle were achieved on two Reconfigurable Arrays that resulted in high-performance kernel mapping. Additionally, by mapping critical code on the reconfigurable logic, speedups ranging from 1.27 to 3.18 relative to an all-processor execution were achieved.

Exploring the Design Space of an Optimized Compiler Approach for Mesh-Like Coarse-Grained Reconfigurable Architectures

Gregory Dimitroulakos, Michalis D. Galanis and Costas E. Goutis

ELECTRICAL AND COMPUTER ENG. DEPARTMENT VLSI LABORATORY
UNIVERSITY OF PATRAS
RIO-PATRAS, ACHAIA, GREECE
{*dhmhgre, mgalanis, goutis*}*@ee.upatras.gr*

In this paper we study the performance improvements and trade-offs derived from an optimized mapping approach applied on a parametric coarse grained reconfigurable array architecture. The processing elements local register files and the processing elements interconnection network is exploited for caching memory data values with data reuse opportunities. The data reused values are transferred through the processing elements interconnection network hence, relieving the bus from the burden of transferring these values. A novel mapping algorithm is also proposed that uses a modulo scheduling technique. This algorithm targets on a flexible architecture template which permits experimental exploration over different architecture alternatives. The experimental results showed that the operation parallelism was significantly improved by our mapping approach. Additionally, we have outlined the relation that exists between the performance improvements and the memory access latency, the interconnection network and the processing elements register file size.

Empowering a Helper Cluster through Data-Width Aware Instruction Selection Policies

Osman S. Unsal[1], Xavier Vera[1], Antonio González[1] and Oguz Ergin[2]

[1]*Intel Barcelona Research Center*
Intel/UPC
Barcelona, Spain
{*osmanx.unsal, xavier.vera, antonio.gonzalez*}*@intel.com*

[2]*Department of Computer Engineering*
TOBB Univ. of Economics and Technology
Ankara, Turkey
oergin@etu.edu.tr

Narrow values that can be represented by less number of bits than the full machine width occur very frequently in programs. On the other hand, clustering mechanisms enable cost- and performance-effective scaling of processor back-end features. Those attributes can be combined synergistically to design special clusters operating on narrow values (a.k.a. Helper Cluster), potentially providing performance benefits.

We complement a 32-bit monolithic processor with a low-complexity 8-bit Helper Cluster. Then, in our main focus, we propose various ideas to select suitable instructions to execute in the data-width based clusters. We add data-width information as another instruction steering decision metric and introduce new data-width based selection algorithms which also consider dependency, inter-cluster communication and load imbalance. Utilizing those techniques, the performance of a wide range of workloads are substantially increased; Helper Cluster achieves an average speedup of 11% for a wide range of 412 apps. When focusing on integer applications, the speedup can be as high as 22% on average.

Session 24

PROGRAMMING ABSTRACTIONS

Algorithmic Skeletons for Stream Programming in Embedded Heterogeneous Parallel Image Processing Applications

Wouter Caarls[1], Pieter Jonker[1] and Henk Corporaal[2]

[1]*Quantitative Imaging Group*
Delft University of Technology
Delft, The Netherlands
{W.Caarls, P.P.Jonker}@tudelft.nl

[2]*Dept. of Electrical Engineering*
Eindhoven University of Technology
Eindhoven, The Netherlands
H.Corporaal@tue.nl

Algorithmic skeletons can be used to write architecture independent programs, shielding application developers from the details of a parallel implementation. In this paper, we present a C-like skeleton implementation language, PEPCI, that uses term rewriting and partial evaluation to specify skeletons for parallel C dialects. By using skeletons to control the iteration of kernel functions, we provide a stream programming language that is better tailored to the user as well as the underlying architecture. Skeleton merging allows us to reduce the overheads usually associated with breaking an application into small kernels.

We have implemented an example image processing application on a heterogeneous embedded prototype platform consisting of an SIMD and ILP processor, and show that a significant speedup can be achieved without requiring knowledge of data parallel processing.

Incrementally Developing Parallel Applications with AspectJ

Joao Luis Ferreira Sobral

Departamento de Informática
Universidade do Minho
Braga, Portugal
jls@di.uminho.pt

This paper presents a methodology to develop more modular parallel applications, based on aspect oriented programming. Traditional object oriented mechanisms implement application core functionality and parallelisation concerns are plugged by aspect oriented mechanisms. Parallelisation concerns are separated into four categories: functional or/and data partition, concurrency, distribution and optimisation. Modularising these categories into separate modules using aspect oriented programming enables (un)pluggability of parallelisation concerns. This approach leads to more incremental application development, easier debugging and increased reuse of core functionality and parallel code, when compared with traditional object oriented approaches. A detailed analysis of a simple parallel application - a prime number sieve - illustrates the methodology and shows how to accomplish these gains.

Auto-Pipe and the X Language: A Pipeline Design Tool and Description Language

Mark A. Franklin[1], Eric J. Tyson[2], James Buckley[3] and Patrick Crowley[4]

[1]*Department of Computer Science and Engineering*
Washington University
St. Louis, MO, USA
jbf@cse.wustl.edu

[2]*Department of Computer Science and Engineering*
Washington University
St. Louis, MO, USA
etyson@wustl.edu

[3]*Department of Physics*
Washington University
St. Louis, MO, USA
buckley@wuphys.wustl.edu

[4]*Department of Computer Science and Engineering*
Washington University
St. Louis, MO, USA
pcrowley@wustl.edu

Auto-Pipe is a tool that aids in the design, evaluation and implementation of applications that can be executed on computational pipelines (and other topologies) using a set of heterogeneous devices including multiple processors and FPGAs. It has been developed to meet the needs arising in the domains of communications, computation on large datasets, and real time streaming data applications. This paper introduces the *Auto-Pipe* design flow and the *X* design language, and presents sample applications. The applications include the Triple-DES encryption standard and a subset of the signal-processing pipeline for VERITAS, a high-energy gamma-ray astrophysics experiment. These applications are discussed and their description in *X* is presented. From the *X* description, simulations of alternative system designs and stage-to-device assignments are obtained and analyzed, and the optimal assignment is presented. The complete system will permit production of executable code and bit maps that may be downloaded onto real devices. Future work required to complete the *Auto-Pipe* design tool is discussed.

Enabling Efficient and Flexible Coupling of Parallel Scientific Applications

Li Zhang and Manish Parashar

The Applied Software Systems Laboratory (TASSL)
Rutgers University
Piscataway, NJ, U.S.A.
{*emmalily, parashar*}*@caipclassic.rutgers.edu*

Emerging scientific and engineering simulations are presenting challenging requirements for coupling between multiple physics models and associated parallel codes that execute independently and in a distributed manner. Realizing coupled simulations requires an efficient, flexible and scalable coupling framework and simple programming abstractions. This paper presents a coupling framework that addresses these requirements. The framework is based on the Seine geometry-based interaction model. It enables efficient computation of communication schedules, supports low-overheads processor-to-processor data streaming, and provides high-level abstraction for application developers. The design, CCA-based implementation, and experimental evaluation of the Seine based coupling framework are presented.

Session 25

RESOURCE ALLOCATION

Skewed Allocation of Non-Uniform Data for Broadcasting over Multiple Channels

A.a. Bertossi[1] and C.m. Pinotti[2]

[1]*Dept. of Computer Science*
University of Bologna
Bologna, Italy
bertossi@cs.unibo.it

[2]*Dept. of Math & Computer Science*
University of Perugia
Perugia, Italy
pinotti@unipg.it

The problem of data broadcasting over multiple channels consists in partitioning data among channels, depending on data popularities, and then cyclically transmitting them over each channel so that the average waiting time of the clients is minimized. Such a problem is known to be polynomially time solvable for uniform length data items, while it is computationally intractable for non-uniform length data items. In this paper, two new heuristics are proposed which exploit a novel characterization of optimal solutions for the special case of two channels and data items of uniform lengths. Sub-optimal solutions for the most general case of an arbitrary number of channels and data items of non-uniform lengths are provided. The first heuristic, called Greedy+, combines the novel characterization with the known greedy approach, while the second heuristic, called Dlinear, combines the same characterization with the dynamic programming technique. Such heuristics have been tested on benchmarks whose popularities are characterized by Zipf distributions. The experimental tests reveal that Dlinear finds optimal solutions almost always, requiring good running times, while Greedy+ is faster and scales well when changes occur on the input parameters, but provides worse solutions than Dlinear.

Comparative Study of Price-based Resource Allocation Algorithms for Ad Hoc Networks

Marcel Luethi, Simin Nadjm-tehrani and Calin Curescu

Department of Computer and Information Science
Linkoping
Linkoping, Sweden
{g-marlu, simin, calcu}@ida.liu.se

As mobile ad hoc networks provide a wide range of possibly critical services, providing quality of service guarantees becomes an essential element. Yet there is a limited understanding of the performance characteristics of different resource allocation algorithms. In particular, there is little work that comparatively studies different algorithms in the same traffic environment. Therefore we study two algorithms, adhoc-TARA and an algorithm based on the gradient projection method, for optimised bandwidth allocation in ad hoc networks under overload situations. The focus is on convergence properties and performance measured in terms of accumulated utility. The simulation results show that the gradient projection algorithm converges to an optimal solution even in large, dynamic networks, but that in such dynamic environments the convergence time can significantly influence the overall performance. In comparison, the near-optimal algorithm adhoc-TARA, which quickly adapts to changes in the state of the network, can exhibit superior performance. Further we illustrate how different parameter settings influence the performance of the algorithms. We conclude that finding an optimal allocation comes at a high price in the rapidly changing environments of ad hoc networks and that near-optimal allocation can be an ample alternative.

Oblivious Parallel Probabilistic Channel Utilization without Control Channels

Christian Schindelhauer[1] and Kerstin Voss[2]

[1]*Heinz Nixdorf Institute*
Paderborn University
Paderborn, 33102
schindel@upb.de

[2]*Parallel Center for Parallel Computing*
Paderborn University
Paderborn, 33102
kerstinv@upb.de

The research interest in sensor nets is still growing because they simplify data acquisition in many applications. If hardware resources are very sparse, routing algorithms cannot use data gathering. However, if a large number of channels can be used, then parallel transmission can compensate this drawback. If the senders and receivers are not known in advance, then a control channel poses a bottleneck for communication. We present an oblivious MAC protocol, called the Funnel protocol, where the channels are nearly optimally utilized in parallel. In this, senders and receivers choose for a polylogarithmic number of rounds (several sending attempts) a decreasing number of channels which are selected equiprobably.

Then, we show that a previously presented approach using only one round and therefore one type of probability distribution is optimal up to some constant factor, and considerably worse than the Funnel protocol. The protocol works with few resources if an sufficient number of channels is available. The Funnel protocol is simple, elegant, and does not need to know the number of senders and receivers, thus being oblivious.

On the bottom line we prove that small messages can be efficiently transmitted by the MAC layer in parallel without a control channel if more than one channel for communication can be used.

Non-cooperative, Semi-cooperative, and Cooperative Games-based Grid Resource Allocation

Samee Ullah Khan and Ishfaq Ahmad

Department of Computer Science and Engineering
University of Texas at Arlington
Arlington, Texas, USA
{sakhan, iahmad}@cse.uta.edu

In this paper we consider, compare and analyze three game theoretical Grid resource allocation mechanisms. Namely, 1) the non-cooperative sealed-bid method where tasks are auctioned off to the highest bidder, 2) the semi-cooperative n-round sealed-bid method in which each site delegate its work to others if it cannot perform the work itself, and 3) the cooperative method in which all of the sites deliberate with one another to execute all the tasks as efficiently as possible.

To experimentally evaluate the above mentioned techniques, we perform extensive simulation studies that effectively encapsulate the task and machine heterogeneity. The tasks are assumed to be independent and bear multiple execution time deadlines. The simulation model is built around a hierarchical Grid infrastructure where machines are abstracted into larger computing centers labeled "federations," each of which are responsible for managing their own resources independently. These federations are then linked together with a primary portal to which Grid tasks would be submitted. To measure the effectiveness of these game theoretical techniques, the recorded performance is evaluated against a conventional baseline method in which tasks are randomly assigned to the sites without any task execution guarantee.

Session 26

PARTITIONING and REFINEMENT

Parallel Hypergraph Partitioning for Scientific Computing

Karen D Devine[1], Erik G Boman[1], Robert T Heaphy[1], Rob H Bisseling[2] and Umit V Catalyurek[3]

[1]*Discrete Algorithms and Math.*
Sandia National Labs
Albuquerque, NM, USA
{*kddevin, egboman, rheaphy*}*@sandia.gov*

[2]*Dept. of Mathematics*
Utrecht University
Utrecht, the Netherlands
rob.bisseling@math.uu.nl

[3]*Dept. of Biomedical Informatics*
Ohio State Univ.
Columbus, OH, USA
umit@bmi.osu.edu

Graph partitioning is often used for load balancing in parallel computing, but it is known that hypergraph partitioning has several advantages. First, hypergraphs more accurately model communication volume, and second, they are more expressive and can better represent nonsymmetric problems. Hypergraph partitioning is particularly suited to parallel sparse matrix-vector multiplication, a common kernel in scientific computing. We present a parallel software package for hypergraph (and sparse matrix) partitioning developed at Sandia National Labs. The algorithm is a variation on multilevel partitioning. Our parallel implementation is novel in that it uses a two-dimensional data distribution among processors. We present empirical results that show our parallel implementation achieves good speedup on several large problems (up to 33 million nonzeros) with up to 64 processors on a Linux cluster.

Multilevel Algorithms for Partitioning Power-Law Graphs

Amine Abou-rjeili and George Karypis

Department of Computer Science & Engineering, Army HPC Research Center, & Digital Technology Center
University of Minnesota
Minneapolis, MN, USA
{*amin, karypis*}*@cs.umn.edu*

Graph partitioning is an enabling technology for parallel processing as it allows for the effective decomposition of unstructured computations whose data dependencies correspond to a large sparse and irregular graph. Even though the problem of computing high-quality partitionings of graphs arising in scientific computations is to a large extent well-understood, this is far from being true for emerging HPC applications whose underlying computation involves graphs whose degree distribution follows a power-law curve. This paper presents new multilevel graph partitioning algorithms that are specifically designed for partitioning such graphs. It presents new clustering-based coarsening schemes that identify and collapse together groups of vertices that are highly connected. An experimental evaluation of these schemes on 10 different graphs show that the proposed algorithms consistently and significantly outperform existing state-of-the-art approaches.

Effective Out-of-Core Parallel Delaunay Mesh Refinement using Off-the-Shelf Software

Andriy Kot, Andrey Chernikov and Nikos Chrisochoides

Computer Science Department
The College of William and Mary
Williamsburg, VA, USA
{*kot, ancher, nikos*}*@cs.wm.edu*

We present two cost-effective and high-performance out-of-core parallel mesh generation algorithms and their implementation on Cluster of Workstations (CoWs). The total wall-clock time including wait-in-queue delays for the out-of-core methods on a small cluster (16 processors) is three times shorter than the total wall-clock time for the in-core generation of the same size mesh (about a billion elements) using 121 processors. Our best out-of-core method, for mesh sizes that fit completely in the core of the CoWs, is about 5% slower than its in-core parallel counterpart method. This is a modest performance penalty for savings of many hours in response time. Both the in-core and out-of-core methods use the best publicly available off-the-shelf sequential in-core Delaunay mesh generator.

Fast Distributed Graph Partition and Application (Extended Abstract)

Bilel Derbel, Mohamed Mosbah and Akka Zemmari

LaBRI
University Bordeaux 1
33405 Talence, France
{*derbel, mosbah, zemmari*}*@labri.fr*

This paper presents efficient deterministic and randomized distributed algorithms for decomposing a graph with n nodes into a disjoint set of connected clusters with small radius and few intercluster edges. Our algorithms can be easily implemented in the distributed $\mathcal{CONGEST}$ model of computation i.e., limited message size, improving the time complexity of previous algorithms from linear to sublinear. One important application of our algorithms is efficient construction of sparse graph spanners. In fact, given a parameter k, we show that there exists a sublinear deterministic distributed algorithm that constructs a graph spanner of stretch $2k-1$ with at most $\mathcal{O}(n^{1+1/k})$ edges in the $\mathcal{CONGEST}$ model.

Session 27

COLLECTIVE COMMUNICATION

Application-Oriented Adaptive MPIBcast for Grids

Rakhi Gupta and Sathish Vadhiyar

Supercomputer Education and Research Centre
Indian Institute of Science
Bangalore, Karnataka, India
rakhi@rishi.serc.iisc.ernet.in, vss@serc.iisc.ernet.in

Due to the importance of collective communications in scientific parallel applications, many strategies have been devised for optimizing collective communications for different kinds of parallel environments. Recently, there has been an increasing interest to evolve efficient broadcast algorithms for computational Grids. In this paper, we present application-oriented adaptive techniques that take into account recent resource characteristics as well as the application's usage of broadcasts for deriving efficient broadcast trees. In particular, we consider two broadcast parameters used in the application, namely, the broadcast message sizes and the time interval between the broadcasts. The results indicate that our adaptive strategies can provide 20% average improvement in performance over the popular MPICH-G2's MPI_Bcast implementation for loaded network conditions.

Pipelined Broadcast on Ethernet Switched Clusters

Pitch Patarasuk[1], Ahmad Faraj[2] and Xin Yuan[3]

[1]*Dept. of Computer Science*
Florida State University
Tallahassee, FL, United State
patarasu@cs.fsu.edu

[2]*Dept. of Computer Science*
Florida State University
Tallahassee, FL, United State
faraj@cs.fsu.edu

[3]*Dept. of Computer Science*
Florida State University
Tallahassee, FL, United State
xyuan@cs.fsu.edu

We consider unicast-based pipelined broadcast schemes for clusters connected by multiple Ethernet switches. By splitting a large broadcast message into segments and broadcasting the segments in a pipelined fashion, pipelined broadcast may achieve very high performance. We develop algorithms for computing various contention-free broadcast trees on Ethernet switched clusters that are suitable for pipelined broadcast, and evaluate the schemes through experimentation. The conclusions drawn from our theoretical and experimental study include the following. First, pipelined broadcast can be more effective than other common broadcast schemes including the ones used in the latest versions of MPICH and LAM/MPI when the message size is sufficiently large. Second, contention-free broadcast trees are essential for pipelined broadcast to achieve high performance. Finally, while it is difficult to determine the optimal message segment size for pipelined broadcast, finding one size that gives good performance is relatively easy.

k-anycast Routing Schemes for Mobile Ad Hoc Networks

Bing Wu and Jie Wu

Computer Science and Engineering
Florida Atlantic University
Boca Raton, FL, USA
bing.florida@gmail.com, jie@cse.fau.edu

Anycast is a communication paradigm that was first introduced to the suit of routing protocols in IPv6 networks. In anycast, a packet is intended to be delivered to one of the nearest group hosts. k-anycast, however, is proposed to deliver a packet to any threshold k members of a set of hosts. In this paper, we propose three k-anycast routing schemes for mobile ad hoc networks. Our research work is motivated by the distributed key management services using threshold cryptography in mobile ad hoc networks in which the certification authority's functionality is distributed to any k servers. However, security is not the main focus of this paper. Our goal is to reduce the routing control messages and network delay to reach any k servers. The first scheme is called controlled flooding. The increase of flooding radius is based on the number of responses instead of increasing radius linearly or exponentially. The second scheme, called component-based scheme I, is to form multiple components such that each component has at least k members. We can treat each component as a virtual server as in anycast, thus, we simplify the k-anycast routing problem into an anycast routing problem. For the highly dynamic network environment, we introduce the third scheme, called component-based scheme II, in which the membership a component maintains is relaxed to be less than k. The performances of the proposed schemes are evaluated through simulations.

DVoDP2P: Distributed P2P Assisted Multicast VoD Architecture

Xiaoyuan Yang[1], Porfidio Hernández[1], Fernando Cores[2], Leandro Souza[1], Ana Ripoll[1], Remo Suppi[1] and Emilio Luque[1]

[1]*Universitat Autònoma de Barcelona*
ETSE, Computer Science Department
Barcelona, Spain
xiao.yuan@aomail.uab.es

[2]*Universitat de Lleida*
Computer Science & Industrial Engineering, EPS
25001-Lleida, Spain
fcores@diei.udl.es

For a high scalable VoD system, the distributed server architecture (DVoD) with more than one server-node is a cost-effective design solution. However, such a design is highly vulnerable to workload variations because the service capacity is limited. In this paper, we propose a new and efficient VoD architecture that combines DVoD with a P2P system. The DVoD's server-nodes is able to offer a minimum required quality of service (QoS) and the P2P system is able to provide the mechanism to increase the system service capacity according to client demands. Our P2P system is able to synchronize a group of clients in order to create multicast channels in local networks to replace server-nodes in the delivery process. Our client collaboration scheme is designed to take into account the P2P system's efficiency and the network overhead. We compared the new VoD architecture with DVoD architecture based on classic multicast and P2P delivery policies (Patching and Chaining). The experimental results showed that our design is better than previous solutions in terms of server-node load, inter-connection network load, local-network overhead and scalability. Compared with the multicast-DVoD, our architecture reduced server-load by up to 37%.

Session 28

DISTRIBUTED COORDINATION

Composite Abortable Locks

Virendra J. Marathe[1], Mark Moir[2] and Nir Shavit[2]

[1]Computer Science
Univ of Rochester
Rochester, NY, USA
vmarathe@cs.rochester.edu

[2]Sun Microsystems Labs
Burlington, MA, USA
{mark.moir, nir.shavit}@sun.com

The need to allow threads to abort an attempt to acquire a lock (sometimes called a timeout) is an interesting new requirement driven by state-of-the-art database applications with soft real-time constraints. This paper presents a new *composite abortable lock* (CAL), a combination of abortable queue-based (QL) and test-and-set based backoff (BL) lock mechanisms, which provides non-blocking aborts while ensuring low space requirements without need for a memory reclamation scheme. The key observation motivating our approach is that the fast lock hand-off achieved by QLs only requires the first few threads to be queued (not *all* waiting threads), and that the remaining threads can run as in a BL. We developed an algorithm that uses only a short fixed size structure for queueing, allowing most threads to back-off. This reduces worst-case space overhead dramatically, and improves performance by eliminating the need for expensive and complicated memory management mechanisms.

Experimental results show that our new CAL algorithm not only saves on space, it actually outperforms Scott's state-of-the-art nonblocking abortable QL under contention, and even more so when there are more threads than processors. Moreover, as the rate of lock aborts increases, the CAL continues to perform well, while Scott's algorithm deteriorates rapidly.

Cooperative Checkpointing Theory

Adam J. Oliner[1], Larry Rudolph[2] and Ramendra K. Sahoo[3]

[1]Computer Science Department
Stanford University
Palo Alto, CA, United States
oliner@cs.stanford.edu

[2]Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, United States
rudolph@csail.mit.edu

[3]IBM T. J. Watson Research Center
Hawthorne, NY, United States
rsahoo@us.ibm.com

Cooperative checkpointing uses global knowledge of the state and health of the machine to improve performance and reliability by dynamically deciding when to skip checkpoint requests made by applications. Using results from cooperative checkpointing theory, this paper proves that periodic checkpointing is not expected to be competitive with the offline optimal. By leveraging probabilistic information about the future, cooperative checkpointing gives flexible algorithms that are optimally competitive. The results prove that simulating periodic checkpointing, by performing only every d^{th} checkpoint, is not competitive with the offline optimal in the worst case; a simple modification gives a provably competitive algorithm. Calculations using failure traces from a prototype of IBM's Blue Gene/L show an application using cooperative checkpointing may make progress 4 times faster than one using periodic checkpointing, under realistic conditions. We contribute an approach to providing large-scale system reliability through cooperative checkpointing and techniques for analyzing the approach.

Structural and Algorithmic Issues of Dynamic Protocol Update

Olivier Rütti[1], Pawel T. Wojciechowski[2] and André Schiper[1]

[1]*School of Computer and Communication Sciences*
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
{*olivier.rutti, andre.schiper*}*@epfl.ch*

[2]*Institute of Computing Science*
Poznan University of Technology
60-965 Poznan, Poland
ptw@cs.put.poznan.pl

In this paper, we study dynamic protocol update (DPU). Contrary to local code updates on-the-fly, DPU requires global coordination of local code replacements. We propose a novel solution to DPU. The key idea is to add a level of indirection between the service callers and the service provider. This indirection level facilitates an implementation of simple and efficient algorithms for DPU. For example, we describe an experimental implementation of adaptive group communication middleware. It can switch between different atomic broadcast protocols on-the-fly. All middleware protocols, including those that depend on the updated protocols, provide service correctly and with negligible delay while the global update takes places. The switching algorithm introduces very low overhead that we illustrate by showing example measurement results.

On Efficient Distributed Deadlock Avoidance for Real-Time and Embedded Systems

Cesar Sanchez[1], Henny B. Sipma[1], Zohar Manna[1], Venkita Subramonian[2] and Christopher Gill[2]

[1]*Computer Science Dept.*
Stanford University
Stanford, CA, USA
{*cesar, sipma, zm*}*@CS.Stanford.EDU*

[2]*Dept. of Computer Science and Engineering*
Washington University
St. Louis, MO, USA
{*venkita, cdgill*}*@CSE.wustl.EDU*

Thread allocation is an important problem in distributed real-time and embedded (DRE) systems. A thread allocation policy that is too liberal may cause deadlock, while a policy that is too conservative limits potential parallelism, thus wasting resources. However, achieving (globally) optimal thread utilization, while avoiding deadlock, has been proven impractical in distributed systems: it requires too much communication between components.

In previous work we showed that efficient local thread allocation protocols are possible if the protocols are parameterized by global static data, in particular by an annotation of the global call graph of all tasks to be performed by the system. We proved that absence of cyclic dependencies in this annotation guarantees absence of deadlock.

In this paper we present an algorithm to compute optimal annotations, that is annotations that maximize parallelism while satisfying the condition of acyclicity. Moreover, we show that the condition of acyclicity is in fact tight and exhibits a rather surprising anomaly: if a cyclic dependency is present in the annotation of the call graph and a certain minimum number of threads is provided, deadlock is reachable. Thus, in the presence of cyclic dependencies, increasing the number of threads may introduce the possibility of deadlock in an originally deadlock free system.

Session 29

SYMBOLIC COMPUTING APPLICATIONS

A Dynamic Firing Speculation to Speedup Distributed Symbolic State-space Generation

Ming-ying Chung and Gianfranco Ciardo

Department of Computer Science and Engineering
University of California, Riverside
Riverside, CA, USA
{*chung, ciardo*}*@cs.ucr.edu*

The *saturation* strategy for symbolic state-space generation is very effective for globally-asynchronous locally-synchronous discrete-state systems. Its inherently sequential nature, however, makes it difficult to parallelize on a NOW. An initial attempt that utilizes idle workstations to recognize event firing patterns and then speculatively compute firings conforming to these patterns is at times effective but can introduce large memory overheads. We suggest an implicit method to encode the firing history of decision diagram nodes, where patterns can be shared by nodes. By preserving the actual firing history efficiently and effectively, the speculation is more informed. Experiments show that our implicit encoding method not only reduces the memory requirements but also enables dynamic speculation schemes that further improve runtime.

Parallelizing Post-Placement Timing Optimization

Jiyoun Kim[1], Marios C. Papaefthymiou[1] and Jose L. Neves[2]

[1]*EECS/ACAL*
University of Michigan
Ann Arbor, MI, USA
{*jiyoun, marios*}*@eecs.umich.edu*

[2]*IBM Server Group*
Poughkeepsie, NY, USA
jneves@us.ibm.com

This paper presents an efficient modeling scheme and a partitioning heuristic for parallelizing VLSI post-placement timing optimization. Encoding the paths with timing violations into a task graph, our novel modeling scheme provides an efficient representation of the timing and spatial relations among timing optimization tasks. Our new partitioning algorithm then assigns the task graph into multiple sessions of parallel processes, so that interprocessor communication is completely eliminated during each session. This partitioning scheme is especially useful for parallelizing processes with heavily connected tasks and, therefore, high communication requirements. For circuits with 20–130 thousand cells, the partitioning heuristic achieves speedups in excess of 5× without degrading solution quality by dynamically utilizing 1–8 processors.

Sim-X: Parallel System Software for Interactive Multi-Experiment Computational Studies

Siu-man Yau[1], Eitan Grinspun[2], Vijay Karamcheti[1] and Denis Zorin[1]

[1]*Courant Institute of Mathematical Sciences*
New York University
New York, NY, USA
{*smyau, vijayk, dzorin*}*@cs.nyu.edu*

[2]*Department of Computer Science*
Columbia University
New York, NY, USA
eitan@cs.columbia.edu

Advances in high-performance computing have led to the broad use of *computational studies* in everyday engineering and scientific applications. A single study may require thousands of *computational experiments*, each corresponding to individual runs of simulation software with different parameter settings; in complex studies, the pattern of parameter changes is complex and may have to be adjusted by the user based on partial simulation results. Unfortunately, existing tools have limited high-level support for managing large ensembles of simultaneous computational experiments.

In this paper, we present a system architecture for interactive computational studies targeting two goals. The first is to provide a framework for high-level user interaction with computational studies, rather than individual experiments; the second is to maximize the size of the studies that can be performed at close to interactive rates.

We describe a prototype implementation of the system and demonstrate performance improvements obtained using our approach for a simple model problem.

Session 30

MULTITHREADING

Exploiting Unbalanced Thread Scheduling for Energy and Performance on a CMP of SMT Processors

Matthew Devuyst, Rakesh Kumar and Dean M. Tullsen

Computer Science and Engineering
University of California, San Diego
San Diego, CA, USA
{*mdevuyst, rakumar, tullsen*}*@ucsd.edu*

This paper explores thread scheduling on an increasingly popular architecture: chip multiprocessors with simultaneous multithreading cores. Conventional multiprocessor scheduling, applied to this architecture, will attempt to balance the thread load across cores. This research demonstrates that such an approach eliminates one of the big advantages of this architecture – the ability to use unbalanced schedules to allocate the right amount of execution resources to each thread. However, accommodating unbalanced schedules creates several difficulties, the biggest being the fact that the search space of all schedules (both balanced and unbalanced) is much greater than that of the balanced schedules alone. This work proposes and evaluates scheduling policies that allow the system to identify and migrate toward good thread schedules, whether the best schedules are balanced or unbalanced.

Helper Thread Prefetching for Loosely-Coupled Multiprocessor Systems

Changhee Jung[1], Daeseob Lim[2], Jaejin Lee[3] and Yan Solihin[4]

[1]*Embedded Software Research Division*
Electronics and Telecommunications Research Institute
Daejeon, 305-530, Korea
chjung@etri.re.kr

[2]*Department of Computer Science and Engineering*
University of California
San Diego, CA 92093, USA
dalim@cse.ucsd.edu

[3]*School of Computer Science and Engineering*
Seoul National University
Seoul, 151-744, Korea
jlee@cse.snu.ac.kr

[4]*Department of Electrical and Computer Engineering*
North Carolina State University
Raleigh, NC 27695, USA
solihin@eos.ncsu.edu

This paper presents a helper thread prefetching scheme that is designed to work on loosely-coupled processors, such as in a standard chip multiprocessor (CMP) system or an intelligent memory system. Loosely-coupled processors have an advantage in that fine-grain resources, such as processor and L1 cache resources, are not contended by the application and helper threads, hence preserving the speed of the application. However, interprocessor communication is expensive in such a system. We present techniques to alleviate this. Our approach exploits large loop-based code regions and is based on a new synchronization mechanism between the application and helper threads. This mechanism precisely controls how far ahead the execution of the helper thread can be with respect to the application thread. We found that this is important in ensuring prefetching timeliness and avoiding cache pollution. To demonstrate that prefetching in a loosely-coupled system can be done effectively, we evaluate our prefetching in a standard, unmodified CMP system, and in an intelligent memory system where a simple processor in memory executes the helper thread. Evaluating our scheme with nine memory-intensive applications with the memory processor in DRAM achieves an average speedup of 1.25. Moreover, our scheme works well in combination with a conventional processorside sequential L1 prefetcher, resulting in an average speedup of 1.31. In a standard CMP, the scheme achieves an average speedup of 1.33.

Compatible Phase Co-Scheduling on a CMP of Multi-Threaded Processors

Ali El-moursy[1], Rajeev Garg[2], David H. Albonesi[3] and Sandhya Dwarkadas[2]

[1]*Electrical & Computer Engineering*
University of Rochester
Rochester, NY, USA
elmours@ece.rochester.edu

[2]*Computer Science*
University of Rochester
Rochester, NY, USA
{garg, sandhya}@cs.rochester.edu

[3]*Computer Systems Laboratory*
Cornell University
Ithaca, NY, USA
albonesi@csl.cornell.edu

The industry is rapidly moving towards the adoption of Chip Multi-Processors (CMPs) of Simultaneous Multi-Threaded (SMT) cores for general purpose systems. The most prominent use of such processors, at least in the near term, will be as job servers running multiple independent threads on the different contexts of the various SMT cores. In such an environment, the co-scheduling of phases from different threads plays a significant role in the overall throughput. Less throughput is achieved when phases from different threads that conflict for particular hardware resources are scheduled together, compared with the situation where compatible phases are co-scheduled on the same SMT core. Achieving the latter requires precise per-phase hardware statistics that the scheduler can use to rapidly identify possible incompatibilities among phases of different threads, thereby avoiding the potentially high performance cost of inter-thread contention.

In this paper, we devise phase co-scheduling policies for a dual-core CMP of dual-threaded SMT processors. We explore a number of approaches and find that the use of ready and in-flight instruction metrics permits effective co-scheduling of compatible phases among the four contexts. This approach significantly outperforms the worst static grouping of threads, and very closely matches the best static grouping, even outperforming it by as much as 7%.

Session 31

RUNTIME OPTIMIZATIONS

Selecting the Tile Shape to Reduce the Total Communication Volume

Nikolaos Drosinos, Georgios Goumas and Nectarios Koziris

School of Electrical and Computer Engineering
National Technical University of Athens
Zografou, Greece
{*ndros, goumas, nkoziris*}*@cslab.ece.ntua.gr*

In this paper we revisit the tile-shape selection problem, that has been extensively discussed in bibliography. An efficient approach is proposed for the selection of a suitable tile shape, based on the minimization of the process communication volume. We consider the large family of applications that arise from the discretization of partial differential equations (PDEs). Practical experience has shown that for such applications and distributed memory architectures, minimizing the total communication volume is more important than minimizing the total number of parallel execution steps. We formulate a new method to determine an appropriate communication-aware tile shape, i.e. the one that reduces the communication volume for a fixed number of processes. Our approach is equivalent to defining a proper Cartesian process grid with MPI_Cart_Create, which means that it can be incorporated in applications in a straightforward manner. Our experimental results illustrate that by selecting the tile shape with the proposed method, the total parallel execution time is significantly reduced due to the minimization of the communication volume, despite the fact that a few more parallel execution steps are required.

Application Classification through Monitoring and Learning of Resource Consumption Patterns

Jian Zhang and Renato Figueiredo

Electrical & Computer Engineering / ACIS Lab
University of Florida
Gainesville, FL, USA
{*jianzh, renato*}*@acis.ufl.edu*

Application awareness is an important factor of efficient resource scheduling. This paper introduces a novel approach for application classification based on the Principal Component Analysis (PCA) and the k-Nearest Neighbor (k-NN) classifier. This approach is used to assist scheduling in heterogeneous computing environments. It helps to reduce the dimensionality of the performance feature space and classify applications based on extracted features. The classification considers four dimensions: CPU-intensive, I/O and paging-intensive, network-intensive, and idle. Application class information and the statistical abstracts of the application behavior are learned over historical runs and used to assist multi-dimensional resource scheduling. This paper describes a prototype classifier for application-centric Virtual Machines. Experimental results show that scheduling decisions made with the assistance of the application class information, improved system throughput by 22.11% on average, for a set of three benchmark applications.

Topology-aware Task Mapping for Reducing Communication Contention on Large Parallel Machines

Tarun Agarwal, Amit Sharma and Laxmikant V. Kale

Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, USA
tarun.agarwal@microsoft.com, asharma6@uiuc.edu, kale@cs.uiuc.edu

Communication latencies constitute a significant factor in the performance of parallel applications. With techniques such as wormhole routing, the variation in no-load latencies became insignificant, i.e., the no-load latencies for far-away processors were not significantly higher (and too small to matter) than those for nearby processors. Contention in the network is then left as the major factor affecting latencies. With networks such as Fat-Trees of hypercubes, with number of wires growing as $P \log P$, even this is not a very significant factor. However, for torus and grid networks now being used in large machines such as BlueGene/L and the Cray XT3, such contention becomes an issue. We quantify the effect of this contention with benchmarks that vary the number of hops traveled by each communicated byte. We then demonstrate a process mapping strategy that minimizes the impact of topology by heuristically minimizing the total number of hop-bytes communicated. This strategy, and its variants, are implemented in an adaptive runtime system in Charm++ and Adaptive MPI, so it is available for a broad class of applications.

Session 32

DISTRIBUTED SYSTEMS

A Virtual Network (ViNe) Architecture for Grid Computing

Maurício Tsugawa and José A. B. Fortes

Dept. of Electrical and Computer Engineering, ACIS Laboratory
University of Florida
Gainesville, FL, USA
{*tsugawa, fortes*}*@ufl.edu*

This paper describes a virtual networking approach for Grids called ViNe. It enables symmetric connectivity among Grid resources and allows existing applications to run unmodified. Novel features of the ViNe architecture include: easy virtual networking administration; support for physical private networks and support for multiple independent virtual networks in the same infrastructure. The requirements of an application-friendly virtual network environment are presented and it is shown how the proposed solution meets them. Qualitative arguments are provided to justify all design decisions. Also presented is an experimental evaluation of the round-trip latencies and bandwidths achieved by a reference implementation. Measurements are reported for WAN-scenarios involving three different institutions. Under favorable conditions, ViNe bandwidths are within 90 to 100% of the available physical network bandwidth.

Wire-Speed Total Order

Tal Anker[1,2], Greegory Greenman[2], Danny Dolev[2] and Ilya Shnayderman[2]

[1] *Radlan*
A Marvell Company
Tel-Aviv, Israel
tala@marvell.com

[2] *The School of Engineering and Computer Science*
The Hebrew University of Jerusalem
Jerusalem, Israel
{*gregory, dolev, ilia*}*@cs.huji.ac.il*

Many distributed systems may be limited in their performance by the number of transactions they are able to support per unit of time. In order to achieve fault tolerance and to boost a system's performance, active state machine replication is frequently used. It employs total ordering service to keep the state of replicas synchronized. In this paper, we present an architecture that enables a drastic increase in the number of ordered transactions in a cluster, using off-the-shelf network equipment. Performance supporting nearly one million ordered transactions per second has been achieved, which substantiates our claim.

Free Network Measurement For Adaptive Virtualized Distributed Computing

Ashish Gupta[1], Marcia Zangrilli[2], Ananth I. Sundararaj[1], Anne I. Huang[2], Peter A. Dinda[1] and Bruce B. Lowekamp[2]

[1]*Dept. of Electrical Engineering and Computer Science*
Northwestern University
Evanston, IL, USA
{*ashish, ais, pdinda*}*@cs.northwestern.edu*

[2]*Computer Science Department*
College of William and Mary
Williamsburg, VA, USA
{*mazang, lowekamp*}*@cs.wm.edu, anne.huang@yale.edu*

An execution environment consisting of virtual machines (VMs) interconnected with a virtual overlay network can use the naturally occurring traffic of an existing, unmodified application running in the VMs to measure the underlying physical network. Based on these characterizations, and characterizations of the application's own communication topology, the execution environment can optimize the execution of the application using application-independent means such as VM migration and overlay topology changes. In this paper we demonstrate the feasibility of such free automatic network measurement by fusing the Wren passive monitoring and analysis system with Virtuoso's virtual networking system. We explain how Wren has been extended to support online analysis, and we explain how Virtuoso's adaptation algorithms have been enhanced to use Wren's physical network level information to choose VM-to-host mappings, overlay topology, and forwarding rules.

Workshop 1

Heterogeneous Computing Workshop HCW 2006

HCW 2006 is sponsored by the U.S. Office of Naval Research and by the IEEE Computer Society, through the Technical Committee on Parallel Processing (TCPP).

The 15th Heterogeneous Computing Workshop (HCW'2006)

Heterogeneous computing systems are those with a range of diverse computing resources that can be local to one another or geographically distributed. The pervasive use of networks and the Internet by all segments of modern society means that the number of connected computing resources is growing tremendously. Hence, the opportunity and need for heterogeneous computing systems to effectively utilize these resources in new, novel ways is growing concomitantly. This has given rise, for instance, to the notions of cluster computing, grid computing, and peer-to-peer computing. The effective implementation of efficient applications in these environments, however, requires that a host of issues be addressed that simply do not occur in "single-chassis" sequential or parallel machines. Thus, the topics of interest concerning heterogeneous systems and environments include, but are not limited to: programming paradigms and tools, resource discovery and management, task and communication scheduling, task coordination and workflow, performance management, heterogeneous cluster computing, grid computing, peer-to-peer computing, adaptive computing, ubiquitous computing, mobile computing, real-time distributed systems, security, fault tolerance, and application case studies.

General Chair:
Arnold L. Rosenberg
University of Massachusetts
Email: rsnbrg@cs.umass.edu

Program Chair:
José A. B. Fortes
University of Florida
Email: forteshcw@acis.ufl.edu

Message from the Heterogeneous Computing Workshop Steering Committee Chair

These are the proceedings of the 15th Heterogeneous Computing Workshop, also known as HCW 2006. Heterogeneous computing is a very important research area with great practical impact, and it includes a large range of systems. A heterogeneous computing system may be a set of machines interconnected by a wide-area network and used to support the execution of jobs submitted by a large variety of users to process data that is distributed throughout the system. It may be a suite of high-performance machines tightly interconnected by a fast, dedicated network and used to process a set of production tasks, where the communicating subtasks of each task may execute on different machines in the suite. It may also be a special-purpose embedded system, such as a set of different types of processors working together to perform a particular application. In one extreme, it may consist of a single machine that can reconfigure itself to operate in different ways (e.g., in different modes of parallelism). All of these types of heterogeneous computing systems (as well as others, e.g., grids and clusters) are appropriate topics for this workshop series. I hope you find the contents of these proceedings informative and interesting, and I encourage you to look also at the proceedings of past and future Heterogeneous Computing Workshops.

Many people have worked very hard to make this year's workshop happen. Jose A. B. Fortes, from the University of Florida, was this year's Program Committee Chair, and he assembled the excellent program for the workshop and collection of papers in these workshop proceedings. Jose did this with the assistance of his Program Committee, which is listed in these proceedings. Arnold L. Rosenberg, from the University of Massachusetts Amherst, was the General Chair. Arny was responsible for the overall organization and administration of this year's workshop, and he did a fine job. I thank Jose, Arny, and the Program Committee for their efforts. I also thank the workshop Steering Committee, listed in these proceedings, for their guidance and assistance.

The workshop is once again cosponsored by the IEEE Computer Society and the US Office of Naval Research. I thank the Office of Naval Research for their support of this workshop's proceedings.

This workshop is held in conjunction with the International Parallel and Distributed Processing Symposium (IPDPS), which is a merger of the symposia formerly known as the International Parallel Processing Symposium (IPPS) and the Symposium on Parallel and Distributed Processing (SPDP). The Heterogeneous Computing Workshop series is very appreciative of the cooperation and assistance we have received from the IPDPS/IPPS organizers for all of the workshop's 15 years.

H. J. Siegel
Colorado State University

Message from the HCW General Chair

Welcome to HCW 2006, the 15th Heterogeneous Computing Workshop. The workshop aims at providing an international forum for researchers who study the phenomenon of heterogeneity, which has become a challenging characteristic of virtually all modern computing platforms. We hope that the workshop will prove a stimulating opportunity for interaction for all speakers and participants.

The HCW workshop series has been associated with the IPDPS meeting since its founding. It is a joy and pleasure to observe the synergy created by this association. It is also a pleasure to acknowledge the IPDPS organizing team, who have established and maintained the infrastructure that enables both IPDPS and all of its satellite workshops.

Many people and organizations deserve our profound gratitude. Jose Fortes, our program chair, has done an outstanding job in managing the entire "life cycle" of HCW: selecting an outstanding program committee, soliciting submissions, coordinating the review process, organizing the keynote speaker and panel, and maintaining the workshop website. In short, Jose has been the type of program chair that general chairs dream of. I wish to extend personal thanks to our Steering Committee Chair, H.J. Siegel, who invited me to be this year's general chair, and who provided decisive help and valuable input during several stages of the organization process. It is a pleasure to thank the workshop sponsors: ONR, the U.S. Office of Naval Research; the IEEE Computer Society, through the Technical Committee on Parallel Processing (TCPP); and the University of Florida.

I wish you all a very productive meeting, and a memorable stay on the beautiful Isle of Rhodes.

Arnold L. Rosenberg
University of Massachusetts Amherst

Message from the HCW Program Chair

I join the Steering Committee and General Chairs of HCW 2006 in welcoming you to Greece and an exciting technical program. This year's HCW program features twelve technical papers, a keynote address and a panel. The Program Committee and additional reviewers evaluated, discussed and selected the papers out of a total of twenty four submissions. Each paper had at least three reviews, the common case being four reviews. Reviewers were asked to consider both the quality of the submissions and their relevance to heterogeneous computing. Program scheduling constraints precluded the acceptance of additional papers of good quality which we hope will be improved as a result of the review process and presented at other venues.

In addition to submitted papers, the technical program continues the HCW tradition of a Keynote Speaker and a panel session. This year's Keynote Speaker is Richard Graham from the Los Alamos National Laboratory. His timely presentation entitled "Aspects of Heterogeneous Computing in the Open MPI environment" discusses the recently finalized Open MPI specification and implementation which include user-transparent support for processor and network heterogeneity.

This year's panel session is entitled "Programming Heterogeneous Systems -Less pain! Better performance!." It considers classical questions of whether languages should require code developers to optimize individual resource usage or should hide such details from programmers, and whether compilers and language features can support models that deliver good performance and programming convenience. These questions have renewed importance as heterogeneity becomes pervasive at all levels of granularity, from multi-core chips and clusters to local- and wide-area networks of computers. The Panelists are Richard Graham (Los Alamos National Laboratory), Alexey Lastovetsky, (University College Dublin) and Samuel Midkiff (Purdue University). Their combined expertise and experience in heterogeneous systems programming bring to the panel session the perspectives of language designers, compiler researchers and applications developers for both distributed and shared memory systems.

On behalf of the Program Committee of HCW I thank both the Keynote Speaker and the Panelists for accepting our invitation to participate in the HCW technical program. I also thank Arnold Rosenberg and H. J. Siegel, HCW 2006 General Chair and HCW Steering Committee Chair, respectively, for their valuable advice and direction. Their support, timely answers to my questions and gentlemanly reminders to stay on schedule deserve much credit for the success of the technical program. Thanks are also due to Shoukat Ali for his great job as Proceedings Chair and willingness to work with different schedules of very busy people. Credit should also be given to my assistant, Ms. Catherine Sembajwe-Reeves, who did an excellent job of taking care of correspondence and web-related matters in support of advertising, paper submissions and program committee activities. Last but not least, I personally wish to thank all the Program Committee members and external reviewers for their efforts in reviewing papers submitted to HCW. The quality of the technical program is entirely to their credit!

José A. B. Fortes, Program Chair of HCW 2006
University of Florida, Gainesville

HCW Keynote: Aspects of Heterogeneous Computing in the Open MPI Environment

Richard L. Graham

Los Alamos National Laboratory
LA-UR-06-0777
NM, USA

There are several aspects to heterogeneous computing, with this talk focusing on several of these, as they relate to execution of a single parallel job. This includes processor and network heterogeneity, as well as the support for making this heterogeneity transparent in the run-time system. Design and implementation choices made by the Open MPI/Open RTE collaboration will be discussed, with an emphasis placed on the effort to make these choices transparent to the end users - both for MPI and non-MPI parallel jobs.

The main standard libraries used for scientific simulation over the last decade and a half are the Message Passing Interface (MPI) and the Parallel Virtual Machine (PVM) libraries. Early implementations of PVM supported processor and network heterogeneity, allowing a single application to run in a hybrid environment. However, this came with a significant performance penalty. Implementors of the MPI standard, the de-facto communications standard for scientific computing, have tended to focus most of their efforts on highly optimized, single system implementations, largely ignoring the challenges posed with the goal of implementing an efficient MPI for a heterogeneous environment. In addition, implementations that enable running an application in a heterogeneous environment have tended to expose these details at the application level, requiring the applications to deal explicitly with these environment, limiting the extent to which heterogeneous computing has caught on.

Building on experience gained with the LA-MPI, LAM/MPI, FT-MPI, and PACX-MPI, the Open MPI project is to enabling applications to run effectively on available hardware. This effort is aimed at removing the practical requirement restricting single application runs to a single type of hardware, and increasing the ability to utilize available hardware. Special attention is given to hiding these details from the end-user, so that, from an applications perspective, running on a homogeneous system is essentially the same as running on a homogenous system. Support is included for high performance communications in a hybrid environment, as well as run-time support for heterogeneous environments.

This talk will describe the component architecture used by the Open MPI project which forms the foundation for providing instance specific implementations of a particular functionality, such as point-to-point communications between two specific end-points. The design choice to provide fine level control over the algorithms deployed within a single job provides a key architectural feature enabling optimal use of system resources. This talk focuses on point-to-point communications and the run-time (Open RTE) environment used to create, monitor, and terminate parallel job execution - both MPI and non-MPI jobs. The point-to-point communications discussion will focus on the architectural features enabling pair-wise communications tailored to the requirements posed only by the specific pair, as well as those that enable the simultaneous use of different network types between a given pair of communication end points. Performance data will also be presented. Aspects of Open RTE that enable distributed process monitoring and control in a heterogeneous environment will be discussed.

Speaker biography: Richard Graham is the Computer Systems and Software Environment (ASC) Program manager, and the Advanced Computing Laboratory acting group leader at the Los Alamos National Laboratory. He joined LANL's Advanced Computing Laboratory (ACL) as a technical staff member in 1999. As team leader for the Resilient Technologies Team he started the LA-MPI project, and is one of the founders of the Open MPI collaboration. Prior to joining the ACL, he spent seven years working at Cray Research and SGI.

Rich obtained his PhD in Theoretical Chemistry from Texas A&M University in 1990 and did post-doctoral work at the James Franck Institute of the University of Chicago. His BS in chemistry was from Seattle Pacific University.

HCW Panel: Programming heterogeneous systems - Less pain! Better performance!

José Fortes

University of Florida
Gainesville, FL, USA
fortes@ufl.edu

Chair: José Fortes

Panelists:
Richard Graham (Los Alamos National Laboratory)
Alexey Lastovetsky, (University College Dublin)
Samuel Midkiff (Purdue University)

Abstract: Heterogeneity in computing systems has been driven by, among others, one or more of the following factors: need for better performance (e.g. in multi-core chips), applications' requirements (e.g. in digital processing systems), timing and logistics of computer facility development (e.g. clusters that are extended and upgraded over time) and emergent systems of systems (e.g. for Grid-computing). The premise that these heterogeneous computing systems (HCS's) offer cost and performance benefits is true only if they can be efficiently programmed. The perspectives and questions on programming HCS's considered by this panel include the following:

1. Should programmers be exposed to heterogeneity so that they can squeeze all the necessary performance by taking advantage of the best resources for the jobs that need them? How can we handle the glut of programmers wishing to program such complex and extensive systems?
2. Should we design compilers that schedule and optimize programs written for homogeneous systems for execution on heterogeneous systems? Will the resulting programs run with better performance than they could achieve in a homogeneous system? How should better be defined in this context?
3. Are programming languages irrelevant in the sense that one can always connect components and/or services to accomplish any task? How generally applicable is this programming model?
4. Are HCS's inherently distributed memory entities where shared-memory programming models cannot succeed? Can shared- and distributed-memory models coexist? How do currently available languages fare in this regard?

The impact of heterogeneity on master-slave on-line scheduling

Jean-francois Pineau, Yves Robert and Frédéric Vivien

LIP, CNRS-INRIA
Ecole Normale Supérieure
Lyon, France
{*Jean-Francois.Pineau, Yves.Robert, Frederic.Vivien*}*@ens-lyon.fr*

In this paper, we assess the impact of heterogeneity for scheduling independent tasks on master-slave platforms. We assume a realistic one-port model where the master can communicate with a single slave at any time-step. We target on-line scheduling problems, and we focus on simpler instances where all tasks have the same size. While such problems can be solved in polynomial time on homogeneous platforms, we show that there does not exist any optimal deterministic algorithm for heterogeneous platforms. Whether the source of heterogeneity comes from computation speeds, or from communication bandwidths, or from both, we establish lower bounds on the competitive ratio of any deterministic algorithm. We provide such bounds for the most important objective functions: the minimization of the makespan (or total execution time), the minimization of the maximum response time (difference between completion time and release time), and the minimization of the sum of all response times. Altogether, we obtain nine theorems which nicely assess the impact of heterogeneity on on-line scheduling.

These theoretical contributions are complemented on the practical side by the implementation of several heuristics on a small but fully heterogeneous MPI platform. Our (preliminary) results show the superiority of those heuristics which fully take into account the relative capacity of the communication links.

Wrekavoc: a Tool for Emulating Heterogeneity

Louis-claude Canon[1] and Emmanuel Jeannot[2]

[1]*ESEO*
Amiens, France
louis-claude.canon@eseo.fr

[2]*INRIA-LORIA*
Vandœuvre les Nancy, France
emmanuel.jeannot@loria.fr

Computer science and especially heterogeneous distributed computing is an experimental science. Simulation, emulation, or in-situ implementation are complementary methodologies to conduct experiments in this context. In this paper we address the problem of defining and controlling the heterogeneity of a platform. We evaluate the proposed solution, called Wrekavoc, with micro-benchmark and by implementing algorithms of the literature.

Scheduling Multiple DAGs onto Heterogeneous Systems

Henan Zhao and Rizos Sakellariou

School of Computer Science
University of Manchester
Manchester, UK
{*hzhao, rizos*}*@cs.man.ac.uk*

The problem of scheduling a single DAG onto heterogeneous systems has been studied extensively. In this paper, we focus on the problem of scheduling more than one DAG at the same time onto a set of heterogeneous resources. The aim is not only to optimize the overall makespan, but also to achieve fairness, defined on the basis of the slowdown that each DAG would experience as a result of competing for resources with other DAGs. Two policies particularly focussing to deliver fairness are presented and evaluated along with another four policies that can be used to schedule multiple DAGs.

Scheduling of Tasks with Batch-shared I/O on Heterogeneous Systems

Nagavijayalakshmi Vydyanathan[1], Gaurav Khanna[1], Umit Catalyurek[2,3], Tahsin Kurc[2], P. Sadayappan[1] and Joel Saltz[1,2]

[1]*Dept. of Computer Science and Engineering*
Ohio State University
Columbus, OH, USA
{*vydyanat, khannag, saday*}*@cse.ohio-state.edu,*
jsaltz@bmi.osu.edu

[2]*Dept. of Biomedical Informatics*
Ohio State University
Columbus, OH, USA
{*umit, kurc*}*@bmi.osu.edu*

[3]*Dept. of Electrical and Computer Engineering*
Ohio State University
Columbus, OH, USA

This paper proposes a novel strategy that uses hypergraph partitioning and K-way iterative mapping-refinement heuristics for scheduling a batch of data-intensive tasks with batch-shared I/O behavior on heterogeneous collections of storage and compute clusters. The strategy formulates the sharing of files among tasks as a hypergraph to minimize the I/O overheads due to transferring of the same set of files multiple times and employs a K-way iterative mapping-refinement scheme to adapt to the heterogeneity of compute clusters and storage networks in the system. We evaluate the proposed approach through real experiments and simulations on application scenarios from two application domains; satellite data processing and biomedical imaging. Our experimental results show that our approach can achieve significant performance improvement over algorithms such as HPS, Shortest Job First, MinMin, MaxMin and Sufferage for workloads with high degree of shared I/O among tasks.

A Task Duplication Based Bottom-Up Scheduling Algorithm for Heterogeneous Environments

Doruk Bozdağ[1], Umit Catalyurek[2] and Füsun Ozgüner[1]

[1]Dept. of Electrical and Computer Engineering
The Ohio State University
Columbus, OH, USA
{bozdagd, ozguner}@ece.osu.edu

[2]Dept. of Biomedical Informatics
The Ohio State University
Columbus, OH, USA
umit@bmi.osu.edu

We propose a new duplication-based DAG scheduling algorithm for heterogeneous computing environments. Contrary to the traditional approaches, proposed algorithm traverses the DAG in a bottom-up fashion while taking advantage of task duplication and task insertion. Experimental results on random DAGs and three different application DAGs show that the makespans generated by the proposed DBUS algorithm are much better than those generated by the existing algorithms, HEFT, HCPFD and HCNF.

FIFO scheduling of divisible loads with return messages under the one-port model

Olivier Beaumont[1], Loris Marchal[2], Veronika Rehn[2] and Yves Robert[2]

[1]LaBri, UMR CNRS 5800
Bordeaux, France
Olivier.Beaumont@labri.fr

[2]LIP, UMR CNRS-INRIA-UCBL 5668
ENS Lyon
Lyon, France
{Loris.Marchal, Veronika.Rehn,
Yves.Robert}@ens-lyon.fr

This paper deals with scheduling divisible load applications on star networks, in presence of return messages. This work is a follow-on of previous studies, where the same problem was considered under the two-port model, where a given processor can simultaneously send and receive messages. Here, we concentrate on the one-port model, where a processor can either send or receive a message at a given time step. The problem of scheduling divisible load on star platforms turns out to be very difficult as soon as return messages are involved. Unfortunately, we have not been able to assess its complexity, but we provide an optimal solution in the special (but important) case of FIFO communication schemes. We also provide an explicit formula for the optimal number of load units that can be processed by a FIFO ordering on a bus network. Finally, we provide a set of MPI experiments to assess the accuracy and usefulness of our results in a real framework.

Using SCTP to hide latency in MPI programs

Humaira Kamal, Brad Penoff, Mike Tsai, Edith Vong and Alan Wagner

Department of Computer Science
University of British Columbia
Vancouver, BC, Canada
{*kamal, penoff, myct, vongpsq, wagner*}*@cs.ubc.ca*

A difficulty in using heterogeneous collections of geographically distributed machines across wide area networks for parallel computing is the huge variability in message latency that is orders of magnitude larger than parallel programs executing on dedicated systems. This variability is in part due to the underlying network bandwidth and latency which can vary dramatically according to network conditions. Although such an environment is not suitable for many message passing programs there are those programs that can take advantage of it.

Using SCTP (Stream Control Transmission Protocol) for MPI, we show how to reduce the effect of latency on task farm programs to allow them to effectively execute in high latency environments. SCTP is a recently standardized transport level protocol that has a number of features that make it well-suited to MPI and our goal is to reduce the effect of latency on MPI programs in wide area networks. We take advantage of SCTP's improved congestion control as well as its ability to have multiple independent message streams over a single connection to eliminate the head of line blocking that can occur in TCP-based middleware.

The use of streams required a novel use of MPI tags to identify independent streams rather than different types of messages. We describe the design of a task farm template that exploits streams, uses buffering and pipelining of task requests to improve its performance under network loss and variable latency. We use these techniques to improve the performance of two real-world MPI programs: a robust correlation matrix computation and mpiBLAST.

A Brokering Framework for Large-Scale Heterogeneous Systems

Xin Bai[1], Ladislau Boloni[1], Dan C. Marinescu[1], Howard Jay Siegel[2], Rose A. Daley[3] and I-jeng Wang[3]

[1]*School of Electrical Engineering and Computer Science*
University of Central Florida
Orlando, Florida, USA
{*xbai, lboloni, dcm*}*@cs.ucf.edu*

[2]*Department of Electrical and Computer Engineering and Department of Computer Science*
Colorado State University
Fort Collins, Colorado, USA
HJ@colostate.edu

[3]*Applied Physics Laboratory*
Johns Hopkins University
Baltimore, Maryland, USA
{*Rose.Daley, I-Jeng.Wang*}*@jhuapl.edu*

In this paper we discuss the role of a broker in a market-oriented resource llocation model for large-scale heterogeneous systems. The simplified model is based upon a three party ystem, provider-broker-consumer. The allocation of resources is determined by their price, their utility to the onsumer, and by the satisfaction of the consumer. The role of the broker is to add societal objectives to resource llocation algorithms and to mediate between greedy consumers and selfish providers. A simulation experiment was onducted to study the transient and the steady-state behavior of several performance measures, including the verage consumer satisfaction, the average utility, and the hourly revenue.

Cooperative Load Balancing for a Network of Heterogeneous Computers

Satish Penmatsa and Anthony T. Chronopoulos

Dept. of Computer Science
The University of Texas at San Antonio
6900 N Loop, 1604 W, San Antonio, TX 78249, USA
{*spenmats, atc*}*@cs.utsa.edu*

In this paper we present a game theoretic approach to solve the static load balancing problem in a distributed system which consists of heterogeneous computers connected by a single channel communication network. We use a cooperative game to model the load balancing problem. Our solution is based on the Nash Bargaining Solution (NBS) which provides a Pareto optimal solution for the distributed system and is also a fair solution. An algorithm for computing the NBS is derived for the proposed cooperative load balancing game. Our scheme is compared with that of other existing schemes under simulations with various system loads and configurations. We show that the solution of our scheme is near optimal and is superior to the other schemes in terms of fairness.

An Economy-driven Mapping Heuristic for Hierarchical Master-Slave Applications in Grid Systems

Nadia Ranaldo[1] and Eugenio Zimeo[2]

[1]*Department of Engineering*
University of Sannio
Benevento, Italy
ranaldo@unisannio.it

[2]*Research Centre on Software Technology*
University of Sannio
Benevento, Italy
zimeo@unisannio.it

In heterogeneous distributed systems, such as Grids, a resource broker is responsible of automatically selecting resources, and mapping application tasks to them. A crucial aspect of resource broker design, especially in a next commercial exploitation of grid systems, in which economy theories for resource management will be applied, is the support to task mapping based on the fulfilment of Quality of Service (QoS) constraints. The paper presents an economy-driven mapping heuristic, called time minimization, for mapping and scheduling the tasks assigned to the slaves of a master-slave application in a hierarchical and heterogeneous distributed system. The validity and accuracy of such heuristic are tested by implementing it in a resource broker of a hierarchical grid middleware used for running a real world application.

Plan Switching: An Approach to Plan Execution in Changing Environments

Han Yu[1], Dan C. Marinescu[1], Annie S. Wu[1], Howard Jay Siegel[2], Rose A. Daley[3] and I-jeng Wang[3]

[1] *School of Electrical Engineering and Computer Science*
University of Central Florida
Orlando, Florida, USA
{hyu, dcm, aswu}@cs.ucf.edu

[2] *Department of Electrical and Computer Engineering and and Department of Computer Science*
Colorado State University
Fort Collins, Colorado, USA
HJ@ColoState.edu

[3] *Applied Physics Laboratory*
Johns Hopkins University
Laurel, Maryland, USA
{Rose.Daley, I-Jeng.Wang}@jhuapl.edu

The execution of a complex task in any environment requires planning. Planning is the process of constructing an activity graph given by the current state of the system, a goal state, and a set of activities. If we wish to execute a complex computing task in a heterogeneous computing environment with autonomous resource providers, we should be able to adapt to changes in the environment. A possible solution is to construct a family of activity graphs beforehand and investigate the means of switching from one member of the family to another when the execution of one activity graph fails. In this paper, we study the conditions when plan switching is feasible. Then we introduce an approach for plan switching and report the simulation results of this approach.

Integrating heterogeneous information services using JNDI

Dirk Gorissen, Piotr Wendykier, Dawid Kurzyniec and Vaidy Sunderam

Dept. of Math and Computer Science
Emory University
Atlanta, GA, USA
dgorissen@gmail.com, {wendyk, dawidk, vss}@mathcs.emory.edu

The capability to announce and discover resources is a foundation for heterogene ous computing systems. Independent projects have adopted custom implementations of information services, which are not interoperable and induce substantial maintenance costs. In this paper, we propose an alternative methodology. We suggest that it is possible to reuse existing naming service deployments and combine them into complex, scalable, hierarchical, distributed federations, by using appropriate client-side integration middleware that unifies service access and hides heterogeneity behind a common API. We investigate a JNDI-based approach, and describe in detail two newly implemented JNDI service providers, which enable unified access to 1) Jini lookup services, and 2) Harness Distributed Naming Services. We claim that these two technologies, along with others already accessible through JNDI such as e.g. DNS and LDAP, offer features suitable for use in hierarchical heterogeneous information systems.

Workshop 2

Workshop on Parallel and Distributed Real-Time Systems WPDRTS 2006

2 WPDRTS • Workshop on Parallel and Distributed Real-Time Systems

Workshop Description:

WPDRTS is a forum for the presentation and discussion of approaches, research findings, and experiences in the domain of large-scale parallel and distributed real-time systems. Both research and development of relevant technologies are of interest, as well as the applications built using such technologies.

Topics of interest include but are not limited to:

- Resource Management: Value-based, Feedback-based, and Power-aware scheduling, Combined scheduling of hard and soft real-time tasks, Dynamic real-time systems, Real-time servers, Fault-tolerance and Security in real-time systems.
- Operating Systems and Middleware: Run-time systems, Middleware architectures, Real-Time Linux, Real-Time CORBA, Real-time Databases.
- Programming Environments: Software design, Parallelization methods/tools for DSP-based, reconfigurable, and mixed-computation-paradigm architectures, Real-Time Java.
- Algorithms and Applications: Signal/image processing, Vision/robotic systems, Sensor Web, Industrial automation, Vehicle guidance, Command and control.
- Architectures: Special-purpose processors, mixed-computation-paradigm, size/weight/power modeling and management.
- Specification, Modeling, and Analysis: Formal methods, Object orientation, Benchmarking, Tools and environments.
- Networking and Communications: Real-time communication protocols, Sensor Networks, allocation control mechanisms, and performance analysis.

General Chairs:
Lisa DiPippo, University of Rhode Island, USA
Vana Kalogeraki, University of California, Riverside, USA

Program Co-chairs:
Zdenek Hanzalek, Czech Technical University in Prague, Czech Republic
Chenyang Lu, Washington University in St. Louis, USA

Program Committee:
Karl-Erik Arzen, Lund Institute of Technology, Sweden
Sanjoy Baruah, University of North Carolina, USA
Ed Brinksma, University of Twente, Netherlands
Maryline Chetto, Universite de Nantes, France
Chris D. Gill, Washington University in St. Louis, USA
Michael G. Harbour, University of Cantabria, Spain
Xenofon D. Koutsoukos, Vanderbilt University, USA
Victor Lee, City University of Hong Kong, China
Giuseppe Lipari, Scuola Superiore S.Anna, Italy
Daniel Mossé, University of Pittsburgh, USA
Paulo Pedreiras, University of Aveiro, Portugal
John Regehr, University of Utah, USA
Ismael Ripoll, Polytechnic University of Valencia, Spain
Douglas C. Schmidt, Vanderbilt University, USA
Oleg Sokolsky, University of Pennsylvania, USA
Francisco Vasques, University of Porto, Portugal
Wei Zhao, NSF and Texas A&M University, USA

Publicity Chairs
Eduardo Tovar, Polytechnic Institute of Porto, Portugal
Victor Lee, City University of Hong Kong, China
Ying Lu, University of Nebraska-Lincoln, USA

Publication Chair
Xenofon D. Koutsoukos, Vanderbilt University, USA

Submission Chair
Xiaorui Wang, Washington University in St. Louis, USA

Special Session Chairs
Tarek F. Abdelzaher, University of Illinois at Urbana-Champaign, USA
Karl-Erik Arzen, Lund Institute of Technology, Sweden
Angelika Mader, University of Twente, Netherlands
Ansgar Fehnker, University of New South Wales, Australia
Jeffery Hansen, Carnegie Mellon University, USA
Frank Drews, Ohio University, USA
Steering Committee Chairs
Klaus Ecker , TU Clausthal, Germany
G. Manimaran, Iowa State University, USA

Steering Committee
David Andrews, University of Kansas, USA
Scott Brandt, University of California at Santa Cruz, USA
Chris Gill, Washington University in St. Louis, USA
Guenter Hommel, Technische Universität Berlin, Germany
Doug Locke, Timesys Corporation, USA
Priya Narasimhan, Carnegie-Mellon University, USA
Barbara Pfarr, NASA Goddard, USA
Viktor Prasanna, University of Southern California, USA
Behrooz Shirazi, University of Texas at Arlington, USA
Lonnie R. Welch, Ohio University, USA
Paul R. Work, Raytheon Company, USA
Armin Zimmerman, Technische Universität Berlin, Germany

WPDRTS Keynote: Component-based Construction of Embedded Systems

Joseph Sifakis

Verimag & ARTIST2 NoE
Centre Equation
38610 GIERES, France
sifakis@imag.fr

We present a framework for the component-based construction of embedded systems. The framework is based on a general semantic model, encompassing various models of computation for real-time systems. It is characterized by the combined use of models for behavior, interaction and dynamic priorities. Interaction models describe interactions between components by using connectors with synchronization types. Dynamic priorities are used to specify controllers and schedulers in particular.

We also present a methodology for model-based composition of real-time systems using this semantic model. The methodology enables correct-by-construction development for properties such as deadlock-freedom and progress, as well as incremental construction and associativity of composition operators. We present two implementations of the framework in system modeling and validation tools developed at Verimag:
- A partial implementation in the state exploration platform of the IF tool suite dedicated to the validation of asynchronous system modeling languages such as UML and SDL;
- A more recent full implementation in a platform for the execution of both synchronous and asynchronous components.

The methodology is illustrated by the use of these tools on case studies for real-time systems modeling and validation.

Decentralized and Dynamic Bandwidth Allocation in Networked Control Systems

Ahmad T. Al-hammouri[1], Michael S. Branicky[1], Vincenzo Liberatore[1] and Stephen M. Phillips[2]

[1]*Dept. of Electrical Engineering and Computer Science*
Case Western Reserve University
Cleveland, Ohio, USA
{ata5, mb, vl}@case.edu

[2]*Dept. of Electrical Engineering*
Arizona State University
Tempe, Arizona, USA
stephen.phillips@asu.edu

In this paper, we propose a bandwidth allocation scheme for networked control systems that have their control loops closed over a geographically distributed network. We first formulate the bandwidth allocation as a convex optimization problem. We then present an allocation scheme that solves this optimization problem in a fully distributed manner. In addition to being fully distributed, the proposed scheme is asynchronous, scalable, dynamic and flexible. We further discuss mechanisms to enhance the performance of the allocation scheme. We present analytical and simulation results.

The Robot Software Communications Architecture (RSCA): Embedded Middleware for Networked Service Robots

Seongsoo Hong[1], Jaesoo Lee[1], Hyeonsang Eom[2] and Gwangil Jeon[3]

[1]*Real-Time Operating Systems Laboratory, School of Electrical Engineering and Computer Science*
Seoul National University
Seoul 151-744, Korea
{*sshong, jslee*}*@redwood.snu.ac.kr*

[2]*Distributed Information Processing Laboratory, School of Computer Science and Computer Engineering*
Seoul National University
Seoul 151-744, Korea
hseom@cse.snu.ac.kr

[3]*Department of Computer Engineering*
Korea Polytechnic University
2121 Jungwang-Dong, Siheung-Si, Gyunggi-Do 429-793, Korea
gijeon@kpu.ac.kr

In this paper, we present a robot middleware technology named Robot Software Communications Architecture (RSCA) for its use in networked home service robots. The RSCA provides a standard operating environment for the robot applications together with a framework that expedites the development of such applications. The operating environment is comprised of a real-time operating system, a communication middleware, and a deployment middleware. Particularly, the deployment middleware supports the reconfiguration of component-based robot applications including installation, creation, start, stop, tear-down, and un-installation. In designing RSCA, we have adopted a middleware called SCA from the software defined radio domain and extend it since the original SCA lacks the real-time guarantees and appropriate event services. We have fully implemented RSCA and performed measurements to quantify its run-time performance. Our implementation clearly shows the viability of RSCA.

Schedulability Analysis of AR-TP, a Ravenscar Compliant Communication Protocol for High-Integrity Distributed Systems

Santiago Urueña[1], Juan Zamorano[1], Daniel Berjón[2], José A. Pulido[2] and Juan A. De La Puente[2]

[1]*Dept. of Comp. Architecute and Technology*
Technical University of Madrid
Boadilla del Monte, Spain
{*suruena, jzamora*}*@datsi.fi.upm.es*

[2]*Dept. of Telematic Systems Engineering*
Technical University of Madrid
Madrid, Spain
berjon@dit.upm.es, {*pulido, jpuente*}*@di.upm.es*

A new token-passing algorithm called AR-TP for avoiding the non-determinism of some networking technologies is presented. This protocol allows the schedulability analysis of the network, enabling the use of standard Ethernet hardware for Hard Real-Time behavior while adding congestion management. It is specially designed for High-Integrity Distributed Hard Real-Time Systems, being fully compliant with the Ravenscar Profile.

Realization of Virtual Networks in the DECOS Integrated Architecture

Roman Obermaisser[1] and Philipp Peti[2]

[1]*Real-Time Systems Group*
Vienna University of Technology
Vienna, Vienna, Austria
romano@vmars.tuwien.ac.at

[2]*Real-Time Systems Group*
Vienna University of Technology
Vienna, Vienna, Austria
php@vmars.tuwien.ac.at

Due to the better utilization of computational and communication resources and the improved coordination of application subsystems, designers of large distributed embedded systems (e.g., in the automotive domain) are eager to replace existing federated architectures with integrated ones. This paper focuses on the communication infrastructure of the DECOS integrated system architecture, which realizes for each application subsystem a so-called virtual network as an overlay network on top of a time-triggered communication protocol. Since all virtual networks share a single physical network, virtual networks promise massive cost savings through the reduction of physical networks and reliability improvements with respect to wiring and connectors. Furthermore, virtual networks support application subsystems that range from ultra-dependable control applications (e.g., an X-by-wire system) to non safetycritical applications such as comfort systems. For this reason, two classes (event-triggered and time-triggered) of virtual networks are realized. Event-triggered virtual networks provide high flexibility for non safetycritical application subsystems, while the predictability of the time-triggered paradigm is better suited for safety-critical application subsystems. Encapsulation mechanisms ensure that the temporal properties of each virtual network are known a priori and independent from the communication activities in other virtual networks. In order to ensure that the virtual network abstractions hold also in the case of software faults, each application subsystem possesses a dedicated virtual network with statically assigned resources at the underlying time-triggered communication service.

A Portable Real-time Emulator for Testing Multi-Radio MANETs

Weirong Jiang[1] and Chao Zhang[2]

[1]*Research Institute of Information Technology*
Tsinghua University
Beijing 100084, China
jwr2000@mails.tsinghua.edu.cn

[2]*Tsinghua-TCB Institute of Applied Communication Systems*
Beijing 100084, China
chao.zhch@gmail.com

In building a real-life mobile ad-hoc network (MANET), network emulation has been appraised as an efficient approach for testing the real implementations of routing algorithms and protocol stacks. Most existing MANET emulators can hardly support both real-time scene construction for proof-of-concept test and real-time traffic recording for performance evaluation simultaneously. They also lack the ability to emulate the multi-radio environment. This paper presents a flexible TCP/IP-based real-time MANET emulator that can be portably deployed to facilitate the development of real multi-radio MANET routing protocols. It friendly provides visual interaction of topology control and rich configuration of emulation conditions to enable a real-time and comprehensive examination of protocol implementations.

Battery Aware Dynamic Scheduling for Periodic Task Graphs

Venkat Rao[1], Gaurav Singhal[2], Nicolas Navet[1], Anshul Kumar[3] and G.s Visweswaran[4]

[1] *TRIO Team*
LORIA INRIA
Nancy, France
venkat174@gmail.com, nicolas.navet@loria.fr

[2] *Dept of Electrical and Computer Engineering*
University of Texas
Austin, USA
gauravsinghal58@gmail.com

[3] *Dept of Computer Science and Engineering*
Indian Institute of Technology
Delhi, India
anshul@cse.iitd.ernet.in

[4] *Dept of Electrical Engineering*
Indian Institute of Technology
Delhi, India
gswaran@ee.iitd.ernet.in

Battery lifetime, a primary design constraint for mobile embedded systems, has been shown to depend heavily on the load current profile. This paper explores how scheduling guidelines from battery models can help in extending battery capacity. It then presents a 'Battery-Aware Scheduling' methodology for periodically arriving taskgraphs with real time deadlines and precedence constraints. Scheduling of even a single taskgraph while minimizing the weighted sum of a cost function has been shown to be NP-Hard. The presented methodology divides the problem in to two steps. First, a good DVS algorithms dynamically determines the minimum frequency of execution. Then, a greedy algorithm allows a near optimal priority function to choose the task which would maximize slack recovery. The methodology also ensures adherence of real time deadlines independent of the choice of the DVS algorithm and priority function used, while following battery guidelines to maximize battery lifetime. Battery simulations carried out on the profile generated by our methodology for a large set of taskgraphs show that battery life time is extended up to 23.3% as compared to existing dynamic scheduling schemes.

Scheduling of Tasks with Precedence Delays and Relative Deadlines - Framework for Time-optimal Dynamic Reconfiguration of FPGAs

Premysl Sucha and Zdenek Hanzalek

Department of Control Engineering
Czech Technical University
Prague, Czech Republic, Czech
{suchap, hanzalek}@fel.cvut.cz

This paper is motivated by existing architectures of field programmable gate arrays (FPGAs). To facilitate the design process we present an optimal scheduling algorithm using a very universal framework, where tasks are constrained by precedence delays and relative deadlines. The precedence relations are given by an oriented graph, where tasks are represented by nodes. Edges in the graph are related either to the minimum time or to the maximum time elapsed between the start times of the tasks. This framework is used to model the runtime dynamic reconfiguration, synchronization with an on-chip processor and simultaneous availability of arithmetic units and SRAM memory. The NP-hard problem of finding an optimal schedule satisfying the timing and resource constraints while minimizing the makespan C_{max}, is solved using two approaches. The first one is based on Integer Linear Programming and the second one is implemented as a Branch and Bound algorithm. Experimental results show the efficiency comparison of the ILP and Branch and Bound solutions.

A Hierarchical Scheduling Model for Component-Based Real-Time Systems

Josè L. Lorente[1], Giuseppe Lipari[2] and Enrico Bini[2]

[1] *Universidad de Cantabria*
Spain
lorentejl@unican.es

[2] *Scuola Superiore Sant'Anna*
Italy
{*lipari, e.bini*}*@sssup.it*

In this paper, we propose a methodology for developing component-based real-time systems based on the concept of hierarchical scheduling. Recently, much work has been devoted to the schedulability analysis of hierarchical scheduling systems, in which real-time tasks are grouped into components, and it is possible to specify a different scheduling policy for each component. Until now, only independent components have been considered.

In this paper, we extend this model to tasks that interact through remote procedure calls. We introduce the concept of abstract computing platform on which each component is executed. Then, we transform the system specification into a set of real-time transactions and present a schedulability analysis algorithm. Our analysis is a generalization of the holistic analysis to the case of abstract computing platforms. We demonstrate the use of our methodology on a simple example.

Schedulability Analysis of Non-Preemptive Recurring Real-Time Tasks

Sanjoy K. Baruah[1] and Samarjit Chakraborty[2]

[1] *Department of Computer Science*
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA
baruah@cs.unc.edu

[2] *Department of Computer Science*
National University of Singapore
3 Science Drive 2, SG 117543, Singapore
samarjit@comp.nus.edu.sg

The recurring real-time task model was recently proposed as a model for real-time processes that contain code with conditional branches. In this paper, we present a necessary and sufficient condition for uniprocessor non-preemptive schedulability analysis for this task model. We also derive a polynomial-time approximation algorithm for testing this condition. Preemptive schedulers usually have a larger schedulability region compared to their non-preemptive counterparts. Further, for most realistic task models, schedulability analysis for the non-preemptive version is computationally more complex compared to the corresponding preemptive version. Our results in this paper show that (surprisingly) the recurring real-time task model does not fall in line with these intuitive expectations, i.e. there exists polynomial-time approximation algorithms for both preemptive and non-preemptive versions of schedulability analysis. This has important implications on the applicability of this model, since fully preemptive scheduling algorithms often have significantly larger runtime overheads.

Towards an Analysis of Race Carrier Conditions in Real-time Java

M. T. Higuera-toledano

DACYA, Facultad de Informatica
Universidad Complutense de Madrid
Madrid, Spain
mthiguer@dacya.ucm.es

The RTSJ memory model propose a mechanism based on a scope three containing all region-stacks in the system and a reference-counter collector. In order to avoid reference cycles among regions on the region-stack, RTSJ defines the single parent rule. The given algorithms to maintain the region-stack tructure are not compliant with the defined parentage relation. More over, the suggested algorithms to maintain the single parent rule introduces race carrier conditions on the application behaviour. This paper proposes alternative approaches in order to avoid this problem.

Fault Tolerance with Real-Time Java

Damien Masson and Serge Midonnet

Institut Gaspard-Monge
Université de Marne-La-Valle
Champs sur Marne, France
damien.masson@univ-mlv.fr, serge.midonnet@esigetel.fr

After having drawn up a state of the art on the theoretical feasibility of a system of periodic tasks scheduled by a preemptive algorithm at fixed priorities, we show in this article that temporal faults can occur all the same within a theoretically feasible system, that these faults can lead to a failure of the system and that we can use the data calculated during control of admission to install detectors of faults and to define a factor of tolerance. We show then the results obtained on a system of periodic tasks coded with Java Real-Time and carried out with the virtual machine *jRate*. These results show that the installation of the detectors and the tolerance to the faults makes an improvement of the behavior of the system in the presence of faults.

A Probabilistic Approach for Fault Tolerant Multiprocessor Real-time Scheduling

Vandy Berten[1], Joël Goossens[1] and Emmanuel Jeannot[2]

[1]*Département d'Informatique*
Université Libre de Bruxelles
Bruxelles, Belgium
{*vandy.berten, joel.goossens*}*@ulb.ac.be*

[2]*INRIA-LORIA*
Université H. Poincar, Nancy 1
Vandoeuvre les Nancy, France
emmanuel.jeannot@loria.fr

In this paper we tackle the problem of scheduling a periodic real-time system on identical multiprocessor platforms, moreover the tasks considered may fail with a given probability. For each task we compute its duplication rate in order to (1) given a maximum tolerated probability of failure, minimize the size of the platform such at least one replica of each job meets its deadline (and does not fail) using a variant of EDF namely $\mathsf{EDF}^{(k)}$ or (2) given the size of the platform, achieve the best possible reliability with the same constraints. Thanks to our probabilistic approach, no assumption is made on the number of failures which can occur. We propose several approaches to duplicate tasks and we show that we are able to find solutions always very close to the optimal one.

A Real-Time PES Supporting Runtime State Restoration after Transient Hardware-Faults

Skambraks

Dept. of Electrical and Computer Engineering
FernUniversität in Hagen
58084 Hagen, Germany,
martin.skambraks@fernuni-hagen.de

Controlling safety-critical real-time applications that cannot immediately be transferred to a safe state requires highly reliable Programmable Electronic Systems (PESs). This demand for fault-tolerance is usually satisfied by applying redundant processing structures inside each PES and, additionally, configuring multiple PES redundantly. Instead of minimising the failure probability of single PESs, it is also desirable to provide a redundant configuration of PESs with the capability to re-start single units at runtime. This requires copying a PES's internal state at runtime, since a re-started unit must equalise its internal state with that of its redundant counterparts before the redundant processing can be rejoined. As a result, redundancy attrition due to transient faults is prevented, since failed channels can be brought back on line. This article states the problems concerned with runtime state restoration of real-time systems, discusses the advantages and disadvantages of existing techniques and introduces a hardware-supported state restoration concept.

Honeybees: Combining Replication and Evasion for Mitigating Base-station Jamming in Sensor Network

Sherif Khattab, Daniel Mossé and Rami Melhem

Department of Computer Science
University of Pittsburgh
Pittsburgh, PA, USA
{*skhattab, mosse, melhem*}*@cs.pitt.edu*

By violating MAC-layer protocols, the jamming attack aims at blocking successful communication among wireless nodes. Wireless sensor networks (WSNs) are highly vulnerable to jamming because of reliance on shared wireless medium, constrained per-sensor resources, and high risk of sensor compromise. Moreover, base stations of WSNs are single points of failure and, thus, attractive jamming targets. To tackle base-station jamming, replication of base stations as well as jamming evasion, by relocation to unjammed locations, have been proposed. In this paper, we propose Honeybees, an energy-aware defense framework against base-station jamming attack in WSNs. Honeybees efficiently combines replication and evasion to allow WSNs to continue delivering data for a long time during a jamming attack.

We present three defense strategies: reactive, proactive, and hybrid, in the context of multi-hop WSN deployment. Through simulation, we show the interaction of these strategies with different attack tactics as well as the effect of system and attack parameters. We found that our honeybees framework struck an energy-efficient balance between replication and evasion that outperformed both separate mechanisms. Specifically, hybrid honeybees outperformed replication and evasion at low and intermediate number of attackers and gracefully degraded to high attack intensity.

Murphy Loves Potatoes: Experiences from a Pilot Sensor Network Deployment in Precision Agriculture

Koen Langendoen, Aline Baggio and Otto Visser

Faculty of Electrical Engineering, Mathematics, and Computer Science
Delft University of Technology
Delft, The Netherlands
{*K.G.Langendoen, A.Baggio, O.W.Visser*}*@tudelft.nl*

We report on preliminary experiences with deploying a large-scale sensor network (about 100 nodes) for a pilot in precision agriculture. The pilot did not answer the initial research questions, but instead revealed many engineering problems typically overlooked by (computer) scientists evaluating their work by means of simulation. The deployment prompted us to rethink our development process and includes important lessons for the WSN research community as a whole.

An Overview of Data Aggregation Architecture for Real-Time Tracking with Sensor Networks

Tian He[1], Lin Gu[2], Liqian Luo[3], Ting Yan[2], John A. Stankovic[2] and Sang H. Son[2]

[1]*Department of Computer Science and Engineering*
University of Minnesota
Minneapolis, Minnesota, USA
Tianhe@cs.umn.edu

[2]*Department of Computer Science*
University of Virginia
Charlottesville, Virginia, USA
{lg6e, ty4k, stankovic, son}@cs.virginia.edu

[3]*Department of Computer Science*
University of Illinois at Urbana Champaign
Urbana, Illinois, USA
lluo@cs.uiuc.edu

Since sensor nodes normally have limited resources in terms of energy, bandwidth and computation capability, efficiency is a key design goal in sensor network research. As one of techniques to achieve efficiency, data aggregation has been extensively investigated in recent literature. Previous research on data aggregation has demonstrated its effectiveness in reducing traffic, easing congestion and decreasing the energy consumption. However few are actually designed for a real-world application and implemented in a running system. This paper describes our design and implementation of a physical tracking system, using an aggressive data aggregation architecture as one of building blocks. This architecture can be generally applied to other sensor systems, where communication efficiency is a paramount concern and networking resources are limited.

Formal Modeling and Analysis of Wireless Sensor Network Algorithms in Real-Time Maude

Peter Csaba Olveczky and Stian Thorvaldsen

Department of Informatics
University of Oslo
Oslo, Norway
{peterol, stianth}@ifi.uio.no

Advanced wireless sensor network algorithms pose challenges to their formal modeling and analysis, such as modeling probabilistic and real-time behaviors and novel forms of communication, and analyzing both correctness and performance. In this paper, we propose using Real-Time Maude to formally model, simulate, and further analyze such algorithms. The Real-Time Maude formalism is expressive yet intuitive, and the tool provides a spectrum of analysis methods, including simulation, reachability analysis, and temporal logic model checking.

We have used Real-Time Maude to formally model and analyze the sophisticated OGDC algorithm. We could perform all the analyses performed by the OGDC developers using the simulation tool ns-2, as well as further analyses which are beyond the capabilities of simulation tools. To the best of our knowledge, this is the first time a formal tool has been applied to such a complex wireless sensor network algorithm.

GTS Allocation Analysis in IEEE 802.15.4 for Real-Time Wireless Sensor Networks

Anis Koubaa[1], Mario Alves[2] and Eduardo Tovar[3]

[1]IPP-HURRAY! Research Group
ISEP/Polytechnic Institute of Porto
Porto, Portugal
akoubaa@dei.isep.ipp.pt

[2]IPP-HURRAY! Research Group
ISEP/Polytechnic Institute of Porto
Porto, Portugal
mjf@isep.ipp.pt

[3]IPP-HURRAY! Research Group
ISEP/Polytechnic Institute of Porto
Porto, Portugal
emt@dei.isep.ipp.pt

The IEEE 802.15.4 protocol proposes a flexible communication solution for Low-Rate Wireless Personal Area Networks including sensor networks. It presents the advantage to fit different requirements of potential applications by adequately setting its parameters. When enabling its beacon mode, the protocol makes possible real-time guarantees by using its Guaranteed Time Slot (GTS) mechanism. This paper analyzes the performance of the GTS allocation mechanism in IEEE 802.15.4. The analysis gives a full understanding of the behavior of the GTS mechanism with regards to delay and throughput metrics. First, we propose two accurate models of service curves for a GTS allocation as a function of the IEEE 802.15.4 parameters. We then evaluate the delay bounds guaranteed by an allocation of a GTS using Network Calculus formalism. Finally, based on the analytic results, we analyze the impact of the IEEE 802.15.4 parameters on the throughput and delay bound guaranteed by a GTS allocation. The results of this work pave the way for an efficient dimensioning of an IEEE 802.15.4 cluster.

Power-Aware Data Dissemination Protocols in Wireless Sensor Networks

Sotiris Nikoletseas

Computer Engineering and Informatics Department and CTI
University of Patras
Patras, Greece
nikole@cti.gr

Recent rapid technological developments have led to the development of tiny, low-power, low-cost sensors. Such devices integrate sensing, limited data processing and communication capabilities.The effective distributed collaboration of large numbers of such devices can lead to the efficient accomplishment of large sensing tasks.

This invited talk focuses on several aspects of energy efficiency. Two protocols for data propagation are studied: the first creates probabilistically optimized redundant data transmissions to combine energy efficiency with fault tolerance, while the second guarantees (in a probabilistic way) the same per sensor energy dissipation, towards balancing the energy load and prolong the lifetime of the network.

A third protocol (in fact a power saving scheme) is also presented, that directly and adaptively affects power dissipation at each sensor. This lower level scheme can be combined with data propagation protocols to further improve energy efficiency.

Algorithmic Models for Sensor Networks

Stefan Schmid and Roger Wattenhofer

Computer Engineering and Networks Laboratory
ETH Zurich
Zurich, Switzerland
{*schmiste, wattenhofer*}*@tik.ee.ethz.ch*

Developing algorithms for sensor networks—and proving their correctness and performance—, requires simplifying but still realistic models. This paper surveys various models in use today and puts them into perspective. In addition, we propose interesting models which are not widely adopted by the community so far.

Solving Generic Role Assignment Exactly

Christian Frank and Kay Römer

Department of Computer Science
ETH Zurich
Zurich, Switzerland
{*chfrank, roemer*}*@inf.ethz.ch*

Generic role assignment is a programming abstraction that supports the assignment of user-defined *roles* to sensor nodes such that certain conditions are met. Many common network configuration problems such as coverage (assign roles ON and OFF to sensor nodes such that ON nodes cover a physical area with their sensors), clustering, or in-network data aggregation can be formulated as role assignment problems. Building on our previous work in this area, we propose an extended role specification language that supports the minimization or maximization of the use of a given role. Moreover, we provide a mapping of this language to integer linear programs and implement this mapping. We show how the resulting tool can be used analyze aspects of role specifications such as feasibility and optimality.

Similarity-Aware Query Processing in Sensor Networks

Ping Xia, Panos K. Chrysanthis and Alexandros Labrinidis

Department of Computer Science
University of Pittsburgh
Pittsburgh, PA, USA
{*pxia, panos, labrinid*}*@cs.pitt.edu*

We assume a sensor network with data-centric storage, where sensor data is stored within the sensor network and ad hoc queries are disseminated and processed inside the network. In such an environment, there are often similarities among submitted queries. Using current solutions, similar queries may have to go through the same expensive query processing steps thus wasting energy. In this paper, we propose a similarity-aware query processing scheme (SAQP) that materializes previous query results within the sensor network and utilizes these materialized results to answer future similar queries. Through simulation, we demonstrate that our SAQP scheme reduces energy consumption on queries with negligible increase in response time, and without compromising the quality of data.

An Optimal Approach to the Task Allocation Problem on Hierarchical Architectures

Alexander Metzner[1], Martin Fraenzle[2], Christian Herde[2] and Ingo Stierand[2]

[1]*Division Safety Critical Systems*
OFFIS e.V.
Oldenburg, Germany
metzner@informatik.uni-oldenburg.de

[2]*Dep. of Computer Science*
Carl-von-Ossietzky University
Oldenburg, Germany
{*fraenzle, herde, stierand*}*@informatik.uni-oldenburg.de*

We present a SAT-based approach to the task and message allocation problem of distributed real-time systems with hierarchical architectures. In contrast to the heuristic approaches usually applied to this problem, our approach is guaranteed to find an optimal allocation for realistic task systems running on complex target architectures. Our method is based on the transformation of such scheduling problems into nonlinear integer optimization problems. The core of the numerical optimization procedure we use to discharge those problems is a solver for arbitrary Boolean combinations of integer constraints. Optimal solutions are obtained by imposing a binary search scheme on top of that solver. Experiments show the applicability of our approach to industrial-size task systems, which are mapped to heterogeneous hierarchical hardware architectures.

Schedulability Analysis of AADL Models

Oleg Sokolsky[1], Insup Lee[1] and Duncan Clarke[2]

[1]*Dept. of Computer and Info. Science*
University of Pennsylvania
Philadelphia, PA, USA
{*sokolsky, lee*}*@cis.upenn.edu*

[2]*Fremont Associates*
Camden, SC, USA
dclarke@fremontassociates.com

The paper discusses the use of formal methods for the analysis of architectural models expressed in the modeling language AADL. AADL describes the system as a collection of interacting components. The AADL standard prescribes semantics for the thread components and rules of interaction between threads and other components in the system. We present a semantics-preserving translation of AADL models into the real-time process algebra ACSR, allowing us to perform schedulability analysis of AADL models.

Timed Automata Based Analysis of Embedded System Architectures

Martijn Hendriks[1] and Marcel Verhoef[1,2]

[1]*ICIS*
Radboud University Nijmegen
Nijmegen, The Netherlands
Martijn.Hendriks@cs.ru.nl, Marcel.Verhoef@chess.nl

[2]*Chess Information Technology BV*
Haarlem, The Netherlands

We show that timed automata can be used to model and to analyze timeliness properties of embedded system architectures. Using a case study inspired by industrial practice, we present in detail how a suitable timed automata model is composed. Exact upper bounds on the timeliness properties can be found with the Uppaal model checker for a number of usage scenarios. We compare our results with a few other performance modeling techniques. This comparison shows that, if the state space of the model is tractable, Uppaal gives the most accurate results at similar cost. The proposed modeling strategy can be automated, which alleviates the difficulty and error-proneness of manually constructing timed automata models.

Time Abstraction in Timed μCRL à la Regions

Jan Friso Groote, Michel A. Reniers and Yaroslav S. Usenko

Department of Mathematics and Computer Science
Technical University of Eindhoven
Eindhoven, The Netherlands
{*J.F.Groote, M.A.Reniers, Y.S.Usenko*}*@tue.nl*

In this paper we present an idea to combine the best parts of the real-time verification methods based on timed automata (the use of *regions* and *zones*), and of the process-algebraic approach of the languages like LOTOS and μCRL. μCRL targets the specification of system behavior in a process-algebraic (ACP) style and deals with data elements in the form of abstract data types. In order to combine the two approaches we propose the following scheme. Both zones and regions, as well as the operations on them could be specified as the abstract data types in μCRL, either as *clock constraints* or as *difference-bound matrices*. A timed automata specification is a parallel composition of timed automata. We use the existing results to translate it to a parallel composition of timed μCRL processes. This translation uses a very simple sort *Time* to represent the real-time clock values. As the result we obtain a semantically equivalent specification in timed μCRL.

As the next step in our scheme, we aim at replacing all parameters of sort Time occurring in the resulting process equation by the parameters of sort Region or Zone. This can be done in a similar way as for timed automata. These data types are countable and because of the decidability results for timed automata only finitely many different values of these parameters will be reached. We could even go further, i.e. due to the fact that infinite state spaces can also be analyzed in μCRL, we could go beyond timed automata verification.

In the final step we transform the resulting timed process equation with regions to an untimed process equation with a finite underlying state space. This is achieved by applying *time-free abstraction* and *relativization* techniques. As a result, the existing untimed analysis tools in the μCRL Toolset could become applicable to the analysis of real-time systems.

Schedulability analysis of flows scheduled with FIFO: Application to the Expedited Forwarding class

Steven Martin[1] and Pascale Minet[2]

[1]*LRI, Paris-Sud University*
91405 Orsay, France
steven.martin@lri.fr

[2]*INRIA Rocquencourt*
78153, France
pascale.minet@inria.fr

In this paper, we are interested in real-time flows requiring quantitative and deterministic QoS (*Quality of Service*) guarantees. We focus more particularly on two QoS parameters: the worst case end-to-end response time and jitter. We consider a FIFO (*First In First Out*) scheduling of flows. The FIFO scheduling is the simplest one to implement and very used. We first establish a bound on the worst case end-to-end response time of any flow in the network, using the trajectory approach. We present an example illustrating our results. Finally, we show how to apply these results to the EF (*Expedited Forwarding*) class in a DiffServ (*Differentiated Services*) architecture.

Real-Time Systems for Multi-Processor Architectures

Éric Piel, Philippe Marquet, Julien Soula and Jean-luc Dekeyser

Laboratoire dinformatique fondamentale de Lille
Université des sciences et technologies de Lille
Lille, France
{*Eric.Piel, Philippe.Marquet, Julien.Soula, Jean-Luc.Dekeyser*}*@lifl.fr*

The ARTiS system is a real-time extension of the GNU/Linux scheduler dedicated to SMP (Symmetric Multi-Processors) systems. It allows to mix High Performance Computing and Real-Time. ARTiS exploits the SMP architecture to guarantee the preemption of a processor when the system has to schedule a real-time task. The implementation is available as a modification of the Linux kernel.

The basic idea of ARTiS is to assign a selected set of processors to real-time operations. A migration mechanism of non-preemptible tasks insures a latency level on these real-time processors. Furthermore, specific loadbalancing strategies permit ARTiS to benefit from the full power of the SMP systems: the real-time reservation, while guaranteed, is not exclusive and does not imply a waste of resources.

QoS-based Management of Multiple Shared Resource in Dynamic Real-Time Systems

Klaus Ecker, Frank Drews and Jens Lichtenberg

School of EECS
Ohio University
Athens, OH, USA
{*ecker, drews, lichtenj*}*@ohio.edu*

Dynamic real-time systems require adaptive resource management to accommodate varying processing needs. We address the problem of resource management with multiple shared resources for soft real-time systems consisting of tasks that have discrete QoS settings that correspond to varying resource usage and varying utility. Given an amount of available resource, the problem is to provide on-line control of the tasks' QoS settings so as to optimize the overall system utility. We propose several heuristic algorithms that will be shown to be compatible with the requirements imposed by our control theoretical resource management framework: (1) By only making incremental adjustments to QoS settings as available resources change, they provide low run-time complexity, making them suitable for use in on-line resource managers (2) Differences between actual utility and optimal utility do not accumulate over time, so there is no long-term degradation in performance. (3) The lower and upper bound on actual utility can be calculated dynamically based on current system conditions, and absolute bounds can be calculated statically in advance. (4) It is possible to respond to the actual resource possible, allowing all resources to be used and tolerating misspecification of task resource requirements.

Adaptability Management and Deterministic Scheduling of Media Flows on Parallel Storage Servers

Costas Mourlas

Dept. of Communication and Media Studies
University of Athens
Athens, Greece
mourlas@media.uoa.gr

We study a new design strategy for the implementation of ParallelMedia Servers with a predictable behavior. This strategy makes the timing properties and the quality of presentation of a set of media streams predictable. The proposed strategy provides deterministic guarantees and service reliability for each stream that cant be compromised by server contention. Our real-time scheduling approach exploits the performance of parallel environments and seems a promising method that brings the advantages of parallel computation in media servers. The proposed mechanism provides deterministic service for both Constant Bit Rate (CBR) and Variable Bit Rate (VBR) streams. We present an efficient placement strategy for data frames as well as an adaptability strategy that allows appropriate frames to be dropped without sacrificing the ability to present multimedia applications predictably in time. A prototype implementation of the proposed parallel media server illustrates the concepts of server allocation and scheduling of continuous media streams.

Workshop 3

Reconfigurable Architectures Workshop RAW 2006

3 RAW • Workshop on Reconfigurable Architectures

Workshop Description:
Run-Time and Dynamic Reconfiguration are characterized by the ability of underlying hardware architectures or devices to rapidly alter the functionalities of its components and the interconnection between them to suit the problem. Key to this ability is reconfiguration handling and speed. Though theoretical models and algorithms for them have established reconfiguration as a very powerful computing paradigm, practical considerations make these models difficult to realize. On the other hand, commercially available devices appear to have more room for exploiting run-time reconfiguration. An appropriate mix of the theoretical foundations of dynamic reconfiguration, and practical considerations, including architectures, technologies and tools supporting RTR is essential to fully reveal and exploit the possibilities created by this powerful computing paradigm.

Topics of interest:

- Models & Architectures
  - Theoretical Interconnect & Computational Models
  - RTR Models and Systems
  - RTR Hardware Architectures
  - Optical Interconnect Models
  - Simulation and Prototyping
  - Bounds and Complexity
- Algorithms & Applications
  - Algorithmic Techniques
  - Mapping Parallel Algorithms
  - Distributed Systems & Networks
  - Fault Tolerance Issues
  - Wireless and Mobile Systems
  - Automotive Applications
  - Infotainment & Multimedia
  - Biology Inspired Applications
- Technologies & Tools
  - Configurable Systems-on-Chip Energy Efficiency
  - Devices and Circuits
  - Reconfiguration Techniques
  - High Level Design Methods
  - System support
  - Adaptive Runtime Systems
  - Organic Computing

Workshop Chair:
Serge Vernalde, IMEC, Belgium

Program Chair:
Juergen Becker, Universitat Karlsruhe (TH), Germany

Steering Chair:
Viktor K. Prasanna, USC, USA

Publicity Chair (USA):
Ramachandran Vaidyanathan, Louisiana State University, USA

Publicity Chair (Europe, Asia):
Reiner Hartenstein, Kaiserslautern University of Technology, Germany

Program Committee:
Jeffrey Arnold, Adaptive Silicon Inc.,
Helena Krupnova, ST Microelectronics
Sergio Bampi, Univ. Federal do Rio Grande
Rudy Lauwereins, IMEC, Leuven
Jürgen Becker, Universität Karlsruhe
Philip Leong, Chinese Univ. Hong Kong
Pascal Benoit, LIRMM, Montpellier
Marnane Liam, University College
Mladen Berekovic, IMEC
Rong Lin, SUNY at Genesco
Neil Bergmann, Univ. of Queensland
Wayne Luk, Imperial College
Don Bouldin, University of Tennessee
Juergen Luka, DaimlerChrysler AG
Gordon Brebner, Univ. of Edinburgh
Patrick Lysaght, Xilinx
Klaus Buchenrieder, Universität der Bundeswehr München
M. Marek-Sadowska, UC Santa Barbara
Thomas Buechner, IBM
John McHenry, National Security Agency
Stephen Chappell, Celoxica
Martin Middendorf, Univ. of Leipzig
Brendan Cremen, Xilinx
Dietmar Mueller, Technische Universitaet Chemnitz
Luigi Carro, Univ. Federal do Rio Grande
Amar Mukherjee, U. Central Florida
Peter Y. K. Cheung, Imperial College
Vincent Mooney III, Georgia Tech.
Adreas Dandalis, Philips
Koji Nakano, Hiroshima University
Jose T. de Sousa, Tech. Univ. Lisbon
Ranjani Parthasarathi, Anna University, Chennai
Oliver Diessel, U. New South Wales
Steven Perry, Altera
Adam Donlin, Xilinx
Marco Platzner, Universität Paderborn
Pedro C. Diniz, USC//ISI
Cameron Patterson, Virginia Tech
Gilbert Edelin, Thales Research & Technology
Bernard Pottier, Université de Bretagne Occidentale
Hossam ElGindy,U. New South Wales
Franz Rammig, Universität Paderborn
Patrick Girard, LIRMM
Ricardo Reis, Univ. Federal do Rio Grande
Manfred Glesner, Darmstadt Univ. Technology
Hartmut Schmeck, Univ. Karlsruhe
Steve Guccione, Cmpware Inc.
Sakir Sezer, Queen's University
Klaus Harbich, Bosch
Gerard Smit, Univ. Twente
Reiner Hartenstein, Kaiserslautern, Univ. of Technology
V. Sridhar, Satyam Comp. Services Ltd.
Ulrich Heinkel, Lucent Technologies
Jürgen Teich, Univ. Erlangen-Nuremberg
Andreas Herkersdorf, Institute for Integrated Systems
Lionel Torres, LIRMM, Montpellier
Christian Hochberger, Dresden University of Technology
Jim Tørresen, Univ. of Oslo
Thomas Hollstein, Darmstadt Univ. of Technology
Jerry L. Trahan, Louisiana State Univ.
Mike Hutton, Altera, USA
R. Vaidyanathan, Louisiana State Univ.
Michael Hübner, Universität Karlsruhe (TH)
Milan Vasilko, Bournemouth Univ.
Mark Jones, Virginia Tech
Stamatis Vassiliadis Delft University of Technology
Theodore Karoubalis, Atmel
Brian Veale, University of Oklahoma
Udo Kebschull, Universität Heidelberg
Martin Vorbach, PACT Informationstechnologie
Andreas Koch, Technische Universität Braunschweig
Klaus Waldschmidt, Universität Frankfurt
Rainer Kress, Infineon Technologies
Norbert Wehn, Kaiserslautern Univ. of Technology
Wido Kruitzer, Philips

RAW Keynote 1: The Outer Limits: Reconfigurable Computing in Space and In Orbit

Maya Gokhale

Los Alamos National Laboratory
Los Alamos, NM, USA

Programmable hardware offers unique opportunities for flexible control and processing on board spacecrafts and satellites. Space missions dictate stringent requirements on size, weight, power, versatility and performance of the on-board data acquisition and computing resources. Reconfigurable FPGA-based devices, with the programmability of software and and speed/size approaching application-specific integrated circuits, make it possible to control and communicate with sensors, as well as process scientific data right on the spacecraft, sending only relevant information back home over the low bandwidth communications link.

Challenges to computing in harsh space environments abound: vibration, thermal cycling, heat dissipation, and radiation all take their toll on space electronics. In spite of these barriers, notable experiments in reconfigurable computing for space applications are being undertaken. These include NASA's reconfigurable scalable computing project, intended for planetary rovers, cameras, and other sensors; the Queensland University FedSat and its successors, using FPGAs for near real-time image processing, communications, and navigation; and the Cibola Flight Experiment, a Los Alamos National Laboratory experiment in on-orbit signal processing using radiation-tolerant FPGAs.

This talk will discuss the perils and possibilities of reconfigurable computing at the outer limits.

RAW Keynote 2: New Horizons of Very High Performance Computing (VHPC): Hurdles and Chances

Reiner Hartenstein

TU Kaiserslautern
Kaiserslautern, Germany
reiner@hartenstein.de

Reconfigurable Computing (RC) delivers the success story of the century. First launched by the hardware / software co-design scene by adopting FPGAs for embedded system design, now a huge second wave has reached a wide variety of scientific computing communities. Google's yawdropping hit rates illustrate the pervasiveness of Reconfigurable Computing, now also being adopted by supercomputing (Cray, sgi, etc.). From FPGA usage as accelerators, speed-up factors by up to four orders of magnitude and more are reported, as well as floor space requirements and electricity invoice amounts reduced by one order of magnitude and more. This is astonishing, since FPGAs and rDPAs have a substantially lower clock speed than microprocessors and an effective integration density being lower by four orders of magnitude: the Reconfigurable Computing Paradox. Algorithmic cleverness is the secret of success, based on software to configware migration mechanisms, striving away from memory-cycle-hungry instruction stream-based computing paradigms. Even higher speedup is achievable by using coarse-grained reconfigurable datapath arrays (rDPAs) available from a number of start-ups. With automatically partitioning configware / software cocompilers the desktop personal supercomputer is near.

The main benefit of RC, having replaced the use of hardwired accelerators, is their flexibility by non-procedural programmability. This also contributes to more recent developments in system architecture, which rely on processes of evolution, self-organization, adaptation and fault tolerance. The main hurdles on the way to heart-stopping new horizons of cheap highest performance are CS-related educational deficits causing the configware / software chasm and a methodology fragmentation between the different cultures of application domains. Since the von Neumann paradigm is loosing its dominance by emerging reconfigurable main processors using hardwired von Neumann coprocessors as auxiliary clerks, it is time for a curricular upgrade. Current CS curricula do not sufficiently meet their transdisciplinary responsibility. The talk gives a survey on fundamental issues in RC and on new directions in CS-related curricula, focused on a dual paradigm organic computing approach.

Analysis of a Reconfigurable Network Processor

Christoforos Kachris and Stamatis Vassiliadis

Computer Engineering Lab
Delft University of Technology
Delft, The Netherlands
{*kachris, stamatis*}*@ce.et.tudelft.nl*

In this paper an analysis of a dynamically reconfigurable processor is presented. The network processor incorporates a processor and a number of co-processors that can be connected to the processor either directly or using a shared bus. The analysis investigates the configuration (in terms of co-processor distributions and interface), formulates the throughput that meets the network demands and the constraints of the platform (area, bus bandwidth, etc.) and takes into account the reconfiguration overhead. To find the configuration that meets the constraints, the platform is formulated into integer linear programming equations. Furthermore, the results of two case studies are presented, for a soft- and a hard- IP core processor, that uses three flows with different processing requirements (IP forward, encryption and media processing). In each case the number and the type of co-processors is shown in terms of the network distribution and the average packet size. Finally, the mapping of the framework in the Xilinx FPGA platform is discussed.

Performance and Power Analysis of Time-multiplexed Execution on Dynamically Reconfigurable Processor

Yohei Hasegawa[1], Shohei Abe[1], Shunsuke Kurotaki[1], Vu Manh Tuan[1], Naohiro Katsura[1], Takuro Nakamura[2], Takashi Nishimura[2] and Hideharu Amano[2]

[1]*Graduate School of Science and Technology*
Keio University
Yokohama, Japan
{*hasegawa, syouhei, kurotaki, vmtuan, katsura*}*@am.ics.keio.ac.jp*

[2]*Faculty of Science and Technology*
Keio University
Yokohama, Japan
{*nakamura, takashi, hunga*}*@am.ics.keio.ac.jp*

Dynamically Reconfigurable Processor (DRP) developed by NEC Electronics is a coarse grain reconfigurable processor that selects a datapath called a context from the on-chip repository of sixteen circuit configurations at run-time. The time-multiplexed execution based on the multicontext functionality is expected to drastically improve area and power efficiency. To demonstrate the impact of the time-multiplexed execution, we have implemented several stream applications on DRP with various context sizes. Throughout the evaluation based on real application designs, we analyzed the impact of the time-multiplexed execution on performance and power dissipation quantitatively.

2D Defragmentation Heuristics for Hardware Multitasking on Reconfigurable Devices

Julio Septién[1], Hortensia Mecha[1], Daniel Mozos[1] and Jesús Tabero[2]

[1]Dept. of Arquitectura de Computadores y Automatica
Universidad Complutense de Madrid
Madrid, Spain
{jseptien, horten, mozos}@dacya.ucm.es

[2]Dept. of Space Programs
Instituto Nacional de Tecnica Aeroespacial
Madrid, Spain
taberogj@inta.es

This paper focuses on the fragmentation problem produced in 2D run-time reconfigurable FPGAs when hardware multitasking management is considered. Though allocation heuristics can take fragmentation into account when a new task arrives, the free area becomes inevitably fragmented as the tasks finish and exit the FPGA. The main contributions of our work are a fragmentation metric able to estimate when the FPGA fragmentation status has become critical, and several heuristics to decide when to perform defragmentation and how to perform it. This defragmentation heuristics can be of a preventive kind, driven by alarms that fire when isolated islands appear or a high fragmentation status is reached. It can be also an on-demand process produced when a task allocation fails though there is enough free area in the FPGA to accommodate it.

A Cost-Effective Context Memory Structure for Dynamically Reconfigurable Processors

Masayasu Suzuki, Yohei Hasegawa, Vu Manh Tuan, Shohei Abe and Hideharu Amano

Graduate School of Science and Technology
Keio University
Yokohama, Japan
{masayasu, hasegawa, vmtuan,
shohei, hunga}@am.ics.keio.ac.jp

Multicontext reconfigurable processors can switch its configuration in a single clock cycle by providing a context memory in each of the processing elements. Although these processors have proven to be powerful in many applications, the number of contexts is often not enough.

The context translation table which translates the global instruction pointer, or the global logical context number, into a local physical context number is proposed to realize a larger application while reducing the actual context memories. Our evaluation using NEC Electronics' DRP-1 shows that the proposed method is effective when the size of the tile is small and the number of context is large. In the most efficient case, the required number of contexts is reduced to 25%, and the total amount of configuration data becomes 6.9%.

The template configuration method which extends this idea harnesses the power of multicontext devices by storing basic contexts as *templates* and combining them to form the actual contexts. While effective in theory, our evaluation shows that the return in adopting such mechanisms in more finer processors as the DRP-1 is minimal where the size of the context memory adds up relative to the number of processing units.

Performance of FPGA Implementation of Bit-split Architecture for Intrusion Detection Systems

Hong-Jip Jung, Zachary K. Baker and Viktor K. Prasanna

Electrical Engineering Systems
University of Southern California
Los Angeles, CA, USA
{*hongjung, zbaker, prasanna*}*@usc.edu*

The use of reconfigurable hardware for network security applications has recently made great strides forward as Field-Programmable Gate Array (FPGA) devices have provided larger and faster resources. The performance of an Intrusion Detection System is dependent on two metrics: throughput and the total number of patterns that can fit on a device.

In this paper, we consider the FPGA implementation details of the bit-split string-matching architecture. The bit-split algorithm allows large hardware state machines to be converted into a form with much higher memory efficiency. We have extended the architecture to satisfy the requirements of the IDS state-of-the-art.

We show that the architecture can be effectively optimized for FPGA implementation by making some changes to the parameters governing the pattern loading within the modules as well new interface hardware for communicating with an external controller. The overall performance (bandwidth * number of patterns) is competitive against other memory-based FPGA string matching architectures.

A Configuration Memory Hierarchy for Fast Reconfiguration with Reduced Energy Consumption Overhead

Elena Perez Ramo[1], Javier Resano[1], Daniel Mozos[1] and Francky Catthoor[2,3]

[1]*DACYA*
Complutense
Madrid, Spain
{*eperez, javier1, mozos*}*@dacya.ucm.es*

[2]*IMEC*
IMEC
Leuven, Belgium
catthoor@imec.be

[3]*ESAT*
K.U.Leuven
Leuven, Belgium

Currently run-time reconfigurable hardware offers really attractive features for embedded systems, such as flexibility, reusability, high performance and, in some cases, low-power consumption. However, the reconfiguration process often introduces significant overheads in performance and energy consumption. In our previous work we have developed a reconfiguration manager that minimizes the execution time overhead. Nevertheless, since the energy overhead is equally important, in this paper we propose a configuration memory hierarchy that provides fast reconfiguration while achieving energy savings. To take advantages of this hierarchy we have developed a configuration mapping algorithm and we have integrated it in our reconfiguration manager. In our experiments we have reduced the energy consumption 22.5% without introducing any performance degradation.

Maximum Edge Matching for Reconfigurable Computing

Markus Rullmann and Renate Merker

Department of EE and IT/Circuits and Systems Laboratory
Dresden University of Technology
Dresden, Germany
rullmann@iee.et.tu-dresden.de, merker@iee1.et.tu-dresden.de

Reconfiguration of tasks implies considerable overhead on the amount of configuration data and time. Much overhead is caused by redundant configuration generated by the design tools which implement similar structures in the designs on different resources. In this paper we propose a new method to identify structural similarities in tasks. Based on this information, we are able to generate automatically constraints to ensure that the place and route tools use identical resources. Thus we ensure that less redundant configuration is produced. In this paper we give a formal description of the underlaying *maximum edge matching problem* and show a method to solve it optimally. We derive a truncation criteria to restrict the search space efficiently. We also propose an Ant Colony Optimization based solution with a problem specific local heuristic and show that it performs optimal as well in our examples, but with considerable lower computational effort.

FPGA implementation of a license plate recognition SoC using automatically generated streaming accelerators

Nikolaos Bellas, Sek Chai, Malcolm Dwyer and Dan

Embedded System Research Lab
Motorola
Schaumburg, IL, USA
{bellas, chai}@labs.mot.com, {Malcolm.Dwyer, Dan.Linzmeier}@motorola.com

Modern FPGA platforms provide the hardware and software infrastructure for building a bus-based System on Chip (SoC) that meet the applications requirements. The designer can customize the hardware by selecting from a large number of pre-defined peripherals and fixed IP functions and by providing new hardware, typically expressed using RTL. Hardware accelerators that provide application-specific extensions to the computational capabilities of a system is an efficient mechanism to enhance the performance and reduce the power dissipation. What is missing is an integrated approach to identify the computationally critical parts of the application and to create accelerators starting from a high level representation with a minimal design effort. In this paper, we present an automation methodology and a tool that generates accelerators. We apply the methodology on an FPGA-based license plate recognition (LPR) system used in law enforcement. The accelerators process streaming data and support a programming model which can naturally express a large number of embedded applications resulting in efficient hardware implementations. We show that we can achieve an overall LPR application speed up from 1.2x to 2.6x, thus enabling real-time functionality under realistic road scenes.

A High-level Target-precise Model for Designing Reconfigurable HW Tasks

Maik Boden[1], Steffen Ruelke[1] and Juergen Becker[2]

[1]*Fraunhofer IIS, Branch Lab EAS*
Dresden, Germany
{*maik.boden, steffen.ruelke*}*@eas.iis.fraunhofer.de*

[2]*ITIV*
University of Karlsruhe
Karlsruhe, Germany
becker@itiv.uni-karlsruhe.de

The increasing complexity of embedded digital HW/SW systems, rising chip development and fabrication costs, and a shortened time-to-market require system-level design methods and the use of reconfigurable architectures. Our design method concerns the modelling of a system and its HW tasks at a high abstraction level. Using design patterns and macros, our library-based approach provides a consistent flow from an executable specification to its implementation. These templates ease the efficient application of partially run-time reconfigurable architectures. A case study depicts the high-level modelling of a HW task and its implementation in detail.

Rapid Development of High Performance Floating-Point Pipelines for Scientific Simulation

Gerhard Lienhart, Andreas Kugel and Reinhard Maenner

Dept. for Computer Science V
University of Mannheim
D-68131 Mannheim, Baden-Wuerttemberg, Germany
{*lienhart, kugel, maenner*}*@ti.uni-mannheim.de*

In the last years, FPGAs became capable of performing complex floating-point based calculations. For many applications, highly parallel calculation units can be implemented which deliver a better performance than general-purpose processors. This paper focuses on applications where the calculations can be done in a pipeline, as it is often the case for simulations. A framework for rapid design of such calculation pipelines is described. The central part is a Perl based code generator, which automatically assembles floating-point operators into synthesizable hardware description code where the generator is directed by a pipeline description file. The framework is supplemented by various floating-point operators and support modules, which allow generating ready-to-use pipelines. The code generator dramatically reduces development time and produces high-quality results. The performance of the framework is demonstrated by the implementation of pipelines for gravitational forces and hydrodynamics.

An Optimal Architecture for a DDC

Tjerk Bijlsma, Pascal T. Wolkotte and Gerard J. M. Smit

University of Twente
Department EEMCS
Enschede, The Netherlands
Bijlsma@cs.utwente.nl, {P.T.Wolkotte, G.J.M.Smit}@utwente.nl

Digital Down Conversion (DDC) is an algorithm, used to lower the amount of samples per second by selecting a limited frequency band out of a stream of samples. A possible DDC algorithm consists of two simple Cascading Integrating Comb (CIC) filters and a Finite Input Response (FIR) filter preceded by a modulator that is controlled with a Numeric Controlled Oscillator (NCO). Implementations of the algorithm have been made for five architectures, two Application Specific Integrated Circuits (ASIC), a General Purpose Processor (GPP), a Field Programmable Gate Array (FPGA), and the Montium Tile Processor (TP). All architectures are functionally capable of performing the algorithm. The differences between the architectures are their performance, flexibility and energy consumption. In this paper we compared the energy consumption of the architectures when performing the DDC algorithm. The ASIC is the best solution if digital down conversion is constantly required. When digital down conversion is needed only parts of the time, the Altera Cyclone II is the best solution due to its smaller technology size. In the spare time the reconfigurable architectures can be reconfigured for other tasks of today's multimedia devices.

Reconfigurable Memory Based AES Co-Processor

Ricardo Chaves[1,2], Georgi Kuzmanov[2], Stamatis Vassiliadis[2] and Leonel Sousa[1]

[1]*DEEC/INESC-ID*
IST
Lisbon, Portugal
{ricardo.chaves, las}@inesc-id.pt

[2]*Computer Engineering Lab.*
TUDelft
Delft, Nederlands
{G.Kuzmanov, s.vassiliadis}@ewi.tudelft.nl

We consider the AES encryption/decryption algorithm and propose a memory based hardware design to support it. The proposed implementation is mapped on the Xilinx Virtex II Pro technology. Both the byte substitution and the polynomial multiplication of the AES algorithm are implemented in a single dual port on-chip memory block (BRAM). Two AES encryption/decryption cores have been designed and implemented on a prototyping XC2VP20-7 FPGA: a completely unrolled loop structure capable of achieving a throughput above 34 Gbits/s, with an implementation cost of 3513 slices and 80 BRAMs; and a fully folded structure, requiring only 515 slices and 12 BRAMs, capable of a throughput above 2 Gbits/s. To evaluate the proposed AES design, its has been embedded in a polymorphic processor organization, as a reconfigurable co-processor. Comparisons to state-of-the-art AES cores indicate that the proposed unfolded core outperforms the most recent works by 34% in throughput and requires 68% less reconfigurable area. Experimental results of both folded and unfolded AES cores suggest over 560% improvement in the throughput/slice metric when compared to the recent AES related art.

Communication Concept for Adaptive Intelligent Run-Time Systems Supporting Distributed Reconfigurable Embedded Systems

Michael Ullmann and Jürgen Becker

Institut für Technik der Informationsverarbeitung (ITIV)
Universität Karlsruhe (TH)
Karlsruhe, Baden-Württemberg, Germany
{*ullmann, becker*}*@itiv.uni-karlsruhe.de*

Reconfigurable computing systems have already shown their abilities to accelerate embedded hardware/ software systems. Since standard processor-based embedded applications have come to their limits we need new concepts for controlling and managing embedded, possibly distributed, reconfigurable hardware/ software computing systems. Succeeding to previous papers which dealt with management aspects of run-time reconfigurable systems and related AI-approaches this contribution describes an approach and proof of concept of a transparent communication mechanism between the application layer and its possibly distributed and reconfigurable hardware/ software sub-function modules.

FPGA based Architecture for DNA Sequence Comparison and Database Search

Euripides Sotiriades, Christos Kozanitis and Apostolos Dollas

Microprocessor and Hardware Laboratory
Technical University of Crete
Chania, Greece
{*esot, kozanit, dollas*}*@mhl.tuc.gr*

DNA sequence comparison is a computationally intensive problem, known widely since the competition for human DNA decryption. Database search for DNA sequence comparison is of great value to computational biologists. Several algorithms have been developed and implemented to solve this problem efficiently, but from a user base point of view the BLAST algorithm is the most widely used one. In this paper we present a new architecture for the BLAST algorithm. The new architecture was fully designed, placed and routed. The post place-and-route cycle-accurate simulation, accounting for the I/O, shows a better performance than a cluster of workstations running highly optimized code over identical datasets. The new architecture and detailed performance results are presented in this paper.

Accelerating DTI Tractography using FPGAs

Aditya Kwatra[1], Viktor Prasanna[2] and Manbir Simgh[3]

[1]Dept. of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
kwatra@usc.edu

[2]Dept. of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
prasanna@usc.edu

[3]Depts. of Radiology and Biomedical Engineering
University of Southern California
Los Angeles, CA, USA
msingh@usc.edu

Diffusion Tensor Imaging (DTI) tractography in Magnetic Resonance Imaging (MRI) is a computationally intensive procedure, requiring on the order of tens of minutes to complete tractography of the entire brain. Tractography computations can be accelerated significantly by use of reconfigurable hardware, such as Field Programmable Gate Arrays (FPGAs). Such acceleration has the potential to lead to real-time tractography, which would greatly facilitate on-site diagnosis and acquisition of additional scans while the patient is still inside the scanner. In this paper we report the development of an FPGA based architecture to accelerate DTI tractography. We identify computationally intensive kernels and design pipelined implementations. Our performance analysis based on the developed architecture gives on the order of 100x speed-up over an optimized C-code based implementation of tractography on a state-of-the-art processor.

An Adaptive System-on-Chip for Network Applications

Roman Koch, Thilo Pionteck, Carsten Albrecht and Erik Maehle

Institute of Computer Engineering
University of Luebeck
Luebeck, Germany
{koch, pionteck, albrecht, maehle}@iti.uni-luebeck.de

This paper presents the hardware architecture of DynaCORE, a dynamically reconfigurable system-on-chip for network applications. DynaCORE is an application specific coprocessor for offloading computationally intensive tasks from a network processor. The system-on-chip architecture is based on an adaptable network-on-chip which allows the dynamic replacement of hardware modules as well as the adaptation of the on-chip communication structure. The coprocessor leverages the active partial reconfiguration feature of modern FPGAs in order to adapt to shifting demand patterns. An embedded general-purpose processor core within the coprocessor runs software which manages the configurations of the device. With reference to a prototypical implementation targeting a Xilinx Virtex-II Pro FPGA, this paper focuses on on-chip communication issues. Topics include the integration of PowerPC processor cores into the configurable logic as well as the mode of operation of the network-on-chip.

Dedicated Module Access in Dynamically Reconfigurable Systems

Hagemeyer, Jens[1], Kettelhoit, Boris[2] and Porrmann, Mario[3]

[1]*Heinz Nixdorf Institute*
University of Paderborn
Paderborn, Germany
jenze@hni.upb.de

[2]*Heinz Nixdorf Institute*
University of Paderborn
Paderborn, Germany
kettelhoit@hni.upb.de

[3]*Heinz Nixdorf Institute*
University of Paderborn
Paderborn, Germany
porrmann@hni.upb.de

Modern FPGAs, such as the Xilinx Virtex-II Series, offer the feature of partial and dynamic reconfiguration, allowing to load various hardware configurations (i.e., HW modules) during run-time. To enable communication with these modules and for controlling purposes, dedicated access to each module as well as dedicated signals to control the global communication are required. This paper discusses several ways of implementing dedicated signals and addresses the impact on dynamically reconfigurable systems. Two new approaches are introduced, which allow a permanent access to the modules and to the communication infrastructure even during reconfiguration.

Exploiting dynamic reconfiguration of platform FPGAs: Implementation issues

Miguel L. Silva[1] and João Canas Ferreira[1,2]

[1]*FEUP/DEEC*
University of Porto
Porto, Portugal
{mlms, jcf}@fe.up.pt

[2]*Inesc Porto*
Porto, Portugal

The effective use of dynamic reconfiguration requires the designer to address many implementation issues. The market introduction of feature-full platform FPGAs equipped with embedded CPU blocks expands the number of situations where dynamic reconfiguration may be applied to improve overall performance and logic utilization. The paper compares the design of two similar systems supporting dynamic reconfiguration and the issues that were addressed in their implementation. The first system supports 32-bit data transfers between CPU and the dynamically reconfigurable circuits. The other implementation supports 64-bit transfers, but its effective use is more complicated and several restrictions must be taken into account. The work includes a performance comparison of the two designs on several simple tasks, including pattern matching, image processing and hashing.

A Distributed Object System Approach for Dynamic Reconfiguration

Ronald Hecht, Stephan Kubisch, Harald Michelsen, Elmar Zeeb and Dirk Timmermann

Institute of Applied Microelectronics and Computer Engineering
University of Rostock
Rostock, Germany
{*ronald.hecht, stephan.kubisch, harald.michelsen, elmar.zeeb, dirk.timmermann*}*@uni-rostock.de*

Managing reconfigurable hardware resources at runtime is expected to be a new task for future operating systems. But due to the mixture of parallel and sequential parts of dynamically reconfigurable applications, it is not entirely clear so far, how to use and to program such systems. A new interpretation of dynamically reconfigurable applications is presented. It will be shown, that the parallel computing concept of distributed object systems may be adapted for dynamically reconfigurable architectures. This approach answers many open questions concerning communication, interruption, and relocation of reconfigurable modules. It is explored by means of an extended Linux operating system in conjunction with a SystemC model of a dynamically reconfigurable FPGA.

Elementary Block Based 2-Dimensional Dynamic and Partial Reconfiguration for Virtex-II FPGAs

Michael Hübner[1], Christian Schuck[2] and Jürgen Becker[3]

[1]*ITIV*
Karlsruhe (TH)
Karlsruhe, Baden Wuerttemberg, Germany
huebner@itiv.uni-karlsruhe.de

[2]*ITIV*
Karlsruhe (TH)
Karlsruhe, Baden Wuerttemberg, Germany
schuck@itiv.uni-karlsruhe.de

[3]*ITIV*
Karlsruhe (TH)
Karlsruhe, Baden Wuerttemberg, Germany
becker@itiv.uni-karlsruhe.de

The development of Field Programmable Gate Arrays (FPGAs) had tremendous improvements in the last few years. They were extended from simple logic circuits to complex Systems-on-Chip which enable the integration of complete microcontroller systems and their peripheral devices. Virtex-II FPGAs from Xilinx provide the possibility of dynamic and partial reconfiguration. This can be taken advantage of to substitute inactive parts of a hardware system and adapt the complete chip to a different requirement of an application while run-time. Existing approaches allow reconfiguration of slot based systems while run-time. Unfortunately such systems suffer from the fact, that fixed sized reconfigurable slots are not completely utilized by all functional blocks. Therefore a new 2-dimensional approach is necessary to optimize the placement of functions on the reconfiguration area for the FPGA. Benefit is a reduced chip size which leads to a reduction of power dissipation. This paper describes the method and procedure to include a 2-dimensional placement of reconfigurable blocks and the integration to a run-time system.

Physically-aware Exploitation of Component Reuse in a Partially Reconfigurable Architecture

Love Singhal and Elaheh Bozorgzadeh

Center for Embedded Computer Systems
University of California, Irvine
Irvine, California, United States
{*lsinghal, eli*}*@ics.uci.edu*

The major drawback of partial dynamic reconfiguration is the reconfiguration delay overhead. To reduce the reconfiguration bits between two consecutive implementations, design components are reused. In this paper, we propose a floorplanner to support two-dimensional partial reconfiguration. Our floorplanner handles many features like mapping, selection and placement of the fixed components, and interconnect planning between the fixed and reconfigurable components. We implemented a sequence of dataflow graphs on Xilinx Virtex-4 devices. The component reuse results in more than 50% savings in reconfiguration bits. The results show a need to tune the physical design tools for minimizing runtime reconfiguration delay overhead.

Partitioned Scheduling of Periodic Real-Time Tasks onto Reconfigurable Hardware

Klaus Danne and Marco Platzner

Dept. of Computer Science
Paderborn University
Paderborn, NRW, Germany
{*danne, platzner*}*@upb.de*

Reconfigurable hardware devices, such as FPGAs, are increasingly used in embedded systems. To utilize these devices for real-time work loads, scheduling techniques are required that generate predictable task timings.

In this paper, we present a partitioning-EDF (earliest deadline first) approach to find such schedules. The FPGA area is partitioned along one dimension into slots. The tasks are partitioned into groups. Then, each group is scheduled to exactly one slot using the EDF rule. We show that the problem of finding an optimal partitioning is related to the well-known 2-dimensional level bin-packing problem. We extend a previously reported ILP model to solve our partitioning problem to optimality. By a simulation study we demonstrate that the partitioning-EDF approach is able to find feasible schedules for most task sets with a system utilization of up to 70%. Additionally, we allow a task to be realized in alternative implementations. A simulation study reveals that the scheduling performance increases considerably if three instead of one task variants are considered. Finally, we model and study the impact of the device reconfiguration time on the scheduling performance.

A Pattern Selection Algorithm for Multi-Pattern Scheduling

Yuanqing Guo, Cornelis Hoede and Gerard J.m. Smit

Faculty of EEMCS
University of Twente
Enschede, the Netherlands
{*y.guo, c.hoede, g.j.m.smit*}*@utwente.nl*

The multi-pattern scheduling algorithm is designed to schedule a graph onto a coarse-grained reconfigurable architecture, the result of which depends highly on the used patterns. This paper presents a method to select a near-optimal set of patterns. By using these patterns, the multi-pattern scheduling will result in a better schedule in the sense that the schedule will have fewer clock cycles.

Mapping DSP Applications on Processor Systems with Coarse-Grain Reconfigurable Hardware

Michalis D. Galanis, Gregory Dimitroulakos and Costas E. Goutis

VLSI Design Laboratory, ECE Department
University of Patras
Patras, Achaia, Greece
{*mgalanis, dhmhgre, goutis*}*@ee.upatras.gr*

In this paper, we present performance results from mapping five real-world DSP applications on an embedded system-on-chip that incorporates coarse-grain reconfigurable logic with an instruction-set processor. The reconfigurable logic is realized by a 2-Dimensional Array of Processing Elements. A mapping flow for improving applications performance by accelerating critical software parts, called kernels, on the Coarse-Grain Reconfigurable Array is proposed. Profiling is performed for detecting critical kernel code. For mapping the detected kernels on the reconfigurable logic a priority-based mapping algorithm has been developed. The experiments for three different instances of a generic system show that the speedup from executing kernels on the Reconfigurable Array ranges from 9.9 to 151.1, with an average value of 54.1, relative to the kernels execution on the processor. Important overall application speedups, due to the kernels acceleration, have been reported for the five applications. These overall performance improvements range from 1.3 to 3.7, with an average value of 2.3, relative to an all-software execution.

VoC: A Reconfigurable Matrix for Stereo Vision Processing

Ricardo Pezzuol Jacobi[1], Renato Barreto Cardoso[2] and Geovany Borges[2]

[1]*Dept. of Computer Science*
University of Brasilia
Brasilia, DF, Brazil
ricardo@exatas.unb.br

[2]*Dept. of Electrical Engineering*
University of Brasilia
Brasilia, DF, Brazil
renbarcar@ig.unb.br, gaborges@ene.unb.br

This paper presents a reconfigurable matrix VoC that can be applied to stereo vision computation. VoC accelerates block pixel matching by providing a highly parallel implementation of the Sum of Absolute Differences metric. Reconfigurability allows VoC to deal with different block sizes, ranging from a single 7x7 SAD computation to 9 simultaneous 5x5 block computations. The pipelined version mapped to Xilinx FPGA could be simulated at 158 MHz, producing 1,42 billion matchings per second.

Selection of Instruction Set Extensions for an FPGA Embedded Processor Core

Brian F. Veale[1], John K. Antonio[1], Monte P. Tull[2] and Sean A. Jones[1]

[1]*School of Computer Science*
University of Oklahoma
Norman, OK, USA
{veale, antonio, sean.jones}@ou.edu

[2]*School of Electrical and Computer Engineering*
University of Oklahoma
Norman, OK, USA
tull@ou.edu

A design process is presented for the selection of a set of instruction set extensions for the PowerPC 405 processor that is embedded into the Xilinx Virtex Family of FPGAs. The instruction set of the PowerPC 405 is extended by selecting additional instructions from the full 32-bit PowerPC instruction set architecture (ISA), of which the PowerPC 405 ISA is a subset. The selected instructions are supported in hardware using the reconfigurable resources of the FPGA. The proposed design process gathers execution statistics for a target application through profiling or simulation. These statistics are then used to estimate the speedup that would be achieved if selected instructions from the full PowerPC ISA are added to the ISA of the PowerPC 405 processor. An experimental study of two embedded benchmarks show significant speedup when this approach is used to extend the PowerPC 405 processor to support various floating-point operations through the use of floating-point cores developed by QinetiQ.

Dynamic Configuration Steering for a Reconfigurable Superscalar Processor

Nick A. Mould[1], Brian F. Veale[2], Monte P. Tull[1] and John K.antonio[2]

[1] School of Electrical and Computer Engineering
University of Oklahoma
Norman, OK, USA
{nick_mould, tull}@ou.edu

[2] School of Computer Science
University of Oklahoma
Norman, OK, USA
{veale, antonio}@ou.edu

A new dynamic vector approach for the selection and management of the configuration of a reconfigurable superscalar processor is proposed. This new method improves on previous work that used steering vectors to guide the selection of functional units to be loaded into the processor. Dependencies among instructions in the instruction buffer are analyzed to enable a new scoring method. The dynamic vector technique is shown to reduce the amount of reconfiguration required while preserving execution resources. Simulation results reveal that, given enough configurable space, the configuration of the processor approaches a stable state.

Automatic Application-Specific Microarchitecture Reconfiguration

Shobana Padmanabhan, Ron K. Cytron, Roger D. Chamberlain and John W. Lockwood

Department of Computer Science and Engineering
Washington University
St. Louis, MO, USA
{shobana, lockwood}@arl.wustl.edu, cytron@cs.wustl.edu, roger@wustl.edu

Applications for constrained embedded systems are subject to strict time constraints and restrictive resource utilization. With soft core processors, application developers can customize the processor for their application, constrained by resources but aimed at high application performance. With such freedom in the design space of the processor, however, comes complexity. We present here an automatic optimization technique that helps the developers with the processor microarchitecture customization.

A naive approach exploring all possible configurations is exponential with the number of parameters and hence is clearly infeasible, even with only tens of reconfigurable parameters. Instead, our approach runs in time that is linear with the number of parameter values, based on an assumption of parameter independence. This makes the approach feasible and scalable. For the dimensions that we customize, namely application runtime and hardware resources, we formulate their costs as a constrained binary integer nonlinear optimization program. Though the results are not guaranteed to be optimal, we find they are near-optimal in practice. Our technique itself is general and can be applied to other design-space exploration problems.

Accelerating CABAC Encoding for Multi-standard Media with Configurability

Oskar Flordal, Di Wu and Dake Liu

Department of Electrical Engineering
Linköping University
Linköping, Sweden
oskar@flordal.net, {diwu, dake}@isy.liu.se

This paper presents the study of how to accelerate CABAC encoding for emerging heterogeneous multimedia applications. The latest image and video compression standards such as JPEG2000 and H.264 both have adopted Context Adaptive Binary Arithmetic Coding to achieve performance enhancement. However, CABAC requires high computing power. After investigating computational complexity of CABAC coding, firstly, instruction level acceleration is elaborated. Secondly, a configurable accelerator for CABAC encoding in multiple standards is proposed. Benchmarking performance and implementation cost is also addressed.

Exploiting Processing Locality through Paging Configurations in Multitasked Reconfigurable Systems

Mohamed Taher and Tarek El-ghazawi

ECE
George Washington University
Washington, DC, USA
{mtaher, tarek}@gwu.edu

FPGA chips in reconfigurable computer systems are used as malleable coprocessors where components of a hardware library of functions can be configured as needed. As the number of hardware functions to be configured typically exceeds the underlying chip area during the execution of an application, previous efforts have introduced configuration caching. Those efforts, however, can exploit either spatial or temporal processing locality. In this work, we propose a technique suitable for multitasking and for cases of single applications that can change the course of processing in a non-deterministic fashion based on data. In order to exploit both spatial and temporal processing locality, simultaneously, the proposed model groups hardware functions into hardware configuration blocks (pages) of fixed size, where multiple pages can be configured on a chip simultaneously. By grouping only related functions that are typically requested together, processing spatial locality can be exploited. Temporal locality is exploited through page replacement techniques. Data mining techniques were used to group related functions into pages. Standard, replacement algorithms as those found in caching were considered. Simulations, as well as emulation using the Cray XD1 reconfigurable high-performance computer were used in the experimental study. The results show a significant improvement in performance using the proposed paging technique.

Investigation into Programmability for Layer 2 Protocol Frame Delineation Architectures

Ciaran Toal and Sakir Sezer

ECIT
Queens University Belfast
Belfast, N. Ireland
ciaran.toal@ee.qub.ac.uk, s.sezer@qub.ac.uk

This paper presents the design and study of reconfigurable architectures for two data-link layer frame delineation techniques used for ATM and GFP. The architectures are targeted to Altera Stratix II FPGA technology and are investigated in terms of performance and area. This work addresses the potential for incorporating programmability into custom purpose architectures that could enable the same processing hardware to be used for processing multiple protocols.

Multi-level Reconfigurable Architectures in the Switch Model

Sebastian Lange and Martin Middendorf

Department of Computer Science
University of Leipzig
Leipzig, Germany
langes@informatik.uni-leipzig.de

In this paper we study multi-level dynamically reconfigurable architectures. These are extensions of standard reconfigurable architectures where ordinary reconfiguration operations correspond to the lowest reconfiguration level. On each higher reconfiguration level the reconfiguration capabilities of the reconfigurable resources that are available on the level directly below can be reconfigured. We show that the problem to find optimal reconfigurations with an arbitrary number of reconfiguration levels can be found in polynomial time for the switch cost model. The problem of finding the optimal number of reconfiguration levels is shown to be solvable in polynomial time on homogenous multi-level architectures but it becomes NP-hard for heterogenous multi-level architectures. Moreover, we present experimental results for some example problems on a simple test architecture.

Platform-based FPGA Architecture: Designing High-Performance and Low-Power Routing Structure for Realizing DSP Applications

Konstantinos Siozios, Konstantinos Tatas, Dimitrios Soudris and Antonios Thanailakis

VLSI Design and Testing Center, Department of Electrical and Computer Engineering
Democritus University of Thrace
Xanthi, Xanthi, Greece
{*ksiop, ktatas, dsoudris, athanail*}*@ee.duth.gr*

The novel design of an efficient FPGA interconnection architecture with multiple Switch Boxes (SB) and hardwired connections for realizing data intensive applications (i.e. DSP applications), is introduced. For that purpose, after exhaustive exploration, we modify the routing architecture through efficient selection of the appropriate switch box with hardwired connections, taking into account the statistical and spatial routing restrictions of DSP applications mapped onto FPGA. More specifically, we propose a new technique for selecting the appropriate combination of switch boxes, depending on the localized performance and power consumption requirements of each specific region of FPGA architecture. In order to perform the mapping, we developed a novel algorithm, which takes into account the modified architectural routing features. This algorithm was implemented within a new tool called EX-VPR. Using a number of DSP applications, extensive comparison study of various combinations of switch boxes in terms of total power consumption, performance, PowerDelay product prove the effectiveness of the proposed approach.

Multi-Clock Pipelined Design of an IEEE 802.11a Physical Layer Transmitter

Maryam Mizani[1] and Daler Rakhmatov[2]

[1]*Department of Electrical and Computer Engineering*
University of Victoria
Victoria, BC, Canada
mmizani@ece.uvic.ca

[2]*Department of Electrical and Computer Engineering*
University of Victoria
Victoria, BC, Canada
daler@ece.uvic.ca

Among different wireless LAN technologies 802.11a has recently become popular due to its high throughput, large system capacity, and relatively long range. In this paper, we propose a reconfigurable architecture for the 802.11a physical layer transmitter, which has low latency and low power consumption due to its pipelined structure. Data from the MAC layer can continuously flow through the pipeline without excessive buffering and handshaking within the physical layer. Dynamically reconfiguring this architecture to work at any data rate supported by 802.11a (eight different modes) can be performed within a few cycles, simply by adjusting the period of two clock signals and changing the value of a 3-bit control signal. Our architecture, prototyped on a Xilinx Virtex-II Pro FPGA, occupies the area of 2059 slices and is estimated to consume 500 mW. These figures can be improved substantially in custom ASIC implementations.

Posters: Reconfigurable Architectures Workshop

On-chip and On-line Self-Reconfigurable Adaptable Platform: the Non-Uniform Cellular Automata Case

Andres Upegui and Eduardo Sanchez

Reconfigurable Digital Systems Group
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
{*andres.upegui, eduardo.sanchez*}*@epfl.ch*

In spite of the high parallelism exhibited by cellular automata architectures, most implementations are usually run in software. For increasing execution parallelism, hardware implementations on FPGAs have been proposed, under the cost of being un-flexible, and inefficient in terms of resource utilization. In this paper we present a platform for evolving CA by exploiting the partial re-configurability of current commercial FPGAs. Our implementation includes an on-chip soft-processor that generates a partial bitstream, reconfigures the FPGA, and computes the fitness. After finding a good individual, the evolved CA can be used as a peripheral for performing useful computation. As case study we present CA co-evolution for a random number generator and for the firefly synchronization problem.

Increasing Analog Programmability in SoCs

Erik Schüler and Luigi Carro

Electrical Engineering
Universidade Federal do Rio Grande do Sul
Porto Alegre, RS, Brazil
{*eschuler, carro*}*@eletro.ufrgs.br*

The use of programmability in Systems-on-Chip (SoC) brings as the main advantage the possibility of reducing the time-to-market and the cost of design, specially when different systems and functions must cover different markets, going from low-power and low-frequency instrumentation to high frequency communication. This paper presents a technique that can be used to increase the analog programmability in a SoC, also allowing one to integrate more analog functions, while guaranteeing the use of the analog part in a larger range of applications. Practical results are presented showing that the proposed technique can be used from DC to RF applications.

Partial and dynamic Reconfiguration of FPGAs : a top down design methodology for an automatic implementation

Florent Berthelot, Fabienne Nouvel and Dominique Houzet

IETR
INSA
Rennes, Bretagne, FRANCE
florent.berthelot@ens.insa-rennes.fr, {fabienne.nouvel, dominique.houzet}@insa-rennes.fr

Dynamic reconfiguration of FPGAs enables systems to adapt to changing demands. This paper concentrates on how to take into account specificities of partially reconfigurable components during the high level Adequation Algorithm Architecture process. We present a method which generates automatically the design for both partially and fixed parts of FPGAs. The runtime reconfiguration manager which monitors dynamic reconfigurations, uses prefetching technic to minimize reconfiguration latency of runtime reconfiguration. We demonstrate the benefits of this approach through the design of a dynamic reconfigurable MC-CDMA transmitter implemented on a Xilinx Virtex2. This methodology is architecture's manufacturers independant and can be applied to different FPGAs.

Architecture of a Multi-Context FPGA Using a hybrid Multiple-Valued/Binary Context Switching Signal

Yoshihiro Nakatani, Masanori Hariyama and Michitaka Kameyama

Graduate School of Information Sciences
Tohoku University
Sendai, Miyagi, Japan
yoshi@kameyama.ecei.tohoku.ac.jp, {hariyama, kameyama}@ecei.tohoku.ac.jp

Multi-context FPGAs have multiple memory bits per configuration bit forming configuration planes for fast switching between contexts. Large amount of memory causes significant overhead in area and power consumption. This paper presents two key technologies. The first is a floating-gate-MOS functional pass gate that merges storage and switching functions area-efficiently. The second is the use of a hybrid multiple-valued/binary context switching signal that eliminates redundancy of a conventional multi-context (MC) switch with high scalability. The transistor count of the proposed MC-switch is reduced to 7% in comparison with that of a SRAM-based one.

A High Level SoC Power Estimation Based on IP Modeling

David Elleouet[1], Nathalie Julien[2] and Dominique Houzet[1]

[1]*IETR*
INSA
RENNES, FRANCE
david.elleouet@ens.insa-rennes.fr,
dominique.houzet@insa-rennes.fr

[2]*LESTER*
University of South Britany
LORIENT, FRANCE
nathalie.julien@univ-ubs.fr

Current electronic system design requires to be concerned with power consumption consideration. However, in a lot of design tools, the application power consumption budget is estimated after RTL synthesis. We propose in this article a methodology based on measurements which allows to model the application power consumption with architectural and algorithmic parameters. So, the modeled applications can be added in a library in order to help the system designer to determine early in the design flow the best adequacy between high performances and low power consumption.

Implementation of a Reconfigurable Hard Real-Time Control System for Mechatronic and Automotive Applications

Steffen Toscher, Roland Kasper and Thomas Reinemann

Institute of Mobile Systems
University Magdeburg
Magdeburg, Germany
{*steffen.toscher, roland.kasper, thomas.reinemann*}*@mb.uni-magdeburg.de*

Control algorithms implemented directly in hardware take advantage of parallel signal processing. Furthermore, implementing controller functionality in reconfigurable hardware facilitates modification of controller structure and parameters during run-time. In this paper, we introduce an implemented and tested reconfigurable hard real-time control system based on an FPGA device. It supports dynamic partial reconfiguration of controller functionality by self-reconfiguration mechanisms. Self-reconfiguration is performed using an internal configuration interface. We also present sophisticated on-chip and off-chip communication solutions. Specification of controller functionality involves Finite State Machines (FSMs) and comprises parts of the distributed communication and reconfiguration solution.

Run-Time Reconfiguration of Communication in SIMD Architectures

Hamed Fatemi[1], Bart Mesman[1], Henk Corporaal[1], Twan Basten[1] and Pieter Jonker[2]

[1]*Electrical Department*
Eindhoven
Eindhoven, The Netherlands
h.fatemi@tue.nl, b.mesman@tue.nl , h.corporaal@tue.nl , a.a.basten@tue.nl

[2]*Imaging Science and Technology Department*
Delft
Delft, The Netherlands
p.p.jonker@tnw.tudelft.nl

SIMD processors are increasingly used in embedded systems for multi-media applications because of their advantages with regard to area- and energy-efficiency. Communication between the processing elements in an SIMD processor has remained a cause of inefficiency however, the SIMD concept prescribes that all processing elements communicate in the same clock cycle. Existing SIMD architectures solve this problem either by multi-hop communication (causing cycle overhead), or by a fully connected communication network (causing area overhead).

In order to solve the communication bottleneck we have introduced a new SIMD architecture (RC-SIMD) with a set of delay-lines in the instruction bus, causing the accesses to the communication network to be distributed over time. We can (re-)configure the size and number of delay-lines, where a specific configuration represents a trade-off between number of clock cycles and size of a clock period. The reconfiguration process is simple, the reconfiguration time is typically around 30 clock cycles (which is far less than 1% of the typical execution time of algorithms), and the added configuration hardware is less than 2%. Furthermore, an experimental study shows that our reconfigurable architecture achieves (on average) more than 10% performance improvement over a non-reconfigurable architecture.

Coupling of a Reconfigurable Architecture and a Multithreaded Processor Core with Integrated Real-Time

Sascha Uhrig[1], Stefan Maier[2], Georgi Kuzmanov[3] and Theo Ungerer[1]

[1]*Institute of Computer Science*
Augsburg
Augsburg, Germany
{uhrig, ungerer}@informatik.uni-augsburg.de

[2]*Infineon*
München, Germany
stefan.maier@infineon.com

[3]*Computer Engineering Laboratory and Computer Science Department*
TU Delft
Delft, The Netherlands
G.Kuzmanov@ewi.tudelft.nl

This paper defines a real-time capable interface between the simultaneous multithreaded CarCore processor and a MOLEN-based reconfigurable unit. CarCore is an IP core that enables simultaneous execution of one hard-real-time thread and further non-real-time threads. The coupling described in this paper extends CarCore by a reconfigurable hardware such that both can execute different threads simultaneously, while the real-time behavior of the hard-real-time thread is not harmed. The challenge is the design of a common memory interface for both, the CarCore and the reconfigurable hardware, such that memory operations fulfil hard-real-time constraints. Experimental results with an MJPEG benchmark show an overall application speedup of 2.75 which approaches the theoretically attainable maximum speedup of 2.78.

Reconfiguration of Embedded Java Applications

João Cláudio Soares Otero, Flávio Rech Wagner and Luigi Carro

Instituto de Informática
Universidade Federal do Rio Grande do Sul
Porto Alegre, Rio Grande do Sul, Brazil
{*jcotero, flavio, carro*}*@inf.ufrgs.br*

This work presents the development of a coarse grain reconfigurable unit to be coupled to a native Java microcontroller, which is designed for an optimized execution of the embedded application. Code fragments to be accelerated through this unit are identified by profiling the application. The unit is able to explore ILP in a simple way and allows for Java compatibility, while also reducing the number of executed instructions, thus improving the performance with simultaneous energy savings. In many cases, as demonstrated by experiments, it also allows for smaller power consumption.

Speech Silicon AM: An FPGA-Based Acoustic Modeling Pipeline for Hidden Markov Model based Speech Recognition

Jeffrey W. Schuster, Raymond Hoare and Kshitij Gupta

Digital Prototyping Laboratory
University of Pittsburgh
Pittsburgh, PA, USA
{*jws52, hoare, ksg3*}*@pitt.edu*

This paper presents the design of a FPGA-based hardware co-processor capable of performing continuous speech recognition on medium sized vocabularies in real-time. The system is based on models derived through analysis of the SPHINX 3 large vocabulary continuous speech recognition engine designed by CMU. By creating a custom, input-driven pipeline for performing the calculations we were able to maximize the throughput of the system while simultaneously minimizing the number of pipeline stalls. By using embedded multiply-accumulate ASIC cells in the FPGA and using advanced placement techniques we were able to reach post place-and-route speeds even greater than those necessary for real-time operation while operating at maximum workload. Further, we use input control vectors, rather than internal finite state machines, to shut down portions of the pipeline when they were not in use to help mitigate power consumption. These results combined with the ability to reprogram the system for different recognition tasks serve to create a system capable of performing real-time speech recognition in a vast array of environments. We synthesized our hardware to a Xilinx Virtex 4 SX and a Xilinx Spartan 3 FPGA. Functional verification was through post place-and-route simulations.

Implementation of a Programmable Array Processor Architecture for Approximate String Matching Algorithms on FPGAs

Panagiotis D. Michailidis and Konstantinos G. Margaritis

Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
{*panosm, kmarg*}*@uom.gr*

Approximate string matching problem is a common and often repeated task in information retrieval and bioinformatics. This paper proposes a generic design of a programmable array processor architecture for a wide variety of approximate string matching algorithms to gain high performance at low cost. Further, we describe the architecture of the array and the architecture of the cell in detail in order to efficiently implement for both the preprocessing and searching phases of most string matching algorithms. Further, the architecture performs approximate string matching for complex patterns that contain don't care, complement and classes symbols. We also implement and evaluate the proposed architecture on a field programmable gate array (FPGA) device using the JHDL tool for synthesis and the Xilinx Foundation tools for mapping, placement, and routing. Finally, our programmable implementation achieves about 9-340 times faster than a desktop computer with a Pentium 4 3.5 GHz for all algorithms when the length of the pattern is 1024.

ReConfigME: A Detailed Implementation of an Operating System for Reconfigurable Computing

Grant Wigley, David Kearney and Mark Jasiunas

Computer and Information Science
University of South Australia
Adelaide, South Australia, Australia
{*Grant.Wigley, David.Kearney, Mark.Jasiunas*}*@unisa.edu.au*

Reconfigurable computing applications have traditionally had the exclusive use of the field programmable gate array, primarily because the logic densities of the available devices have been relatively similar in size compared to the application. But with the modern FPGA expanding beyond 10 million system gates, and through the use of dynamic reconfiguration, it has become feasible for several applications to share a single high density device. However, developing applications that share a device is difficult as the current design flow assumes the exclusive use of the FPGA resources. As a consequence, the designer must ensure that resources have been allocated for all possible combinations of loaded applications at design time. If the sequence of application loading and unloading is not known in advance, all resource allocation cannot be performed at design time because the availability of resources changes dynamically. In this paper we present an implementation of an operating system that has the ability to share its FPGA resources dynamically among multiple executing applications.

An Automated Development Framework for a RISC Processor with Reconfigurable Instruction Set Extensions

Nikolaos Vassiliadis, George Theodoridis and Spiridon Nikolaidis

Section of Electronics and Computers, Department of Physics
Aristotle University of Thessaloniki
Thessaloniki, Greece
{nivas, theodor, snikolaid}@physics.auth.gr

By coupling a reconfigurable hardware to a standard processor, high levels of flexibility and adaptability are achieved. However, this approach requires modifications to the compiler of the processor to take into account reconfigurable aspects. In this paper, a development framework for a RISC processor with reconfigurable instruction set extensions is presented. The framework is fully automated, hiding all reconfigurable related issues from the user and can be used for both program and fine-tune the architecture at design time. We demonstrate the above issues using a set of benchmarks. Experimental results show an x2.9 average speedup in addition to potential energy reduction.

High-Level Synthesis with Reconfigurable Datapath Components

George Economakos

Microprocessors and Digital Systems Laboratory
National Technical University of Athens
Athens, Greece
geconom@microlab.ntua.gr

High-level synthesis is becoming more popular as design densities keep increasing, especially in the ASIC design world. Although FPGA design follows ASIC design methodologies and FPGA densities are increasing too, programmable devices also offer the advantage of partial reconfiguration, which allows an algorithm to be partially mapped into a small and fixed FPGA device that can be reconfigured at run time, as the mapped application changes its requirements. This paper presents a novel resource constrained high-level synthesis scheduling heuristic, which utilizes reconfigurable datapath components. The resulting schedule can be shortened so as the gain in clock cycles can overcome the timing overhead of reconfiguration. The main advantage of the proposed methodology is that through run time reconfiguration, more complicated algorithms can be mapped into smaller devices without speed degradation.

An Optically Differential Reconfigurable Gate Array with a Holographic Memory

Minoru Watanabe, Mototsugu Miyano and Fuminori Kobayashi

Department of Systems Innovation and Informatics
Kyushu Institute of Technology
Fukuoka, Japan
{*watanabe, fkoba*}*@ces.kyutech.ac.jp, none@email*

cally Reconfigurable Gate Arrays (ORGAs) offer the possibility of providing a virtual gate count that is much larger than those of currently available VLSIs by exploiting the large storage capacity of holographic memory. We developed an Optically Differential Reconfigurable Gate Array (ODRGA-VLSI) with no overhead and fast reconfiguration capability. This paper presents the results of development of a perfect optical reconfigurable system with the ODRGA-VLSI chip and holographic memory. Experimental results of the reconfiguration procedure and circuit performance on a gate array are also presented.

A Stochastic Multi-Objective Algorithm for the Design of High Performance Reconfigurable Architectures

Wing On Fung[1] and Tughrul Arslan[1,2]

[1]*School of Electronics and Engineering*
University of Edinburgh
Edinburgh, Scotland, United Kingdom
{*wing.fung, tughrul.arslan*}*@ed.ac.uk*

[2]*Institute for System Level Integration*
The Alba Centre, Alba Campus
Livingston, Scotland, United Kingdom

The increasing demand for FPGAs and reconfigurable hardware targeting high performance low power applications has lead to an increasing requirement for new high performance reconfigurable embedded FPGA cores. This paper presents a multi-objective population based algorithm which given a library of basic blocks and a list of constraints, identifies an optimum reconfigurable embedded reconfigurable core suitable for the target application.

Reconfigurable Communications for Image Processing Applications

André Borin Soares, Luigi Carro and Altamiro Amadeu Susin

Instituto de Informatica
UFRGS
Porto Alegre, RS, Brasil
{*borin, carro, susin*}*@inf.ufrgs.br*

This work tries to reuse programmable communication resources like a Network-on-Chip (NoC) in the acceleration of image applications. We show a mathematical model for the computation and communication pattern of two distributed motion estimation algorithms, Full Search Block Matching Algorithm and Multi-Resolution Block Matching Algorithm. Experimental results show that the use of the Multi-Resolution method reduces not only the computation time but also the traffic of messages on the NoC. This leads to a lower power consumption in the NoC during the processing time of each image. The studied examples show the importance of the link between algorithms and their mapping onto a programmable fabric, not only regarding computation, but facing communication as well.

Design and Analysis of Matching Circuit Architectures for a Closest Match Lookup

Kieran Mclaughlin[1], Friederich Kupzog[2], Holger Blume[2], Sakir Sezer[1], Tobias Noll[2] and John Mccanny[1]

[1] *The Institute of Electronics Communications and Information Technology at QUB*
Belfast, United Kingdom
kieran.mclaughlin@ee.qub.ac.uk

[2] *The Institute of Electrical Engineering and Computer Systems*
RWTH Aachen University
Aachen, Germany

This paper investigates the implementation of a number of circuits used to perform a high speed closest value match lookup. The design is targeted particularly for use in a search trie, as used in various networking lookup applications, but can be applied to many other areas where such a match is required. A range of different designs have been considered and implemented on FPGA. A detailed description of the architectures investigated is followed by an analysis of the synthesis results.

RTOS Extensions for dynamic hardware / software monitoring and configuration management.

Yvan Eustache, Jean-philippe Diguet and Milad El Khodary

LESTER laboratory
UBS / CNRS
Lorient, France
{*yvan.eustache, jean-philippe.diguet, elkhodary*} *@univ-ubs.fr*

We present our solution for a flexible and unified implementation of self-adaptive systems on reconfigurable architectures. This approach is based on a couple of local and global reconfiguration managers. In this paper we describe how the managers are implemented in the context of an usual RTOS and the new services we add for hardware and software monitoring, reconfiguration decision and reconfiguration control which also includes hardware and software interface modeling.

Securing Embedded Programmable Gate Arrays in Secure Circuits

Nicolas Valette[1], Lionel Torres[2], Gilles Sassatelli[2] and Frederic Bancel[1]

[1]*Smartcard Division*
STMicroelectronics
ROUSSET, France
{*nicolas.valette, frederic.bancel*} *@st.com*

[2]*Microelectronics*
LIRMM / UMR 5506
MONTPELLIER, France
{*torres, sassatelli*} *@lirmm.fr*

The purpose of this article is to propose a survey of possible approaches for implementing embedded reconfigurable gate arrays into secure circuits. A standard secure interfacing architecture is proposed and motivations justifying such an approach are discussed. This paper also lists all features offered by FPGA vendors (Field Programmable Gate Array) aiming at securing those circuits according to different concerns. This article emphasizes on configuration memory programming which is probably the weakest point of using programmable devices on a secure context.

Design Space Exploration for Low-Power Reconfigurable Fabrics

Gayatri Mehta, Raymond R. Hoare, Justin Stander and Alex K. Jones

Electrical and Computer Engr.
University of Pittsburgh
Pittsburgh, PA, USA
gmehta@engr.pitt.edu, hoare@ece.pitt.edu, jns36@pitt.edu akjones@ece.pitt.edu

This paper presents a parameterizable, coarse-grained, reconfigurable fabric model that attempts to maintain Field Programmable Gate Array (FPGA)-like programmability and Computer Aided Design (CAD), with Application Specific Integrated Circuit (ASIC)-like power characteristics for Digital Signal Processing (DSP) style applications. Using this model, architectural design space decisions are explored in order to define an energy-efficient fabric. The impact on energy and performance due to the variation of different parameters such as datawidth and interconnection flexibility has been studied. The multiplexer cardinality usage has also been studied by mapping some of the signal and image processing applications onto the fabric. The results point to the use of power optimized 32-bit width computational elements interconnected by low cardinality multiplexers like 4:1 multiplexers.

Exploiting Dynamic Reconfiguration Techniques: The 2D-VLIW Approach

Ricardo Santos[1,2], Rodolfo Azevedo[1] and Guido Araujo[1]

[1] *Institute of Computing*
State University of Campinas
Campinas, SP, Brazil
{ricrs, rodolfo, guido}@ic.unicamp.br

[2] *Department of Computer Engineering*
Dom Bosco Catholic University
Campo Grande, MS, Brazil

Fast reconfiguration is a mandatory feature for reconfigurable computing architectures. Research in this area has been increasingly focusing on new reconfiguration techniques that can sustain the target performance goal. For reconfigurable pipelined architectures, the challenge is to allow the simultaneous execution, at the same stage, of configuration and computation tasks. In this context, this paper presents a new dynamic reconfiguration technique, based on a configuration cache, that tackles this challenge by configuring and executing operations on functional units during the execution stage. This approach is implemented in a pipelined reconfigurable multiple-issue architecture called 2D-VLIW. Our dynamic reconfiguration technique takes advantage of the 2D-VLIW pipelined execution by starting reconfiguration concurrently to activities like reading operand registers and executing operations.

Applying Single Processor Algorithms to Schedule Tasks on Reconfigurable Devices Respecting Reconfiguration Times

Florian Dittmann and Marcelo Götz

Heinz Nixdorf Institute
University Paderborn
Paderborn, Germany
{roichen, goetz}@upb.de

In the single machine environment, several scheduling algorithms exist that allow to quantify schedules with respect to feasibility, optimality, etc. In contrast, reconfigurable devices execute tasks in parallel, which intentionally collides with the single machine principle and seems to require new methods and evaluation strategies for scheduling. However, the reconfiguration phases of adaptable architectures usually take place sequentially. Run-time adaptation is realized using an exclusive port, which again is occupied for some reasonable time during reconfiguration. We have to handle the duration and the sequential exclusiveness of reconfiguration phases. Here, we can find an analogy to the single machine environment, as both scenarios must derive a sequential schedule for an exclusive resource. Thus, we investigate the appliance of single processor scheduling algorithms to task reconfiguration on reconfigurable systems in this paper. We determine necessary adaptations and propose methods to evaluate the scheduling algorithms.

Dynamically Reconfigurable Cache Architecture Using Adaptive Block Allocation Policy

Milene Barbosa Carvalho, Luís Fabrcio Wanderley Góes and Carlos Augusto Paiva Da Silva Martins

Computational and Digital Systems Laboratory (LSDC)
Pontifical Catholic University of Minas Gerais, PUC Minas
Belo Horizonte, Minas Gerais, Brazil
milene@ieee.org, lfwgoes@yahoo.com.br, capsm@pucminas.br

In this paper, we present a dynamically reconfigurable cache architecture using adaptive block allocation policy analyzed by means of simulation. Our main objectives are: to propose a reconfigurable cache architecture and to propose, implement and analyze the performance of an adaptive cache block allocation policy. First, we present a proposal of the reconfigurable cache architecture that can adapt according to the workload. Then we present our adaptive policy and do performance tests comparing our cache architecture with some set associative configurations. In these tests, we use some traces from BYU Trace Distribution Center of SPEC 2000 Benchmark. Finally, we analyze the results based on metrics like cache miss ratio, response time, etc. Our main contributions are: the proposal of a dynamically reconfigurable cache architecture; proposal, development and implementation of an adaptive cache block allocation policy.

Practical Design of a Computation and Energy Efficient Hardware Task Scheduler in Embedded Reconfigurable Computing Systems

Tyrone Tai-on Kwok and Yu-kwong Kwok

Department of Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong
tokwok@eee.hku.hk, ykwok@hku.hk

By utilizing massively parallel circuit design in FPGAs, the overall system efficiency, in terms of computation efficiency and energy efficiency, can be greatly enhanced by offloading some computation-intensive tasks which are originally executed in the instruction set processor to the FPGA fabric. In essence, a hardware task scheduler is needed. However, most of the work in the literature considers scheduling algorithms which are unable or difficult to be implemented using the design flows in current development platform. Moreover, little of the work takes energy consumption into consideration. In this paper, we present the design of a hardware task scheduler which takes energy consumption into consideration, and can be readily implemented using current design flows.

Reconfigurable Context-Free Grammar Based Data Processing Hardware with Error Recovery

James Moscola, Young H. Cho and John W. Lockwood

Department of Computer Science and Engineering
Washington University
St. Louis, MO, USA
{jmm5, young, lockwood}@arl.wustl.edu

This paper presents an architecture for context-free grammar (CFG) based data processing hardware for reconfigurable devices. Our system leverages on CFGs to tokenize and parse data streams into a sequence of words with corresponding semantics. Such a tokenizing and parsing engine is sufficient for processing grammatically correct input data. However, most pattern recognition applications must consider data sets that do not always conform to the predefined grammar. Therefore, we augment our system to detect and recover from grammatical errors while extracting useful information. Unlike the table look up method used in traditional CFG parsers, we map the structure of the grammar rules directly onto the Field Programmable Gate Array (FPGA). Since every part of the grammar is mapped onto independent logic, the resulting design is an efficient parallel data processing engine. To evaluate our design, we implement several XML parsers in an FPGA. Our XML parsers are able to process the full content of the packets up to 3.59 Gbps on Xilinx Virtex 4 devices.

Power Consumption Advantage of a Dynamic Optically Reconfigurable Gate Array

Minoru Watanabe and Fuminori Kobayashi

Department of Systems Innovation and Informatics
Kyushu Institute of Technology
Fukuoka, Japan
{*watanabe, fkoba*}*@ces.kyutech.ac.jp*

Recently, various types of ORGAs have been developed. However, their gate counts were not satisfactory compared with those of FPGAs. Therefore, to improve the gate density of conventional ORGAs, a dynamic ORGA (DORGA) architecture that can remove static memory functions to store a configuration context has been proposed. However, the DORGA architecture offers not only the advantages of a high gate count, but also the advantage of low reconfiguration power consumption.This paper presents measurement results of the optical reconfiguration power consumption of a DORGA-VLSI chip and shows the power consumption advantages of the DORGA architecture through comparison with other ORGAs.

VHDL to FPGA automatic IPCore generation: A case study on Xilinx design flow

Fabrizio Ferrandi[1], Giovanna Ferrara[2], Roberto Palazzo[1], Vincenzo Rana[1] and Marco Domenico Santambrogio[1]

[1]*Dipartimento di Elettronica e Informazione*
Politecnico di Milano
Milano, Italy
ferrandi@elet.polimi.it, {*roberto.palazzo, vincenzo.rana*}*@email.it, marco.santambrogio@polimi.it*

[2]*Siemens S.p.A*
Cinisello Balsamo, Italy
giovanna.ferrara@siemens.com

This paper aims at introducing a methodology that allows an easy implementation of IP-Cores focusing only on their functionalities rather than their interfaces and their integration in a given architecture. The proposed approach implements all the communication infrastructure needed by a component, described in VHDL, to be finally inserted into a real architecture that can be implemented on FPGAs, reducing the time to market of the final implementation of the system. To validate the entire methodology, we have performed a comparison based on the CoreConnect communication infrastructure, between our results with the classical Xilinx design flow using EDK and ISE.

Workshop 4

Workshop on High-Level Parallel Programming Models and Supportive Environments
HIPS 2006

4 HIPS • Workshop on High-Level Parallel Programming Models and Supportive Environments

Workshop Description:

HIPS is a full-day workshop focusing on high-level programming of parallel and grid architectures. Its goal is to bring together researchers working in the areas of applications, computational models, language design, compilers, system architecture, and programming tools to discuss new developments in programming such systems. One of the keys for a (commercial) breakthrough of parallel processing and grid computing are techniques that facilitate efficient usage of such environments. This covers, for example, programming languages, programming tools, system middleware, as well as communication libraries. These techniques have to be integrated to provide a full solution to the problem. This integration is the topic HIPS is devoted too. The technical program will consist of presentations on the following topics:

- Concepts and languages for parallel and Grid programming
  - Concurrent object-oriented programming
  - Hybrid programming, e.g. OpenMP/MPI, components/MPI
  - Extensions to traditional programming models, e.g. MPI and OpenMP
  - Grid workflow programming
  - Application development environment for service-oriented architectures
  - Mobile agents
  - Component-based programming
- Supportive techniques on system level
  - Architectural and communication support
  - Compiler techniques
  - Runtime systems
  - System monitoring
  - Languages and tools for resource management
- Supportive techniques on application level
  - Application monitoring
  - Performance analysis, and optimization
  - Automatic performance analysis support
  - Integrated programming environments

Workshop Chairs:
Michael Gerndt, Technische Universität München, Germany

Steering Committee:
Rudolf Eigenmann, Purdue University, USA
Michael Gerndt, Technische Universität München, Germany
Frank Müller, North Carolina State University, USA
Craig Rasmussen, Los Alamos National Laboratory, USA
Martin Schulz, Cornell University, USA

Program Committee:
Sigfried Benkner, Universität Wien, Austria
Marian Bubak, Stanislaw Staszic University of Mining and Metallurgy in Cracow, Poland
Marios Dikaiakos, University of Cyprus, Cyprus
Rudolf Eigenmann, Purdue University, USA
Thilo Ernst FhG, Germany
Vincent Freeh, North Carolina State University, USA
Edgar Gabriel, HLRS, Germany
Michael Gerndt, Technische Universität München, Germany
Vladimir Getov, University of Westminster, UK
Fabrice Huet, Inria-CNRS-Univ. Nice, France
Peter Kacsuk, SZTAKI, Hungary
Dieter Kranzlmüller, Johannes Kepler University Linz, Austria
Craig Lee, The Aerospace Corporation, USA
Manish Parashar, Rutgers, The State University of New Jersey, USA
Craig Rasmussen, Los Alamos National Laboratory, USA
Martin Schulz, CASC/Lawrence Livermore National Laboratory, USA
Domenico Talia, Universita della Calabria, Italy
Ian Taylor, Cardiff University, UK
Greg Watson, Los Alamos National Laboratory, USA
Roland Wismüller, Universität Siegen, Germany

HIPS Keynote: Towards a Sophisticated Grid Workflow Development and Computing Environment

Thomas Fahringer

Institute for Computer Science
University of Innsbruck
Innsbruck, Austria
Thomas.Fahringer@uibk.ac.at

While Grid infrastructures can provide massive compute and data storage power, it is still an art to effectively harness the power of Grid computing. Current application development for Grid commonly requires the programmer to deal with many low level and complex details such as selecting software components on specific Grid computers, mapping applications onto the Grid, explicitly specify data transfer operations, etc.

In this talk we will present the ASKALON environment whose goal is to create an invisible Grid for both Grid users and application developers. ASKALON is centered around a set of high-level services for transparent and effective Grid access, including a Scheduler for optimized mapping of workflows onto the Grid, an Enactment Engine for reliable application execution, a Resource Manager covering both computers and application components, and a Performance Prediction and Analysis service based on a traing phase, analytical models and dynamic measurements. A sophisticated XML-based programming interface that shields the user form the Grid middleware details, allows the high-level composition of workflow applications. ASKALON is used to develop and port scientific applications as workflows in the Austrian Grid. Experimental results using several real-world scientific applications to demonstrate the effectiveness of ASKALON will be demonstrated.

Tree-based Overlay Networks for Scalable Applications

Dorian C. Arnold, Gary D. Pack and Barton P. Miller

Computer Sciences Department
University of Wisconsin-Madison
Madison, Wisconsin, USA
{*darnold, pack, bart*}*@cs.wisc.edu*

The increasing availability of high-performance computing systems with thousands, tens of thousands, and even hundreds of thousands of computational nodes is driving the demand for programming models and infrastructures that allow effective use of such large-scale environments. Tree-based Overlay Networks (TBŌNs) have proven to provide such a model for distributed tools like performance profilers, parallel debuggers, system monitors and system administration tools.

We demonstrate that the extensibility and flexibility of the TBŌN distributed computing model, along with its performance characteristics, make it surprisingly general, particularly for applications outside the tool domain. We describe many interesting applications and commonly-used algorithms for which TBŌNs are well-suited and provide a new (non-tool) case study, a distributed implementation of the *mean-shift* algorithm commonly used in computer vision to delineate arbitrarily shaped clusters in complex, multi-modal feature spaces.

Towards a Universal Client for Grid Monitoring Systems: Design and Implementation of the Ovid Browser

Marios D. Dikaiakos, Artemakis Artemiou and George Tsouloupas

Computer Science
University of Cyprus
Nicosia, Cyprus
{*mdd, cs01aa2, georget*}*@cs.ucy.ac.cy*

In this paper, we present the design and implementation of Ovid, a browser for Grid-related information. The key goal of Ovid is to support the seamless navigation of users in the Grid information space. Key aspects of Ovid are: (i) A set of navigational primitives, which are designed to cope with problems such as network disorientation and information overloading; (ii) A small set of Ovid views, which present the enduser with high-level, visual abstractions of Grid information; these abstractions correspond to simple models that capture essential aspects of a Grid infrastructure. (iii) Support for embedding and implementing hyperlinks that connect related entities represented within different information views; (iv) A plug-in mechanism, which enables the seamless integration with Ovid of third-party software that retrieves and displays data from various Grid information sources, and (v) a modular software design, which allows the easy integration of different visualization algorithms that support the graphical representation of large amounts of Gridrelated information in the context of Ovids views.

The Monitoring Request Interface (MRI)

Edmond Kereku and Michael Gerndt

Institut fuer Informatik
Technische Universitaet Muenchen
Garching bei Muenchen, Germany
{*kereku, gerndt*}*@in.tum.de*

In this paper we present MRI, a high level interface for selective monitoring of code regions and data structures in single and multiprocessor environments. MRI keeps transparent the available monitoring resources from the performance analysis tools and can electively generate monitoring results as online profile information, or as postmortem traces. MRI is the first step toward a standard monitoring interface which can be used by a broad range of performance analysis tools, from profiler tools, trace producers and visualizers, up to complex automatic performance analyzers. We also present an implementation of MRI for SMPs which transparently use a simulation backend and a PAPI backend to obtain performance data.

Modeling and executing Master-Worker applications

Hinde Lilia Bouziane, Christian Pèrez and Thierry Priol

IRISA/INRIA
Campus de Beaulieu
35042 Rennes cedex, France
{*Hinde.Bouziane, Christian.Perez, Thierry.Priol*}*@irisa.fr*

This paper describes work in progress to extend component models to support Master-Worker applications and to let them to be executed on Grid infrastructures. The proposed approach is generic enough to be applied to existing component models such as the OMG CORBA and the ObjectWeb FRACTAL component models. One objective of our research is to relieve Grid application designers of managing low level programming and implementation aspects. With the proposed approach, a designer has only to cope with the description of an abstract view of the application architecture in which he has to specify what the master and the workers have to do while leaving the system environment to manage the low level aspects such as communication between the master and the workers.

Towards MPI progression layer elimination with TCP and SCTP

Brad Penoff and Alan Wagner

Department of Computer Science
University of British Columbia
Vancouver, BC, Canada
{*penoff, wagner*}*@cs.ubc.ca*

MPI middleware glues together the components necessary for execution. Almost all implementations have a communication component also called a message progression layer that progresses outstanding messages and maintains their state. The goal of this work is to thin or eliminate this communication component by pushing the functionality down onto the standard IP stack in order to take advantage of potential advances in commodity networking. We introduce a TCP-based design that successfully eliminates the communication component. We discuss how this eliminated TCP-based design doesn't scale and show a more scalable design based on the Stream Control Transmission Protocol (SCTP) that has a thinned communication component. We compare the designs showing why SCTP one-to-many sockets in their current form can only thin the communication component. We show what additional features would be required of SCTP to enable a practical design with a fully eliminated communication component.

Babylon v2.0:Middleware for Distributed, Parallel, and Mobile Java Applications

Willem Van Heiningen[1], Tim Brecht[2] and Steve Macdonald[2]

[1]Integrative Biology
Hospital for Sick Children
Toronto, Ontario, Canada
willem@sickkids.ca

[2]David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
{brecht, stevem}@uwaterloo.ca

Babylon v2.0 is a collection of tools and services that provide a 100% Java compatible environment for developing, running and managing parallel, distributed and mobile Java applications. It incorporates features like object migration, asynchronous method invocation and remote class loading while providing an easy-to-use interface. Additionally, Babylon v2.0 enables Java applications to seamlessly create and interact with remote objects while protecting those objects from other applications by implementing access restrictions and separate name spaces.

This paper describes the most important programming features of the Babylon v2.0 system, using a heat diffusion example to show how they are used in practice. The potential cluster computing benefits of the system are demonstrated with experimental results which show that sequential Java applications can achieve significant performance benefits from using Babylon v2.0 to parallelize their work across a cluster of workstations.

Iterators in Chapel

Mackale Joyner[1], Bradford L. Chamberlain[2] and Steven J. Deitz[2]

[1]Dept. of Computer Science
Rice University
Houston, TX, USA
mjoyner@cs.rice.edu

[2]Cray Inc.
Seattle, WA, USA
{bradc, deitz}@cray.com

A long-held tenet of software engineering is that algorithms and data structures should be specified orthogonally in order to minimize the impact that changes to one will have on the other. Unfortunately, this principle is often not well-supported in scientific and parallel codes due to the lack of abstractions for factoring iteration away from computation in traditional scientific languages. The result is a fragile situation in which complex loop nests are used to express parallelism and maximize performance, yet must be maintained individually as the algorithm and data structures evolve. In this paper, we introduce the iterator concept in the Chapel parallel programming language, designed to address this problem and provide a means for factoring iteration away from computation. The paper illustrates iterators using several examples, compares our approach with those taken in other languages, and describes our implementation in the Chapel compiler.

Automatic Code Generation for Distributed Memory Architectures in the Polytope Model

Michael Claßen and Martin Griebl

FMI
University of Passau
Passau, Germany
{*michael.classen, martin.griebl*}*@uni-passau.de*

The polytope model has been used successfully as a tool for program analysis and transformation in the field of automatic loop parallelization. However, for the final step of automatic code generation, the generated code is either only usable on shared memory architectures or severely restricts the parallelization methods that can be applied. In this paper, we present a fully automated method for generating efficient target code, which is executable on clusters that are based on a distributed memory architecture. We also provide speedup results of experiments on a local cluster.

Techniques Supporting threadprivate in OpenMP

Xavier Martorell, Marc Gonzalez, Alejandro Duran, Jairo Balart, Roger Ferrer, Eduard Ayguade and Jesus Labarta

Barcelona Supercomputing Center
Technical University of Catalunya
Barcelona, Barcelona, Spain
{*xavim, marc, aduran, jbalart, rferrer, eduard, jesus*}*@ac.upc.edu*

This paper presents the alternatives available to support threadprivate data in OpenMP and evaluates them. We show how current compilation systems rely on custom techniques for implementing thread-local data. But in fact the ELF binary specification currently supports data sections that become threadprivate by default. ELF naming for such areas is Thread-Local Storage (TLS). Our experiments demonstrate that implementing threadprivate based on the TLS support is very easy, and more efficient. This proposal goes in the same line as the future implementation of OpenMP on the GNU compiler collection. In addition, our experience with the use of threadprivate in OpenMP applications shows that usually it is better to avoid it. This is because threadprivate variables reside in common blocks and they impede the compiler to fully optimize the code. So it is better to keep threadprivate as a temporary technique only to ease porting MPI codes to OpenMP.

A Configurable Framework for Stream Programming Exploration in Baseband Applications

Jerker Bengtsson and Bertil Svensson

Centre for Research on Embedded Systems
Halmstad University
Halmstad, Sweden
{*Jerker.Bengtsson, Bertil.Svensson*}*@ide.hh.se*

This paper presents a configurable framework to be used for rapid prototyping of stream based languages. The framework is based on a set of design patterns defining the elementary structure of a domain specific language for high-performance signal processing. A stream language prototype for baseband processing has been implemented using the framework. We introduce language constructs to efficiently handle dynamic reconfiguration of distributed processing parameters. It is also demonstrated how new language specific primitive data types and operators can be used to efficiently and machine independently express computations on bit-fields and data-parallel vectors. These types and operators yield code that is readable, compact and amenable to a stricter type checking than is common practice. They make it possible for a programmer to explicitly express parallelism to be exploited by a compiler. In short, they provide a programming style that is less error prone and has the potential to lead to more efficient implementations.

Workshop 5

Java for Parallel and Distributed Computing Workshop
JAVAPDC 2006

5 JAVAPDC • Workshop on Java for Parallel and Distributed Computing

Workshop Description:

This workshop focuses on Java for parallel and distributed computing and supportive environments. One of its aims is to bring together the IPDPS community around Java and Java based technologies, and to provide an opportunity to share experience and views of current trends and activity in the domain.

Topics of interest include but are not limited to:

- Java for parallel and distributed computing;
- Internet for parallel and distributed computing;
- Programming/communication/distribution libraries;
- Software tools and environments;
- Code transformations, compilers, optimizations, etc.;
- Real world distributed and parallel applications based on Java;
- Reflection;
- Meta-computing;
- Theoretical foundations and formal methods;
- Compiler technology and performance issues;
- Real-time applications;
- Multi-agent systems;
- Data mining and financial applications;
- Software portability, components, and reuse;
- Standards for object interoperability;
- Embedded Java and wireless devices, seamless distributed computing environment;
- Java for global computing, the Web and the Grid;
- Java extensions for distributed computing.

Program Co-chairs:
Denis Caromel, Université de Nice Sophia Antipolis, France

Serge Chaumette, Université Bordeaux I, France

Geoffrey Fox, Community Grids Laboratory, USA

Peter Graham, University of Manitoba, Canada

Program Committee:

Denis Caromel, Université de Nice Sophia Antipolis, France

Serge Chaumette, Université Bordeaux I, France

Geoffrey Fox, Community Grids Laboratory, USA

Peter Graham, University of Manitoba, Canada

Jack Dongarra, University of Tennessee

Doug Lea, State University of New York at Oswego

Vladimir Getov, University of Westminster, London, U.K.

George K. Thiruvathukal, Loyola University and Northwestern University

David Walker, Cardiff University, UK

More on JACE: New Functionalities, New Experiments

Jacques Mohcine Bahi, Stephane Domas and Kamel Mazouzi

Laboratoire d'Informatique de Franche-Comté
Université de Franche-Comte
Belfort, France
{*bahi, sdomas, mazouzi*}*@iut-bm.univ-fcomte.fr*

Java is often criticized for its poor performances compared to native codes. Nevertheless, this language provides lots of interesting functionalities to easily implement scientific applications on a widely distributed architecture (i.e. grid). The context of this paper is that of iterative algorithms. In order to increase the efficiency of the code, we suggest to use a special class of algorithms called AIACs (Asynchronous Iterations, Asynchronous Computations). This paper presents new results on our works to combine Java and asynchronism within a programming/execution environment called JACE. New functionalities have been added and interesting comparisons with C/MPI and on the impact of overlap techniques are given.

Exploiting Dynamic Proxies in Middleware for Distributed, Parallel, and Mobile Java Applications

Willem Van Heiningen[1], Tim Brecht[2] and Steve Macdonald[2]

[1]*Integrative Biology*
Hospital for Sick Children
Toronto, Ontario, Canada
willem@sickkids.ca

[2]*David R. Cheriton School of Computer Science*
University of Waterloo
Waterloo, Ontario, Canada
{*brecht, stevem*}*@uwaterloo.ca*

Babylon v2.0 is a collection of tools and services that provide a 100% Java compatible environment for developing, running and managing parallel, distributed and mobile Java applications. It incorporates features like object migration, asynchronous method invocation and remote class loading while providing an easy-to-use interface. The implementation of Babylon v2.0 exploits *dynamic proxies*, a feature added to Java 1.3 that allows runtime creation of proxy objects. This paper shows how Babylon v2.0 exploits dynamic proxies to implement several key features without the need for special language or virtual machine extensions, preprocessors, or compilers. The resulting Babylon programs are portable across all Java virtual machines, and the development process is simplified by removing the extra steps needed to invoke external stub compilers and incorporate the generated code into an application. This simplification also allows remote objects to be created for any class that supports an interface to its methods, even if source code is not available.

Performance Analysis of Java Concurrent Programming: A Case Study of Video Mining System

Wenlong Li, Eric Li, Ran Meng, Tao Wang and Carole Dulong

Intel China Research Center
Intel Corporation
Beijing, China
{*wenlong.li, eric.q.li, ran.meng, tao.wang, carole.dulong*}*@intel.com*

As multi/many core processors become prevalent, programming language is important in constructing efficient parallel applications. In this work, we build a multithreaded video mining application with Java, examine the thread profiling information and micro-architecture metrics to identify the factors limiting the scalability, and employ a number of ways to improve performance. Besides, we conduct some thread scheduling experiments. According to the experiments and detailed analysis, we conclude that for this video mining application: (1) Java is a good parallel language candidate for many core processors in terms of performance, scalability, and ease of programming; (2) Thread affinity mechanism is effective in improving data locality, but brings little benefit to multithreaded Java application due to its conservative memory model in JVM.

High-Level Execution and Communication Support for Parallel Grid Applications in JGrid

Szabolcs Pota and Zoltan Juhasz

Dept. of Information Systems
University of Veszprem
Veszprem, Hungary
{*pota, juhasz*}*@irt.vein.hu*

This paper describes the high-level execution and communication support provided in JGrid, a service-oriented dynamic grid framework. One of its core services, the Compute Service, is the key component in creating dynamic computational grid systems that enable the execution of sequential and parallel interactive grid applications. A fundamental set of program execution modes supported by the service is described, then a programming model and its corresponding application programming interface is presented. The execution support of the service architecture is described in detail illustrating how remote evaluation and run-time task spawning are provided. The paper also shows in detail how task spawning and dynamic proxies can be used for a service-oriented communication mechanism for coarse-grain parallel grid applications.

Fault Injection in Distributed Java Applications

William Hoarau, Sébastien Tixeuil and Fabien Vauchelles

LRI - CNRS 8623 & INRIA Grand Large
Paris Sud-XI
Orsay, France
{*hoarau, tixeuil, vauchel*}*@lri.fr*

In a network consisting of several thousands computers, the occurrence of faults is unavoidable. Being able to test the behaviour of a distributed program in an environment where we can control the faults (such as the crash of a process) is an important feature that matters in the deployment of reliable programs.

In this paper, we investigate the possibility of injecting software faults in distributed java applications. Our scheme is by extending the FAIL-FCI software. It does not require any modification of the source code of the application under test, while retaining the possibility to write high level fault scenarios. As a proof of concept, we use our tool to test FreePastry, an existing java implementation of a Distributed Hash Table (DHT), against node failures.

Saburo, a tool for I/O and concurrency management in servers

Gautier Loyauté[1], Rémi Forax[2] and Gilles Roussel[3]

[1]*Institut Gaspard-Monge*
Université de Marne-la-Vallée
Champs-sur-Marne, Marne-la-Vallée, France
loyaute@univ-mlv.fr

[2]*Institut Gaspard-Monge*
Université de Marne-la-Vallée
Champs-sur-Marne, Marne-la-Vallée, France
forax@univ-mlv.fr

[3]*Institut Gaspard-Monge*
Université de Marne-la-Vallée
Champs-sur-Marne, Marne-la-Vallée, France
roussel@univ-mlv.fr

This paper presents a Java framework based on **separation of concerns** and **code generation** concepts that facilitates development of concurrency and I/O in servers. In this approach, the application is modeled by a graph whose vertices correspond to units of treatment connected by channels. It allows to build all kind of servers: multi-threaded, **S**ingle-**P**rocess **E**vent-**D**riven, **S**taged **E**vent **D**riven **A**rchitecture, etc. without modification of the functional part. This architecture also permits to extend very easily an application, adding vertices and edges to the graph. The aim of our development tool is to improve programmer productivity and portability, decreasing development time, and reducing bugs or deadlock problems.

Chedar: Peer-to-Peer Middleware

Annemari Auvinen, Mikko Vapa, Matthieu Weber, Niko Kotilainen and Jarkko Vuori

Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland
{*annauvi, mikvapa, mweber, npkotila, jarkko.vuori*}*@jyu.fi*

In this paper we present a new peer-to-peer (P2P) middleware called CHEap Distributed ARchitecture (Chedar). Chedar is totally decentralized and can be used as a basis for P2P applications. Chedar tries to continuously optimize its overlay network topology for maximum performance. Currently Chedar combines four different topology management algorithms and provides functionality to monitor how the peer-to-peer network is self-organizing. It also contains basic search algorithms for P2P resource discovery. Chedar has been used for building a data fusion prototype and a P2PDisCo distributed computing application, which provides an interface for distributing the computation of Java applications. To allow Chedar to be used in mobile devices, the Mobile Chedar middleware has also been developed.

Workflow Fine-grained Concurrency with Automatic Continuation

Giancarlo Tretola[1] and Eugenio Zimeo[2]

[1]*Department of Engineering*
University of Sannio
Benevento, ITALY
tretola@unisannio.it

[2]*Research Centre on Software Technology*
University of Sannio
Benevento, ITALY
zimeo@unisannio.it

Workflow enactment systems are becoming an effective solution to ease programming, deployment and execution of distributed applications in several domains such as telecommunication, manufacturing, e-business, e-government and grid computing. In some of these fields, efficiency and traffic optimization represent key aspects for a wide diffusion of workflow engines and modeling tools. This paper focuses on a technique that enables fine-grained concurrency in compute and data-intensive workflows and reduces the traffic on the network by limiting the number of interactions to the ones strictly needed to bring the data where they are really necessary for continuing the flow of computations. We implemented this technique by using the concepts of wait by necessity and automatic continuation and we integrated it in a flexible, Java workflow engine that through the new mechanisms is able to navigate a workflow anticipating the enactment of sequential activities.

Distributed Monte Carlo Simulation of Light Transportation in Tissue

Andrew J. Page[1], Shirley Coyle[2], Thomas M. Keane[1], Thomas J. Naughton[1], Charles Markham[1] and Tomas Ward[2]

[1]*Dept. of Computer Science*
National University of Ireland Maynooth
Maynooth, Co. Kildare, Ireland
{*andrew.j.page, thomas.m.keane, tom.naughton*}*@nuim.ie, cmarkham@cs.nuim.ie*

[2]*Dept. of Electronic Engineering*
National University of Ireland Maynooth
Maynooth, Co. Kildare, Ireland
shirley.coyle@eeng.may.ie, tomas.ward@eeng.nuim.ie

A distributed Monte Carlo simulation which models the propagation of light through tissue has been developed. It will allow for improved calibration of medical imaging devices for investigating tissue oxygenation in the white matter of the cerebral cortex. The application can distribute the simulation over an unbounded number of processors in parallel. We have found that this application is highly parallelisable resulting in up to 97We found that the source illumination footprint has an effect on the distribution of photons in the head and that lasers do produce a small beam in a highly scattering medium. This application will help researchers to improve the accuracy of their experiments.

The Benefits of Java and Jini in the JGrid System

Szabolcs Pota and Zoltan Juhasz

Dept. of Information Systems
University of Veszprem
Veszprem, Hungary
{*pota, juhasz*}*@irt.vein.hu*

The Java language and platform have been considered by many as natural candidate for creating grid systems. The platform-independent runtime environment, safe and high-level language and its built-in support for networking and security are very valuable features. Despite its potential and the many proof-of-concept systems developed, the grid community is turning to web services technology as its implementation base. In this paper, we show that Java, by joining forces with Jini Technology can provide a very appealing technology base for highly dynamic grid systems. The key properties of Java and Jini technology are examined with reference to their role in grids. Then, the JGrid Jini-based service-oriented grid system is overviewed describing its key concepts, services and how it extends Jini to address some of the unique requirements of grid systems.

Workshop 6

Workshop on Nature Inspired Distributed Computing
NIDISC 2006

6 NIDISC • Workshop on Nature Inspired Distributed Computing

Workshop Description:

Techniques based on metaheuristics and nature-inspired paradigms can provide efficient solutions to a wide variety of problems. Moreover, parallel and distributed metaheuristics can be used to provide more powerful problem solving environments in a variety of fields, ranging, for example, from finance to bio- and health-informatics. This workshop seeks to provide an opportunity for researchers to explore the connection between metaheuristics and the development of solutions to problems that arise in operations research, parallel computing, telecommunications, and many others.

Topics of interest include but are not limited to:

- Nature-inspired methods (e.g. ant colonies, GAs, cellular automata, DNA and molecular computing, local search, etc) for problem solving environments.
- Parallel and distributed metaheuristics techniques (algorithms, technologies and tools).
- Applications combining traditional parallel and distributed computing and optimization techniques as well as theoretical issues (convergence, complexity, etc).
- Other algorithms and applications relating the above mentioned research areas.

General Chairs:
Albert Y. Zomaya, The University of Sydney, Australia
Fikret Ercal, University of Missouri, Rolla, USA

Program Co-chairs:
El-ghazali Talbi, Lab d'Informatique Fondam. de Lille, France
Enrique Alba, University of Málaga, Spain

Program Committee:
Azzedine Boukerche, University of Ottawa, Canada
Martin Middendorf, University of Leipzig, Germany
Pascal Bouvry, University of Luxembourg, Luxembourg
Michelle D. Moore, Texas A & M - Corpus Christi, USA
Juergen Branke, University of Karlsruhe, Germany
G. Spezzano, University of Calabria, Italy
Erick Cantú-Paz, Lawrence Livermore National Laboratory, USA
Franciszek Seredynski, Polish Academy of Sciences, Poland
Tarek El-Ghazawi, George Washington University, USA
Marco Tomassini, University of Lausanne, Switzerland
Nordine Melab, University of Lille, France

A nature-inspired algorithm for the disjoint paths problem

Maria J. Blesa and Christian Blum

ALBCOM research group, Dept. de Llenguatges i Sistemes Informatics
Universitat Politecnica de Catalunya
Barcelona, Spain
{*mjblesa, cblum*}*@lsi.upc.edu*

One of the basic operations in communication networks consists in establishing routes for *connection requests* between physically separated network nodes. In many situations, either due to technical constraints or to quality-of-service and survivability requirements, it is required that no two routes interfere with each other. These requirements apply in particular to routing and admission control in large-scale, high-speed and optical networks. The same requirements also arise in a multitude of other applications such as real-time communications, VLSI design, scheduling, bin packing, and load balancing. This problem can be modeled as a combinatorial optimization problem as follows. Given a graph G representing a network topology, and a collection $T = \{(s_1, t_1) \dots (s_k, t_k)\}$ of pairs of vertices in G representing connection request, the maximum *edge-disjoint paths problem* is an NP-hard problem that consists in determining the maximum number of pairs in T that can be routed in G by mutually edge-disjoint $s_i - t_i$ paths.

We propose an *ant colony optimization* (ACO) algorithm to solve this problem. ACO algorithms are approximate algorithms that are inspired by the foraging behavior of real ants. The decentralized nature of these algorithms makes them suitable for the application to problems arising in large-scale environments. First, we propose a basic version of our algorithm in order to outline its main features. In a subsequent step we propose several extensions of the basic algorithm and we conduct an extensive parameter tuning in order to show the usefulness of those extensions. In comparison to a multi-start greedy approach, our algorithm generates in general solutions of higher quality in a shorter amount of time. In particular the run-time behaviour of our algorithm is one of its important advantages.

A Parallel Memetic Algorithm Applied to the Total Tardiness Machine Scheduling Problem

Vinícius Jacques Garcia[1], Paulo Morelato França[1], Alexandre De Sousa Mendes[2] and Pablo Moscato[2]

[1]*Faculdade de Engenharia Elétrica e de Computação*
State University of Campinas
Campinas, São Paulo, Brazil
{*jacques, franca*}*@densis.fee.unicamp.br*

[2]*School of Electrical Engineering and Computer Science*
The University of Newcastle
Newcastle, New South Wales, Australia
{*mendes, moscato*}*@cs.newcastle.edu.au*

This work proposes a parallel memetic algorithm applied to the total tardiness single machine scheduling problem. Classical models of parallel evolutionary algorithms and the general structure of memetic algorithms are discussed. The classical model of global parallel genetic algorithm was used to model the global parallel memetic analogue where the parallelization is only applied to the individual optimization phase of the algorithm. Computational tests show the efficiency of the parallel approach when compared to the sequential version. A set of eight instances, with sizes ranging from 56 up to 323 jobs and with known optimal solutions, is used for the comparisons.

Sharing resources with artificial ants

Christophe Guéret, Nicolas Monmarché and Mohamed Slimane

Laboratoire d'informatique
Université François Rabelais de Tours
Tours, France
{*christophe.gueret, nicolas.monmarche, mohamed.slimane*}*@univ-tours.fr*

As networks are growing up , more and more information becomes available every day. Despite the presence of software enabling communications and content sharing, they are not always shared among people inside networks. We present here an architecture aimed at helping people to share informations and find collaborators inside an organization. It is part of our PIAF framework, an intelligent agent system we use to develop recommender and personalization software. The main contribution of this paper is the introduction of principles of stigmergy and artificial ants to model data flows in a social network.

Ant Stigmergy on the Grid: Optimizing the Cooling Process in Continuous Steel Casting

Peter Korosec[1], Jurij Silc[1], Bogdan Filipic[2] and Erkki Laitinen[3]

[1]*Computer Systems Department*
Jozef Stefan Institute
Ljubljana, Slovenia
{*peter.korosec, jurij.silc*}*@ijs.si*

[2]*Department of Intelligent Systems*
Jozef Stefan Institute
Ljubljana, Slovenia
bogdan.filipic@ijs.si

[3]*Department of Mathematical Sciences*
University of Oulu
Oulu, Finland
erkki.laitinen@oulu.fi

Most of the world steel production is nowadays based on continuous casting. This is a complex metallurgical process in which liquid steel is cooled and shaped into semi-manufactures. To achieve proper quality of cast steel, it is essential to control the metal flow and heat transfer during the casting process. They depend on numerous parameters, such as the casting temperature, casting speed and coolant flows. The paper presents a new distributed metaheuristic algorithm in an optimal control problem related to the cooling process in the continuous casting of steel. The optimization task is to tune 18 coolant flows in the caster secondary cooling system to achieve the target surface temperatures along the slab. Sequential search algorithms are proved inefficient for this problem because they take too much time to compute an appropriate solution. For this reason a new distributed search algorithm based on stigmergy perceived in ant colony was developed. The algorithm was run on the Grid that allows us to solve this optimization problem in much shorter time. As a matter of fact, the computation time can be decreased from half a day to a few hours without any decrease in the solution quality.

Distributed Workflow Coordination: Molecules and Reactions

Zsolt Nemeth[1], Christian Perez[2] and Thierry Priol[2]

[1]*MTA SZTAKI*
Budapest, Hungary
zsnemeth@sztaki.hu

[2]*IRISA*
Rennes, France
{*christian.perez, thierry.priol*}*@irisa.fr*

Workflow execution on large-scale heterogeneous distributed computing systems, such as Grids, requires a complex coordination. Activities of complex workflow patterns must be matched with entities of the computing system that possesses highly dynamic properties. We pinpoint the key concept of such workflow coordination as actions according to actual and local conditions – analogously to chemical reactions. Modeling workflow enactment as molecules and reactions, formalized in the nature inspired γ-calculus, yielded an autonomously evolving, distributed, decentralized coordination model that can adapt to a dynamically changing environment.

A metaheurisitc based on fusion and fission for partitioning problems

Charles-edmond Bichot

Laboratoire d'Optimisation Globale (LOG)
ENAC — DSNA/DTI
Toulouse, France
bichot@recherche.enac.fr

Metaheuristics are very useful methods because they can find (approximate) solutions of a great variety of problems. One of them, which interests us, is graph partitioning. We present a new metaheuristic based on nuclear fusion and fission of atoms. This metaheuristic, called fusion fission, is compared to other classical algorithms. First, we present spectral and multilevel algorithms which are used to solve partitioning problems. Secondly, we present two metaheuristics applied to partitioning problems : simulated annealing and ant colony algorithms. We will show that fusion fission gives good results, compared to the other algorithms. We demonstrate on a problem of Air Traffic Control that metaheuristics methods can give better results than specific methods.

A Nonself Space Approach to Network Anomaly Detection

Marek Ostaszewski[1], Franciszek Seredynski[1,2,3] and Pascal Bouvry[4]

[1]Institute of Computer Science
University of Podlasie
Siedlce, Poland
marekostaszewski@o2.pl, sered@ipipan.waw.pl

[2]Department of Computer Networks
Polish-Japanese Institute of Information Technology
Warsaw, Poland

[3]Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland

[4]Faculty of Sciences, Technology and Communication
Luxembourg University
Luxembourg, Luxembourg
pascal.bouvry@uni.lu

The paper presents an approach for the anomaly detection problem based on principles of immune systems. Flexibility and efficiency of the anomaly detection system are achieved by building a model of network behavior based on self-nonself space paradigm. Covering both self and nonself spaces by hyperrectangular structures is proposed. Structures corresponding to self-space are built using a training set from this space. Hyperrectangular detectors covering nonself space are created using niching genetic algorithm. Coevolutionary algorithm is proposed to enhance this process. Results of conducted experiments show a high quality of intrusion detection which outperforms the quality of recently proposed approach based on hypersphere representation of self-space.

Parallel Implementation of Evolutionary Strategies on Heterogeneous Clusters with Load Balancing

Juan Francisco Garamendi and Jose Luis Bosque

Escuela Superior de Ciencias Experimentales y Tecnologia
Universidad Rey Juan Carlos
Madrid, Spain
jf.garamendi@alumnos.urjc.es, joseluis.bosque@urjc.es

This paper presents a load balancing algorithm for a parallel implementation of an evolutionary strategy on heterogeneous clusters. Evolutionary strategies can efficiency solve a diverse set of optimization problems. Due to cluster heterogeneity and in order to improve the speedup of the parallel implementation a load balancing algorithm has been implemented. This load balancing algorithm takes into account cluster heterogeneity and it is based on an optimal intial distribution. This initial distribution is determined based on the cluster nodes' computational powers, that are dinamically measured in each slave node by an ad hoc load-bechmark. The implementation presents very satisfactory parallelization results, both in performance and scalability and Super-linear speedup is reached for several tests configurations. Experimental results show excellent perfomence, increasing the improvements with the load balancing algorithm.

Placement and Routing of Boolean Functions in constrained FPGAs using a Distributed Genetic Algorithm and Local Search.

Manuel Rubio Del Solar[1], Juan Manuel Sánchez Pérez[2], Juan Antonio Gómez Pulido[3] and Miguel Ángel Vega Rodríguez[4]

[1]*Dep. de Informática*
Universidad de Extremadura
Cáceres, Cáceres, Spain
mrubio@unex.es

[2]*Dep. de Informática*
Universidad de Extremadura
Cáceres, Cáceres, Spain
sanperez@unex.es

[3]*Dep. de Informática*
Universidad de Extremadura
Cáceres, Cáceres, Spain
jangomez@unex.es

[4]*Dep. de Informática*
Universidad de Extremadura
Cáceres, Cáceres, Spain
mavega@unex.es

In this work we present a system for implementing the placement and routing stages in the FPGA cycle of design, into the physical design stage. We start with the ISCAS benchmarks, on EDIF format, of Boolean functions to be implemented. They are processed by a parser in order to obtain an internal representation which is able to be processed by a Genetic Algorithm (GA) tool. This tool develops the Placement and Routing tasks, considering possible restricted area into the FPGA. In order to help to the GA to make the Routing stage we have added a local search procedure. That local search gets a path between two points without considering neither their placement nor the restricted areas among them. The GA is fully customizable, featuring the ability to work with one or several islands. The experiments have verified that using distributing execution improves the costs and speeds up the convergence towards better results in smaller slots of time.

Evaluating Parallel Simulated Evolution Strategies for VLSI Cell Placement

Sadiq M. Sait, Mustafa Imran Ali and Ali Mustafa Zaidi

Computer Engineering Department
King Fahd University of Petroleum and Minerals
Dhahran-31261, Saudi Arabia
{sadiq, mustafa, alizaidi}@ccse.kfupm.edu.sa

Simulated Evolution (SimE) is an evolutionary metaheuristic that has produced results comparable to well established stochastic heuristics such as SA, TS and GA, with shorter runtimes. However, for problems with a very large set of elements to optimize, such as in VLSI placement and routing, runtimes can still be very large and parallelization is an attractive option. Compared to other metaheuristics, parallelization of SimE has not been extensively explored. This paper presents a comprehensive set of parallelization approaches for SimE when applied to multiobjective VLSI cell placement problem. Each of these approaches are evaluated with respect to SimE characteristics and the constraints imposed by the problem instance. Conclusions drawn can be extended to parallelization of other SimE based optimization problems.

A Proposal of Metaheuristics Based in the Cooperation between Operators in Combinatorial Optimization Problems

Alejandro Sancho-royo, David Pelta and José L. Verdegay

Depto de Ciencias de la Computación e IA
Universidad de Granada
Granada, Spain
alejandrosanchoroyo@gmail.com, {dpelta, verdegay}@decsai.ugr.es

In the context of optimization problems, metaheuristics are tools that stand out by its excellent results and generality. A lot of metaheuristics are formed by a population of agents that operates in a search space. A frame of metaheuristics inspired in cooperation between unrelated individuals is proposed and three different methods of cooperation are suggested. The implementation of the cooperation between agents is made using Soft Computing techniques. A fuzzy rules system has been designed concretely to perform the cooperation. Details about the implementation of three methods of cooperation and the computation of the fuzzy rules are offered for the models considered. A framework of experimentation over the combinations of methods and models is proposed.

Advances in Applying Genetic Programming to Machine Learning, Focussing on Classification Problems

Stephan Winkler[1,2], Michael Affenzeller[1] and Stefan Wagner[1]

[1] *Department for Software Engineering*
Upper Austrian University for Applied Sciences
Hagenberg, 4232, Austria
{stephan, michael, stefan}@heuristiclab.com

[2] *Institute for Design and Control of Mechatronical Systems*
Johannes Kepler University
Linz, 4040, Austria

A Genetic Programming based approach for solving classification problems is presented in this paper. Classification is understood as the act of placing an object into a set of categories, based on the object's properties; classification algorithms are designed to learn a function which maps a vector of object features into one of several classes. This is done by analyzing a set of input-output examples ("training samples") of the function. Here we present a method based on the theory of Genetic Algorithms and Genetic Programming that interprets classification problems as optimization problems: Each presented instance of the classification problem is interpreted as an instance of an optimization problem, and a solution is found by a heuristic optimization algorithm. The major new aspects presented in this paper are suitable genetic operators for this problem class (mainly the creation of new hypotheses by merging already existing ones and their detailed evaluation) we have designed and implemented. The experimental part of the paper documents the results produced using new hybrid variants of genetic algorithms as well as investigated parameter settings.

A Parallel Exact Hybrid Approach for Solving Multi-Objective Problems on the Computational Grid

Mohand Mezmaz, Nouredine Melab and El-ghazali Talbi

Laboratoire d'Informatique Fondamentale de Lille
University of Lille1
Villeneuve d'Ascq, France
{*mezmaz, melab, talbi*}*@lifl.fr*

This paper presents a parallel hybrid exact multi-objective approach which combines two metaheuristics - a genetic algorithm (GA) and a memetic algorithm (MA), with an exact method - a branch and bound (B&B) algorithm. Such approach profits from both the exploration power of the GA, the intensification capability of the MA and the ability of the B&B to provide optimal solutions with proof of optimality. To fully exploit the resources of a computational grid, the hybrid method is parallelized according to three well-known parallel models - the island model for the GA, the multi-start model for the MA and the parallel tree exploration model for the B&B. The obtained method has been experimented and validated on a bi-objective flow-shop scheduling problem. The approach allowed to solve exactly for the first time an instance of the problem - 50 jobs on 5 machines. More than 400 processors belonging to 4 administrative domains have contributed to the resolution process during more than 6 days.

A Combined Genetic-Neural Algorithm for Mobility Management

Javid Taheri[1] and Albert Y. Zomaya[2]

[1]*School of Information Technologies*
The University of Sydney
Sydney, NSW, Australia
JavidT@it.usyd.edu.au

[2]*School of Information Technologies*
The University of Sydney
Sydney, NSW, Australia
Zomaya@it.usyd.edu.au

This work presents a new approach to solve the location management problem by using the location areas approach. A combination of a genetic algorithm and the Hopfield neural network is used to find the optimal configuration of location areas in a mobile network. Toward this end, the location areas configuration of the network is modeled so that the general condition of all the chromosomes of each population improves rapidly by the help of a Hopfield neural network. The Hopfield neural network is incorporated into the genetic algorithm optimization process, to expedite its convergence, since the generic genetic algorithm is not fast enough. Simulation results are very promising and they lead to network configurations that are unexpected but very efficient.

Workforce Planning with Parallel Algorithms

Enrique Alba, Gabriel Luque and Francisco Luna

Department of Languages and Computational Sciences
University of Malaga
Malaga, Spain
{*eat, gabriel, flv*}*@lcc.uma.es*

Workforce planning is an important activity that enables organizations to determine the workforce needed for continued success. A workforce planning problem is a very complex task that requires modern techniques to be solved adequately. In this work, we describe the development of two parallel metaheuristic methods, a parallel genetic algorithm and a parallel scatter search, which can find high-quality solutions to 20 different problem instances. Our experiments show that parallel versions do not only allow to reduce the execution time but they also improve the solution quality.

Self-Organized Task Allocation for Computing Systems with Reconfigurable Components

Daniel Merkle, Martin Middendorf and Alexander Scheidler

Department of Computer Science
University of Leipzig
Leipzig, Germany
{*merkle, middendorf, scheidler*}*@informatik.uni-leipzig.de*

A self-organized allocation scheme for service tasks in computing systems is proposed in this paper. Usually components of a computing system need some service from time to time in order perform their work efficiently. In adaptive computing systems the components and the necessary tasks adapt to the needs of users or the environment. Since in such cases the type of service tasks will often change it is attractive to use reconfigurable hardware to perform the service tasks. The studied system consists of normal worker components and helper components which have reconfigurable hardware and can perform different service tasks. The speed with which a service tasks is executed by a helper depends on its actual configuration. Different strategies for the helpers to decide about service task acceptance and reconfiguration are proposed. These strategies are inspired by stimulus-threshold models that are used to explain task allocation in social insects.

A Multiple Task Allocation Framework for Biological Sequence Comparison in a Grid Environment

Azzedine Boukerche[1], Marcelo S. Sousa[2] and Alba C. M. A. De Melo[2]

[1] *SITE*
University of Ottawa
Ottawa, Canada
boukerch@site.uottawa.ca

[2] *Department of Computer Science*
University of Brasilia
Brazil
albamm@cic.unb.br

The evolution of DNA sequencing techniques generated huge sequence repositories and hence the need for efficient algorithms to compare them. The increase search speed, heuristic algorithms like BLAST were developed and are widely used. In order to further reduce BLAST execution time, this paper evaluates an adaptive task allocation framework to perform BLAST searches in a grid environment against segmented genetic databases segments. Our results present very good speedups and also show that no single task allocation strategy is able to achieve the lowest execution times for all scenarios. Also, our results show that the proposed adaptive strategy was able to deal with the heterogeneous and non-dedicated nature of a grid.

A Physical Particle and Plane Framework for Load Balancing in Multiprocessors

Navid Imani[1] and Hamid Sarbazi Azad[1,2]

[1] *School of Computer Science*
IPM
Tehran, Iran
navid_imani@softhome.net

[2] *Department of Computer Engineering*
Sharif University of Tech.
Tehran, Iran

Different models for load balancing have been proposed before, each of which has its own features and advantages when considered for a specific scenario. Yet, nearly all of the existing techniques have assumed an oversimplified model of the system which is often not the case of the real world. In this paper, a new gradient based algorithm for dynamic load balancing on multiprocessors is proposed. This algorithm is an analogy of a classical physical model of a Particle & Plane system which operates based on the classic laws of physics dictated by the nature.

Workshop 7

Workshop on High Performance Computational Biology HiCOMB 2006

7 HiCOMB • Workshop on High Performance Computational Biology

Workshop Description:

Computational Biology is fast emerging as an important discipline for academic research and industrial application. The large size of biological data sets, inherent complexity of biological problems and the ability to deal with error-prone data all result in large run-time and memory requirements. The goal of this workshop is to provide a forum for discussion of latest research in developing high-performance computing solutions to problems arising from molecular biology. The workshop is especially interested in parallel algorithms, memory-efficient algorithms, large scale data mining techniques, and design of high-performance software.

Topics of interest include but are not limited to:

- Bioinformatic databases
- Computational genomics
- Computational proteomics
- DNA assembly, clustering, and mapping
- Gene expression and microarrays
- Gene identification and annotation
- Parallel algorithms for biological analysis
- Parallel architectures for biological applications
- Molecular evolution
- Molecular sequence analysis
- Phylogeny reconstruction algorithms
- Protein structure prediction and modeling
- String data structures and algorithms

Workshop Co-Chairs:
David A. Bader, Georgia Institute of Technology, USA

Lawrence H. Baker, Iowa State University, USA

Program Chair:

Chau-Wen Tseng, University of Maryland, USA

Program Committee:

Mike Cummings, University of Maryland
Art Delcher, University of Maryland
Nathan Edwards, University of Maryland
Wu-Chun Feng, Los Alamos National Laboratory
Guang Gao, Univerity of Delaware
Attila Gursoy, Koc University
Sorin Istrail, Brown University
Luay Nakhleh, Rice University
Jan Prins, University of North Carolina at Chapel Hill
Joel Saltz, Ohio State University
Alejandro A. Schäffer, National Institutes of Health
Alexandros Stamatakis, FORTH-ICS
Michela Taufer, University of Texas at El Paso
Thomas Wu, Genentech
Albert Y. Zomaya, University of Western Australia

Bio-Sequence Database Scanning on a GPU

Weiguo Liu, Bertil Schmidt, Gerrit Voss, Andre Schroder and Wolfgang Muller-wittig

School of Computer Engineering, Centre for Advanced Media Technology
Nanyang Technological University
639798, Singapore
{*liuweiguo, asbschmidt, askwmwittig*}*@ntu.edu.sg, vossg@camtech.ntu.edu.sg, habicht@orlen.de*

Protein sequences with unknown functionality are often compared to a set of known sequences to detect functional similarities. Efficient dynamic programming algorithms exist for this problem, however current solutions still require significant scan times. These scan time requirements are likely to become even more severe due to the rapid growth in the size of these databases. In this paper, we present a new approach to bio-sequence database scanning using computer graphics hardware to gain high performance at low cost. To derive an efficient mapping onto this type of architecture, we have reformulated the Smith-Waterman dynamic programming algorithm in terms of computer graphics primitives. Our OpenGL implementation achieves a speedup of approximately sixteen on a high-end graphics card over available straightforward and optimized CPU Smith-Waterman implementations.

Some Initial Results on Hardware BLAST Acceleration with a Reconfigurable Architecture

Euripides Sotiriades, Christos Kozanitis and Apostolos Dollas

Department of Electronic and Computer Engineering
Technical University of Crete
Chania, Crete, Greece
{*esot, kozanit, dollas*}*@mhl.tuc.gr*

The BLAST algorithm is the prevalent tool that is used by molecular biologists for DNA Sequence Matching and Database Search. In this work we demonstrate that with an appropriate reconfigurable architecture, BLAST performance can be improved with a single-chip solution 5 times over a specialized and optimized computer cluster, or 37 times over a single computer. These initial results account for I/O and are very encouraging for the development of a large scale,reconfigurable BLAST engine.

Phylospaces: Reconstructing Evolutionary Trees in Tuple Space

Marc L. Smith[1] and Tiffani L. Williams[2]

[1]Department of Computer Science
Colby College
Waterville, ME, USA
mlsmith@colby.edu

[2]Department of Computer Science
Texas A&M University
College Station, TX, USA
tlw@cs.tamu.edu

Phylospaces is a novel framework for reconstructing evolutionary trees in tuple space, a distributed shared memory that permits processes to communicate and coordinate with each other. Our choice of tuple space as a concurrency model is somewhat unusual, given the prominence and success of pure message passing models, such as MPI. We use Phylospaces to devise Cooperative Rec-I-DCM3, a population-based strategy for navigating tree space. Cooperative Rec-I-DCM3 is based on Rec-I-DCM3, the fastest sequential algorithm under maximum parsimony. We compare the performance of the algorithms on two datasets consisting of 2,000 and 7,769 taxa, respectively. Our results demonstrate that Cooperative Rec-I-DCM3 outperforms its sequential counterpart by at least an order of magnitude.

Parallel Implementation of a Quartet-Based Algorithm for Phylogenetic Analysis

B. B. Zhou[1], D. Chu[1], M. Tarawneh[1], P. Wang[1], C. Wang[1], A. Y. Zomaya[1] and R. P. Brent[2]

[1]School of Information Technologies
University of Sydney
Sydney, NSW, Australia
{bbz, dchu, monther, pwan, cwan, zomaya}@it.usyd.edu.au

[2]Mathematical Science Institute
Australian National University
Canberra, ACT, Australia
rpb@rpbrent.co.uk

This paper describes a parallel implementation of our recently developed algorithm for phylogenetic analysis on the IBM BlueGene/L cluster. This algorithm constructs evolutionary trees for a given set of DNA or protein sequences based on the topological information of every possible quartet trees. Our experimental results showed that it has several advantages over many popular algorithms. By distributing the quartet weights evenly across the processing nodes and making effective use of a fast collective network on the IBM BlueGene/L cluster, we are able to achieve a close to linear speedup even when the number of processors involved in the computation is large.

Phylogenetic Models of Rate Heterogeneity: A High Performance Computing Perspective

Alexandros Stamatakis

Foundation for Research and Technology Hellas
Heraklion, Crete, Greece
stamatak@ics.forth.gr

Inference of phylogenetic trees using the maximum likelihood (ML) method is NP-hard. Furthermore, the computation of the likelihood function for huge trees of more than 1,000 organisms is computationally intensive due to a large amount of floating point operations and high memory consumption. Within this context, the present paper compares two competing mathematical models that account for evolutionary rate heterogeneity: the Γ and CAT models. The intention of this paper is to show that—from a purely empirical point of view—CAT can be used instead of Γ. The main advantage of CAT over Γ consists in significantly lower memory consumption and faster inference times. An experimental study using RAxML has been performed on 19 real-world datasets comprising 73 up to 1,663 DNA sequences. Results show that CAT is on average 5.5 times faster than Γ and—surprisingly enough—also yields trees with slightly superior **Γ likelihood values**. The usage of the CAT model decreases the amount of average L2 and L3 cache misses by factor 8.55.

Parallel Multiple Sequence Alignment with Local Phylogeny Search by Simulated Annealing

Jaroslaw Zola[1], Denis Trystram[1], Andrei Tchernykh[2] and Carlos Brizuela[2]

[1]*Laboratoire ID-IMAG*
Montbonnot, France
{zola, trystram}@imag.fr

[2]*CICESE*
Ensenada, Mexico
{chernykh, cbrizuel}@cicese.mx

The problem of multiple sequence alignment is one of the most important problems in computational biology. In this paper we present a new method that simultaneously performs multiple sequence alignment and phylogenetic tree inference for large input data sets. We describe a parallel implementation of our method that utilises simulated annealing metaheuristic to find locally optimal phylogenetic trees in reasonable time. To validate the method, we perform a set of experiments with synthetic as well as real-life data.

MT-ClustalW: Multithreading Multiple Sequence Alignment

Kridsadakorn Chaichoompu[1], Surin Kittitornkun[1] and Sissades Tongsima[2]

[1]*Dept.of Computer Engineering, Faculty of Engineering*
King Mongkuts Institute of Technology Ladkrabang
Ladkrabang, Bangkok, Thailand
{s7060809, kksurin}@kmitl.ac.th

[2]*National Center for Genetic Engineering and Biotechnology*
Klong Luang, Pathumtani, Thailand
sissades@biotec.or.th

ClustalW is the most widely used tool for aligning multiple protein or nucleotide sequences. The alignment is achieved via three stages: pairwise alignment, guide tree generation and progressive alignment. This paper analyzes and enhances a multithreaded implementation of ClustalW called ClustalW-SMP for higher throughput. Our goal is to multithread ClustalW maximize the degree of parallelism on multithreading ClustalW called MultiThreading-ClustalW (MT-ClustalW). As a result, bioinformatics laboratories are able to use this MT-ClustalW with much less energy consumption on multicore and SMP (Symmetric MultiProcessor) machines than that of PC clusters. The experiment results show that the MT-ClustalW framework can achieve a considerable speedup over the sequential ClustalW and original multithreaded ClustalW-SMP implementations.

Parallel Implementation of the Replica Exchange Molecular Dynamics Algorithm on Blue Gene/L

M. Eleftheriou[1], A. Rayshubski[1], J. W. Pitera[2], B. G. Fitch[1], R. Zhou[1] and R. S. Germain[1]

[1]*IBM T. J. Watson Research Center*
Yorktown Heights, NY, 10598-0218
{mariae, arayshu, bgf, ruhongz, rgermain}@us.ibm.com

[2]*IBM Almaden Research Center*
San Jose, CA, 95120-6099
pitera@us.ibm.com

The Replica Exchange method is a popular approach for studying the folding thermodynamics of small to modest size proteins in explicit solvent, since it is easily parallelized. However, Replica Exchange can become computationally expensive for large-scale studies, due to the number of replicas needed as well as interprocessor communication requirements both between and within replicas. In this paper we discuss an implementation of Replica Exchange Molecular Dynamics on Blue Gene/L for performing large scale simulation studies of systems of biological interest. The algorithm is tuned with an awareness of the physical network topology and hardware performance features of the Blue Gene/L architecture. Performance measurements for Replica Exchange using the Blue Matter Molecular Dynamics application are presented on Blue Gene/L hardware with up to 256 replicas simulated on 8,192 compute nodes. Both scalability and performance are achieved with this implementation.

Application Re-Structuring and Data Management on a GRID Environment: a Case Study for Bioinformatics

Giovanni Ciriello[1], Matteo Comin[1] and Concettina Guerra[1,2]

[1]*Dept. of Information Engineering*
University of Padova
Padova, Italy
{*ciriello, ciompin, guerra*}*@dei.unipd.it*

[2]*College of Computing*
Georgia Institute of Technology
Atlanta, GA, US

This paper describes a distributed implementation of PROuST, a method for protein structure comparison, that involves a major restructuring of the application for an efficient grid immersion. PROuST consists of several components that perform different tasks at different stages. Given a target protein, an index-based search retrieves from a database a list of proteins that are good candidates for similarity, then a dynamic programming algorithm aligns the target protein with each candidate protein. The same geometric properties of secondary structure elements of proteins are used by different components of PROuST. Thus, an important issue of the distributed implementation is data transfer vs. data recomputation tradeoffs. Our implementation avoids recomputation by re-using the hash table data as much as possible, once they are accessed. The algorithmic changes to the application allow to reduce the number of data accesses to storage elements and consequently the execution time. In addition this paper discusses data replication strategies on a grid environment to optimize the data transfer time.

A Method to Improve Structural Modeling Based on Conserved Domain Clusters

Fa Zhang[1], Lin Xu[1] and Bo Yuan[2]

[1]*Institute of Computing Technology*
Chinese Academy of Sciences
Beijing, P.R.China
{*zf, xulin*}*@ncic.ac.cn*

[2]*Biomedical Informatics and Pharmacology*
The Ohio State University
Columbus, OH, USA
yuan.33@osu.edu

Homology modeling requires an accurate alignment between a query sequence and its homologs with known three-dimensional (3D) information. Current structural modeling techniques largely use entire protein chains as templates, which are selected based only on their sequence alignments with the queries. Protein can be largely described as combinations of conserved domains, and already more than two-third of the known protein domains can be found in the Protein Data Bank (PDB). We presented a method to improve structural modeling based on conserved domain clusters. First, we searched and mapped all the InterPro domains in the entire PDB, partitioned and clustered homologous domains into the domain-based template library. For each of the resulting clusters created, a multiple structural alignment was generated based only on the 3D coordinates of all the residues involved. Then we used the structural alignments as anchors to increase the alignment accuracy between a query and its templates, and consequently improve the quality of predicted structure for query protein. We implemented the method on DAWNING 4000A cluster system. The preliminary results show that our domain-based template library and the structure-anchored alignment protocol can be used for the partial prediction for a majority of known protein sequences with better qualities.

An Experimental Study of Optimizing Bioinformatics Applications

Guangming Tan[1,2], Lin Xu[1,2], Shengzhong Feng[1] and Ninghui Sun[1]

[1] *Institute of Computing Technology*
Chinese Academy of Sciences
Beijing, P. R. China
{*tgm, xulin, fsz, snh*}*@ncic.ac.cn*

[2] *Graduate School of Chinese Academy of Sciences*
Chinese Academy of Sciences
Beijing, P. R. China

As bioinformatics is an emerging application of high performance computing, this paper first evaluates the memory performance of several representative bioinformatics applications so that some appropriate optimization methods can be applied. Based on the computational behavior of these bioinformatics applications, we propose two optimized algorithms on high performance computer architectures. 1) For the data(I/O) intensive program, MegaBlast, we overlap computation with I/O to produce an improved high-throughput algorithm with reduced time and memory requirements. 2) For a CPU-intensive RNA secondary structure prediction algorithm, we propose a fine-grain parallel $O(N^3)$ algorithm based on reconfigurable arrays (FPGAs). In order to optimize the FPGA architecture, we evaluate the performance in different architectures using cycle-by-cycle simulator.

Workshop 8

Advances in Parallel and Distributed Computing Models
APDCM 2006

8 APDCM • Workshop on Advances in Parallel and Distributed Computing Models

Workshop Description:

The past twenty years have seen a flurry of activity in the arena of parallel and distributed computing. In recent years, novel parallel and distributed computational models have been proposed in the literature, reflecting advances in new computational devices and environments such as optical interconnects, programmable logic arrays, networks of workstations, radio communications, mobile computing, DNA computing, quantum computing, sensor networks etc. It is very encouraging to note that the advent of these new models has lead to significant advances in the resolution of various difficult problems of practical interest. The main goal of this workshop is to provide a timely forum for the exchange and dissemination of new ideas, techniques and research in the field of the parallel and distributed computational models. The workshop is meant to bring together researchers and practitioners interested in all aspects of parallel and distributed computing taken in an inclusive, rather than exclusive, sense.

Topics of interest include but are not limited to:

- Models of Parallel and Distributed Computing
  - BSP and LogP models
  - Radio communication models
  - Mobile computing models
  - Sensor network models
  - Hardware-specific models
  - Systolic arrays and cellular automata
  - Biologically-based computing models
  - Quantum models
  - Reconfigurable models
  - Optical models
- Algorithms and Applications
  - Geometric and graph algorithms
  - Combinatorial algorithms
  - Randomized and approximation techniques
  - Numerical algorithms
  - Network algorithms
  - Localized algorithms
  - Distributed algorithms
  - Image processing
  - High-performance computing
- Practical Aspects
  - Architectural and implementation issues
  - Performance analysis and simulation
  - PVM/MPI
  - Programmable logic arrays
  - Design of network protocols
  - Ad-hoc networks
  - Development tools
  - Fault tolerance

Workshop Chair:

Oscar H. Ibarra, University of California, Santa Barbara, USA

Program Co-chairs:

Koji Nakano, Hiroshima University, Japan

Jacir L. Bordim, Brasilia University, Brazil

Program Committee:

Anu Bourgeois, Georgia State University, USA
Satoshi Fujita, Hiroshima University, Japan
Akihiro Fujiwara, Kyushu Institute of Technology, Japan
Shuichi Ichikawa, Toyohashi University of Technology, Japan
Yasushi Inoguchi, JAIST, Japan
Chuzo Iwamoto, Hiroshima University, Japan
Xiaohong Jiang, Tohoku University, Japan
Hirotsugu Kakugawa, Osaka University, Japan
Weifa Liang, Australian National University, Australia
Rong Lin, State University of New York, USA
Susumu Matsumae, Tottori University of Environmental Studies, Japan
Eiji Miyano, Kyushu Institute of Technology, Japan
Mitsuo Motoki, JAIST, Japan
Sanguthevar Rajasekaran, University of Connecticut, USA
Ivan Stojmenovic, University of Ottawa, Canada
Yasuhiko Takenaga, University of Electro-communications, Japan
Jerry L. Trahan, Louisiana State University, USA
Ramachandran Vaidyanathan, Lousiana State University, USA
Biing-Feng Wang, National Tsinhua University, Taiwan
Dajin Wang, MontClair State University, USA
José Alberto Fernández Zepeda, CICESE, Mexico
Jingyuan Zhang, University of Alabama, USA

APDCM Keynote: Learning Computing Models from Cells and Tissues: P Systems

Gheorghe Paun

Department of Computer Science and Artificial Intelligence
University of Sevilla
Sevilla, Andalucia, Spain
gpaun@us.es

This is intended to be a quick introduction to membrane computing, a branch of natural computing inspired in the structure and functioning of living cells and in their organization in tissues. The corresponding models, called P systems, are parallel distributed computing devices, handling multisets of abstract objects in a compartmentalized architecture defined by cell-like or tissue-like membrane arrangements. Most classes of P systems are Turing universal and in certain cases can provide polynomial solutions to computationally hard problems (by means of a time-space trade-off). Several applications in biology/medicine, computer science, linguistics, economics were reported. The talk will present only basic ideas and (types of) results and of applications. Details can be found at the Web site `http://psystems.disco.unimib.it`.

Optimal Map Contruction of an Unknown Torus

Hanane Becha[1] and Paola Flocchini[2]

[1] *School of Information Technology and Engineering*
University of Ottawa
Ottawa, Ontario, Canada
hbecha@site.uottawa.ca

[2] *School of Information Technology and Engineering*
University of Ottawa
Ottawa, Ontario, Canada
flocchin@site.uottawa.ca

In this paper we consider the map construction problem in the case of an *anonymous, unoriented* torus of unknown size. An agent that can move from node to neighbouring node in the torus is initially placed in an arbitrary node and has to construct an edge-labeled map. In other words, it has to draw, in its local memory, an edge-labeled torus isomorphic to the one it is moving on. The agent has enough local memory to represent the torus and one or two tokens that can be dropped on and picked up from nodes. Efficiency is measured in terms of number of moves performed by the agent.

When the agent has no token available, the problem is clearly unsolvable. In the paper we show that, when the agent has one token available there exists an optimal algorithm for constructing the map of the torus; the agent, in fact, performs $\Theta(N)$ moves (where N is the number of nodes of the torus). Before showing the optimal solution with the optimal number of tokens, we describe a simpler solution that works when two tokens are available, we then modify it to obtain the same bound when the agent has only one token available.

Ant-inspired Query Routing Performance in Dynamic Peer-to-Peer Networks

Mojca Ciglaric and Tone Vidmar

Faculty of Computer and Information Science
University of Ljubljana
Ljubljana, Slovenia
{*mojca.ciglaric, tone.vidmar*}*@fri.uni-lj.si*

P2P Networks are highly dynamic structures since their nodes peer users keep joining and leaving continuously. In the paper, we study the effects of network change rates on query routing efficiency. First, the problem background is described and abstract system model is defined. The system characteristics and behavior are analyzed and abstracted with a set of measurable metrics. The paper studies ant-inspired Mute query routing protocol and compares its behavior to previously suggested routing protocols. The chosen routing technique makes use of cached metadata from previous answer messages (analogy to ants laying feromone on their trail when searching for food). The paper also discusses mechanisms for broken path detection and metadata maintenance. Further, simulations in various dynamic network environments are presented and discussed: the degree of network dynamics varies from one node departure and node join per ten queries generated to five node departures and joins per one generated query. Several metrics are used to clarify the protocol behavior even with high rate of node departures, but it is shown that above a certain threshold it literally breaks down and exhibits considerable efficiency degradation.

Decontamination of Chordal Rings and Tori

Paola Flocchini[1], Miaojun Huang[1] and Flaminia Luccio[2]

[1] *School of Information Technology and Engineering*
University of Ottawa
Ottawa, Canada
{*flocchin, mhuang*}*@site.uottawa.ca*

[2] *Dipartimento di Matematica e Informatica*
University of Trieste
Trieste, Italy
luccio@dsm.univ.trieste.it

In this paper we consider the problem of decontaminating a network, i.e., protecting it from unwanted and dangerous intrusions. Initially all nodes are contaminated and a team of agents is deployed to clean the entire network. When an agent transits on a node, it can clean it, when the node is left unguarded, however, it will be recontaminated as soon as at least one of its neighbour is contaminated. We study the problem in asynchronous chordal ring networks with n nodes and chord lengths $d_1 = 1, d_2, ..., d_k$, and in tori.

We consider two variations of the model: one where an agent has only local knowledge, the other in which it has "visibility", i.e., it can "see" the state of its neighbouring nodes.

We first show that, when the largest chord d_k is not too large ($d_k \leq \sqrt{n}$), the number of agents necessary to perform the task in chordal rings does not depend on the size of the network but only on the length of the longest chord. We also show a lower bound on the number of agents for the torus topology. We then propose tight strategies for decontamination. We analyse the number of moves and the time complexity of the decontamination algorithms showing that the visibility assumption allows us to decrease substantially both complexity measures. Another advantage of the "visibility model" is that agents move independently and autonomously without requiring any coordination.

Reducing the Associativity and Size of Step Caches in CRCW Operation

Martti Forsell

Platform Architectures Team
VTT Technical Research Center of Finland
Oulu, Finland
Martti.Forsell@VTT.Fi

Step caches are caches in which data entered to an cache array is kept valid only until the end of ongoing step of execution. Together with an advanced pipelined multithreaded architecture they can be used to implement concurrent read concurrent write (CRCW) memory access in shared memory multiprocessor systems on chip (MP-SOC) without cache coherency problems. Unfortunately obvious step cache architectures assume full associativity, which can become expensive since the size and thus associativity of caches equal the number of threads per processor being at least the square root of the number of processors. In this paper, we describe a technique to radically reduce the associativity and even size of step caches in CRCW operation. We give a short performance evaluation of limited associativity step cache systems with different settings using simple parallel programs on a parametrical MP-SOC framework. According to the evaluation, the performance of limited associativity step cache systems comes very close to that of fully associative step cache systems, while decreasing the size of caches decreases the performance gradually.

Simulating a PR-Mesh on an LARPBS

Mathura Gopalan[1], Anu Goel Bourgeois[1] and José Alberto Fernández Zepeda[2]

[1]*Department of Computer Science*
Georgia State University
Atlanta, GA, USA
mathura.gopalan@gmail.com, abourgeois@cs.gsu.edu

[2]*Department of Computer Science*
CICESE
Ensenada, B. C., Mexico
fernan@cicese.mx

The unidirectional nature of propagation and predictable delays are two characteristics of optically pipelined buses that have made them popular in recent years. Many models have been proposed that use reconfigurable optically pipelined buses. In this paper we establish a relationship between a one dimensional and a two dimensional model of this type. This simulation shows that the challenge is to map the processors so that those belonging to a two-dimensional bus segment are contiguous and in the same order on the simulating one-dimensional model. We focus on the Linear Array with a Reconfigurable Pipelined Bus System (LARPBS) and its two dimensional counterpart the Pipelined Reconfigurable Mesh (PR-Mesh).

A Strategyproof Mechanism for Scheduling Divisible Loads in Bus Networks without Control Processors

Thomas E. Carroll and Daniel Grosu

Department of Computer Science
Wayne State University
Detroit, MI, USA
{*tec, dgrosu*}*@cs.wayne.edu*

Divisible Load Theory (DLT) considers the scheduling of arbitrarily partitionable loads in distributed systems. The underlying assumption of DLT is that the processors are obedient (*i.e.*, they do not "cheat" the protocol), which is unrealistic when the processors are owned by autonomous, self-interested organizations that have no *a priori* motivation for cooperation and which strive to maximize their own welfare. In this scenario, they will manipulate the algorithm if it is beneficial to do so. In this paper we propose a strategyproof mechanism for scheduling divisible loads in bus networks *without* control processors. We augment DLT with incentives so that it is to the benefit of a processor to truthfully report its processing capacity and to process its assignment at full capacity. The mechanism provides incentives to processors for reporting deviants and issues fines to deviants, which results in abated willingness to deviate.

Efficient Hardware Algorithms for n Choose k Counters

Yasuaki Ito, Koji Nakano and Youhei Yamagishi

Graduate School of Engineering
Hiroshima University
Higashi-Hiroshima, Hiroshima, Japan
{*yasuaki, nakano*}*@cs.hiroshima-u.ac.jp, youhei@cs.hirohsima-u.ac.jp*

An "n choose k" counter (C(n,k) counter for short) is a counter which lists all n-bit numbers with (n-k) 0's and k 1's. The "n choose k" counter has applications to solving combinatorial optimization problems and image processing. The main contribution of this work is to present an efficient hardware implementation of the C(n,k) counter. In some applications, C(n,k) counters are used only for small k. The second contribution is to show more efficient implementations that support C(n,k) counters only for small k. We evaluate the performance of our new implementation and known implementations in terms of the number of used slices and the clock frequency for the Xilinx VirtexII family FPGA XC2V3000-4. Although the theoretical analysis shows that our implementation is not the best, it runs in higher clock frequency using fewer number of slices than the other implementations.

A Self-Stabilizing Minimal Dominating Set Algorithm with Safe Convergence

Hirotsugu Kakugawa and Toshimitsu Masuzawa

Department of Computer Science
Osaka Universtiy
Osaka, Japan
{*kakugawa, masuzawa*}*@ist.osaka-u.ac.jp*

A self-stabilizing distributed system is a fault-tolerant distributed system that tolerates any kind and any finite number of transient faults, such as message loss and memory corruption. In this paper, we formulate a concept of safe convergence in the framework of self-stabilization. An ordinary self-stabilizing algorithm has no safety guarantee while it is in converging from any initial configuration. The safe convergence property guarantees that a system quickly converges to a safe configuration, and then, it gracefully moves to an optimal configuration without breaking safety. Then, we propose a minimal independent dominating set algorithm with safe convergence property. Especially, the proposed algorithm computes the lexicographically first minimal independent dominating set according to the process identifier as a priority. The priority scheme can be arbitrarily changed such as stability, battery power and/or computation power of node.

A Framework for Developing Distributed Location Based Applications

Andrej Krevl and Mojca Ciglaric

Faculty of computer and Information Science
University of Ljubljana
Ljubljana, Slovenia
andrej.krevl@fri.uni-lj.si

Location based services and applications are buzzwords nowadays, yet they have been around for quite some time in a variety of applications. However these applications are scarce because of the high costs associated with the positioning equipment. This paper presents different options for determining location of mobile devices such as mobile phones and Pocket PCs. It describes positioning possibilities using WiFi networks, GSM networks, Bluetooth beacons and the GPS system. Furthermore, it proposes a framework for developing distributed location based applications. The paper specifies which components comprise the framework, data structures that are used for spatial data interchange and Web Services that are used for communication between components. It also describes a location aware application prototype built on top of the proposed framework. It concludes that building applications on top of the proposed framework is feasible and discusses benefits and drawbacks of this approach.

A Calculus of Functional BSP Programs with Projection

Frédéric Loulergue

Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)
University of Orléans
Orléans, France
frederic.loulergue@univ-orleans.fr

Bulk Synchronous Parallel ML (BSML) is an extension of the functional language Objective Caml to program Bulk Synchronous Parallel (BSP) algorithms. It is deterministic, deadlock free and performances are good and predictable. Parallelism is expressed with a set of 4 primitives on a parallel data structure called parallel vector. These primitives are pure functional ones: they have no side-effect. It is thus possible to prove the correctness of BSML programs using a proof assistant like Coq. The BSλ-calculus is an extension of the λ-calculus which models the core semantics of BSML. Nevertheless some principles of BSML are not well captured by this calculus. This paper presents a new calculus, with a projection primitive, which provides a better model of the core semantics of BSML.

Network Decontamination with Local Immunization

Fabrizio Luccio[1], Linda Pagli[1] and Nicola Santoro[2]

[1] *Dipartimento di Informatica*
University of Pisa
Pisa, Italy
{*luccio, pagli*}*@di.unipi.it*

[2] *School of Computer Science*
Carleton University
Ottawa, Ontario, Canada
santoro@scs.carleton.ca

We consider the problem of decontaminating a network infected by a mobile virus. The task is to be carried out by a team of antiviral system agents (*cleaners*), able to disinfect visited sites, avoiding any recontamination of disinfected areas. The goal is to perform the task using as small a team of antiviral agents as possible and minimizing the amount of agents' movements across the network. In all the existing literature, it is assumed that a disinfected site, in absence of a cleaner, becomes recontaminated if just one of its neighbours is contaminated. In other words, it is assumed that the immunity level of a disinfected site is *nil*. This assumption is quite strong and not necessarily realistic, e.g., in systems that employ local majority-based rules to enhance reliability and fault-tolerance. In this paper we consider the network decontamination problem under a new model of *immunity* to recontamination: we consider the case when a disinfected vertex, after the cleaning agent has gone, will become recontaminated only if a weak majority of its neighbours are infected. We study the effects of this level of immunity on the nature of the problem, in particular on the number of antiviral agents necessary to decontaminate the entire network. We focus on tori and on trees, and establish lower-bounds on the team size; we also establish lower bounds on the number of moves performed by an optimal-size time of cleaners. We design and present strategies for disinfecting tori and trees; we prove that these strategies are optimal in terms of both team size and number of moves. In particular, the upper and lower bounds are are tight for tree networks and for synchronous tori; the bounds are within a constant factor of each other in the case of asynchronous tori.

An Advanced Performance Analysis of Self-stabilizing Protocols : Stabilization Time with Transient Faults during Convergence

Yoshihiro Nakaminami, Hirotsugu Kakugawa and Toshimitsu Masuzawa

Information Science and Technology
Osaka University
Toyonaka-city, Osaka, Japan
{*nakaminm, kakugawa, masuzawa*}*@ist.osaka-u.ac.jp*

A self-stabilizing protocol is a brilliant framework for fault tolerance. It can recover from any number and any type of transient faults and eventually converge to its intended behavior. Performance of a self-stabilizing protocol is usually measured by stabilization time: the time required to complete the convergence to its intended behavior under the assumption that no new fault occurs during the convergence. But a self-stabilizing protocol has no guarantee to complete the convergence if faults are frequently occurred.

This paper brings new light to efficiency analysis of stabilization. The efficiency is evaluated with consideration for faults occurring during the convergence. We propose a new performance measure of self-stabilizing protocols, a stabilization time with $\sharp F$ faults. It is the worst case time to converge to intended behaviors with consideration for $\sharp F$ faults occurring during the convergence. To show the feasibility and effectiveness of the approach, this paper applies the approach to the maximal matching protocol proposed by Hsu and Huang and show that its stabilization time with $\sharp F$ faults is $2m + n + 4\Delta \cdot \sharp F$ where m, n and Δ are the number of links, the number of vertices and maximum degree respectively.

Cache-Oblivious Simulation of Parallel Programs

Andrea Pietracaprina, Geppino Pucci and Francesco Silvestri

Department of Information Engineering
University of Padova
Padova, Italy
{*capri, geppo, silvest1*}*@dei.unipd.it*

This paper explores the relation between the structured parallelism exposed by the Decomposable BSP (D-BSP) model through submachine locality and locality of reference in multi-level cache hierarchies. Specifically, an efficient cache-oblivious algorithm is developed to simulate D-BSP programs on the Ideal Cache Model (ICM). The effectiveness of the simulation is proved by showing that optimal cache-oblivious algorithms for prominent problems can be obtained from D-BSP algorithms. Finally, a tight relation between optimality in the D-BSP and ICM models is established.

Enhancing the Performance of HLA-Based Simulation Systems via Software Diversity and Active Replication

Francesco Quaglia

Dipartimento di Informatica e Sistemistica
"La Sapienza"
Roma, Italy
quaglia@dis.uniroma1.it

In this paper we explore active replication based on software diversity for improving the responsiveness of simulation systems. Our proposal is framed by the High-Level-Architecture (HLA), namely the emerging standard for interoperability of simulation packages, and results in the design and implementation of an Active Replication Management Layer (ARML), which supports the execution of multiple software diversity-based replicas of a same simulator in a totally transparent manner. Beyond presenting the replication framework and the design/implementation of ARML, we also report the results of an experimental evaluation on a case study, quantifying the benefits from our proposal in terms of execution speed.

Broadcasting and Routing in Faulty Mesh Networks

Milos Stojmenovic and Amiya Nayak

SITE
University of Ottawa
Ottawa, Ontario, Canada
milos22@gmail.com, anayak@site.uottawa.ca

Broadcasting is a data communication task in which one processor sends the same message to all other processors. Routing is a task where a source processor sends a message to a destination processor. A faulty node is in an error state and cannot participate in the activities or the communication in a given network. In this paper, we consider the family of mesh networks, which include the mesh connected computer (MCC), k-dimensional mesh, torus, and k-ary n-cube. Our goal is to design routing and broadcasting algorithms which will use local knowledge of faults, no additional resources, will work for an arbitrary number and structure of faults, will guarantee delivery to all nodes connected to the source, and will remain optimal in a fault free mesh. We did not find any solution in literature to satisfy these desirable properties. Our routing and broadcasting schemes for MCCs and tori, and our broadcasting algorithm for the all-port model on any faulty mesh network satisfy all of these properties. For routing and broadcasting in a one-port model in higher dimensions, a condition on fault structure needs to be met. We propose a new broadcasting algorithm which guarantees delivery to all processors connected to the source in the all-port model of faulty meshes. We then describe a routing algorithm that guarantees delivery in faulty MCCs and tori, the connectivity of the source and destination being the only obvious requirement. The algorithm can be extended to faulty k-D meshes and k-ary n-cubes, where the delivery will be guaranteed if healthy nodes in every 2-D submesh (sub-tori) remain connected. We then describe broadcasting algorithms for the one-port model, which again guarantee delivery to all connected processors in two-dimensional cases, and guarantee delivery in k-dimensional cases if healthy processors in every 2-D submesh (sub-tori) remain connected.

Self-Stabilizing Distributed Algorithms for Graph Alliances

Pradip Srimani and Zhenyu Xu

Computer Science
Clemson University
Clemson, SC, USA
srimani@cs.clemson.edu, zxu@clemson.edu

Graph alliances are recently developed global properties of any symmetric graph. Our purpose in the present paper is to design self-stabilizing fault tolerant distributed algorithms for the global offensive and the global defensive alliance in a given arbitrary graph. We also provide complete analysis of the convergence time of both the algorithms.

Workshop 9

Communication Architecture for Clusters CAC 2006

9 CAC • Workshop on Communication Architecture for Clusters

Workshop Description:

Many of the world's fastest computer systems are PC or workstation clusters. Numerous research groups in academia, industry, and government are currently engaged in cluster research, seeking new ways to advance the state of the art of cluster communication. The goal of the CAC workshop is to bring together researchers and practitioners working in the areas of communication hardware and software to discuss their latest findings as well as future trends in the design of scalable, high-performance, and cost-effective communication architectures for clusters.

Topics of interest include but are not limited to:

- novel network-interface and switch architectures for supporting efficient point-to-point and collective communication,
- design, implementation, and optimization of low-level communication protocols (e.g., VAPI, Tports, GM, and IP) and higher-level communication layers (e.g., MPI, sockets, put/get, and distributed shared memory),
- tools and techniques for evaluating cluster and application performance, and
- communication and architectural issues related to router/switch organization, flow control, congestion control, routing and deadlock handling, load balancing, reliability, QoS support, topology discovery, and dynamic reconfiguration.

Co-Chairs:

Scott Pakin (LANL)
Mazin Yousif (Intel)

Program Committee:

Angelos Bilas (FORTH & U. Crete)
Ron Brightwell (SNL)
Darius Buntinas (ANL)
Wu-Chun Feng (VT)
José Flich (UPV)
Mitchell Gusat (IBM)
Ben Lee (OSU)
Olav Lysne (Oslo)
Jarek Nieplocha (PNNL)
Greg Pfister (IBM)
Timothy Pinkston (USC)
Vikram Saletore (Intel)
Cris Simpson (Intel)
Evan Speight (IBM)
Pete Wyckoff (OSC)

Seekable Sockets: A Mechanism to Reduce Copy Overheads in TCP-based Messaging

Chase Douglas and Vijay S. Pai

School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN, 47907
{*cndougla, vpai*}*@purdue.edu*

This paper extends the traditional socket interface to TCP/IP communication with the ability to seek rather than simply receive data in order. Seeking on a TCP socket allows a user program to receive data without first receiving all previous data on the connection. Through repeated use of seeking, a messaging application or library can treat a TCP socket as a list of messages with the potential to receive and remove data from any arbitrary point rather than simply the head of the socket buffer. Seeking facilitates copy-avoidance between a messaging library and user code by eliminating the need to first copy unwanted data into a library buffer before receiving desired data that appears later in the socket buffer.

The seekable sockets interface is implemented in the Linux 2.6.13 kernel. Experimental results are gathered using a simple microbenchmark that receives data out-of-order from a given socket, yielding up to a 40% reduction in processing time. The code for seekable sockets is now available for patching into existing Linux kernels and for further development into messaging libraries.

Asynchronous Zero-copy Communication for Synchronous Sockets in the Sockets Direct Protocol (SDP) over InfiniBand

P. Balaji, S. Bhagvat, H. W. Jin and D. K. Panda

Computer Science and Engineering
Ohio State University
Columbus, Ohio, USA
{*balaji, bhagvat, jinhy, panda*}*@cse.ohio-state.edu*

Sockets Direct Protocol (SDP) is an industry standard pseudo sockets-like implementation to allow existing sockets applications to directly and transparently take advantage of the advanced features of current generation networks such as InfiniBand. The SDP standard supports two kinds of sockets semantics, viz., Synchronous sockets (e.g., used by Linux, BSD, Windows) and Asynchronous sockets (e.g., used by Windows, upcoming support in Linux). Due to the inherent benefits of asynchronous sockets, the SDP standard allows several intelligent approaches such as *source-avail and sink-avail based zero-copy* for these sockets. Unfortunately, most of these approaches are not beneficial for the synchronous sockets interface. Further, due to its portability, ease of use and support on a wider set of platforms, the synchronous sockets interface is the one used by most sockets applications today. Thus, a mechanism by which the approaches proposed for asynchronous sockets can be used for synchronous sockets is highly desirable. In this paper, we propose one such mechanism, termed as *AZ-SDP (Asynchronous Zero-Copy SDP)*, where we memory-protect application buffers and carry out communication asynchronously while maintaining the synchronous sockets semantics. We present our detailed design in this paper and evaluate the stack with an extensive set of benchmarks. The experimental results demonstrate that our approach can provide an improvement of close to 35% for medium-message uni-directional throughput and up to a factor of 2 benefit for computation-communication overlap tests and multi-connection benchmarks.

Fast Barrier Synchronization for InfiniBand

Torsten Hoefler, Torsten Mehlan, Frank Mietke and Wolfgang Rehm

Dept. of Computer Science
Chemnitz University of Technology
Chemnitz, 09107, Germany
{*htor, tome, mief, rehm*}*@cs.tu-chemnitz.de*

The MPI_Barrier() call can be crucial for several applications and has been target of different optimizations since several decades. The best solution to the barrier problem scales with $O(log_2 N)$ and uses the dissemination principle. A new method using an enhanced dissemination principle and inherent network parallelism will be demonstrated in this paper. The new approach was able to speedup the barrier performance by 40% in relation to the best published algorithm. It is shown that it is possible to leverage the inherent hardware parallelism inside the InfiniBand™network to lower the latency of the MPI_Barrier() operation without additional costs. The principle of sending multiple messages in (pseudo-) parallel can be implemented into a well known algorithm to decrease the number of rounds and speed the overall operation up.

Efficient SMP-Aware MPI-Level Broadcast over InfiniBand

Amith Rajith Mamidala, Lei Chai, Hyun-wook Jin and Dhabaleswar K Panda

Computer Science and Engineering
The Ohio State University
Columbus, Ohio, USA
{*mamidala, chail, jinhy, panda*}*@cse.ohio-state.edu*

Most of the high-end computing clusters found today feature multi-way SMP nodes interconnected b an ultra-low latency and high bandwidth network. InfiniBand is emerging as a high-speed network for such systems. InfiniBand provides a scalable and efficient hardware multicast primitive to efficiently implement many MPI collective operations. However, employing hardware multicast as the communication method may not perform well in all cases. This is true especially when more than one process is running per node. In this context, shared memory channel becomes the desired communication medium within the node as it delivers latencies which are of an order of magnitude lower than the inter-node message latencies. Thus, to deliver optimal collective performance, coupling hardware multicast with shared memory channel becomes necessary. In this paper we propose mechanisms to address this issue. On a 16-node 2-way SMP cluster, the Leader-based scheme proposed in this paper improves the performance of the MPI_Bcast operation by a factor of as much as 2.3 and 1.8 when compared to the point-to-point and original solution employing only hardware multicast. We have also evaluated our designs on NUMA based system and obtained a performance improvement of 1.7 using our designs on 2-node 4-way system. We also propose a Dynamic Attach Policy as an enhancement to this scheme to mitigate the impact of process skew on the performance of the collective operation.

Efficient RDMA-based Multi-port Collectives on Multi-rail QsNetII Clusters

Ying Qian and Ahmad Afsahi

Electrical and Computer Engineering
Queen's University
Kingston, ON, Canada
qiany@ee.queensu.ca, ahmad.afsahi@queensu.ca

Many scientific applications use MPI collective communications intensively. Therefore, efficient and scalable implementation of collective operations is critical to the performance of such applications running on clusters. Quadrics QsNetII is a high-performance interconnect for clusters that implements some collectives at the Elan level. These collectives are directly used by their corresponding MPI collectives. Quadrics software supports point-to-point striping over multi-rail QsNetII networks. However, multi-rail collectives have not been supported. In this work, we propose a number of RDMA-based multi-port collectives over multi-rail QsNetII clusters directly at the Elan level. Our performance results indicate that the proposed multi-port gather gains an improvement of up to 6.35 for 1MB message over the native elan_gather. The proposed multi-port all-to-all performs better than the native elan_alltoall by a factor of 2.19 for 16KB message. Moreover, we have also proposed two algorithms for the scatter operation.

Benefits of High Speed Interconnects to Cluster File Systems: A Case Study with Lustre

Weikuan Yu[1], Ranjit Noronha[2], Shuang Liang[3] and Dhabaleswar K. Panda[4]

[1]*Computer Science & Engineering*
The Ohio State University
Columbus, Ohio, USA
yuw@cse.ohio-state.edu

[2]*Computer Science & Engineering*
The Ohio State University
Columbus, Ohio, USA
noronha@cse.ohio-state.edu

[3]*Computer Science & Engineering*
The Ohio State University
Columbus, Ohio, USA
liangs@cse.ohio-state.edu

[4]*Computer Science & Engineering*
The Ohio State University
Columbus, Ohio, USA
panda@cse.ohio-state.edu

Cluster file systems and Storage Area Networks (SAN) make use of network IO to achieve higher IO bandwidth. Effective integration of networking mechanisms is important to their performance. In this paper, we perform an evaluation of a popular cluster file system, Lustre, over two of the leading high speed cluster interconnects: InfiniBand and Quadrics. Our evaluation is performed with both sequential IO and parallel IO benchmarks in order to explore the capacity of Lustre under different communication characteristics. Experimental results show that direct implementations of Lustre over both interconnects can improve its performance, compared to an IP emulation over InfiniBand (IPoIB). The performance of Lustre over Quadrics is comparable to that of Lustre over InfiniBand with the platforms we have. Latest InfiniBand products can embrace latest technologies, such as PCI-Express and DDR, and provide higher capacity. Our results show that over a Lustre file system with two Object Storage Servers (OSSs), InfiniBand with PCI-Express technology can improve Lustre write performance by 24%. Furthermore, our experimental results indicate that Lustre meta-data operations do not scale with an increasing number of OSSs, in spite of using high performance interconnects.

iWarp Protocol Kernel Space Software Implementation

Dennis Dalessandro[1], Ananth Devulapalli[2] and Pete Wyckoff[3]

[1]Ohio Supercomputer Center
Springfield, OH, USA
dennis@osc.edu

[2]Ohio Supercomputer Center
Springfield, OH, USA
ananth@osc.edu

[3]Ohio Supercomputer Center
Columbus, OH, USA
pw@osc.edu

Zero-copy, RDMA, and protocol offload are three very important characteristics of high performance interconnects. Previous networks that made use of these techniques were built upon proprietary, and often expensive, hardware. With the introduction of iWarp, it is now possible to achieve all three over existing low-cost TCP/IP networks.

iWarp is a step in the right direction, but currently requires an expensive RNIC to enable zero-copy, RDMA, and protocol offload. While the hardware is expensive at present, given that iWarp is based on a commodity interconnect, prices will surely fall. In the meantime only the most critical of servers will likely make use of iWarp, but in order to take advantage of the RNIC both sides must be so equipped.

It is for this reason that we have implemented the iWarp protocol in software. This allows a server equipped with an RNIC to exploit its advantages even if the client does not have an RNIC. While throughput and latency do not improve by doing this, the server with the RNIC does experience a dramatic reduction in system load. This means that the server is much more scalable, and can handle many more clients than would otherwise be possible with the usual sockets/TCP/IP protocol stack.

A Look at Application Performance Sensitivity to the Bandwidth and Latency of Infiniband Networks.

Darren J Kerbyson

Performance and Architecture Lab (PAL)
Los Alamos National Laboratory
Los Alamos, NM, USA
djk@lanl.gov

This work explores the expected performance of three applications on a High Performance Computing cluster interconnected using Infiniband. In particular, the expected performance across a range of configurations is analyzed notably Infiniband 4x, 8x and 12x representing link-speeds of 10Gb/s, 20Gb/s, and 30Gb/s respectively as well as near-neighbor MPI message latencies of 4μs and 1.5μs. In addition we also consider the impact of node size, from one to eight processors that share a single network connection. The performance analysis is based on the use of detailed performance models of the three applications developed at Los Alamos. The results of the analysis show that the application performance can range by as much as 60% from best to worst. The relative importance of bandwidth, latency and node size differs between the applications.

Communication Patterns

Rolf Riesen

Scalable Computing Systems
Sandia National Laboratories
Albuquerque, NM, USA
rolf@cs.sandia.gov

Parallel applications have message-passing patterns that are important to understand. Network topology, routing decisions, and connection and buffer management need to match the communication patterns of an application for it to run efficiently and scale well. These patterns are not easily discerned from the source code of an application, and even when the data is available it is not easy to categorize it appropriately such that meaningful knowledge emerges.

We describe a novel system to gather the information we need to discover an application's communication pattern. We create five categories that help us analyze that data and explain how information from each category can be useful in the design of networking hardware and software. We use the NAS parallel benchmarks as examples on how to apply our techniques.

A Preliminary Analysis of the InfiniPath and XD1 Network Interfaces

Ron Brightwell, Doug Doerfler and Keith D. Underwood

Center for Computation, Computers, Information, and Math
Sandia National Laboratories
Albuquerque, NM, USA
{rbbrigh, dwdoerf, kdunder}@sandia.gov

Two recently delivered systems have begun a new trend in cluster interconnects. Both the InfiniPath network from PathScale, Inc., and the RapidArray fabric in the XD1 system from Cray, Inc., leverage commodity network fabrics while customizing the network interface in an attempt to add value specifically for the high performance computing (HPC) cluster market. Both network interfaces are compatible with standard InfiniBand (IB) switches, but neither use the traditional programming interfaces to support MPI. Another fundamental difference between these networks and other modern network adapters is that much of the processing needed for the network protocol stack is performed on the host processor(s) rather than by the network interface itself. This approach stands in stark contrast to the current direction of most high-performance networking activities, which is to offload as much protocol processing as possible to the network interface. In this paper, we provide an initial performance comparison of the two partially custom networks (PathScale's InfiniPath and Cray's XD1) with a more commodity network (standard IB) and a more custom network (Quadrics Elan4). Our evaluation includes several micro-benchmark results as well as some initial application performance data.

Workshop 10

NSF Next Generation Software Program Meeting
NSFNGS 2006

Workshop Description:

This workshop provides a forum for an overview, project presentations, and discussion of the research fostered and funded by the NSF Next Generation Software (NGS) Program, and the Advanced Execution Systems (AES) and the Systems Modeling and Analysis (SMA) components of the follow-up Computer Systems Research (CSR) Program. The present Workshop is part of the Next Generation Software workshop series that started in 2001 and has been conducted yearly in conjunction with IPDPS. The topics addressed in the workshop are on research in systems' software technology in the scope of NGS and of the AES and the SMA components, namely: systems modeling, analysis and performance engineering methods, programming environments, enhanced compiler capabilities, tools for the development, dynamic runtime support and dynamic composition of complex applications executing on heterogeneous, parallel and distributed computing platform assemblies, such as computational grids, encompassing high-end platforms, clusters, embedded and sensor systems, and special purpose processing systems.

Workshop Organizer:
Frederica Darema
Computer and Information Science and Engineering Directorate
National Science Foundation

Techniques and Tools for Dynamic Optimization

Jason D. Hiser[1], Naveen Kumar[2], Min Zhao[2], Shukang Zhou[1], Bruce R. Childers[2], Jack W. Davidson[1] and Mary Lou Soffa[1]

[1]*Department of Computer Science*
University of Virginia
Charlottesville, VA, USA
{*hiser, zhou, jwd, soffa*}*@cs.virginia.edu*

[2]*Department of Computer Science*
University of Pittsburgh
Pittsburgh, PA, USA
{*naveen, lilyzhao, childers*}*@cs.pitt.edu*

Traditional code optimizers have produced significant performance improvements over the past forty years. While promising avenues of research still exist, traditional static and profiling techniques have reached the point of diminishingreturns. The main problem is that these approaches have only a limited view of the program and have difficulty taking advantage of the actual run-time behavior of a program. We are addressing this problem through the development of a dynamic optimization system suited for aggressive optimizationusing the full power of the most beneficial optimizations. We have designed our optimizer to operate using a software dynamic translation (SDT) execution system. Difficult challenges in this research include reducing SDT overhead and determining what optimizations to apply and where in the code to apply them. Another challenge is having the necessary tools to ensure the reliability of software that is dynamically optimized. In this paper, we describe our efforts in reducing overhead in SDT and efficient techniques for instrumenting the application code. We also describe our approach to determine what and where an optimization should be applied. We discuss other fundamental issues in developing a dynamic optimizer and finally present a basic debugger for SDT systems.

Program Phase Detection and Exploitation

Chen Ding[1], Sandhya Dwarkadas[1], Michael C. Huang[1], Kai Shen[1] and John B. Carter[2]

[1]*Department of Computer Science*
University of Rochester
Rochester, NY, USA
{*cding, sandhya, mihuang, kshen*}*@cs.rochester.edu*

[2]*Department of Computer Science*
University of Utah
Salt Lake City, Utah, USA
retrac@cs.utah.edu

Studies of application behavior reveal the nested repetition of large and small program phases, with significant variation among phases in such characteristics as memory reference patterns, memory and energy usage, I/O activity, and occupancy of micro-architectural resources. In this project, we study theories and techniques for reliably predicting and exploiting phased behavior, so an advanced execution environment may allocate resources in a way that better matches program needs, or to transform programs so that their needs better match the available resources. In this paper, we present the basic components of the study and report the progress in the past half year.

An overview of the ECO project

Jacqueline Chame, Chun Chen, Pedro Diniz, Mary Hall, Yoon-ju Lee and Robert F. Lucas

Information Sciences Institute
USC
Marina Del Rey, CA, USA
{*jchame, chunchen, pedro, mhall, yoonju, rflucas*}*@isi.edu*

In this paper, we describe a compilation system that automates much of the process of performance tuning that is currently done manually by application programmers interested in high performance. Our approach combines compiler models and heuristics with guided empirical search to take advantage of their complementary strengths. The models and heuristics limit the search to a small number of candidate implementations, and the empirical results provide the most accurate information to the compiler to select among candidates and tune optimization parameter values. The overall approach can be employed to alleviate some of the performance problems that lead to inefficiencies in key applications today: register pressure, cache conflict misses, and the trade-off between synchronization, parallelism and locality in SMPs.

The main focus of the paper is an algorithm for simultaneously optimizing across multiple levels of the memory hierarchy for dense-matrix computations. We have developed an initial compiler implementation, and present automatically-generated results on matrix multiply. Results on two architectures, SGI R10000 and Sun UltraSparc IIe, outperform the native compiler, and either outperform or achieve comparable performance as the ATLAS self-tuning library and the hand-tuned vendor BLAS library.

This paper describes other components of the ECO system, including supporting tools and experiments with programmer-guided performance tuning. This approach has provided a foundation for a general framework for systematic optimization of domain-specific applications. Specifically, we are developing an optimization system for signal and image processing that exploits signal properities, and we are using machine learning and a knowledge-rich representation can be exploited to optimize molecular dynamics simulation.

Dynamic Program Phase Detection in Distributed Shared-Memory Multiprocessors

Engin Ipek[1], José F. Martínez[1], Bronis R. De Supinski[2], Sally A. Mckee[1] and Martin Schulz[2]

[1]*Computer Systems Laboratory*
Cornell University
Ithaca, NY, USA
{*engin, martinez, sam*}*@csl.cornell.edu*

[2]*Center for Advanced Scientific Computing*
Lawrence Livermore National Laboratory
Livermore, CA, USA
{*bronis, schulz*}*@llnl.gov*

We present a novel hardware mechanism for dynamic program phase detection in distributed shared memory (DSM) multiprocessors. We show that successful hardware mechanisms for phase detection in uniprocessors do not necessarily work well in DSM systems, since they lack the ability to incorporate the parallel applications global execution information and memory access behavior based on data distribution. We then propose a hardware extension to a well-known uniprocessor mechanism that significantly improves phase detection in the context of DSM multiprocessors. The resulting mechanism is modest in size and complexity, and is transparent to the parallel application.

Hierarchically Tiled Arrays for Parallelism and Locality

Jia Guo[1], Ganesh Bikshandi[1], Daniel Hoeflinger[1], Gheorghe Almasi[2], Basilio Fraguela[3], María Jesús Garzarán[1], David Padua[1] and Christoph Von Praun[2]

[1]*University of Illinois at Urbana-Champaign*
USA
{*jiaguo, bikshand, hoefling, garzaran, padua*}*@cs.uiuc.edu*

[2]*IBM T.J. Watson Research Center*
Yorktown Heights, USA
{*gheorghe, praun*}*@us.ibm.com*

[3]*Universidade da Coruña*
Spain
basilio@udc.es

Parallel programming is facilitated by constructs which, unlike the widely used SPMD paradigm, provide programmers with a global view of the code and data structures. These constructs could be compiler directives containing information about data and task distribution, language extensions specifically designed for parallel computation, or classes that encapsulate parallelism. In this paper, we describe a class developed at Illinois and its MATLAB implementation. This class can be used to conveniently express both parallelism and locality. A C++ implementation is now underway. Its characteristics will be reported in a future paper. We have implemented most of the NAS benchmarks using our HTA MATLAB extensions and found during that HTAs enable the fast prototyping of parallel algorithms and produce programs that are easy to understand and maintain.

Hierarchical Multithreading: Programming Model and System Software

Guang R. Gao[1], Thomas Sterling[2,3], Rick Stevens[4], Mark Hereld[4] and Weirong Zhu[1]

[1]*Department of Electrical and Computer Engineering*
University of Delaware
Delaware, USA
{*ggao, weirong*}*@capsl.udel.edu*

[2]*Center for Advanced Computing Research*
California Institute of Technology
California, USA
tron@cacr.caltech.edu

[3]*Department of Computer Science*
Louisiana State University
Louisiana, USA

[4]*Mathematics and Computer Science Division*
Argonne National Laboratory
Illinois, USA
{*stevens, hereld*}*@mcs.anl.gov*

This paper addresses the underlying sources of performance degradation (e.g. latency, overhead, and starvation) and the difficulties of programmer productivity (e.g. explicit locality management and scheduling, performance tuning, fragmented memory, and synchronous global barriers) to dramatically enhance the broad effectiveness of parallel processing for high end computing. We are developing a hierarchical threaded virtual machine (HTVM) that defines a dynamic, multithreaded execution model and programming model, providing an architecture abstraction for HEC system software and tools development. We are working on a prototype language, LITL-X (pronounced little-X) for Latency Intrinsic-Tolerant Language, which provides the application programmers with a powerful set of semantic constructs to organize parallel computations in a way that hides/manages latency and limits the effects of overhead. This is quite different from locality management, although the intent of both strategies is to minimize the effect of latency on the efficiency of computation. We will work on a dynamic compilation and runtime model to achieve efficient LITL-X program execution. Several adaptive optimizations will be studied. A methodology of incorporating domain-specific knowledge in program optimization will be studied. Finally, we plan to implement our method in an experimental testbed for a HEC architecture and perform a qualitative and quantitative evaluation on selected applications.

Recent Advances in Checkpoint/Recovery Systems

Greg Bronevetsky, Rohit Fernandes, Daniel Marques, Keshav Pingali and Paul Stodghill

Department of Computer Science
Cornell University
Ithaca, NY, USA
{*bronevet, rohitf, marques, pingali, stodghil*}*@cs.cornell.edu*

Checkpoint and Recovery (CPR) systems have many uses in high-performance computing. Because of this, many developers have implemented it, by hand, into their applications. One of the uses of checkpointing is to help mitigate the effects of interruptions in computational service (both planned and unplanned) In fact, some supercomputing centers expect their users to use checkpointing as a matter of policy. And yet, few centers provide fully automatic checkpointing systems for their high-end production machines.

The paper is a status report on our work on the family of C^3 systems for (almost) fully automatic checkpointing for scientific applications. To date, we have shown that our techniques can be used for checkpointing sequential, MPI and OpenMP applications written in C, Fortran, and several other languages. A novel aspect of our work is that we have not built a single checkpointing system, rather, we have developed a methodology and a set of techniques that have enabled us to develop a number of systems, each meeting different design goals and efficiency requirements.

Dynamic Aspects for Runtime Fault Determination and Recovery

Jeremy Manson, Jan Vitek and Suresh Jagannathan

Department of Computer Science
Purdue University
Indiana, USA
{*jmanson, jv, suresh*}*@cs.purdue.edu*

One of the most promising applications of Aspect Oriented Programming (AOP) is the area of fault tolerance and recovery. In traditional programming languages, error handling code must be closely interwoven with program logic. AOP allows the programmer to take a more modular approach - error handling code can be woven into the code by expressing it as an aspect. One major impediment to handling error code in this way is that while errors are a dynamic, runtime property, most research on AOP has focused on static properties. In this paper, we propose a method for handling a variety of run-time faults as dynamic aspects. First, we separate fault handling into two different notions: fault determination, or the discovery of faults within a program, and and fault recovery, or the logic used to recover from a fault. Our position is that fault determination should be expressed as dynamic aspects. We propose a system, called Rescue, that exposes underlying features of the virtual machine in order to express faults as variety of run-time constraints. We show how our methodology can be used to address several of the flaws in state of the art fault fault handling techniques. This includes their limitations in handling parallel and distributed faults, their obfuscated nature and their overly simplistic notion of what a fault actually may comprise.

An Extensible Global Address Space Framework with Decoupled Task and Data Abstractions

Sriram Krishnamoorthy[1], Umit Catalyurek[2], Jarek Nieplocha[3], Atanas Rountev[1] and P. Sadayappan[1]

[1]*Dept. of Computer Science and Engineering*
The Ohio State University
Columbus, OH, USA
{*krishnsr, rountev, saday*}*@cse.ohio-state.edu*

[2]*Dept. of Biomedical Informatics*
The Ohio State University
Columbus, OH, USA
umit@bmi.osu.edu

[3]*Computational Sciences and Mathematics*
Pacific Northwest National Laboratory
Richland, WA, USA
jarek.nieplocha@pnl.gov

Although message passing using MPI is the dominant model for parallel programming today, the significant effort required to develop high-performance MPI applications has prompted the development of several parallel programming models that are more convenient. Programming models such as Co-Array Fortran, Global Arrays, Titanium, and UPC provide a more convenient global view of the data, but face significant challenges in delivering high performance over a range of applications. It is particularly challenging to achieve high performance using global-address-space languages for unstructured applications with irregular data structures.

In this paper, we describe a global-address-space parallel programming framework with decoupled task and data abstractions. The framework centers around the use of task pools, where tasks specify operands in a distributed, globally addressable pool of data chunks. The data chunks can be addressed in a logical multidimensional tuple space, and are distributed among the nodes of the system. Locality-aware load balancing of tasks in the task pool is achieved through judicious mapping via hyper-graph partitioning, as well as dynamic task/data migration. The framework implements a transparent interface for out-of-core data, so that explicit orchestration of movement of data between disks and memory is not required of the programmer. The use of the framework for implementation of parallel block-sparse tensor computations in the context of a quantum chemistry application is illustrated.

Toward Reliable and Efficient Message Passing Software Through Formal Analysis

Ganesh Gopalakrishnan and Robert M. Kirby

University of Utah
School of Computing
Salt Lake City, UT, USA
{*ganesh, kirby*}*@cs.utah.edu*

The quest for high performance drives parallel scientific computing software design. Well over 60% of the high-performance computing (HPC) community writes programs using the MPI library; to gain performance, they are known to perform many manual optimizations. Even tools that accept high level descriptions often generate MPI code, due to its eminent portability. However, since the overall performance of a program does not usually port (due to variations in the target architecture, cluster size, etc.), manual changes to the code are inevitable in today's approaches to MPI programming and optimization. This, together with the vastness and evolving nature of the MPI standard, and the innate complexity of concurrent programming introduces costly bugs.

Our research addresses these challenges through specific efforts in the following broad areas: (i) high level expression of the parallel algorithm and compilation thereof into optimized MPI programs, (ii) optimizations of user-written detailed MPI programs through localized transformations such as barrier removal, (iii) formal modeling of complex communication standards, such as the MPI-2 standard and a facility for answering putative queries (this need arises when standard documents are impossibly difficult to manually study in order to answer questions that are not explicitly addressed in the standard), (iv) formal modeling of new (and hence relatively less well understood) features of communication libraries, such as the one-sided communication facility of MPI-2, and (v) formal modeling of intricate control algorithms in these libraries such as the progress engine for TCP and/or shared memory in MPICH2 (a formal model can explicate commonalities, help formally verify, as well as help create better future implementations). Our research gains focus through numerous collaborations.

Compiler-Assisted Software Verification Using Plug-Ins

Sean Callanan, Radu Grosu, Xiaowan Huang, Scott A. Smolka and Erez Zadok

Stony Brook University
USA
spyffe@cs.sunysb.edu, grosu@cs.sunysb.edu, xhuang@cs.sunysb.edu, sas@cs.sunysb.edu, ezk@cs.sunysb.edu

We present Protagoras, a new plug-in architecture for the GNU compiler collection that allows one to modify GCC's internal representation of the program under compilation. We illustrate the utility of Protagoras by presenting plug-ins for both compile-time and runtime software verification and monitoring. In the compile-time case, we have developed plug-ins that interpret the GIMPLE intermediate representation to verify properties statically. In the runtime case, we have developed plug-ins for GCC to perform memory leak detection, array bounds checking, and reference-count access monitoring.

An Overview of the Jahob Analysis System: Project Goals and Current Status

Viktor Kuncak and Martin Rinard

MIT Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology
Cambridge, MA, USA
{vkuncak, rinard}@csail.mit.edu

We present an overview of the Jahob system for modular analysis of data structure properties. Jahob uses a subset of Java as the implementation language and annotations with formulas in a subset of Isabelle as the specification language. It uses monadic second-order logic over trees to reason about reachability in linked data structures, the Isabelle theorem prover and Nelson-Oppen style theorem provers to reason about high-level properties and arrays, and a new technique to combine reasoning about constraints on uninterpreted function symbols with other decision procedures. It also incorporates new decision procedures for reasoning about sets with cardinality constraints. The system can infer loop invariants using new symbolic shape analysis. Initial results in the use of our system are promising; we are continuing to develop and evaluate it.

Verification of Software via Integration of Design and Implementation

Andrew S. Miner and Samik Basu

Department of Computer Science
Iowa State University
Ames, IA, USA
{*asminer, sbasu*}*@cs.iastate.edu*

Model checking is usually applied at the design phase to verify that preliminary highlevel design specifications conform to their requirements. Source code analysis, on the other hand, is used to check for correctness of implementation once it is realized from the design specifications. However, the current practice of validating a design and its implementation in isolation makes it necessary to employ rigorous testing analysis to empirically ensure that the implementation satisfies the design specification. This article describes a formal framework that allows design models to contain embedded partial implementations as components; these models are then formally analyzed to ensure that global requirements are satisfied. This framework can be utilized to incrementally develop and ensure correctness of the design and the corresponding implementation. Realization of this framework requires consolidation and expansion of traditional formal verification techniques by integration of model checking, program analysis and constraint solving.

Unification of Verification and Validation Methods for Software Systems: Progress Report and Initial Case Study Formulation

James C. Browne[1], Calvin Lin[1], Kevin Kane[1], Yoonsik Cheon[2] and Patricia Teller[2]

[1]*Department of Computer Sciences*
University of Texas at Austin
Austin, Texas, USA
{*browne, lin, kane*}*@cs.utexas.edu*

[2]*Department of Computer Science*
University of Texas at El Paso
El Paso, Texas, USA
{*cheon, pteller*}*@cs.utep.edu*

This paper presents initial research on unification of methods for verification and validation (V&V)of software systems. The synergism among methods for V&V are described. The requirements for a unification are defined. The initial steps of a case study of application of the unified approach to V&V is sketched including definition of the problem domain, the approach and some details of a property specification language. An undergraduate course introducing the unified approach to V&V is described. The relationship of this research to other efforts toward unification of V&V are discussed.

Vision for Liquid Architecture

Roger D. Chamberlain[1], Ron K. Cytron[1], Jason E. Fritts[2] and John W. Lockwood[1]

[1]*Dept. of Computer Science and Engineering*
Washington University
Saint Louis, MO, USA
{*roger, cytron, lockwood*}*@wustl.edu*

[2]*Dept. of Mathematics and Computer Science*
Saint Louis University
Saint Louis, MO, USA
jfritts@slu.edu

In the liquid architecture project, we are exploring ways in which architectural flexibility can be exploited to improve the execution properties of individual applications. Here, we report on successes we have had to date in this area, and present our vision of where this research should proceed into the future.

Statistical Sampling of Microarchitecture Simulation

Thomas F. Wenisch, Roland E. Wunderlich, Babak Falsafi and James C. Hoe

Computer Architecture Laboratory (CALCM)
Carnegie Mellon University
Pittsburgh, PA, USA
{*twenisch, rolandw, babak, jhoe*}*@ece.cmu.edu*

Current software-based microarchitecture simulators are many orders of magnitude slower than the hardware they simulate. Hence, most microarchitecture design studies draw their conclusions from drastically truncated benchmark simulations that are often inaccurate and misleading. The Sampling Microarchitecture Simulation (SMARTS) framework is an approach to enable fast and accurate performance measurements of full-length benchmarks. SMARTS accelerates simulation by selectively measuring in detail only an appropriate benchmark subset. SMARTS prescribes a statistically sound procedure for configuring a systematic sampling simulation run to achieve a desired quantifiable confidence in estimates.

Analysis of the SPEC CPU2000 benchmark suite shows that CPI can be estimated to within $\pm$ 3% with 99.7% confidence by measuring fewer than 50 million instructions per benchmark. In practice, inaccuracy in microarchitectural state initialization introduces an additional uncertainty which we empirically bound to $\sim$2% for the tested benchmarks. We present two implementations of SMARTS that both achieve an average error of only 0.64% on CPI. SMARTSim constructs accurate model state through functional warmingcontinuously warming large microarchitectural structures (e.g., caches and the branch predictor) while functionally simulating the billions of instructions between measurementsreducing average simulation turnaround from 5.5 days to 7.0 hours. TurboSMARTSim replaces functional warming with live-pointscheckpoints that store a bare minimum of functionally-warmed state for accurate simulation of a limited execution windowfurther reducing average turnaround to 91 seconds.

Designing Next Generation Data-Centers with Advanced Communication Protocols and Systems Services

P. Balaji, K. Vaidyanathan, S. Narravula, H. -w. Jin and D. K. Panda

Department of Computer Science and Engineering
The Ohio State University
, USA
{*balaji, vaidyana, narravul, jinhy, panda*}*@cse.ohio-state.edu*

Current data-centers rely on TCP/IP over Fast- and Gigabit-Ethernet for data communication even within the cluster environment for cost-effective designs, thus limiting their maximum capacity. Together with raw performance, such data-centers also lack in efficient support for intelligent services, such as requirements for caching documents, managing limited physical resources, load-balancing, controlling overload scenarios, and prioritization and QoS mechanisms, that are becoming a common requirement today. On the other hand, the System Area Network (SAN) technology is making rapid advances during the recent years. Besides high performance, these modern interconnects are providing a range of novel features and their support in hardware (e.g., RDMA, atomic operations, QoS support). In this paper, we address the capabilities of these current generation SAN technologies in addressing the limitations of existing data-centers. Specifically, we present a novel framework comprising of three layers (communication protocol support, data-center service primitives and advanced data-center services) that work together to tackle the issues associated with existing data-centers. We also present preliminary results in the various aspects of the framework, which demonstrate close to an order of magnitude performance benefits achievable by our framework as compared to existing data-centers in several cases.

I/O Conscious Algorithm Design and Systems Support for Data Analysis on Emerging Architectures

G. Buehrer[1], A. Ghoting[1], Xi Zhang[1], S. Tatikonda[1], S. Parthasarathy[1,2], T. Kurc[2] and J. Saltz[1,2]

[1]*Department of Computer Science and Engineering*
The Ohio State University
Columbus, OH, USA

[2]*Department of Biomedical Informatics*
The Ohio State University
Columbus, OH, USA

Advances in data collection and storage technologies have given rise to large dynamic data stores. In order to effectively manage and mine such stores on modern and emerging architectures, one must consider both designing effective middleware support and re-architecting algorithms, to derive performance that commensurates with technological advances. In this article, we present a top-down view of how one can achieve this goal for next generation data analysis centers. Specifically, we present a case study on frequent pattern algorithms, and show how such algorithms can be restructured to be cache, memory and I/O conscious. Furthermore, motivated by such algorithms, we present a services oriented middleware framework for the derivation of high performance on next generation architectures.

Virtual Playgrounds: Managing Virtual Resources in the Grid

K. Keahey[1,2], J. Chase[3] and I. Foster[1,2]

[1] *University of Chicago*
USA
{*keahey, foster*}*@mcs.anl.gov*

[2] *Argonne National Laboratory*
USA

[3] *Duke University*
USA
chase@cs.duke.edu

Large Grid deployments increasingly require abstractions and methods decoupling the work of resource providers and resource consumers to implement scalable management methods. We proposed the abstraction of a Virtual Workspace (VW) describing a virtual execution environment that can be made dynamically available to authorized Grid clients by using well-defined protocols. Virtual workspaces provide resources in controllable ways that are independent of how a resource is consumed. A Virtual Playground may combine many such workspaces, as well as other aspects of virtual environments, such as networking and storage, to form virtual Grids. In this paper, we report on the goals and progress of the Virtual Playground Project and put in context the research to date.

The GHS Grid Scheduling System: Implementation and Performance Comparison

Ming Wu and Xian-he Sun

Department of Computer Science
Illinois Institute of Technology
Chicago, Illinois, USA
{*wuming, sun*}*@iit.edu*

Effective task scheduling and deployment is hard to achieve in a Grid environment, where computing resources are heterogamous and shared between local and Grid users without a central control. Current scheduling systems, such as AppLeS, use NWS (Network Weather Service) for short-term estimation of resource availability and do not address the influence of the variation of resource availability in task scheduling. These inherent limitations prevent existing scheduling systems from working effectively to solve large-scale tasks in a Grid environment. Adopting APST (AppLeS Parameter Sweep Template) as the deployment environment, we have developed a task scheduling system for large-scale applications based on our recent results in performance prediction and task scheduling. Preliminary experimental results show that the newly developed system works well and is significantly more appropriate for large applications than existing systems.

On Improving Performance and Energy Profiles of Sparse Scientific Applications

Konrad Malkowski, Ingyu Lee, Padma Raghavan and Mary Jane Irwin

Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA, USA
{*malkowsk, inlee, raghavan, mji*}*@cse.psu.edu*

In many scientific applications, the majority of the execution time is spent within a few basic *sparse* kernels such as sparse matrix vector multiplication (SMV). Such sparse kernels can utilize only a fraction of the available processing speed because of their relatively large number of data accesses per floating point operation, and limited data locality and data re-use. Algorithmic changes and tuning of codes through blocking and loop unrolling schemes can improve performance but such tuned versions are typically not available in benchmark suites such as the SPEC CFP 2000. In this paper, we consider sparse SMV kernels with different levels of tuning that are representative of this application space. We emulate certain memory subsystem optimizations using SimpleScalar and Wattch to evaluate improvements in performance and energy metrics. We also characterize how such an evaluation can be affected by the interplay between code tuning and memory subsystem optimizations. Our results indicate that the optimizations reduce execution time by over 40%, and the energy by over 85%, when used with power control modes of CPUs and caches. Furthermore, the relative impact of the same set of memory subsystem optimizations can vary significantly depending on the level of code tuning. Consequently, it may be appropriate to augment traditional benchmarks by tuned kernels typical of high performance sparse scientific codes to enable comprehensive evaluations of future systems.

An Automated Approach to Improve Communication-Computation Overlap in Clusters

Lewis Fishgold, Anthony Danalis, Lori Pollock and Martin Swany

Department of Computer and Information Sciences
University of Delaware
Newark, DE, USA
{*fishgold, danalis, pollock, swany*}*@cis.udel.edu*

Applications that execute on parallel clusters face scalability concerns due to the high communication overhead that is usually associated with such environments. Modern network technologies that support Remote Direct Memory Access (RDMA) can offer true zero copy communication and reduce communication overhead by overlapping it with computation. For this approach to be effective the parallel application using the cluster must be structured in a way that enables communication computation overlapping. Unfortunately, the trade-off between maintainability and performance often leads to a structure that prevents exploiting the potential for communication computation overlapping. This paper describes a sourceto- source optimizing transformation that can be performed by an automatic (or semi-automatic) system in order to restructure MPI codes towards maximizing communication-computation overlapping.

Decentralized Runtime Analysis of Multithreaded Applications

Koushik Sen, Abhay Vardhan, Gul Agha and Grigore Rosu

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, USA
{ksen, vardhan, agha, grosu}@cs.uiuc.edu

Violations of a number of common safety properties of multithreaded programs–such as atomicity and absence of dataraces–cannot be observed by looking at the linear execution trace. We characterize a class of such properties, called *robust properties*, and define a simple but expressive epistemic logic to specify them. We then develop an efficient algorithm to automatically monitor and predict violations of robust safety properties. Our algorithm is based on capturing the causal structure of a computation through a mechanism similar to vector clock updates. The algorithm automatically synthesizes decentralized monitors to evaluate the information at each thread and to detect and predict safety violations. Based on this approach, a tool named DAME has been developed and evaluated on some simple examples.

Aligning Traces for Performance Evaluation

Todd Mytkowicz[1], Amer Diwan[1], Matthias Hauswirth[2] and Peter F. Sweeney[3]

[1] *University of Colorado at Boulder*
CO, USA
Todd.Mytkowicz@colorado.edu, diwan@cs.colorado.edu

[2] *University of Lugano*
Switzerland
Matthias.Hauswirth@unisi.ch

[3] *IBM Thomas J. Watson Research Center*
NY, USA
pfs@us.ibm.com

For many performance analysis problems, the ability to reason across traces is invaluable. However, due to non-determinism in the OS and virtual machines, even two identical runs of an application yield slightly different traces. For example, it is unlikely that two identical runs of an application will suffer context switches at exactly the same points. These sorts of variations across traces make it difficult to reason across traces. This paper describes and evaluates an algorithm, Dynamic Time Warping (DTW), that can be used to align traces, thus enabling us to reason across traces. While DTW comes from prior work our use of DTW is novel. Also we describe and evaluate an enhancement to DTW that significantly improves the quality of its alignments. Our results show that for applications whose performance varies significantly over time, DTW does a great job at aligning the traces. For applications whose performance stays largely constant for significant periods of time, the original DTW does not perform well; however, our enhanced DTW performs much better.

Model-driven Generative Techniques for Scalable Performability Analysis of Distributed Systems

Arundhati Kogekar[1], Dimple Kaul[1], Aniruddha Gokhale[1], Paul Vandal[2], Upsorn Praphamontripong[2], Swapna Gokhale[2], Jing Zhang[3], Yuehua Lin[3] and Jeffrey Gray[3]

[1]*Electrical Engineering and Computer Science*
Vanderbilt University
Nashville, Tennessee, USA
{*akogekar, dkaul, gokhale*}*@dre.vanderbilt.edu*

[2]*Computer Science and Engineering*
University of Connecticut
Storrs, Connecticut, USA
{*paul.vandal, upsorn.praphamontripong*}*@uconn.edu,*
ssg@engr.uconn.edu

[3]*Computer and Information Science*
University of Alabama at Birmingham
Birmingham, Alabama, USA
{*zhangj, liny, gray*}*@cis.uab.edu*

The ever increasing societal demand for the timely availability of newer and feature-rich but highly dependable network-centric applications imposes the need for these applications to be constructed by the composition, assembly and deployment of off-the-shelf infrastructure and domain-specific services building blocks. Service Oriented Architecture (SOA) is an emerging paradigm to build applications in this manner by defining a choreography of loosely coupled building blocks. However, current research in SOA does not yet address the performability (i.e., performance and dependability) challenges of these modern applications. Our research is developing novel mechanisms to address these challenges. We initially focus on the composition and configuration of the infrastructure hosting the individual services. We illustrate the use of domain-specific modeling languages and model weavers to model infrastructure composition using middleware building blocks, and to enhance these models with the desired performability attributes. We also demonstrate the use of generative tools that synthesize metadata from these models for performability validation using analytical, simulation and empirical benchmarking tools.

Engineering Reliability into Hybrid Systems via Rich Design Models: Recent Results and Current Directions

Somo Banerjee[1], Leslie Cheung[1], Leana Golubchik[1,2], Nenad Medvidovic[1], Roshanak Roshandel[3] and Gaurav Sukhatme[1]

[1]*Computer Science Department*
University of Southern California
Los Angeles, CA, USA
{*sbanerje, lccheung, leana, neno, gaurav*}*@usc.edu*

[2]*EE-Systems Dept, IMSC*
University of Southern California
Los Angeles, CA, USA

[3]*Dept. of Comp. Sci. & Software Engr.*
Seattle University
Seattle, WA, USA
roshanak@seattleu.edu

Software reliability techniques are aimed at reducing or eliminating failures in software systems. Reliability in software systems has traditionally been measured during or after system implementation. However, software engineering methodology lays stress on doing the "correct things" early on in the software development lifecycle in order to curb development and maintenance costs. In this paper, we argue that reliability of a software system should be assessed throughout the systems life span, starting with the software architecture level. Our research goal is to estimate the reliability of software systems in early design stages, which we believe involves the ability to reason about numerous uncertainties that exist in this stage, including uncertainty due to lack of execution artifacts. Our proposed approach is to develop techniques that will couple software architectural models with a suite of stochastic reliability estimation models and allow us to reason about these uncertainties. In this paper, we present our recent results using our technique for reliability estimation of software components at the level of software architecture. Another important part of this paper is the discussion of our ongoing research efforts and open research problems in this area.

Workshop 11

High-Performance Power-Aware Computing HPPAC 2006

11 HPPAC • Workshop on High-Performance, Power-Aware Computing

Workshop Description:

High-performance computing is and has always been performance oriented. However, a consequence of the push towards maximum performance is increased energy consumption, especially at supercomputing centers. Moreover, as peak performance is rarely attained, some of this energy consumption results in little or no performance gain. In addition, large energy consumption costs supercomputing centers a significant amount of money and wastes natural resources.

The main goal of this workshop is to provide a timely forum for the exchange and dissemination of new ideas, techniques, and research in power-aware, high-performance computing. HP-PAC will present research that reduces (1) power, (2) energy consumption, or (3) heat generation, with little or no performance penalty. This workshop differs from other power-aware workshops in that it is specifically interested in saving energy in large scale, scientific applications, rather than in small mobile devices.

Topics of interest include:

- Novel power-aware architectures for HPC
- Power-aware middleware for HPC
- Power-aware runtime systems for HPC
- Reduced power/energy/heat algorithms & applications
- Surveys or studies of power/energy/heat usage of HPC applications

Workshop Co-chairs:

Vincent W. Freeh, North Carolina State University, USA

David K. Lowenthal, University of Georgia, USA

Program Committee:

Frank Bellosa, University of Karlsruhe, Germany
Kirk Cameron, University of South Carolina, USA
Bronis de Supinski, Lawenrce Livermore National Laboratory, USA
Wu-Chun Feng, Los Alamos National Laboratory, USA
Vincent W. Freeh, North Carolina State University, USA
Soraya Ghiasi, IBM Austin Research Laboratory, USA
Dirk Grunwald, University of Colorado, USA
Chung-Hsing Hsu, Los Alamos National Laboratory, USA
Jesus Labarta, Technical University of Catalonia, Spain
David K. Lowenthal, University of Georgia, USA
Satoshi Matsuoka, Tokyo Institute, Japan
Padma Raghavan, Pennsylvania State University, USA
Fred Weber, AMD Corporation, USA

Conjugate Gradient Sparse Solvers: Performance-Power Characteristics

Konrad Malkowski, Ingyu Lee, Padma Raghavan and Mary Jane Irwin

Computer Science and Engineering
The Pennsylvania State University
University Park, PA, USA
{*malkowsk, inlee, raghavan, mji*}*@cse.psu.edu*

We characterize the performance and power attributes of the conjugate gradient (CG) sparse solver which is widely used in scientific applications. We use cycle-accurate simulations with SimpleScalar and Wattch, on a processor and memory architecture similar to the configuration of a node of the BlueGene/L. We first demonstrate that substantial power savings can be obtained without performance degradation if low power modes of caches can be utilized. We next show that if Dynamic Voltage Scaling (DVS) can be used, power and energy savings are possible, but these are realized only at the expense of performance penalties. We then consider two simple memory subsystem optimizations, namely memory and level-2 cache prefetching. We demonstrate that when DVS and low power modes of caches are used with these optimizations, performance can be improved significantly with reductions in power and energy. For example, execution time is reduced by 23%, power by 55% and energy by 65% in the final configuration at 500MHz relative to the original at 1GHz. We also use our codes and the CG NAS benchmark code to demonstrate that performance and power profiles can vary significantly depending on matrix properties and the level of code tuning. These results indicate that architectural evaluations can benefit if traditional benchmarks are augmented with codes more representative of tuned scientific applications.

Integrated Link/CPU Voltage Scaling for Reducing Energy Consumption of Parallel Sparse Matrix Applications

Seung Woo Son, Konrad Malkowski, Guilin Chen, Mahmut Kandemir and Padma Raghavan

Computer Science and Engineering
The Pennsylvania State University
University Park, PA, USA
{*sson, malkowsk, guilchen, kandemir, raghavan*}*@cse.psu.edu*

Reducing power consumption is quickly becoming a first-class optimization metric for many high-performance parallel computing platforms. One of the techniques employed by many prior proposals along this direction is voltage scaling and past research used it on different components such as networks, CPUs, and memories. In contrast to most of the existent efforts on voltage scaling that target a single component (CPU, network or memory components), this paper proposes and experimentally evaluates a voltage/frequency scaling algorithm that considers CPU and communication links in a mesh network at the same time. More specifically, it scales voltages/frequencies of both CPUs in the network and the communication links among them in a coordinated fashion (instead of one after another) such that energy savings are maximized without impacting execution time. Our experiments with several tree-based sparse matrix computations reveal that the proposed integrated voltage scaling approach is very effective in practice and brings 13% and 17% energy savings over the pure CPU and pure communication link voltage scaling schemes, respectively. The results also show that our savings are consistent with the different network sizes and different sets of voltage/frequency levels.

Profile-based Optimization of Power Performance by using Dynamic Voltage Scaling on a PC cluster

Yoshihiko Hotta[1], Mitsuhisa Sato[1], Hideaki Kimura[1], Satoshi Matsuoka[2], Taisuke Boku[1] and Daisuke Takahashi[1]

[1] 1Graduate School of Systems and Information Engineering
University of Tsukuba
Tennoudai Tsukuba Ibaraki, Japan
{hotta, msato, kimura, taisuke, daisuke}@hpcs.cs.tsukuba.ac.jp

[2] Dept. of Mathematical and Computing Sciences
Tokyo Institute of Technology
Tokyo, Japan
matsu@is.titec.ac.jp

Currently, several of the high performance processors used in a PC cluster have a DVS (Dynamic Voltage Scaling) architecture that can dynamically scale processor voltage and frequency. Adaptive scheduling of the voltage and frequency enables us to reduce power dissipation without a performance slowdown during communication and memory access. In this paper, we propose a method of profiled-based power-performance optimization by DVS scheduling in a high-performance PC cluster. We divide the program execution into several regions and select the best gear for power efficiency. Selecting the best gear is not straightforward since the overhead of DVS transition is not free. We propose an optimization algorithm to select a gear using the execution and power profile by taking the transition overhead into account. We have built and designed a power-profiling system, PowerWatch. With this system we examined the effectiveness of our optimization algorithm on two types of power-scalable clusters (Crusoe and Turion). According to the results of benchmark tests, we achieved almost 40% reduction in terms of EDP (energy-delay product) without performance impact (less than 5%) compared to results using the standard clock frequency.

Online Strategies for High-Performance Power-Aware Thread Execution on Emerging Multiprocessors

Matthew Curtis-maury, James Dzierwa, Christos D. Antonopoulos and Dimitrios S. Nikolopoulos

Department of Computer Science
College of William and Mary
Williamsburg, VA, USA
{mfcurt, jadzie, cda, dsn}@cs.wm.edu

Granularity control is an effective means for trading power consumption with performance on dense shared memory multiprocessors, such as multi-SMT and multi-CMP systems. With granularity control, the number of threads used to execute an application, or part of an application, is changed, thereby also changing the amount of work done by each active thread. In this paper, we analyze the energy/performance trade-off of varying thread granularity in parallel benchmarks written for shared memory systems. We use physical experimentation on a real multi-SMT system and a power estimation model based on the die areas of processor components and component activity factors obtained from a hardware event monitor. We also present HPPATCH, a runtime algorithm for live tuning of thread granularity, which attempts to simultaneously reduce both execution time and processor power consumption.

Dynamic Power Saving in Fat-Tree Interconnection Networks Using On/Off Links

Marina Alonso[1], Salvador Coll[2], Juan-miguel Martinez[1], Vicente Santonja[1], Pedro Lopez[1] and Jose Duato[1]

[1]*Dept. Computer Engineering*
Universidad Politecnica de Valencia
Valencia, Spain
{*malonso, jmmr, visan, plopez, jduato*}*@disca.upv.es*

[2]*Dept. Electronic Engineering*
Universidad Politecnica de Valencia
Valencia, Spain
scoll@eln.upv.es

Current trends in high-performance parallel computers show that fat-tree interconnection networks are one of the most popular topologies. The particular characteristics of this topology, that provide multiple alternative paths for each source/destination pair, make it an excellent candidate for applying power consumption reduction techniques. Such techniques are being increasingly applied in computer systems and the interconnection network is not an exception, since its contribution to the system power budget is not negligible. In this paper, we present a mechanism that dynamically switches on and off network links as a function of traffic. The mechanism is designed to guarantee network connectivity, according to the underlying routing algorithm. In this way, the default routing algorithm can be used regardless of the power saving actions taken, thus simplifying router design. Our simulation results show that significant network power consumption reductions can be obtained at no cost. Latency remains the same although the number of operating network links is dynamically adjusted.

Making a Case for a Green500 List

Sushant Sharma[1], Chung-hsing Hsu[1] and Wu-chun Feng[2]

[1]*Advanced Computing Lab*
Los Alamos National Laboratory
Los Alamos, NM, USA
{*sushant, chunghsu*}*@lanl.gov*

[2]*Department of Computer Science*
Virginia Polytechnic Institute and State University
Blacksburg, VA, USA
feng@cs.vt.edu

For decades now, the notion of "performance" has been synonymous with "speed" (as measured in FLOPS, short for floating-point operations per second). Unfortunately, this particular focus has led to the emergence of supercomputers that consume egregious amounts of electrical power and produce so much heat that extravagant cooling facilities must be constructed to ensure proper operation. In addition, the emphasis on speed as the performance metric has caused other performance metrics to be largely ignored, e.g., reliability, availability, and usability. As a consequence, all of the above has led to an extraordinary increase in the total cost of ownership (TCO) of a supercomputer.

Despite the importance of the TOP500 List, we argue that the list makes it much more difficult for the high-performance computing (HPC) community to focus on performance metrics other than speed. Therefore, to raise awareness to other performance metrics of interest, e.g., energy efficiency for improved reliability, we propose a Green500 List and discuss the potential metrics that would be used to rank supercomputing systems on such a list.

Power-Performance Efficiency of Asymmetric Multiprocessors for Multi-threaded Scientific Applications

Ryan E. Grant and Ahmad Afsahi

Electrical and Computer Engineering
Queen's University
Kingston, ON, Canada
ryan.grant@ece.queensu.ca, ahmad.afsahi@queensu.ca

Recently, under a fixed power budget, asymmetric multiprocessors (AMP) have been proposed to improve the performance of multi-threaded applications compared to symmetric multiprocessors. An AMP is a multiprocessor system in which its processors are not operating at the same frequency.

Power consumption has become an important design constraint in servers and high-performance server clusters. This paper explores the power-performance efficiency of Hyper-Threaded (HT) AMP servers, and proposes a new scheduling algorithm that can be used to reduce the overall power consumption of a server while maintaining a high level of performance. Prototyping AMPs on a commercial 4-way SMP server, we show that on average 15.6% energy savings and 6.1% slowdown for the HT-disabled case, and 7.1% energy savings and 4.8% slowdown for the HT-enabled case can be achieved across NAS and SPEC OpenMP applications.

Compiler And Runtime Support For Predictive Control Of Power And Cooling

Henry G. Dietz and William R. Dieter

Electrical and Computer Engineering Department
University of Kentucky
Lexington, KY, USA
{hankd, dieter}@engr.uky.edu

The low cost of clusters built using commodity components has made it possible for many more users to purchase their own supercomputer. However, even modest-sized clusters make significant demands on the power and cooling infrastructure. Minimizing impact of problems after they are detected is not as effective as avoiding problems altogether. This paper is about achieving the best system performance by predicting and avoiding power and cooling problems.

Although measuring power and thermal properties of a code is not trivial, the primary issue is making predictions sufficiently in advance so that they can be used to drive predictive, rather than just reactive, control at runtime. This paper presents new compiler analysis supporting interprocedural power prediction and a variety of other compiler and runtime technologies making feed-forward control feasible. The techniques apply to most computer systems, but some properties specific to clusters and parallel supercomputing are used where appropriate.

MegaProto/E: Power-Aware High-Performance Cluster with Commodity Technology

Taisuke Boku[1], Mitsuhisa Sato[1], Daisuke Takahashi[1], Hiroshi Nakashima[2], Hiroshi Nakamura[3], Satoshi Matsuoka[4] and Yoshihiko Hotta[1]

[1]*Graduate School of Systems and Information Engineering*
University of Tsukuba
Tsukuba, Ibaraki, Japan
{*taisuke, msato, daisuke*}*@cs.tsukuba.ac.jp,*
hotta@hpcs.cs.tsukuba.ac.jp

[2]*Department of Information and Computer Sciences*
Toyohashi University of Technology
Toyohashi, Aichi, Japan
nakasima@tutics.tut.ac.jp

[3]*Research Center for Advanced Science and Technology*
The University of Tokyo
Tokyo, Japan
nakamura@hal.rcast.u-tokyo.ac.jp

[4]*Global Scientific Information and Computing Center*
Tokyo Institute of Technology
Tokyo, Japan
matsu@is.titech.ac.jp

In our research project named "Mega-Scale Computing Based on Low-Power Technology and Workload Modeling", we have been developing a prototype cluster not based on ASIC or FPGA but instead only using commodity technology. Its packaging is extremely compact and dense, and its performance/power ratio is very high. Our latest prototype cluster unit named "MegaProto/E" with 16 Transmeta Efficeon processors achieves 32 GFlops of peak performance, which is 2.2-fold greater than that of the old one. The cluster unit is equipped with an independent dual network of Gigabit Ethernet, including dual 24-port switches. The maximum power consumption of the cluster unit is 320 W, which is comparable with that of today's high-end PC servers for high performance clusters. Performance evaluation using NPB kernels and HPL shows that the performance of MegaProto/E exceeds that of a dual-Xeon server in all the benchmarks, and its performance ratio ranges from 1.3 to 3.7. These results reveal that our solution of implementing a number of ultra low-power processors in compact packaging is an excellent way to achieve extremely high performance in applications with a certain degree of parallelism.

Workshop 12

Workshop on Parallel and Distributed Scientific and Engineering Computing PDSEC 2006

• Workshop on Parallel and Distributed Scientific and Engineering Computing

Workshop Description:
This workshop is to bring together computer scientists, applied mathematicians and researchers to present, discuss and exchange ideas, results, work in progress and experiences in the area of parallel and distributed computing for problems in science and engineering applications and inter-disciplinary applications.

General Co-Chairs:
Laurence T. Yang, St. Francis Xavier Univ, Canada
Gudula Rünger, Chemnitz Univ of Technology, Germany

Program Co-chairs:
Thomas Rauber, Univ of Bayreuth, Germery
Nectarios Koziris, National Technical Univ of Athens, Greece

Program Committee:
Jemal Abawajy, Deakin Univ
Hamid Arabnia, Univ of Georgia
Eric Aubanel, Univ New Brunswick
David Bader, Univ of New Mexico
Zhong-Zhi Bai, Chinese Academy Sci
Prith Banerjee, Northwestern Univ
Ioana Banicescu, Mississippi State Univ, USA
Subhash Bhalla, Univ of Aizu, Japan
Virendra Bhavsar, Univ of New Brunswick, Canada
Angelos Bilas, Univ of Crete and FORTH, Greece
Rupak Biswas, NASA, USA
Petter Bjorstad, Univ of Bergen
John Boisseau, UT Austin, USA
Anu Bourgeois, Georgia State Univ
Marian Bubak, AGH & Cyfronet
Martin Buecker, Aachen Univ of Technology, Germany
J. Carlos Cabaleiro, Universidade Santiago de Compostela, Spain
Jesus Carretero, Universidad Carlos III de Madrid, Spain
Xing Cai, Univ of Oslo, Norway
Barbarra Chapman, Univ of Houston
Vipin Chaudhary, Wayne State Univ
Ling Chen, Yangzhou Univ, China
Po-Jen Chuang, Tamkang Univ
Raphael Couturier, LIFC, Belfort
Dave Curkendall, JPL, Caltech, USA
Yuanshun Dai, Indiana Univ-Purdue
Rodrigo de Mello, Univ of Sao Paulo
Beniamino Di Martino, Second Univ of Naples, Italy
Ramon Doallo, Universidade da Coruna, Spain
Andrei Doncescu, Univ of West French Indies, France
Nahid Ehmad, PRISM, France
Tarek El-Ghazawi, George Washington Univ, USA
Len Freeman, Univ Manchester, UK
Michael Gerndt, Technical Univ of Munich, Germany
Luc Giraud, CERFACS, France
George Gravvanis, Democritus Univ of Thrace, Greece
Minyi Guo, Univ of Aizu, Japan
Anshul Gupta, IBM, USA
Yanbo Han, Chinese Academy of Sciences, China
Pao Yoh Han, Case Western Reserve
Chun-Hsi Huang, Univ Connecticut
Jung-Chang Huang, Univ of Houston
Tsung-Chuan Huang, National Sun Yat-sen Univ, Taiwan
Constantinos Ierotheou, Univ Greenwich, UK
Mehaut Jean-Francois, INRIA Grenoble, France
Wei-Min Jeng, Soochow Univ, Taiwan
Weijia Jia, CityU of Hong Kong
David Kaeli, Northeastern Univ, USA
Helen Karatza, Aristotle Univ of Thessaloniki, Greece
Daniel S. Katz, JPL, Caltech, USA
Alexey Lastovetsky, Univ College of Dublin, Ireland
Choi-Hong Lai, Univ Greenwich, UK
Chen Li, Univ of California at Irvine
Keqin Li, SUNY, USA
Lei Li, Hosei Univ, Japan
Sanli Li, Tsinghua Univ, P. R. China
Yiming Li, National Chiao Tung Univ, Taiwan
David Lilja, Univ of Minnesota, USA
Xiaola Lin, Univ of Hong Kong
Linzhang Lu, Xiamen Univ, China
Graham Megson, Univ Reading, UK
Combacau Michel, LAAS CNRS
Russ Miller, SUNY at Buffalo, USA
Jun Ni, Univ of Iowa, USA
John O'Donnell, Univ of Glasgow, UK
Manish Parashar, Rutgers Univ, USA
Tomas F. Pena, Universidade Santiago de Compostela, Spain
Bernard Philippe, IRISA, France
Janusz Pipin, National Research Council of Canada
Constantine Polychronopoulos, UIUC
Xiangzhen Qiao, Chinese Aca. Sci.
Enrique Quintana-Orti, Univ Jaime I
Jose D. P. Rolim, Univ of Geneva, Switzerland
Sartaj Sahni, Univ of Florida, USA
Ahmed Sameh, Purdue Univ, USA
Olaf Schenk, Univ of Basel, Switzerland
Stanislav G. Sedukhin, Univ of Aizu
Ruth E. Shaw, Univ New Brunswick
Hongchi Shi,Univ Missouri-Columbia
T.E. Simos, Democritus Univ of Thrace, Greece
Tony Skjellum, Mississippi State Univ
Peter Strazdins, Australian National Univ
Eric de Sturler, UIUC, USA
Sabin Tabirca, Univ College Cork, Ireland
Luciano Tarricone, Univ Lecce, Italy
David Taniar, Monash Univ, Australia
Parimala Thulasiraman, Univ of Manitoba, Canada
Ruppa Thulasiram, Univ of Manitoba
Karen Tomko, Univ of Cincinnati
Xinmin Tian, INTEL, USA
Lorenzo Verdoscia, ICAR, Italian National Research Council, Italy
Layne Watson, Virginia Tech, USA
Mateo Valero, Universidad Politecnica de Catalunya, Spain
Robert van de Geijn, UT Austin, USA
Hui Wang, Univ of Aizu, Japan
Jie Wu, Florida Atlantic Univ, USA
Bin Xiao, Hong Kong Polytechnic
Chengzhong Xu, Wayne State Univ
Zhiwei Xu, Chinese Academy of Sci.
Jingling Xue, Univ of New South Wales, Australia
Zahari Zlatev, National Environmental Research Institute, Denmark
Jun Zhang, Univ of Kentucky, USA
Weimin Zheng, Tsinghua Univ, China
Yao Zheng, Zhejiang Univ, China
Bingbing Zhou, Univ of Sydney
Wanlei Zhou, Deakin Univ, Australia
Xiaobo Zhou, Univ of Colorado, USA
Jianping Zhu, Univ of Akron, USA
Ming Zhu, Drexel Univ, USA
Hans Zima, JPL, Caltech, USA
Albert Y. Zomaya, Univ of Sydney

PDSEC Keynote: Facing the Challenges of Multicore Processor Technologies using Autonomic System Software

Dimitris Nikolopoulos

Department of Computer Science
College of William and Mary
Williamsburg, VA, USA
dsn@cs.wm.edu

Multicore processor technologies, which appear to dominate the processor design landscape, require a shift of paradigm in the development of programming models and supporting environments for scientific and engineering applications. System software for multicore processors needs to exploit fine-grain concurrent execution capabilities and cope with deep, non- uniform memory hierarchies. Software adaptation to multicore technologies needs to happen even as hardware platforms change underneath the software. Last but not least, due to the extremely high compute density of chip multiprocessing components, system software needs to increase its energy-awareness and treat energy and temperature distribution as first-class optimization targets. Unfortunately, energy awareness is most often at odds with high performance.

In the first part of this talk I will discuss some of the major challenges of software adaptation to multicore technologies and motivate the use of autonomic, self- optimizing system software, as a vehicle for both high performance portability and energy-efficient program execution. In the second part of the talk I will present ongoing research in runtime environments for dense parallel systems built from multicore and SMT components, and focus on two topics, polymorphic multithreading, and power-aware concurrency control with quality-of-service guarantees. In the same context, I will discuss enabling technologies for improved software autonomy via dynamic runtime optimization, including continuous hardware profilers, and online power-efficiency predictors.

Simulation of a Hybrid Model for Image Denoising

Ricolindo Carino[1], Ioana Banicescu[1,2], Hyeona Lim[3], Neil Williams[3] and Seongjai Kim[3]

[1]*Center for Computational Sciences*
Mississippi State University
Mississippi State, MS, U.S.A
rlc@erc.msstate.edu, ioana@cse.msstate.edu

[2]*Dept. of Computer Science and Engineering*
Mississippi State University
Mississippi State, MS, U.S.A

[3]*Dept. of Mathematics and Statistic*
Mississippi State University
Mississippi State, MS, U.S.A
{hlim, skim}@math.msstate.edu, tnw7@msstate.edu

We propose a new model for image denoising which is a hybrid of the total variation model and the Laplacian mean-curvature model. An efficient numerical procedure to compute the hybrid model is also presented. The hybrid model and its computational procedure introduce a number of parameters. As a preliminary step to the synthesis of a method for selecting optimal parameters, the hybrid model was simulated on a number of known images with synthetically added noise. The parallel simulation code was easily composed from existing serial code and a dynamic load balancing tool. The estimated parallel efficiency of the simulation is in excess of 96% on 32 processors of a general-purpose Linux cluster

Parallelisation of a Simulation Tool for Casting and Solidification Processes on Windows Platforms

Carsten Clauss, Silke Schuch, Rainer Finocchiaro, Stefan Lankes and Thomas Bemmerl

Chair for Operating Systems
RWTH Aachen University
Aachen, Germany
{*carsten, silke, rainer, lankes, bemmerl*}*@lfbs.rwth-aachen.de*

Since the beginning of computational engineering, the numerical simulation of physical processes is an essential element in the area of high performance computing. Thus, also the domain of metal foundry demands the computational simulation of casting and solidification processes. A popular software tool for this purpose has been developed by the RWP GmbH in Roetgen, Germany. This tool, named WinCast, is a complete software suite, which contains modules for pre-, main- and post-processing of simulation data sets. A core module of WinCast is TFB, which determines the chronological temperature distribution of a casting process based on a finite-element-method and a Gauss-Seidel solver. With the increasing demand for even higher precision of the simulation results on one hand, and a growing need for even larger data sets on the other hand, the parallelisation of this module became inevitable. In this paper, we present our work accomplished to parallelise the solving algorithm of this module. We have chosen an MPI based master-slave approach for compute clusters by using a self-developed MPI library for Windows platforms.

High-Performance Computing in Remotely Sensed Hyperspectral Imaging: The Pixel Purity Index Algorithm as a Case Study

Antonio Plaza, David Valencia and Javier Plaza

Department of Computer Science
University of Extremadura
Avda. de la Universidad s/n, E-10071 Cáceres, Spain
{*aplaza, davalec, jplaza*}*@unex.es*

The incorporation of last-generation sensors to airborne and satellite platforms is currently producing a nearly continual stream of high-dimensional data, and this explosion in the amount of collected information has rapidly created new processing challenges. For instance, hyperspectral imaging is a new technique in remote sensing that generates hundreds of spectral bands at different wavelength channels for the same area on the surface of the Earth. The price paid for such a wealth of spectral information available from latest-generation sensors is the enormous amounts of data that they generate. In recent years, several efforts have been directed towards the incorporation of high-performance computing (HPC) models in remote sensing missions. This paper explores three HPC-based paradigms for efficient information extraction from remote sensing data using the Pixel Purity Index (PPI) algorithm (available from the popular Kodaks Research Systems ENVI software) as a case study for algorithm optimization. The three considered approaches are: 1) Commodity cluster-based parallel computing; 2) Distributed computing using heterogeneous networks of workstations; and 3) FPGA-based hardware implementations. Combined, these parts deliver an excellent snapshot of the state-of-the-art in those areas, and offer a thoughtful perspective on the potential and emerging challenges of adapting HPC models to remote sensing problems.

Parallel Calculation of Volcanoes for Cryptographic Uses

Santi Martinez[1], Rosana Tomas[2], Concepcio Roig[1], Magda Valls[1] and Ramiro Moreno[2]

[1]*Dept. d'Informatica i Enginyeria Industrial*
Universitat de Lleida
Lleida, Spain
{*santi, roig, magda*}*@eps.udl.es*

[2]*Dept. de Matematica*
Universitat de Lleida
Lleida, Spain
{*rosana, ramiro*}*@eps.udl.es*

Elliptic curve cryptosystems are nowadays widely used in the design of many security devices. Nevertheless, since not every elliptic curve is useful for cryptographic purposes, mechanisms for providing good curves are highly needed. The generation of the volcano graph of elliptic curves can help to provide such good curves. However, this procedure turns out to be very expensive when performed sequentially. Hence, a parallel application for the calculation of such volcano graphs is proposed in this paper. In order to obtain high efficiency, a theoretical analysis is provided for obtaining an accurate granularity and for giving the appropriate number of tasks to be created. Experimental results show the benefits obtained in the speedup when executing the application in a cluster of workstations with message-passing for the generation of different volcano graphs. By the use of simulation, we study the scalability of the implementation and show that a speedup of more than 80 can be achieved in some cases.

Parallel Genetic Algorithm for SPICE Model Parameter Extraction

Yiming Li and Yen-yu Cho

Department of Communication Engineering
National Chiao Tung University
Hsinchu, Taiwan
{*ymli, yycho*}*@ymlabcad02.eic.nctu.edu.tw*

Models of simulation program with integrated circuit emphasis (SPICE) are currently playing a central role in the connection between circuit design and chip fabrication communities. An automatic model parameter extraction system that simultaneously integrates evolutionary and numerical optimization techniques for optimal characterization of very large scale integration (VLSI) devices has recently been advanced. In this paper, to accelerate the extraction process, a parallelization of the genetic algorithm (GA) for VLSI device equivalent circuit model parameter extraction is developed. The GA implemented in the extraction system is mainly parallelized with a diffusion scheme on a PC-based Linux cluster with message passing interface libraries. Parallelization of GA is governed by many factors, which affect the quality of extracted parameters and its efficiency. The diffusion GA is superior to an isolated GA, and the superiority of the diffusion GA is significant when the number of devices to be optimized is increased. Theoretical estimation and preliminary implementation show that there is an optimal number of processors with respect to the number of devices to be extracted. Benchmark results, such as speedup and efficiency including accuracy of extraction are presented and discussed for different sets of realistic multiple VLSI devices to show the robustness and efficiency of the method. We believe that the practical implementation of the parallel GA approach benefits the engineering of SPICE model parameter extraction in modern electronic industry.

Parallelization of Module Network Structure Learning and Performance Tuning on SMP

Hongshan Jiang[1], Chunrong Lai[2], Wenguang Chen[1], Yurong Chen[2], Wei Hu[2], Weimin Zheng[1] and Yimin Zhang[2]

[1]*Dept. of Computer Science*
Tsinghua University
Beijing, P.R.China
jhs03@mails.tsinghua.edu.cn, {cwg, zwm-dcs}@tsinghua.edu.cn

[2]*Intel China Research Center Ltd.*
Beijing, P.R.China
{chunrong.lai, yurong.chen, wei.hu, yimin.zhang}@intel.com

As an extension of Bayesian network, module network is an appropriate model for inferring causal network of a mass of variables from insufficient evidences. However learning such a model is still a time-consuming process. In this paper, we propose a parallel implementation of module network learning algorithm using OpenMP. We propose a static task partitioning strategy which distributes sub-search-spaces over worker threads to get the tradeoff between load-balance and software-cache-contention. To overcome performance penalties derived from shared-memory contention, we adopt several optimization techniques such as memory pre-allocation, memory alignment and static function usage. These optimizations have different patterns of influence on the sequential performance and the parallel speedup. Experiments validate the effectiveness of these optimizations. For a 2,200 nodes dataset, they enhance the parallel speedup up to 88%, together with a 2X sequential performance improvement. With resource contentions reduced, workload imbalance becomes the main hurdle to parallel scalability and the program behaviors more stable in various platforms.

Reducing Reconfiguration Time of Reconfigurable Computing Systems in Integrated Temporal Partitioning and Physical Design Framework

Farhad Mehdipour[1], Morteza Saheb Zamani[1], Hamid Reza Ahmadifar[3], Mehdi Sedighi[1] and Kazuaki Murakami[2]

[1]*IT and Computer Engineering Department*
Amirkabir University of Technology
Tehran, Iran
{mehdipur, szamani, msedighi}@ce.aut.ac.ir

[2]*Dep. of Informatics*
Kyushu University
Fukuoka, Japan
murakami@i.kyushu-u.ac.jp

[3]*Guilan University*
Rasht, Iran

In reconfigurable systems, reconfiguration latency is a very important factor impact the system performance. In this paper, a framework is proposed that integrates the temporal partitioning and physical design phases to perform a static compilation process for reconfigurable computing systems. A temporal partitioning algorithm is proposed which attempts to decrease the time of reconfiguration on a partially reconfigurable hardware. This algorithm attempts to find similar single or pair of operations between subsequent partitions. Considering similar pairs instead of single nodes brings about less complexity for routing process. By using this technique, smaller reconfiguration bit-stream is obtained, which directly decreases the reconfiguration overhead time at the run-time. A complementary algorithm attempts to increase the similarity of subsequent partitions by searching for similar pairs and using a technique called dummy node insertion. An incremental physical design process based on similar configurations produced in the partitioning stage improves the metrics over iterations.

On the Performance of Parallel Normalized Explicit Preconditioned Conjugate Gradient Type Methods

George A. Gravvanis and Konstantinos M. Giannoutakis

Department of Electrical and Computer Engineering
Democritus University of Thrace
GR 67100 Xanthi, Greece
{*ggravvan, kgiannou*}*@ee.duth.gr*

A new class of parallel normalized preconditioned conjugate gradient type methods in conjunction with normalized approximate inverses algorithms, based on normalized approximate factorization procedures, for solving sparse linear systems of irregular structure, which are derived from the finite element method of a two dimensional boundary value problem, is introduced. Parallel normalized explicit preconditioned conjugate gradient - type methods for distributed memory systems based on the block row distribution (for the vectors and the explicit approximate inverse), using Message Passing Interface (MPI) communication library, is also presented with theoretical estimates on speedups and efficiency, in order to examine the parallel behavior of these methods using normalized explicit approximate inverses as the suitable pre-conditioner. Collective communications have been utilized at the synchronization points and non blocking communications have been used, where the exchanging of messages can be overlapped with computations, where applicable. Application of the methods on a two dimensional boundary value problem is discussed and numerical results are given, concerning the parallel performance in terms of speedups and efficiency.

The General Matrix Multiply-Add Operation on 2D Torus

Ahmed Sherif Zekri and Stanislav G. Sedukhin

Graduate School of Computer Science and Engineering
The University of Aizu
Aizu Wakamatsu, Fukushima, Japan
{*d8062103, sedukhin*}*@u-aizu.ac.jp*

In this paper, the index space of the (n×n)-matrix multiply-add problem $C = C + A \cdot B$ is represented as a 3D n×n×n torus. All possible modular time-scheduling functions to activate the computation and data rolling inside the 3D torus index space are determined. To maximize efficiency when solving a single problem, we mapped the computations at the index points into the 2D n×n toroidal array processor. All optimal 2D data allocations that solve the problem in n multiply-add-roll steps are obtained. The well known Cannon's algorithm is one of the 2D resulting allocations. We used the optimal data allocations to describe all variants of the general matrix multiply-add operation (GEMM) on the 2D toroidal array processor. By controling the movement of data, the transposition operation is avoided in 75% of the GEMM variants. However, only one explicit matrix transpose is needed for the remaining 25%. Ultimately, we described four versions of the GEMM operation covering the possible layouts of the initially loaded data into the array processor.

Towards a parallel framework of grid-based numerical algorithms on DAGs

Zeyao Mo[1], Aiqing Zhang[2] and Xiaolin Cao[3]

[1]High Performance Computing Center
Institite of Applied Physics and Computational Mathematics
Beijing, P.R.China
zeyao_mo@iapcm.ac.cn

[2]High Performance Computing Center
Institite of Applied Physics and Computational Mathematics
Beijing, P.R.China
aiqing_zhang@iapcm.ac.cn

[3]High Performance Computing Center
Institite of Applied Physics and Computational Mathematics
Beijing, P.R.China
xiaolincao@iapcm.ac.cn

This paper presents a parallel framework of grid-based numerical algorithms where data dependencies between grid zones can be modeled by a directed acyclic graph (DAG). The construction of DAG for numerical algorithms for solution of partial differential equations varying from the Boltzmann transport equation to the linearly convection-dominated fluids is presented. The framework consists of three parts on how to partition, order and calculate the vertices of digraph. Numerical results using hundreds of processors on two parallel machines show the efficiencies and moderate scalability of this framework.

Efficient Parallel Implementation of a Weather Derivatives Pricing Algorithm based on the Fast Gauss Transform

Yusaku Yamamoto

Dept. of Computational Science & Engineering
Nagoya University
Nagoya Aichi, Japan
yamamoto@na.cse.nagoya-u.ac.jp

CDD weather derivatives are widely used to hedge weather risks and their fast and accurate pricing is an important problem in financial engineering. In this paper, we propose an efficient parallelization strategy of a pricing algorithm for the CDD derivatives. The algorithm uses the fast Gauss transform to compute the expected payoff of the derivative and has proved faster and more accurate than the conventional Monte Carlo method. However, speeding up the algorithm on a distributed-memory parallel computer is not straight-forward because naïve parallelization will require a large amount of inter-processor communication. Our new parallelization strategy exploits the structure of the fast Gauss transform and thereby reduces the amount of inter-processor communication considerably. Numerical experiments show that our strategy achieves up to 50% performance improvement over the naïve one on an 16-node Mac G5 cluster and can compute the price of a representative CDD derivative in 7 seconds. This speed is adequate for almost any applications.

Parallel implementation and performance characterization of MUSCLE

Xi Deng[1], Eric Li[2], Jiulong Shan[2] and Wenguang Chen[1]

[1]*Dept. of Computer Science*
Tsinghua University
Beijing, China
dengx03@mails.tsinghua.edu.cn, cwg@tsinghua.edu.cn

[2]*Intel China Research Center Ltd.*
Beijing, China
{eric.q.li, jiulong.shan}@intel.com

Multiple sequence alignment is a fundamental and very computationally intensive task in molecular biology. MUSCLE, a new algorithm for creating multiple alignments of protein sequences, achieves a highest rank in accuracy and the fastest speed compared to ClustalW as well as T-Coffee, some widely used tools in multiple sequence alignment. To further accelerate the computations, we present the parallel implementation of MUSCLE in this paper. It is decomposed into several independent modules, which are parallelized with different OpenMP paradigms. We also conduct detailed performance characterization on symmetric multiple processor systems. The experiments show that MUSCLE scales well with the increase of processors, and achieves up to 15.x speedup on 16-way shared memory multiple processor system.

Multiple Sequence Alignment by Quantum Genetic Algorithm

Layeb Abdesslem, Meshoul Souham and Batouche Mohamed

LIRE laboratory, PRAI group
University of Mentouri
Constantine, Constantine, Algeria
layeb@yahoo.fr, {meshoul, batouche}@wissal.dz

In this paper we describe a new approach for the well known problem in bioinformatics: Multiple Sequence Alignment (MSA). MSA is fundamental task as it represents an essential platform to conduct other tasks in bioinformatics such as the construction of phylogenetic trees, the structural and functional prediction of new protein sequences. Our approach merges between the classical genetic algorithm and some principles of the quantum computing like interference, measure, superposition, etc. It differs from other genetic methods of the literature by using a small population size and a less iteration required to find good quality alignments thanks to the used quantum principles: state superposition, interference, quantum mutation and quantum crossover. Another attractive feature of this method is its ability to provide an extensible platform for evaluating different objective functions. Experiments on a wide range of data sets have shown the effectiveness of the proposed approach and its ability to achieve good quality solutions comparing to those given by other popular multiple alignment programs.

Node-Disjoint Paths in Hierarchical Hypercube Networks

Ruei Yu Wu[1], Gerard J. Chang[2] and Gen Huey Chen[3]

[1]Department of Management Information Systems
Hwa Hsia Institute of Technology
Jhonghe, Taiwan
fish@inrg.csie.ntu.edu.tw

[2]Department of Mathematics
National Taiwan University
Taipei, Taiwan
gjchang@math.ntu.edu.tw

[3]Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
ghchen@csie.ntu.edu.tw

The hierarchical hypercube network is suitable for massively parallel systems. An appealing property of this network is the low number of connections per processor, which can facilitate the VLSI design and fabrication of the system. Other alluring features include symmetry and logarithmic diameter, which imply easy and fast algorithms for communication. In this paper, a maximal number of node-disjoint paths are constructed between every two distinct nodes of the hierarchical hypercube network. Their maximal length is not greater than max2^{m+1} + 2m+ 1, 2^{m+1} +m+ 4, where 2^{m+1} is the diameter.

Coordinated Checkpoint from Message Payload in Pessimistic Sender-Based Message Logging

Mehdi Aminian, Mohammad K. Akbari and Bahman Javadi

Department of Computer Eng. and Information Technology
Amirkabir University of Technology
Tehran, Tehran, Iran
{maminian, akbari, javadi}@ce.aut.ac.ir

Execution of MPI applications on Clusters and Grid deployments suffers from node and network failure that motivates the use of fault tolerant MPI implementations. Two category techniques have been introduced to make these systems fault-tolerant. The first one is checkpoint-based technique and the other one is called log-based recovery protocol. Sender-based pessimistic logging which falls in the second category is harnessing from huge amount of messages payloads which must be kept in volatile memory. In this paper we present a Coordinated Checkpoint from Message Payload (CCMP) to reduce the aforementioned overhead. The proposed method was examined by MPICH-V2, a public domain platform implementing pessimistic logging with uncoordinated checkpoint. Experimental results demonstrated the reduction of run-time for NPB benchmarks in both fault-free and faulty environments.

Tree Partition based Parallel Frequent Pattern mining on Shared Memory Systems

Dehao Chen[1], Chunrong Lai[2], Wei Hu[2], Wenguang Chen[1], Yimin Zhang[2] and Weimin Zheng[1]

[1]*Dept. of Computer Science*
Tsinghua University
Beijing, China
chendh05@mails.tsinghua.edu.cn, {cwg, zwm-dcs}@tsinghua.edu.cn

[2]*Intel China Research Center*
Intel Corporation
Beijing, China
{chunrong.lai, wei.hu, yimin.zhang}@intel.com

in this paper, we present a tree-partition algorithm for parallel mining of frequent patterns. Our work is based on FP-Growth algorithm, which is constituted of tree-building stage and mining stage. The main idea is to build only one FP-Tree in the memory, partition it into several independent parts and distribute them to different threads. A heuristic algorithm is devised to balance the workload. Our algorithm can not only alleviate the impact of locks during the tree-building stage, but also avoid the overhead that do great harm to the mining stage. We present the experiments on different kinds of datasets and compare the results with other parallel approaches. The results suggest that our approach has great advantage in efficiency, especially on certain kinds of datasets. As the number of processors increases, our parallel algorithm shows good scalability.

Workshop 13

Performance Modeling, Evaluation, and Optimization of Parallel and Distributed Systems
PMEO 2006

13 PMEO • Workshop on Performance Modeling, Evaluation, and Optimization of Parallel and Distributed Systems

Workshop Description:

The performance modeling, evaluation, and optimization of parallel, distributed, and grid systems have been an important research topic over the past years and poses challenging problems that require new tools and methods to keep up with the rapid evolution and increasing complexity of such systems. This workshop brings together scientists, engineers, practitioners, and computer users to share and exchange their experiences, discuss challenges, and report state-of-the-art and in-progress research on all aspects of performance modeling, evaluation, and optimization of parallel, distributed, and grid systems.

Topics of interest include but are not limited to:

- Predictive performance models of parallel and distributed systems
- Performance measurement and monitoring tools
- Tracing and trace analysis
- Simulation
- Analytical modeling
- Software tools for system performance and evaluation
- Automatic performance analysis
- Performance comparison
- Performance of memory and I/O interconnect
- Performance of communication networks
- Performance of mobile distributed systems
- Performance analysis and evaluation of parallel and distributed applications
- Improvement in system performance through optimization and tuning
- Case studies showing the role of evaluation in the design of systems

Workshop Co-Chairs:

M. Ould-Khaoua, University of Glasgow, U.K.
G. Min, University of Bradford, U.K.

Publicity Chair:

Mirela Sechi Moretti Annoni Notare, Barddal University, Brazil

Program Committee:

K. Al-Begain, Univ. of Glamorgan (UK)
A. Al-Dubai, Napier University (UK)
M. Ajmone-Marsan, Politechnico di Torino (Italy)
H. R. Arabnia, Univ. of Georgia (USA)
I. Awan, Univ. of Bradford (UK)
M. Baker, Univ. of Portsmouth (UK)
A. Benslimane, Avignon Univ. (France)
P. Bose, IBM T. J. Watson Research Center (USA)
A Boukerche, Univ. of North Texas (USA)
J. Bradley, Imperial College London (UK)
P. Cockshott, Univ. of Glasgow (UK)
M. Colajanni, Univ. of Modena (Italy)
K. Day, Sultan Qaboos Univ. (Oman)
S. Dharmaraja, IIT (India)
R. Fatoohi, San Jose State University
E. Gelenbe, Imperial College London (UK)
P. Harrison, Imperial College London (UK)
R. Ibbett, Univ. of Edinburgh (UK)
S. Jarvis, Univ. of Warwick (UK)
H. Karatza, Univ. of Thessaloniki (Greece)
A. Khonsari, IPM (Iran)
K. Li, State Univ. of New York at New Paltz (USA)
S. Loucif, Emirates University, (UAE)
L. M. Mackenzie, Univ. of Glasgow (UK)
M. Naimi, University of Cergy-Pontoise (France)
Y. Pan, Georgia State Univ. (USA)
D. K. Pradhan, Univ. of Bristol (UK)
M. S. M. A. Notare, Barddal University, (Brazil)
H. Sarbazi-Azad, Sharif Univ. & IPM (Iran)
N. Thomas, Univ. of Newcastle (UK)
A. Touzene, Sultan Qaboos Univ. (Oman)
M. E. Woodward, Univ. of Bradford (UK)
J. Wu, Florida Atlantic Univ. (USA)
Q. Yang, Univ. of Rhode Island (USA)
A. Zomaya, Univ. of Sydney (Australia)

PMEO Keynote: Remove the Memory Wall: From performance modeling to architecture optimization

Xian-he Sun

Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616, USA
sun@iit.edu

Data access is a known bottleneck of high performance computing (HPC). The prime sources of this bottleneck are the performance gap between the processor and memory storage and the large memory requirements of ever-hungry applications. Although advanced memory hierarchies and parallel file systems have been developed in recent years, they only provide high bandwidth for contiguous, well-formed data streams, performing poorly for accessing small, noncontiguous data. Unfortunately, many HPC applications make a large number of requests for small and noncontiguous pieces of data, as do high-level I/O libraries such as HDF-5. The problematic memory wall remains after years of study and, in fact, is becoming the most important issue of HPC. We propose a new I/O architecture for HPC. Unlike traditional I/O designs where data is stored and retrieved by request, our architecture is based on a novel Server-Push model in which a data access server proactively pushes data from a file server to the compute nodes memory or to its cache directly based on the architecture design. Simulation results show that with the new approach the cache hit rates increase well above 90% for various benchmark applications that are notorious for poor cache performance.

Performance evaluation is the driven force of the push-based model. Mechanisms of performance modeling, evaluation, and optimization are applied to data access pattern identification, prefetching algorithm design, data replacement strategy development, and architecture optimization to enable the Server-Push model. Our current success illustrates the power and unique role of performance evaluation in computing.

Performance Evaluation of Supercomputers using HPCC and IMB Benchmarks

Subhash Saini[1], Robert Ciotti[1], Brian T. N. Gunney[2], Thomas E. Spelce[2], Alice Koniges[2], Don Dossa[2], Panagiotis Adamidis[3], Rolf Rabenseifner[3], Sunil R. Tiyyagura[3], Matthias Mueller[4], and Rod Fatoohi[5]

[1] *Advanced Supercomputing Division*
NASA Ames Research Center
Moffett Field, California, USA
Subhash.Saini@nasa.gov, ciotti@nas.nasa.gov

[2] *Center for Applied Scientific Computing*
Lawrence Livermore National Laboratory
Livermore, California, USA
{*gunney, spelce1, koniges, dossa1*}*@llnl.gov*

[3] *High-Performance Computing-Center (HLRS)*
University of Stuttgart
Allmandring, Stuttgart, Germany
{*adamidis, rabenseifner, sunil*}*@hlrs.de*

[4] *ZIH*
TU Dresden
Zellescher Weg, Dresden, Germany
matthias.mueller@tu-dresden.de

[5] *Computer Engineering*
San Jose State University
San Jose, California, USA

The HPC Challenge (HPCC) benchmark suite and the Intel MPI Benchmark (IMB) are used to compare and evaluate the combined performance of processor, memory subsystem and interconnect fabric of five leading supercomputers - SGI Altix BX2, Cray X1, Cray Opteron Cluster, Dell Xeon cluster, and NEC SX-8. These five systems use five different networks (SGI NUMALINK4, Cray network, Myrinet, InfiniBand, and NEC IXS). The complete set of HPCC benchmarks are run on each of these systems. Additionally, we present Intel MPI Benchmarks (IMB) results to study the performance of 11 MPI communication functions on these systems.

Multiprocessor on Chip : Beating the Simulation Wall Through Multiobjective Design Space Exploration with Direct Execution

Riad Ben Mouhoub[1] and Omar Hamami[2]

[1]*Laboratoire Electronique et Informatique*
École Nationale Supérieure de Techniques Avances
Paris, Ile de France, FRANCE
riad.benmouhoub@ensta.fr

[2]*Laboratoire Electronique et Informatique*
École Nationale Supérieure de Techniques Avances
Paris, Ile de France, FRANCE
hammami@ensta.fr

Design space exploration of multiprocessors on chip requires both automatic performance analysis techniques and efficient multiprocessors configuration performance evaluation. Prohibitive simulation time of single multiprocessor configuration makes large design space exploration impossible without massive use of computing resources and still implementation issues are not tackled. This paper proposes a new performance evaluation methodology for multiprocessors on chip which conduct a multiobjective design space exploration through emulation. The proposed approach is validated on a 4 way multiprocessor on chip design space exploration where a 6 order of magnitude improvement have been achieved over cycle accurate simulation.

LogfP - A Model for small Messages in InfiniBand

Torsten Hoefler, Torsten Mehlan, Frank Mietke and Wolfgang Rehm

Dept. of Computer Science
Chemnitz University of Technology
Chemnitz, 09107, GERMANY
{*htor, tome, mief, rehm*}*@cs.tu-chemnitz.de*

Accurate models of parallel computation are often crucial to optimize parallel algorithms for their running time. In general the easier the model's use and the smaller the number of parameters and interdependencies among them, the more inaccuarcies are introduced by simplification. On the other hand a too complex model is unusable. We show that it is possible to derive a relatively accurate and easy model for small message performance over the InfiniBand network. This model allows the developer to gain knowledge about the inherent parallelism of a specific InfiniBand hardware and encourages him to use this parallelism efficiently. Several well known models hide this feature and some of them even penalize the use of parallelism because the model designers were not aware of new emerging architectures like InfiniBand.

A Framework to Develop Symbolic Performance Models of Parallel Applications

Sadaf R Alam and Jeffrey S Vetter

Oak Ridge National Laboratory
Oak Ridge, TN-37831, USA
{alamsr, vetter}@ornl.gov

Performance and workload modeling has numerous uses at every stage of the high-end computing lifecycle: design, integration, procurement, installation and tuning. Despite the tremendous usefulness of performance models, their construction remains largely a manual, complex, and time-consuming exercise. We propose a new approach to the model construction, called modeling assertions (MA), which borrows advantages from both the empirical and analytical modeling techniques. This strategy has many advantages over traditional methods: incremental construction of realistic performance models, straightforward model validation against empirical data, and intuitive error bounding on individual model terms. We demonstrate this new technique on the NAS parallel CG and SP benchmarks by constructing high fidelity models for the floating-point operation cost, memory requirements, and MPI message volume. These models are driven by a small number of key input parameters thereby allowing efficient design space exploration of future problem sizes and architectures.

Cost Evaluation from Specifications for BSP Programs

Virginia Niculescu

Dept. of Computer Science
Babes-Bolyai University
Cluj-Napoca, Romania
vniculescu@cs.ubbcluj.ro

BSP has shown that structured parallel programming is not only a performance win, but it is also a program construction win, especially if we add a formal method for designing. Maybe the most important advantage that BSP brings is the effective cost model that allows a good evaluation of the performance.

The paper presents a technique for cost evaluation from specifications for BSP programs. We consider parameterized specifications and processes for BSP programs, and the parameters are the number of processes, the index of the local process, and the data distribution. The possibility of counting the number of communications from postconditions, allows us to make a cost evaluation even at the early stages of the design, and so it leads us to the right decisions.

Performance analysis of Stochastic Process Algebra models using Stochastic Simulation

Jeremy T. Bradley[1], Stephen T. Gilmore[2] and Nigel Thomas[3]

[1]*Department of Computing*
Imperial College London
London, UK
jb@doc.ic.ac.uk

[2]*Laboratory for the Foundations of Computer Science*
The University of Edinburgh
Edinburgh, UK
Stephen.Gilmore@ed.ac.uk

[3]*School of Computing Science*
University of Newcastle-upon-Tyne
Newcastle-upon-Tyne, UK
Nigel.Thomas@ncl.ac.uk

We present a translation of a generic stochastic process algebra model into a form suitable for stochastic simulation. By systematically generating rate equations from a process description, we can use tools developed for chemical and biochemical reaction analysis to provide time-series output for models with state spaces of $O(10^{10000})$ and beyond. We apply these techniques to a significant case study: that of a secure electronic voting protocol.

An Adaptive Dynamic Grid-based Approach to Data Distribution Management

Azzedine Boukerche[1], Yunfeng Gu[1] and Gen Huey Chenregina Araujo[1,2]

[1]*SITE*
University of Ottawa
Ottawa, Canada
{*boukerch, yungu*}*@site.uottawa.ca,*
rba.ufscar@uol.com.br

[2]*Federal University of Sao Carlos*
Sao Carlos, Brazil

This paper presents a novel Adaptive Dynamic Grid-based Data Distribution Management (DDM) scheme, which we refer to as ADGB. The main objective of our protocol is to optimize DDM time through matching probability (MP) and federates' performance. A Distribution Rate (DR) along with MP are used as part of the ADGB method to select, throughout the simulation, from different devised advertisement schemes, the best scheme to achieve maximum gain with acceptable network traffic overhead. As opposed to previous protocols, the novelty of our ADGB scheme is its focus on improving overall performance, an important goal for DDM strategy. In this paper, we present our scheme and highlight its performance analysis.

Modelling job allocation where service duration is unknown

Nigel Thomas

School of Computing Science
University of Newcastle
Newcastle upon Tyne, UK
nigel.thomas@ncl.ac.uk

In this paper a novel job allocation scheme in distributed systems (TAG) is modelled using the Markovian process algebra PEPA. This scheme requires no prior knowledge of job size and has been shown to be more efficient than round robin and random allocation when the job size distribution is heavy tailed and the load is not high. In this paper the job size distribution is assumed to be of a phase-type and the queues are bounded. Numerical results are derived and compared with those derived from models employing random allocation and the shortest queue strategy. It is shown that TAG can perform well for a range of performance metrics.

A simulator for parallel applications with dynamically varying compute node allocation

Basile Schaeli, Sebastian Gerlach and Roger D. Hersch

School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
{*basile.schaeli, sebastian.gerlach, rd.hersch*}*@epfl.ch*

Dynamically allocating computing nodes to parallel applications is a promising technique for improving the utilization of cluster resources. We introduce the concept of dynamic efficiency which expresses the resource utilization efficiency as a function of time. We propose a simulation framework which enables predicting the dynamic efficiency of a parallel application. It relies on the DPS parallelization framework to which we add direct execution simulation capabilities. The high level flow graph description of DPS applications enables the accurate simulation of parallel applications without needing to modify the application code. Thanks to partial direct execution, simulation times and memory requirements may be reduced. In simulations under partial direct execution, the application's parallel behavior is simulated thanks to direct execution, and the duration of individual operations is obtained from a performance prediction model or from prior measurements. We verify the accuracy of our simulator by comparing the effective running time, respectively the dynamic efficiency, of parallel program executions with the running time, respectively the dynamic efficiency, predicted by the simulator. These comparisons are performed for an LU factorization application under different parallelization and dynamic node allocation strategies.

Comparison of MPI Benchmark Programs on an SGI Altix ccNUMA Shared Memory Machine

Nor Asilah Wati Abdul Hamid, Paul Coddington and Francis Vaughan

School of Computer Science
University of Adelaide
Adelaide, South Australia, Australia
{*asilah, paulc, francis*}*@cs.adelaide.edu.au*

The results produced by five different MPI benchmark programs on an SGI Altix 3700 are analyzed and compared. There are significant differences in the results for some MPI operations. We investigate the reasons for these discrepancies, which are due to differences in the measurement techniques, implementation details and default configurations of the different benchmarks. The variation in results on the Altix are generally much greater than on a distributed memory machine, due primarily to the ccNUMA architecture and the importance of cache effects, as well as some implementation details of the SGI MPI libraries.

Interconnect Performance Evaluation of SGI Altix 3700 BX2, Cray X1, Cray Opteron Cluster, and Dell PowerEdge

Rod Fatoohi[1], Subhash Saini[2] and Robert Ciotti[3]

[1]*Computer Engineering*
San Jose State University
San Jose, CA, USA
rfatoohi@email.sjsu.edu

[2]*NASA Ames Research Center*
Moffett Field, CA, USA
Subhash.Saini@nasa.gov

[3]*NASA Ames Research Center*
Moffett Field, CA, USA
ciotti@nas.nasa.gov

We study the performance of inter-process communication on four high-speed multiprocessor systems using a set of communication benchmarks. The goal is to identify certain limiting factors and bottlenecks with the interconnect of these systems as well as to compare these interconnects. We measured network bandwidth using different numbers of communicating processors and communication patterns - such as point-to-point communication, collective communication, and dense communication patterns. The four platforms are: a 512-processor SGI Altix 3700 shared-memory machine using Itanium-2 1.6 GHz processors and interconnected by SGI NUMAlink-4 switch with 3.2 GB/s bandwidth per node; a 64-processor (single-streaming) Cray X1 shared-memory machine using 800 MHz processor with 16 processors per node and 32 1.6 GB/s full duplex links; a 128-processor Cray Opteron cluster using 2 GHz AMD Opteron processors and interconnected by a Myrinet network; and a 1280-node Dell PowerEdge cluster with Intel Xeon 3.6 GHz processors interconnected by an InfiniBand network. Our results show the impact of the network bandwidth and topology on the overall performance of each interconnect.

Towards Building a Highly-Available Cluster Based Model for High Performance Computing

Azzedine Boukerche[1], Raed Al-shaikh[2] and Mirela Sechi[3]

[1]PARADISE Research Laboratory
University of Ottawa
Ottawa, Ontario, Canada
boukerch@site.uottawa.ca

[2]PARADISE Research Laboratory
University of Ottawa
Ottawa, Ontario, Canada
rshaikh@site.uottawa.ca

[3]Computer Science
Barddal University
Brasil, Brasil, Brasil
mnotare@ieee.org

In recent years, we have witnessed a growing interest in high performance computing (HPC) using a cluster of workstations. However, many challenges remain to be resolved before these systems become dependable. One of the challenges in a clustered environment is to keep system failure to the minimum level and while achieving the highest possible level of system availability. High-Availability (HA) computing attempts to avoid the problems of unexpected failures through active redundancy and preemptive measures. In this paper, we propose to build HA-clusters based model for high performance computing. Our model is based on combination of both HPC and HA concepts, we also propose to investigate further the hardware and the management layers of the HA-HPC cluster design, and the parallel-applications layer (i.e. FT-MPI implementations). In this work, we focus upon the latter layer. We discuss our model, and present our simulation experiments we have carried out to evaluate our proposed model.

Scheduling Heuristics for Efficient Broadcast Operations on Grid Environments

Luiz Angelo Barchet- Steffenel[1] and Grégory Mounie[2]

[1]LORIA, Université Nancy-2
Nancy, France
barchet@loria.fr

[2]ID-IMAG
Montbonnot St-Martin, France
mounie@imag.fr

The popularity of large-scale parallel environments like computational grids has emphasised the influence of network heterogeneity on the performance of parallel applications. Collective communication operations are especially concerned by this problem, as heterogeneity interferes directly on the performance of the communication strategies. In this paper we focus on the development of scheduling techniques to minimise the total communication time (makespan) of a broadcast operation on a grid environment. We observed that most optimisation techniques present in the literature are unable to deal with the complexity of a large network environment. In our work we propose the use of hierarchical communication levels to reduce the optimisation complexity, while keeping high performance levels. Indeed, we propose three heuristics designed to meet the requirements of a hierarchically structured grid composed of tenths of clusters, a tendency for the next years.

Performance Evaluation of Scheduling Applications with DAG Topologies on Multiclusters with Independent Local Schedulers

Ligang He, Stephen A. Jarvis, Daniel P. Spooner and Graham R. Nudd

Department of Computer Science
University of Warwick
Coventry, United Kingdom
{*liganghe, saj, dps, grn*}*@dcs.warwick.ac.uk*

Before an application modelled as a Directed Acyclic Graph (DAG) is executed on a heterogeneous system, a DAG mapping policy is often enacted. After mapping, the tasks (in the DAG-based application) to be executed at each computational resource are determined. The tasks are then sent to the corresponding resources, where they are orchestrated in the pre-designed pattern to complete the work. Most DAG mapping policies in the literature assume that each computational resource is a processing node of a single processor, i.e. the tasks mapped to a resource are to be run in sequence. Our studies demonstrate that if the resource is actually a cluster with multiple processing nodes, this assumption will cause a misperception in the tasks execution time and execution order. This will disturb the pre-designed cooperation among tasks so that the expected performance cannot be achieved. In this paper, a DAG mapping algorithm is presented for multicluster architectures. Each constituent cluster in the multicluster is shared by background workload (from other users) and has its own independent local scheduler. The multicluster DAG mapping policy is based on theoretical analysis and its performance is evaluated through extensive experimental studies. The results show that compared with conventional DAG mapping policies, the new scheme that we present can significantly improve the scheduling performance of a DAG-based application in terms of the schedule length.

On the Performance Analysis of Recursive Data Replication Scheme for File Sharing in Mobile Peer-to-Peer Devices Using the HyMIS Scheme

Constandinos X. Mavromoustakis and Helen D. Karatza

Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece, Greece
{*cmavrom, karatza*}*@csd.auth.gr*

Advances in wireless networks enable high rates interaction between mobile devices. Short-range wireless communication technologies such as wearable PCs demand low latency and reliability as the first thing for considering QoS. Mobile Peer-to-Peer devices as an autonomous system of mobile routers that are self-organized, self-configured and completely decentralized are characterized by bounded resource sharing reliability. Due to the uncertainty in available resources wireless networks could rarely host file sharing applications in a reliable manner. This paper examines the response of a gossip-based data replication scheme for reliable file sharing under specified patterns and conditions, using the Hybrid Mobile Infostation System (HyMIS). This scheme is based on the advantages of mobile Infostations. Combining the strengths of autonomic gossiping and the hybrid entirely mobile- Infostation concept, this scheme enables end to end reliability. Examination is performed for the response, the robustness and the offered reliability while examining the effectiveness of the proposed scheme for facing mobility limitations using the gossip-based selection of users.

A design environment for mobile applications

Stephen Gilmore, Valentin Haenel, Jane Hillston and Jennifer Tenzer

Laboratory for Foundations of Computer Science
The University of Edinburgh
Edinburgh, Midlothian, Scotland
{*stg, jeh*}*@inf.ed.ac.uk, valentin.haenel@gmx.de, j.n.tenzer@sms.ed.ac.uk*

In this paper we show how high-level UML models of mobile computing applications can be analysed for classical performance measures such as throughput. The approach proceeds by compiling the UML model into a representation in the formally-defined modelling language of PEPA nets. The compilation process and subsequent performance analysis based on numerical solution of a Continuous-Time Markov Chain is supported by a software tool, the Choreographer design platform. Choreographer interoperates with popular UML tools by reading and writing UML models in the XML Metadata Interchange format (XMI). Specifically we extract a PEPA net model from a UML activity diagram, analyse the PEPA net and report the results back as a modified activity diagram. We present an example use of the Choreographer design platform to investigate the throughput of activities in a UML activity diagram. The example which we model represents both physical and logical mobility. The scenario is of a PDA user on board a moving train connecting to a remote Web site and loading pages of dynamically-generated HTML content. With little overhead the modelling language allows the modeller to precisely record the mobile and immobile components of the system and to distinguish location-changing events from changes of computational state. The extractor-workbench-reflector tool chain powers the performance analysis of high-level model descriptions, returning results in the language in which they were submitted.

Efficient Broadcasting of Safety Messages in Multihop Vehicular Networks

Carla-fabiana Chiasserini[1], Rossano Gaeta[2], Michele Garetto[1], Marco Gribaudo[2] and Matteo Sereno[2]

[1]*Dipartimento di Elettronica*
Politecnico di Torino
Torino, Italy
{*chiasserini, garetto*}*@polito.it*

[2]*Dipartimento di Informatica*
Università di Torino
Torino, Italy
{*rossano, marcog, matteo*}*@di.unito.it*

We focus on a vehicular network supporting safety applications, and we present an application and a channel access mechanism for efficient multihop broadcasting. We study the performance of the proposed solution by developing an analytical framework, which provides several metrics relevant to message dissemination. Analytical results are compared with the performance obtained through *ns*.

Performance Analysis of the Reactor Pattern in Network Services

Swapna Gokhale[1], Aniruddha Gokhale[2], Jeff Gray[3], Paul Vandal[1] and Upsorn Praphamontripong[1]

[1]*Dept. of Computer Science and Engineering*
Univ. of Connecticut
Storrs, CT, USA
{*ssg, pvandal*}*@engr.uconn.edu,*
upsorn.praphamontripong@huskymail.uconn.edu

[2]*Dept. of Electrical Engineering and Computer Science*
Vanderbilt University
Nashville, TN, USA
a.gokhale@vanderbilt.edu

[3]*Dept. of Computer and Information Science*
Univ. of Alabama at Birmingham
Birmingham, AL, USA
gray@cis.uab.edu

The growing reliance on services provided by software applications places a high premium on the reliable and efficient operation of these applications. A number of these applications follow the event-driven software architecture style since this style fosters evolvability by separating event handling from event demultiplexing and dispatching functionality. The event demultiplexing capability, which appears repeatedly across a class of event-driven applications, can be codified into a reusable pattern, such as the Reactor pattern. In order to enable performance analysis of event-driven applications at design time, a model is needed that represents the event demultiplexing and handling functionality that lies at the heart of these applications. In this paper, we present a model of the Reactor pattern based on the well-established Stochastic Reward Net (SRN) modeling paradigm. We discuss how the model can be used to obtain several performance measures such as the throughput, loss probability and upper and lower bounds on the response time. We illustrate how the model can be used to obtain the performance metrics of a Virtual Private Network (VPN) service provided by a Virtual Router (VR). We validate the estimates of the performance measures obtained from the SRN model using simulation.

Performance Evaluation of an Enhanced Distributed Channel Access Protocol under Heterogeneous Traffic

Mamun I. Abu-tair and Geyong Min

Department of Computing, School of Informatics
University of Bradford
Bradford, U.K.
{*m.i.a.abu-tair, g.min*}*@brad.ac.uk*

Recently there have been considerable interests focusing on the performance evaluation of IEEE 802.11e Medium Access Control (MAC) protocols, which were proposed for supporting Quality of Services (QoS) in Wireless Local Area Networks (WLANs). Different from most existing work, this study has conducted comprehensive performance evaluation and analysis of the IEEE 802.11e Enhanced Distributed Channel Access (EDCA) protocol in the presence of heterogeneous network traffic including non-bursty Poisson, bursty ON/OFF, and self-similar traffic generated by wireless multimedia applications. The performance results on throughput, access delay and medium utilization have demonstrated that the protocol is able to achieve satisfying QoS differentiation for heterogeneous multimedia traffic. On the other hand the results have showed that IEEE 802.11e EDCA suffering from the low medium utilization due to the overhead generated by transmission collisions and back-off processes.

Performance Evaluation of Wormhole Routed Network Processor-Memory Interconnects

Taskin Kocak and Jacob Engel

School of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL, USA
{*tkocak, jengel*}*@cs.ucf.edu*

Network line cards are experiencing ever increasing line rates, random data bursts, and limited space. Hence, they are more vulnerable than other processor-memory environments, to create data transfer bottlenecks and hot-spots. Solutions to the memory bandwidth bottleneck are limited by the area available on the line card and network processor I/O pins. As a result, we propose to explore more suitable off-chip interconnect and communication mechanisms that will replace the existing systems and that will provide extraordinary high throughput. We utilize our custom-designed, event-driven, interconnect simulator to evaluate the performance of wormhole routed packet-based off-chip k-ary n-cube interconnect architectures for line cards. Our performance results show that wormhole routed k-ary n-cube based interconnect topologies significantly outperform the existing line card interconnects and they are able to sustain higher traffic loads.

On the Probability Distribution of Busy Virtual Channels

Nasser Alzeidi[1], Ahmed Khonsari[2,3], Mohamed Ould-khaoua[1] and Lewis Mackenzie[1]

[1]*Computing Science Depaertment*
University of Glasgow
Glasgow, UK
{*zeidi, mohamed, lewis*}*@dcs.gla.ac.uk*

[2]*Department of ECE*
University of Tehran
Tehran, Iran
ak@ipm.ir

[3]*School of Computer Science*
IPM
Tehran, Iran

A major issue in modelling the performance merits of interconnection network is dealing with virtual channels. Some analytical models chose not to deal with this issue at all i.e. one virtual channel per physical channel. More sophisticated models, however, relayed on a method proposed by Dally to capture the effect of arranging the physical channel into many virtual channels. In this study, we investigate the accuracy of Dallys method and propose an alternative approach to deal with virtual channels in analytical performance modelling. The new method is validated via simulation experiments and results reveal its accuracy under different traffic conditions.

A Comparative Performance Analysis of n-Cubes and Star Graphs

Abbas Eslami Kiasari[1,2] and Hamid Sarbazi-azad[1,2]

[1]*IPM School of Computer Science*
Tehran, Iran
{*kiasari, azad*}*@ipm.ir*

[2]*Dept. of Computer Engineering*
Sharif University of Technology
Tehran, Iran

Many theoretical-based comparison studies, relying on the graph theoretical viewpoints with using structural and algorithmic properties, have been conducted for the hypercube and the star graph. None of these studies, however, considered real working conditions and implementation limits. We have compared the performance of the star and hypercube networks for different message length and virtual channels and considered two implementation constraints, namely the constant bisection bandwidth and constant node pin-out. We use two accurate analytical models already proposed for the star graph and hypercube and implement the parameter changes imposed by technological implementation constraints. The comparison results reveal that the star graph has a better performance compared to the equivalent hypercube under light traffic loads while the opposite conclusion is reached for heavy traffic loads. The hypercube with more channels compared to its equivalent star graph saturates later showing that it can bear heavier traffic loads.

Software-Based Fault-Tolerant Routing Algorithm in Multi-Demensional Networks

F. Safaei[1,3], M. Rezazad[1], A. Khonsari[1,2], M. Fathy[3], M. Ould-khaoua[4] and N. Alzeidi[4]

[1]*IPM School of Computer Science*
Tehran, Iran
{*safaei, rezazad*}*@ipm.ir, ak@imp.ir*

[2]*Dept. of Electrical and Computer Engineering*
University of Tehran
Tehran, Iran

[3]*Dept. of Computer Engineering*
Iran University of Science and Technology
Tehran, Iran
mahfathy@iust.ac.ir

[4]*Dept. of Computing Science*
University of Glasgow
Glasgow, United Kingdom
{*mohamed, zeidi*}*@dcs.gla.ac.uk*

Massively parallel computing systems are being built with hundreds or thousands of components such as nodes, links, memories, and connectors. The failure of a component in such systems will not only reduce the computational power but also alter the networks topology. The Software-Based fault-tolerant routing algorithm is a popular routing to achieve fault-tolerance capability in networks. This algorithm is initially proposed only for two dimensional networks. Since, higher dimensional networks have been widely employed in many contemporary massively parallel systems; this paper proposes an approach to extend this routing scheme to these indispensable higher dimensional networks. Deadlock and livelock freedom and the performance of presented algorithm, have been investigated for networks with different dimensionality and various fault regions. Furthermore, performance results have been presented through simulation experiments.

A Systematic Multi-step Methodology for Performance Analysis of Communication Traces of Distributed Applications based on Hierarchical Clustering

Gaby Aguilera[1], Patricia J. Teller[1], Michela Taufer[1] and Felix Wolf[2]

[1]*Computer Science*
University of Texas - El Paso
El Paso, TX, USA
{*maguilera, pteller, mtaufer*}*@utep.edu*

[2]*Forschungszentrum Jülich*
Jülich, Germany
f.wolf@fz-juelich.de

Often parallel scientific applications are instrumented and traces are collected and analyzed to identify processes with performance problems or operations that cause delays in program execution. The execution of instrumented codes may generate large amounts of performance data, and the collection, storage, and analysis of such traces are time and space demanding. To address this problem, this paper presents an efficient, systematic, multi-step methodology, based on hierarchical clustering, for analysis of communication traces of parallel scientific applications. The methodology is used to discover potential communication performance problems of three applications: TRACE, REMO, and SWEEP3D.

TPCC-UVa: An Open-Source TPC-C Implementation for Parallel and Distributed Systems

Diego R. Llanos and Belén Palop

Departamento de Informática
Universidad de Valladolid
Valladolid, Spain
{*diego, b.palop*}*@infor.uva.es*

This paper presents TPCC-UVa, an open-source implementation of the TPC-C benchmark intended to be used in parallel and distributed systems. TPCC-UVa is written entirely in C language and it uses the PostgreSQL database engine. This implementation includes all the functionalities described by the TPC-C standard specification for the measurement of both uni- and multiprocessor systems performance. The major characteristics of the TPC-C specification are discussed, together with a description of the TPCC-UVa implementation and architecture and real examples of performance measurements.

An Entropy-Based Algorithm for Time-Driven Software Instrumentation in Parallel Systems

Ahmet Özmen

Dept. of Electrical and Electronics Engineering
Dumlupinar University
Kutahya, Turkey
ozmen@dumlupinar.edu.tr

While monitoring, instrumented long running parallel applications generate huge amount of instrumentation data. Processing and storing this data incurs overhead, and perturbs the execution. Techniques that eliminates unnecessary instrumentation data and lower the intrusion without loosing any performance information is valuable to tool developers. This paper presents a new algorithm for software instrumentation to measure the amount of information content of instrumentation data to be collected. The algorithm is based on entropy concept introduced in information theory, and it makes selective data collection for a time-driven software monitoring system possible.

Analytical Performance Modelling of Partially Adaptive Routing in Hypercubes

Ahmad Patooghy[1,2] and Hamid Sarbazi-azad[1,2]

[1]*IPM school of computer science*
Tehran, Iran
{patooghy, azad}@ipm.ir

[2]*Sharif University of Technology*
Tehran, Iran

Although several analytical models have been proposed in the literature for different interconnection networks with different routing algorithms, there is only one work dealing with partially adaptive routing algorithms. This paper proposes an accurate analytical model to predict message latency in wormhole-routed hypercube based networks using the partially adaptive routing algorithm. The results obtained from simulation experiments confirm that the proposed model exhibits a good accuracy for various network sizes and under different operating conditions.

Approximated Tensor Sum Preconditioner for Stochastic Automata Networks

Abderezak Touzene

Computer Science
Sultan Qaboos University
Al Khod, Muscat, OMAN
touzene@squ.edu.om

Some iterative and projection methods for SAN have been tested with a modest success. Several preconditioners for SAN have been developed to speedup the convergence rate. Recently Langville and Stewart proposed the Nearest Kronecker Product (NKP) preconditioner for SAN with a great success. Encouraged by their work, we propose a new preconditioning method, called Approximated Tensor Sum Preconditioner (ATSP), which uses tensor sum preconditioner rather than Kronecker product preconditioner. In ATSP, we take into account the effect of the synchronizations using an approximation technique. Our preconditioner outperforms the NKP preconditioner for the tested SAN Model.

Using Stochastic Petri Nets for Performance Modelling of Application Servers

Fábio N. Souza, Roberto D. Arteiro, Nelson S. Rosa and Paulo R. M. Maciel

Centro de Informática
Universidade Federal de Pernambuco
Recife, PE, Brasil
{fns, rda, nsr, prmm}@cin.ufpe.br

Application servers have been widely adopted as distributed infrastructure (or middleware) for developing distributed systems. Current approaches for performance evaluation of application servers have mainly concentrated on the adoption of measurement techniques. This paper, however, focuses on the use of simulation techniques and presents an approach for performance modelling and evaluation of application servers using Petri nets. In order to illustrate how the proposed approach may be applied, Petri net models of JBoss application server are presented and their performance results are compared with ones that have been measured.

Workshop 14

High-Performance Grid Computing Workshop HPGC 2006

• Workshop on High Performance Grid Computing

Workshop Description:

Grids are becoming ubiquitous, and are now incorporating a wireless dimension. Grids provide enormous computational potential for science, engineering, medicine, finance, and entertainment. The High Performance Grid Computing workshop provides a forum for presenting research results on most aspects of grid computing, with a focus on performance, in the following areas: Applications, Benchmarking, Infrastructure, Management and Scheduling, Partitioning and Load Balancing, and Programming Models.

Topics of interest include but are not limited to:

- Applications: Theory and practice of composing grid applications consisting of multiple interacting tasks.
- Benchmarking: Grid measurement technology for evaluating performance of grid hardware and middleware; benchmark results.
- Infrastructure: Implementation and evaluation of computational grid middleware.
- Management and Scheduling: Management, monitoring, resource allocation, scheduling, and metascheduling.
- Partitioning and Load Balancing: Partitioning applications for computational grids for achieving high performance, and load balancing of grid applications.
- Programming Models: Methods for remote execution and intertask communications.

Workshop Organizers:

Eric Aubanel, University of New Brunswick, Canada

Virendra C. Bhavsar, University of New Brunswick, Canada

Michael Frumkin, Intel Corporation, USA

Program Committee:

Akshai Aggarwal, University of Windsor, Windsor, ON, Canada
Rupak Biswas, Nasa Ames Research, Moffett Field, CA, USA
Henri Casanova, San Diego Supercomputing Center, CA, USA
Nikos P. Chrisochoides, College of William and Mary, Wlliamsburg, VA, USA
Anthony T. Chronopoulos, Univ. of Texas at San Antonio, USA
Weichang Du, University of New Brunswick, Canada
S.S. Iyengar, Louisianna State University, Baton Rouge, LA, USA
George Karypis, University of Minnesota, Minneapolis, MN, USA
Thuy T. Le, San Jose State University, USA
Gabriel Mateescu, National Research Council, Ottawa, Canada
Rodrigo Fernandes de Mello, University of São Paulo
Francois Pellegrini, INRIA and LaBRI, Universite Bordeaux, France
Sushil Prasad, Georgia State University, USA
Thierry Priol, IRISA, France
Andrew Rau-Chaplin, Dalhousie University, Canada
Thomas Rauber, University of Bayreuth, Germany
Ruth Shaw, University of New Brunswick, Canada
Simon Chong-Wee See, Sun Asia Pacific Science and Technology Center and Nanyang Technological University
Allan Snavely, San Diego Supercomputing Center, La Jolla, CA, USA
Laurence Tianruo Yang, St. Francis Xavier University, Antigonish, NS, Canada
Rob F. Van Der Wijngaart, NASA Ames, USA

HPGC Keynote: Major Grid Projects Around the World

Wolfgang Gentzsch

D-Grid Initiative
Munich, Germany
wgentzsch@d-grid.de

This talk will present and compare several major grid projects, with a focus on measuring and achieving performance of applications on grids. For these purposes we will investigate and compare the areas of applications, infrastructure, management, scheduling, load balancing, and benchmarking, for the Teragrid project and the North Carolina Statewide Grid effort in the US, Naregi in Japan, and EGEE and the German D-Grid project in Europe.

Multisite Co-allocation Algorithms for Computational Grid

Weizhe Zhang[1], Albert M. K. Cheng[2] and Mingzeng Hu[1]

[1]*School of Computer Science and Technology*
Harbin Institute of Technology
Harbin, Heilongjiang, P.R.China
zwz@pact518.hit.edu.cn, mzhu@hit.edu.cn

[2]*Department of Computer Science*
University of Houston
Houston, TX, USA
cheng@cs.uh.edu

Efficient multisite job scheduling facilitates the cooperation of multi-domain massively parallel processor systems in a computing grid environment. However, co-allocation, heterogeneity, adaptability, and scalability emerge as tough challenges for the design of multisite job scheduling models and algorithms. This paper presents a new multisite job scheduling schema based on the multisite job scheduling model and the performance model for a heterogeneous grid environment. There are three key components: resource selection, reservation, and backfilling. The optimal and greedy-heuristic adaptive resource selection strategies are introduced. The conservative and easy backfilling are incorporated into the backfilling procedure. Experiments indicate that the scheduler and the algorithm are effective and perform better than a non-adaptive algorithm.

Price-based User-optimal Job Allocation Scheme for Grid Systems

Satish Penmatsa and Anthony T. Chronopoulos

Dept. of Computer Science
The University of Texas at San Antonio
6900 N Loop, 1604 W, San Antonio, TX 78249, USA
{*spenmats, atc*}*@cs.utsa.edu*

In this paper we propose a price-based user-optimal job allocation scheme for grid systems whose nodes are connected by a communication network. The job allocation problem is formulated as a noncooperative game among the users who try to minimize the expected cost of their own jobs. We use the concept of Nash equilibrium as the solution of our noncooperative game and derive a distributed algorithm for computing it. The prices that the grid users has to pay for using the computing resources owned by different resource owners are obtained using a pricing model based on a game theory framework. Finally, our scheme is compared with a system-optimal job allocation scheme under simulations with various system loads and configurations and conclusions are drawn.

An Evaluation of Heuristics for SLA Based Parallel Job Scheduling

Viktor Yarmolenko and Rizos Sakellariou

School of Computer Science
The University of Manchester
Manchester, UK
{*viktor.yarmolenko, rizos.sakellariou*}*@manchester.ac.uk*

In the context of SLA based job scheduling for high performance grid computing, this paper investigates the behaviour of various scheduling heuristics to schedule SLA-bounded jobs onto a parallel computing resource. The key objective of this investigation is to evaluate the effectiveness of simple scheduling heuristics using as criteria the maximization of resource utilization (both in terms of time and SLAs serviced) and income. Our results suggest how each SLA constraint ought to be prioritized in order to improve the income.

Speeding up NGB with Distributed File Streaming Framework

Bingchen Li[1], Kang Chen[1], Zhiteng Huang[1], Hrabri L. Rajic[2] and Robert H. Kuhn[2]

[1]*Intel China Research Center Ltd.*
Beijing, China
{*bingchen.li, kang.chen, zhiteng.huang*}*@intel.com*

[2]*KSL, Software Products Division, Intel*
Champaign, Il, USA
hrabri.rajic@intel.com, bob.kuhn@intelcom

Grid computing provides a very rich environment for scientific calculations. In addition to the challenges it provides, it also offers new opportunities for optimization. In this paper we have utilized DFS (Distributed File Streaming) framework to speed up NAS Grid Benchmark workflows. By studying I/O patterns of NGB codes we have identified program locations where it is possible to overlap computation and data workflow phases. By integrating DFS into NGB, we demonstrate a useful method of improving overall workflow efficiency by streaming the output of the current process to make an input of the following stage, reducing a workflow to a series of distributed producer consumer stages. DFS framework eliminates file transfers and in the process makes process scheduling more efficient, leading to overall performance improvements in the turnaround time for HC (Helical Chain) data flow graph under Globus grid environment with the embedded DFS over the original version of the benchmark.

Anticipated Distributed Task Scheduling for Grid Environments

Thomas Rauber[1] and Gudula Rünger[2]

[1]*Computer Science Department*
University Bayreuth
Bayreuth, Germany
rauber@uni-bayreuth.de

[2]*Computer Science Department*
Chemnitz University of Technology
Chemnitz, Germany
ruenger@informatik.tu-chemnitz.de

Heterogeneous distributed environments or grid environments provide large computing resources for the execution of large scientific applications. The effective use of those platforms requires a suitable representation of the application algorithm which makes a distribution of parts of the application across the distributed environment possible. A representation of an application algorithm in form of interacting tasks has been shown to be a suitable programming model for those distributed environments, where tasks can be shipped to remote computing resources for execution. The efficient execution of an application also depends on the time for sending tasks and data to remote resoures, which adds an additional overhead to the distributed execution time. In this paper, we propose a method to overlap the execution of current tasks with the shipping time for tasks to be executed later. The efficient overlapping is achieved by an anticipated scheduling algorithm for the placement of future task executions.

Loosely-coupled Loop Scheduling in Computational Grid

José Herrera[1], Eduardo Huedo[2], Rubén Santiago Montero[1] and Ignacio Martín Llorente[1,2]

[1]*Departamento de Arquitectura de Computadores y Automática*
Universidad Complutense de Madrid
Madrid, Spain
jherrera@fdi.ucm.es, {rubensm, llorente}@dacya.ucm.es

[2]*Laboratorio de Computación Avanzada Simulacin y Aplicaciones Telematicas*
Centro de Astrobiologia (CSIC-INTA)
Torrejón, Madrid, Spain
huedoce@inta.es

Loop distribution is one of the most useful techniques to reduce the execution time of parallel applications. Traditionally, loop scheduling algorithms are implemented based on parallel programming paradigms such as MPI. This approximation presents three main disadvantages when applied in a Grid environment, namely: (i) all resources must be simultaneously allocated to begin execution of the application; (ii) it is necessary to restart the whole application when a resource fails; (iii) it is not possible to add new resources to a currently running application. To overcome these limitations, we propose a new approach to implement loop distribution schemes in computational Grids. This approach is implemented using the Distributed Resource Management Application API (DRMAA) standard and the GridWay metascheduling framework. The efficiency of this approach to solve the Mandelbrot set problem is analyzed in a Globus-based research testbed.

Execution and Composition of E-Science Applications using the WS-Resource Construct

Evangelos Floros and Yannis Cotronis

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Athens, GREECE
{floros, cotronis}@di.uoa.gr

Service Oriented Architectures are emerging as the recommended paradigm for developing dispersed e-science environments. In this paper we analyze the characteristics and requirements of a common class of scientific applications, namely Computational Simulation Models, and define a generic service-oriented framework for their execution and composition. Finally we present the work done so far towards the implementation of such framework based on the WSRF set of specifications and the Globus Toolkit.

A Job Monitoring System for the LCG Computing Grid

Ahmad Hammad[1], Torsten Harenberg[2], Dimitri Igdalov[1], Peter Mättig[2], David Meder[2] and Peer Ueberholz[1]

[1] *Department of Computer Science*
Niederrhein University of Applied Sciences
47805 Krefeld, Germany
{*haah0001, igdi0001, peer.ueberholz*}*@hsnr.de*

[2] *Department of Physics*
University of Wuppertal
42097 Wuppertal, Germany
{*harenberg, maettig, meder*}*@physik.uni-wuppertal.de*

Experience with generating simulation data of high energy physics experiments has shown that a job monitoring system (JMS) is essential to understand failures of jobs within the Grid. Such a system can give information about the status of the user job as well as the worker node in parallel while a user job is running. It should support the user directly by allowing the user to interact with the running job and should be able to make an automatic error correction. Furthermore, such a system can be extended for an automatic classification of errors which can improve the stability and performance of the Grid environment. To increase the acceptance of the Grid, a graphical user interface (GUI) has been developed and integrated with the job monitoring system. Both components are currently integrated in the computing environment for generating data for the DO Experiment. In this paper we want to describe the basic components of the job monitoring software.

SmartNetSolve: High-Level Programming System for High Performance Grid Computing

Thomas Brady[1], Eugene Konstantinov[2] and Alexey Lastovetsky[1]

[1] *School of Computer Science and Informatics*
University College Dublin
Belfield, Dublin 4, Ireland
thomas.brady@ucd.ie , alexey.lastovetsky@ucd.ie

[2] *IBM Ireland*
IDA Business Park, Ballycoolin Industrial Estate
Blanchardstown, Dublin, Ireland
ekonstan@ie.ibm.com

The paper presents SmartNetSolve, an extension of NetSolve, the programming system for high performance Grid computing. The extension is aimed at higher performance of Grid applications by improving the mapping of remote tasks and allowing them to communicate directly. To achieve more optimal mapping SmartNetSolve allows a group of tasks to be scheduled collectively, meanwhile NetSolve only allows for individual and independent mapping of remote tasks. SmartNetSolve also extends the communication model of the application by allowing remote tasks to communicate directly. The paper presents the overall design of the SmartNetSolve programming system with particular focus on its motivation and the underlying execution and communication models.

IMAGE: An approach to building standards-based enterprise Grids

Gabriel Mateescu[1] and Masha Sosonkina[2]

[1] *Research Computing Support Group*
National Research Council
Ottawa, Ontario, Canada
gabriel.mateescu@nrc.gc.ca

[2] *Scalable Computing Laboratory*
USDOE Ames Laboratory
Ames, Iowa, USA
masha@scl.ameslab.gov

We describe a system for aggregating heterogeneous resources from distinct administrative domains into an enterprise-wide compute grid, such that the aggregated resource provides the services of reliable and flexible queuing, scheduling, execution, and monitoring of batch applications. The system provides scheduling across multiple cluster Grids, user account mapping across domains, and file staging, thereby enabling the consolidation of organization-wide distributed resources into a virtual resource, while preserving local control of resources. The concept of abstract queue, as the unit of aggregating heterogeneous resources, is introduced and instantiated for distributed resource scheduling. The proposed system is an open source, standards-based alternative to similar commercial systems.

Workshop 15

Dependable Parallel, Distributed and Network-Centric Systems DPDNS 2006

15 DPDNS • Workshop on Dependable Parallel, Distributed and Network-Centric Systems

Workshop Description:

Increasingly large and complex parallel, distributed and network-centric computing systems provide unique challenges to the researchers in dependable computing, especially because of the high failure rates intrinsic to these systems. The goal of this workshop in continuation of the FTPDS (Fault-Tolerant Parallel and Distributed Systems) workshop series is to provide a forum for researchers and practitioners to discuss all aspects of dependability including reliability, availability, safety and security for parallel, distributed and network-centric systems. All aspects of design, theory and realization are of interest.

Topics of interest include but are not limited to:

- Dependable parallel, distributed and network-centric systems
- High availability in parallel, distributed and network-centric computing systems
- Safety and security in distributed and network-centric computing systems
- Dependable high-speed wide, local, and system area networks
- Dependable mobile computing
- Dependable clusters of workstations and PCs
- Dependable internet servers
- Dependability in distributed embedded systems
- Using COTS for designing dependable network-centric computing systems
- Dependable protocols for distributed and network-centric systems
- Protocol verification and validation
- Practical experiences and prototypes
- Dependability evaluation of parallel, distributed and network-centric systems
- Dependable quantum computing
- Dependable organic computing
- Dependable biocomputing

Program Chair:

Karl-Erwin Grosspietsch, Fraunhofer AIS, Germany

Steering Committee Chair:

D. Avresky

Program Committee:

B. Ciciani, University of Roma, Italy
G. Deconinck, University of Leuven, Belgium
A. Doering, IBM Research Zurich, Switzerland
R. Ekwall, EPFL, Switzerland
O. Frieder, IIT, Chicago, USA
K. Kanoun, LAAS-CNRS, France
C. Katsinis, Drexel University, USA
R. Khazan, MIT Lincoln Lab, USA
T. Kikuno, Osaka University, Japan
L. Lipsky, University of Connecticut, USA
M. Malek, Humboldt University, Germany
E. Nett, University of Magdeburg, Germany
D. Nikolos, University of Patras, Greece
M. Roy, LAAS-CNRS, France
J. G. Silva, University of Coimbra, Portugal

Scalable Resilience – The ReSIST Network of Excellence

Jean-claude Laprie

LAAS-CNRS
Toulouse, France
laprie@laas.fr

ReSIST is a Network of Excellence that integrates leading researchers active in the multidisciplinary domains of Dependability, Security, and Human Factors, in order that Europe will have a well-focused coherent set of research activities aimed at ensuring that future ubiquitous computing systems (the immense systems of ever-evolving networks of computers and mobile devices which are needed to support and provide Ambient Intelligence), have the necessary resilience and survivability, despite any residual development and physical faults, interaction mistakes, or malicious attacks and disruptions.

At the heart of ReSIST is the Joint Programme of Research (JPR). Two main steps will take place, according to the structuring of the research activities:
1) first according to the basic resilience building technologies for the survivability of information infrastructures, i.e., resilience design, resilience verification and resilience evaluation
2) then according to the resilience scaling technologies: evolvability, assessability, usability and diversity. This move from resilience building technologies towards resilience scaling technologies will be accompanied and facilitated by the resilience integration technologies: a resilience knowledge base, and the development of a resilience-explicit computing approach.

The Joint Programme of Excellence Spreading (JPES) contributes to integration via the production of documents incorporating results from the JPR, e.g., a) common courseware for training activities, and b) best practices for dissemination activities.

The Joint Steering Programme
a) guides integration in assigning and updating the activities of the JPR and the JPES,
b) favours integration via the allocation of the resources of ReSIST, and
c) assesses integration.

Construction of Efficient OR-based Deletion-tolerant Coding Schemes

Peter Sobe and Kathrin Peter

Institute of Computer Engineering
Univ. of Luebeck
Luebeck, Germany
sobe@iti.uni-luebeck.de, kathrin.peter@zib.de

Fault–tolerant data layouts for storage systems are based on the principle to add redundancy to groups of data blocks and store them in different fault regions. Commonly, XOR-based codes are used with an optimal redundancy overhead but with the disadvantage of relatively high calculation costs. We present a scheme that encodes input data in a highly redundant code and exploits that redundancy for a fault tolerance scheme. It allows to recalculate missed bits in fewer steps than needed for XOR-based schemes. This simple and efficient en- and decoding requires an appropriate hardware architecture or a highly parallel microprocessor architecture. Particularly, disjunctions over many input bits must be calculated, e.g. by wide OR-gates or busses that are driven by multiple logic input lines. The high redundant encoding is combined with data compression for separated data streams, each stream dedicated to a storage device. The compression not only eliminates the introduced redundancy of the used code, it also eliminates redundancy in the input data.

Analysis of Checksum-Based Execution Schemes for Pipelined Processors

Bernhard Fechner

Computer Science
FernUniversität in Hagen
Hagen, NRW, Germany
Bernhard.Fechner@fernuni-hagen.de

The performance requirements for contemporary microprocessors are increasing as rapidly as their number of applications grows. By accelerating the clock, performance can be gained easily but only with high additional power consumption. The electrical potential between logic 0 and 1 is decreased as integration and clock rates grow, leading to a higher susceptibility for transient faults, caused e.g. by power fluctuations or Single Event Upsets (SEUs). We introduce a technique which is based on the well-known cyclic redundancy check codes (CRCs) to secure the pipelined execution of common microprocessors against transient faults. This is done by computing signatures over the control signals of each pipeline stage including dynamic out-of-order scheduling. To correctly compute the checksums, we resolve the timedependency of instructions in the pipeline. We will first discuss important physical properties of Single Event Upsets (SEUs). Then we present a model of a simple processor with the applied scheme as an example. The scheme is extended to support n-way simultaneous multithreaded systems, resulting in two basic schemes. A cost analysis of the proposed SEU-detection schemes leads to the conclusion that both schemes are applicable at reasonable costs for pipelines with 5 to 10 stages and maximal 4 hardware threads. A worst-case simulation using software fault-injection of transient faults in the processor model showed that errors can be detected with an average of 83% even at a fault rate of 10^{-2}. Furthermore, the scheme is able to detect an error within an average of only 5.05 cycles.

Web Server Protection by Customized Instruction Set

Bernhard Fechner, Jörg Keller and Andreas Wohlfeld

FB Informatik, LG Parallelität und VLSI
FernUniversität
Hagen, Germany
{*bernhard.fechner, joerg.keller, andreas.wohlfeld*}*@fernuni-hagen.de*

We present a novel technique to secure the execution of a processor against the execution of malicious code (trojans, viruses). The main idea is to permute parts of the opcode values so that it gets a different semantic meaning. A virus which does not know the permutation is not able to execute and will cause a failure such as segmentation violation, whereby the execution of malicious code is prevented. The permutation is realized by a lookup table. We develop several variants that require only small changes to microprocessors. We sketch how to bootstrap a system such that all intended applications (including operating system) are reversely permuted, and can execute as intended. While this will be cumbersome for typical personal computers, it will work for web servers, because the number of applications and frequency of installation is lower. Furthermore, web servers are particularly endangered: they cannot be protected as good as personal computers, because by the very nature of their duty they are more openly connected with the internet than any other computer in an organization's network.

Evaluating a Clock Synchronization for Dependable Sensor Networks

Spiro Trikaliotis and Georg Lukas

Institute for Distributed Systems
University of Magdeburg
Magdeburg, Germany
{*spiro, glukas*}*@ivs.cs.uni-magdeburg.de*

A synchronized clock is an important prerequisite for many distributed algorithms. This clock is used to give an "occurred before" relationship, as well as for synchronizing distributed actions. There are many clock synchronization algorithms with varying precisions and assumptions on the underlying network topology. In this paper, a synchronization protocol is presented which achieves a high precision in the order of $20\mu s$ to $30\mu s$ in a one-hop wireless environment, and a multiple of this value for multi-hop wireless networks, such as sensor networks. The protocol works reliably even if message losses occur, which is very likely in wireless networks. For this, it utilizes redundancy in the sent time information. This protocol is implemented and evaluated on standard PC hardware running RT-Linux/Free, and an outline of the extension for multi-hop scenarios is given.

Power-Dependable Transactions in Mobile Networks

Ami Marowka[1] and David Semé[2]

[1]*Software Engineering Department*
Shenkar College of Engineering and Design
Ramat Gan, Israel
amimar2@yahoo.com

[2]*LaRIA : Laboratoire de Recherche en Informatique d*
Université de Picardie Jules Verne
Amiens, France
David.Seme@u-picardie.fr

We define a Quality-of-Power-Service (QoPS) metric to evaluate the efficiency of power-aware routing protocols in wireless ad-hoc networks. The aim of power management of routing protocols is to prolong the life-time of individual nodes in wireless network and thus to increase the delivery rate of Unicast transactions.

QoPS metric is applied to different location-based Unicast transaction protocols. The results confirm that powerrelative distribution of data streams in multi-paths Unicast transaction protocols consume substantially less energy from individual nodes than from other distribution methods. The locality distribution phenomenon discovered by the simulations explains, on the one hand, the long lifetime of large, dense, and highly degree wireless networks, and on the other hand, the short lifetime of small, sparse, and low degree networks.

Power Consumption Comparison for Regular Wireless Topologies using Fault-Tolerant Beacon Vector Routing

Luke Demoracski and Dimiter R. Avresky

Network Computing Laboratory
Northeastern University
Boston, MA, USA
lukedemo@yahoo.com, avresky@ece.neu.edu

Fault-tolerant Beacon Vector Routing (FBVR) is an efficient technique for routing in the presence of node failures. Several common wireless topologies exist that can be used with this technique. This paper compares the power consumption of various regular topologies using FBVR and makes appropriate recommendations. The topology types include Mesh, Torus, Communication Graph, and F-Cycle Ring (FCR).

An existing analytical method for power consumption prediction is used. The results of this analytical method are compared against simulation results, which match closely, showing a high level of confidence in the power consumption results.

A Simulation Study of the Effects of Multi-path Approaches in e-Commerce Applications

Paolo Romano, Francesco Quaglia and Bruno Ciciani

Dipartimento di Informatica e Sistemistica
Universita' di Roma "La Sapienza"
Roma, Italy
{romanop, quaglia, ciciani}@dis.uniroma1.it

Response time is a key factor of any e-Commerce application, and a set of solutions have been proposed to provide low response time despite network congestions or failures. Being them mostly based on caching of Web objects and replication of DBMS managed data at the edges, or at intermediate points, of the Web infrastructure, they reveal effective when handling client requests only performing read access to application data. However, any update request typically needs to be redirected to the origin DBMSs, hence not taking advantage from data replication and related client proximity. In order to alleviate the effects of network congestions or failures, we have proposed a multi-path protocol that increases the likelihood for the update request to be processed along a responsive (e.g. failure free) network path in between the client location and the origin DBMS sites. In this paper we present an extensive simulation study of the effects of such a multi-path approach on the client perceived response time. The study relies on both Brite generated network topologies and the NLANR graph. Also, well known realistic TCP models are used to capture the effects of network delays during both normal and anomalous (i.e. packet loss affected) operation mode.

Plan-Based Replication for Fault-Tolerant Multi-Agent Systems

Alessandro De Luna Almeida, Samir Aknine, Jean-pierre Briot and Jacques Malenfant

Laboratoire d'Informatique
Université Pierre et Marie Curie - Paris 6
Paris, France
{*Alessandro.Luna-Almeida, Samir.Aknine, Jean-Pierre.Briot, Jacques.Malenfant*}*@lip6.fr*

The growing importance of multi-agent applications and the need for a higher quality of service in these systems justify the increasing interest in fault-tolerant multi-agent systems. In this article, we propose an original method for providing dependability in multi-agent systems through replication. Our method is different from other works because our research focuses on building an automatic, adaptive and predictive replication policy where critical agents are replicated to avoid failures. This policy is determined by taking into account the criticality of the plans of the agents, which contain the collective and individual behaviors of the agents in the application. The set of replication strategies applied at a given moment to an agent is then fine-tuned gradually by the replication system so as to reflect the dynamicity of the multi-agent system.

User Perceived Unavailability due to Long Response Times

Magnos Martinello, Mohamed Kaaniche, Karama Kanoun and Carlos Aguilar Melchor

LAAS/CNRS
Toulouse, France
{*magnos, kaaniche, kanoun, caguilar*}*@laas.fr*

In this paper, we introduce a simple analytical modeling approach for computing service unavailability due to long response time, for infinite and finite single-server systems as well as multi-server systems. Closed-form equations of system unavailability based on the conditional response time distributions are derived and sensitivity analyses are carried out to analyze the impact of long response time on service unavailability. The evaluation provides practical quantitative results that can help distributed system developers in design decisions.

Predicting Failures of Computer Systems: A Case Study for a Telecommunication System

Felix Salfner, Michael Schieschke and Miroslaw Malek

Institut für Informatik
Humboldt-Universität zu Berlin
Berlin, Germany
{*salfner, schiesch, malek*}*@informatik.hu-berlin.de*

The goal of online failure prediction is to forecast imminent failures while the system is running. This paper compares Similar Events Prediction (SEP) with two other well-known techniques for online failure prediction: a straightforward method that is based on a reliability model and Dispersion Frame Technique (DFT). SEP is based on recognition of failure-prone patterns utilizing a semi-Markov chain in combination with clustering. We applied the approaches to real data of a commercial telecommunication system. Results are presented in terms of precision, recall, F-measure and accumulated runtime-cost. The results suggest a significantly improved forecasting performance.

Dynamic Resource Allocation of Computer Clusters with Probabilistic Workloads

Marwan Sleiman[1], Lester Lipsky[2] and Robert Sheahan[3]

[1]*Department of Computer Science and Engineering*
University of Connecticut
Storrs, CT, USA
marwan@engr.uconn.edu

[2]*Department of Computer Science and Engineering*
University of Connecticut
Storrs, CT, USA
lester@engr.uconn.edu

[3]*Department of Computer Science and Engineering*
University of Connecticut
Storrs, CT, USA
roberts@engr.uconn.edu

Real-time resource scheduling is an important factor for improving the performance of cluster computing. In many distributed and parallel processing systems, particularly real-time systems, it is desirable and more efficient for jobs to finish as close to a target time as possible. This work models the execution time for such a stochastic environment and proposes a dynamic algorithm for optimizing the job completion times by dynamically allocating resources to jobs that are behind schedule and taking resources from jobs that are ahead of schedule. We validate our analytical model with simulations that represent the real computing environment. The results of our simulations show that our alternative is the best estimate to predict the time remaining by using earlier data. Emphasis is placed on where variance enters the system and how well it can be controlled. Also our dynamic algorithm involves modifying the architecture to help reduce the peak number of servers used to execute a job and thus optimize the computation cost.

Workshop 16

International Workshop on Security in Systems and Networks
SSN 2006

• Workshop on Security in Systems and Networks

Workshop Description:

The proliferation of Internet services and applications is bringing systems and network security issues to the fore. The past few years have seen significant increase in cyber attacks on the Internet, resulting in degraded confidence and trusts in the use of Internet and computer systems. There is an increasing demand for measures to guarantee the privacy, integrity, and availability of resources in distributed systems, such as Grid and P2P systems. The attacks, including DDoS, email virus, and worms, are getting more sophisticated, spreading faster, and causing more damages. The attacks originally exploited the weakness of the individual protocols and operating systems, but now also have started to attack the basic infrastructure of the Internet. There is a consensus that a key contributing factor leading to cyber threats is the lack of integrated and cohesive strategies that extend beyond the network level, to protect the applications and devices at system level as well. Many techniques, algorithms, protocols and tools have been developed in the different aspects of cybersecurity, namely, authentication, access control, availability, integrity, privacy, confidentiality and non-repudiation as they apply to both networks and systems. This workshop aims to bring together the technologies and researchers who share interest in the area of network and distributed system security. The main purpose is to promote discussions of research and relevant activities in security-related subjects. It also aims at increasing the synergy between academic and industry professionals working in this area.

Topics of interest include but are not limited to:

- Ad hoc and sensor network security
- Cryptographic algorithms and distributed digital signatures
- Distributed denial of service attacks
- Distributed intrusion detection and protection systems
- Firewall and distributed access control
- Grid computing security
- Key management
- Network security issues and protocols
- Mobile codes security and Internet Worms
- Security in e-commerce
- Security in peer-to-peer and overlay networks
- Security in mobile and pervasive computing
- Security architectures in distributed and parallel systems
- Security theory and tools in distributed and parallel systems
- Video surveillance and monitoring systems
- Information hiding and multimedia watermarking in distributed systems
- Web content secrecy and integrity

General Co-Chairs:

Cheng-Zhong Xu,Wayne State University, USA

Xiaobo Zhou, University of Colorado at Colorado Springs, USA

Program Chair:

Weisong Shi, Wayne State University, USA

Program Committee:

Bill Ayen, Network Information and Space Security Center, USA
Terry Boult, University of Colorado at Colorado Springs, USA
David Chadwick, University of Salford, UK
Shigang Chen, University of Florida, USA
Huirong Fu, Oakland University, USA
Yong Guan, Iowa State University, USA
Minaxi Gupta, Indiana University, USA
John Ioannidis, Columbia University, USA
Anca Ivan, IBM T. J. Watson Research Center, USA
James B. D. Joshi, University of Pittsburgh, USA
Donggang Liu, Univeristy of Texas at Arlington, USA
Jianfeng Ma, Xidian University, China
Daniel Massey, Colorado State University, USA
Patrick McDaniel, Penn State University, USA
Geyong Min, University of Bradford, UK
Bernhard Plattner, ETH Zurich, Switzerland
Vassilis Prevelakis, Drexel University, USA
Sanjeev Setia, George Mason University, USA
Sean W. Smith, Dartmouth College, USA
Wietse Venema, IBM T.J. Watson Research Center, USA
S. Felix Wu, University of California at Davis, USA
David K.Y. Yau, Purdue University, USA
Bin Xiao, Hong Kong Polytechnic University
Yunquan Zhang, Chinese Academy of Sciences, China
Sheng Zhong, State University of New York at Buffalo, USA

Advisory Committee:

Kai Hwang, University of Southern California, USA
George Cybenko, Dartmouth College, USA
Xiaodong Zhang, Ohio State University, USA
C. Edward Chow, University of Colorado at Colorado Springs, USA

Honeypot Back-propagation for Mitigating Spoofing Distributed Denial-of-Service Attacks

Sherif Khattab[1], Rami Melhem[1], Daniel Mossé[1] and Taieb Znati[1,2]

[1]*Department of Computer Science*
University of Pittsburgh
Pittsburgh, PA, USA
{*skhattab, melhem, mosse, znati*}*@cs.pitt.edu*

[2]*Department of Information Science and Telecommunications*
University of Pittsburgh
Pittsburgh, PA, USA

The Denial-of-Service (DoS) attack remains a challenging problem in the current Internet. In a DoS defense mechanism, a honeypot acts as a decoy within a pool of servers, whereby any packet received by the honeypot is most likely an attack packet. We have previously proposed the roaming honeypots scheme to enhance this mechanism by camouflaging the honeypots within the server pool, thereby making their locations highly unpredictable. In roaming honeypots, each server acts as a honeypot for some periods of time, or honeypot epochs, the duration of which is determined by a pseudo-random schedule shared among servers and legitimate clients.

In this paper, we propose a honeypot back-propagation scheme to trace back attack sources when attacks occur. Based on this scheme, the reception of a packet by a roaming honeypot triggers the activation of a DAG of honeypot sessions rooted at the honeypot under attack towards attack sources. The formation of this tree is achieved in a hierarchical fashion: first at the Autonomous system (AS) level and then at the router level within an AS if needed. The proposed scheme supports incremental deployment and provides deployment incentives for ISPs. Through ns-2 simulations, we show how the proposed scheme enhances the performance of a vanilla Pushback defense by obtaining accurate attack signatures and acting promptly once an attack is detected.

Detecting Selective Forwarding Attacks in Wireless Sensor Networks

Bo Yu[1,2] and Bin Xiao[1]

[1]*Dept. of Computing*
Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
{*csbyu, csbxiao*}*@comp.polyu.edu.hk*

[2]*Dept. of Computer Science and Engineering*
Fudan University
Shanghai, China

Selective forwarding attacks may corrupt some mission-critical applications such as military surveillance and forest fire monitoring. In these attacks, malicious nodes behave like normal nodes in most time but selectively drop sensitive packets, such as a packet reporting the movement of the opposing forces. Such selective dropping is hard to detect. In this paper, we propose a lightweight security scheme for detecting selective forwarding attacks. The detection scheme uses a multi-hop acknowledgement technique to launch alarms by obtaining responses from intermediate nodes. This scheme is efficient and reliable in the sense that an intermediate node will report any abnormal packet loss and suspect nodes to both the base station and the source node. To the best of our knowledge, this is the first paper that presents a detailed scheme for detecting selective forwarding attacks in the environment of sensor networks. The simulation results show that even when the channel error rate is 15%, simulating very harsh radio conditions, the detection accuracy of the proposed scheme is over 95%.

A Case for Exploit-Robust and Attack-Aware Protocol RFCs

Venkat Pathamsetty[1] and Prabhaker Mateti[2]

[1]Critical Infrastructure Assurance Group
Cisco Systems
Austin, TX, USA
vpothams@cisco.com

[2]Dept. of Computer Science and Engineering
Wright State University
Dayton, OH, USA
pmateti@wright.edu

A large number of vulnerabilities occur because protocol implementations failed to anticipate illegal packets. RFCs typically define what constitute "right" packets relevant to the protocol and they specify what the response should be for such packets. They are often ambiguous and remain silent on what the protocol implementation should do for packets which deviate from the specification.

Implementers must and, by and large, do faithfully implement an RFC. However, implementers usually take any silence in a specification as "design freedom". Even though the protocol implementers are network specialists, they often are not knowledgeable in network security and cryptography issues, past exploits and common attack techniques that can impact the security of a protocol module, and consequently, the whole system.

This paper systematically discusses vulnerabilities that can be attributed to protocol designs, inadequacies of RFCs, and omissions of the protocol implementers. Using specific examples, we point out how ambiguities in protocol RFCs have lead to security vulnerabilities. We correlate various types of security vulnerabilities with the way the RFCs are written. We make a case for such exploit-robust and attack-aware RFCs, and recommend the features for a better RFC, called eRFC (Enhanced RFC). We offer advice to RFC writers, implementers and RFC approval bodies. The most effective solution to reducing network security incidents is to fix the RFCs in such a way that the implementers are forced to write an exploit-robust implementation, irrespective of their security knowledge and expertise.

Fault and Intrusion Tolerance of Wireless Sensor Networks

Liang-min Wang[1], Jian-feng Ma[1], Chao Wang[1] and Alex Chichung Kot[2]

[1]Xidian University
Key Laboratory of CNIS of Education Ministry
Xi'an Shaanxi, China
{liangminwang, ejfma}@hotmail

[2]Nanyang Technological University
School of Electrical and Electronic Engineering
Singapore
eackot@ntu.edu.sg

The following three questions should be answered in developing new topology with more powerful ability to tolerate node-failure in wireless sensor network. First, what is node-failure tolerance of topologies? Second, how to evaluate this tolerance ability? Third, which type of topologies is more efficient in tolerating node-failure? Without giving the answers, the existing work regards fault-tolerance topology as the multiply connected graph, and use the connectivity of the graph as the standard to evaluate tolerance ability. In this paper, we argue that fault tolerance of topologies is not equivalent to the connectivity of multiply connected graph by illustrating two concrete examples. Then the definition of node-failure tolerance is presented. According fault and intrusion, the two sources of failure nodes, we define fault tolerance and intrusion tolerance as the standards to evaluate the tolerance ability of topologies, and analyze the tolerance performance of hierarchical structure of wireless sensor network by using these standards. Finally, the function relation between hierarchical topology and its tolerance abilities of fault and intrusion is obtained, and an obvious corollary is that fault tolerance increase with the ratio of cluster head hierarchical structure, but with the intrusion tolerance decreasing.

Network Intrusion Detection with Semantics-Aware Capability

Walter Scheirer and Mooi Choo Chuah

Dept. of Computer Science and Engineering
Lehigh University
Bethlehem, PA, USA
{*wjs3, mcc7*}*@cse.lehigh.edu*

Malicious network traffic, including widespread worm activity, is a growing threat to Internet-connected networks and hosts. In this paper, we propose a network intrusion detection system (NIDS) with semantics-aware capability. Our NIDS segregates suspicious traffic from the regular traffic flow, extracts binary code from the suspicious traffic, and performs semantic analysis on it to identify potential threats. Our contributions in this work are threefold: (a) we believe our prototype is the first NIDS that provides semantics-aware capability, (b) our implementation is more efficient than what is reported in previously published semantic detection work, (c) our designed templates can capture polymorphic shellcodes with added sequences of stack and mathematic operations.

Analysis of BGP Prefix Origins During Google's May 2005 Outage

Tao Wan and Paul C. Van Oorschot

School of Computer Science
Carleton University
Ottawa, Ontario, Canada
{*twan, paulv*}*@scs.carleton.ca*

Google went down for 15 to 60 minutes around 22:10, May 07, 2005 UTC. This was explained by Google as having been caused by internal DNS misconfigurations. Another vulnerable protocol which could have caused such service outage is BGP. To pursue the latter possibility further, we explore how BGP was functioning during that period of time using the RouteViews BGP data set. Interestingly, our investigation reveals that one Autonomous System (i.e., AS174 operated by Cogent), which is apparently independent from Google, mysteriously originated routes for one of the IP prefixes assigned to Google (64.233.161.0/24) immediately prior to the service outage. As a result, 49.1% of ASes re-advertising routes for 64.233.161.0/24 switched to the incorrect path. Those poisoned ASes directly serve 1500 IP prefixes, and span a broad range of geographic locations. Since this erroneous prefix origination apparently has not occurred previously, or after this specific instance, we consider that it might have been the result of malicious activity (e.g., compromise of one or more BGP speakers) and contributed at least partially to Google's service outage.

A Note on Broadcast Encryption Key Management with Applications to Large Scale Emergency Alert Systems

Guoqiang Shu[1], David Lee[1] and Mihalis Yannakakis[2]

[1]Dept. of Computer Science and Engineering
The Ohio State University
Columbus, OH, USA
{shug, lee}@cse.ohio-state.edu

[2]Dept. of Computer Science
Columbia University
New York, NY, USA
mihalis@cs.columbia.edu

Emergency alerting capability is crucial for the prompt response to natural disasters and terrorist attacks. The emerging network infrastructure and secure broadcast techniques enable prompt and secure delivery of emergency notification messages. With the ubiquitous deployment of alert systems, scalability and heterogeneity pose new challenges for the design of secure broadcast schemes. In this paper we discuss the key generation problem with the goal of minimizing the total number of keys which need to be generated by the alert center and distributed to the users. Two encryption schemes, zero message scheme and extended header scheme, are modeled formally. For both schemes we show the equivalence of the general optimal key generation (OKG) problem and the bipartite clique cover (BCC) problem, and show that OKG problem is NP-Hard. The result is then generalized to the case with resource constraints, and we provide a heuristic algorithm for solving the restricted BCC (and OKG) problem.

Coordinate Transformation – A Solution for the Privacy Problem of Location Based Services?

Andreas Gutscher

Institute of Communication Networks and Computer Engineering (IKR)
Universität Stuttgart
Stuttgart, Germany
gutscher@ikr.uni-stuttgart.de

Protecting location information of mobile users in Location Based Services (LBS) is a very important but quite difficult and still largely unsolved problem. Location information has to be protected against unauthorized access not only from users but also from service providers storing and processing the location data, without restricting the functionality of the system. This paper discusses why existing privacy enhancing techniques are insufficient to solve this problem and proposes a new approach basing on coordinate transformations. It shows how location information can be rendered illegible in such a way that it is still possible to perform processing operations required by LBS.

Preserving Source Location Privacy in Monitoring-Based Wireless Sensor Networks

Yong Xi, Loren Schwiebert and Weisong Shi

Department of Computer Science
Wayne State University
Detroit, MI 48202, USA
{*yongxi, loren, weisong*}*@wayne.edu*

While a wireless sensor network is deployed to monitor certain events and pinpoint their locations, the location information is intended only for legitimate users. However, an eavesdropper can monitor the traffic and deduce the approximate location of monitored objects in certain situations. We first describe a successful attack against the flooding-based phantom routing, proposed in the seminal work by Celal Ozturk, Yanyong Zhang, and Wade Trappe. Then, we propose GROW (Greedy Random Walk), a two-way random walk, i.e., from both source and sink, to reduce the chance an eavesdropper can collect the location information. We improve the delivery rate by using local broadcasting and greedy forwarding. Privacy protection is verified under a backtracking attack model. The message delivery time is a little longer than that of the broadcasting-based approach, but it is still acceptable if we consider the enhanced privacy preserving capability of this new approach. At the same time, the energy consumption is less than half the energy consumption of flooding-base phantom routing, which is preferred in a low duty cycle, environmental monitoring sensor network.

Shubac: A Searchable P2P Network Utilizing Dynamic Paths for Client/Server Anonymity

Aharon Brodie and Cheng-zhong Xu

Department of Electrical and Computer Engineering
Wayne State University
Detroit, Michigan, USA
asb@brodie.com, czxu@wayne.edu

A general approach to achieve anonymity on P2P networks is to construct an indirect path between client and server for each data transfer. The indirection, together with randomness in the selection of intermediate nodes, provides a guarantee of anonymity to some extent. It, however, comes at the cost of a large communication overhead. In this paper, we present Shubac, a searchable, anonymous peer to peer (P2P) overlay network. It implements a flexible dynamic path approach that shrinks paths in size to reduce overhead and delays and meanwhile reconfigures paths dynamically throughout a communication to maintain a high level of privacy. This dynamic path approach enables Shubac to make a good tradeoff between anonymity and efficiency.

Energy-Efficient ID-based Group Key Agreement Protocols for Wireless Networks

Chik How Tan[1] and Joseph Chee Ming Teo[2]

[1] *Department of Computer Science and Media Technology*
Gjøvik University College
Gjøvik, NO, Norway
chik.tan@hig.no

[2] *School of Electrical and Electronic Engineering*
Nanyang Technological University
Singapore, SG, Singapore
jose0002@ntu.edu.sg

One useful application of wireless networks is for secure group communication, which can be achieved by running a Group Key Agreement (GKA) protocol. One well-known method of providing authentication in GKA protocols is through the use of digital signatures. Traditional certificate-based signature schemes require users to receive and verify digital certificates before verifying the signatures but this process is not required in ID-based signature schemes. In this paper, we present an energy-efficient ID-based authenticated GKA protocol and four energy-efficient ID-based authenticated dynamic protocols, namely Join, Leave, Merge and Partition protocol, to handle dynamic group membership events, which are frequent in wireless networks. We provide complexity and energy cost analysis of our protocols and show that our protocols are more energy-efficient and suitable for wireless networks.

Base Line Performance Measurements of Access Controls For Libraries and Modules

Jason W Kim and Vassilis Prevelakis

Department of Computer Science
Drexel University
Philadelphia, PA, USA
{jkim, vp}@cs.drexel.edu

Having reliable security in systems is of the utmost importance. However, the existing framework of writing, distributing and linking against code in the form of libraries and/or modules does a very poor job of keeping track of who has access to what code and who can call what function.

The status-quo is insufficient for a variety of reasons. As the amount of code written that represents some kind of a rights-protected entity increases, we need a systematic, easily adopted framework for designating who has access to what code, and under which conditions.

While adding access controls to libraries and modules (as well as functions held securely within them), we also give regard to the performance characteristics and ease-of-use considerations. In this vein, we discuss the design and implementation of a framework (called SecModule) used for generating (and using) libraries under access controls, as well as performance measurements of invoking functions that are held inside the protected library.

Automated Refinement of Security Protocols

Anders M. Hagalisletto

Department of Computer Science
University of Oslo
Oslo, Norway
andersmo@ifi.uio.no

The design of security protocols is usually performed manually by pen and paper, by experts in security. Assumptions are rarely specified explicitly. We present a new way to approach security specification: The protocol is refined fully automated into a specification that contains assumptions sufficient to execute the protocol. As a result, the protocol designer using our method does not have to be a security expert to design a protocol, and can learn immediately how the protocol should work in practice.

A Correctness Proof of the SRP Protocol

Huabing Yang, Xingyuan Zhang and Yuanyuan Wang

Institute of Command Automation
PLA University of Science and Technology
Nanjing, Jiangsu, P.R. China
{*yanghuabing, wangyy2005*}*@gmail.com, xyzhang@public1.ptt.js.cn*

The correctness of a routing protocol can be divided into two parts, a liveness property proof and a safety property proof. The former requires that route(s) should be discovered and data be transmitted successfully, while the latter requires that the discovered routes have some desired characters such as containing only benign nodes. While safety properties are relatively easier to prove, the proof of liveness properties is usually harder. This paper presented a liveness proof of a secure routing protocol, SRP in Isabelle/HOL. The liveness property proved says that if a data package needs to be sent, then it will be sent and then received, and finally, the sender will receive an acknowledgement sent back by the receiver. There are three main contributions in this paper. Firstly, a liveness property is proved for a secure routing protocol, and this has never been done before. Secondly, our validation model can deal with arbitrarily many nodes including malicious ones, and nodes are allowed to move randomly. Thirdly, a *fail* set is defined to restrict the attackers' actions, so that the safety properties used to prove the liveness property can be established. The paper explains why it is reasonable to prevent malicious nodes from performing the events in *fail* set.

Checkpointing and Rollback-Recovery Protocol for Mobile Systems with MW Session Guarantee

Jerzy Brzezinski[1], Anna Kobusinska[2] and Michal Szychowiak[3]

[1]Institute of Computing Science
Poznan University of Technology
Poznan, Poland
Jerzy.Brzezinski@cs.put.poznan.pl

[2]Institute of Computing Science
Poznan University of Technology
Poznan, Poland
Anna.Kobusinska@cs.put.poznan.pl

[3]Institute of Computing Science
Poznan University of Technology
Poznan, Poland
Michal.Szychowiak@cs.put.poznan.pl

In the mobile environment, weak consistency replication of shared data is the key to obtaining high data availability, good access performance, and good scalability. Therefore new class of consistency models, called session guarantees, recommended for mobile environment, has been introduced. Session guarantees, called also client-centric consistency models, have been proposed to define required properties of the system regarding consistency from the clients point of view. Unfortunately, none of proposed consistency protocols providing session guarantees is resistant to server failures. Therefore, in this paper checkpointing and rollback-recovery protocol rVsMW, which preserves Monotonic Writes session guarantee is presented. The recovery protocol is integrated with the underlying consistency protocol by integrating operations of taking checkpoints with coherence operations of VsSG protocol.

Workshop 17

Workshop on System Management Tools for Large-Scale Parallel Systems
SMTPS 2006

17 SMTPS • Workshop on System Management Tools for Large-Scale Parallel Systems

Workshop Description:

We are entering a new era in computing where the size and complexity of scientific and engineering simulations is growing at a speed that has never been observed before. In order to satisfy the needs of these applications, parallel systems with an "extreme-scale" are being designed and deployed. Although the progress in hardware and architecture design has made it possible to build machines with tens of thousands of processors, the development of software tools for such systems is still lagging behind. To name just a few examples, new operating system level modifications are needed to efficiently utilize the massive computing and networking power. In addition, sophisticated fault-tolerant tools are in great need to minimize the performance loss under a faulty condition and to automate the recovery process which can further reduce management costs. The scale of the systems also demands advanced power management tools. For both commodity supercomputing clusters and custom-designed supercomputers, system maintenance, reliability, fault isolation, prevention and control pose huge challenges. There is a great need of research not only in terms of scale of the machine, but also in terms of their implications on system performance and utilization. This workshop is intended to bring together researchers and practitioners to begin identifying the new challenges imposed by this trend and investigating efficient software tools to improve the performance, reliability and operation of large scale parallel systems.

Topics of interest include but are not limited to:

- Scalable operating system design
- Scalable resource management tools
- Efficient failure diagnosis, failure prediction and failure recovery tools
- Scalable job scheduling tools
- Scalable parallel check-pointing tools
- Self-healing and self-management tools
- Power management for large scale machines
- System bring-up and control tools
- Ease of system maintenance, services including system management experiences
- Performance, system utilization implications
- Scalable I/O and file system management

Workshop Co-Chairs:

Fabrizio Petrini, Pacific Northwest National Lab, USA

Ramendra Sahoo, IBM Research, USA

Yanyong Zhang, Rutgers University, USA

Program Chair:

Kyung Dong Ryu, IBM Research, USA

Program Committee:

Ricardo Bianchini, Rutgers
Henri Casanova, UCSD
Dick Epema, Delft
Dror Feitelson, Hebrew University
Rahul Garg, IBM India
Ravishankar Iyer, Intel
John Janakiraman, HP
Joefon Jann, IBM Research
Jose E. Moreira, IBM
Manish Parashar, Rutgers
Anand Sivasubramaniam, Penn State
Rajeev Thakur, Argonne
Andy Yoo, LLNL

SMTPS Keynote: Research and Technology Advances in Systems Software for Large Scale Computing Systems

Frederica Darema

NSF/CISE
Arlington, VA, USA
fdarema@nsf.gov

The talk will address research and technology advances for optimized and dependable execution in large scale computing environments. Applications in nearly all sectors, scientific, engineering, and commercial, are becoming more encompassing in including the behaviors of the systems of the systems they represent, and becoming at the same time more powerful but also more complex. At the same time, driven by application requirements and enabled by hardware technology advances, computational platforms are becoming as well increasingly more powerful but also more complex. Efficient and effective development of applications, optimized use of the computational resources, and guaranteeing quality of service and dependability at all layers of the computational system, requires systems software advances, such as in programming environments, application composition systems, optimized application mapping and dynamic runtime technologies, debugging and check-pointing methods, and performance-engineered hardware and software capabilities at all layers. An overarching consideration, and thesis of this talk, is that these advances need to be made in a synergistic and integrated manner, taking a systems-view in developing these enabling technologies, rather than advancing each of the individual technologies in an isolated manner.

On-the-Fly Kernel Updates for High-Performance Computing Clusters

Kristis Makris[1] and Kyung Dong Ryu[2]

[1]*Dept. of Computer Science and Engineering*
Arizona State University
Tempe, AZ, USA
kristis.makris@asu.edu

[2]*IBM T. J. Watson Research Center*
Yorktown Heights, NY, USA
kryu@us.ibm.com

High-performance computing clusters running long-lived tasks currently cannot have kernel software updates applied to them without causing system downtime. These clusters miss opportunities for increased performance via specialized kernel support, cannot benefit from new kernel features, and continue to operate with kernel security holes unpatched, at least until the next scheduled maintenance date. We developed a system enabling dynamic kernel updates in parallel computing clusters to address these problems. Our system, DynAMOS, is founded on execution flow high-jacking through *function cloning*. It enables commodity operating systems popularly used in clusters gain adaptive and mutative capabilities.

To demonstrate the efficacy of our system, we illustrate our experience in dynamically updating and extending a Linux cluster. We introduce adaptive memory paging for efficient gang-scheduling, extend the kernel's process scheduler to support unobtrusive fine-grain cycle stealing, apply public security fixes, and inject performance monitoring functionality to a selection of kernel functions. Our benchmarks show that the overhead imposed by DynAMOS is mostly in the range of 1-8% for common Linux kernel functions.

A Tool for Environment Deployment in Clusters and light Grids

Yiannis Georgiou, Julien Leduc, Brice Videau, Johann Peyrard and Olivier Richard

Laboratoire Informatique et Distribution ID-IMAG (UMR5132)
GRENOBLE, FRANCE
{*Yiannis.Georgiou, Julien.Leduc, Brice.Videau, Johann.Peyrard, Olivier.Richard*}*@imag.fr*

Focused around the field of the exploitation and the administration of high performance large-scale parallel systems , this article describes the work carried out on the deployment of environment on high computing clusters and grids. We initally present the problems involved in the installation of an environment (OS, middleware, libraries, applications...) on a cluster or grid and how an effective deployment tool, *Kadeploy2*, can become a new form of exploitation of this type of infrastructures. We present the tool's design choices, its architecture and we describe the various stages of the deployment method, introduced by *Kadeploy2*. Moreover, we propose methods on the one hand, for the improvement of the deployment time of a new environment; and in addition, for the support of various operating systems. Finally, to validate our approach we present tests and evaluations realized on various clusters of the experimental grid *Grid5000*.

Lossless Compression for Large Scale Cluster Logs

Raju Balakrishnan[1] and Ramendra K. Sahoo[2]

[1]*India Software Laboratory, IBM*
Koramangala Rd
Bangalore, Karnataka, India
rajubala@in.ibm.com

[2]*IBM TJ Watson Research*
19 Skyline Dr.
Hawthorne, NY 10532, USA
rsahoo@us.ibm.com

The growing computational and storage needs of several scientific applications mandate the deployment of extreme-scale parallel machines, such as IBMs Blue Gene/L which can accommodate as many as 128K processors. One of the biggest challenges these systems face, is to manage generated system logs while deploying in production environments. Large amount of log data is created over extended period of time, across thousands of processors. These logs generated can be voluminous because of the large temporal and spatial dimensions, and containing records which are repeatedly entered to the log archive. Storing and transferring such large amount of log data is a challenging problem. Commonly used generic compression utilities are not optimal for such large amount of data considering a number of performance requirements. In this paper we propose a compression algorithm which preprocesses these logs before trying out any standard compression utilities. The compression ratios and times for the combination shows 28.3% improvement in compression ratio and 43.4% improvement in compression time on average over different generic compression utilities. The test data used is log data produced by 64 racks, 65536 processor Blue Gene/L installation at Lawrence Livermore National Laboratory.

Evaluating Cooperative Checkpointing for Supercomputer Systems

Adam J. Oliner[1] and Ramendra K. Sahoo[2]

[1]*Computer Science Department*
Stanford University
Palo Alto, CA, United States
oliner@cs.stanford.edu

[2]*IBM T. J. Watson Research Center*
Hawthorne, NY, United States
rsahoo@us.ibm.com

Cooperative checkpointing, in which the system dynamically skips checkpoints requested by applications at runtime, can exploit system-level information to improve performance and reliability in the face of failures. We evaluate the applicability of cooperative checkpointing to large-scale systems through simulation studies considering real workloads, failure logs, and different network topologies. We consider two cooperative checkpointing algorithms: *work-based* cooperative checkpointing uses a heuristic based on the amount of unsaved work and *risk-based* cooperative checkpointing leverages failure event prediction. Our results demonstrate that, compared to periodic checkpointing, risk-based checkpointing with event prediction accuracy as low as 10% is able to significantly improve system utilization and reduce average bounded slowdown by a factor of 9, without losing any additional work to failures. Similarly, work-based checkpointing conferred tremendous performance benefits in the face of large checkpoint overheads.

Easy and Reliable Cluster Management: The Self-management Experience of Fire Phoenix

Zhang Zhi-hong, Meng Dan, Zhan Jian-feng, Wang Lei, Wu Lin-ping and Huang Wei

Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
{*zzh, dm, jfzh, wl, wlp, hw*}*@ncic.ac.cn*

High-Performance clusters are rapidly becoming an important computing platform for both scientific and business applications. To fulfill the new demands and challenges, cluster system software is inevitably complex. Even for experienced administrators, the management of a cluster system is an exhausting job. This paper introduces Fire Phoenix, a scalable and self-managing cluster system software that supports both scientific and commercial applications. With the self-configuring and self-healing features, much of the machine configuration and error recovery can be done automatically. Our design has been proven effective in the operations of the Dawning 4000A supercomputer, which is the biggest cluster system in China.

Resource Management with Stateful Support for Analytic Applications

Liana L. Fong[1], Catherine H. Crawford[2] and Hidayatullah Shaikh[1]

[1]*IBM T.J. Watson Research Center*
Yorktown Heights, New York, USA
{*llfong, hshaikh*}*@us.ibm.com*

[2]*IBM System Technology Group*
IBM T.J. Watson Research Center
Yorktown Heights, New York, USA
catcraw@us.ibm.com

Analytic applications from various industrial sectors have specific attributes and requirements including relatively long processing time, parallelization, multiple interactive invocations, web services, and expected quality of service objectives. Current parallel resource management systems for batch-oriented jobs lack the effective support for multiple interactive invocations with consideration in quality of service objectives, while transaction processing systems do not support dynamic creation of parallel application instances. To better serve the analytic applications, a set of additional resource management services, defined as stateful support, introduces the concept of Service Instance and Service Instance Management. This set of stateful support services can be implemented as extension to existing parallel resource management to serve these analytic applications that rapidly increase in the demand of computing power.

Improving Cluster Utilization through Intelligent Processor Sharing

Gary Stiehr and Roger D. Chamberlain

Department of Computer Science and Engineering
Washington University
Saint Louis, MO, USA
{*garystiehr, roger*}*@wustl.edu*

A dedicated cluster is often not fully utilized even when all of its processors are allocated to jobs. This occurs any time that a running job does not use 100% of each of the processors allocated to it. We increase the throughput and efficiency of the cluster by scheduling background jobs to run concurrently with the primary jobs originally scheduled on the cluster. We do this while maintaining the quality of service provided to the primary jobs. Our results come from empirical measurements using production applications.

A Database-centric Approach to System Management in the Blue Gene/L Supercomputer

Ralph Bellofatto[1], Paul G. Crumley[1], David Darrington[2], Brant Knudson[2], Mark Megerian[2], Jose E. Moreira[1], Alda S. Ohmacht[1], John Orbeck[2], Don Reed[2] and Greg Stewart[2]

[1]*IBM Thomas J. Watson Research Center*
Yorktown Heights, NY, USA
{*ralphbel, pgc, jmoreira, sanomiya*}*@us.ibm.com*

[2]*IBM Systems and Technology Group*
Rochester, MN, USA
{*ddarring, bknudson, megerian, orbeck, donreed, gregstew*}*@us.ibm.com*

In designing the management system for Blue Gene/L, we adopted a database-centric approach. All configuration and operational data for a particular Blue Gene/L system are stored in a relational database that is kept in the systems service node. The database also serves as the communication bus for the various processes implementing the management system. This design offers many advantages, including the ability to use SQL commands to retrieve reliability, availability, and serviceability (RAS) information about the system. Information about machine partitioning and user jobs can be obtained the same way. Leveraging the database, we have developed a web interface for system management. This management system has been successfully implemented and deployed in all 19 Blue Gene/L installations at the time of this writing.

OVIS: A Tool for Intelligent, Real-time Monitoring of Computational Clusters

J. M. Brandt, A. C. Gentile, D. J. Hale and P. P. Pebay

Sandia National Laboratories
Livermore, CA, USA
{*brandt, gentile, djhale, pppebay*}*@sandia.gov*

Traditional cluster monitoring approaches consider nodes in singleton, using manufacturer-specified extreme limits as thresholds for failure "prediction". We have developed a tool, OVIS, for monitoring and analysis of large computational platforms which, instead, uses a statistical approach to characterize single device behaviors from those of a large number of statistically similar devices.

Baseline capabilities of OVIS include the visual display of deterministic information about state variables (*e.g.*, temperature, CPU utilization, fan speed) and their aggregate statistics. Visual consideration of the cluster as a comparative ensemble, rather than as singleton nodes, is an easy and useful method for tuning cluster configuration and determining effects of real-time changes.

Additionally, OVIS incorporates a novel Bayesian inference scheme to dynamically infer models for the normal behavior of a system and to determine bounds on the probability of values evinced in the system. Individual node values that are unlikely given the current applicable model are flagged as aberrant. This can be a much earlier indicator of problems than waiting for the crossing of some threshold that is necessarily set high to preclude too many false alarms.

We present OVIS and discuss its applications in cluster configuration and environmental tuning and to abnormality and problem discovery in our production clusters.

A Study of MPI Performance Analysis Tools on Blue Gene/L

I-hsin Chung, Robert E. Walkup, Hui-fang Wen and Hao Yu

IBM Thomas J. Watson Research Center
Yorktown Heights, NY, USA
{*ihchung, walkup, hfwen, yuh*}*@us.ibm.com*

Applications on todays massively parallel supercomputers rely on performance analysis tools to guide them toward scalable performance on thousands of processors. However, conventional tools for parallel performance analysis have serious problems due to the large data volume that may be required. In this paper, we discuss the scalability issue for MPI performance analysis on Blue Gene/L, the worlds fastest supercomputing platform. We present an experimental study of existing MPI performance tools that were ported to BG/L from other platforms. These tools can be classified into two categories: profiling tools that collect timing summaries, and tracing tools that collect a sequence of time-stamped events. Profiling tools produce small data volumes and can scale well, but tracing tools tend to scale poorly. The experimental study discusses the advantages and disadvantages for the tools in the two categories and will be helpful in the future performance tools design.

A Multiprocessor Architecture for the Massively Parallel Model GCA

Wolfgang Heenes, Rolf Hoffmann and Johannes Jendrsczok

FG Rechnerarchitektur, FB Informatik
TU Darmstadt
Darmstadt, Hessen, Germany
{*heeens, hoffmann, jendrsczok*}*@ra.informatik.tu-darmstadt.de*

The GCA (Global Cellular Automata) model consists of a collection of cells which change their states synchronously depending on the states of their neighbors like in the classical CA model. In differentiation to the CA model the neighbors are not fixed and local, they are variable and global. The GCA model is applicable to a wide range of parallel algorithms. In this paper a multiprocessor architecture for the massively parallel GCA model is presented. In contrast to a special purpose implementation of a GCA algorithm the multiprocessor system allows the implementation in a flexible way through programming. The architecture mainly consists of a number of cell processors and a network. The cell processors are dedicated RISC processors, the network is a crossbar implemented with multiplexers. Only read-accesses through the network are necessary in the GCA model leading to a simplified structure. A system with 32 processors was implemented as a prototype on a FPGA. The analysis and implementation results have shown that the performance of the system scales very well with the number of processors.

Dynamic Performance Prediction of an Adaptive Mesh Application

Mark M Mathis and Darren J Kerbyson

Performance and Architecture Laboratory (PAL)
Los Alamos National Laboratory
Los Alamos, NM, USA
{mmathis, djk}@lanl.gov

While it is possible to accurately predict the execution time of a given iteration of an adaptive application, it is not generally possible to predict the data-dependent adaptive behavior the application will take and therefore to predict the total execution time for a given execution. To remedy this situation we have developed an executable performance model that can be utilized dynamically at runtime directly from the application of interest. In this manner, the application itself can rapidly predict the expected execution time for its next iteration based on current information on the data layout and level of adaptivity. This enables the application itself to determine: if an optimum level of performance is being achieved (i.e. by comparing measured and predicted times); when to perform a checkpoint (if the next iteration will exceed a predefined time limit between checkpoints); or when to terminate (if the next iteration will exceed the application's system time allocation for instance). The dynamic model is shown to have high accuracy over a number of test cases, even in the presence of interference (system activities that are not a part of application activities).

Workshop 18

International Workshop on Hot Topics in Peer-to-Peer Systems
HOTP2P 2006

18 HOTP2P • Workshop on Hot Topics in Peer-to-Peer Systems

Workshop Description:

Peer-to-Peer (P2P) systems are decentralized, self-organizing distributed systems that cooperate to exchange data. These systems have emerged as the dominant consumer of residential Internet subscribers' bandwidth, and are being increasingly used in many different application domains. In the last few years, research on P2P systems has been quite intensive, and has produced remarkable results in scalability, robustness, location, distributed storage, and system measurements. Consequently, P2P systems continue to evolve, differentiating today's state-of-the-art from earlier instantiations such as Napster, KaZaA, Gnutella, and Morpheus. The International Workshop on Hot Topics in Peer-to-Peer Systems (Hot-P2P) aims to bring together researchers and practitioners, from both industry and academia, in the fields of systems, networking, and theory, and to represent an occasion to share latest research results and ideas on P2P systems, thereby promoting research activities in this area.

Topics of interest include but are not limited to:

- Applications of P2P systems
- P2P systems and infrastructures
- Performance evaluation of P2P systems
- Workload characterization for P2P systems
- Trust and Security issues in P2P systems
- Network support for P2P systems
- Protocols for resource managements/discovery/scheduling and their evaluation
- Fault tolerance in P2P systems
- DHT and other scalable lookup algorithms
- Self-organization and self-management in Grid-like environments

Program Chair:

Giovanni Chiola, Universita' di Genova (Italy)

Publicity Chair:

Marina Ribaudo, Universita' di Genova (Italy)

Program Committee:

Cosimo Anglano, Universita' del Piemonte Orientale "A. Avogadro" (Italy)
Giuseppe Ateniese, Johns Hopkins University (USA)
Julien Bourgeois, LIFC, Universite' Franche-Comte, (France)
Franck Cappello, INRIA/Universite' Paris Sud (France)
Walfredo Cirne, Universidade Federal de Campina Grande (Brasil)
Michele Colajanni, Universita' di Modena e Reggio Emilia (Italy)
Antonio Corradi, Universita' di Bologna (Italy)
Paul Ezhilchelvan, University of Newcastle (UK)
Luisa Gargano, Universita' di Salerno (Italy)
Giulio Iannello, Universita' Campus Biomedico, Roma (Italy)
Mario Lauria, Ohio State University (USA)
Laurent Lefevre, INRIA (France)
Luigi Mancini, Universita' di Roma "La Sapienza" (Italy)
Keith Marzullo, University of California, San Diego (USA)
Manish Parashar, Rutgers University, New Jersey (USA)
Giancarlo Ruffo, Universita' di Torino (Italy)
Sanjeev Setia, George Mason University (USA)
Gene Tsudik, University of California, Irvine (USA)
Geoffrey M. Voelker, University of California, San Diego (USA)
Rich Wolski, University of California, Santa Barbara (USA)

Neighbourhood Maps: Decentralised Ranking in Small-World P2P Networks

Matteo Dell'amico

Dipartimento di Informatica e Scienze dell'Informazione
Università di Genova
Genova, Italy
dellamico@disi.unige.it

Reputation in P2P networks is an important tool to encourage cooperation among peers. It is based on ranking of peers according to their past behaviour.

In large-scale real world networks, a global centralised knowledge about all nodes is neither affordable nor practical. For this reason, reputation ranking is often based on local history knowledge available on the evaluating node. This criterion is not optimal, since it ignores useful data about interactions with other peers.

We propose a simple, scalable and decentralised method, called "neighbourhood maps", that approximates rankings calculated using link-analysis techniques, exploiting the short-distance characteristics of small-world networks.

We test our algorithms using data from the OpenPGP web-of-trust, a real-world network of trust relationships.

Improving Cooperation in Peer-to-Peer Systems Using Social Networks

Wenyu Wang[1], Li Zhao[2] and Ruixi Yuan[3]

[1]*Automation Department*
Tsinghua Univ.
Beijing, China
wangwengyu@mails.tsinghua.edu.cn

[2]*Automation Department*
Tsinghua Univ.
Beijing, China
zhaoli04@mails.tsinghua.edu.cn

[3]*Automation Department*
Tsinghua Univ.
Beijing, China
ryuan@tsinghua.edu.cn

Rational and selfish nodes in P2P systems usually lack effective incentives to cooperate, contributing to the increase of free-riders, and degrading the system performance. Various attacks such as whitewashing, collusion, and software cracking pose great challenges on distributed reputation management. To tackle these problems, we propose to build a social network on P2P system, and use the strength of social connections to facilitate transactions in P2P system. The "small world" character of social networks makes it feasible for nodes to locate resources and conduct transactions while maintain limited local memory history. Such distributed memory combined by relationship between peers constructs a powerful reputation management network, which could have better performance than shared history system and is more robust under various attacks. Our simulation and analysis show that the social network model can greatly incent cooperation in P2P networks and enormously reduce the memory cost.

Modeling Malware Propagation in Gnutella Type Peer-to-Peer Networks

Krishna Kumar Ramachandran and Biplab Sikdar

Department of Electrical, Computer and Systems Engineering
Rensselaer Polytechnic Institute
Troy, New York, USA
{*ramak, sikdab*}*@rpi.edu*

A key emerging and popular communication paradigm, primarily employed for information dissemination, is peer-to-peer (P2P) networking. In this paper, we model the spread of malware in decentralized, Gnutella type of peer-to-peer networks. Our study reveals that the existing bound on the spectral radius governing the possibility of an epidemic outbreak needs to be revised in the context of a P2P network. We formulate an analytical model that emulates the mechanics of a decentralized Gnutella type of peer network and study the spread of malware on such networks. We show analytically, that a framework which does not incorporate the behavioral characteristics of peers ends up over estimating the epidemic threshold metric, $\mathcal{R}_0$. This in turn results in false positives, an undesirable feature.We also characterize the conditions under which the network may reach a malware free equilibrium and validate our theoretical results with numerical simulations.

Privacy-aware Presence Management in Instant Messaging Systems

Karsten Loesing, Markus Dorsch, Martin Grote, Knut Hildebrandt, Maximilian Röglinger, Matthias Sehr, Christian Wilms and Guido Wirtz

Distributed and Mobile Systems Group
Otto-Friedrich-Universität Bamberg
Bamberg, Germany
{*karsten.loesing, guido.wirtz*}*@wiai.uni-bamberg.de*, {*markus-dorsch, max.roeglinger, christian.wilms*}*@gmx.de*,
{*martin.grote, knut.hildebrandt, matthias-michael.sehr*}*@stud.uni-bamberg.de*

Information about online presence allows participants of instant messaging (IM) systems to determine whether their prospective communication partners will be able to answer their requests in a timely manner, or not. This makes IM more personal and closer than other forms of communication such as e-mail. On the other hand, revelation of presence constitutes a potential of misuse by untrustworthy entities, e.g. generation of presence logs. We argue that current IM systems do not take reasonable precautions to protect presence information. We propose an IM system designed to be robust against attacks to disclose a user's presence. It stores presence information in a distributed hash table (DHT) in a way that is only detectable and applicable for intended users and even not comprehensible for the DHT nodes. We apply an anonymous communication network to protect the users' physical addresses.

Using incentives to increase availability in a DHT

Fabio Picconi[1] and Pierre Sens[2]

[1]*Laboratoire d'Informatique de Paris 6*
Paris, France
fabio.picconi@lip6.fr

[2]*INRIA*
Rocquencourt, France
pierre.sens@lip6.fr

Distributed Hash Tables (DHTs) provide a means to build a completely decentralized, large-scale persistent storage service from the individual storage capacities contributed by each node of the peer-to-peer overlay. However, persistence can only be achieved if nodes are highly available, that is, if they stay most of the time connected to the overlay.

In this paper we present an incentives-based mechanism to increase the availability of DHT nodes, thereby providing better data persistence for DHT users. High availability increases a nodes reputation, which translates into access to more DHT resources and a better Quality-of-Service. The mechanism required for tracking a nodes reputation is completely decentralized, and is based on certificates reporting a nodes availability which are generated and signed by the nodes neighbors. An audit mechanism deters collusive neighbors from generating fake certificates to take advantage of the system.

Optimizing the finger table in Chord-like DHTs

Giovanni Chiola[1], Gennaro Cordasco[2], Luisa Gargano[2], Alberto Negro[2] and Vittorio Scarano[2]

[1]*DISI*
University of Genoa
Genoa, Italy
chiolag@acm.org

[2]*DIA*
University of Salerno
Baronissi, Italy
{*cordasco, lg, alberto, vitsca*}*@dia.unisa.it*

The Chord protocol is the best known example of implementation of logarithmic complexity routing for structured peer-to-peer networks. Its routing algorithm, however, does not provide an optimal trade-off between resources exploited (the size of the "finger table") and performance (the average/worst-case number of hops to reach destination). Cordasco et al. showed that a finger table based on Fibonacci distances provides lower number of hops with fewer table entries. In this paper we generalize this result, showing how to construct an improved finger table when the objective is to reduce the number of hops, possibly at the expense of an increased size of the finger table. Our results can also be exploited to guarantee low routing time in case a fraction of nodes is assumed to fail.

Linyphi: An IPv6-Compatible Implementation of SSR

Pengfei Di, Massimiliano Marcon and Thomas Fuhrmann

IBDS System Architecture
University of Karlsruhe
Karlsruhe, Germany
{*di, marcon, fuhrmann*}*@ira.uka.de*

Scalable Source Routing (SSR) is a self-organizing routing protocol designed for supporting peer-to-peer applications. It is especially suited for networks that do not have a well crafted structure, e.g. ad-hoc and mesh-networks. SSR is based on the combination of source routes and a virtual ring structure. This ring is used in a Chord-like manner to obtain source routes to destinations that are not yet in the respective router cache. This approach makes SSR more efficient than flooding-based, ad-hoc routing protocols like AODV or DSR. As a consequence, SSR can provide routing for very large mesh network clouds without requiring any centralized administration. Moreover, SSR directly provides the semantics of a structured routing overlay.

In this paper we present Linyphi, an implementation of SSR for wireless access routers. Linyphi combines IPv6 and SSR so that unmodified IPv6 hosts have transparent connectivity to both the Linyphi mesh network and the IPv4/v6 Internet. This allows peer-to-peer applications to directly benefit from other peers in the neighborhood without the need to route through the respective Internet service provider.

We give a basic outline of the implementation and demonstrate its suitability in real-world mesh network scenarios. Linyphi is available for download.

Interceptor: Middleware-level Application Segregation and Scheduling for P2P Systems

Cosimo Anglano

Dipartimento di Informatica
Universita' del Piemonte Orientale
Alessandria, Italy
cosimo.anglano@unipmn.it

Very large size Peer-to-Peer systems are often required to implement efficient and scalable services, but usually they can be built only by assembling resources contributed by many independent users. Among the guarantees that must be provided to convince these users to join the P2P system, particularly important is the ability of ensuring that P2P applications and services run on their nodes will not unacceptably degrade the performance of their own applications because of an excessive resource consumption. In this paper we present *Interceptor*, a middleware-level application segregation and scheduling system that is able to strictly enforce quantitative limitations on node resource usage and, at same time, to make P2P applications achieve satisfactory performance even in face of these limitations.

A Scalable Algorithm to Monitor Chord-based P2P Systems at Runtime

Andreas Binzenhöfer[1], Gerald Kunzmann[2] and Robert Henjes[1]

[1] *Institute of Computer Science*
University of Würzburg
Würzburg, Bavaria, Germany
{*binzenhoefer, henjes*}*@informatik.uni-wuerzburg.de*

[2] *Institute of Communication Networks*
Technische Universität München
München, Bavaria, Germany
gerald.kunzmann@tum.de

Peer-to-peer (p2p) systems are a highly decentralized, fault tolerant, and cost effective alternative to the classic client-server architecture. Yet companies hesitate to use p2p algorithms to build new applications. Due to the decentralized nature of such a p2p system the carrier does not know anything about the current size, performance, and stability of its application. In this paper we present an entirely distributed and scalable algorithm to monitor a running p2p network. The snapshot of the system enables a telecommunication carrier to gather information about the current performance parameters of the running system as well as to react to discovered errors.

Lightweight Emulation to Study Peer-to-Peer Systems

Lucas Nussbaum and Olivier Richard

Laboratoire Informatique et Distribution - IMAG
ENSIMAG - Antenne de Montbonnot - ZIRST
38330 Montbonnot Saint-Martin, France
{*Lucas.Nussbaum, Olivier.Richard*}*@imag.fr*

The current methods used to test and study peer-to-peer systems (namely modeling, simulation, or execution on real testbeds) often show limits regarding scalability, realism and accuracy. This paper describes and evaluates P2PLab, our framework to study peer-to-peer systems by combining emulation (use of the real studied application within a configured synthetic environment) and virtualization. P2PLab is scalable (it uses a distributed network model) and has good virtualization characteristics (many virtual nodes can be executed on the same physical node by using process-level virtualization). Experiments with the BitTorrent file-sharing system complete this paper and demonstrate the usefulness of this platform.

Simulating and Optimizing A Peer-to-Peer Computing Framework

Jean-baptiste Ernst-desmulier[1], Julien Bourgeois[1], Minh Thanh Ngo[1], François Spies[1] and Jérome Verbeke[2]

[1]*Laboratoire d'Informatique de Franche-Comté*
University of Franche-Comté
Montbéliard, France
{*ernst, bourgeoi, ngo, spies*}*@lifc.univ-fcomte.fr*

[2]*Lawrence Livermore National Laboratory*
Livermore, CA, USA
verbeke2@llnl.gov

The aim of P2P computing is to build virtual computing systems dedicated to large-scale computational problems.JXTA proposes an underlying infrastructure on which JNGI, one of the first P2P decentralized computing frameworks is built. In order to test this framework, we have built a tool named P2PPerf, which allows us to study the behavior of JNGI and to optimize it according to our simulation results.

Model-based Evaluation of Search Strategies in peer-to-peer Networks

Rossano Gaeta and Matteo Sereno

Dipartimento di Informatica
Universita
Torino, Italia
{*rossano, matteo*}*@di.unito.it*

This paper exploits a previously developed analytical modeling framework to compare several variations of the basic flooding search strategy in unstructured decentralized peer-to-peer (P2P) networks. The model predictions are used to compute system-oriented performance indexes (the average and the coefficient of variation of the number of query messages) as well as user-oriented measures (the probability of finding at least one replica of a resource, the average search time). The trade-off between the optimization of system-oriented measures and the improvement of user-oriented quality indexes is investigated for several variations of the basic flooding strategy suggesting that adding control parameters to the basic flooding mechanism might prove beneficial in this class of systems.

A formal framework for the performance analysis of P2P networks protocols

Angelo Spognardi and Roberto Di Pietro

Dipartimento di Informatica
Università di Roma "La Sapienza"
Rome, Italy
{*spognardi, dipietro*}*@di.uniroma1.it*

In this paper we propose a formal framework based on the Markov Chains to prove the performance of P2P protocols. Despite the proposal of several protocols for P2P networks, sometimes there is a lack of a formal demonstration of their performance: experimental simulations are the most used method to evaluate their performance, such as the average length of a lookup. In this paper we introduce a versatile model for the analysis of P2P protocols. We employ this model to formally prove which is the average lookup length for two sample protocols: BaRT and Koorde. We verify the effectiveness of the proposed framework also via extensive simulations.

Workshop 19

Workshop on Performance Optimization for High-Level Languages and Libraries POHLL 2006

19 POHLL • Workshop on Performance Optimization for High-Level Languages and Libraries

Workshop Description:

The complexity of software development has led to many efforts aimed at raising the level of abstraction for the programmer. This includes both object-oriented general-purpose approaches as well as domain-specific languages and libraries. While performance considerations are not paramount for all domains, there are many domains where high performance is essential. This workshop aims to bring together researchers from different domains, who have addressed performance optimization issues in the context of high-level languages/libraries and problem solving environments, to share their successes as well as the challenges they face. This workshop is of interest to researchers and graduate students in several areas such as compilation technology, domain-specific languages, library development, problem-solving environments, etc.

Topics of interest include but are not limited to:

- program synthesis to facilitate the development of high-performance programs for specific application domains such as signal processing, computational chemistry, etc.
- compile/runtime techniques for scalable implementation of "high-productivity" high-performance languages like Chapel, Fortress, X10
- compiler techniques for optimization of high-level mathematical languages like MATLAB.
- compile/runtime techniques for scalable implementations of parallel global-address space languages and libraries, such as Co-Array Fortran, Global Arrays, OpenMP, SHMEM, Titanium, UPC etc.
- development of high-performance implementations of algorithms (e.g. FFT) for a variety of architectures, by exploiting special structural properties of the algorithms.
- automatic optimization of library implementations together with the optimization of programs that use them.
- efficient synthesis of recursive linear algebra codes that exploit deep memory hierarchies in current computer systems.
- problem solving environments for high-performance computing applications
- high-performance computing with object-oriented and component-based frameworks

General/Program Co-Chairs:

Gerald Baumgartner, Louisiana State University

J. (Ram) Ramanujam, Louisiana State University

P. (Saday) Sadayappan, The Ohio State University

Program Committee:

Eduard Ayguadé, Universitat de Politècnica de Catalunya
Gerald Baumgartner, Louisiana State University
David Bernholdt, Oak Ridge National Laboratory
Daniel Chavarria Miranda, Pacific Northwest National Laboratory
Jack Dongarra, University of Tennessee
Robert van de Geijn, The University of Texas at Austin
John Gilbert, University of California, Santa Barbara
Jeremy Johnson, Drexel University
Ricky Kendall, Oak Ridge National Laboratory
Calvin Lin, The University of Texas at Austin
John Mellor-Crummey, Rice University
Jarek Nieplocha, Pacific Northwest National Laboratory
David Padua, University of Illinois at Urbana-Champaign
Keshav Pingali, Cornell University
Marcus Pueschel, Carnegie Mellon University
J. (Ram) Ramanujam, Louisiana State University
P. (Saday) Sadayappan, The Ohio State University
Rob Schreiber, Hewlett Packard Laboratories
Rich Vuduc, Lawrence Livermore National Laboratory
Trey White, Oak Ridge National Laboratory
Qing Yi, The University of Texas at San Antonio

POHLL Keynote: New Parallel Programming Abstractions and the Role of Compilers

Laxmikant V. Kale

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, USA
kale@cs.uiuc.edu

Most of the parallel programming, especially in applications in Computational Science and Engineering (CSE), is done using MPI. OpenMP is used on some shared memory platforms. However, it is becoming increasingly evident that new higher level parallel programming abstractions are needed if we have to increase programming productivity further.

Here, I present my views on what kinds of high level languages and abstractions one should look for, what research is needed to develop them, what obstacles I see in their development and adoption, and what role compilers can and should play in their development. In particular, I argue that adaptive run-time systems to separate the issues of resource management and abstractions for supporting global (but disciplined) view of data and global view of control are needed. Further, the role of compiler research needs to be directed to supporting such models, even though that requires a paradigm shift (toward simpler problems!) for the compiler research community.

Automatically Translating a General Purpose C++ Image Processing Library for GPUs

Jay L. T. Cornwall, Olav Beckmann and Paul H. J. Kelly

Department of Computing
Imperial College London
London, United Kingdom
{jay.cornwall, o.beckmann, p.kelly}@imperial.ac.uk

This paper presents work-in-progress towards a C++ source-to-source translator that automatically seeks parallelisable code fragments and replaces them with code for a graphics co-processor. We report on our experience with accelerating an industrial image processing library. To increase the effectiveness of our approach, we exploit some domain-specific knowledge of the library's semantics.

We outline the architecture of our translator and how it uses the ROSE source-to-source transformation library to overcome complexities in the C++ language. Techniques for parallel analysis and source transformation are presented in light of their uses in GPU code generation.

We conclude with results from a performance evaluation of two examples, image blending and an erosion filter, hand-translated with our parallelisation techniques. We show that our approach has potential and explain some of the remaining challenges in building an effective tool.

Memory Minimization for Tensor Contractions using Integer Linear Programming

A. Allam[1], J. Ramanujam[2], G. Baumgartner[3] and P. Sadayappan[4]

[1]Dept. of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA, USA
atef@ece.lsu.edu

[2]Dept. of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA, USA
jxr@ece.lsu.edu

[3]Dept. of Computer Science
Louisiana State University
Baton Rouge, LA, USA
gb@cse.ohio-state.edu

[4]Dept. of Computer Science and Engineering
The Ohio State University
Columbus, OH, USA
saday@cse.ohio-state.edu

This paper presents a technique for memory optimization for a class of computations that arises in the field of correlated electronic structure methods such as coupled cluster and configuration interaction methods in quantum chemistry. In this class of computations, loop computations perform a multi-dimensional sum of product of input arrays. There are many different ways to get the same final results that differ in the required number of arithmetic operations required. In addition, for a given number of arithmetic operations, different expressions of the loop have different memory requirements. Loop fusion is a plausible solution for reducing memory usage. By fusing loops between the producer and consumer loop nests, the required storage of intermediate array is reduced by the range of the fused loop. Because resultant loops have to be legal after fusion, some loops can not be fused at the same time. This paper develops a novel integer linear programming (ILP) formulation that is shown to be highly effective on a number of test cases producing the optimal solutions using very small execution times. The main idea in the ILP formulation is the encoding of legality rules for loop fusion of a special class of loops using logical constraints over binary decision variables and a highly effective approximation of memory usage.

Improving Locality of Nonserial Polyadic Dynamic Programming

Guangming Tan[1,2], Ninghui Sun[1] and Dongbo Bu[1]

[1]Institue of Computing Technology
Chinese Academy of Sciences
Beijing, P. R. China
{tgm, snh, bdb}@ncic.ac.cn

[2]Graduate School of Chinese Academy of Sciences
Chinese Academy of Sciences
Beijing, P. R. China

Dynamic programming (DP) is a commonly used technique for solving a wide variety of discrete optimization problems, which have different variants of dynamic programming formulation. This paper investigated one important DP formulation, which called nonserial polyadic dynamic programming formulation and time complexity is $O(n^3)$. We exploit the property of the algorithm to develop a high performance implementation using the combination of cache-oblivious and cache-conscious strategy. The efficiency in our improved algorithm comes from two sources: reducing the number of cache misses and TLB misses. Experiments on three modern computing platforms show a performance improvement of 2-10 times over a standard implementation of DP formulation.

An Approach to Locality-Conscious Load Balancing and Transparent Memory Hierarchy Management with a Global-Address-Space Parallel Programming Model

Sriram Krishnamoorthy[1], Umit Catalyurek[2], Jarek Nieplocha[3] and P. Sadayappan[1]

[1]*Department of Computer Science and Engineering*
The Ohio State University
Columbus, OH, USA
{*krishnsr, saday*}*@cse.ohio-state.edu*

[2]*Department of Biomedical Informatics*
The Ohio State University
Columbus, OH, USA
umit@bmi.osu.edu

[3]*Computational Sciences and Mathematics*
Pacific Northwest National Laboratory
Richland, WA, USA
jarek.nieplocha@pnl.gov

The development of efficient parallel out-of-core applications is often tedious, because of the need to explicitly manage the movement of data between files and data structures of the parallel program. Several large-scale applications require multiple passes of processing over data too large to fit in memory, where significant concurrency exists within each pass. This paper describes a global-address-space framework for the convenient specification and efficient execution of parallel out-of-core applications operating on block-sparse data. The programming model provides a global view of block-sparse matrices and a mechanism for the expression of parallel tasks that operate on block-sparse data. The tasks are automatically partitioned into phases that operate on memory-resident data, and mapped onto processors to optimize load balance and data locality. Experimental results are presented that demonstrate the utility of the approach.

Support for Adaptivity in ARMCI Using Migratable Objects

Chao Huang, Chee Wai Lee and Laxmikant V. Kale

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, USA
{*chuang10, cheelee, kale*}*@cs.uiuc.edu*

Many new paradigms of parallel programming have emerged that compete with and complement the standard and well-established MPI model. Most notable, and successful, among these are models that support some form of global address space. At the same time, approaches based on migratable objects (also called virtualized processes) have shown that resource management concerns can be separated effectively from the overall parallel programming effort. For example, Charm++ supports dynamic load balancing via an intelligent adaptive runtime system. It is also becoming clear that a multi-paradigm approach that allows modules written in one or more paradigms to coexist and co-operate will be necessary to tame the parallel programming challenge.

ARMCI is a remote memory copy library that serves as a foundation of many global address space languages and libraries. This paper presents our preliminary work on integrating and supporting ARMCI with the adaptive run-time system of Charm++ as a part of our overall effort in the multi-paradigm approach.

A Decomposition Approach for Optimizing the Performance of MPI Libraries

Olaf Hartmann[1], Matthias Kühnemann[1], Thomas Rauber[2] and Gudula Rünger[1]

[1]Department of Computer Science
Chemnitz University of Technology
Chemnitz, Germany
{ruenger, kumat, ruenger}@informatik.tu-chemnitz.de

[2]Department of Computer Science
University Bayreuth
Bayreuth, Germany
rauber@uni-bayreuth.de

MPI provides a portable message passing interface for many parallel execution platforms but may lead to inefficiencies for some platforms and applications. In this article we show that the performance of both, standard libraries and vendor-specific libraries, can be improved by an orthogonal organization of the processors in 2D or 3D meshes and by decomposing the collective communication operations into several phases. We describe an adaptive approach with a configuration phase to determine for a specific execution platform and a specific MPI library which decomposition leads to the best performance. This may also depend on the number of processors and the size of the messages to be transferred. The decomposition approach has been implemented in the form of a library extension which is called for each activation of a collective MPI operation. This has the advantage that neither the application programs nor the MPI library need to be changed while leading to significant performance improvements for many collective MPI operations.

Annotating User-Defined Abstractions for Optimization

Dan Quinlan[1], Markus Schordan[2], Richard Vuduc[1] and Qing Yi[3]

[1]Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
Livermore, CA, USA
{dquinlan, richie}@llnl.gov

[2]Institute of Computer Languages
Vienna University of Technology
Vienna, Austria
markus@complang.tuwien.ac.at

[3]Dept. of Computer Science
University of Texas at San Antonio
San Antonio, TX, USA
qingyi@cs.utsa.edu

Although conventional compilers implement a wide range of optimization techniques, they frequently miss opportunities to optimize the use of abstractions, largely because they are not designed to recognize and use the relevant semantic information about such abstractions. In this position paper, we propose a set of annotations to help communicate high-level semantic information about abstractions to the compiler, thereby enabling the large body of traditional compiler optimizations to be applied to the use of those abstractions. Our annotations explicitly describe properties of abstractions that are needed to guarantee the applicability and profitability of a broad variety of such optimizations, including memoization, reordering, data layout transformations, and inlining and specialization.

Effecting Parallel Graph Eigensolvers Through Library Composition

Alex Breuer, Peter Gottschling, Douglas Gregor and Andrew Lumsdaine

Open Systems Laboratory
Indiana University
Bloomington, IN, USA
{*abreuer, pgottsch, dgregor, lums*}*@osl.iu.edu*

Many interesting problems in graph theory can be reduced to solving an eigenproblem of the adjacency matrix or Laplacian of a graph. Given the availability of high-quality linear algebra and graph libraries, one might expect that one could merely use a graph data structure within a eigensolver. However, conventional libraries are rigidly constructed, requiring conversion to library-specific data structures or using heavyweight abstraction methods that prevent efficient composition.

The Generic Programming methodology addresses the problems of reusability and composability by careful factorization of a domain into efficient library abstractions. We describe the composition process that makes the data structures from a library supporting one domain usable with the algorithms of another library for a disjoint domain without conversion or heavyweight abstractions. To illustrate the process, we compose two separately-developed libraries, one for solving eigenproblems sequentially and the other for solving graph problems in parallel, effecting an efficient, scalable parallel graph eigensolver.

On the impact of data input sets on statistical compiler tuning

Masayo Haneda[1], Peter M. W. Knijnenburg[2] and Harry A. G. Wijshoff[3]

[1]*LIACS*
Leiden University
Leiden, The Netherlands
haneda@liacs.nl

[2]*Insitute for Informatics*
University of Amsterdam
Amsterdam, The Netherlands
peterk@science.uva.nl

[3]*LIACS*
Leiden University
Leiden, The Netherlands
harryw@liacs.nl

In recent years, several approaches have been proposed to use profile information in compiler optimization. This profile information can be used at the source level to guide loop transformations as well as in the backend to guide low level optimizations. At the same time, profile guided library generators have been proposed also, like Atlas, Spiral, or FFTW, that tune their routines for the underlying hardware. These approaches have led to excellent performance improvements. However, a possible drawback of these approaches is that applications are optimized using a single or a limited set of data inputs. It is well known that programs can exhibit vastly differing behaviors for different inputs. Therefore, it is not clear whether the performance numbers reported are still valid for other input than the input used to optimize the program. In this paper, we address this problem for a specific statistical compiler tuning method. We use three different platforms and several SPECint2000 benchmarks. We show that when we tune the compiler using train data, we obtain a compiler setting that still performs well for reference data. These results suggest that profile guided optimization may be more stable than is sometimes believed and that a limited number of train data sets is sufficient to obtain a well optimized program for all inputs.

A General Data Dependence Analysis to Nested Loop Using Integer Interval Theory

Zhou Jing[1] and Zeng Guosun[2]

[1] *Department of computer Science*
Tongji University
Shanghai, Shanghai, China
dinese@163.com

[2] *Tongji Branch, National Engineering*
Tongji University
Shanghai, Shanghai, China
gszeng@mail.tongji.edu.cn

Many dependence tests have been proposed for loop parallelization in the case of arrays with linear subscripts, but little work has been done on the arrays with non-linear subscripts, which sometimes occur in parallel benchmarks and scientific and engineering applications. This paper focuses on array subscripts coupled integer power index variables. We attempt to use the integer interval theory to solve the above difficult dependence test problem. Some interval solution rules for polynomial equations have been proposed in this paper. Furthermore, based on the proposed rules, we present a novel approach to loop dependence analysis, which is termed the Polynomial Variable Interval test or PVI-test, and also develop a related algorithm. Some case studies show that the PVI-test is effective and efficient. Compared to the VI test, the PVI-test makes significant improvement, and is therefore a more general scheme of dependence test.

Index