Proceedings
21st International Parallel and Distributed Processing Symposium

IPDPS 2007
Abstracts and CD-ROM

IEEE Catalog Number: 07TH8938
ISBN: 1-4244-0909-8
Library of Congress: 2006940135

Proceedings
21st International Parallel and Distributed Processing Symposium

March 26–30, 2007
Long Beach, California USA

Sponsored by
IEEE Computer Society Technical Committee on Parallel Processing

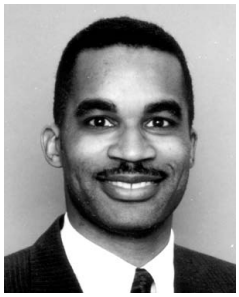
In Cooperation with
IEEE Computer Society Technical Committee on Computer Architecture (TCCA)
IEEE Computer Society Technical Committee on Distributed Processing (TCDP)
ACM SIGARCH

Summary of Contents

Detailed Table of Contents

International Parallel and Distributed Processing Symposium
IPDPS 2007

Message from the General Chair



Welcome to the 21st International Parallel and Distributed Processing Symposium (IPDPS'07) taking place here in Long Beach, California, located within the greater Los Angeles metropolitan area. This marks the first time in twelve years that the symposium has been held in southern California, the entertainment capital of the world and home of many highly-ranked universities and top Fortune-500 companies. We are very pleased you are taking part in this technical forum centered on parallel and distributed processing which brings together top research scientists and engineers from all around the world from academia, industry, and research laboratories. This area of computing is growing in importance and pervasiveness, particularly given that we are entering an era in which computer systems are built from emerging multicore processors. We have worked very hard to make every facet of this symposium interesting for attendees by carefully selecting 109 general session technical papers (including four "best papers"), twenty-one workshops and associated papers, three keynote speakers, a banquet speaker, a commercial track and associated papers, a symposium panel, and a symposium tutorial.

The success of this symposium is a direct result of the hard work and contributions made by many, including the organizing committee members, the steering committee, the authors and sponsors-all of whom I gratefully acknowledge. I am especially indebted to D.K. Panda, the Program Chair, who assembled an excellent group of Program Committee members and Program Vice-Chairs for the four main areas of the symposium: Yves Robert for "Algorithms," Srinivas Aluru for "Applications," Per Stenstrom for "Architecture," and Jose Moreira for "Software." I greatly appreciate the efforts made by D.K. Panda and his team in putting together a tremendous, high-quality technical program, including a panel on the topic of "Is the MultiCore Roadmap Going to Live Up to Its Promises?" moderated by Per Stenstrom. I thank our panelists for participating: Michel Dubois, Tim Mattson, Kunle Olukotun, David Padua, and Marc Tremblay. I also thank our keynote and banquet speakers for contributing to the symposium: Christopher Johnson, Michael Flynn, Mark Seager, and Umesh Vazirani.

It has been my pleasure working with other members of the organizing committee. Among them, Alan Sussman, the Workshops Chair, worked tirelessly in putting together an excellent set of workshops-with the help of the Workshops Vice-Chair Yuanyuan Yang. I thank them, along with the organizers of each workshop, for the marvelous job they did in assembling a variety of emerging topics of interest to the parallel and distributed processing community. I also thank Sushil Prasad, the Tutorials Chair, for his efforts in selecting a timely tutorial of broad interest to the IPDPS audience. I especially thank Shoukat Ali, the Proceedings Chair, who did an outstanding job in pulling together the conference and workshop proceedings which includes the tutorial presentation. I acknowledge Nalini Venkatasubramanian and Gary Augusta, the Commercial Presentations and Exhibits Co-Chairs, along with Debbie Nielsen, for attracting commercial participation and sponsorship. I am also very grateful to the Publicity Coordinators Ioannis Chatzigiannakis, Bo Hong and Cho-Li Wang, and to the IPDPS Webmaster Anna Brown (www.mediagirl.com), for the major effort they put forth in publicizing this symposium. The support of Jie Wu, the General Vice-Chair, is also acknowledged.

I thank many other vital supporting cast members who, with their help and timely work, have made this symposium into the success that it is. Headliners in this list include Production Chair Sally Jelinek, Finance Chair Bill Pitts, Local Arrangements Chair Susamma Barua, and Steering Co-Chairs Viktor Prasanna and George Westrom. Without these key individuals, the process of running this symposium would not have gone nearly as smoothly. As has been echoed by my predecessors, these individuals shielded me from having to handle a myriad of important administrative details which require great competence and skill. I benefited greatly from their wealth of experience and unparalleled dedication. I especially thank Sally Jelinek for pitching in any number of ways above and beyond the call of her defined duties, particularly in her assistance with generating interest in commercial participation. I also greatly appreciate Viktor Prasanna

for being very generous with his time and useful guidance.

Last, I express my sincere gratitude to the steering committee who invited me to serve as General Chair and who provided the organizing committee useful direction throughout. This has been a very rewarding experience for all of us. I trust that you, too, will find this year's symposium to be an enriching experience, both informative and stimulating. Enjoy your time here at IPDPS'07 and indulge in the uniquenesses offered by this beautiful and exciting city of Long Beach!

IPDPS 2007 General Chair
Timothy Mark Pinkston
University of Southern California and National Science Foundation

Message from the Program Chair

Welcome to the 21st International Parallel and Distributed Processing Symposium (IPDPS '07). We have an excellent conference program featuring 109 technical papers (including four "best papers"), three keynote talks, one banquet talk and one panel.

A total of 419 papers were submitted for the conference from around the world. After submission, each paper was assigned to one or more of the Vice-Chairs. The Vice-Chairs then assigned each paper to three program committee members. To maintain a high quality of the reviews, the program committee members were responsible for reading the papers themselves and submitting the reviews. They were allowed to request help from additional reviewers, as needed. However, the responsibility for entering the final reviews and recommendations in EDAS was left solely to the program committee members.

A record number of 1234 reviews (an average of 2.95 reviews per paper) were collected from 100 program committee members. Before the physical program committee meeting, electronic discussions were held among the program committee members through EDAS to focus on papers with high variant reviews and generate consensus. The program committee meeting was held on December 8, 2006 in Columbus, Ohio. The meeting was attended by the Program Chair, all Vice Chairs, the General Chair and a number of Program Committee members. We deliberated on the papers over a period of nine hours and finally selected 109 papers (an acceptance rate of only 26%) for the final program. A set of papers were accepted under "shepherding" and several Vice Chairs and Program Committee members worked closely with the authors of these "shepherded" papers to improve their quality. Based on the recommendations by the Vice-Chairs, a discussion also took place during the program committee meeting to select the four "best papers," one from each track. Congratulations to the authors of all accepted papers including the "best papers."

The 109 papers have been organized into 27 sessions and one "Best Papers" session. In addition, there will be three keynote talks, one banquet talk and one panel. I am very grateful to our keynote and banquet speakers for accepting our invitations: Christopher Johnson, Michael Flynn, Mark Seager and Umesh Vazirani. I am also grateful to Per Stenstrom for heading the panel on "Is the Multi-Core Roadmap going to Live Up to its Promises?" and assembling an outstanding set of panelists: Michel Dubois, Tim Mattson, Kunle Olukotun, David Padua and Marc Tremblay. The keynote talks, banquet talk and the panel will try to complement the technical papers to provide an outstanding perspective on the state-of-art research challenges and solutions for the parallel and distributed computing field.

I am indebted to many individuals for contributing to the success of the technical program. First of all, I would like to thank all four Vice Chairs (Yves Robert - Algorithms, Srinivas Aluru - Applications, Per Stenstrom - Architectures and Jose Moreira - Software) for shouldering the major responsibilities in handling the submissions in their respective areas. My sincere thanks is extended to all 100 program committee members who accepted our invitations, served on the committee while handling a high load of submissions, and were able to turn-in high quality reviews on time. I would also like to thank an outstanding set of reviewers (their names appearing in the proceedings as a separate list) who extended help to the program committee members by providing high quality reviews. I would also like to thank all of the authors who submitted papers to the conference. Without help from all of these dedicated individuals, it would not have been possible for me to put together this outstanding program.

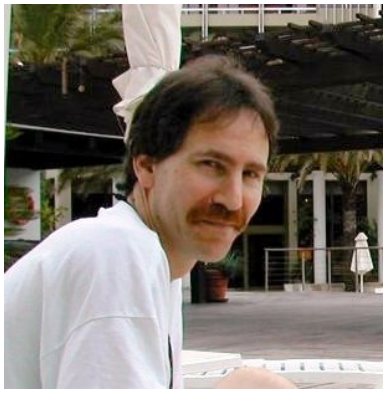
It was a great honor to be asked to perform the function of the Program Chair for this prestigious and leading conference in the field. I would like to extend my sincere thanks to Viktor Prasanna and George Westrom (Steering Committee Co-Chairs) and Timothy Pinkston (General Chair) for giving me this opportunity to serve the community and providing

appropriate advice, guidance, help, support and encouragement throughout the last year during many difficult phases of organizing this conference. It was also a great pleasure to work with a set of highly dedicated and experienced individuals: Jie Wu (General Vice Chair), Sally Jelinek (Production Chair), Shoukat Ali (Proceedings Chair), Bill Pitts (Finance Chair) and Susamma Barua (Local Arrangements Chair). Finally, I would like to express my sincere thanks to Carrie Casto of the Computer Science and Engineering Department at the Ohio State University for extending tremendous help to me during the last several months in handling the submissions and providing logistical support for the program committee meeting held in Columbus. My sincere thanks to all of these individuals. Their timely help and guidance on many issues made my life much easier.

I hope that you will enjoy the strong technical program of this conference while exploring beautiful Long Beach and California!!

IPDPS 2007 Program Chair
Dhabaleswar K. (DK) Panda
The Ohio State University

Message from the Workshops Chair



Welcome to IPDPS 2007 in Long Beach. This year's workshops program includes 21 workshops with a total of 333 papers. Many of the workshops have grown steadily in strength and several are now operating with parallel sessions or on multiple days. We are pleased to welcome two new workshops this year, one on Multithreaded Architectures and Applications and the other on Large-Scale, Volatile Desktop Grids. As always, we are looking for new workshop proposals for the next IPDPS.

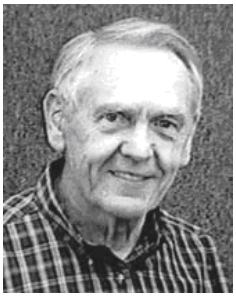
It is my great pleasure to work with such a diverse and energetic group of workshop organizers. It is impossible to over-emphasize our appreciation for all these people who invest so much time and energy to create and publicize the workshops, collect and review submissions, and then organize the programs. Their contribution to the community is enormous, and I urge every participant to personally thank these individuals for their hard work and dedication. I've also been fortunate to have the assistance of Yuanyuan Yang as my workshop vice chair again this year. Her effort, especially in dealing with EDAS system issues, has been most welcome.

In the IPDPS organization, I must first thank Shoukat Ali for his amazingly efficient management of the camera-ready paper submission process. Unless you've been on his end of this frantic endeavor, you can't begin to imagine what he and his helpers go through in the week that papers all pour in. Thanks to Sally Jelinek for her wide-ranging efforts - "production chair" doesn't begin to describe all of the jobs she tackles for this conference, including the workshops. Besides designing the proceedings and overseeing the proceedings production process, many of our workshop attendees can thank her for arranging invitation letters for visas. She will also be one of the smiling faces at the registration desk, so when you see her, be sure to say thanks. Susamma Barua has once again handled the very complicated local arrangements for the workshops, including arranging space, A/V needs, and refreshments. While you drink your coffee during a break, stop by the registration desk and offer your compliments. As usual, Bill Pitts has kept the financial wheels turning smoothly. Thanks also to Anna Brown for keeping the workshop web page and links up to date.

Long Beach is a wonderful location for a conference, and I hope that everyone has both great technical interactions at the workshops and conference, and a fine time taking advantage of all that Long Beach has to offer.

IPDPS 2007 Workshops Chair
Alan Sussman
University of Maryland, College Park

Message from the Steering Co-Chairs

We extend attendees at IPDPS 2007 a warm welcome back to Southern California. As part of the founding group for this conference - the Orange County chapter of the IEEE Computer Society - we have had the privilege to see their vision for an international symposium on parallel processing more than fulfilled. In addition to the Southern California organizers who have been involved since the beginning 21 years ago, there are others worldwide who have stayed the course and helped to build the technical excellence and volunteer spirit of the conference. We want to extend full credit and gratitude to these volunteers who share our pride in the special character of this event.

Each year, the general chair and program chair are charged with fashioning a technical program that reflects the state-of-the art in our field. It is also their job to guide the other technical chairs and support all of the volunteers involved in producing the conference. The IPDPS 2007 chairs - Timothy Pinkston and D.K. Panda - have teamed to muster all the elements of this year's technical program and to continue the record of excellence for this conference. Our thanks and congratulations to them as well as those of you who submitted research papers reflecting your important contributions to parallel and distributed processing technology; they are the lifeblood of the conference.

IPDPS has had the benefit of sustained involvement of committed volunteers, at both the organizational and technical levels. We have worked to build and refresh this corps of volunteers, so that our organization remains dynamic and responsive to developments in our field and the world generally. Those who step forward to volunteer know the hard work and investment of time that will be required, but we trust they have discovered the rewards both immediate and future of this experience.

Serving as technical organizers, at every level including program committee members and workshop organizers, is a challenging and learning experience which we hope supports professional objectives and networking goals of our community. We are always planning at least five years in advance for future conferences and are already committed to Miami in 2008 and Rome in 2009. The success of these events will depend upon attracting new and energetic volunteers. We invite participation by researchers new to the field and new to this event, so that we may continue to evolve and reinvent the conference.

We have looked forward to returning this year to our home base and hope that while you are here, you have an opportunity to take full advantage of all that Southern California has to offer.

IPDPS 2007 Steering Co-Chairs
Viktor Prasanna, University of Southern California
George Westrom, Discovery Science Center & FSEA

IPDPS 2007 Organization

General Chair

Timothy Mark Pinkston, University of Southern California and National Science Foundation

General Vice-Chair

Jie Wu, Florida Atlantic University, USA

Program Chair

Dhabaleswar K. (DK) Panda, The Ohio State University, USA

Steering Co-Chairs

Viktor K. Prasanna, University of Southern California
George Westrom, Discovery Science Center & FSEA

Steering Committee

K. Mani Chandy, California Institute of Technology, 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 M. Pinkston, Univeristy of Southern California and National Science Foundation, USA
Viktor K. Prasanna, 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
George B. Westrom, Discovery Science Center and Future Scientists & Engineers of America, USA
Jie Wu, Florida Atlantic University, USA

Workshops Committee

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

Tutorials Chair

Sushil Prasad, Georgia State University, USA

Commercial Presentations & Exhibits Co-Chairs

Nalini Venkatasubramanian, University of California at Irvine, USA
Gary Augusta, OCTANe, USA

Proceedings Chair

Shoukat Ali, University of Missouri-Rolla, USA

Finance Chair

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

Local Arrangements Chair

Susamma Barua, California State University, Fullerton, USA

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

Dhabaleswar K. (DK) Panda, The Ohio State University, USA

Program Vice-Chairs

Algorithms

Yves Robert, Ecole Normale Supérieure de Lyon, France

Applications

Srinivas Aluru, Iowa State University, USA

Architectures

Per Stenstrom, Chalmers University of Technology, Sweden

Software

Jose Moreira, IBM Thomas J. Watson Research Center, USA

Program Committee

Mikhail ATALLAH (Purdue U.) USA
David BADER (Georgia Inst. of Tech.) USA
Ioana BANICESCU (Mississippi State U.) USA
Olivier BEAUMONT (LaBRI Bordeaux) France
Michael BENDER (State U. New York at Stony Brook) USA
Ricardo BIANCHINI (Rutgers U.) USA
Gianfranco BILARDI (U. Padova) Italy
Angelos BILAS (U. Crete/FORTH) Greece
Alain BUI (U. Reims) France
John CARTER (U. Utah) USA
Umit CATALYUREK (Ohio State U.) USA
Alok CHOUDHARY (Northwestern U.) USA
Nikos CHRISOCHOIDES (College of William & Mary) USA
Almadena CHTCHELKANOVA (NSF) USA
Marcelo CINTRA (U. Edinburgh) UK
Andrea CLEMATIS (CNR Genoa) Italy
Toni CORTES (Barcelona Supercomputing Center) Spain
Chita DAS (Penn. State U.) USA
Bronis R. DE SUPINSKI (Lawrence Livermore National Lab.) USA
Erik DIRKX (Vrije U. Brussels) Belgium
Maria ELEFTHERIOU (IBM T.J. Watson Res. Center) USA
Robert ELSAESSER (U. Paderborn) Germany
Thomas FAHRINGER (U. Innsbruck) Austria
Akihiro FUJIWARA (Kyushu Institute of Technology) Japan
Rahul GARG (IBM India Research Lab) India
Maria Jesus GARZARAN (U. Illinois at Urbana-Champaign) USA
Michael GERNDT (Technical U. Munich) Germany
Vladimir GETOV (U. of Westminster) UK
Domingo GIMENEZ (U. Murcia) Spain
Michael T. GOODRICH (U. California) USA
Manimaran GOVINDARASU (Iowa State U.) USA
Håkan GRAHN (Blekinge Inst. Tech.) Sweden
Sandeep K.S. GUPTA (Arizona State U.) USA
John GUSTAFSON (Clearspeed Technology Inc.) USA
Mark HEINRICH (U. Central Florida) USA
Bruce HENDRICKSON (Sandia National Labs) USA
Adolfy HOISIE (Los Alamos National Lab.) USA
Bo HONG (Drexel U.) USA
Ananth KALYANARAMAN (Washington State U.) USA
Helen KARATZA (Aristotle U. of Thessaloniki) Greece
Hironori KASAHARA (Waseda U.) Japan
Manolis KATEVENIS (U. Crete/FORTH) Greece
Daniel S. KATZ (Louisiana State U. & JPL) USA
Stefanos KAXIRAS (U. of Patras) Greece
Paul H J KELLY (Imperial College London), UK
Suresh KOTHARI (Iowa State U.) USA
Kuan-Ching LI (Providence U.) Taiwan
Calvin LIN (U. of Texas at Austin) USA
Olav LYSNE (Oslo U.) Norway
Muthucumaru MAHESWARAN (McGill U.) Canada
Allen MALONY (U. of Oregon) USA
Fredrik MANNE (U. of Bergen) Norway

Pierre MANNEBACK (Faculte Polytechnique de Mons) Belgium
Tomàs MARGALEF (Autonomous U. Barcelona) Spain
Milo MARTIN (U. of Pennsylvania) USA
Jose MARTINEZ (Cornell U.) USA
Xavier MARTORELL (Technical U. of Catalunya) Spain
Pedro MEDEIROS (New U. Lisbon) Portugal
Celso MENDES (U. Illinois at Urbana-Champaign) USA
Samuel MIDKIFF (Purdue U.) USA
Edson Toshimi MIDORIKAWA (U. Sao Paulo) Brazil
Bernd MOHR (Research Centre Juelich) Germany
Andreas MOSHOVOS (U. Toronto) Canada
Rajeev MURALIDHAR (Intel) India
Kengo NAKAJIMA (U. Tokyo) Japan
David O'HALLARON (Carnegie Mellon U.) USA
Marcin PAPRZYCKI (SWPS and IBS PAN) Poland
Manish PARASHAR (Rutgers U.) USA
Franck PETIT (LARIA Amiens) France
Cynthia A. PHILLIPS (Sandia National Labs) USA
Alex POTHEN (Old Dominion U.) USA
Sushil K. PRASAD (Georgia State U.) USA
Padma RAGHAVAN (Pennsylvania State U.) USA
Soumyendu RAHA (Indian Institute of Science) India
Alex RAMIREZ (UPC Barcelona) Spain
Sanjay RANKA (U. of Florida) USA
Lawrence RAUCHWERGER (Texas A&M U.) USA
Jose RENAU (Univ. Santa Cruz) USA
P. SADAYAPPAN (Ohio State U.) USA
Yanos SAZEIDES (U. Cyprus) Cyprus
Bertil SCHMIDT (Nanyang Technological U.) Singapore
Martin SCHULZ (Lawrence Livermore National Lab.) USA
Hong SHEN (Manchester Metropolitan U.) UK
Yefim SHUF (IBM T.J. Watson Research Center) USA
Siang SONG (U. Sao Paulo) Brazil
Yong Ho SONG (Hanyang U.) Korea
Masha SOSONKINA (Ames Laboratory) USA
Leonel SOUSA (TU Lisbon) Portugal
Srikanta TIRTHAPURA (Iowa State U.) USA
Sivan TOLEDO (Tel-Aviv U.) Israel
Theo UNGERER (U. Augsburg) Germany
Sathish VADHIYAR (Indian Institute of Science) India
Stamatis VASSILIADIS (TU Delft) The Netherlands
Frédéric VIVIEN (INRIA) France
Biing-Feng WANG (Tsing Hua U.) Taiwan
Ramin YAHYAPOUR (U. of Dortmund) Germany
Sudhakar YALAMANCHILI (Georgia Inst. of Tech.), USA
Yuanyuan YANG (State U. New York at Stony Brook) USA
Craig ZILLES (Univ. Illinois at Urbana-Champaign) USA
Albert ZOMAYA (Univ. Sydney) Australia

IPDPS 2007 Technical Program

March 26-30, 2007– Long Beach, California, USA

Monday, March 26, 2007

Workshops 1-12

TCPP Presentation and Invited Speech
Reinventing Computing
Burton Smith, Technical Fellow, Advanced Strategies and Policy, Microsoft Corporation

Abstract: The many-core inflection point presents a new challenge for our industry, namely general-purpose parallel computing. Unless this challenge is met, the continued growth and importance of computing itself and of the businesses engaged in it are at risk. We must make parallel programming easier and more generally applicable than it is now, and build hardware and software that will execute arbitrary parallel programs on whatever scale of system the user has. The changes needed to accomplish this are significant and affect computer architecture, the entire software development tool chain, and the army of application developers that will rely on those tools to develop parallel applications. This talk will point out a few of the hard problems that face us and some prospects for addressing them.

Bio: Dr. Burton J. Smith, Technical Fellow for Microsoft Corporation, works with various groups within the company to help expand efforts in the areas of parallel and high performance computing. He reports directly to Craig Mundie, chief technical officer and senior vice president for Advanced Strategies and Policy.

Burton is recognized as an international leader in high performance computer architecture and programming languages for parallel computers. Before joining Microsoft, he served at Cray Inc., formerly Tera Computer Company, as chief scientist and a member of the board of directors from its inception in 1988 to December 2005, and was its chairman from 1988 to 1999. Prior to founding Tera Computer Company in 1988 Burton spent six years with Denelcor, Inc. and three years with the Institute for Defense Analyses. From 19701979 he taught at the Massachusetts Institute of Technology and the University of Colorado.

In 2003, Burton received the Seymour Cray Computing Engineering Award from the IEEE Computer Society and was elected to the National Academy of Engineering. He received the Eckert-Mauchly Award in 1991 given jointly by the Institute for Electrical and Electronic Engineers and the Association for Computing Machinery and was elected a fellow of each organization in 1994. Burton attended the University of New Mexico, where he earned a BSEE degree, and the Massachusetts Institute of Technology, where earned SM, EE, and Sc.D degrees.

Tuesday, March 27, 2007

Keynote Speech: Large-Scale Bioimaging and Visualization
Christopher Johnson, University of Utah

Abstract: The next decades will see an explosion in the use and the scope of medical imaging, fueled by advanced computing and visualization techniques. In my opinion, advanced, multimodal imaging and visualization techniques, powered by new computational methods, will change the face of biology and medicine and provide comprehensive views of the human body in progressively greater depth and detail. As the resolution of imaging devices continue to increase, image sizes grow accordingly. Multi-modal and/or longitudinal imaging studies result in large-scale data sets requiring parallel computing and visualization. In this presentation, I will discuss the state-of-the-art in large-scale biomedical imaging and visualization research, present examples of their vital roles in neuroscience, neurosurgery, radiology, and biology and discuss future challenges.

Bio: Professor Johnson directs the Scientific Computing and Imaging Institute at the University of Utah where he is a Distinguished Professor of Computer Science and holds faculty appointments in the Departments of Physics and Bioengineering. His research interests are in the areas of scientific computing and scientific visualization. Dr. Johnson founded the SCI research group in 1992, which has since grown to become the SCI Institute employing over 100 faculty, staff and students. Professor Johnson serves on several international journal editorial boards, as well as on advisory boards to several national and international research centers. Professor Johnson has received several awards, including the the

NSF Presidential Faculty Fellow (PFF) award from President Clinton in 1995 and the Governor's Medal for Science and Technology from Governor Michael Leavitt in 1999. In 2003 he received the Distinguished Professor Award from the University of Utah. In 2004 he was elected a Fellow of the American Institute for Medical and Biological Engineering (AIMBE) and in 2005 he was elected a Fellow of the American Association for the Advancement of Science (AAAS).

Session 1: Peer-to-Peer Algorithms

Session 2: Science, Finance and Combinatorial Applications

Session 3: Cluster and Server Architectures

Session 4: Software Support for Large Scale Scientific Computing

Session 5: Scheduling Algorithms

Session 6: Search, Text and Web Applications

Session 7: Processor Architecture

Session 8: Performance Analysis and Optimization

Session 9: Complexity of Algorithms

Session 10: Power and Energy Aware Computing

Session 11: Performance Modeling and Evaluation

Session 12: Middleware and Tools

Symposium Evening Tutorial: High-performance Computing Methods for Computational Genomics
Presenters: Srinivas Aluru, David A. Bader, and Ananth Kalyanaraman

Abstract: As biomolecular sequence data continue to be amassed at unprecedented rates, the design of effective computational methods and capabilities that can derive biologically significant information from them has become both increasingly challenging and imperative. In this tutorial, the audience will be first introduced to the different types of biomolecular sequence data and the wealth of information they encode. Following this technical grounding, high-performance computing approaches developed to address some of the most computationally challenging problems in genomics will be described. The contents will be presented in three parts: (i) In the first part, we will describe methods that were designed to query a sequence against a large sequence database. Two popular parallel approaches, mpiBLAST and ScalaBLAST, implementing the NCBI BLAST suite of programs will be described. (ii) Next, we will describe PaCE, which is a parallel DNA sequence clustering algorithm. As direct applications, we will discuss the clustering of large-scale Expressed Sequence Tag data and the assembly of complex genomes. (iii) Finally, we describe GRAPPA, which is a high-performance software suite developed for phylogenetic reconstruction of a collection of genomes or genes.

Throughout the tutorial, emphasis will be on both scalability and effectiveness in exploiting large-scale state-of-the-art supercomputing technologies. The intended audience are academic and industry researchers, educators, and/or commercial application developers, with a computational background. No background in biology is assumed.

Wednesday, March 28, 2007

Keynote Speech: Avoiding the Memory Bottleneck through Structured Arrays
Michael J. Flynn, Stanford University

Abstract: Basic to parallel program speedup is dealing with memory bandwidth requirements. One solution is an architectural arrangement to stream data across multiple processing elements before storing the result in memory. This MISD type of configuration provides multiple operations per data item fetched from memory. One realization of this streamed approach uses FPGAs. We'll discuss both the general memory problem and some results based on work at Maxeler using FPGAs for acceleration.

Bio: Michael Flynn is Senior Advisor to the Maxeler Corporation, an acceleration solutions company based in London. He received his Ph.D. from Purdue University and joined IBM working there for ten years in the areas of computer organization and design. He was design manager System 360 Model 91 Central Processing Unit. Between 1966 and 1974 Prof.

Flynn was a faculty member of Northwestern University and the Johns Hopkins University. From 1975 until 2000, he was a Professor of Electrical Engineering at Stanford University and served as the Director of the Computer Systems Laboratory from 1977 to 1983. He was founding chairman of both the ACM Special Interest Group on Computer Architecture and the IEEE Computer Society's Technical Committee on Computer Architecture. Prof. Flynn was the 1992 recipient of the ACM/IEEE Eckert-Mauchley Award for his technical contributions to computer and digital systems architecture. He was the 1995 recipient of the IEEE-CS Harry Goode Memorial Award in recognition of his outstanding contribution to the design and classification of computer architecture. In 1998 he received the Tesla Medal from the International Tesla Society (Belgrade), and an honorary Doctor of Science from Trinity College (University of Dublin), Ireland. He is the author of three books and over 250 technical papers, and he is also a fellow of the IEEE and the ACM.

Plenary Session: Best Papers

Session 13: Wireless, Adhoc and Sensor Algorithms

Session 14: Applications on Emerging Architectures

Session 15: Interconnection Networks

Session 16: Performance Prediction and Distributed Systems

IPDPS Panel: Is the Multi-Core Roadmap going to Live Up to its Promises?
Moderator: Per Stenstrom, Chalmers University of Technology
Panelists: Michel Dubois, USC; Tim Mattson, Intel; Kunle Olukotun, Stanford; David Padua, UIUC; and Marc Tremblay, Sun Microsystems

Abstract: Multi-cores are here to stay, whether we like it or not. With a quadrupling of the core count every three years, chips with hundreds of processor cores are projected in the next decade. The question is, how much of their computational power can be unleashed, what it will take to unleash it, and how best can research accelerate progress? Several decades of research in multiprocessing has not really made the case. On the other hand, now that coarse-grain parallelism seems to be our only hope and the computing landscape is arguably different, opportunities may arise. The following cross-cutting issues will be debated in this panel with the hope of distilling new avenues for parallelism exploitation:

- Is the computing landscape (technology, applications, and market) today sufficiently different to exploit multiprocessors from what it was in the past? If yes, in what sense? If not, why?
- Do we need more research in multiprocessing given past work? If yes, what are the biggest challenges? If not, state the reasons.
- Will progress in software/architecture make it possible to make sequential languages prevail? If yes, what are the top priorities in research to make that happen? If not, what are the visions for a parallel-language paradigm shift and what are the major challenges in software/architecture research to accelerate uptake in the programming community?
- Would multi-disciplinary research (across the applications, algorithms, software, and architecture areas) be a good way to accelerate developments? Then, what areas should interact more closely and with what goals in mind?

Banquet and Invited Speech
Why Peta-Scale is Different: An Ecosystem Approach to Predictive Scientific and Engineering Simulation
Mark Seager, Lawrence Livermore National Labs

Abstract: With the recent advent of 100s of teraFLOP/s-scale simulations capability at Lawrence Livermore National Laboratory and other sites, it has become clear that the scientific method has changed. This transition has taken us from theory and experiment to theory and experiment being tightly integrated by simulation. With the advent of peta-scale simulations on the horizon it is appropriate to take stock of the recent advances and to look forward to the coming wave of future systems.

In this talk we focus on some areas of science that open up with peta-scale systems and how this is VERY different from the science one can accomplish with a single workstation (giga-scale simulation). In actual fact, the science enabled by tera-scale and peta-scale systems require a whole new approach to the scientific method. One of the things we are starting to realize being at the leading edge of applying this new technology, is that with the coming onset of peta-scale simulations (systems, visualization, and applications) is that we may be headed for huge scientific breakthroughs enabled

and driven by simulation. This is not just hype (there is already plenty of that), we give specific examples of where these breakthroughs are may occur and why. Another major ramification of this scientific simulation transformation is that the development, deployment and gainful employment of tera-scale→peta-scale simulations requires a vastly different approach from a professor and a few graduate students writing a code, doing scaling studies and publishing a few papers. There is still a place for this type of research. Indeed, it is the development foundation of techniques employed in larger simulations. However, the real world problems that are now becoming tractable to solve with peta-scale simulations require a multi-disciplinary, multi-physics, multi-scale approach that is way beyond what a single researcher and graduate students can accomplish.

This scale of computation is driving fundamental changes in platform design, infrastructure and the methods by which applications are developed. In the area of platform development, highly scalable systems are being proposed by multiple vendors to achieve a sustained petaFLOP/s on real scientific simulations. These systems are characterized by massive numbers of cores, memory and corresponding huge power requirements. Simulation environments proposed for these systems rely on shared file systems, visualization, and data assessment assets tightly coupled with the platform. We discuss techniques that may be used to utilize these complex system of systems.

Bio: Dr. Seager received his B.S. Degree in Mathematics and Astrophysics at the University of New Mexico at Albuquerque in 1979 and received his PhD in Numerical Analysis from the University of Texas at Austin in 1984. Mark started working at Lawrence Livermore National Laboratory in 1983 and has been working in the field of parallel processing ever since. He manages the Platforms Program for the Advanced Simulation and Computing (ASC) Program at LLNL and has managed multiple vendor partnerships to successfully deploy architectures such as ASCI Blue Pacific (3.9 TF/s in 1998), ASCI White (12.3 TF/s in 2000) and Purple (100TF/s in 2005) and BlueGene/L (360 TF/s in 2005). In addition, Dr. Seager developed the LLNL Linux strategy and helped deploy multiple generations of leading edge clusters (MCR at 11.3 TF/s in 2002 and Thunder at 23 TF/s in 2004, Multiple Peloton SU clusters at 70TF/s in 2006). Dr. Seager is now focused on the challenges of peta-scale systems, simulation environments and applications development strategies.

Thursday, March 29, 2007

Keynote Speech: Quantum Physics and the Nature of Computation
Umesh Vazirani, University of California Berkeley

Abstract: Quantum physics is a fascinating area from a computational viewpoint. The features that make quantum systems prohibitively hard to simulate classically are precisely the aspects exploited by quantum computation to obtain exponential speedups over classical computers. In this talk I will survey our current understanding of the power (and limits) of quantum computers, and prospects for experimentally realizing them in the near future. I will also touch upon insights from quantum comuptation that have resulted in new classical algorithms for efficient simulation of certain important quantum systems.

Bio: Umesh Vazirani received his B.Tech in Computer Science from M.I.T. in 1981 and his PhD in Computer Science from U.C. Berkeley in 1985. He is on the faculty in Computer Science at U.C. Berkeley, where he is director of the Berkeley Quantum Computing Center, and holder of the Strauch Chair in Computer Science. Prof. Vazirani is a theoretician with broad interests in algorithms, complexity theory and novel models of computation. He has done seminal work in quantum computation and on the computational foundations of randomness. His books include "An Introduction to Computational Learning Theory" (with Michael Kearns, MIT Press, 1995), and "Algorithms" (with Sanjoy Dasgupta and Christos Papadimitriou, MIT press, 2006).

Session 17: Network Algorithms

Session 18: Peer-to-Peer Systems and Applications I

Session 19: Networks and Storage Systems

Session 20: Compiler Optimization and Software Environment

Session 21: Distributed Algorithms

Session 22: Peer-to-Peer Systems and Applications II

Session 23: Job Scheduling

Session 24: Fault Tolerance and Checkpointing

Session 25: Load Balancing Algorithms

Session 26: Distributed and Mobile Applications

Session 27: Algorithms for Parallel Execution

Commercial Track

New Developments in Intel Chipset Architecture
– *Presenters: Sivakumar Radhakrishnan and Sundaram Chinthamani, Intel*

The Impact of MultiCore Processors on HPC Cluster Design
– *Presenter: Marc Hamilton, Sun Microsystems*

Friday, March 30, 2007

Workshops 13-21

IPDPS 2007 Reviewers

Tariq Abdullah
Carmelo Acosta
Ali-Reza Adl-Tabatabai
Mauricio Alvarez
Rouzbeh Amini
Claudio L. Amorim
Jonathan Appavoo
Ernest Artiaga
Clement R. Attanasio
Faruk Bagci
Kevin Baker
Amnon Barak
Jorge Barbosa
Lee Baugh
Alan Bivens
Yvonne Bleischwitz
Carlos Boneti
Christian Boulinier
Anu Bourgeois
Jeremy Bradley
Greg Bronevestsky
Ashley Brown
Franck Butelle
Bogdan Carbunar
Nick Carter
Antonio Carzaniga
Francisco J. Cazorla
Brad Chamberlain
Walter Chang
Zhijiang Chang
Songqing Chen
Trishul Chilimbi
Avery Ching
Sung-Eun Choi
Nikos Chrysos
Catalin Ciobanu
Paul Coddington
Kenin Coloma
William Cook
Julita Corbalan
Ajoy Kumar Datta
Kei Davis
Sylvie Delaët
Gilles Dequen
Karen Devine
Stéphane Devismes
Akshaye Dhawan
Yoann Dieudonné
Cheng Ding
Fred Douglis
Jose Duato
Cezary Dubnicki
Bertrand Ducourthial
Dominic Dumrauf
Schahram Dustdar
Yaakoub El-Khamra
Tom Engelsiepen
Brett Estrade
Rainer Feldmann
Robert A. Fiedler
Tony Field
Olivier Flauzac
Stefan Freitag
Martin Gairing
Jorge Garcia
Georgi Gaydadjiev
Assefaw Gebremedhin
Isaac Gelado
George Georgakopoulos
Roberto Gioiosa
Daniel Gmach
Karl M. Göschka
Maria Gradinariu
Christian Grimme
Sven Grothklags
Damrong Guoy
Audun Fosselie Hansen
Uli Harder
Eric Horlait
Mark Hulber
Cruz Izu
Christophe Jaillet
Ricardo Jimenez-Peris
Sven Karlson
Irit Katriel
Eli Katsiri
Kamil Kedzierski
Alfons Kemper
Ralf Klasing
Florian Kluge
Christos Kozyrakis
Michael Krajecki
Amund Kvalbein
Charles Lakos
Zhou Lei
Joachim Lepping
Kin Leung
Qun Li
Wei-Keng Liao
Beng-Hong Lim
Bruce Lowekamp
Xiaosong Ma
Vikram Makhija
Srilaxmi Malladi
Biju T. Maniampadavathu
Manolis Marazakis
Pedro Marcuello
Daniel Meister
Henning Meyerhenke
Lotfi Mhamdi
Maged Michael
Eiji Miyano
Jean-Frédéric Myoupo
Kiran Nagaraja
Vijay K. Naik
Jeff Napper
Ramanathan Narayanan
Naveen Neelakantam
Thu Nguyen
Dimitris Nikolopoulos
Florent Nolot
Nils Agne Nordbotten
Alison Norman
Adam J. Oliner
Fabio Oliveira
Fukuhito Ooshita
Beatriz Otero
Scott Pakin
Jairo Panetta
Alexander Papaspyrou
George Passas
Chao Peng
Miquel Pericas
Andreas Pietzowski
Eduardo Pinheiro
Evaggelia Pitoura
Dionisis Pnevmatikatos
Cheryl Pope
Mario Porrmann
Behnaz Pourebrahimi
Christoph von Praun
Jon Preston
Daji Qiao
Sven-Arne Reinemo
Nicholas Riley
Irina Rish
Rob Ross
Yaoping Ruan
Gudula Ruenger
Matthew Sackman

Bratin Saha
Ramendra Sahoo
Rizos Sakellariou
Pierre Salverda
Friman Sanchez
Jose Carlos Sancho
Benjamin Satzger
Thomas Sauerwald
Lars Schley
Florian Schoppmann
David Semé
Yingpeng Shang
Michael Sheng
Kamana Sigdel
Oliver Sinnen
Tor Skeie
Thomas Sodring
Devan Sohier
Ioannis Sourdis
Olivier Soyez
Evan Speight
Bogdan Spinean
Radu Stefan
Trond Steihaug
Paul Stodghill
Richard Strelitz
Gong Su
Riky Subrata
Alexandre Tabbal
Takeyuki Tamura
Dimitris Theodoropoulos
Mathew S. Thoennes
Karsten Tiemann
Suzuki Tomoko
Jordi Torres
John M. Tracey
Tobias Tscheuschner
Mayank Tyagi
Xavi Verdu
Vincent Villain
Spyros Voulgaris
Chen Wang
Haining Wang
Hsiao-Hsi Wang
Cathy Xia
Qin Xin
Shinichi Yamagiwa
Emmanuel Yashchin
Hao Yu
Haibo Zhang
Yanyong Zhang

Session 1

Peer-to-Peer Algorithms

VoroNet: A scalable object network based on Voronoi tessellations

Olivier Beaumont[1], Anne-Marie Kermarrec[2], Loris Marchal[3] and Etienne Riviere[4]

[1]*LaBRI/INRIA*
Bordeaux, France
obeaumon@labri.fr

[2]*IRISA/INRIA*
Rennes, France
akermarr@irisa.fr

[3]*ENS Lyon*
Lyon, France
loris.marchal@ens-lyon.fr

[4]*IRISA/Universite de Rennes 1*
Rennes, France
eriviere@irisa.fr

In this paper, we propose the design of VoroNet, an object based peer to peer overlay network relying on Voronoi tessellations, along with its theoretical analysis and experimental evaluation. VoroNet differs from previous overlay networks in that peers are application objects themselves and get identifiers reflecting the semantics of the application instead of relying on hashing functions. This enables a scalable support for efficient search in large collections of data. In VoroNet, objects are organized in an attribute space according to a Voronoi diagram. VoroNet is inspired from the Kleinbergs small-world model where each peer gets connected to close neighbours and maintains an additional pointer to a long-range neighbour. VoroNet improves upon the original proposal as it deals with general object topologies and therefore copes with skewed data distributions. We show that VoroNet can be built and maintained in a fully decentralized way. The theoretical analysis of the system proves that routing in VoroNet can be achieved in a poly-logarithmic number of hops in the size of the system. The analysis is fully confirmed by our experimental evaluation by simulation.

Almost Peer-to-Peer Clock Synchronization

Ahmed Sobeih[1], Michel Hack[2], Zhen Liu[2] and Li Zhang[2]

[1]*Department of Computer Science*
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
sobeih@uiuc.edu

[2]*IBM T.J. Watson Research Center*
Hawthorne, NY 10532, USA
{hack, zhenl, zhangli}@us.ibm.com

In this paper, an almost peer-to-peer (AP2P) clock synchronization protocol is proposed. AP2P is almost peer-to-peer in the sense that it provides the desirable features of a purely hierarchical (client/server) clock synchronization protocol while avoiding the undesirable consequences of a purely peer-to-peer one. In AP2P, a unique node is elected as a leader in a distributed manner. Each non-leader node adjusts its clock rate based on message exchanges with its neighbors, taking into consideration that neighbors that are closer to the leader have more effect on the adjustment than the neighbors that are further away from the leader. We compare the performance of AP2P with that of the Server Time Protocol (STP), which is a purely hierarchical clock synchronization protocol. Simulation results, which have been conducted on several network topologies, have shown that AP2P can provide a clock synchronization accuracy that is indistinguishable from that of STP. Furthermore, AP2P is more fault-tolerant because it can recover from certain types of failures that STP cannot recover from.

Locality-Aware Consistency Maintenance for Heterogeneous P2P Systems

Zhenyu Li[1,2], Gaogang Xie[1,3] and Zhongcheng Li[1]

[1]*Institute of Computing Technology*
Chinese Academy of Sciences
Beijing, China
{*zyli, xie, zcli*}*@ict.ac.cn*

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

[3]*INRIA-Rocquencourt*
LE CHESNAY, France

Replication and caching have been deployed widely in current P2P systems. In update-allowed P2P systems, a consistency maintenance mechanism is strongly demanded. Several solutions have been proposed to maintain the consistency of P2P systems. However, they either use too much redundant update messages, or ignore the heterogeneity nature of P2P systems. Moreover, they propagate updated contents on a locality-ignorant structure, which could consume unnecessary backbone bandwidth and delay the convergence of consistency maintenance. This paper presents a locality-aware consistency maintenance scheme for heterogeneous P2P systems. Taking the heterogeneity nature, we form the replica nodes into a locality-aware hierarchical structure: the upper layer is DHT-based and a node in the lower layer attaches to a physically close node in the upper layer. An efficient update tree is built dynamically upon the upper layer to propagate the updated contents. Theoretical analyses and simulation results demonstrate the effectiveness of our scheme. Specially, experiment results show that, compared with gossip-based scheme, our approach reduces the cost by about one order of magnitude.

Benefits of Targeting in Trusted Gossiping for Peer-to-Peer Information Sharing

Arindam Mitra[1] and Muthucumaru Maheswaran[2]

[1]*Dept. Computer Science*
University of Manitoba
Winnipeg, Manitoba, Canada
arindam@cs.umanitoba.ca

[2]*School of Computer Science*
McGill University
Montreal, Quebec, Canada
maheswar@cs.mcgill.ca

In a recent study, we proposed a trusted gossip protocol for rumor resistant information sharing in peer-to-peer networks. While trust aware gossiping significantly reduced the rumor spread on the network, we observed that the random message spraying in trusted gossip creates too many redundant messages increasing the message overhead and error rate. In this paper, we propose a message targeting scheme that can significantly improve the performance of the trusted gossip. Our targeting scheme can be easily implemented in a social network setting. We performed large-scale simulations using traces collected from the Flickr social network and other data sets to estimate the performance of targeting in trusted gossip. Our experiments show that significant performance gains can be achieved.

Session 2

Science, Finance and Combinatorial Applications

Building the Tree of Life on Terascale Systems

Xizhou Feng[1], Kirk W. Cameron[1], Carlos P. Sosa[2] and Brian Smith[2]

[1]*Computer Science*
Virginia Tech
Blacksburg, VA, USA
{*fengx, cameron*}*@cs.vt.edu*

[2]*BlueGene/L Development*
IBM
Rochester, MN, USA
{*cpsosa, smithbr*}*@us.ibm.com*

Bayesian phylogenetic inference is an important alternative to maximum likelihood-based phylogenetic method. However, inferring large trees using the Bayesian approach is computationally demandingrequiring huge amounts of memory and months of computational time. With a combination of novel parallel algorithms and latest system technology, terascale phylogenetic tools will provide biologists the computational power necessary to conduct experiments on very large dataset, and thus aid construction of the tree of life.

In this work we evaluate the performance of PBPI, a parallel application that reconstructs phylogenetic trees using MCMC-based Bayesian methods, on two terascale systems, Blue Gene/L at IBM Rochester and System X at Virginia Tech. Our results confirm that for a benchmark dataset with 218 taxa and 10000 characters, PBPI can achieve linear speedup on 1024 or more processors for both systems.

Inverse Space-Filling Curve Partitioning of a Global Ocean Model

John M. Dennis

Computational & Information Systems Laboratory
National Center for Atmospheric Research
Boulder, CO, USA
dennis@ucar.edu

In this paper, we describe how inverse space-filling curve partitioning is used to increase the simulation rate of a global ocean model. Space-filling curve partitioning allows for the elimination of load imbalance in the computational grid due to land points. Improved load balance combined with code modifications within the conjugate gradient solver significantly increase the simulation rate of the Parallel Ocean Program at high resolution. The simulation rate for a high resolution model nearly doubled from 4.0 to 7.9 simulated years per day on 28,972 IBM Blue Gene/L processors. We also demonstrate that our techniques increase the simulation rate on 7545 Cray XT3 processors from 6.3 to 8.1 simulated years per day. Our results demonstrate how minor code modifications can have significant impact on resulting performance for very large processor counts.

A Parallel Workflow for Real-time Correlation and Clustering of High-Frequency Stock Market Data

Camilo Rostoker, Alan Wagner and Holger Hoos

Department of Computer Science
University of British Columbia
Vancouver, BC, Canada
{*rostokec, wagner, hoos*}*@cs.ubc.ca*

We investigate the design and implementation of a parallel workflow environment targeted towards the financial industry. The system performs real-time correlation analysis and clustering to identify trends within streaming high-frequency intra-day trading data. Our system utilizes state-of-the-art methods to optimize the delivery of computationally-expensive real-time stock market data analysis, with direct applications in automated/algorithmic trading as well as knowledge discovery in high-throughput electronic exchanges. This paper describes the design of the system including the key online parallel algorithms for robust correlation calculation and clique-based clustering using stochastic local search. We evaluate the performance and scalability of the system, followed by a preliminary analysis of the results using data from the Toronto Stock Exchange.

A Grid-enabled Branch and Bound Algorithm for Solving Challenging Combinatorial Optimization Problems

Mohand Mezmaz, Nouredine Melab and El-Ghazali Talbi

Laboratoire d'Informatique Fondamentale de Lille
Université des Sciences et Technologies de Lille
Villeneuve d'Ascq, France
{*mezmaz, melab, talbi*}*@lifl.fr*

Solving optimally large instances of combinatorial optimization problems requires a huge amount of computational resources. In this paper, we propose an adaptation of the parallel Branch and Bound algorithm for computational grids. Such gridification is based on new ways to efficiently deal with some crucial issues, mainly dynamic adaptive load balancing, fault tolerance, global information sharing and termination detection of the algorithm. A new efficient coding of the work units (search sub-trees) distributed during the exploration of the search tree is proposed to optimize the involved communications. The algorithm has been implemented following a large scale idle time stealing paradigm(Farmer-Worker). It has been experimented on a Flow-Shop problem instance (Ta056) that has never been optimally solved. The new algorithm allowed to realize a success story as the optimal solution has been found with proof of optimality, within 25 days using about 1900 processors belonging to 9 Nation-wide distinct clusters (administration domains). During the resolution, the worker processors were exploited with an average of 97% while the farmer processor was exploited only 1.7% of the time. These two rates are good indicators on the efficiency of the proposed approach and its scalability.

Session 3

Cluster and Server Architectures

MultiEdge: An Edge-based Communication Subsystem for Scalable Commodity Servers

Sven Karlsson[1], Stavros Passas[2], George Kotsis[2] and Angelos Bilas[2]

[1]*FORTH-ICS*
Heraklion, Greece
svenka@ics.forth.gr

[2]*FORTH-ICS & Department of Computer Science,*
University of Crete
Heraklion, Greece
{stabat, kotsis, bilas}@ics.forth.gr

At the core of contemporary high performance computer systems is the communication infrastructure. For this reason, there has been a lot of work on providing low-latency, high-bandwidth communication subsystems for clusters. In this paper, we introduce *MultiEdge*, a connection oriented communication system designed for high-speed commodity hardware. *MultiEdge* provides support for end-to-end flow-control, ordering, and reliable transmission. It transparently supports multiple physical links within a single connection.

We use *MultiEdge* to examine the behavior of edge-based protocols using both micro-benchmarks and real-life shared memory applications. Our results show that *MultiEdge* is able to deliver about 88% of the nominal link throughput with a single 10-GBit/s link and more than 95% with multiple 1-GBit/s links. Our application results show that performing all of the communication protocol at the edge does not seem to cause any degradation in performance.

Efficient Block Device Sharing over Myrinet with Memory Bypass

Evangelos Koukis[1] and Nectarios Koziris[2]

[1]*Computing Systems Laboratory*
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
vkoukis@cslab.ece.ntua.gr

[2]*Computing Systems Laboratory*
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
nkoziris@cslab.ece.ntua.gr

Efficient sharing of block devices over an interconnection network is an important step in deploying a shared-disk parallel filesystem on a cluster of SMPs. In this paper we present gmbock, a client/server system for network sharing of storage devices over Myrinet, which uses an optimized data path in order to transfer data directly from the storage medium to the NIC, bypassing the host CPU and main memory bus. Its design enhances existing programming abstractions, combining the user level networking characteristics of Myrinet with Linux's virtual memory infrastructure, in order to construct the data path in a way that is independent of the type of block device used. Experimental evaluation of a prototype system shows that remote I/O bandwidth can improve up to 36.5%, compared to an RDMA-based implementation. Moreover, interference on the main memory bus of the host is minimized, leading to an up to 41% improvement in the execution time of memory-intensive applications.

Achieving Reliable Parallel Performance in a VoD Storage Server Using Randomization and Replication

Yung Ryn Choe and Vijay S. Pai

Electrical and Computer Engineering
Purdue University
West Lafayette, IN, USA
{*yung, vpai*}*@purdue.edu*

This paper investigates randomization and replication as strategies to achieve reliable performance in disk arrays targeted for video-on-demand (VoD) workloads. A disk array can provide high aggregate throughput, but only if the server can effectively balance the load on the disks. Such load balance is complicated by two key factors: workload hotspots caused by differences in popularity among media streams, and "fail-stutter" faults that arise when the performance of one or more devices drops below expectations due to manufacturing variations, hardware problems, or geometry-related variations.

This paper focuses on the random duplicate assignment (RDA) data allocation policy which places each data block on two disks chosen at random, independent of other blocks in the same media stream or other streams. This strategy is compared to traditional single-disk file allocation, disk striping (RAID-0), disk mirroring (RAID-1), and randomization without duplication. The various allocation schemes are implemented and tested using a prototype VoD server with 2 dual-core Opteron processors, 8 SATA disks, and 4 Gigabit Ethernet interfaces running the Linux 2.6 kernel. The results indicate that combining randomization and replication allows RDA to effectively tolerate both workload hotspots and fail-stutter faults better than previous schemes.

A Cost-Effective, High Bandwidth Server I/O network Architecture for Cluster Systems

Hsing-Bung Chen, Gary Grider and Parks Fields

Los Alamos National LAB
Los Alamos, New Mexico, USA
{*hbchen, ggrider, parks*}*@lanl.gov*

In this paper we present a cost-effective, high bandwidth server I/O network architecture, named PaScal (Parallel and Scalable). We use the PaScal server I/O network to support data-intensive scientific applications running on very large-scale Linux clusters. PaScal server I/O network architecture provides (1) bi-level data transfer network by combining high speed interconnects for computing Inter-Process Communication (IPC) requirements and low-cost Gigabit Ethernet interconnect for global IP based storage/file access, (2) bandwidth on demand I/O network architecture without re-wiring and reconfiguring the system, (3) Multi-path routing scheme, (4) reliability improvement through reducing large number of network components in server I/O network, and (5) global storage/file systems support in heterogeneous multi-cluster and Grids environments. We have compared the PaScal server I/O network architecture with the Federated server I/O network architecture (FESIO). Concurrent MPI-I/O performance testing results and deployment cost comparison demonstrate that the PaScal server I/O network architecture can outperform the FESIO network architecture in many categories: cost-effectiveness, scalability, and manageability and ease of large-scale I/O network.

Session 4

Software Support for Large Scale Scientific Computing

Babel Remote Method Invocation

Gary Kumfert, James Leek and Thomas Epperly

Lawrence Livermore National Laboratory
Livermore, CA, USA
{*kumfert, leek2, tepperly*}*@llnl.gov*

Babel is a high-performance, n-way language interoperability tool for the HPC community that now includes support for distributed computing via Remote Method Invocation (RMI). We describe the design and implementation of Babel RMI, including its specification in our Scientific Interface Definition Language (SIDL), modifications to Babel's code generators, and support for third-party wire protocols. Babel RMI's programming model consistency, functional capabilities, and runtime performance are compared in context with COM, CORBA, Grid/Web Services, and Java RMI. Babel RMI's current features and performance uniquely recommend it for "short-haul" distributed computing within a machine room or single cluster. We describe the experience of some early adopters who use Babel RMI to couple and coordinate multiple MPI jobs on a single cluster to perform multiscale material science calculations.

Nonuniformly Communicating Noncontiguous Data: A Case Study with PETSc and MPI

Pavan Balaji, Darius Buntinas, Satish Balay, Barry Smith, Rajeev Thakur and William Gropp

Mathematics and Computer Science
Argonne National Laboratory
Argonne, IL, USA
{*balaji, buntinas, balay, bsmith, thakur, gropp*}*@mcs.anl.gov*

Due to the complexity associated with developing parallel applications, scientists and engineers rely on high-level software libraries such as PETSc, ScaLAPACK and PESSL to ease this task. Such libraries assist developers by providing abstractions for mathematical operations, data representation and management of parallel layouts of the data, while internally using communication libraries such as MPI and PVM. With high-level libraries managing data layout and communication internally, it can be expected that they organize application data suitably for performing the library operations optimally. However, this places additional overhead on the underlying communication library by making the data layout noncontiguous in memory and communication volumes (data transferred by a process to each of the other processes) nonuniform. In this paper, we analyze the overheads associated with these two aspects (noncontiguous data layouts and nonuniform communication volumes) in the context of the PETSc software toolkit over the MPI communication library. We describe the issues with the current approaches used by MPICH2 (an implementation of MPI), propose different approaches to handle these issues and evaluate these approaches with micro-benchmarks as well as an application over the PETSc software library. Our experimental results demonstrate close to an order of magnitude improvement in the performance of a 3-D Laplacian multi-grid solver application when evaluated on a 128 processor cluster.

CCA-LISI: On Designing A CCA Parallel Sparse Linear Solver Interface

Fang (cherry) Liu and Randall Bramley

Computer Science Department
Indiana University
Bloomington, IN, U.S.A
{*fangliu, bramley*}*@cs.indiana.edu*

Sparse linear solvers account for much of the execution time in many high-performance computing (HPC) applications, and not every solver will work on all problems. Hence choosing a suitable solver is crucial step for an efficient application. Unfortunately, the best linear solver cannot be determined in the early stage of application development. Experiments on finding best suitable solver require a plug and play mechanism.

This work is part of the Common Component Architecture (CCA) effort on designing common interface among various parallel high performance linear solver libraries, hence componenizing them and enabling dynamical switching. The implementation of this interface provides a CCA toolkit and is reusable under CCA-compliant framework such as Ccaffeine.

Optimizing Distributed Application Performance Using Dynamic Grid Topology-Aware Load Balancing

Gregory A. Koenig and Laxmikant V. Kale

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois, United States
{*koenig, kale*}*@cs.uiuc.edu*

Grid computing offers a model for solving large-scale scientific problems by uniting computational resources owned by multiple organizations to form a single cohesive resource for the duration of individual jobs. Despite the appeal of using Grid computing to solve large problems, its use has been hindered by the challenges involved in developing applications that can run efficiently in Grid environments. One substantial obstacle to deploying Grid applications across geographically distributed resources is cross-site latency. While certain classes of applications, such as master-slave style or functional decomposition type applications, lend themselves well to running in Grid environments due to inherent latency tolerance, other classes of applications, such as tightly-coupled applications in which each processor regularly communicates with its neighboring processors, represent a significant challenge to deployment on Grids.

In this paper, we present a dynamic load balancing technique for Grid applications based on graph partitioning. This technique exploits knowledge of the topology of the Grid environment to partition the computation's communication graph in such a way as to reduce the volume of cross-site communication, thus improving the performance of tightly-coupled applications that are co-allocated across distributed resources. Our technique is particularly well suited to codes from disciplines like molecular dynamics or cosmology due to the non-uniform structure of communication in these types of applications. We evaluate the effectiveness of our technique when used to optimize the execution of a tightly-coupled classical molecular dynamics code called LeanMD deployed in a Grid environment.

Session 5

Scheduling Algorithms

On the Design of Online Scheduling Algorithms for Advance Reservations and QoS in Grids

Claris Castillo, George N. Rouskas and Khaled Harfoush

Department of Computer Science
North Carolina State University
Raleigh, NC, USA
{*ccastil, rouskas*}*@ncsu.edu, harfoush@csc.ncsu.edu*

We consider the problem of providing QoS guarantees to Grid users through advance reservation of resources. Advance reservation mechanisms provide the ability to allocate resources to users based on agreed-upon QoS requirements and increase the predictability of a Grid system, yet incorporating such mechanisms into current Grid environments has proven to be a challenging task due to the resulting resource fragmentation. We use concepts from computational geometry to present a framework for tackling the resource fragmentation, and for formulating a suite of scheduling strategies. We also develop efficient implementations of the scheduling algorithms that scale to large Grids. We conduct a comprehensive performance evaluation study using simulation, and we present numerical results to demonstrate that our strategies perform well across several metrics that reflect both user- and system-specific goals. Our main contribution is a timely, practical, and efficient solution to the problem of scheduling resources in emerging on-demand computing environments.

Reconfigurable Resource Scheduling with Variable Delay Bounds

Charles Gregory Plaxton, Yu Sun, Mitul Tiwari and Harrick Vin

Department of Computer Science
The University of Texas at Austin
Austin, TX, USA
{*plaxton, sunyu, mitult, vin*}*@cs.utexas.edu*

Certain emerging network applications involve dynamically allocating shared resources to a variety of services to provide QoS guarantees for each service. Motivated by such applications, we address the following online scheduling problem belonging to the recently introduced class of reconfigurable resource scheduling problems: unit jobs of different categories arrive over time and need to be completed within category-specific delay bounds, or else they are dropped at a unit drop cost; processors can be reconfigured to process jobs of a certain category at a fixed reconfiguration cost; the goal is to minimize the total cost. We study this problem in the framework of competitive analysis. Through a novel combination of the EDF and LRU scheduling principles, we obtain an online algorithm that is constant competitive when given a constant factor resource advantage over an optimal offline algorithm.

A Strategyproof Mechanism for Scheduling Divisible Loads in Linear Networks

Thomas E. Carroll and Daniel Grosu

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

In this paper we augment DLT (Divisible Load Theory) with incentives such that it is beneficial for processors to report their true processing capacity and compute their assignments at full capacity. We propose a strategyproof mechanism with verification for scheduling divisible loads in linear networks with boundary load origination. The mechanism provides incentives to processors for reporting deviants. The deviants are penalized which abates their willingness to deviate in the first place. We prove that the mechanism is strategyproof and satisfies the voluntary participation condition.

Scheduling in the $\mathcal{Z}$-Polyhedral Model

Gautam[1,2], Daegon Kim[1] and Sanjay Rajopadhye[1]

[1]*Computer Science Department*
Colorado State University
Fort Collins, CO, USA
{*gautam, kim*}*@cs.colostate.edu,*
Sanjay.Rajopadhye@colostate.edu

[2]*IRISA*
Universite de Rennes I
Rennes, France

The polyhedral model is extensively used for analyses and transformations of regular loop programs, one of the most important being automatic parallelization. The model, however, is limited in expressivity and the need for the generalization to more general class of programs has been widely known. Analyses and transformations in the polyhedral model rely on certain closure properties. Recently, these closure properties were extended to programs where variables may be defined over unions of $\mathcal{Z}$-polyhedra which are the intersection of polyhedra and lattices.

We present the scheduling analysis for the automatic parallelization of programs in the $\mathcal{Z}$-polyhedral model, and obtain multidimensional schedules through an ILP formulation that minimizes latency. The resultant schedule can then be used to construct a space-time transformation to obtain an equivalent program in the $\mathcal{Z}$-polyhedral model.

Session 6

Search, Text and Web Applications

A Landmark-based Index Architecture for General Similarity Search in Peer-to-Peer Networks

Xiaoyu Yang[1] and Yiming Hu[2]

[1]Dept. of Electrical & Computer Engineering and Computer Science
University of Cincinnati
Cincinnati, OH, USA
yangxu@ececs.uc.edu

[2]Dept. of Electrical & Computer Engineering and Computer Science
University of Cincinnati
Cincinnati, OH, USA
yhu@ececs.uc.edu

The indexing of complex data and similarity search plays an important role in many application areas. Traditional centralized index structure can not scale with the rapid proliferation of data volume. In this paper, we propose a scalable index architecture built on top of distributed hash tables (DHT), to support similarity search in the general metric space. Based on efficient space mapping and query routing mechanisms, our architecture can provide a general platform to support arbitrary number of indexes on different data types. Significantly, it does not need to generate or maintain any search trees. Instead, the embedded trees in the underlying distributed hash tables are exploited to deliver queries. To deal with skewed data distribution, we also provide load-balancing mechanisms to ensure that no node in the system is unduly loaded. The performance of the proposed design is evaluated through simulations with a variety of metrics. The experimental results demonstrate that out approach can efficiently solve similarity query at a low cost.

Optimized Inverted List Assignment in Distributed Search Engine Architectures

Jiangong Zhang and Torsten Suel

CIS Department
Polytechnic University
Brooklyn, NY, USA
zjg@cis.poly.edu, suel@poly.edu

We study efficient query processing in distributed web search engines with global index organization. The main performance bottleneck in this case is due to the large amount of index data that is exchanged between nodes during the processing of a query, and previous work has proposed several techniques for significantly reducing this cost. We describe an approach that provides substantial additional improvement over previous techniques. In particular, we analyze search engine query traces in order to optimize the assignment of index data to the nodes in the system, such that terms frequently occurring together in queries are also often collocated on the same node. Our experiments show that in return for a modest factor increase in storage space, we can achieve a reduction in communication cost of an order of magnitude over the previous best techniques.

Scalable Visual Analytics of Massive Textual Datasets

Manojkumar Krishnan, Shawn Bohn, Wendy Cowley, Vernon Crow and Jarek Nieplocha

Computational Sciences and Mathematics Division
Pacific Northwest Lational Laboratory
Richland, WA, USA
{*manoj, shawn.bohn, wendy, vern.crow, jarek.nieplocha*}*@pnl.gov*

This paper describes the first scalable implementation of a text processing engine used in visual analytics tools. These tools aid information analysts in interacting with and understanding large textual information content through visual interfaces. By developing a parallel implementation of the text processing engine, we enabled visual analytics tools to exploit cluster architectures and handle massive datasets. The paper describes key elements of our parallelization approach and demonstrates virtually linear scaling when processing multi-gigabyte data sets such as Pubmed. This approach enables interactive analysis of large datasets beyond capabilities of existing state-of-the art visual analytics tools.

Spam-Resilient Web Rankings via Influence Throttling

James Caverlee, Steve Webb and Ling Liu

College of Computing
Georgia Institute of Technology
Atlanta, GA, USA
{*caverlee, webb, lingliu*}*@cc.gatech.edu*

Web search is one of the most critical applications for managing the massive amount of distributed Web content. Due to the overwhelming reliance on Web search, there is a rise in efforts to manipulate (or spam) Web search engines. In this paper, we develop a spam-resilient ranking model that promotes a source-based view of the Web. One of the most salient features of our spam-resilient ranking algorithm is the concept of influence throttling. We show how to utilize influence throttling to counter Web spam that aims at manipulating link-based ranking systems, especially PageRank-like systems. Through formal analysis and experimental evaluation, we show the effectiveness and robustness of our spam-resilient ranking model in comparison with existing Web algorithms such as PageRank.

Session 7

Processor Architecture

Conserving Memory Bandwidth in Chip Multiprocessors with Runahead Execution

Martin Karlsson and Erik Hagersten

Department of Information Technology
Uppsala University
Uppsala, Sweden
{*martin.karlsson, erik.hagersten*}*@it.uu.se*

The introduction of chip multiprocessors (CMPs) presents new challenges and trade-offs to computer architects. Architects must now strike a balance between the number of cores per chip versus the amount of on-chip cache and the cost-efficient amount of pin bandwidth. Technology projections indicate that the cost of pin bandwidth will increase significantly and may therefore inhibit the number of processor cores per CMP.

Runahead execution is a very promising approach to tolerate long memory latencies. In this paper we study the memory access characteristics of runahead execution. We show that temporal and data dependency aspects of runahead execution makes it possible to conserve bandwidth through the use of smaller cache blocks in the cache. We demonstrate, using execution-driven full system simulation, that our method of fine-grained fetching can obtain significant performance speedups in bandwidth constrained systems but also yield stable performance in systems that are not bandwidth limited.

Simulating Red Storm: Challenges and Successes in Building a System Simulation

Keith D. Underwood, Michael Levenhagen and Arun F. Rodrigues

Scalable Computing Systems
Sandia National Laboratories
Albuquerque, NM, USA
{*kdunder, mjleven, afrodri*}*@sandia.gov*

Supercomputers are increasingly complex systems merging conventional microprocessors with system on a chip level designs that provide the network interface and router. At Sandia National Labs, we are developing a simulator to explore the complex interactions that occur at the system level. This paper presents an overview of the simulation framework with a focus on the enhancements needed to transform traditional simulation tools into a simulator capable of modeling system level hardware interactions and running native software. Initial validation results demonstrate simulated performance that matches the Cray Red Storm system installed at Sandia. In addition, we include a "what if" study of performance implications on the Red Storm network interface.

Architectural Support for Network Applications on Simultaneous MultiThreading Processors

Kyueun Yi[1] and Jean-Luc Gaudiot[2]

[1]*Electrical Engineering and Computer Science*
University of California, Irvine
Irvine, CA, USA
kyueuny@uci.edu

[2]*Electrical Engineering and Computer Science*
University of California, Irvine
Irvine, CA, USA
gaudiot@uci.edu

As network applications become increasingly sophisticated and internet traffic is getting heavier, future network processors must continue processing computation-intensive network applications at line rates. Most programmable network processors on the market today, such as the Intel IXP2800, target low performance (from 100 Mbps to 10 Gbps). However, low cost edge routers will find it hard to cope with the forthcoming sophistication of network applications to be processed at those speeds. Hence, new architectures should be designed for the programmable network processors of the future. The goal of this paper is to evaluate the applicability and efficiency of Simultaneous MultiThreaded (SMT) as a network processor. Indeed, the SMT model inherently allows the multiple parallel threads which must be dealt with in network processor applications. In this paper, we investigate the architectural implications of network applications on the SMT architecture. We demonstrate that, when executed as independent threads, applications chosen from different network layers show an improved IPC and cache behavior when compared with the situation where the program executed comes from a single network application. Finally, a new architectural solution to cope with packet dependency is proposed and evaluated.

Microarchitectural Support for Speculative Register Renaming

Jesús Alastruey[1], Teresa Monreal[1], Víctor Viñals[1] and Mateo Valero[2]

[1]*Dpto. Informática e Ingeniería de Sistemas-I3A*
Universidad de Zaragoza
Zaragoza, Spain
{jalastru, tmonreal, victor}@unizar.es

[2]*Dept. d'Arquitectura de Computadors*
Universitat Politècnica de Cataluña
Barcelona, Spain
mateo@ac.upc.edu

This paper proposes and evaluates a new microarchitecture for out-of-order processors that supports speculative renaming. We call speculative renaming to the speculative omission of physical register allocation along with the speculative early release of physical registers. These renaming policies may cause a register operand not to be kept in the Physical Register File (PRF). Thus, we add a low-ported Auxiliary Register File (XRF) located outside the processor core that keeps the values absent in PRF and supplies them at higher latency. To support the location of register operands being either in PRF or XRF, we use virtual registers. We consider omission and release policies directed by hardware prediction. Namely, we will use a single Last-Use Predictor that directs both speculative omission and release. We call this mechanism SR-LUP (Speculative Renaming based on Last-Use Prediction). Two Last-Use predictor designs of incremental complexity and performance are analyzed. In a 256-ROB, 8-way processor with an 80int+80fp PRF, SR-LUP with an 11-port 256int+256fp XRF, speeds up computations up to 11.5% and 29% for INT and FP SPEC2K benchmarks, respectively. For FP benchmarks, if the PRF limits the clock frequency, a conventionally managed 128int+128fp PRF can be replaced using SR-LUP by a 64int+64fp PRF backed up with a 10-port 224int+224fp XRF, showing 19% IPS gain.

Session 8

Performance Analysis and Optimization

Automatic Trace-Based Performance Analysis of Metacomputing Applications

Daniel Becker[1,2], Felix Wolf[1,2], Wolfgang Frings[1], Markus Geimer[1], Brian J. N. Wylie[1] and Bernd Mohr[1]

[1]Forschungszentrum Juelich
John von Neumann Institute for Computing (NIC)
Juelich, Germany
{D.Becker, F.Wolf, W.Frings, M.Geimer, B.Wylie, B.Mohr}@fz-juelich.de

[2]Department of Computer Science
RWTH Aachen University
Aachen, Germany

The processing power and memory capacity of independent and heterogeneous parallel machines can be combined to form a single parallel system that is more powerful than any of its constituents. However, achieving satisfactory application performance on such a metacomputer is hard because the high latency of inter-machine communication as well as differences in hardware of constituent machines may introduce various types of wait states. In our earlier work, we have demonstrated that automatic pattern search in event traces can identify the sources of wait states in parallel applications running on a single computer. In this article, we describe how this approach can be extended to metacomputing environments with special emphasis on performance problems related to inter-machine communication. In addition, we demonstrate the benefits of our solution using a real-world multi-physics application.

An Implementation and Evaluation of Client-Side File Caching for MPI-IO

Wei-Keng Liao[1], Avery Ching[1], Kenin Coloma[1], Alok Choudhary[1] and Lee Ward[2]

[1]Dept. of Electrical Engineering and Computer Science
Northwestern University
Evanston, Illinois, United States
{wkliao, aching, kcoloma, choudhar}@ece.northwestern.edu

[2]Dept. of Scalable Computing Systems
Sandia National Laboratories
Albuquerque, New Mexico, United States
lee@sandia.gov

Client-side file caching has long been recognized as a file system enhancement to reduce the amount of data transfer between application processes and I/O servers. However, caching also introduces cache coherence problems when a file is simultaneously accessed by multiple processes. Existing coherence controls tend to treat the client processes independently and ignore the aggregate I/O access pattern. This causes a serious performance degradation for parallel I/O applications. In our earlier work, we proposed a caching system that enables cooperation among application processes in performing client-side file caching. The caching system has since been integrated into the MPI-IO library. In this paper we discuss our new implementation and present an extended performance evaluation on GPFS and Lustre parallel file systems. In addition to comparing our methods to traditional approaches, we examine the performance of MPI-IO caching under direct I/O mode to bypass the underlying file system cache. We also investigate the performance impact of two file domain partitioning methods to MPI collective I/O operations: one which creates a balanced workload and the other which aligns accesses to the file system stripe size. In our experiments, alignment results in better performance by reducing file lock contention. When the cache page size is set to a multiple of the stripe size, MPI-IO caching inherits the same advantage and produces significantly improved I/O bandwidth.

A Utility-based Approach to Cost-Aware Caching in Heterogeneous Storage Systems

Liton Chakraborty and Ajit Singh

Dept. of Electrical and Computer Engineering
University of Waterloo
Waterloo, ON, Canada
litonc@swen.uwaterloo.ca, asingh@etude.uwaterloo.ca

Modern single and multi-processor computer systems incorporate, either directly or through a LAN, a number of storage devices with diverse performance characteristics. These storage devices have to deal with workloads with unpredictable burstiness. Storage aware caching scheme—that partitions the cache among the disks, and aims at balancing the work across the disks — is necessary in this environment. Moreover, maintaining proper size for these partitions is crucial. The existing storage aware caching schemes assume linear relationship between cache size and hit ratio. But, in practice a (disk) partition may accumulate cache blocks (thus, *choke* the remaining disks) without increasing the hit ratio significantly. This disk choking phenomenon may degenerate the performance of the disk system. In this paper, we address this issue of disk choking and present a repartitioning framework based on the notion of marginal gains. Experimental results shows the effectiveness of our approach. We show that our scheme outperforms the existing storage-aware caching schemes while supplied with a workload showing the non-linear relationship between cache size and hit ratio.

Integrated Risk Analysis for a Commercial Computing Service

Chee Shin Yeo[1] and Rajkumar Buyya[2]

[1]*Department of Computer Science and Software Engineering*
The University of Melbourne
Parkville, VIC 3010, Australia
csyeo@csse.unimelb.edu.au

[2]*Department of Computer Science and Software Engineering*
The University of Melbourne
Parkville, VIC 3010, Australia
raj@csse.unimelb.edu.au

Utility computing has been anticipated to be the next generation of computing usage. Users have the freedom to easily switch to any commercial computing service to complete jobs whenever the need arises and simply pay only on usage, without any investment costs. A commercial computing service however has certain objectives or goals that it aims to achieve. In this paper, we identify three essential objectives for a commercial computing service: (i) meet SLA, (ii) maintain reliability, and (iii) earn profit. This leads to the problem of whether a resource management policy implemented in the commercial computing service is able to meet the required objectives or not. So, we also develop two evaluation methods that are simple and intuitive: (i) separate and (ii) integrated risk analysis to analyze the effectiveness of resource management policies in achieving the required objectives. Evaluation results based on five policies successfully demonstrate the applicability of separate and integrated risk analysis to assess policies in terms of the required objectives.

Session 9

Complexity of Algorithms

Max-Min Fair Bandwidth Allocation Algorithms for Packet Switches

Deng Pan[1] and Yuanyuan Yang[2]

[1]Dept. of Computer Science
State University of New York at Stony Brook
Stony Brook, NY 11794, USA
pandeng@cs.sunysb.edu

[2]Dept. of Electrical & Computer Engineering
State University of New York at Stony Brook
Stony Brook, NY 11794, USA
yang@ece.sunysb.edu

With the rapid development of broadband applications, the capability of networks to provide quality of service (QoS) has become an important issue. Fair scheduling algorithms are a common approach for switches and routers to support QoS. All fair scheduling algorithms are running based on a bandwidth allocation scheme. The scheme should be feasible in order to be applied in practice, and should be efficient to fully utilize available bandwidth and allocate bandwidth in a fair manner. However, since a single input port or output port of a switch has only the bandwidth information of its local flows (i.e., the flows traversing itself), it is difficult to obtain a globally feasible and efficient bandwidth allocation scheme. In this paper, we show how to fairly allocate bandwidth in packet switches based on the max-min fairness principle. We first formulate the problem, and give the definitions of feasibility and max-min fairness for bandwidth allocation in packet switches. As the first step to solve the problem, we consider the simpler unicast scenarios, and present the max-min fair bandwidth allocation algorithm for unicast traffic. We then extend the analysis to the more general multicast scenarios, and present the max-min fair bandwidth allocation algorithm for multicast traffic. We prove that both algorithms achieve max-min fairness, and analyze their complexity. The proposed algorithms are universally applicable to any type of switches and scheduling algorithms.

Network-Oblivious Algorithms

Gianfranco Bilardi[1,2], Andrea Pietracaprina[1], Geppino Pucci[1] and Francesco Silvestri[1]

[1]Dept. of Information Engineering
University of Padova
Padova, Italy
{bilardi, capri, geppo, silvest1}@dei.unipd.it

[2]IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA

The design of algorithms that can run unchanged yet efficiently on a variety of machines characterized by different degrees of parallelism and communication capabilities is a highly desirable goal. We propose a framework for *network-obliviousness* based on a model of computation where the only parameter is the problem's input size. Algorithms are then evaluated on a model with two parameters, capturing parallelism and granularity of communication. We show that, for a wide class of network-oblivious algorithms, optimality in the latter model implies optimality in a block-variant of the Decomposable BSP model, which effectively describes a wide and significant class of parallel platforms. We illustrate our framework by providing optimal network-oblivious algorithms for a few key problems, and also establish some negative results.

Minimum number of wavelengths equals load in a DAG without internal cycle

Michel Cosnard[1] and Jean Claude Bermond[2]

[1] *INRIA Sophia Antipolis*
INRIA
Sophia Antipolis, FRANCE
Michel.Cosnard@inria.fr

[2] *Mascotte Project, I3S*
CNRS-UNSA-INRIA
Sophia Antipolis, FRANCE
Jean-Claude.Bermond@inria.fr

Let P be a family of dipaths. The load of an arc is the number of dipaths containing this arc. Let p(G,P) be the maximum of the load of all the arcs and let w(G,P) be the minimum number of wavelengths (colors) needed to color the family of dipaths P in such a way that two dipaths with the same wavelength are arc-disjoint.

Let G be a DAG (Directed Acyclic Graph). An internal cycle is an oriented cycle such that all the vertices have at least one predecessor and one successor in G (said otherwise every cycle contain neither a source nor a sink of G). Here we prove that if G is a DAG without internal cycle, then for any family of dipaths P, w(G,P) = p(G,P). On the opposite we give examples of DAGs with internal cycles such that the ratio between w(G,P) and p(G,P) cannot be bounded.

We also consider an apparently new class of DAGs, which is of interest in itself, those for which there is at most one dipath from a vertex to another. We call these digraphs UPP-DAGs. For these UPP-DAGs we show that the load is equal to the maximum size of a clique of the conflict graph. We show that if an UPP-DAG has only one internal cycle, then for any family of dipaths w(G,P) = 4/3 p(G,P) and we exhibit an UPP-DAG and a family of dipaths reaching the bound. We conjecture that the ratio between w(G,P)and p(G,P) cannot be bounded.

A Comparison of Dag-Scheduling Strategies for Internet-Based Computing

Robert Hall[1], Arnold Rosenberg[2] and Arun Venkataramani[3]

[1] *Computer Science*
University of Massachusetts
Amherst, MA, USA
rwhall@cs.umass.edu

[2] *Computer Science*
University of Massachusetts
Amherst, MA, USA
rsnbrg@cs.umass.edu

[3] *Computer Science*
University of Massachusetts
Amherst, MA, USA
arun@cs.umass.edu

A fundamental challenge in Internet computing (IC) is to efficiently schedule computations having complex interjob dependencies, given the unpredictability of remote machines, in availability and time of access. The recent IC Scheduling theory focuses on these sources of unpredictabil- ity by crafting schedules that maximize the number of exe- cutable jobs at every point in time. In this paper, we ex- perimentally investigate the key question: does IC Schedul- ing yield significant positive benefits for real IC? To this end, we develop a realistic computation model to match jobs to client machines and conduct extensive simulations to com- pare IC-optimal schedules against popular, intuitively com- pelling heuristics. Our results suggest that for a large range of computation-dags, client availability patterns, and two quite different performance metrics, IC-optimal schedules signifi- cantly outperform schedulesproducedbypopularheuristics,by as much as 1020%.

Session 10

Power and Energy Aware Computing

Power-Aware Speedup

Rong Ge and Kirk W. Cameron

Department of Computer Science
Virginia Tech
Blacksburg, VA, 24061
{*ge, cameron*}*@cs.vt.edu*

Power-aware processors operate in various power modes to reduce energy consumption with a corresponding decrease in peak processor throughput. Recent work has shown power-aware clusters can conserve significant energy (>30%) with minimal performance loss (<1%) running parallel scientific workloads. Nonetheless, such savings are typically achieved using a priori knowledge of application performance. Accurate prediction of parallel power consumption and performance is an open problem. However, such techniques would improve our understanding of power-aware cluster tradeoffs and enable identification of system configurations optimized for performance and power ("sweet spots"). Speedup models are powerful analytical tools for evaluating and predicting the performance of parallel applications. Unfortunately, existing speedup models do not quantify parallel overhead for simplicity. Consequently, these models are incapable of accurately accounting for performance and power. We propose power-aware speedup to model and predict the scaled execution time of power-aware clusters. The new model accounts for parallel overhead and predicts (within 7%) the power-aware performance and energy-delay products for various system configurations (i.e. processor counts and frequencies) on NAS Parallel benchmark codes.

A Near-optimal Solution for the Heterogeneous Multi-processor Single-level Voltage Setup Problem

Tai-Yi Huang, Yu-Che Tsai and Edward T.-H. Chu

Computer Science
National Tsing Hua University
Hsinchu, Taiwan (R.O.C)
{*tyhuang, yctsai, edward*}*@eos.cs.nthu.edu.tw*

A heterogeneous multi-processor (HeMP) system consists of several heterogeneous processors, each of which is specially designed to deliver the best energy-saving performance for a particular category of applications. A low power real-time scheduling algorithm is required to schedule tasks on such a system to minimize its energy consumption and complete all tasks by their deadline. The problem of determining the optimal speed for each processor to minimize the total energy consumption is called the voltage setup problem. This paper provides a near-optimal solution for the HeMP single-level voltage setup problem. To our best knowledge, we are the first work that addresses this problem. Initially, each task is assigned to a processor in a local-optimal manner. We next propose a couple of solutions to reduce energy by migrating tasks between processors. Finally, we determine each processors speed by its final workload and the deadline. We conducted a series of simulations to evaluate our algorithms. The results show that the local-optimal partition leads to a considerably better energy-saving schedule than a commonly-used homogeneous multi-processor scheduling algorithm. Furthermore, at all measurable configurations, our energy consumption is at most 3% more than the optimal value obtained by an exhaustive iteration of all possible task-to-processor assignments. In summary, our work is shown to provide a nearoptimal solution at its polynomial-time complexity.

Optimal Energy Balanced Data Gathering in Wireless Sensor Networks

Haibo Zhang[1,2], Hong Shen[1] and Yasuo Tan[2]

[1] *School of Computer Science*
University of Adelaide
Adelaide, South Australia, Australia
haibozh@gmail.com, hong@cs.adelaide.edu.au

[2] *Graduate School of Information Science*
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
ytan@jaist.ac.jp

Unbalanced energy consumption is an inherent problem in wireless sensor networks where some nodes may be overused and die out early, resulting in a short network lifetime. In this paper, we investigate the problem of balancing energy consumption for data gathering sensor networks. Our key idea is to exploit the tradeoff between hop-by-hop transmission and direct transmission to balance energy dissipation among sensor nodes. By assigning each node a transmission probability which controls the ratio between hop-by-hop transmission and direct transmission, we formulate the energy consumption balancing problem as an optimal transmission probability allocation problem. We discuss this problem for both chain networks and general networks. Moreover, we present the solution to compute the optimal number of sections in terms of maximizing the network lifetime. Numerical results demonstrate that our methods outperform the traditional hop-by-hop and direct transmission schemes and achieve significant lifetime extension especially for dense sensor networks.

Verifiable Credit Based Transfers in Wireless Ad Hoc Networks

Bogdan Carbunar, Brett Lindsley, Michael Pearce and Venu Vasudevan

Pervasive Platforms and Architectures Lab
Motorola Labs
Schaumburg, IL, USA
{*carbunar, brett.lindsley, michael.pearce, cvv012*}*@motorola.com*

Encouraging cooperation between users of mobile devices operating in ad hoc mode is a difficult task mostly because the scarce battery and bandwidth resources of devices suggest that selfish behavior may be most beneficial. The insecure usage of credits to reward cooperation can easily provide an incentive for cheating, thus, on the long term only leading to selfishness. In this paper we propose several secure credit based mechanisms enforcing fairness in a hybrid wireless content retrieval system operating both in cellular and ad hoc connectivity modes. Our solution consists of mechanisms for securely and privately discovering desired content on neighboring devices, simultaneously exchanging credit and content shares in a verifiable manner and for generating and expiring non-forgeable credits. We present experimental results of a partial prototype of our system implemented on MPx and E680i cellular phones and HP iPaq hx4700 PDAs, along with extensive simulation results showing that our solution significantly reduces the effectiveness of selfish behavior, making it an unattractive strategy.

Session 11

Performance Modeling and Evaluation

Towards A Better Understanding of Workload Dynamics on Data-Intensive Clusters and Grids

Hui Li[1] and Lex Wolters[2]

[1]Leiden Institute of Advanced Computer Science
Leiden University
Leiden, Zuid Holland, The Netherlands
hui.li@computer.org

[2]Leiden Institute of Advanced Computer Science
Leiden University
Leiden, Zuid Holland, The Netherlands
llexx@liacs.nl

This paper presents a comprehensive statistical analysis of workloads collected on data-intensive clusters and Grids. The analysis is conducted at different levels, including Virtual Organization (VO) and user behavior. The aggregation procedure and scaling analysis are applied to job arrival processes, leading to the identification of several basic patterns, namely, pseudo-periodicity, long range dependence (LRD), and (multi)fractals. It is shown that statistical measures based on interarrivals are of limited usefulness and count based measures should be trusted instead when it comes to correlations. We also study workload characteristics like job run time, memory consumption, and cross correlations between these characteristics. A "bag-of-tasks" behavior is empirically proved, strongly indicating temporal locality. We argue that pseudo-periodicity, LRD, and "bag-of-tasks" behavior are important workload properties on data-intensive clusters and Grids, which are not present in traditional parallel workloads. This study has important implications on workload modeling and performance predictions in data-intensive Grid environments.

Efficient Statistical Performance Modeling for Autonomic, Service-Oriented Systems

Rui Zhang[1], Alan Bivens[2] and Iead Rezek[3]

[1]Computing Laboratory
University of Oxford
Oxford, England
ruiz@comlab.ox.ac.uk

[2]IBM T.J. Watson Research Center
Hawthorne, NY, US
jbivens@us.ibm.com

[3]Department of Engineering Science
University of Oxford
Oxford, England
irezek@robots.ox.ac.uk

As service-oriented environments grow in size and complexity, managing their performance becomes increasingly difficult. To assist administrators, autonomic techniques have been adopted to permit these environments to be self-managing (problem localization, workload management, etc.). These techniques need a sense of system state and the ability to project a new state given some change within the environment. Recent work addressing this issue frequently used statistically learned models which were derived entirely from data. However, many environments already have management facilities in place that could provide precise and useful insights (e.g. workflows) into the system. This paper introduces a method of modeling service-oriented system performance using Bayesian networks and specifically addresses the benefits obtained by incorporating these insights into the model learning process. To further minimize model building costs, we devise a decentralized method to concurrently learn parts of the model where knowledge inclusion is impossible. Simulations and applications in actual environments show significant reductions in learning time, better accuracy and stronger tolerance to small learning data sets.

Multicore Surprises: Lessons Learned from Optimizing Sweep3D on the Cell Broadband Engine

Fabrizio Petrini[1], Gordon Fossum[2], Juan Fernandez[3], Ana Lucia Varbanescu[4], Mike Kistler[2] and Michael Perrone[5]

[1]*Pacific Northwest National Laboratory*
Richland, WA, USA
fabrizio.petrini@pnl.gov

[2]*IBM Austin Research Lab*
Austin, TX, USA
{fossum, mkistler}@us.ibm.com

[3]*Engineering Department*
University of Murcia
Murcia, Spain
juanf@um.es

[4]*Engineering Department*
Delft University of Technology
Delft, The Netherlands
A.L.Varbanescu@tudelft.nl

[5]*IBM TJ Watson*
Yorktown Heights, NY, USA
mpp@us.ibm.com

The Cell Broadband Engine (BE) processor provides the potential to achieve an impressive level of performance for scientific applications. This level of performance can be reached by exploiting several dimensions of parallelism, such as thread-level parallelism using several Synergistic Processing Elements, data streaming parallelism, vector parallelism in the form of 128-bit SIMD operations, and pipeline parallelism by issuing multiple instructions in the same clock cycle. In our exploration to achieve the optimum level of performance for Sweep3D, we have enjoyed many pleasant surprises, such as a very high floating point performance, reaching 64% of the theoretical peak in double precision, and an overall performance speedup ranging from 4.5 times when compared with "heavy iron" processors, up to over 20 times with conventional processors.

Challenges in Mapping Graph Exploration Algorithms on Advanced Multi-core Processors

Oreste Villa[2], Daniele Paolo Scarpazza[1], Fabrizio Petrini[1] and Juan Fernandez Peinador[1]

[1]*Computational & Information Sciences Division*
Pacific Northwest National Laboratory
Richland, WA, United States of America
{daniele.scarpazza, fabrizio.petrini}@pnl.gov,
juanf@ditec.um.es

[2]*Dipartimento di Elettronica e Informazione*
Politecnico di Milano
Milano, Italy
ovilla@elet.polimi.it

Multi-core processors are a shift of paradigm in computer architecture that promises a dramatic increase in performance. But multi-core processors also bring an unprecedented level of complexity in algorithmic design and software development. In this paper we describe the challenges and design choices involved in parallelizing a breadth-first search (BFS) algorithm on a state-of-the-art multi-core processor, the Cell Broadband Engine (Cell BE). Our experiments obtained on a pre-production Cell BE board running at 3.2 GHz show almost linear speedups when using multiple synergistic processing units, and an impressive level of performance when compared to other processors. The Cell BE is typically an order of magnitude faster than conventional processors, such as the AMD Opteron and the Intel Pentium 4 and Woodcrest, an order of magnitude faster than the MTA-2 multi-threaded processor, and two orders of magnitude faster than a BlueGene/L processor.

Session 12

Middleware and Tools

Stack Trace Analysis for Large Scale Debugging

Dorian C. Arnold[1], Dong H. Ahn[2], Bronis R. de Supinski[2], Gregory L. Lee[2], Barton P. Miller[1] and Martin Schulz[2]

[1]*Computer Sciences Department*
University of Wisconsin
Madison, WI, USA
{*darnold, bart*}*@cs.wisc.edu*

[2]*Lawrence Livermore National Laboratory*
Livermore, CA, USA
{*ahn1, bronis, lee218, schulzm*}*@llnl.gov*

We present the Stack Trace Analysis Tool (STAT) to aid in debugging extreme-scale applications. STAT can reduce problem exploration spaces from thousands of processes to a few by sampling stack traces to form *process equivalence classes*, groups of processes exhibiting similar behavior. We can then use full-featured debuggers on representatives from these behavior classes for root cause analysis.

STAT scalably collects stack traces over a sampling period to assemble a profile of the application's behavior. STAT routines process the samples to form a call graph prefix tree that encodes common behavior classes over the program's process space and time. STAT leverages MRNet, an infrastructure for tool control and data analyses, to overcome scalability barriers faced by heavy-weight debuggers.

We present STAT's design and an evaluation that shows STAT gathers informative process traces from thousands of processes with sub-second latencies, a significant improvement over existing tools. Case studies of production codes verify that STAT can quickly identify errors that were previously difficult to locate.

Single IP Address Cluster for Internet Servers

Hiroya Matsuba[1] and Yutaka Ishikawa[2]

[1]*Information Technology Center*
The University of Tokyo
Bunkyo, Tokyo, Japan
matsuba@cc.u-tokyo.ac.jp

[2]*Graduate School of Information Science and Technology*
The University of Tokyo
Bunkyo, Tokyo, Japan
ishikawa@is.s.u-tokyo.ac.jp

Operating a cluster on a single IP address is required when the cluster is used to provide certain Internet services. This paper proposes SAPS, a new method to assign a single IP address to a cluster. The TCP/IP protocol is handled at a single node called the I/O server. The other nodes, called applications nodes, provide the socket interface to applications. The I/O server and applications nodes are connected using a cluster-dedicated network, such as the Myrinet network. The key benefit of the proposed method is that the TCP/IP protocol does not care about congestion and packet loss in the cluster, which often happens if multiple nodes send packets to the bottleneck router. Instead, the cluster-dedicated network manages the packet congestion more efficiently than the TCP/IP protocol. The result of the bandwidth benchmark shows SAPS fully utilizes the bandwidth of the Gigabit Ethernet. The result of the SPEC Web benchmark shows SAPS handles 7.9% more requests than the existing method.

RF2ID: A Reliable Middleware Framework for RFID Deployment

Nova Ahmed[1], Rajnish Kumar[2], Robert Steven French[3] and Umakishore Ramachandran[4]

[1]*College of Computing*
Georgia Institute of Technology
Atlanta, Ga, USA
nova@cc.gatech.edu

[2]*College of Computing*
Georgia Institute of Technology
Atlanta, Ga, USA
rajnish@cc.gatech.edu

[3]*College of Computing*
Georgia Institute of Technology
Atlanta, Ga, USA
robert.steven@cc.gatech.edu

[4]*College of Computing*
Georgia Institute of Technology
Atlanta, Ga, USA
rama@cc.gatech.edu

The reliability of RFID systems depends on a number of factors including: RF interference, deployment environment, configuration of the readers, and placement of readers and tags. While RFID technology is improving rapidly, a reliable deployment of this technology is still a significant challenge impeding wide-spread adoption. This paper investigates system software solutions for achieving a highly reliable deployment that mitigates all sources of inherent unreliability in RFID technology. We have developed (1) a virtual reader abstraction to improve the potentially error-prone nature of data produced by the physical readers and antennas; and (2) a novel path abstraction to capture the logical flow of information among virtual readers as RFID-tagged objects move throughout the environment. Utilizing these abstractions, we have designed and implemented an RFID middleware, RF2ID (Reliable Framework for Radio Frequency Identification), to organize and support queries over data streams in an efficient manner. Prototype implementation using both RFID readers and simulated readers using an empirical model of RFID readers show that RF2ID is able to provide high reliability and support path-based object detection.

A WSRF-Compliant Debugger for Grid Applications

Donny Kurniawan and David Abramson

Monash e-Science and Grid Engineering Lab
Monash University
Caulfield, Victoria, Australia
{*donny.kurniawan, david.abramson*}*@infotech.monash.edu.au*

Grid computing allows the utilization of vast computational resources for solving complex scientific and engineering problems. However, development tools for Grid applications are not as mature as their traditional counterparts, especially in the area of debugging and testing. Debugging Grid applications typically requires a programmer to address non-trivial issues such as heterogeneity, job scheduling, hierarchical resources, and security. This paper presents the design and implementation of a Grid service debug architecture that is compliant with the Web Service Resource Framework standard. The debugger provides a library with a set of well-defined debug APIs.

Plenary Session

Best Papers

Hypergraph-based Dynamic Load Balancing for Adaptive Scientific Computations

Umit V. Catalyurek[1], Erik G. Boman[2], Karen D. Devine[2], Doruk Bozdağ[1], Robert Heaphy[2] and Lee Ann Riesen[2]

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

[2]*Discrete Algorithms and Math. Dept.*
Sandia National Laboratories
Albuquerque, NM, USA
{*egboman, kddevin, rheaphy, lafisk*}*@sandia.gov*

Adaptive scientific computations require that periodic repartitioning (load balancing) occur dynamically to maintain load balance. Hypergraph partitioning is a successful model for minimizing communication volume in scientific computations, and partitioning software for the static case is widely available. In this paper, we present a new hypergraph model for the dynamic case, where we minimize the sum of communication in the application plus the migration cost to move data, thereby reducing total execution time. The new model can be solved using hypergraph partitioning with fixed vertices. We describe an implementation of a parallel multilevel repartitioning algorithm within the Zoltan load-balancing toolkit, which to our knowledge is the first code for dynamic load balancing based on hypergraph partitioning. Finally, we present experimental results that demonstrate the effectiveness of our approach on a Linux cluster with up to 64 processors. Our new algorithm compares favorably to the widely used ParMETIS partitioning software in terms of quality, and would have reduced total execution time in most of our test cases.

Scientific Application Performance on Candidate PetaScale Platforms

Leonid Oliker[1], Andrew Canning[1], Jonathan Carter[1], Costin Iancu[1], Michael Lijewski[1], Shoaib Kamil[1], John Shalf[1], Hongzhang Shan[1], Erich Strohmaier[1], Stephane Ethier[2] and Tom Goodale[3]

[1]*CRD/NERSC*
LBNL
Berkeley, CA, USA
{*loliker, acanning, jtcarter, cciancu, mjlijewski, sakamil, jshalf, hshan, estrohmaier*}*@lbl.gov*

[2]*PPPL*
Princeton University
Princeton, NJ, USA
ethier@pppl.gov

[3]*Computer Science*
Cardiff University
The Parade, CF24 4QJ, UK

After a decade where HEC (high-end computing) capability was dominated by the rapid pace of improvements to CPU clock frequency, the performance of next-generation supercomputers is increasingly differentiated by varying interconnect designs and levels of integration. Understanding the tradeoffs of these system designs, in the context of high-end numerical simulations, is a key step towards making effective petascale computing a reality. This work represents one of the most comprehensive performance evaluation studies to date on modern HEC systems, including the IBM Power5, AMD Opteron, IBM BG/L, and Cray X1E. A novel aspect of our study is the emphasis on full applications, with real input data at the scale desired by computational scientists in their unique domain. We examine six candidate ultra-scale applications, representing a broad range of algorithms and computational structures. Our work includes the highest concurrency experiments to date on five of our six applications, including 32K processor scalability for two of our codes and describe several successful optimizations strategies on BG/L, as well as improved X1E vectorization. Overall results indicate that our evaluated codes have the potential to effectively utilize petascale resources; however, several applications will require reengineering to incorporate the additional levels of parallelism necessary to achieve the vast concurrency of upcoming ultra-scale systems.

Speculative Flow Control for High-Radix Datacenter Interconnect Routers

Cyriel Minkenberg and Mitchell Gusat

IBM Research, Zurich Research Laboratory
Rüschlikon, Switzerland
{*sil, mig*}*@zurich.ibm.com*

High-radix switches are desirable building blocks for large computer interconnection networks, because they are more suitable to convert chip I/O bandwidth into low latency and low cost than low-radix switches. Unfortunately, most existing switch architectures do not scale well to a large number of ports. For example, the complexity of the buffered crossbar architecture scales quadratically with the number of ports. Compounded with support for long round-trip times and many virtual channels, the overall buffer requirements limit the feasibility of such switches to modest port counts. Compromising on the buffer sizing leads to a drastic increase in latency and reduction in throughput, as long as traditional credit flow control is employed at the link level. We propose a novel link-level flow control protocol that enables high-performance scalable routers based on the increasingly popular buffered crossbar architecture to scale to higher port counts without sacrificing performance. By combining credited and speculative transmission, this scheme achieves reliable delivery, low latency, and high throughput, even with crosspoint buffers that are significantly smaller than the round-trip time.

Scalable Compression and Replay of Communication Traces in Massively Parallel Environments

Michael Noeth[1], Frank Mueller[1], Martin Schulz[2] and Bronis R. de Supinski[2]

[1]*Computer Science*
North Carolina State University
Raleigh, NC, USA
mjnoeth@yahoo.com, mueller@cs.ncsu.edu

[2]*Center for Applied Scientific Computing*
Lawrence Livermore National Laboratory
Livermore, CA, USA
{*schulzm, bronis*}*@llnl.gov*

Characterizing the communication behavior of large-scale applications is a difficult and costly task due to code/system complexity and their long execution times. An alternative to running actual codes is to gather their communication traces and then replay them, which facilitates application tuning and future procurements. While past approaches lacked lossless scalable trace collection, we contribute an approach that provides orders of magnitude smaller, if not near constant-size, communication traces regardless of the number of nodes while preserving structural information. We introduce intra- and inter-node compression techniques of MPI events and present results of our implementation for BlueGene/L. Given this novel capability, we discuss its impact on communication tuning and beyond. To the best of our knowledge, such a concise representation of MPI traces in a scalable manner combined with deterministic MPI call replay are without any precedence.

Session 13

Wireless, Adhoc and Sensor Algorithms

Distributed, Reliable Restoration Techniques using Wireless Sensor Devices

Yannis Drougas and Vana Kalogeraki

Dept. of Computer Science and Engineering
University of CA, Riverside
Riverside, CA, USA
{*drougas, vana*}*@cs.ucr.edu*

Wireless sensor networks are small, inexpensive and flexible computational platforms, that have found popular applications in various areas including environmental monitoring, health care and disaster recovery. One fundamental question is how to place the nodes in the network so that complete coverage of the monitored area is achieved. In this paper, we use techniques from discrepancy theory that accurately represent the uncovered area using just a few discrete points, to make sure that every point in the network is covered by at least k sensors, where k is calculated based on user reliability requirements. Our technique is fully distributed, deploying a low number of sensors, and minimizes the communication costs. Our experiments demonstrate that our technique is highly effective in achieving a reliable restoration of a given sensor network area.

Topology-Transparent Duty Cycling for Wireless Sensor Networks

Yu Chen[1], Eric Fleury[2] and Violet R. Syrotiuk[3]

[1]*ARES/INRIA, INSA de Lyon*
Villeurbanne, France
Yu.Chen@inrialpes.fr

[2]*ARES/INRIA, INSA de Lyon*
Villeurbanne, France
Eric.Fleury@inria.fr

[3]*Computer Science & Engineering*
Arizona State University
Tempe, AZ, U.S.A
syrotiuk@asu.edu

Our goal is to save energy in *wireless sensor networks* (WSNs) by periodic duty-cycling of sensor nodes. We schedule sensor nodes between active (transmit or receive) and sleep modes while bounding packet latency in the presence of collisions. In order to support a dynamic WSN topology, we focus on topology-transparent approaches to scheduling; these are independent of detailed topology information. Much work has been done on topology-transparent scheduling in which all nodes are active. In this work, we examine the connection between topology-transparent duty-cycling and such *non-sleeping* schedules. This suggests a way to construct topology-transparent duty-cycling schedules. We analyse the performance of topology-transparent schedules with a focus on throughput in the worst case. A construction of topology-transparent duty-cycling schedules based on a topology-transparent non-sleeping schedule is proposed. The constructed schedule achieves the maximum average throughput in the worst case if the given non-sleeping schedule satisfies certain properties.

Average-Case Performance Evaluation of Online Algorithms for Routing and Wavelength Assignment in WDM Optical Networks

Keqin Li

Department of Computer Science
State University of New York
New Paltz, New York 12561, USA
lik@newpaltz.edu

We investigate the problem of online routing and wavelength assignment and the related throughput maximization problem in wavelength division multiplexing optical networks. It is pointed out that these problems are highly inapproximable. We evaluate the average-case performance of several online algorithms, which have no knowledge of future arriving connection requests when processing the current connection request. Our experimental results on a wide range of optical networks demonstrate that the average-case performance of these algorithms are very close to optimal.

Energy-Aware Self-Stabilization in Mobile Ad Hoc Networks: A Multicasting Case Study

Sandeep K. S. Gupta, Tridib Mukherjee and Ganesh Sridharan

The IMPACT Laboratory, School of Computing and Informatics
Arizona State University
Tempe, AZ, USA
{*sandeep.gupta, tridib, ganesh.sridharan*}*@asu.edu*

Dynamic networks, e.g. Mobile Ad Hoc Networks (MANETs), call for adaptive protocols that can tolerate topological changes due to nodes' mobility and depletion of battery power. Also proactivity in these protocols is essential to ensure low latency. Self-stabilization techniques for distributed systems provide both adaptivity and proactivity to make it suitable for the MANETs. However, energy-efficiency - a prime concern in MANETs with battery-powered nodes - is not guaranteed by self-stabilization. In this paper, we propose a node-based energy metric that minimizes the energy consumption of the multicast tree by taking into account the overhearing cost.tabilization. We apply the metric to Self-Stabilizing Shortest Path Spanning Tree (SS-SPST) protocol to obtain energy-aware SS-SPST (SS-SPST-E). Using simulations, we study the energy-latency tradeoff by comparing SS-SPST-E with SS-SPST and other MANET multicast protocols, such as ODMRP and MAODV.

Session 14

Applications on Emerging Architectures

On the Design and Analysis of Irregular Algorithms on the Cell Processor: A Case Study of List Ranking

David A. Bader, Virat Agarwal and Kamesh Madduri

College of Computing
Georgia Institute of Technology
Atlanta, GA, USA
{*bader, virat, kamesh*}*@cc.gatech.edu*

The Sony-Toshiba-IBM Cell Broadband Engine is a heterogeneous multicore architecture that consists of a traditional microprocessor (PPE), with eight SIMD co-processing units (SPEs) integrated on-chip. We present a complexity model for designing algorithms on the Cell processor, along with a systematic procedure for algorithm analysis. To estimate the execution time of the algorithm, we consider the computational complexity, memory access patterns (DMA transfer sizes and latency), and the complexity of branching instructions. This model, coupled with the analysis procedure, simplifies algorithm design on the Cell and enables quick identification of potential implementation bottlenecks. Using the model, we design an efficient implementation of list ranking, a representative problem from the class of combinatorial and graph-theoretic applications. Due to its highly irregular memory patterns, list ranking is a particularly challenging problem to parallelize on current cache-based and distributed memory architectures. We describe a generic work-partitioning technique on the Cell to hide memory access latency, and apply this to efficiently implement list ranking. We run our algorithm on a 3.2 GHz Cell processor using an IBM QS20 Cell Blade and demonstrate a substantial speedup for list ranking on the Cell in comparison to traditional cache-based microprocessors. For a random linked list of 1 million nodes, we achieve an an overall speedup of 8.34 over a PPE-only implementation. This work is supported in part by NSF Grants CNS-0614915, CAREER CCF-0611589, ITR EF/BIO 03-31654, and DARPA Contract NBCH30390004.

RAxML-Cell: Parallel Phylogenetic Tree Inference on the Cell Broadband Engine

Filip Blagojevic[1], Alexandros Stamatakis[2], Christos D. Antonopoulos[3] and Dimitris S. Nikolopoulos[4]

[1]*Computer Scinece*
Virginia Tech
Blacksburg, VA, USA
filip@cs.vt.edu

[2]*School of Computer & Communication Sciences*
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
Alexandros.Stamatakis@epfl.ch

[3]*Computer Scinece*
College of William & Mary
Williamsburg, VA, USA
cda@cs.wm.edu

[4]*Computer Scinece*
Virginia Tech
Blacksburg, VA, USA
dsn@cs.vt.edu

Computational phylogeny is a challenging application even for the most powerful supercomputers. It is also an ideal candidate for benchmarking emerging multiprocessor architectures, because it exhibits fine- and coarse-grain parallelism at multiple levels. In this paper, we present the porting, optimization, and evaluation of RAxML on the Cell Broadband Engine. RAxML is a provably efficient, hill climbing algorithm for computing phylogenetic trees, based on the Maximum Likelihood (ML) method. The Cell Broadband Engine, a heterogeneous multi-core processor with SIMD accelerators which was initially marketed for set-top boxes, is currently being deployed on supercomputers and high-end server architectures. We present both conventional and unconventional, Cell-specific optimizations for RAxML's search algorithm on a real Cell multiprocessor. While exploring these optimizations, we present solutions to problems related to floating point code execution, complex control flow, communication, scheduling, and multi-level parallelization on the Cell.

Hardware/Software Co-Design for Matrix Computations on Reconfigurable Computing Systems

Ling Zhuo and Viktor K. Prasanna

Department of Electrical Engineering
University of Southern California
Los Angeles, California, United States
{*lzhuo, prasanna*}*@usc.edu*

Recently, reconfigurable computing systems have been built which employ Field-Programmable Gate Arrays (FPGAs) as hardware accelerators for general-purpose processors. These systems provide new opportunities for scientific computations. However, the co-existence of the processors and the FPGAs in such systems also poses new challenges to application developers. In this paper, we investigate a design model for hybrid designs, that is, designs that utilize both the processors and the FPGAs. The model characterizes a reconfigurable computing system using various system parameters, including the floating-point computing power of the processor and the FPGA, the number of nodes, the memory bandwidth and the network bandwidth. Using the model, we investigate hardware/software co-design for two computationally intensive applications: matrix factorization and all-pairs shortest-paths problem. Our designs balance the load between the processor and the FPGA, as well as overlap the computation time with memory transfer time and network communication time. The proposed designs are implemented on 6 nodes in a Cray XD1 chassis. Our implementations achieve 20 GFLOPS and 6.6 GFLOPS for these two applications, respectively.

Masked Queries for Search Accuracy in Peer-to-Peer File-Sharing Systems

Wai Gen Yee, Linh Thai Nguyen and Ophir Frieder

Department of Computer Science
Illinois Institute of Technology
Chicago, IL, USA
{*yee, nguylin*}*@iit.edu, ophir@ir.iit.edu*

Peer-to-peer file-sharing systems suffer from the overspecification of query results due to the fact that queries are conjunctive and the descriptions of shared files are sparse. Ultimately, longer queries, which should yield more accurate results, actually do the opposite. The judicious masking of query terms circumvents the shortcomings of conjunctive query processing, significantly improving query accuracy.

Session 15

Interconnection Networks

Mixed-radix Twisted Torus Interconnection Networks

José M. Cámara[1], Miquel Moretó[2], Enrique Vallejo[3], Ramón Beivide[3], Carmen Martínez[3], José Miguel-Alonso[4] and Javier Navaridas[4]

[1]Dept. of Electromechanical Engineering
University of Burgos
Burgos, Castilla y León, Spain
checam@ubu.es

[2]Dept. of Computer Architecture
Technical University of Catalonia
Barcelona, Catalonia, Spain
mmoreto@ac.upc.edu

[3]Computer Architecture Group
University of Cantabria
Santander, Cantabria, Spain
{enrique, mon, carmenmf}@atc.unican.es

[4]Dept. of Computer Architecture and Technology
University of the Basque Country
San Sebastián, Basque Country, Spain
{j.miguel, javier-navaridas}@ehu.es

Many parallel computers use Tori interconnection networks. Machines from Cray, HP and IBM, among others, exploit these topologies. In order to maintain full network symmetry, 2D and 3D Tori (k-ary 2-cubes and k-ary 3-cubes) must have the same number of nodes (k) per dimension resulting in square or cubic topologies. Nevertheless, for practical reasons, computer engineers have designed and built 2D and 3D Tori having a different number of nodes per dimension. These mixed-radix topologies are not edge-symmetric which translates into poor performance provoked by an unbalanced use of the network links.

In this paper, we propose and analyze twisted 2D and 3D Tori which remove the network bottlenecks present in mixed-radix standard Tori. These new topologies recover edge-symmetry and, consequently, balance the utilization of their links. We describe the distance-related parameters of these twisted networks and use simulation to asses their performance under synthetic loads. The obtained results show noticeable and consistent performance gains (up to a 88% increase in accepted load). In addition, we propose scalable and practicable packet routing and folding techniques for these interconnection subsystems. The complexity of the resulting architectural solutions is similar to the one exhibited by traditional routing and folding mechanisms employed in standard Tori. This fact together with the performance improvements obtained could justify the use of these twisted topologies in the future.

Performance, Cost, and Energy Evaluation of Fat H-Tree: A Cost-Efficient Tree-Based On-Chip Network

Hiroki Matsutani[1], Michihiro Koibuchi[2] and Hideharu Amano[1]

[1]Department of Information and Computer Science
Keio University
Yokohama, JAPAN
{matutani, hunga}@am.ics.keio.ac.jp

[2]National Institute of Informatics
Tokyo, JAPAN
koibuchi@nii.ac.jp

Fat H-Tree is a novel tree-based interconnection network providing a torus structure, which is formed by combining two folded H-Tree networks, and is an attractive alternative to tree-based networks such as Fat Trees in a microarchitecture domain. In this paper, we introduce Fat H-Tree and its deadlock-free routing algorithms. The performance of Fat H-Tree is evaluated using real application traces, and the result is compared with those of other tree-based networks. The network logic area and wire resources for Fat H-Tree are computed based on a typical implementation of on-chip routers using a 0.18um standard cell library. In addition, the energy consumption is estimated based on the gate-level power analysis. The results show that 1) Fat H-Tree outperforms Fat Tree with two upward and four downward connections in terms of throughput and average hop count; 2) Fat H-Tree requires 19.3%-26.4% smaller network logic area compared with the Fat Tree; 3) Fat H-Tree consumes 8.3%-8.6% less energy compared with the Fat Tree due to its short average hop count; 4) Fat H-Tree uses slightly more wire resources compared with the Fat Tree, but the current process technology can provide sufficient wire resources for implementing Fat H-Tree based on-chip networks.

Table-lookup based Crossbar Arbitration for Minimal-Routed, 2D Mesh and Torus Networks

Daeho Seo[1] and Mithuna Thottethodi[2]

[1] *School of Electrical and Computer Engineering*
Purdue University
West Lafayette, Indiana, United States
seod@purdue.edu

[2] *School of Electrical and Computer Engineering*
Purdue University
West Lafayette, Indiana, United States
mithuna@purdue.edu

Crossbar arbitration—which determines the allocation of output ports to packets in the input queues—is a performance-critical stage in the overall performance of routers for input-queued networks. The overall performance of crossbar arbitration depends on two metrics: (a) *matching power* – the ability of the arbiter to maximize the number of matches between requesting inputs and free outputs and (b) arbitration throughput – the number of such matches per unit time. Ideally, crossbar arbitration should maximize both metrics. Unfortunately, implementing high performance matching schemes compromises arbitration throughput. Similarly, simpler arbitration mechanisms that deliver high arbitration throughput offer lower matching power.

The major contribution of this paper is the design of a table-lookup based crossbar arbitration mechanism—TabArb—that delivers superior matching and high arbitration throughput for minimal-routed, two dimensional mesh and torus networks. The two key innovations of TabArb are: (a) it forwards multiple requests from each input port to multiple output ports to expose adequate matching potential and (b) it employs precomputed tables that store maximum cardinality matches for all possible request combinations. Our technique improves the saturation throughput of adaptive routed mesh network by 14.8%. It offers little improvement for the DOR router due to limited opportunity.

Power-Aware Bandwidth-Reconfigurable Optical Interconnects for High-Performance Computing (HPC) Systems

Avinash Karanth Kodi[1] and Ahmed Louri[2]

[1] *Department of Electrical and Computer Engineering*
University of Arizona
Tucson, Arizona, USA
avinashk@ece.arizona.edu

[2] *Department of Electrical and Computer Engineering*
University of Arizona
Tucson, Arizona, USA
louri@ece.arizona.edu

As communication distances and bit rates increase, opto-electronic interconnects are becoming de-facto standard for designing high-bandwidth low-latency interconnection networks for high performance computing (HPC) systems. While bandwidth scaling with efficient multiplexing techniques (wavelengths, time and space) are available, static assignment of wavelengths can be detrimental to network performance for adversial traffic patterns. Dynamic bandwidth reconfiguration based on actual traffic pattern can lead to improved network performance by utilizing idle resources. While dynamic bandwidth re-allocation (DBR) techniques can alleviate interconnection bottlenecks, power consumption also increases considerably. In this paper, we propose a dynamically reconfigurable architecture called E-RAPID (Extended-Reconfigurable, All-Photonic Interconnect for Distributed and parallel systems) that not only dynamically re-allocates bandwidth, but also reduces the power consumption for all traffic patterns. Our proposed LS (Lock-Step) reconfiguration technique combines Dynamic Power Management (DPM) with DBR techniques, achieving a reduction in power consumption of 25% - 50% while degrading the throughput by less than 5%.

Session 16

Performance Prediction and Distributed Systems

A Performance Prediction Framework for Grid-Based Data Mining Applications

Leonid Glimcher and Gagan Agrawal

Department of Computer Science and Engineering
The Ohio State University
Columbus, OH, USA
{*glimcher, agrawal*}*@cse.ohio-state.edu*

For a grid middleware to perform resource allocation, *prediction models* are needed, which can determine how long an application will take for completion on a particular platform or configuration. In this paper, we take the approach that by focusing on the characteristics of the class of applications a middleware is suited for, we can develop simple performance models that can be very accurate in practice.

The particular middleware we consider is FREERIDE-G (FRamework for Rapid Implementation of Datamining Engines in Grid), which supports a high-level interface for developing data mining and scientific data processing applications that involve data stored in remote repositories. The FREERIDE-G system needs detailed performance models for performing resource selection, i.e., choosing computing nodes and replica of the dataset. This paper presents and evaluates such a performance model. By exploiting the fact that the processing structure of data mining and scientific data analysis applications developed on FREERIDE-G involves generalized reductions, we are able to develop an accurate performance prediction model.

We have evaluated our model using implementations of three well-known data mining algorithms and two scientific data analysis applications developed using FREERIDE-G. Results from these five applications show that we are able to accurately predict execution times for applications as we vary the number of storage nodes, number of nodes available for computation, the dataset size, the network bandwidth, and the underlying hardware.

Prediction Services for Distributed Computing

Warren Smith

Texas Advanced Computing Center
The University of Texas at Austin
Austin, TX, USA
wsmith@tacc.utexas.edu

Users of distributed systems such as the TeraGrid and Open Science Grid can execute their applications on many different systems. We wish to help such users, or the grid schedulers they use, select where to run applications by providing predictions of when tasks will complete if sent to different systems. We make predictions of file transfer times, batch scheduler queue wait times, and application execution times using historical information and instance-based learning techniques. Our prediction errors for data from the TACC lonestar system are 37 percent of mean file transfer time, 115 percent for mean queue wait time, and 72 percent of mean execution time. Our approach achieves significantly lower prediction error on other workloads. We have wrapped these prediction techniques with web services, making predictions available to users of distributed systems as well as tools such as resource brokers and metaschedulers.

Adaptive Predictor Integration for System Performance Prediction

Jian Zhang and Renato Figueiredo

Advanced Computing and Information Systems (ACIS) Laboratory
University of Florida
Gainesville, FL, USA
{*jianzh, renato*}*@acis.ufl.edu*

The integration of multiple predictors promises higher prediction accuracy than the accuracy that can be obtained with a single predictor. The challenge is how to select the best predictor at any given moment. Traditionally, multiple predictors are run in parallel and the one that generates the best result is selected for prediction. In this paper, we propose a novel approach for predictor integration based on the learning of historical predictions. It uses classification algorithms such as k-Nearest Neighbor (k-NN) based supervised learning to forecast the best predictor for the workload under study. Then only the forecasted best predictor is run for prediction. Our experimental results show that it achieved 20.18% higher best predictor forecasting accuracy than the cumulative MSE based predictor selection approach used in the popular Network Weather Service system. In addition, it outperformed the observed most accurate single predictor in the pool for 44.23% of the performance traces.

Machine Bank: Own Your Virtual Personal Computer

Shuo Tang[1], Yu Chen[2] and Zheng Zhang[2]

[1]*Dept. of Computer Science and Technology*
Tsinghua University
Beijing, P. R. China
ts@mails.tsinghua.edu.cn

[2]*Microsoft Research Asia*
Beijing, P. R. China
{*ychen, zzhang*}*@microsoft.com*

In this paper, we report the design, implementation and experimental results of Machine Bank, a system engineered towards the popular shared-lab scenario, where users outnumber available PCs and may get different PCs in different sessions. Machine Bank allows users to preserve their entire working environment across sessions. Each client runs virtual machine, which is saved to and reinstantiated from a content-addressable backend storage. We carefully designed lightweight hooks at client side that implements caching and tracking logics to improve reinstantiation speed as well as to remove unnecessary network and disk traffic. Our detailed evaluation demonstrates that these techniques are effective, and the overall performance fits well with the shared-lab usage.

Session 17

Network Algorithms

A Semi-Distributed Axiomatic Game Theoretical Mechanism for Replicating Data Objects in Large Distributed Computing Systems

Samee Ullah Khan and Ishfaq Ahmad

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

Replicating data objects onto servers across a system can alleviate access delays. The selection of data objects and servers requires solving a constraint optimization problem, which is NP-complete in general. A majority of conventional replica placement techniques falter on issues of scalability or solution quality. To counteract such issues, we propose a game theoretical replica placement technique, in which computational agents compete for the allocation or reallocation of replicas onto their servers in order to reduce the user perceived access delays. The technique is based upon six well-defined axioms, each guaranteeing certain basic game theoretical properties. This eccentric method of designing game theoretical techniques using axioms is unique in the literature and takes away from the designers the cumbersome mathematical details of game theory. The distinctive feature of these axioms is that when amassed together, their individual properties constrict into one system-wide performance enhancement property, which in our case is the reduction of access time. The control of the proposed technique is "semi-distributed" in nature, wherein all the heavy processing is done on the servers of the distributed system and the central body is only required to take a binary decision: (0) not to replicate or (1) to replicate. This semi-distributed approach makes the technique scalable and helps solutions to converge in a fast turn-around time without loosing much of the solution quality. Experimental comparisons are made against: 1) branch and bound, 2) greedy, 3) genetic, 4) Dutch auction, and 5) English auction. As attested by the results, the proposed technique maintains superior solution quality in terms of lower communication cost and reduced execution time.

Online Aggregation over Trees

C. Greg Plaxton[1], Mitul Tiwari[2] and Praveen Yalagandula[3]

[1]*Department of Computer Science*
University of Texas
Austin, TX, USA
plaxton@cs.utexas.edu

[2]*Department of Computer Science*
University of Texas
Austin, TX, USA
mitult@cs.utexas.edu

[3]*HP Labs*
Palo Alto, CA, USA
praveen.yalagandula@hp.com

Consider a distributed network with nodes arranged in a tree, and each node having a local value. We consider the problem of aggregating values (e.g., summing values) from all nodes to the requesting nodes in the presence of writes. The goal is to minimize the total number of messages exchanged. The key challenges are to define a notion of "acceptable" aggregate values, and to design algorithms with good performance that are guaranteed to produce such values. We formalize the acceptability of aggregate values in terms of certain consistency guarantees. We propose a lease-based aggregation mechanism, and evaluate algorithms based on this mechanism in terms of consistency and performance. With regard to consistency, we adapt the definitions of strict and causal consistency to apply to the aggregation problem. We show that any lease-based aggregation algorithm provides strict consistency in sequential executions, and causal consistency in concurrent executions. With regard to performance, we propose an online lease-based aggregation algorithm, and show that, for sequential executions, the algorithm is constant competitive against any offline algorithm that provides strict consistency. Our online lease-based aggregation algorithm is presented in the form of a fully distributed protocol, and the aforementioned consistency and performance results are formally established with respect to this protocol.

Optimizing Multiple Distributed Stream Queries Using Hierarchical Network Partitions

Sangeetha Seshadri, Vibhore Kumar, Brian F. Cooper and Ling Liu

College of Computing
Georgia Institute of Technology
Atlanta, Georgia, U.S.A
{*sangeeta, vibhore, cooperb, lingliu*}*@cc.gatech.edu*

We consider the problem of query optimization in distributed data stream systems where multiple continuous queries may be executing simultaneously. In order to achieve the best performance, query planning (such as join ordering) must be considered in conjunction with deployment planning (e.g., assigning operators to physical nodes with optimal ordering). However, such a combination involves not only a large number of network nodes but also many query operators, resulting in an extremely large search space for optimal solutions. Our paper aims at addressing this problem by utilizing hierarchical network partitions. We propose two algorithms - Top-Down and Bottom-Up which utilize hierarchical network partitions to provide scalable query optimization. Formal analysis is presented to establish the bounds on the search-space and to show the sub-optimality of our algorithms. We have implemented both algorithms in the IFLOW system, an adaptive distributed stream management system. Through simulations and experiments using a prototype deployed on Emulab we demonstrate the effectiveness of our algorithms.

A Scalable Cluster Algorithm for Internet Resources

Chuang Liu[1] and Ian Foster[2,3]

[1]*Microsoft*
Redmond, WA, U.S.A.
chuangl@microsoft.com

[2]*Argonne National Laboratory*
Argonne, IL, U.S.A.
foster@mcs.anl.gov

[3]*Department of Computer Science*
University of Chicago
Chicago, IL, U.S.A.

Applications such as parallel computing, online games, and content distribution networks need to run on a set of resources with particular network connection characteristics to get good performance. To locate such resource sets, we introduce a scalable algorithm to compute a hierarchical cluster structure for a large number of Internet resources such that resources in a cluster have much smaller latency with each other than with other resource. Using the hierarchical cluster structure, we propose an approximate algorithm to answer queries for a resource set with desired network connections. We evaluate this method in a large distributed Internet environment including 2500 DNS servers, and show that our algorithm can locate required resources with high accuracy in much shorter time than traditional methods.

Session 18

Peer-to-Peer Systems and Applications I

Making Peer-to-Peer Anonymous Routing Resilient to Failures

Yingwu Zhu[1] and Yiming Hu[2]

[1]CSSE
Seattle University
Seattle, WA, USA
zhuy@seattleu.edu

[2]ECECS
University of Cincinnati
Cincinnati, OH, USA
yhu@ececs.uc.edu

One hurdle to using peer-to-peer networks as anonymizing networks is churn. Node churn makes anonymous paths fragile and short-lived: failures of a relay node disrupt the path, resulting in message loss and communication failures. To make anonymous routing resilient to node failures, we take two steps: (1) we use a simple yet powerful idea based on message redundancy by erasure coding and path redundancy to mask node failures; (2) we base mix choices of a path on node lifetime prediction and choose stable nodes as relay nodes, thereby prolonging single path durability. We present an allocation of erasure-coded message segments among multiple paths that provides a guideline on how to maximize routing resilience upon different node availabilities in real-world systems. Via detailed simulations, we compare routing resilience of our approach and existing anonymity protocols, showing that our approach greatly improves routing resilience while consuming modest bandwidth.

Pseudo Trust: Zero-Knowledge Based Authentication in Anonymous Peer-to-Peer Protocols

Li Lu[1], Jinsong Han[2], Lei Hu[1], Jinpeng Huai[3,4], Yunhao Liu[2] and Lionel M. Ni[2]

[1]State Key Lab of Information Security
Graduate School of Chinese Academy of Sciences
Beijing, China
{luli, hu}@is.ac.cn

[2]Dept. of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong, China
{jasonhan, liu, ni}@cse.ust.hk

[3]School of Computer Science
Beihang University
Beijing, China
huaijp@buaa.edu.cn

[4]State Key Lab of Software Developing Environment
Beihang University
Beijing, China

Most of the current trust models in peer-to-peer (P2P) systems are identity based, which means that in order for one peer to trust another, it needs to know the other peers identity. Hence, there exists an inherent tradeoff between trust and anonymity. To the best of our knowledge, there is currently no P2P protocol that provides complete mutual anonymity as well as authentication and trust management. We propose a zero-knowledge authentication scheme called Pseudo Trust (PT), where each peer, instead of using its real identity, generates an unforgeable and verifiable pseudonym using a one-way hash function. A novel authentication scheme based on Zero-Knowledge Proof is designed so peers can be authenticated without leaking any sensitive information. We analyze the levels of security and anonymity in PT, and evaluate its performance using trace-driven simulations and a prototype implementation. The strengths of Pseudo Trust include the lack of need for a centralized trusted party or CA, high scalability and security, low traffic and cryptography processing overheads, and man-in-middle attack resistance.

Gossip-based Reputation Aggregation for Unstructured Peer-to-Peer Networks

Runfang Zhou[1] and Kai Hwang[2]

[1]*Computer Science*
University of Southern California
Los Angeles, CA, USA
rzhou@usc.edu

[2]*Electrical Engineering*
University of Southern California
Los Angeles, CA, USA
kaihwang@usc.edu

Peer-to-Peer (P2P) reputation systems are needed to evaluate the trustworthiness of participating peers and to combat selfish and malicious peer behaviors. The reputation system collects locally generated peer feedbacks and aggregates them to yield global reputation scores. Development of decentralized reputation system is in great demand for unstructured P2P networks since most P2P applications on the Internet are unstructured. In the absence of fast hashing and searching mechanisms, how to perform efficient reputation aggregation is a major challenge on unstructured P2P computing.

We propose a novel reputation aggregation scheme called GossipTrust. This system computes global reputation scores of all nodes concurrently. By resorting to a gossip protocol and leveraging the power nodes, GossipTrust is adapted to peer dynamics and robust to disturbance by malicious peers. Simulation experiments demonstrate the system as scalable, accurate, robust and fault-tolerant. These results prove the claimed advantages in low aggregation overhead, storage efficiency, and scoring accuracy in unstructured P2P networks. With minor modifications, the system is also applicable to structured P2P systems with projected better performance.

Replication Strategy in Unstructured Peer-to-Peer Systems

Guofu Feng[1], Yuquan Jiang[1], Guihai Chen[2], Qing Gu[2], Sanglu Lu[2] and Daoxu Chen[2]

[1]*School of Information Science*
Nanjing Audit University
Nanjing, Jiangsu, China
{fgf, jyq}@nau.edu.cn

[2]*The State Key Lab. for Novel Software Technology*
Nanjing University
Nanjing, Jiangsu, China
{ghchen, guq, sanglu, cdx}@nju.edu.cn

The unstructured Peer-to-Peer (P2P) systems usually use a blind search method to find the requested data object by propagating a query to a number of peers randomly. In order to increase the success rate of blind search, replication techniques are widely used in these systems. Most P2P systems replicate the most frequently accessed data objects to improve system performance. However, existing replication strategies cannot answer the question that how many replicas of an object should be kept in the P2P system. If an object is replicated excessively, it inevitably will affect the average efficiency of a replica, which will decrease the whole search performance. This paper addresses the issue of finding the proper number of replicas for an object according to its query rate. In this paper, we firstly investigate the precise relation among success rate, the allocation of replicas and query rate. Then we propose an approach of the allocation of copies to optimize the success rate. As a benchmark, our result offers a new understanding of replication.

Session 19

Networks and Storage Systems

Packet Reordering in Network Processors

Govind Sreekar Shenoy[1], Ramaswamy Govindarajan[1,2] and Joy Kuri[3]

[1]Supercomputer Education and Research Centre
Indian Institute of Science
Bangalore, Karnataka, India
sgovind@hpc.serc.iisc.ernet.in, govind@serc.iisc.ernet.in

[2]Computer Science and Automation
Indian Institute of Science
Bangalore, Karnataka, India

[3]Centre for Electronics Design and Technology
Indian Institute of Science
Bangalore, Karnataka, India
kuri@cedt.iisc.ernet.in

Network processors today consists of multiple parallel processors (microengines) with support for multiple threads to exploit packet level parallelism inherent in network workloads. With such concurrency, packet ordering at the output of the network processor cannot be guaranteed. This paper studies the effect of concurrency in network processors on packet ordering. We use a validated Petri net model of a commercial network processor, Intel IXP 2400, to determine the extent of packet reordering for IPv4 forwarding application. Our study indicates that in addition to the parallel processing in the network processor, the allocation scheme for the transmit buffer also adversely impacts packet ordering. In particular, our results reveal that these packet reordering results in a packet retransmission rate of up to 61%. We explore different transmit buffer allocation schemes namely, contiguous, strided, local, and global which reduces the packet retransmission to 24%. We propose an alternative scheme, Packet Sort, which guarantees complete packet ordering while achieving a throughput of 2.5 Gbps. Further, Packetsort outperforms the in-built packet ordering schemes in the IXP processor by up to 35%

Deadline-based QoS Algorithms for High-performance Networks

Alejandro Martínez[1], Francisco J. Alfaro[1], José L. Sánchez[1] and José Duato[2]

[1]Computing Systems Department
Univ. of Castilla-La Mancha
Albacete, Spain
{alejandro, falfaro, jsanchez}@dsi.uclm.es

[2]Dept. of Systems Data Processing and Computers
Tech. Univ. of Valencia
Valencia, Spain
jduato@disca.upv.es

Nowadays, low-latency and contention-free interconnection networks are demanded for the execution of parallel applications. Moreover, high bandwidth is also required to access storage devices. In addition to these, there is also a need for administration traffic used to configure and manage the machine. Finally, some low-priority traffic like backup copies is needed. Therefore, there is a great variety of application requirements in such environments.

The usual solution to cope with this variety of communication necessities is to provide more resources than needed to ensure meeting traffic requirements. A subtler approach could be taken in the design of the interconnection for such machines. A single network with some quality of service (QoS) support could be used to provide each kind of traffic with its specific requirements.

The two main types of QoS support are per-traffic-class and per-flow support. Deadline-based algorithms can provide powerful QoS provision. However, the cost associated with keeping ordered lists of packets makes them impractical for high-performance networks.

In this paper, we discuss how to obtain most of the benefits of the per-flow QoS within the restrictions of high-performance switches. More specifically, we will propose a novel strategy to emulate the Earliest Deadline First (EDF) family of algorithms by using a pair of FIFO queues.

Parallel I/O Performance Characterization of Columbia and NEC SX-8 Superclusters

Subhash Saini[1], Dale Talcott[1], Rajeev Thakur[2], Panagiotis Adamidis[3], Rolf Rabenseifner[4] and Robert Ciotti[1]

[1]*NASA Advanced Supercomputing Division*
NASA Ames Research Cente
Moffett Field, California, USA
subhash.saini@nasa.gov, dtalcott@mail.arc.nasa.gov,
ciotti@nas.nasa.gov

[2]*Mathematics and Computer Science Division*
Argonne National Laboratory, Argonne
Argonne, IL, USA
thakur@mcs.anl.gov

[3]*German Climate Computing Center*
Hamburg, Germany
adamidis@dkrz.de

[4]*High Performance Computing Center*
University of Stuttgart
Stuttgart,, Germany
rabenseifner@hlrs.de

Many scientific applications running on todays supercomputers deal with increasingly large data sets and are correspondingly bottlenecked by the time it takes to read or write the data from/to the file system. We therefore undertook a study to characterize the parallel I/O performance of two of todays leading parallel supercomputers: the Columbia system at NASA Ames Research Center and the NEC SX-8 supercluster at the University of Stuttgart, Germany. On both systems, we ran a total of seven parallel I/O benchmarks, comprising five low-level benchmarks: (i) IO_Bench, (ii) MPI Tile IO, (iii) IOR (POSIX and MPI-IO), (iv) b_eff_io (five different patterns), and (v) SPIOBENCH, and two scalable synthetic compact application (SSCA) benchmarks: (a) HPCS (High Productivity Computing Systems) SSCA #3 and (b) FLASH IO (parallel HDF5). We present the results of these experiments characterizing the parallel I/O performance of these two systems.

Design Alternatives for a High-Performance Self-Securing Ethernet Network Interface

Derek L. Schuff and Vijay S. Pai

Department of Electrical and Computer Engineering
Purdue University
West Lafayette, IN, USA
{*dschuff, vpai*}*@purdue.edu*

This paper presents and evaluates a strategy for integrating the Snort network intrusion detection system into a high-performance programmable Ethernet network interface card (NIC), considering the impact of several possible hardware and software design choices. While currently proposed ASIC, FPGA, and TCAM systems can match incoming string content in real-time, the system proposed also supports the stream reassembly and HTTP content transformation capabilities of Snort. This system, called LineSnort, parallelizes Snort using concurrency across TCP sessions and executes those parallel tasks on multiple low-frequency pipelined RISC processors embedded in the NIC. LineSnort additionally exploits opportunities for intra-session concurrency. The system also includes dedicated hardware for high-bandwidth data transfers and for high-performance string matching.

Detailed results obtained by simulating various software and hardware configurations show that the proposed system can achieve intrusion detection throughputs in excess of 1 Gigabit per second for fairly large rule sets. Such performance requires the system to use hardware-assisted string matching and a small shared data cache. The system can extract performance through increases in processor clock frequency or parallelism, allowing additional flexibility for designers to achieve performance within specified area or power budgets. By efficiently offloading the computationally difficult task of intrusion detection to the network interface, LineSnort enables intrusion detection to run directly on PC-based network servers rather than just at powerful edge-based appliances. As a result, LineSnort has the potential to protect servers against the growing menace of LAN-based attacks, whereas traditional edge-based intrusion detection deployments can only protect against external attacks.

Session 20

Compiler Optimization and Software Environment

Towards Optimal Multi-level Tiling for Stencil Computations

Lakshminarayanan Renganarayana[1], Manjukumar Harthikote-Matha[2], Rinku Dewri[3] and Sanjay Rajopadhye[4]

[1]*Computer Science Department*
Colorado State University
Fort Collins, CO, United States
ln@cs.colostate.edu

[2]*Computer Science Department*
Colorado State University
Fort Collins, CO, United States
manjuhmm@cs.colostate.edu

[3]*Computer Science Department*
Colorado State University
Fort Collins, CO, United States
rinku@cs.colostate.edu

[4]*Computer Science Department*
Colorado State University
Fort Collins, CO, United States
sanjay.rajopadhye@colostate.edu

Stencil computations form the performance-critical core of many applications. Tiling and parallelization are two important optimizations to speed up stencil computations. Many tiling and parallelization strategies are applicable to a given stencil computation. The best strategy depends not only on the combination of the two techniques, but also on many parameters: tile and loop sizes in each dimension; computation-communication balance of the code; processor architecture; message startup costs; etc. The best choices can only be determined through design-space exploration, which is extremely tedious and error prone to do via exhaustive experimentation. We characterize the space of multi-level tilings and parallelizations for 2D/3D Gauss-Siedel stencil computation. A systematic exploration of a part of this space enabled us to derive a design which is up to a factor of two faster than the standard implementation.

An Optimizing Compiler for Parallel Chemistry Simulations

Jun Cao[1], Ayush Goyal[2], Samuel P. Midkiff[1] and James M. Caruthers[2]

[1]*School of Electrical and Computer Engineering*
Purdue University
West Lafayette, IN, United States
{caoj, smidkiff}@purdue.edu

[2]*School of Chemical Engineering*
Purdue University
West Lafayette, IN, United States
{ayush, caruther}@purdue.edu

Well designed domain specific languages enable the easy expression of problems, the application of domain specific optimizations, and dramatic improvements in productivity for their users. In this paper we describe a compiler for polymer chemistry, and in particular rubber chemistry, that achieves all of these goals. The compiler allows the development of a system of ordinary differential equations describing a complex rubber reaction - a task that used to require months - to be done in days. The generated code, like much machine generated code, is more complex than human written code, and stresses commercial compilers to the point of failure. However, because of knowledge of the form of ODEs generated, the compiler can perform specialized common sub-expression and other algebraic optimizations that simplifies the code sufficiently to allow it to be compiled (eliminating all but 6.9% of the operations in our largest program) and to provide five times faster performance on our largest benchmark codes.

A Scalable Approach for the Secure and Authorized Tracking of the Availability of Entities in Distributed Systems

Shrideep Pallickara, Jaliya Ekanayake and Geoffrey Fox

Community Grids Lab
Indiana University
Bloomington, IN, USA
{*spallick, jekanaya, gcf*}*@indiana.edu*

As the scale and proliferation of distributed applications continues to increase a need often arises to track the availability of entities that comprise the distributed system. An entity that is part of such a distributed system could be a resource, a service that provides a set of exposed capabilities, an application or a user. In this paper we present a transport-independent scheme for tracking the availability of entities in distributed systems. The scheme enforces the authorized generation and consumption of traces (encapsulating entity availability). The scheme also facilitates the secure distribution of traces while coping with some classes of denial of service attacks.

Architectural Considerations for Efficient Software Execution on Parallel Microprocessors

Srinivas Vadlamani and Stephen Jenks

Dept. of Electrical Engineering and Computer Science
University of California
Irvine, CA, USA
{*srinivas.v, sjenks*}*@uci.edu*

Chip Multiprocessors (CMPs) and Simultaneous Multithreading (SMT) processors provide high performance but put more pressure on the memory interface than their single-thread counterparts. The "memory wall" problem is exacerbated by multiple threads sharing a memory interface, and will get worse as more cores are added. Therefore, communications between cores, using shared caches or fast interconnects between private caches, are needed to keep the CPUs busy without burdening the memory interface. Multiple CMP systems add another dimension to this challenging problem, as the communication mechanism is no longer uniform. To parallelize data-intensive applications for high performance on these systems, one must explore a number of execution behaviors in a complex architecture-dependent exercise that entails identifying key components of the communication subsystem and understanding their behavior under varying workloads. As part of ongoing research into efficient program execution models for parallel microprocessors, we have developed a tool to evaluate the performance of the storage controllers at different levels of the memory hierarchy under varying workloads and measure cache coherence overhead. The tool allows exploration of architectural features of real processors that affect the performance of several parallel execution approaches. Here, we demonstrate its use by evaluating two of our parallel programming models that employ architecture-specific optimizations and compare them to a conventional model for several applications on parallel microprocessors.

Session 21

Distributed Algorithms

File Creation Strategies in a Distributed Metadata File System

Ananth Devulapalli[1] and Pete Wyckoff[2]

[1] *Ohio Supercomputer Center*
Springfield, OH, USA
ananth@osc.edu

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

As computing breaches petascale limits both in processor performance and storage capacity, the only way that current and future gains in performance can be achieved is by increasing the parallelism of the system. Gains in storage performance remain low due to the use of traditional distributed file systems such as NFS, where although multiple clients can access files at the same time, only one node can serve files to the clients. New file systems that distribute load across multiple data servers are being developed; however, most implementations concentrate all the metadata load at a single server still. Distributing metadata load is important to accommodate growing numbers of more powerful clients.

Scaling metadata performance is more complex than scaling raw I/O performance, and with distributed metadata the complexity increases further. In this paper we present strategies for file creation in distributed metadata file systems. Using the PVFS distributed file system as our testbed, we present designs that are able to reduce the message complexity of the create operation and increase performance. Compared to the basecase create protocol implemented in PVFS, our design delivers near constant operation latency as the system scales, does not degenerate under high contention situations, and increases throughput linearly as the number of metadata servers increase. The design schemes are applicable to any distributed file system implementation.

Fast Failure Detection in a Process Group

Xinjie Li and Monica Brockmeyer

Department of Computer Science
Wayne State University
Detroit, MI, USA
{xinjieli, mbrockmeyer}@wayne.edu

Failure detectors represent a very important building block in distributed applications. The speed and the accuracy of the failure detectors is critical to the performance of the applications built on them. In a common implementation of failure detector based on heartbeats, there is a tradeoff between speed and accuracy so it is difficult to be both fast and accurate. Based on the observation that in many distributed applications, one process takes a special role as the leader, we propose a Fast Failure Detection (FFD) algorithm that detects the failure of the leader both fast and accurately. Taking advantage of spatial multiple timeouts, FFD detects the failure of the leader within a time period of just a little more than one heartbeat interval, making it almost the fastest detection algorithm possible based on heartbeat messages. FFD could be used stand alone in a static configuration where the leader process is fixed at one site. In a dynamic setting, where the role of leader has to be assumed by another site if the current leader fails, FFD could be used in collaboration with a leader election algorithm to speed up the process of electing a new leader.

Aggregate Threshold Queries in Sensor Networks

Izchak Sharfman[1], Assaf Schuster[1] and Daniel Keren[2]

[1]*Computer Science Department*
Technion
Haifa, Israel
{*tsachis, assaf*}*@cs.technion.ac.il*

[2]*Computer Science Department*
Haifa University
Haifa, Israel
dkeren@cs.haifa.ac.il

An important class of queries over sensor networks are network-wide aggregation queries. In this work we study a class of aggregation queries which we refer to as *aggregate threshold queries*. The goal of an aggregate threshold query is to continuously monitor the network and give a notification every time an aggregated value crosses a predetermined threshold value. Aggregate threshold queries are of particular importance in a wireless sensor environment, since they allow network-wide events to be detected, with a minimum expenditure of energy. Such network-wide events might include, for example, the variance in sensor readings exceeding a certain threshold.

We present an efficient algorithm for implementing arbitrary aggregate threshold queries over sensor networks. Our algorithm is based on a novel geometric approach by which an arbitrary aggregate threshold query can be split into a set of numerical constraints on the readings of the individual sensors. These constraints are used by the individual sensors to monitor their readings. The constraints are constructed so that as long as none of the constraints are violated, it is guaranteed that the aggregated value has not crossed the threshold. Experiments we performed on real-world data indicate that by employing these constraints, sensors are able to reduce the number of transmissions required for implementing the query by orders of magnitude, thus significantly reducing energy consumption.

A Model for Large Scale Self-stabilization

Thomas Herault, Pierre Lemarinier, Olivier Peres, Laurence Pilard and Joffroy Beauquier

Univ Paris Sud
LRI UMR8623
INRIA
Orsay, France
{*herault, lemarini, peres, pilard, jb*}*@lri.fr*

We introduce a new model for distributed algorithms designed for large scale systems that need a low-overhead solution to allow the processes to communicate with each other. We assume that every process can communicate with any other process provided it knows its identifier, which is usually the case in e.g. a peer to peer system, and that nodes may arrive or leave at any time. To cope with the large number of processes, we limit the memory usage of each process to a small constant number of variables, combining this with previous results concerning failure detectors and resource discovery. We illustrate the model with a self-stabilizing algorithm that builds and maintains a spanning tree topology. We provide a formal proof of the algorithm and the results of experiments on a cluster.

Session 22

Peer-to-Peer Systems and Applications II

Performance scalability of the JXTA P2P framework

Gabriel Antoniu[1], Loïc Cudennec[1], Mathieu Jan[1] and Mike Duigou[2]

[1]*IRISA/INRIA*
Campus de Beaulieau
35042 Rennes Cedex, France
{*Gabriel.Antoniu, Loic.Cudennec, Mathieu.Jan*}*@irisa.fr*

[2]*Project JXTA*
Sun Microsystems
Santa Clara, CA, U.S.A
Mike.Duigou@sun.com

Features of the P2P model, such as scalability and volatility tolerance, have motivated its use in distributed systems. Several generic P2P libraries have been proposed for building distributed applications. However, very few experimental evaluations of these frameworks have been conducted, especially at large scales. Such experimental analyses are important, since they can help system designers to optimize P2P protocols and better understand the benefits of the P2P model. This is particularly important when the P2P model is applied to special use cases, such as grid computing. This paper focuses on the scalability of two main protocols proposed by the JXTA P2P platform. First, we provide a detailed description of the underlying mechanisms used by JXTA to manage its overlay and propagate messages over it: the rendezvous protocol. Second, we describe the discovery protocol used to find resources inside a JXTA network. We then report a detailed, large-scale, multi-site experimental evaluation of these protocols, using the nine clusters of the French Grid'5000 testbed.

Popularity Adaptive Search in Hybrid P2P Systems

Xiaoqiu Shi[1], Jinsong Han[2], Yunhao Liu[2] and Lionel M. Ni[2]

[1]*Dept. of Computer Science*
Wenzhou University
Wenzhou, Zhejiang, China
sxq@wzu.edu.cn

[2]*Dept. of Computer Science and Engineering*
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{*jasonhan, liu, ni*}*@cs.ust.hk*

In a hybrid peer-to-peer (P2P) system, flooding and DHT are both employed for content locating. The decision to use flooding or DHT largely depends on the population of desired data. Previous works either use local information only, or do not consider dynamic factors of P2P systems. In this paper, we propose a Popularity Adaptive Search method for Hybrid (PASH) P2P systems. By dynamically detecting the content popularity, PASH properly selects search methods and efficiently saves query traffic cost and response time. We comprehensively evaluate PASH through synthetic and trace-driven simulations. The results show that PASH outperforms existing approaches and it also scales well.

CoQUOS: Lightweight Support for Continuous Queries in Unstructured Overlays

Lakshmish Ramaswamy, Jianxia Chen and Piyush Parate

Department of Computer Science
University of Georgia
Athens, Georgia, USA
{*laks, chen, parate*}*@cs.uga.edu*

The utility and the effectiveness of peer-to-peer (P2P) content distribution systems can be greatly enhanced by augmenting their ad-hoc content discovery mechanisms with two capabilities, namely a mechanism to enable the peers to register their queries and receive notifications when corresponding data-items are added to the network and a means for the peers to advertise their new content. While P2P-based publish-subscribe systems can infuse these capabilities, developing full-fledged publish-subscribe systems on top of unstructured P2P networks requires complex techniques, and it is often an overkill for many P2P applications. For these applications, we study the alternate continuous query paradigm, which is functionally similar to publish-subscribe systems, but provides best-effort notification guarantees. This paper presents CoQUOS – a scalable and lightweight middleware to support continuous queries in unstructured P2P networks. A key strength of the CoQUOS system is that it can be implemented on any unstructured overlay network. Moreover, CoQUOS preserves the simplicity and flexibility of the overlay network. Central to our design of the CoQUOS middleware is a completely decentralized scheme to register a query at different regions of the P2P network. This mechanism includes two novel components, namely cluster resilient random walk algorithm for propagating query to various regions of the network and dynamic probability-based query registration technique for ensuring that the registrations are well distributed. Our experiments show that the proposed techniques are highly effective and their overheads are low.

RASC: Dynamic Rate Allocation for Distributed Stream Processing Applications

Yannis Drougas and Vana Kalogeraki

Computer Science & Engineering
University of California, Riverside
Riverside, CA, United States
{*drougas, vana*}*@cs.ucr.edu*

In today's world, stream processing systems have become important, as applications like media broadcasting, sensor network monitoring and on-line data analysis increasingly rely on real-time stream processing. In this paper, we propose a distributed stream processing system that composes stream processing applications dynamically, while meeting their rate demands. Our system consists of the following components: (1) a distributed component discovery algorithm that discovers components available at nodes on demand, (2) resource monitoring techniques to maintain current resource availability information, (3) a scheduling algorithm that schedules application execution, and (4) a minimum cost composition algorithm that composes applications dynamically based on component and resource availability and scheduling demands. Our detailed experimental results, over the PlanetLab testbed, demonstrate the performance and efficiency of our approach.

Session 23

Job Scheduling

Provably Efficient Online Non-clairvoyant Adaptive Scheduling

Yuxiong He[1], Wen-Jing Hsu[1] and Charles E. Leiserson[2]

[1]*School of Computer Engineering*
Nanyang Technological University
Nanyang Avenue, Singapore
{*heyu0006, hsu*}*@ntu.edu.sg*

[2]*Computer Science and Artificial Intelligence Laboratory*
Massachusetts Institute of Technology
Cambridge, MA, USA
cel@mit.edu

Scheduling competing jobs on multiprocessors has always been an important issue for parallel and distributed systems. The challenge is to ensure global, system-wide efficiency while offering a level of fairness to user jobs. Various degrees of successes have been achieved over years. However, few existing schemes address both efficiency and fairness over a wide range of work loads. Moreover, in order to obtain analytical results, most of them require prior information about jobs, which may be difficult to obtain in real applications.

This paper presents a novel adaptive scheduling algorithm GRAD that ensures fair allocation under all levels of workload, and it offers provable efficiency without requiring prior information of job's parallelism. Moreover, it provides effective control over the scheduling overhead and ensures efficient utilization of processors. Specifically, we show that GRAD is $O(1)$-competitive against an optimal offline scheduling algorithm with respect to both mean response time and makespan for batched jobs and non-batched jobs respectively.

To the best of our knowledge, GRAD is the first non-clairvoyant scheduling algorithm that offers such guarantees. We also believe that our new approach of resource request-allotment protocol deserves further exploration.

The simulation results show that, for non-batched jobs, the makespan produced by GRAD is no more than 1.39 times of the optimal on average. For batched jobs, the mean response time produced by GRAD is no more than 2.37 times of the optimal on average.

Analysis of Scheduling Algorithms with Reservations

Lionel Eyraud Dubois[1], Grégory Mounié[2] and Denis Trystram[2]

[1]*LIP*
ÉNS Lyon
Lyon, France
Lionel.Eyraud-Dubois@ens-lyon.fr

[2]*LIG*
Grenoble Universités
Grenoble, France
{*mounie, trystram*}*@imag.fr*

In this work, we analyze the problem of scheduling a set of independent jobs on a homogeneous parallel computer. This problem has been widely studied from both a theoretical perspective (complexity analysis, and predictability of scheduling algorithms) and practical side (schedulers in production systems). It is common for some processors of a cluster to become unavailable for a certain period of time corresponding to reservations. These reservations represent blocks of time and quantities of resources set asigned in advance for specific applications.

We propose here to investigate the scheduling problem where there are restricted resource availabilities. Our main result is to provide a deep analysis for this problem (complexity, lower bounds and upper bounds) for several variants of list scheduling algorithms. More precisely, we show that the problem of scheduling with any reservations can not be approximated. This leads to the study of restricted versions of this problem where the amount of reservation is limited.

Our analysis is based on an old bound of Graham for resource constraint list scheduling for which we propose a new simpler proof by considering the continuous version of this problem.

An Adaptive Rescheduling Strategy for Grid Workflow Applications

Zhifeng Yu and Weisong Shi

Department of Computer Science
Wayne State University
Detroit, Michigan, USA
{*zhifeng.yu, weisong*}*@wayne.edu*

Scheduling is the key to the performance of grid workflow applications. Various strategies are proposed, including static scheduling strategies which map jobs to resources before execution time, or dynamic alternatives which schedule individual job only when it is ready to execute. While sizable work supports the claim that the static scheduling performs better for workflow applications than the dynamic one, it is questioned how a static schedule works effectively in a grid environment which changes constantly. This paper proposes a novel adaptive rescheduling concept, which allows the workflow planner works collaboratively with the run time executor and reschedule in a proactive way had the grid environment changes significantly. An HEFT-based adaptive rescheduling algorithm is presented, evaluated and compared with traditional static and dynamic strategies respectively. The experiment results show that the proposed strategy not only outperforms the dynamic one but also improves over the traditional static one. Furthermore we observed that it performs more efficiently with data intensive application of higher degree of parallelism.

Predictive Resource Scheduling in Computational Grids

Clovis Chapman, Mirco Musolesi, Wolfgang Emmerich and Cecilia Mascolo

Department of Computer Science
University College London
London, United Kingdom
{*c.chapman, m.musolesi, w.emmerich, c.mascolo*}*@cs.ucl.ac.uk*

The integration of clusters of computers into computational grids has recently gained the attention of many computational scientists. While considerable progress has been made in building middleware and workflow tools that facilitate the sharing of compute resources, little attention has been paid to grid scheduling and load balancing techniques to reduce job waiting time. Based on a detailed analysis of usage characteristics of an existing grid that involves a large CPU cluster, we observe that grid scheduling decisions can be significantly improved if the characteristics of current usage patterns are understood and extrapolated into the future. The paper describes an architecture and an implementation for a predictive grid scheduling framework which relies on Kalman filter theory to predict future CPU resource utilisation. By way of replicated experiments we demonstrate that the prediction achieves a precision within 15-20% of the utilisation later observed and can significantly improve scheduling quality, compared to approaches that only take into account current load indicators.

Session 24

Fault Tolerance and Checkpointing

A Job Pause Service under LAM/MPI+BLCR for Transparent Fault Tolerance

Chao Wang[1], Frank Mueller[1], Christian Engelmann[2] and Stephen L. Scott[2]

[1]Computer Science
North Carolina State University
Raleigh, NC, USA
wchao@ncsu.edu, mueller@cs.ncsu.edu

[2]Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN, USA
{engelmannc, scottsl}@ornl.gov

Checkpoint/restart (C/R) has become a requirement for long-running jobs in large-scale clusters due to a mean-time-to-failure (MTTF) in the order of hours. After a failure, C/R mechanisms generally require a complete restart of an MPI job from the last checkpoint. A complete restart, however, is unnecessary since all but one node are typically still alive. Furthermore, a restart may result in lengthy job requeuing even though the original job had not exceeded its time quantum.

In this paper, we overcome these shortcomings. Instead of job restart, we have developed a transparent mechanism for job pause within LAM/MPI+BLCR. This mechanism allows live nodes to remain active and roll back to the last checkpoint while failed nodes are dynamically replaced by spares before resuming from the last checkpoint. Our methodology includes LAM/MPI enhancements in support of scalable group communication with fluctuating number of nodes, reuse of network connections, transparent coordinated checkpoint scheduling and a BLCR enhancement for job pause. Experiments in a cluster with the NAS Parallel Benchmark suite show that our overhead for job pause is comparable to that of a complete job restart. A minimal overhead of 5.6% is only incurred in case migration takes place while the regular checkpoint overhead remains unchanged. Yet, our approach alleviates the need to reboot the LAM run-time environment, which accounts for considerable overhead resulting in net savings of our scheme in the experiments. Our solution further provides full transparency and automation with the additional benefit of reusing existing resources. Executing continues after failures within the scheduled job, *i.e.*, the application staging overhead is not incurred again in contrast to a restart. Our scheme offers additional potential for savings through incremental checkpointing and proactive diskless live migration, which we are currently working on.

An optimistic checkpointing and selective message logging approach for consistent global checkpoint collection in distributed systems

Qiangfeng Jiang and D. Manivannan

Department of Computer Science
University of Kentucky
Lexington, KY, USA
{richardj, manivann}@cs.uky.edu

In this paper, we present an asynchronous consistent global checkpoint collection algorithm which prevents contention for network storage at the file server and hence reduces the checkpointing overhead. The algorithm has two phases: In the first phase, a process initiates consistent global checkpoint collection by saving its state tentatively and asynchronously (called tentative checkpoint) in local memory or remote stable storage if there is no contention for stable storage while saving the state; in the second phase, the message log associated with the tentative checkpoint is stored in stable storage (checkpoint finalization phase). The tentative checkpoint together with the associated message log stored in the stable storage becomes part of a consistent global checkpoint. Under our algorithm, two or more processes can concurrently initiate consistent global checkpoint collection. Every tentative checkpoint will be finalized successfully unless a failure occurs. The finalized checkpoints of each process is assigned a unique sequence number in ascending order. Finalized checkpoints with same sequence number form a consistent global checkpoint.

DejaVu: Transparent User-Level Checkpointing, Migration, and Recovery for Distributed Systems

Joseph F. Ruscio[1], Michael A. Heffner[2] and Srinidhi Varadarajan[1]

[1]*Computing Systems Research Laboratory Department of Computer Science*
Virginia Tech
Blacksburg, VA, United States
{jruscio, srinidhi}@cs.vt.edu

[2]*Evergrid*
Blacksburg, VA, United States
mike.heffner@evergrid.com

In this paper, we present a new fault tolerance system called DejaVu for transparent and automatic checkpointing, migration, and recovery of parallel and distributed applications. DejaVu provides a transparent parallel checkpointing and recovery mechanism that recovers from any combination of systems failures without any modification to parallel applications or the OS. It uses a new runtime mechanism for transparent incremental checkpointing that captures the least amount of state needed to maintain global consistency and provides a novel communication architecture that enables transparent migration of existing MPI codes, without source-code modifications. Performance results from the production-ready implementation show less than 5% overhead in real-world parallel applications with large memory footprints.

A Fault Tolerance Protocol with Fast Fault Recovery

Sayantan Chakravorty and Laxmikant V. Kale

Department of Computer Science
University of Illinois
Urbana-Champaign, IL, USA
{schkrvrt, kale}@uiuc.edu

Fault tolerance is an important issue for large machines with tens or hundreds of thousands of processors. Checkpoint-based methods, currently used on most machines, rollback all processors to previous checkpoints after a crash. This wastes a significant amount of computation as all processors have to redo all the computation from that checkpoint onwards. In addition, recovery time is bound by the time between the last checkpoint and the crash. Protocols based on message logging avoid the problem of rolling back all processors to their earlier state. However, the recovery time of existing message logging protocols is no smaller than the time between the last checkpoint and crash. We present a fault tolerance protocol, in this paper, that provides fast restarts by using the ideas of message logging and object-based processor virtualization. We evaluate our implementation of the protocol in the Charm++/Adaptive MPI runtime system. We show that our protocol provides fast restarts and, for many applications, has low fault-free overhead.

Session 25

Load Balancing Algorithms

Route Table Partitioning and Load Balancing for Parallel Searching with TCAMs

Dong Lin[1], Yue Zhang[1], Chengchen Hu[1], Bin Liu[1], Xin Zhang[2] and Derek Pao[3]

[1]*Dept. of Computer Science and Technology*
Tsinghua University
Beijing, China
lindong05@tsinghua.org.cn,
zhang-yue@mails.tsinghua.edu.cn, huc@ieee.org,
liub@tsinghua.edu.cn

[2]*Dept. of Computer Science*
Carnegie Mellon University
Pittsburgh, PA, USA
xzhang1@cmu.edu

[3]*Dept. of Electronic Engineering*
City University of Hong Kong
Hong Kong, China
d.pao@cityu.edu.hk

With the continuous advances in optical communications technology, the link transmission speed of Internet backbone has been increasing rapidly. This in turn demands more powerful IP address lookup engine. In this paper, we propose a power-efficient parallel TCAM-based lookup engine with a distributed logical caching scheme for dynamic load-balancing. In order to distribute the lookup requests among multiple TCAM chips, a smart partitioning approach called pre-order splitting divides the route table into multiple sub-tables for parallel processing. Meanwhile, by virtual of the cache-based load balancing scheme with slow-update mechanism, a speedup factor of N-1 can be guaranteed for a system with N(N>2) TCAM chips, even with unbalanced bursty lookup requests.

Dynamic Multi-User Load Balancing in Distributed Systems

Satish Penmatsa and Anthony T. Chronopoulos

Department of Computer Science
The University of Texas at San Antonio
San Antonio, TX, USA
{*spenmats, atc*}*@cs.utsa.edu*

In this paper, we review two existing static load balancing schemes based on M/M/1 queues. We then use these schemes to propose two dynamic load balancing schemes for multi-user (multi-class) jobs in heterogeneous distributed systems. These two dynamic load balancing schemes differ in their objective. One tries to minimize the expected response time of the entire system while the other tries to minimize the expected response time of the individual users. The performance of the dynamic schemes is compared with that of the static schemes using simulations with various loads and parameters. The results show that, at low communication overheads, the dynamic schemes show superior performance over the static schemes. But as the overheads increase, the dynamic schemes (as expected) yield similar performance to that of the static schemes.

Distributed Aggregation Algorithms with Load-Balancing for Scalable Grid Resource Monitoring

Min Cai and Kai Hwang

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

Scalable resource monitoring and discovery are essential to the planet-scale infrastructures such as Grids and PlanetLab. This paper proposes a scalable Grid monitoring architecture that builds distributed aggregation trees (DAT) on a structured P2P network like Chord. By leveraging Chord topology and routing mechanisms, the DAT trees are implicitly constructed from native Chord routing paths without membership maintenance. To balance the DAT trees, we propose a balanced routing algorithm on Chord that dynamically selects the parent of a node from its finger nodes by its distance to the root.

This paper shows that this balanced routing algorithm enables the construction of almost completely balanced DATs, when nodes are evenly distributed in the Chord identifier space. We have evaluated the performance and scalability of a DAT prototype implementation with up to 8192 nodes. Our experimental results show that the balanced DAT scheme scales well to a large number of nodes and corresponding aggregation trees. Without maintaining explicit parent-child membership, it has very low overhead during node arrival and departure. We demonstrate that the DAT scheme performs well in Grid resource monitoring.

Session 26

Distributed and Mobile Applications

A Performance Analysis of Indirect Routing

J. M. Opos, S. Ramabhadran, A. Terry, J. Pasquale, A. C. Snoeren and A. Vahdat

Dept. of Computer Science and Engineering
University of California, San Diego
La Jolla, CA, USA
{*jopos, sriram, aterry, pasquale, snoeren, vahdat*}*@cs.ucsd.edu*

Indirect routing involves sending messages between Internet end nodes through a specified intermediate node to effect a different end-to-end route than the default direct route. Indirect routing offers the potential for significant improvements in throughput performance by exploiting the Internets richness in throughput diversity. We present a performance analysis of indirect routing, showing that it can result in a 33-49% increase in average throughput performance while incurring low overhead.

Measuring the Robustness of Resource Allocations in a Stochastic Dynamic Environment

Jay Smith[1,2], Luis Briceno[2], Anthony Maciejewski[2], H. J. Siegel[2,3], Timothy Renner[3], Vladimir Shestak[2], Joshua Ladd[4], Andrew Sutton[3], David Janovy[2] and Sudha Govindasamy[2]

[1]*Global Server Systems Operations*
IBM
Boulder, CO, USA
bigfun@us.ibm.com

[2]*Dept. of Electrical and Computer Engineering*
Colorado State University
Ft. Collins, CO, USA
{*ldbricen, aam, hj, shestak*}*@engr.colostate.edu,*
{*David.Janovy, Sudha.Govindasamy*}*@colostate.edu*

[3]*Dept. of Computer Science*
Colorado State University
Ft. Collins, CO, USA
{*Timothy.Renner, Andrew.Sutton*}*@colostate.edu*

[4]*Mathematics Department*
Colorado State University
Ft. Collins, CO, USA
Joshua.Ladd@colostate.edu

Heterogeneous distributed computing systems often must operate in an environment where system parameters are subject to uncertainty. Robustness can be defined as the degree to which a system can function correctly in the presence of parameter values different from those assumed. We present a methodology for quantifying the robustness of resource allocations in a dynamic environment where task execution times are stochastic. The methodology is evaluated through measuring the robustness of three different resource allocation heuristics within the context of a stochastic dynamic environment. A Bayesian regression model is fit to the combined results of the three heuristics to demonstrate the correlation between the stochastic robustness metric and the presented performance metric. The correlation results demonstrated the significant potential of the stochastic robustness metric to predict the relative performance of the three heuristics given a common objective function.

Implementing Replica Placements: Feasibility and Cost Minimization

Thanasis Loukopoulos[1], Nikos Tziritas[1], Petros Lampsas[2] and Spyros Lalis[1]

[1]*Department of Computer and Communications Engineering*
University of Thessaly
Volos, Greece
{*luke, nitzirit, lalis*}*@inf.uth.gr*

[2]*Department of Informatics and Computer Technology*
Technological Educational Institute of Lamia
Lamia, Greece
plam@teilam.gr

Given two replication schemes X^{old} and X^{new}, the Replica Transfer Scheduling Problem (RTSP) aims at reaching X^{new}, starting from X^{old}, with minimal implementation cost. In this paper we generalize the problem description to include special cases, where deadlocks can occur while in the process of implementing X^{new}. We address this impediment by introducing artificial (dummy) transfers. We then prove that RTSP-decision is NP-complete and propose two kinds of heuristics. The first attempts to replace dummy transfers with valid ones, while the second minimizes the implementation cost. Experimental evaluation of the algorithms illustrates the merits of our approach.

Session 27

Algorithms for Parallel Execution

Task-pushing: a Scalable Parallel GC Marking Algorithm without Synchronization Operations

Ming Wu[1] and Xiao-Feng Li[2]

[1]*Institute of Computing Technology, Chinese Academy of Sciences*
Beijing, China
wuming@ict.ac.cn

[2]*Middleware Products Division, Software and Solutions Group, Intel Corp*
Beijing, China
xiao.feng.li@intel.com

This paper describes a scalable parallel marking technique for garbage collection that does not employ any synchronization operation. To achieve good scalability, two major design issues have to be resolved in parallel marking algorithm, i.e., the overhead of synchronization operations and load balance. This paper presents task-pushing, a novel parallel marking algorithm where each thread proactively gives up its spare tasks to other threads. Enlightened by the idea of communicating sequential process (CSP), task-pushing arranges the computation into a process network, eliminating synchronization operations in the whole marking process. Load balance is achieved by dripping tasks from thread local mark-stack for other threads to execute. To the best of our knowledge, this is the first parallel marking algorithm that completely avoids the synchronization primitives. We evaluated task-pushing in aspects of queuing efficiency, load balancing strategy, synchronization overhead, and overall scalability. The results on a 16-way Intel Xeon machine showed that task-pushing has better scalability than work-stealing technique with pseudojbb and GCOld server-kind Java benchmarks.

Taking Advantage of Collective Operation Semantics for Loosely Coupled Simulations

Joe Shang-Chieh Wu and Alan Sussman

UMIACS and Department of Computer Science
University of Maryland
College Park, MD, USA
{meou, als}@cs.umd.edu

Although a loosely coupled component-based framework offers flexibility and versatility for building and deploying large-scale multi-physics simulation systems, the performance of such a system can suffer from excessive buffering of data objects which may or may not be transferred between components. By taking advantages of the collective properties of parallel simulation components, which is common for data-parallel scientific applications, an optimization method, which we call **buddy-help**, can greatly enhance overall performance. Buddy-Help can reduce the time taken for buffering operations in an exporting component, when there are timing differences across processes in the exporting component. The optimization enables skipping unnecessary buffering operations, once another process, which has already performed the collective export operation, has determined that the buffered data will never be needed. Because an analytical study would be very difficult due to the complexity of the overall coupled simulation system, the performance improvement enabled by buddy-help is investigated via a micro-benchmark specifically designed to illustrate the behavior of coupled simulation scenarios under which buddy-help can provide performance gains.

Accelerating Distributed Computing Applications Using a Network Offloading Framework

Yaron Weinsberg[1], Danny Dolev[1], Pete Wyckoff[2] and Tal Anker[3,1]

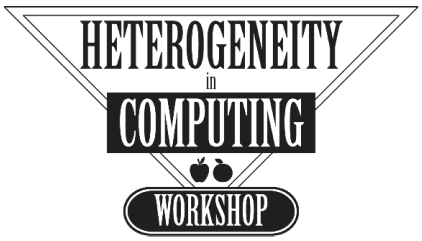
[1]*Computer Science*
The Hebrew University
Jerusalem, Israel
{*wyaron, dolev*}*@cs.huji.ac.il*

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

[3]*Radlan - Marvell*
Tel-Aviv, Israel
tala@marvell.com

During the last two decades, a considerable amount of academic research has been conducted in the field of distributed computing. Typically, distributed applications require frequent network communication, which becomes a dominate factor in the overall runtime overhead. The recent proliferation of *programmable* peripheral devices for computer systems may be utilized in order to improve the performance of such applications. Offloading application-specific network functions to peripheral devices can improve performance and reduce host CPU utilization. Due to the peculiarities of each particular device and the difficulty of programming an outboard CPU, the need for an abstracted offloading framework is apparent. This paper proposes a novel offloading framework, called Hydra, that enables utilization of such devices. The framework enables an application developer to design the offloading aspects of the application by specifying an "offloading layout", which is enforced by the runtime during application deployment. The performance of a variety of distributed algorithms can be significantly improved by utilizing such a framework. We demonstrate this claim by evaluating several offloaded applications: a distributed total message ordering algorithm and a packet generator.

Workshop 1
Heterogeneity in Computing Workshop
HCW 2007



HCW 2007 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 16th Heterogeneity in Computing Workshop (HCW 2007)

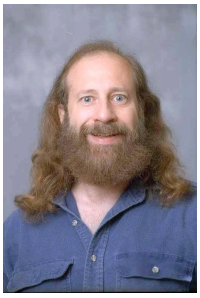
The pervasive use of networks and the Internet has lead the number of connected computing resources to grow tremendously, thus creating opportunity and need for heterogeneous computing systems. Furthermore, heterogeneity is also present in emerging computer architectures. In this context, issues of interest to he HCW workshop include but are not limited to: parallel and distributed computing, programming paradigms and tools, resource discovery and management, task and communication scheduling, task coordination and workflow management, performance management, heterogeneous cluster computing, heterogeneous computer architecture, grid computing, peer-to-peer computing, adaptive computing, ubiquitous computing, mobile computing, fault tolerance, as well as application case studies.

General Chair:
José A. B. Fortes
University of Florida
Email: fortes@ufl.edu

Program Chair:
Henri Casanova
University of Hawai`i at Manoa
Email: henric@hawaii.edu

1-4244-0910-1/07/$20.00 ©2007 IEEE

Message from the HCW Steering Committee Chair



These are the proceedings of the 16th Heterogeneity in Computing Workshop, also known as HCW 2007. The title of the workshop has been changed from the original title of "Heterogeneous Computing Workshop" to reflect the breadth of the impact of heterogeneity, as well as to stress that the focus of the workshop is on the management and exploitation of heterogeneity. All of this is, of course, taken in the context of the parent conference, the International Parallel and Distributed Processing Symposium (IPDPS), and so explores heterogeneity in distributed and parallel computing systems. Arnold "Arny" Rosenberg, from the University of Massachusetts at Amherst, a past HCW Program Committee Chair and General Chair, suggested this name change, and it was approved by the HCW Steering Committee (listed elsewhere in these proceedings).

Heterogeneity in parallel and distributed computing systems is a very important research area with great practical impact for 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. I encourage you to look also at the proceedings of past and future HCWs.

Many people have worked very hard to make this year's workshop happen. Henri Casanova, a well known and respected researcher from the University of Hawai`i, was this year's Program Committee Chair. He worked diligently with the Program Committee (listed elsewhere in these proceedings) to assemble the excellent program for the workshop. José A. B. Fortes, from the University of Florida, was the General Chair. José has been a personal friend and colleague of mine for over 20 years, and I greatly respect him as person and as a professional. José was responsible for the overall organization and administration of this year's workshop, and he did a fine job. Given his experience as the Program Chair of HCW 2006, he provided valuable guidance to Henri (as Henri will do as the General Chair for HCW 2008). I thank Henri, José, and the Program Committee for their efforts. I also thank the workshop Steering Committee for their oversight and assistance.

1-4244-0910-1/07/$20.00 ©2007 IEEE

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 Heterogeneity in 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 16 years.

H. J. Siegel
Colorado State University

1-4244-0910-1/07/$20.00 ©2007 IEEE

Message from the HCW General Chair

Welcome to HCW 2007, the 16th workshop of the HCW series of meetings. Starting this year, on the recommendation of the Steering Committee and feedback from HC researchers, the acronym HCW stands for Heterogeneity in Computing Workshop. This reflects the pervasive nature of heterogeneity, which now impacts all aspects of computing, including not only platforms and algorithms, but also systems software, interfaces, computing models, applications, etc. We hope that HCW 2007 will be the first of a long series of meetings where international scientists gather to present and exchange ideas on this increasingly important area of computing research.

This year's workshop continues to build on the successful association of HCW with the IPDPS meeting. The framework and infrastructure support provided by the IPDPS umbrella for its associated workshops, including HCW, has proven to be a mutually beneficial arrangement. It is a pleasure to acknowledge the success and efforts of the IPDPS and HCW organizing teams in this continuing relationship.

Among the many people who have contributed to HCW 2007, several deserve a special mention and thanks. Henri Casanova, the Program Chair, did an excellent job in all tasks (and more) needed to bring about an exciting technical program: constituting a high-quality program committee, organizing the review process, identifying keynote speakers and supporting the workshop website. H. J. Siegel, the Steering Chair Committee, was instrumental in providing advice and continuity wisdom when and where needed in the process of organizing HCW. Alan Sussman, the IPDPS Workshops Committee Chair, coordinated and provided a clear liaison between HCW and the IPDPS organization.

Last but not least, 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 Advanced Computing and Information Systems (ACIS) laboratory of the University of Florida.

I wish you all a productive and enjoyable workshop in sunny Long Beach, California.

José A. B. Fortes
University of Florida

1-4244-0910-1/07/$20.00 ©2007 IEEE

Message from the HCW Program Chair

Let me add my welcome to HCW 2007 to those of the Steering Committee and General Chair. The international Technical Program Committee (TPC) has invested a lot of time and effort in putting together a program that we hope workshop attendees and participants will find technically valuable as well as intellectually stimulating.

The 10 papers in the proceedings were selected from 21 submissions based on several criteria, including technical soundness, originality, and appropriateness to the workshop. Each submission was reviewed by at least three TCP members, possibly with the help of external reviewers. I wish to thank all TCP members for their hard work in evaluating and discussing the papers. Special thanks are due to external reviews: Vineet Chadha, Kaoutar El Maghraoui, Emmanuel Jeannot, Jean-Sébastien Gay. Finally, I wish to thank all authors who submitted their work to HCW'2007.

The entire HCW'2007 Organizing Committee is delighted to have two outstanding Keynote Speakers this year. The first speaker is Dr. Thomas Sterling, a Professor in the Dept. of Computer Science and the Center of Computation and Technology at Louisiana State University, a Faculty Associate at the Center for Advance Computation Research at the California Institute of Technology, and a Distinguished Visiting Scientist at the Computing and Computational Sciences Directorate at the Oak Ridge National Laboratory. The second speaker is Dr. Dean Tullsen, a Professor in the Computer Science and Engineering at the University of California, San Diego. Both speakers will share with workshop attendees their unique perspectives on heterogeneous computing today and their vision for the future.

I am personally very grateful to H.J. Siegel, Chair of the HCW Steering Committee, and José Fortes, General Chair of HCW'2007, for their sound advice and constant guidance for putting together the HCW program. I also want to thank Alan Sussman, the IPDPS Workshop Chair, for helping with the logistics of the paper reviewing system and with publication of the proceedings, and for always being quick to answer the many questions I had along the way.

Henri Casanova
University of Hawai`i at Manoa

1-4244-0910-1/07/$20.00 ©2007 IEEE

Keynote – ParalleX: An Asynchronous Execution Model for Scalable Heterogeneous Computing

Thomas Sterling[1,2,3]

[1]*Dept. of Computer Science, Center of Computation and Technology*
Louisiana State University
Baton Rouge, Louisiana, U.S.A.
tron@cct.lsu.edu

[2]*Center for Advance Computation Research*
California Institute of Technology
Pasadena, California, U.S.A.

[3]*Computing and Computational Science Directorate*
Oak Ridge National Laboratory
Oak Ridge, Tennessee, U.S.A.

Heterogeneous system architecture has long been appreciated as a potential strategy for achieving super-linear speedup with respect to some normalizing parameter like number of nodes, cost, or power. However the challenge of programming and managing the system resources has been a limiting factor for the application of such systems on a broad scale. The exception has been the use of special purpose processors such as graphics processing units that may yield dramatic increases for such functionality enabling capabilities largely impossible otherwise like realistic high resolution real time interactive games. However, with power emerging as the dominant constraint on high performance computing and the need to make better use of logic and storage resources such components as the ClearSpeed SIMD attached processor and the IBM cell architectures among others is forcing mainstream computing to adopt heterogeneous processing. This keynote presentation will describe a computational model, ParalleX that provides an asynchronous runtime framework for supporting effective execution in an environment comprising heterogeneous elements. ParalleX is based on a message-driven split-phase multithreaded transaction processing paradigm synthesizing a number of concepts represented in prior art that in ensemble will facilitate management of heterogeneous resources and provide the basis for a systematic programming methodology. Also discussed in this presentation is another example of a heterogeneous architecture, Gilgamesh II, that provides separate mechanisms for computations that exhibit disparate locality properties.

Study of an Iterative Technique to Minimize Completion Times of Non-Makespan Machines

Luis Diego Briceño[1], Mohana Oltikar[1], Howard Jay Siegel[1,2] and Anthony A. Maciejewski[1]

[1]*Department of Electrical and Computer Engineering*
Colorado State University
Fort Collins, Colorado, USA
{*ldbricen, mohanar, hj, aam*}*@colostate.edu*

[2]*Department of Computer Science*
Colorado State University
Fort Collins, Colorado, USA

Heterogeneous computing (HC) is the coordinated use of different types of machines, networks, and interfaces to maximize the combined performance and/or cost effectiveness of the system. Heuristics for allocating resources in an HC system have different optimization criteria. A common optimization criterion is to minimize the completion time of the last to finish machine (makespan). In some environments, it is useful to minimize the finishing times of the other machines in the system, i.e., those machines that are not the last to finish. Consider a production environment where a set of known tasks are to be mapped to resources off-line before execution begins. Minimizing the finishing times of all the machines will provide the earliest available ready time for these machines to execute tasks that were not initially considered. In this study, we examine an iterative approach that decreases machine finishing times by repeatedly running a resource allocation heuristic. The goal of this study is to investigate whether this iterative procedure can reduce the finishing time of some machines compared to the mapping initially generated by the heuristic. We show that the effectiveness of the iterative approach is heuristic dependent and study the behavior of the iterative approach for each of the chosen heuristics. This work which identifies heuristics can and cannot attain improvements in the completion time of non-makespan machines using this iterative approach.

Bi-criteria Scheduling Algorithm with Deployment in Cluster

Feryal-Kamila MoulaÏ and Gregory Mounie

LIG
Grenoble University
Grenoble, Isère, France
feryal-kamila.moulai@imag.fr

Computational grids clusters, provide powerful computing resources for executing applications of large scale. In Grid (clusters) usually several applications run simultaneously. The originality of Grid'5000 is that each application has characterized by its own specific requirement such as operating system (OS) or library components. Deploying the adequate OS needs to reboot the processors on which the application is executed. It is time-consuming and moreover frequent reboots may damage machines. In this work we investigate how to minimize the number of deployments, while keeping the running time as short as possible. We present the multiprocssors scheduling with deployment problem and provides a list scheduling algorithm. The analysis details are presented in the worst case performance of the algorithm.

Optimal Assignment of a Tree-Structured Context Reasoning Procedure onto a Host-Satellites System

Hailiang Mei, Pravin Pawar and Ing Widya

Dept. of Computer Science
University of Twente
Enschede, The Netherlands
{H.Mei, P.Pawar, I.A.Widya}@utwente.nl

In this paper, we study the problem of an optimal assignment of a tree-structured context reasoning procedure onto the computation resources in a host-satellites configuration. The objective function to be minimized is the end-to-end processing delay, which is a crucial factor in a number of context- aware applications, e.g. mobile healthcare applications. The presented solution is a modification of an earlier method proposed by Bokhari, in which the optimal assignment problem to minimize the bottleneck processing time is transformed into a path-searching problem in a doubly weighted graph. Due to the incompatible requirements raised in our study, e.g. a-prior known location of the sensors, we propose a colouring scheme and a new search algorithm in this paper to obtain the optimal assignment in order to satisfy our objective.

PFAS: A Resource-Performance-Fluctuation-Aware Workflow Scheduling Algorithm for Grid Computing

Fangpeng Dong and Selim G. Akl

School of Computing
Queen's University
Kingston, ON, Canada
{*dong, akl*} *@cs.queensu.ca*

Resource performance in the Computational Grid is not only heterogeneous, but also changing dynamically. However scheduling algorithms designed for traditional parallel and distributed systems, such as clusters, only consider the heterogeneity of the resources. In this paper, a workflow scheduling algorithm, called PFAS, is proposed and tested in the Grid environment. PFAS considers dynamic resource performance fluctuation in the Grid, and conducts the scheduling according to its knowledge of the fluctuation. This new algorithm works in an offline way which allows it to be easily set up and run with less cost. Simulations show that our approach can achieve better schedules than the HEFT algorithm.

Stochastic Approach to Scheduling Multiple Divisible Tasks on a Heterogeneous Distributed Computing System

Ankur Kamthe[1] and Soo-Young Lee[2]

[1]*Computer Science and Engineering*
University of California, Merced
Merced, CA, USA
akamthe@ucmerced.edu

[2]*Dept. of Electrical and Computer Engineering*
Auburn University
Auburn, AL, USA
leesooy@eng.auburn.edu

Heterogeneity has been considered in scheduling, but without taking into account the temporal variation of completion times of the sub-tasks for a divisible, independent task. In this paper, the problem of scheduling multiple, divisible independent tasks on a heterogeneous distributed computing system is addressed. The "stochastic" approach, which was previously applied to DAG scheduling, is employed for scheduling a group of multiple divisible as well as whole independent tasks. It explicitly considers the standard deviations (temporal heterogeneity) in addition to the mean execution times in deriving a schedule, in order to model more closely what would actually happen "on average" on a temporally heterogeneous system (instead of approximating the random weights by their means only as in other approaches). Through an extensive computer simulation, it has been shown that the proposed approach can improve schedules significantly over those by a scheme which uses the average weights only.

Load Balancing in the Bulk-Synchronous-Parallel Setting using Process Migrations

Olaf Bonorden

Heinz Nixdorf Institute, Computer Science Department
University of Paderborn
Paderborn, Germany
bono@uni-paderborn.de

The Paderborn University BSP (PUB) library is a powerful C library that supports the development of bulk synchronous parallel programs for various parallel machines. To utilize idle times on workstations for parallel computations, we implement virtual processors using processes. These processes can be migrated to other hosts, when the load of the machines changes. In this paper we describe the implementation for a Linux workstation cluster. We focus on process migration and show first benchmarking results. In contrast to other implementations, we do not make use of a kernel module or patch, thus the user does not need administration access to the computers.

Strategies for Replica Placement in Tree Networks

Anne Benoit, Veronika Rehn and Yves Robert

Laboratoire LIP
ENS Lyon
Lyon, France
{*anne.benoit, veronika.rehn, yves.robert*}*@ens-lyon.fr*

In this paper, we discuss and compare several policies to place replicas in tree networks, subject to server capacity constraints. The client requests are known beforehand, while the number and location of the servers are to be determined. The standard approach in the literature is to enforce that all requests of a client be served by the closest server in the tree. We introduce and study two new policies. In the first policy, all requests from a given client are still processed by the same server, but this server can be located anywhere in the path from the client to the root. In the second policy, the requests of a given client can be processed by multiple servers.

One major contribution of this paper is to assess the impact of these new policies on the total replication cost. Another important goal is to assess the impact of server heterogeneity, both from a theoretical and a practical perspective. In this paper, we establish several new complexity results, and provide several efficient polynomial heuristics for NP-complete instances of the problem. These heuristics are compared to an absolute lower bound provided by the formulation of the problem in terms of the solution of an integer linear program.

High-Performance Multi-Rail Support with the NewMadeleine Communication Library

Olivier Aumage, Elisabeth Brunet, Guillaume Mercier and Raymond Namyst

Inria, LaBRI
Université Bordeaux 1
Talence, France
{*aumage, brunet, mercier, namyst*}*@labri.fr*

This paper focuses on message transfers across multiple heterogeneous high-performance networks in the NewMadeleine Communication Library. NewMadeleine features a modular design that allows the user to easily implement load-balancing strategies efficiently exploiting the underlying network but without being aware of the low-level interface. Several strategies are studied and preliminary results are given. They show that performance of network transfers can be improved by using carefully designed strategies that take into account NIC activity.

Enhancing Portability of HPC Applications across High-end Computing Platforms

Magdalena Slawinska, Jaroslaw Slawinski, Dawid Kurzyniec and Vaidy Sunderam

Mathematics and Computer Science
Emory University
Atlanta, Georgia, USA
{*magg, jaross, dawidk, vss*}*@mathcs.emory.edu*

Fast hardware turnover in supercomputing centers, stimulated by rapid technological progress, results in high heterogeneity among HPC platforms, and necessitates that applications are ported and adapted frequently. The cutting-edge nature of the hardware mandates customized performance tuning, which, coupled with continuously growing application complexity, makes the process inherently and increasingly challenging. In this paper, we analyze build procedures of a representative set of HPC applications, and attempt to identify commonalities that can be exploited to enhance cross-platform portability. We then propose a novel method for reducing non-portabilities while preserving high performance. The approach, based on *profiles* that capture and isolate non-portable features at various levels, requires only a moderate amount of changes to the existing makefiles. It leverages expertise of system designers and administrators, reducing burdens placed on application scientists. As a proof of concept, we discuss the application of our methodology to enhancing portability of the Milc application across heterogeneous HPC platforms.

Domain Decomposition vs. Master-Slave in Apparently Homogeneous Systems

Cyril Banino-Rokkones

Dept. of Computer & Information Science
Norwegian University of Science and Technology (NTNU)
Trondheim, NORWAY
banino@idi.ntnu.no

This paper investigates the utilization of the master-slave (MS) paradigm as an alternative to domain decomposition (DD) methods for parallelizing lattice gauge theory (LGT) models within distributed memory environments. The motivations for this investigation are twofold. First, LGT models are inherently difficult to parallelize efficiently with DD methods. Second, DD methods have proven useful for homogeneous environments, but are impractical for heterogeneous and dynamic environments. Besides, many modern supercomputer architectures that look homogeneous (such as multi-core or SMP), are in fact heterogeneous and dynamic environments. We highlight this issue by comparing a traditional first-come first-served MS implementation to a simple but yet efficient selective MS scheduling strategy that automatically accounts for system heterogeneity and variability. Our experimental results with the parallelization of our LGT model, reveal that the selective MS implementation achieves good efficiency, but lacks of scalability. In contrast, the DD method is highly scalable, but at the expense of a poor efficiency. These results open up for a hybrid approach, where the MS and the DD methods would be combined for achieving scalable high performance.

Keynote – Holistic Design of Multi-Core Architectures

Dean Tullsen

Department of Computer Science and Engineering
University of California, San Diego
La Jolla, Louisiana, U.S.A.
tullsen@cs.ucsd.edu

Several forces are driving the market to put multiple execution cores on a single processor chip. But we cannot view (or design) those cores (and the connections between them) in the same way we did when we lived in a uniprocessor world. Previously, we expected each core to provide good performance on virtually any application, with energy efficiency, and without error or failure. Now that the level of interface with the user and the system is a multi-core chip, those requirements need only be met at the chip level – no single core need meet them. This provides the opportunity to think about processor architecture in whole new ways. This talk will describe holistic design of a multi-core architecture – designing cores, caches, interconnect so that the chip as a whole provides maximum performance, high energy efficiency, and high performance per area. We will discuss, in particular, on-chip heterogeneous multiprocessing and conjoined core architectures.

Workshop 2
Workshop on Parallel and Distributed Real-Time Systems
WPDRTS 2007

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:

- General Topics: Adaptive and reflective real-time systems, Applications, benchmarks, and tools, Architectures and hardware/software co-design, Distributed real-time and embedded middleware, Fault-tolerance, security, and robustness, Real-time operating systems, Real-time and embedded databases, Soft real-time and mixed-critical systems, Algorithms and Applications, QoS based resource management and real-time scheduling, Programming languages and environments, Specification, modeling, and analysis of real-time systems, Certification of resource managers, Real-time communication protocols and architecture.
- Formal methods for distributed real-time systems: Verification and validation, Formal techniques for performance evaluation, Formal modeling in systems design, Scheduling and optimization, Schedulability analysis, Case studies.
- Automotive systems: network integration in the automotive domain, design space exploration with respect to realtime issues, analysis techniques for automotive specific application domains, bus systems or operating systems.
- Certification of Dynamic and Adaptive Systems: intelligent instrumentation, statistical analysis, to defining, testing, and analysis of operating boundaries and boundedness.
- Wireless sensor networks: Communication protocols, high-level operating system and programming abstractions, middleware and service architectures, configuration management, testbeds, in-network information processing, security, novel applications and experience reports, resource discovery and management, QoS issues, disconnected and weakly-connected WSNs, tools and methodologies forbuilding WSNs.

General Chairs:
Zdenek Hanzalek, Czech Technical University in Prague, Czech Republic
Chenyang Lu, Washington University

Program Co-chairs:
Frank Drews, Ohio Univ.
Angelika Mader, Univ. of Twente, The Netherlands

Program Committee:
Sanjoy Baruah, Univ. of North Carolina
Ed Brinksma Univ. of Twente & Embedded Systems Institute, Netherlands
Balakrishnan Dasarathy, Telcordia Technologies
Maryline Chetto, Universite de Nantes, France
Chris D. Gill, Washington Univ.
Aniruddah Gokhalé, Vanderbilt Univ.
Jeffery P. Hansen, CMU
Jozef Hooman, Radboud Univ. Nijmegen & Embedded Systems Institute, Netherlands
Pierre Jansen, Univ. of Twente, Netherlands
Robert P. Judd, Ohio Univ.
David Juedes, Ohio Univ.
Odej Kao, Technical Univ. of Berlin, Germany
Xenofon D. Koutsoukos, Vanderbilt Univ.
Patrick Lardieri, Lockheed Martin
Victor Lee, City Univ. of Hong Kong, China
Giuseppe Lipari, Scuola Superiore S.Anna, Italy
Joseph P. Loyall, BBN Technologies
Daniel Mossé Univ. of Pittsburgh
Paulo Pedreiras, Univ. of Aveiro, Portugal
Ismael Ripoll, Polytechnic Univ. of Valencia, Spain
John M. Slaby, Raytheon Integrated Defense Systems
Oleg Sokolsky, Univ. of Pennsylvania
Peter van der Stok, Philips Research Laboratories, Netherlands
Francisco Vasques, Univ. of Porto, Portugal

Publicity Chair
Dazhang Gu, Ohio University

Publication Chair
William Leal, Ohio University

Submission Chair
Jelena Marincic, University of Twente, The Netherlands

Special Session Chairs
Ansgar Fehnker University of New South Wales, Australia
Michaela Huhn, University of Braunschweig, Germany

Steering Committee Chairs
Chris Gill, Washington Univ.
Vana Kalogeraki, Univ. of California

Steering Committee
David Andrews, Univ. of Kansas
Scott Brandt, Univ. of California at Santa Cruz
Lisa DiPippo, Univ. of Rhode Island
Klaus Ecker, TU Clausthal, Germany
G. Manimaran, Iowa State Univ.
Priya Narasimhan, Carnegie-Mellon Univ.
Barbara Pfarr, NASA Goddard
Viktor Prasanna, USC
Behrooz Shirazi, Univ. of Texas at Arlington
Peter van der Stok, Philips Research Laboratories, The Netherlands
Lonnie R. Welch, Ohio Univ.
Paul R. Work, Raytheon Company
Armin Zimmermann, Technical Univ. of Berlin, Germany

Competitive Analysis of Partitioned Scheduling on Uniform Multiprocessors

Björn Andersson and Eduardo Tovar

IPP-HURRAY! Research group
Institute Polytechnic Porto, ISEP/IPP
Porto, Portugal
{*bandersson, emt*}*@dei.isep.ipp.pt*

Consider the problem of scheduling a set of sporadically arriving tasks on a uniform multiprocessor with the goal of meeting deadlines. A processor p has the speed S_p. Tasks can be preempted but they cannot migrate between processors. We propose an algorithm which can schedule all task sets that any other possible algorithm can schedule assuming that our algorithm is given processors that are three times faster.

Integrated Environment for Embedded Control Systems Design

Roman Bartosinski[1], Zdenek Hanzalek[2], Petr Struzka[3] and Libor Waszniowski[2]

[1]*Department of Signal Processing*
Czech Academy of Science
Prague, Czech Republic, Czech Republic
bartosr@utia.cas.cz

[2]*Department of Control Engineering*
Czech Technical University
Prague, Czech Republic, Czech Republic
{*hanzalek, xwasznio*}*@fel.cvut.cz*

[3]*Department of Embedded Systems*
UNIS, Ltd.
Brno, Czech Republic, Czech Republic
pstruzka@unis.cz

The motivation of our work is to make a design tool for distributed embedded systems compliant with HIS and AUTOSAR. The tool is based on Processor Expert, a component oriented development environment supporting several hundreds of microcontrollers, and Matlab Simulink which is the de-facto standard in the rapid prototyping of the control applications but it does not have an adequate HW support. The objective is to provide an integrated development environment for embedded controllers having distributed nature and real-time requirements. Therefore we discuss the advantages of using an automatically generated code in the development cycle of the control embedded software. We present a developed block set and Processor Expert Real-Time Target for Matlab Real-Time Workshop Embedded Coder. The case study shows a development cycle for a servo control design.

Improved Schedulability Analysis of EDF Scheduling on Reconfigurable Hardware

Nan Guan[1], Zonghua Gu[2], Qingxu Deng[1], Weichen Liu[2] and Ge Yu[1]

[1]Computer Science and Engineering
Northeastern University
Shenyang, China
guannan0609@hotmail.com, {dengqx, yuge}@mail.neu.edu.cn

[2]Computer Science and Engineering
HKUST
Hong Kong, China
{zgu, weichen}@ust.hk

Reconfigurable devices, such as Field Programmable Gate Arrays (FPGAs), are very popular in todays embedded systems design due to their low-cost, high-performance and flexibility. Partially Runtime-Reconfigurable (PRTR) FPGAs allow hardware tasks to be placed and removed dynamically at runtime. Hardware task scheduling on PRTR FPGAs brings many challenging issues to traditional real-time scheduling theory, which have not been adequately addressed by the research community compared to software task scheduling on CPUs. In this paper, we consider the schedulability analysis problem of HW task scheduling on PRPR FPGAs. We derive utilization bound tests for two variants of global EDF scheduling, and use synthetic tasksets to compare performance of the tests to existing work and simulation results.

The Design and Implementation of Real-time Event-based Applications with RTSJ

Damien Masson and Serge Midonnet

Laboratoire d'informatique de l'institut Gaspard-Monge
Université de Marne-la-Vallee
Champs sur Marne, France
damien.masson@univ-mlv.fr, serge.midonnet@esigetel.fr

This paper presents a framework to design real-time event-based applications using Java. The Real-Time Specification for Java (RTSJ) is well designed for hard periodic real-time systems. Though it also proposes classes to model asynchronous events and deal with sporadic or aperiodic tasks, it remains insufficient. The literature proposes the use of periodic servers called tasks servers to handle non periodic traffics in real-time systems. Unfortunately, there is no support for task servers in RTSJ. In order to fix this lack, we propose an RTSJ extension model. To validate our design, we adapt and implement two policies: the polling server and the deferrable server policies. To show how these policies are efficient, we compare implementation results and results we obtain with a discrete-event-based simulator.

Using Speed Diagrams for Symbolic Quality Management

Jacques Combaz[1], Jean-Claude Fernandez[2], Joseph Sifakis[3] and Loïc Strus[4]

[1] *Verimag*
Gières, France
jacques.combaz@imag.fr

[2] *Verimag*
Gières, France
jean-claude.fernandez@imag.fr

[3] *Verimag*
Gières, France
joseph.sifakis@imag.fr

[4] *Verimag*
Gières, France
loic.strus@imag.fr

We present a quality management method for multimedia applications. The method takes as input an application software composed of actions. The execution times of actions are unknown increasing functions of quality level parameters. The method allows the construction of a Quality Manager which computes adequate action quality levels so as to meet QoS requirements for a given platform. These include deadlines for the actions as well as quality maximization and smoothness.

We extend and improve results of a previous paper by focusing on the reduction of overhead due to quality management. We propose a symbolic quality management method using speed diagrams, a representation of the system's dynamics. Instead of numerically computing a quality level for each action, the Quality Manager changes action quality levels based on the knowledge of constraints characterizing control relaxation regions. These are sets of states in which quality management for a given number of steps can be relaxed without degrading quality.

We provide experimental results for quality management of an MPEG encoder, in particular performance benchmarks for both numeric and symbolic quality management.

A Flexible Scheme for Scheduling Fault-Tolerant Real-Time Tasks on Multiprocessors

Michele Cirinei[1], Enrico Bini[1], Giuseppe Lipari[1] and Alberto Ferrari[2]

[1] *ReTiS Laboratory*
Scuola Superiore Sant'Anna
Pisa, Italy
{m.cirinei, e.bini, lipari}@sssup.it

[2] *PARADES EEIG*
Roma, Italy
aferrari@parades.cnr.rm.it

The recent introduction of multicore system-on-a-chip architectures for embedded systems opens a new range of possibilities for both increasing the processing power and improving the fault-robustness of real-time embedded applications. Fault-tolerance and performance are often contrasting requirements. Techniques to improve robustness to hardware faults are based on replication of hardware and/or software. Conversely, techniques to improve performance are based on exploiting inherent parallelism of multiprocessor architectures.

In this paper, we propose a technique that allows the user to trade-off parallelism with fault-tolerance in a multicore hardware architecture. Our technique is based on a combination of hardware mechanisms and real-time operating system mechanisms. In particular, we apply hierarchical scheduling techniques to efficiently support fault-tolerant, fault-silent and non-fault-tolerant tasks in the same system.

Expected Time for Obtaining Dependable Data in Real-Time Environment

Yue Yu and Shangping Ren

Department of Computer Science
Illinois Institute of Technology
Chicago, Illinois, USA
{*yyu8, ren*}*@iit.edu*

In real-time environment, data usually has a lifespan associated with it. The semantics and the importance of the data depend on the time when data is utilized. Hence, the process of getting a consensus data from a group of replicated units must not take longer time than the lifespan of the data. However, in real environments, every unit, faulty or non-faulty, may encounter delays when processing and sending their data which inevitably increases the time of acquiring a consensus. The latency for obtaining a valid data hence depends not only on the time when individual voters make their votes, but also on the accuracy and credibility of the votes. Thus, a new metric, i.e. credibility function, need be taken into account in evaluating expected time and deciding replications. This paper presents analytical solutions for expected time under different voting schemes when dependable data can be obtained. We also show that if not all voters are truthful, adding more replications does not improve much on the time of obtaining valid results.

Static-Priority Scheduling and Resource Hold Times

Marko Bertogna[1], Nathan Fisher[2] and Sanjoy K. Baruah[3]

[1]*Computer Science*
Scuola Superiore S.Anna
Pisa, Italy
marko@sssup.it

[2]*Computer Science*
University of North Carolina
Chapel Hill, North Carolina, USA
fishern@cs.unc.edu

[3]*Computer Science*
University of North Carolina
Chapel Hill, North Carolina, USA
baruah@cs.unc.edu

The duration of time for which each application locks each shared resource is critically important in composing multiple independently-developed applications upon a shared "open" platform.

In a companion paper, we formally defined and studied the concept of resource hold time (RHT) — the largest length of time that may elapse between the instant that an application system locks a resource and the instant that it subsequently releases the resource.

We extend the discussion and results to systems scheduled using static-priority scheduling algorithms, with resource access arbitrated using Stack Resource Policy (SRP), or Priority Ceiling Protocol (PCP). We present a method to compute resource hold times for every resource, and an algorithm to decrease them without changing the semantics of the application or compromising application feasibility.

Tiresias: Black-Box Failure Prediction in Distributed Systems

Andrew Williams[1], Soila Pertet[2] and Priya Narasimhan[3]

[1]*Electrical and Computer Engineering*
Carnegie Mellon
Pittsburgh, PA, United States
andrewwi@andrew.cmu.edu

[2]*Electrical and Computer Engineering*
Carnegie Mellon
Pittsburgh, PA, United States
spertet@ece.cmu.edu

[3]*Electrical and Computer Enginnering*
Carnegie Mellon
Pittsburgh, PA, United States
priya@cs.cmu.edu

Faults in distributed systems can result in errors that manifest in several ways, potentially even in parts of the system that are not collocated with the root cause. These manifestations often appear as deviations (or "errors") in performance metrics. By transparently gathering, and then identifying escalating anomalous behavior in, various node-level and system-level performance metrics, the Tiresias system makes black-box failure-prediction possible. Through the trend analysis of performance metrics, Tiresias provides a window of opportunity (look-ahead time) for system recovery prior to impending crash failures. We empirically validate the heuristic rules of the Tiresias system by analyzing fault-free and faulty performance data from a replicated middleware-based system.

Toward a Unified Standard for Worst-Case Execution Time Annotations in Real-Time Java

Trevor Harmon and Raymond Klefstad

Dept. of Electrical Engineering and Computer Science
University of California, Irvine
Irvine, California, USA
{tharmon, klefstad}@uci.edu

As real-time systems become more prevalent, there is a need to guarantee that these increasingly complex systems perform as designed. One technique involves a static analysis to place an upper bound on worst-case execution time (WCET). This temporal analysis cannot be made automatic and normally requires source annotations to assist a WCET analysis tool.

At the same time, there is a growing interest in using Java for real-time systems. Several WCET analysis prototypes for Java have been created, and more are under development. Each relies on a competing and incompatible convention for annotations, resulting in portability problems and duplication of effort.

We propose that Java's own annotation mechanism should be used to address such issues. These built-in annotations provide a common platform for WCET analysis, improving portability and reducing the effort necessary to create these vital tools. We examine the features that make Java's annotation standard attractive for WCET analysis, then discuss its current failings and make recommendations for future improvements.

Hardware Capacity Evaluation in Shared-Nothing Data Warehouses

Ricardo Antunes and Pedro Furtado

Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal
{*rantunes, pnf*}*@dei.uc.pt*

Parallel data warehouses have mosty been seen as dedicated decision support systems, but as the need for cost effective multi-application business solutions grows, non-dedication and adaptation to existing environments become evermore tempting.

Data placement in such systems is a major problem seeing as existing environments may have heterogeneous nodes, which process data at different rates, and non-dedicated environments can not promise full resource commitment to execute a query. Therefore it is critical to distribute a data warehouses information in such a way that no node be overloaded or underestimated.

This paper presents the Capacity Evaluator (CE), an application capable of measuring the systems ability to process information, which in turn helps the Data Warehouse Parallel Architectures (DWPA) automatic data placer determine exactly how much data should be allocated to each node.

Scalable, Distributed, Dynamic Resource Management for the ARMS Distributed Real-Time Embedded System

Kurt Rohloff, Yarom Gabay, Jianming Ye and Richard Schantz

National Intelligence Research and Applications
BBN Technologies
Cambridge, MA, USA
{*krohloff, ygabay, jye, schantz*}*@bbn.com*

We present a scalable, hierarchical control system for the dynamic resource management of a distributed real-time embedded (DRE) system. This DRE is inspired by the DARPA Adaptive and Reflective Middleware Systems (ARMS) program. The goal of the control system is to simultaneously manage multiple resources and QoS concerns using a utility-driven approach for decision making and performance evaluation. At each level of the control hierarchy there are multiple local controllers which autonomously make decisions to optimize their local utility. The controllers in the hierarchy can use different, localized resource control algorithms and the systems user can tune the operations of the local controllers. We discuss how the selections of local control algorithms affect the behavior of the overall system. The control system is designed to be easily adaptable to other multi-tiered DRE systems.

Capacity Sharing and Stealing in Dynamic Server-based Real-Time Systems

Luís Nogueira and Luís Miguel Pinho

IPP-HURRAY
Polytechnic Institute of Porto
Porto, Portugal
{*luis, lpinho*}*@dei.isep.ipp.pt*

This paper proposes a dynamic scheduler that supports the coexistence of guaranteed and non-guaranteed bandwidth servers to efficiently handle soft-tasks' overloads by making additional capacity available from two sources: (i) residual capacity allocated but unused when jobs complete in less than their budgeted execution time; (ii) stealing capacity from inactive non-isolated servers used to schedule best-effort jobs. The effectiveness of the proposed approach in reducing the mean tardiness of periodic jobs is demonstrated through extensive simulations. The achieved results become even more significant when tasks' computation times have a large variance.

A Framework for Modeling Operating System Mechanisms in the Simulation of Network Protocols for Real-Time Distributed Systems

Paolo Pagano, Prashant Batra and Giuseppe Lipari

ReTiS Laboratory
Scuola Superiore Sant'Anna
Pisa, Italy
{*Paolo.Pagano, Prashant.Batra, Giuseppe.Lipari*}*@sssup.it*

In this paper we present a software tool for the simulation of distributed real-time embedded systems. Our tool is based on the popular NS-2 package for simulating the networking aspects, and on the RTSim package for the real-time operating system aspects. By reusing much of the existing code, our simulator covers a very wide range of network protocols and real-time mechanisms.

After describing the architecture of our tool, we tested it in a simple wireless sensor networks scenario, and we measured the latency in transmitting and receiving messages due to the concurrent activities in the nodes. These effects have been tested against two node scheduling policies, and under different load conditions in the CPU of the nodes.

Authentication in Reprogramming of Sensor Networks for Mote Class Adversaries

Limin Wang and Sandeep Kulkarni

Department of Computer Science & Engineering
Michigan State University
East Lansing, MI, USA
{*wanglim1, sandeep*}*@cse.msu.edu*

Reprogramming is an essential service for wireless sensor networks. Authenticating reprogramming process is important as sensors need to verify that the code image is truly from a trusted source. There are two ways to achieve authentication: public key based and symmetric key based. Although previous work has shown that public key authentication is feasible on sensor nodes if used sparingly, it is still quite expensive compared to symmetric key based approach. In this paper, we propose a symmetric key based protocol for authenticating reprogramming process. Our protocol is based on the secret instantiation algorithm from [5,11], which requires only $O(\log n)$ keys to be maintained at each sensor. We integrate this algorithm with the existing reprogramming protocol. Through simulation, we show that it is able to authenticate reprogramming process at very low communication cost, and has very short delay.

Period-Dependent Initial Values for Exact Schedulability Test of Rate Monotonic Systems

Wan-Chen Lu[1], Kwei-Jay Lin[2], Hsin-Wen Wei[1] and Wei-Kuan Shih[1]

[1]*Computer Science*
National Tsing Hua University
Hsinchu, Taiwan
{*wanchen, bertha, wshih*}*@rtlab.cs.nthu.edu.tw*

[2]*Electrical Engineering and Computer Science*
University of California
Irvine, California, USA
klin@uci.edu

Real-time systems using Rate Monotonic fixed priority scheduling can be checked for schedulability either by pessimistic schedulability conditions or exact testing. Exact testing provides a more precise result but cannot always be performed in polynomial time. Audsley et al. proposed one of the earliest methods by iteratively deriving the job response times. Other researchers have improved the efficiency of their exact test method by using different initial values. All currently proposed initial values do not use the relationship between task periods. In this paper we define initial values using the largest and the second largest periods in a system. We show that the new initial values can significantly improve the exact test.

Towards a Distributed Continuous Certification Process

Adam Porter

Computer Science
University of Maryland
College Park, MD, USA
aporter@cs.umd.edu

Software scale and complexity are growing by every measure: more hardware and software, more communication links, more interdependency, more lines of code, more storage and data, etc. At the same time business trends are increasingly squeezing development resources. In particular, development processes are straining under severe cost and time-to-market pressures. Global competition and market deregulation are shrinking profit margins and thus limiting budgets for the development and QA of software.

In response to these trends, developers have begun to change the way they build and validate software systems by (among other things) moving towards more flexible product designs allowing dynamic reconfiguration.

This approach promises to improve cost, quality, and development-time, but creates other problems, especially when used in the context of safety-critical systems.

To realize this promise, however, effective certification becomes more important than ever since as static controls are removed or reduced, it becomes even more vital that (1) problems be caught as quickly as possible and (2) systems not be allowed to drift so far from their intended functional and performance requirements that rework costs overwhelm the hoped-for efficiencies.

This article will present and discuss some of our recent efforts to address these problems.

Special Session on Certification of Dynamic and Adaptive Systems

Paul R Work

Raytheon Company
Portsmouth, RI, USA
paul_r_work@raytheon.com

As part of the 15th International Workshop on Parallel and Distributed Real-Time Systems, 26-27 March 2007, Long Beach, CA, to be held in conjunction with IPDPS 2007, there will be a Special Session focused on the topic of Certification of Dynamic and Adaptive Systems.

The verification, validation, and eventual certification of dynamic and adaptive systems is a challenging set of activities both intellectually and, at this time, physically, due to the limits of the state of research and technology in this area. Many more systems are being built using todays dynamic technologies to achieve significant operational capabilities in a timely manner and yet some will need to operate safely and all will need to perform reliably. This is complicated further if these solutions need to do so in a low latency environment with little to no failure cases. The purpose of this session is to bring to light research and development being performed (or even, just being dreamed of) to look at the scalability problems with certifying dynamic and adaptive solutions.

Formal Analysis of Time-Dependent Cryptographic Protocols in Real-Time Maude

Peter Csaba Ölveczky and Martin Grimeland

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

This paper investigates the suitability of applying the general-purpose Real-Time Maude tool to the formal specification and model checking analysis of time-dependent cryptographic protocols. We restrict the intruders so that they become *non-Zeno*, and propose a complete analysis method for finding attacks that are reachable from the initial state. Our method has been used on the benchmark *wide-mouthed frog* (WMF) and *Kerberos* protocols, on which we can find all the well known flaws in short time. We use the WMF protocol to illustrate formal specification and the use of our method to analyze timed authentication properties.

Improved Output Jitter Calculation for Compositional Performance Analysis of Distributed Systems

Rafik Henia, Razvan Racu and Rolf Ernst

Institute of Computer and Communication Network Engineering
Technical University of Braunschweig
Braunschweig, Germany
{*henia, razvan, ernst*}*@ida.ing.tu-bs.de*

Compositional performance analysis iteratively alternates local scheduling analysis techniques and output event model propagation between system components to enable performance analysis of heterogeneous distributed systems. In spite of its high scalability and adaptability, the compositional approach may suffer from overestimated results compared with other system performance verification techniques. The main reason is an incomplete consideration of event sequence correlations. In this paper we present a new technique that improves the output jitter calculation by correlating jitter and response times and offers significantly tighter analysis bounds.

Generating Efficient Distributed Deadlock Avoidance Controllers

Cesar Sanchez, Henny B. Sipma and Zohar Manna

Computer Science Department
Stanford University
Stanford, CA, USA
{*cesar, sipma, zm*}*@cs.stanford.edu*

General solutions to deadlock avoidance in distributed systems are considered impractical due to the high communication overhead. In previous work we showed that practical solutions exist when all possible sequences of resource requests are known a priori in the form of call graphs; in this case protocols can be constructed that involve no communication. These run-time protocols make use of annotations of the call graph that are computed statically based on the structure of the call graph. If the annotations are acyclic, then deadlocks are unreachable.

This paper focuses on the computation of these annotations. We first show that our algorithm for computing acyclic annotations is complete: every optimal annotation can be generated. We then show that, given a cyclic annotation and a fixed set of resources, checking whether deadlocks are reachable is NP-complete. Finally, we consider the problem of computing minimal annotations that satisfy given constraints on the number of available resources. We show that the problem is NP-complete in the general case, but that it can be solved in polynomial time if the only restrictions are that the number of certain resources is 1, that is, these resources are binary semaphores.

Workshop 3
Reconfigurable Architectures Workshop
RAW 2007

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
Vera Lauer, DaimlerChrysler AG
Juergen Becker, Univ. Karlsruhe (TH)
Rudy Lauwereins, IMEC, Leuven
Pascal Benoit, LIRMM
Philip Leong, Chinese Univ. of HK
Mladen Berekovic, IMEC
Marnane Liam, Univ. College
Neil Bergmann, Univ. of Queensland
Wayne Luk, Imperial College
Don Bouldin, Univ. of Tennessee
Juergen Luka, DaimlerChrysler AG
Elaheh Bozorgzadeh, U. of California
Patrick Lysaght, Xilinx
Gordon Brebner, Xilinx
Thomas Buechner, IBM
John McHenry, National Security Agency
Fabio Campi, Univ. di Bologna
Martin Middendorf, Univ. of Leipzig
Luigi Carro, Univ. Federal do Rio Grande
Amar Mukherjee, U. of Central Florida
Peter Y. K. Cheung, Imperial College, London
Koji Nakano, Hiroshima Univ.
Andreas Dandalis, Philips
Ranjani Parthasarathi, Anna University, Chennai
Marco Platzner, Univ. Paderborn
Oliver Diessel, U. New South Wales
Cameron Patterson, Virginia Tech
Adam Donlin, Xilinx
Thilo Pionteck, Univ. Lübeck
Pedro C. Diniz, USC
Bernard Pottier, Univ. de Bretagne Occidentale
Gilbert Edelin, Thales Research & Technology
Franz Rammig, Univ. Paderborn
Hossam ElGindy, U.New South Wales
Ricardo Reis, Univ. Federal do Rio Grande
Marco Santambrogio, Politecnico di Milano
Manfred Glesner, Darmstadt Univ. of Technology
Hartmut Schmeck, Univ. Karlsruhe (TH)
Steve Guccione, Cmpware, Inc.
Sakir Sezer, Queen's Univ.
Gerard Smit, Univ. of Twente
Masanori Hariyama, Tohoku Univ.
Srinivas Katkoori, U. of South Florida
Reiner Hartenstein, U. of Kaiserslautern
V. Sridhar, Satyam Comp. Services
Ulrich Heinkel, Lucent Technologies
Juergen Teich, Friedrich-Alexander-Universitaet Erlangen
Andreas Herkersdorf, Institute for Integrated Systems
Lionel Torres, LIRMM, Montpellier
Christian Hochberger, Dresden Univ. of Technology
Jim Tørresen, Univ. of Oslo
Thomas Hollstein, Darmstadt Univ. of Technology
Jerry L. Trahan, Louisiana State Univ.
Ramachandran Vaidyanathan, Louisiana State Univ.
Michael Hübner, Univ. Karlsruhe
Carlos Valderrama, Univ. Mons
Mark Jones, Virginia Tech
Milan Vasilko, Bournemouth Univ.
Stamatis Vassiliadis, Delft Univ. of Technology
Udo Kebschull, Univ. Leipzig
Brian Veale, Univ. of Oklahoma
Andreas Koch, Technische Univ. Braunschweig
Martin Vorbach, PACT Informationstechnologie
Rainer Kress, Infineon Technologies
Klaus Waldschmidt, Univ. Frankfurt
Norbert Wehn, Univ.of Kaiserslautern
Joachim Pistorius, Altera
Mauricio Ayala, Universidade de Brasilia

A New Framework to Accelerate Virtex-II Pro Dynamic Partial Self-Reconfiguration

Christopher Claus, Florian H. Mueller, Johannes Zeppenfeld and Walter Stechele

Institute for Integrated Systems
Munich University of Technology
Munich, Bavaria, Germany
{*christopher.claus, Florian.Mueller, Zeppenfe, Walter.Stechele*}*@tum.de*

The Xilinx Virtex family of FPGAs provides the ability to perform partial run-time reconf iguration, also known as dynamic partial reconf iguration (DPR). Taking this concept one step further, partial dynamic self-reconf iguration becomes possible through the Internal Conf iguration Access Port (ICAP). In this paper a framework for lowering reconf iguration times using the combitgen tool to reduce the overhead found within bitstreams, along with a completely new, very simple and area eff icient ICAP controller that is connected directly to the Processor Local Bus (PLB) and is equipped with Direct Memory Access (DMA) capabilities is presented. Using this PLB Master ICAP controller, it is possible to reach the maximum practical throughput that can be achieved with the ICAP interface of Virtex-II Pro devices. Compared to an alternative realization using the OPBHWICAP provided by Xilinx (a slave attachment on the On-Chip Peripheral Bus), it is possible to achieve improvements concerning reconf iguration times by a factor of 20.

Partial Dynamic Reconfiguration in a Multi-FPGA Clustered Architecture Based on Linux

Vincenzo Rana[1], Marco Santambrogio[1], Donatella Sciuto[1], Boris Kettelhoit[2], Markus Koester[2], Mario Porrmann[2] and Ulrich Rueckert[2]

[1]*Dipartimento di Elettronica e Informazione*
Politecnico di Milano
20133 Milano, Italy
vincenzo.rana@microlab-mi.net, {*santambr, sciuto*}*@elet.polimi.it*

[2]*Heinz Nixdorf Institute*
University of Paderborn
33102 Paderborn, Germany
{*kettelhoit, koester, porrmann, rueckert*}*@hni.upb.de*

Dynamically reconfigurable hardware allows for implementing systems that can be adapted at run-time according to the needs of the user. This paper presents an architecture that is composed of multiple FPGAs that are connected to an embedded processor. Thus, the architecture is referred to as a Multi-FPGA Clustered Architecture (MFCA). All FPGAs can be partially and dynamically reconfigured to integrate user-defined IP-Cores into the system at run-time. For the resource management and communication management we have implemented a Linux Operating System on the embedded processor that can be used to control the reconfiguration of the FPGAs by means of simple function calls. Furthermore, the Linux OS completely hides the physical infrastructure of the MFCA from user applications, offering a consistent interface to utilize partial reconfiguration.

Communication Architectures for Dynamically Reconfigurable FPGA Designs

Thilo Pionteck[1], Carsten Albrecht[1], Roman Koch[1], Erik Maehle[1], Michael Hübner[2] and Jürgen Becker[2]

[1] *Institute of Computer Engineering*
University of Lübeck
Lübeck, Germany
{*pionteck, albrecht, koch, maehle*}*@iti.uni-luebeck.de*

[2] *Institut für Technik in der Informationsverarbeitung*
Universität Karlsruhe
Karlsruhe, Germany
{*huebner, becker*}*@itiv.uni-karlsruhe.de*

This paper gives a survey of communication architectures which allow for dynamically exchangeable hardware modules. Four different architectures are compared in terms of reconfiguration capabilities, performance, flexibility and hardware requirements. A set of parameters for the classification of the different communication architectures is presented and the pro and cons of each architecture are elaborated. The analysis takes a minimal communication system for connecting four hardware modules as a common basis for the comparison of the diverse data given in the papers on the different architectures.

Optimization of Area and Performance by Processor-Like Reconfiguration

Tobias Oppold, Sven Eisenhardt and Wolfgang Rosenstiel

Department of Computer Engineering
University of Tuebingen
Tuebingen, Germany
{*oppold, eisenhar, rosenstiel*}*@informatik.uni-tuebingen.de*

It is well known that the area efficiency of a digital circuit can be improved by reconfiguration due to the reuse of resources. In this paper, we show that this benefit can be achieved for a wide range of applications if the reconfiguration can take place within each clock cycle, and we quantify the benefit by area estimations from a synthesizable architecture model. Although reconfiguration typically involves a decrease of performance, we show how performance can actually be increased by redirecting communication through the time domain. This increase is quantified by estimations from a silicon-proven commercial architecture and its associated compiler.

Splice: A Standardized Peripheral Logic and Interface Creation Engine

Justin Thiel and Ron K. Cytron

Department of Computer Science and Engineering
Washington University
Saint Louis, Missouri, USA
{*jthiel, cytron*}*@cs.wustl.edu*

Recent advancements in FPGA technology have allowed manufacturers to place general-purpose processors alongside user-configurable logic gates on a single chip. At first glance, these integrated devices would seem to be the ideal deployment platform for hardware-software co-designed systems, but some issues, such as incompatibility across vendors and confusion over which bus interfaces to support, have impeded adoption of these platforms. This paper describes the design and operation of Splice, a software-based code generation tool designed to address these types of issues by providing a bus-independent structure that allows end-users to integrate their customized peripheral logic easily into embedded systems.

Exploiting Communication Concurrency for Efficient Deadlock Free Routing in Reconfigurable NoC Platforms

Maurizio Palesi[1], Shashi Kumar[2], Rickard Holsmark[2] and Vincenzo Catania[1]

[1]*Dipartimento di Ingegneria Informatica e delle Telecomunicazioni*
University of Catania
Catania, Italy
{*mpalesi, vcatania*}*@diit.unict.it*

[2]*Embedded Systems Department*
Jonkoping University
Jonkoping, Sweden
{*Shashi.Kumar, Rickard.Holsmark*}*@jth.hj.se*

In this paper we make a case for the use of NoC paradigm to develop future FPGAs in which large computational blocks (cores) are connected to each other through a packet switched communication network. We propose a methodology to develop efficient and deadlock free routing algorithms for such NoC platforms which can be specialized for an application or a set of concurrent applications. Application specific topology of communicating cores as well as information about their communication concurrency over time is exploited to maximize communication adaptivity and performance. We demonstrate, both through analysis of adaptivity as well as simulation based evaluation of latency and throughput, that our algorithm gives significantly higher performance as compared to general purpose deadlock free algorithms like XY and Odd-Even.

Power-Aware Routing for Well-Nested Communications On The Circuit Switched Tree

Hatem M. El-Boghdadi

Computer Eng. Dept.
Cairo University
Giza, EGYPT
helboghdadi@eng.cu.edu.eg

Although algorithms that employ dynamic reconfiguration are extremely fast, they need the underlying architecture to change structure very rapidly, possibly at each step of the computation. This increases the power requirement of such algorithms which is not acceptable in nowadays devices that strive to reduce the power requirements. This paper deals with the circuit switched tree (CST), an interconnect used to implement dynamically reconfigurable architectures.

In this paper, we introduce a new technique called Power Aware Dynamic Reconfiguration (PADR). Under this technique, we propose a power-aware algorithm for configuring the CST and scheduling a class of communications, called the well-nested communicatons on the CST. We show that the algorithm is power optimal. The algorithm requires only local information at processing elements (PEs), yet it correctly establishes paths between communicating PEs. We also show that the algorithm is optimal and efficient.

Using Rewriting Logic to Match Patterns of Instructions from a Compiler Intermediate Form to Coarse-Grained Processing Elements

Carlos Morra[1], Joao M. P. Cardoso[2] and Juergen Becker[1]

[1] *Institut für Technik der Informationsverarbeitung (ITIV)*
Universität Karlsruhe (TH)
Karlsruhe, Germany
{morra, becker}@itiv.uni-karlsruhe.de

[2] *UTL/IST, Department of Informatics Engineering*
INESC-ID
Lisbon, Portugal
jmpc@acm.org

This paper presents a new and retargetable method to identify patterns of instructions with direct support in coarsegrained processing elements (PEs). The method uses a three-address code SSA (static single assignment) representation of the kernel being mapped and Rewriting Logic for template matching and algebraic optimizations. This approach is able to identify sets of SSA instructions that can be mapped to different PE complexities available in coarse-grained reconfigurable computing architectures. As a proof of concept, results of the approach with a number of benchmark kernels, as far as coverage of template instructions is concerned, are included.

Interconnect Customization for a Coarse-grained Reconfigurable Fabric

Gayatri Mehta[1], Justin Stander[1], Mustafa Baz[2], Brady Hunsaker[2] and Alex K. Jones[1]

[1]*Electrical and Computer Engineering*
University of Pittsburgh
Pittsburgh, PA, USA
{*gmehta, jstander, akjones*}*@engr.pitt.edu*

[2]*Industrial Engineering*
University of Pittsburgh
Pittsburgh, PA, USA
mub3@pitt.edu, hunsaker@engr.pitt.edu

This paper describes several system-level interconnection strategies for a coarse-grained reconfigurable fabric designed for low-energy hardware acceleration. A small, representative sub-graph for signal and image processing applications is used to predict the success of mapping larger applications onto the fabric device with these different interconnection strategies, which include 32:1, 8:1, 5:1, 4:1, 3553:1 (3:1, 5:1, 5:1, 3:1) and 355:1 (3:1, 5:1, 5:1) cardinalities. Three mapping techniques are presented and used to complete mappings onto several of these fabric instances including a mixed integer linear programming technique, a constraint programming approach, and a greedy heuristic. We present results for area (in number of required rows), power, delay, and energy as well as run times for mapping a set of signal and image processing benchmarks onto each of these interconnects. Our results indicate that the 5:1 interconnect provides the best overall results and does not require any additional hardware resources than the baseline 4:1 technique. When compared with other implementation strategies, the reconfigurable fabric energy consumption, using 5:1-based interconnect, is within 5-10X of a direct ASIC implementation, is 10X better than an Virtex II Pro FPGA and is 100X better than an Intel XScale processor.

A Modulo Scheduling Algorithm for a Coarse-Grain Reconfigurable Array Template

Akira Hatanaka and Nader Bagherzadeh

Electrical Engineering and Computer Science
University of California, Irvine
Irvine, CA, United States
{*ahatanak, nader*}*@uci.edu*

Coarse Grain Reconfigurable Arrays (CGRAs) have been drawing attention due to its programmability and performance. Compilation onto CGRAs is still an open problem. Several groups have proposed algorithms that software pipeline loops onto CGRAs.

In this paper, we present an efficient modulo scheduling algorithm for a CGRA template. The novelties of the approach are the separation of resource reservation and scheduling, use of a compact three-dimensional architecture graph and a resource usage aware relocation algorithm. Preliminary experiments indicate that the proposed algorithm can find schedules with small initiation intervals within a reasonable amount of time.

A CAM Emulator Using Look-Up Table Cascades

Hiroki Nakahara, Tsutomu Sasao and Munehiro Matsuura

Department of Computer Science and Electronics
Kyushu Inst. of Tech.
Iizuka, Fukuoka, Japan
nakahara@aries01.cse.kyutech.ac.jp, {sasao, matsuura}@cse.kyutech.ac.jp

An address table relates k different registered vectors to the addresses from 1 to k. An address generation function represents the address table. This paper presents a realization of an address generation function with an LUT cascade on an FPGA. The address generation function is implemented by BRAMs of an FPGA, while the addition and the deletion of registered vectors are implemented by an embedded processor on the FPGA. Compared with CAMs produced by the Xilinx Core Generator, our implementations are smaller and faster. This paper also shows that the addition and deletion of a registered vector can be done in time that is proportional to the number of cells in the LUT cascade.

A Reconfigurable Computing Engine for Wavelet Transforms

Kang Sun, Xuezeng Pan and Lingdi Ping

College of Computer Science and Technology
Zhejiang University
Hangzhou, Zhejiang, P.R. China
{ksun, panxz, pingld}@zju.edu.cn

In the past a few years, wavelet transforms have become a hot topic of research. Discrete and continuous wavelet transforms have been widely used in signal and multimedia processing. Due to the high performance and flexibility of reconfigurable computing systems, it is very attractive to design a reconfigurable architecture for discrete and continuous wavelet transform of wide range of wavelet filters. In this paper, a unified computation framework for discrete and continuous wavelet transform based on lifting scheme and a reconfigurable architecture that includes reconfigurable lifting step arrays and reconfigurable address generator are proposed. The unified framework is the theory basis of this system. The step array is the computing core of this engine. And the address generator supports several memory scan pattern which is used to generate memory access addresses. In order to validate this architecture, an FPGA prototype is built based on Xilinx VirtexII FPGA to test the reconfiguration of 2-D discrete 5/3 and 9/7 transforms (defined in specification of JPEG2000) and 2-D continuous Haar wavelet transform. Furthermore, a 3-level decomposition for a 512x512 grayscale image is performed and the results show that the decomposition can be finished within 12.16ms when running at 20MHz. It can be concluded that this design is applicable and scalable.

Using an FPGA for Fast Bit Accurate SoC Simulation

Pascal T. Wolkotte, Philip K. F. Hölzenspies and Gerard J. M. Smit

Department of EEMCS
University of Twente
Enschede, The Netherlands
{*P.T.Wolkotte, P.K.F.Holzenspies, G.J.M.Smit*}*@utwente.nl*

In this paper we describe a sequential simulation method to simulate large parallel homo- and heterogeneous systems on a single FPGA. The method is applicable for parallel systems were lengthy cycle and bit accurate simulations are required. It is particularly designed for systems that do not fit completely on the simulation platform (i.e. FPGA). As a case study, we use a Network-on-Chip (NoC) that is simulated in SystemC and on the described FPGA simulator. This enables us to observe the NoC behavior under a large variety of traffic patterns. Compared with the SystemC simulation we achieved a factor 80-300 of speed improvement, without compromising the cycle and bit level accuracy.

A General Purpose Partially Reconfigurable Processor Simulator (PReProS)

Alisson V. Brito[1], Matthias Kuehnle[1], Elmar U. K. Melcher[2] and Juergen Becker[1]

[1]*ITIV*
Universitaet Karlsruhe (TH)
Karlsruhe, BW, Germany
{*brito, kuehnle, becker*}*@itiv.uni-karlsruhe.de*

[2]*DEE*
Federal University of Campina Grande (UFCG)
Campina Grande, PB, Brazil
elmar@dsc.ufcg.edu.br

An innovative technique to model and simulate partial and dynamic reconfigurable processors is presented in this paper. The basis for development is a SystemC kernel modified for dynamic reconfiguration. The presented approach can either be used at transaction-level, which allows the modeling and simulation of higher-level hardware and embedded software, or at register transfer level (RTL), if the dynamic system behavior is desired to be observed at signal level. The reconfigurable processor can be easily set to model the desired architecture in a behavioral but reasonable way. An example is presented where a XPP processor is implemented and simulated, executing typical applications. The resulting statistics assist either in the choice of the best cost/benefit configuration area that should be available on chip, or in the choice of the target architecture itself.

CONFETTI : A reconfigurable hardware platform for prototyping cellular architectures

Pierre- André Mudry, Fabien Vannel, Gianluca Tempesti and Daniel Mange

Cellular Architectures Research Group
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
{*pierre-andre.mudry, fabien.vannel, gianluca.tempesti, daniel.mange*}*@epfl.ch*

In this article, we describe a novel hardware platform aimed at the realization of cellular architectures. The system is built hierarchically from a very simple computing unit, called ECell. Several of these units can then be connected, using a high-speed serial communication protocol, to a more complex structure called the UltraStack. Consisting of four different kinds of interconnected boards computational, routing, power supply, and display), these stacks can then be joined together to form an arbitrarily large parallel network of programmable circuits.

This structure, while theoretically universal in its operation, is however particularly suited for the implementation of cellular computing applications.

A Reconfigurable Load Balancing Architecture for Molecular Dynamics

Jonathan Phillips, Matthew Areno, Chris Rogers, Aravind Dasu and Brandon Eames

ECE
Utah State University
Logan, UT, USA
{*jdphillips, matthewareno, crogers*}*@cc.usu.edu,* {*dasu, beames*}*@engineering.usu.edu*

This paper proposes a novel architecture supporting dynamic load balancing on an FPGA for a Molecular Dynamics algorithm. Load balancing is primarily achieved through the use of specialized processing units, referred to as FLEX units. FLEX units are able to switch between tasks required by a molecular dynamics algorithm as often as needed in order to cater to the nature of the input parameters. This architecture is capable of run-time performance analysis and dynamic resource allocation in order to maximize throughput. Results of a prototype of the architecture targeting an FPGA are presented

Fast SEU Detection and Correction in LUT Configuration Bits of SRAM-based FPGAs

Hamid R. Zarandi[1], Seyed Ghassem Miremadi[2], Costas Argyrides[3] and Dhiraj K. Pradhan[4]

[1]*Department of Computer Engineering*
Sharif University of Technology
Tehran, Tehran, IRAN
zarandi@ce.sharif.edu

[2]*Department of Computer Engineering*
Sharif University of Technology
Tehran, Tehran, IRAN
miremadi@sharif.edu

[3]*Department of Computer Science*
Bristol University
Bristol, Bristol, UK
costas@cs.bris.ac.uk

[4]*Department of Computer Science*
Bristol University
Bristol, Bristol, UK
pradhan@cs.bri.ac.uk

FPGAs are an appealing solution for the space-based remote sensing applications. However, in a low-earth orbit, configuration bits of SRAM-based FPGAs are susceptible to single-event upsets (SEUs). In this paper, a new protected CLB and FPGA architecture are proposed which utilize error detection and correction codes to correct SEUs occurred in LUTs of the FPGA. The fault detection and correction is achieved using online or offline fast detection and correction cycles. In the latter, detection and correction is performed in predefined error-correction intervals. In both of them error detections and corrections of k-input LUTs are performed with a latency of 2^k clock cycle without any required reconfiguration and significant area overhead. The power and area analysis of the proposed techniques show that these methods are more efficient than the traditional schemes such as duplication with comparison and TMR circuit design in the FPGAs.

Radiation Hardened Coarse-Grain Reconfigurable Architecture for Space Applications

Sajid Baloch[1,2], Tughrul Arslan[1,2] and Adrian Stoica[1,3]

[1]*School of electronics and engineering*
The University of Edinburgh, UK
Edinburgh, UK
sajid.baloch@sli-institute.ac.uk, tughrul.arslan@ed.ac.uk,
adrain.stoica@jpl.nasa.gov

[2]*Institute for System Level Integration, UK*
The University of Edinburgh, UK
Livingston, UK

[3]*Jet Propulsion Lab*
NASA
Pasadena, California, USA

Technology trends are such that single event effects (SEE)are likely to become even more of a concern for the future. Decreasing feature sizes, lower operating voltage, and higher speeds, all conspire to increase susceptibility to single event upsets (SEU). Upset in avionics is an established concern. Upset at the ground level is becoming a concern for manufacturers of micro-electronics for terrestrial applications. The use of flip-chip packaging and multiple levels of metals further exacerbate the problem. Typical methods of mitigation that either increase the transistor count or reduce IC performance are not acceptable to commercial manufacturers. SOI technology may help in this regard, but is not a magic bullet to end all SEE concerns. We present unique schemes to model and rectify single event disruption in combinatorial and synchronous parts of a reconfigurable architecture. We compare our scheme with different schemes already introduced and results are reported to prove the efficacy of the proposed radiation hardened reconfigurable architecture.

A Cryptographic Coarse Grain Reconfigurable Architecture Robust Against DPA

Daniel Mesquita[1], Benoît Badrignans[2], Lionel Torres[2], Gilles Sassatelli[2], Michel Robert[2] and Fernando Moraes[3]

[1]*INESC-ID*
Instituto Superior Técnico
Lisboa, Portugal
mesquita@inesc-id.pt

[2]*LIRMM*
Université Montpellier II
Montpellier, France
{*badrignans, torres, sassate, robert*}*@lirmm.fr*

[3]*PPGCC/FACIN*
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, Brasil
moraes@inf.pucrs.br

This work addresses the problem of information leakage of cryptographic devices, by using the reconfiguration technique allied to an RNS based arithmetic. The information leaked by circuits, like power consumption, electromagnetic emissions and time to compute may be used to find cryptographic secrets. The results issue of prototyping shows that our coarse grained reconfigurable architecture is robust against power analysis attacks.

Hierarchical Cluster Assignment for Coarse-Grain Reconfigurable Coprocessors

Martino Sykora[1], Davide Pavoni[1], Joel Cambonie[2], Roberto Costa[3] and Stefano Crespi Reghizzi[1]

[1]*Dip. Elettronica e Informazione*
Politecnico di Milano
Milano, Italy
{*sykora, crespi*}*@elet.polimi.it,*
davide.pavoni@gmail.com

[2]*STMicroelectronics - Grenoble*
Grenoble, France
joel.cambonie@st.com

[3]*STMicroelectronics - AST Manno Lab*
Manno, Switzerland
roberto.costa@st.com

Embedded media applications have to satisfy real-time, low power consumption and silicon area constraints. These applications spend most of the execution time in the iteration of a few kernels; such kernels are typically made of independent operations, which can be executed in parallel. Clustered architectures are a solution designed to exploit the high Instruction Level Parallelism (ILP) of the media kernels, to keep a good level of scalability and to match the strict constraints of the embedded domains. Within this category, architectures with reconfigurable connections between clusters are of particular interest. The enhanced flexibility allows them to handle several different data-paths effectively, hence multiple applications; this is a key economic factor in the semiconductor world, in which the cost of the masks significantly increases at every technological advance. This papers describes Hierarchical Cluster Assignment (HCA), a compilation technique that deals with the problem of mapping the computation of multimedia kernels onto the clusters of the target machine. HCA exploits the hierarchical structure of the clusters of the target architectures; it works by decomposing the problem of cluster assignment into a sequence of simpler sub-problems, each of them involving a subset of the kernel instructions and a subset of the machine clusters. A prototype of this methodology has been implemented in a flexible framework and tested on machine models based on the DSPFabric architecture.

QUKU: A FPGA Based Flexible Coarse Grain Architecture Design Paradigm using Process Networks

Sunil Shukla[1,2], Neil W. Bergmann[1] and Jürgen Becker[2]

[1] *Information Technology & Electrical Engg.*
The University of Queensland
Brisbane, Australia
{*sunil, n.bergmann*}*@itee.uq.edu.au*

[2] *Institut für Technik der Informationsverarbeitung*
Universität Karlsruhe
Karlsruhe, Germany
becker@itiv.uni-karlsruhe.de

DSP applications can be suitably represented using Process Network Models. This paper uses a modification of Kahn Process Network to solve the problem of finding an optimum architectural template for coarse grain array on per application basis. We have proposed an architecture, QUKU, which is a coarse grain reconfigurable PE array overlaid on FPGA fabric. Conventional CGRAs support very fast dynamic reconfiguration using micro-codes as opposed to Megabytes of configuration data in FPGAs. But CGRAs have fixed physical layout which ties their usability to a particular application domain. The uniqueness of QUKU lies in the fact that it supports dual layered reconfiguration. The dual layered reconfiguration capability overcomes the slow reconfiguration of FPGAs as well as overcomes the problem of finding an optimal architecture which suits a wide variety of application domains. The FPGA level reconfiguration is an infrequently happening reconfiguration which typically takes a few milli-seconds. This method, although takes a few milli-seconds, serves the important function of physically reconfiguring the array. The very frequently happening reconfiguration is at PE level which configures the PE array to perform the new functionality, without changing the physical aspects of the array. The very fast occuring dynamic reconfiguration of the PE array provides an excellent option to overcome the slow partial dynamic reconfiguration methodology used in FPGAs.

To generate an application specific PE layout, a modification of Kahn Process Networks is used. The application of process network in architectural exploration presents a lot of options in terms of PE layout. The final layout can be chosen based on parameters like area, power and performance efficiency.

Speedups and Energy Savings of Microprocessor Platforms with a Coarse-Grained Reconfigurable Data-Path

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

VLSI Design Lab., ECE Dept.
University of Patras
Patras, Achaia, Greece
{*mgalanis, dhmhgre, goutis*}*@ece.upatras.gr*

This paper presents the performance improvements and the energy reductions by coupling a high-performance coarse-grained reconfigurable data-path with a microprocessor in a generic platform. The data-path has been previously introduced by the authors. It is composed by computational units able to realize complex operations which aid in improving the performance of time critical application parts, called kernels. A design flow is proposed for mapping high-level software descriptions to the microprocessor system. Eight real-life applications are mapped on three different instances of the system. Significant overall application speedups, relative to a software-only solution, ranging from 1.74 to 3.94 are reported being close to theoretical speedup bounds. Average energy savings of 59% are achieved, while the reduction in the system energy-delay product ranges from 66% to 92%.

Cost-Driven Hybrid Configuration Prefetching for Partial Reconfigurable Coprocessor

Ying Chen[1] and Simon Y. Chen[2]

[1] School of Engineering
San Francisco State University
San Francisco, CA, USA
yingchen@sfsu.edu

[2] Broadband Division
DSP Group, Inc.
Santa Clara, CA, USA
simon.yingchen@gmail.com

Reconfigurable computing systems have developed the capability of changing the configuration of the reconfigurable coprocessor multiple times during the course of a program. However, in most systems the reconfigurable coprocessor wastes computation cycles while waiting for the reconfiguration to complete. Therefore, the high demand for frequent run-time reconfiguration directly translates into higher reconfiguration overhead. Some studies have introduced the concept of prefetching to reduce the reconfiguration overhead. However, these prefetching algorithms are probability-driven. We believe that including configuration size information in the prediction algorithm directly links the training of the predictor with the performance gain. Therefore we proposed a performance-oriented cost-driven algorithm for coarse-grained configuration prefetching. Our cycle accurate simulation results show that the proposed cost-driven algorithm outperforms the probability-driven predictor by 10.8% to 29.6% in reducing reconfiguration overhead.

A Reconfiguration Aware Circuit Mapper for FPGAs

Markus Rullmann and Renate Merker

Department of EE and IT/Circuits and Systems Laboratory
Dresden University of Technology
Dresden, Germany
{markus.rullmann, renate.merker}@tu-dresden.de

Dynamic reconfiguration for fine grained architectures is still associated with significant reconfiguration costs. In this paper we propose a new reconfiguration aware design flow. The tools in this flow implement a set of tasks concurrently. The flow leads to task implementations with minimal costs for routing reconfiguration. This is mainly achieved by our mapping tool which solves two fundamental problems: Our mapping algorithm generates variants for the mapping of netlist cells to logic blocks. From those logic blocks a subset for each task is selected that minimizes the cost for routing reconfiguration. We derive a cost function and formulate an integer linear program to solve this problem. We implemented several task sets with our method and compare the results to previous solutions. We show that the reconfiguration aware mapping leads to better results than early approaches with vendor provided tools.

Miss Ratio Improvement For Real-Time Applications Using Fragmentation-Aware Placement

Ahmed Abou Elfarag, Hatem M. El-Boghdadi and Samir I. Shaheen

Computer Engineering Dept.
Cairo University
Giza, EGYPT
abouelfarag@aast.edu, helboghdadi@eng.cu.edu.eg, sshaheen@ieee.org

Partially reconfigurable Field-Programmable Gate Arrays (FPGAs) allow parts of the chip to be configured at run-time where each part could hold an independent task. Online placement of these tasks result in area fragmentation leading to poor utilization of chip resources.

In this paper, we propose a new metric for measuring area fragmentation. The new fragmentation metric gives an indication to the continuity of the occupied (or free) space and not the amount of occupied space. We show how this metric can be extended for multi-dimensional structures. We also show how this metric can be computed efficiently at run time. Next we use this measure during online placement of tasks on FPGAs, such that the chip fragmentation is reduced. Our results show improvement of chip utilization when using this fragmentation aware placement method over other placement methods with well known Bottom Left First Fit, and Best Fit placement strategies. In real time environment, we achieve an improvement in miss ratio when using the fragmentation aware placement over the bottom left placement strategy.

Managing dynamic reconfiguration on MIMO Decoder

Hongzhi Wang, Jean-Philippe Delahaye, Pierre Leray and Jacques Palicot

SCEE Team
IETR/Supelec
CESSON-SEVIGNE, FRANCE
{hongzhi.wang, Jean-Philippe.Delahaye, pierre.leray, jacques.palicot}@supelec.fr

This paper is about the implementation of a MIMO V-BLAST (Vertical Bell Laboratories Layered Space-Time) square root decoder in a FPGA using dynamic partial reconfiguration. The decoder architecture is based on four CORDIC (CO-ordinate Rotation DIgital Computer) Units. Among these CORDIC units, three are used in rotation mode and the fourth one is used in vectoring mode. The design implementation aims power saving and area efficiency allowing dynamically changing the interconnections between the fixed modules in the reconfigurable modules. This MIMO square root design method shows the configuration time improvement, area efficiency and flexibility of the decoder by using the dynamic partial reconfiguration method.

Model and Methodology For the Synthesis of Heterogeneous and Partially Reconfigurable Systems

Florian Dittmann[1], Marcelo Götz[1] and Achim Rettberg[2]

[1]Heinz Nixdorf Institute
University Paderborn
33102 Paderborn, Germany
{roichen, mgoetz}@upb.de

[2]C-LAB
University Paderborn
33102 Paderborn, Germany
achim@c-lab.de

When reconfigurable devices are used in modern embedded systems and their capability to adapt to changing application requirements becomes an issue, comprehensive modeling and design methods are required. Such methods must respect the whole range of functionality of the reconfigurable fabrics. In particular, the heterogeneity and reconfiguration delay of modern FPGAs are important details. Comprehensive methods to exploit these characteristics within the integrated design of embedded systems are still not available. In this paper, we introduce a synthesis methodology for reconfigurable systems that respects the specific requirements of run-time reconfiguration. The methodology bases on profound concepts, and expands known notations and model techniques.

An Architectural Framework for Automated Streaming Kernel Selection

Nikolaos Bellas, Sek M. Chai, Malcolm Dwyer and Dan Linzmeier

Embedded Systems Research
Motorola
Schaumburg, IL, USA
{nikos.bellas, sek.chai, malcolm.dwyer, dan.linzmeier}@motorola.com

Hardware accelerators are increasingly used to extend the computational capabilities of baseline scalar processors to meet the growing performance and power requirements of embedded applications. The challenge to the designer is the extensive human effort required to identify the appropriate kernels to be mapped to gates, and to implement a network of accelerators to execute the kernels. In this paper, we present a methodology to automate the selection of streaming kernels in a reconfigurable platform based on the characteristics of the application. The methodology is based on a flow graph that describes the streaming computations and communications. The flow graph is used to efficiently identify the most profitable subset of streaming kernels that optimize performance without exceeding the available area of the reconfigurable fabric.

High-Level Synthesis of HW Tasks Targeting Run-Time Reconfigurable FPGAs

Maik Boden[1], Thomas Fiebig[1], Torsten Meissner[1], Steffen Ruelke[1] and Juergen Becker[2]

[1]*Design Automation Division (EAS)*
Fraunhofer Institute for Integrated Circuits (IIS)
Dresden, Germany
{*Maik.Boden, Thomas.Fiebig, Torsten.Meissner, Steffen.Ruelke*}*@eas.iis.fraunhofer.de*

[2]*Institute for Information Processing Technology*
University of Karlsruhe, ITIV
Karlsruhe, Germany
Becker@itiv.uni-karlsruhe.de

This paper presents a novel High-Level Synthesis (HLS) and optimization approach targeting FPGA architectures that are reconfigurable at run-time. To model a reconfigurable system on a high level of abstraction, we use a hierarchical operation (control and data) flow graph. In order to reduce the overhead for reconfiguring the system, we apply resource sharing to our model to deduce reusable design parts for the implementation. A case study compares our HLS approach with a reference design which was manually coded on Register-Transfer-Level.

A multi-context holographic memory recording system for Optically Reconfigurable Gate Arrays

Rio Miyazaki[1], Minoru Watanabe[2] and Fuminori Kobayashi[3]

[1]*Innovation Plaza Fukuoka*
Japan Science and Technology Agency
Sawara-ku, Fukuoka, Japan
miyazaki@fukuoka.jst-plaza.jp

[2]*Systems Innovation and Informatics*
Kyushu Institute of Technology
Iizuka, Fukuoka, Japan
watanabe@ces.kyutech.ac.jp

[3]*Systems Innovation and Informatics*
Kyushu Institute of Technology
Iizuka, Fukuoka, Japan
fkoba@ces.kyutech.ac.jp

Optically 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 a holographic memory. The first ORGA was developed to achieve rapid reconfiguration and a number of reconfiguration contexts; it consisted of a gate array VLSI, a holographic memory, and a laser diode array. The ORGA achieved a 16 μs to 20 μs reconfiguration period that was faster than that of FPGAs, with 100 reconfiguration contexts. However, the ORGA requires the gate array to halt during reconfiguration. Therefore, the ORGA can not be reconfigured frequently because of the associated reconfiguration overhead.

On the other hand, new ORGA-VLSIs that have less than 10 ns reconfiguration capability without any related overhead have already been fabricated. However, to date, a multi-holographic reconfiguration system that is suitable for such rapidly reconfigurable ORGA-VLSIs without any overhead has never been developed. For such realization, this paper proposes a four-context ORGA architecture and a multi-context holographic memory recording system used for it. In addition, experimentally demonstrated results of recording a holographic memory and reconfiguring an ORGA-VLSI are described.

Code Compression and Decompression for Instruction Cell Based Reconfigurable Systems

Nazish Aslam[1], Mark Milward[2], Ioannis Nousias[2], Tughrul Arslan[1,2] and Ahmet Erdogan[1,2]

[1]*Institute for System Level Integration*
Livingston, UK
nazish.aslam@sli-institute.ac.uk, t.arslan@ed.ac.uk,
ahmet.erdogan@ee.ed.ac.uk

[2]*School of Engineering and Electronics*
University of Edinburgh
Edinburgh, UK
mark.milward@ed.ac.uk, s0238762@sms.ed.ac.uk

Code compression has been applied to embedded systems to minimize the silicon area utilized for program memories, and lower the power consumption. More recently, it has become a necessity for multiple-issue architectures, such as VLIW and TTA, to permit a viable realization of these designs. In this paper, a code compression and decompression scheme suitable for newly emerging reconfigurable technologies is presented, which pose further challenges by having an order of magnitude higher memory requirement due to much wider instruction words than typical VLIW/TTA architectures. Two dictionary-based lossless compression schemes are implemented and compared for an example reconfigurable system. This paper looks at several conflicting design parameters, such as the compression ratio, silicon area and speed. Test programs for a 2D DCT, minimum error, wimax and H.264 have been evaluated with compression ratios in the range of 41% to 62% recorded with the best scheme.

C++ based System Synthesis of Real-Time Video Processing Systems targeting FPGA Implementation

Mattias O'Nils, Benny Thornberg and Najeem Lawal

Dept. of Information Technology & Media
Mid Sweden University
Sundsvall, Sweden
{*mattias.onils, benny.thornberg, najeem.lawal*}*@miun.se*

Implementing real-time video processing systems put high requirements on computation and memory performance. FPGAs have proven to be effective implementation architecture for these systems. However, the hardware based design flow for FPGAs make the implementation task complex. The system synthesis tool presented in this paper reduces this design complexity. The synthesis is done from a SystemC based coarse grain data flow graph that captures the video processing system. The data flow graph is optimized and mapped onto an FPGA. The results from real-life video processing systems clearly show that the presented tool produces effective implementations.

A Study of Design Efficiency with a High-Level Language for FPGAs

Zain-Ul-Abdin and Bertil Svensson

Centre for Research on Embedded Systems (CERES)
Halmstad University
Halmstad, Sweden
Zain-ul-Abdin@ide.hh.se

Over the years reconfigurable computing devices such as FPGAs have evolved from gate-level glue logic to complex reprogrammable processing architectures. However, the tools used for mapping computations to such architectures still require the knowledge about architectural details of the target device to extract efficiency.

A study of the Mobius language and tools is presented in this paper, with a focus on generated hardware performance. A number of streaming and memory-intensive applications have been developed and the results have been compared with the corresponding implementations in VHDL and a behavioral hardware description language. Based upon experimental evidences, it is concluded that Mobius, a minimal parallel processing language targeted for reconfigurable architectures, enhances productivity in terms of design time and code maintainability without considerably compromising performance and resources.

Workshop 4
Workshop on High-Level Parallel Programming Models and Supportive Environments
HIPS-TOPMoDRS 2007

4 HIPS - TOPMoDRS • Workshop on High-Level Parallel Programming Models and Supportive Environments

Joint Workshop:
Due to the significant overlap between the interests of **HIPS** and those of **TOPMoDRS**, the two workshops have joined hands on an experimental basis for IPDPS 2007.

HIPS Workshop Description:

HIPS is a full-day workshop focusing on high-level programming of chip multi-processors (multi-core PCs), computing clusters, and massively-parallel machines. Like its predecessors, the workshop seeks cross-fertilizing research in areas of parallel applications, language design, compilers, run-time systems, and programming tools. It provides a timely and lightweight forum for scientists and engineers to present the latest ideas, findings, and tools in these rapidly changing fields.
This year the workshop especially encourages innovative approaches for programming the increasingly popular chip multi-processors and the fast growing large-scale parallel systems.
The technical program will consist of presentations on the following topics:

- new programming constructs for exploiting parallelism and locality
- experience with and improvements for existing parallel languages and run-time environments such as MPI, OpenMP, HPF, Cilk, UPC, and Co-array Fortran.
- parallel programming tools and environments
- (scalable) performance analysis, modeling, and monitoring
- OS and architectural support for parallel programming and debugging
- software and system support for extreme scalability including the issues of fault tolerance

HIPS Workshop Chair:
Chen Ding, University of Rochester, USA

HIPS 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

HIPS Program Committee:
Arun Chauhan, Indiana University, IN, USA
Daniel G. Chavarria, Pacific Northwest National Laboratory, WA, USA
Li Chen, Chinese Academy of Sciences, Beijing, China
Guang R. Gao, University of Delaware, DE, USA
Michael Gerndt, Technische Universität München, Germany
Mary W. Hall, USC Information Sciences Institute, CA, USA
Ami Marowka Shenkar, College of Engineering and Design, Israel
Daniel Quinlan, Lawrence Livermore National Laboratory, CA, USA
Lawrence Rauchwerger, Texas A & M University, TX, USA
Sven-Bodo Scholz, University of Hertfordshire, UK
Allan Snavely, University of California, San Diego, CA, USA
Rich Wolski, University of California, Santa Barbara, CA, USA
Xin Yuan, Florida State University, FL, USA

TOPMoDRS Workshop Description:
The TOPMoDRS workshop is a half-day session held at the IPDPS 2007. The workshop focuses on tools and programming models for designing reliable systems with a special focus on distributed applications. The main goal of the workshop is to bring researchers and practitioners together in a setting where they can discuss interesting research topics, like:

- novel architectures to build reliable distributed systems,
- the design and implementation of new tools, techniques, programming languages, and compilers to increase the reliability of distributed systems,
- bug finding and debugging tools for distributed applications, and
- new developments in formal verification of distributed environments.

Workshop Organizer and Program Chair:
Cristian Tapus, California Institute of Technology (Caltech) and the Center for Advanced Computing Research, USA

Technical Program Committee:
Jason Hickey, Caltech, USA
Ranjit Jhala, University of California, San Diego , USA
Joe Kiniry, University College Dublin, Ireland
Sorin Lerner, University of California, San Diego, USA
Nenad Medvidovic, University of Southern California, USA
Christine Morin, INRIA, France
Cristina Nita-Rotaru, Purdue University, USA
Aleksey Nogin, HRL Laboratories, USA
Nicolae Tapus, ``Politehnica'' University of Bucharest, Romania
Yuval Tamir, University of California, Los Angeles (UCLA), USA

Other Reviewers:
Israel Hsu, UCLA, USA
Michael Le, UCLA, USA

Keynote – Programming Distributed Memory Sytems Using OpenMP

Ayon Basumallik, Seung-Jai Min and Rudolf Eigenmann

Electrical and Computer Engineering
Purdue University
West Lafayette, IN, USA
{*basumall, smin, eigenman*}*@ecn.purdue.edu*

OpenMP has emerged as an important model and language extension for shared-memory parallel programming. On shared-memory platforms, OpenMP offers an intuitive, incremental approach to parallel programming. In this paper, we present techniques that extend the ease of shared-memory parallel programming in OpenMP to distributed-memory platforms as well.

First, we describe a combined compile-time/runtime system that uses an underlying Software Distributed Shared Memory System and exploits repetitive data access behavior in both regular and irregular program sections. We present a compiler algorithm to detect such repetitive data references and an API to an underlying software distributed shared memory system to orchestrate the learning and pro-active reuse of communication patterns.

Second, we introduce a direct translation of standard OpenMP into MPI message-passing programs for execution on distributed memory systems. We present key concepts and describe techniques to analyze and efficiently handle both regular and irregular accesses to shared data. Finally, we evaluate the performance achieved by our approaches on representative OpenMP applications.

Keynote – A Compile-time Cost Model for OpenMP

Chunhua Liao and Barbara Chapman

Computer Science Department
University of Houston
Houston, TX, USA
{*liaoch, chapman*}*@cs.uh.edu*

OpenMP has gained wide popularity as an API for parallel programming on shared memory and distributed shared memory platforms. It is also a promising candidate to exploit the emerging multicore, multithreaded processors. In addition, there is an increasing trend to combine OpenMP with MPI to take full advantage of mainstream supercomputers consisting of clustered SMPs. All of these require that attention be paid to the quality of the compiler's translation of OpenMP and the flexibility of runtime support. Many compilers and runtime libraries have an internal cost model that helps evaluate compiler transformations, guides adaptive runtime systems, and helps achieve load balancing. But existing models are not sufficient to support OpenMP, especially on new platforms. In this paper, we present our experience adapting the cost models in OpenUH, a branch of Open64, to estimate the execution cycles of parallel OpenMP regions using knowledge of both software and hardware. Our OpenMP cost model reuses major components from Open64, along with extensions to consider more OpenMP details. Preliminary evaluations of the model are presented using kernel benchmarks. The challenges and possible extensions for modeling OpenMP on multicore platforms are also discussed.

Optimizing Inter-Nest Data Locality Using Loop Splitting and Reordering

Sofiane Naci

The Computer Laboratory
University of Cambridge
Cambridge, United Kingdom
Sofiane.Naci@cl.cam.ac.uk

With the increasing gap between processor speed and memory latency, the performance of data-dominated programs are becoming more reliant on fast data access, which can be improved using data locality optimization. Most studies in this area focus on optimizing data locality in individual loop nests. However, in many embedded applications, data access patterns exhibit a significant amount of inter-nest reuse. In this paper, we present a compiler strategy that optimizes inter-nest data locality using code restructuring and loop transformations. Our approach captures data reuse between all loop nests in the program and then splits and reorders the nests so that those sharing arrays are closer together. The transformed program is then further optimized using loop transformations. We improve on previous studies by using global program analysis and Integer Linear Programming to find the best nest ordering. The approach has been tested on many data-intensive embedded kernels and our simulation results indicate promising performance improvements.

Explaining StGermain: An aspect oriented environment for building extensible computational mechanics modeling software

Steve Quenette[1], Louis Moresi[2], P. D. Sunter[1] and Bill F. Appelbe[1]

[1]*Computational Software Development*
VPAC
Melbourne, Victoria, Australia
{steve, pds, bill}@vpac.org

[2]*School of Mathematical Sciences*
Monash University
Melbourne, Victoria, Australia
louis.moresi@sci.monash.edu.au

HPC scientific computational models are notoriously difficult to develop, debug, and maintain. The reasons for this are multifaceted — including difficulty of parallel programming, the lack of standard frameworks, and the lack of software engineering skills in scientific software developers.

In this paper we discuss the drivers, design and deployment of StGermain, a software framework that significantly simplifies the development of a spectrum of HPC computational mechanics models. The key distinction between StGermain and conventional approaches to developing computational models is that StGermain decomposes parallel scientific applications into a hierarchical architecture, supporting applications collectively built by a diverse community of scientists, modelers, computational scientists, and software engineers.

Automatic Performance Diagnosis of Parallel Computations with Compoitional Models

Li Li and Allen Malony

Computer and Information Science
University of Oregon
Eugene, OR, United States
{*lili, malony*}*@cs.uoregon.edu*

Performance tuning involves a diagnostic process to locate and explain sources of program inefficiency. A performance diagnosis system can leverage knowledge of performance causes and symptoms that come from expertise with parallel computational models. This paper extends our model-based performance diagnosis approach to programs with multiple models. We study two types of model compositions (nesting and restructuring) and demonstrate how the Hercule performance diagnosis framework can automatically discover and interpret performance problems due to model nesting in the FLASH application.

Reifying Control of Multi-Owned Network Resources

Nadeem Jamali[1] and Chen Liu[2]

[1]*Department of Compuer Science*
University of Saskatchewan
Saskatoon, SK, Canada
n.jamali@agents.usask.ca

[2]*Department of Computing Science*
University of Alberta
Edmonton, AB, Canada
cliu@cs.ualberta.ca

Communication delay is a key source of uncertainty in distributed systems. Existing approaches to reduce this uncertainty focus on maintaining sufficient surplus bandwidth; applications, on their part, are designed in ways to tolerate certain degree of uncertainty in communication delays. This leads to contention between the goals of optimal utilization and acceptable delays.

We argue that the multi-owned nature of today's networks offers opportunities to reason about and scalably control networks at a fine grain. An explicit treatment of network resource ownership and trade allows reasoning about acceptable delays. This can lead to scalable mechanisms for fine-grained accounting and reification of control, which make it possible to quantify and control network utilization.

We bring together ownership, fine-grained accounting, and reification of control in a model for resource acquisition and control called CyberOrgs. Cyberorgs encapsulate distributed computations with resources required for their execution. A cyberorg acquires resources required by its computations by buying them from other cyberorgs using eCash.

We present a novel approach for implementing fine-grained network resource control based on the CyberOrgs model. A prototype implementation is described with experimental results illustrating the effectiveness of control.

Evaluation of Stream Virtual Machine on Raw Processor

Jinwoo Suh[1], Richard Lethin[2], Stephen P. Crago[1], Janice O. Mcmahon[1] and Dong-In Kang[1]

[1]*Information Sciences Institute*
University of Southern California
Arlington, VA, USA
{*jsuh, crago, jmcmahon, dkang*}*@isi.edu*

[2]*Reservoir Labs.*
New York, New York, USA
lethin@reservoir.com

Stream processing exploits the properties of stream applications such as parallelism and throughput-oriented nature of the applications. One of the most recent approaches is community-supported Morphware Stable Interface (MSI) [11] used as a stable abstraction between High-Level Compilers (HLC) and Low-Level architecture-specific Compilers (LLC). We focus on one part of the MSI, the Stream Virtual Machine (SVM) [4][7][11]. We implemented a High-Level Compiler that produces SVM output renderings and SVM implementation. The SVM is implemented with the Raw Compiler as the LLC and an accompanying library. We also implemented stream applications such as matrix multiplication, FIR bank, and Ground Moving Target Indicator (GMTI) using the implemented compilers. These applications are optimized and the results are analyzed. The results show that the SVM framework is generally suitable for streaming applications on Raw processor.

A Multi-Level Parallel Implementation of a Program for Finding Frequent Patterns in a Large Sparse Graph

George Karypis[1] and Steve Reinhardt[2]

[1]*Computer Science Department*
University of Minnesota
Minneapolis, MN, USA
karypis@cs.umn.edu

[2]*Server Products Group*
SGI
Eagan, MN, USA
spr@sgi.com

Graphs capture the essential elements of many problems broadly defined as searching or categorizing. With the rapid increase of data volumes from sensors, many application disciplines need to process larger graphs quickly. This paper presents the results of parallelizing with OpenMP an algorithm that finds, in a single large labeled undirected sparse graph, the connected subgraphs with a given minimum number of edge-disjoint embeddings. Parallelism is exploited at two levels in the algorithm. The lack of a priori knowledge of the extent of parallelism for a given input required use of a dynamic, multi-level approach based on the proposed OpenMP taskq/task extensions. The parallel implementation required the addition of 21 directives and about 50 accompanying lines of code, in an original code of about 15,000 lines. Experimental results show excellent speed-up to 30 processors for the graphs used, with a best speed-up of 26.1 compared to the serial version. The taskq/task constructs show promise for problems exhibiting unstructured parallelism.

Bandwidth Efficient All-reduce Operation on Tree Topologies

Pitch Patarasuk and Xin Yuan

Department of Computer Science
Florida State University
Tallahassee, Florida, USA
{*patarasu, xyuan*}*@cs.fsu.edu*

We consider efficient implementations of the all-reduce operation with large data sizes on tree topologies. We prove a tight lower bound of the amount of data that must be transmitted to carry out the all-reduce operation and use it to derive the lower bound for the communication time of this operation. We develop a topology specific algorithm that is bandwidth efficient in that (1) the amount of data sent/received by each process is minimum for this operation; and (2) the communications do not incur network contention on the tree topology. With the proposed algorithm, the all-reduce operation can be realized on the tree topology as efficiently as on any other topology when the data size is sufficiently large. The proposed algorithm can be applied to several contemporary cluster environments, including high-end clusters of workstations with SMP and/or multi-core nodes and low-end Ethernet switched clusters. We evaluate the algorithm on various clusters of workstations, including a Myrinet cluster with dual-processor SMP nodes, an Infiniband cluster with two dual-core processors SMP nodes, and an Ethernet switched cluster with single processor nodes. The results show that the routines implemented based on the proposed algorithm significantly out-perform the native MPI_Allreduce and other recently developed algorithms for high-end SMP clusters when the data size is sufficiently large.

Runtime Optimization of Application Level Communication Patterns

Edgar Gabriel and Shuo Huang

Department of Computer Science
University of Houston
Houston, TX, USA
{*gabriel, shhuang*}*@cs.uh.edu*

This paper introduces the Abstract Data and Communication Library (ADCL). ADCL is an application level communication library aiming at providing the highest possible performance for application level communication operations on a given execution environment. The library provides for each communication pattern a large number of implementations and incorporates a runtime selection logic in order to choose the implementation leading to the highest performance of the application on the current platform. Two different runtime selection algorithms are currently available within ADCL: the library can either apply a brute force search strategy which tests all available implementations of a given communication pattern; alternatively, a heuristic relying on attributes characterizing an implementation has been developed in order to speed up the runtime decision procedure. The paper also evaluates the performance of a finite difference code using ADCL on an AMD Opteron cluster using InfiniBand and Gigabit Ethernet interconnects.

High Performance MPI on IBM 12x InfiniBand Architecture

Abhinav Vishnu[1], Brad Benton[2] and Dhabaleswar K Panda[3]

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

[2]Brad Benton
IBM
Austin, TX, USA
brad.benton@us.ibm.com

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

InfiniBand is becoming increasingly popular in the area of cluster computing due to its open standard and high performance. I/O interfaces like PCI-Express and GX+ are being introduced as next generation technologies to drive InfiniBand with very high throughput. HCAs with throughput of 8x on PCI-Express have become available. Recently, support for HCAs with 12x throughput on GX+ has been announced. In this paper, we design a Message Passing Interface (MPI) on IBM 12x Dual-Port HCAs, which consist of multiple send/recv engines per port. We propose and study the impact of various communication scheduling policies (binding, striping and round robin). Based on this study, we present a policy, EPC (Enhanced point-to-point and collective), which incorporates different kinds of communication patterns; (point-to-point (blocking, non-blocking), and collective) communication for data transfer. We implement our design and evaluate it with micro-benchmarks, collective communication and NAS parallel benchmarks. Using EPC on a 12x InfiniBand cluster with one HCA and one port, we can improve the performance by 41% with ping-pong latency test and 63-65% with the unidirectional and bi-directional bandwidth tests, compared with the default single-rail MPI implementation. Our evaluation on NAS Parallel Benchmarks shows an improvement of 7-13% in execution time for Integer Sort and Fourier Transform.

Coordinating Data Parallel SAC Programs with S-Net

Clemens Grelck, Sven-Bodo Scholz and Alex Shafarenko

Computer Science
University of Hertfordshire
Hatfield, Hertfordshire, United Kingdom
{C.Grelck, S.Scholz, A.Shafarenko}@herts.ac.uk

We propose a two-layered approach for exploiting different forms of concurrency in complex systems: We specify computational components in our functional array language SAC, which exploits data parallel properties of array processing code. The declarative stream processing language S-NET is used to orchestrate the collaborative behaviour of these components in a streaming network. We illustrate our approach by a hybrid implementation of a sudoku puzzle solver as a representative for more complex search problems.

Decomposing Partial Order Execution Graphs to Improve Message Race Detection

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*

In message-passing parallel applications, messages are not delivered in a strict order. In most applications, the computation results and the set of messages produced during the execution should be the same for all distinct orderings of messages delivery. Finding an ordering that produces a different outcome then reveals a message race. Assuming that the Partial Order Execution Graph (POEG) capturing the causality between events is known for a reference execution, the present paper describes techniques for identifying independent sets of messages and within each set equivalent message orderings. Orderings of messages belonging to different sets may then be re-executed independently from each other, thereby reducing the number of orderings that must be tested to detect message races. We integrated the presented techniques into the Dynamic Parallel Schedules parallelization framework, and applied our approach on an image processing, a linear algebra, and a neighborhood-dependent parallel computation. In all cases, the number of possible orderings is reduced by several orders of magnitudes. In order to further reduce this number, we describe an algorithm that generates a subset of orderings that are likely to reveal existing message races.

Multi-Core Model Checking with SPIN

Gerard J. Holzmann[1] and Dragan Bosnacki[2]

[1] *Laboratory for Reliable Software*
NASA/JPL
Pasadena, CA, USA
gerard@spinroot.com

[2] *Biomedical Engineering*
Eindhoven University
Eindhoven, The Netherlands
dragan@win.tue.nl

We present the first experimental results on the implementation of multi-core model checking algorithms for the SPIN model checker. These algorithms specifically target shared-memory systems, and are initially restricted to dual-core systems. The extensions we have made require only small changes in the SPIN source code, and preserve virtually all existing verification modes and optimization techniques supported by SPIN, including the verification of both safety and liveness properties and the verification of SPIN models with embedded C code fragments.

The Mojave Compiler: Providing Language Primitives for Whole-Process Migration and Speculation for Distributed Applications

Justin D. Smith[1], Cristian Tapus[2] and Jason Hickey[3]

[1]Computer and Information Sciences
University of Pennsylvania
Philadelphia, PA, USA
jyasu@cis.upenn.edu

[2]Caltech's Center for Advanced Computing Research
California Institute of Technology
Pasadena, CA, USA
crt@cs.caltech.edu

[3]Computer Science
California Institute of Technology
Pasadena, CA, USA
jyh@cs.caltech.edu

We present an approach for implementing language-level primitives for whole-process migration and speculative execution in a compiler and associated runtime environment. These primitives are exposed to the user through simple language constructs that do not require the user to manage process state explicitly. With migration and speculation we show how the user can quickly add persistent checkpoints to any large-scale distributed application that requires longevity in a faulty environment. We demonstrate the use of migration and speculation primitives for checkpointing in a canonical grid computation application, and analyze the results of this implementation.

Packet Loss Burstiness: Measurements and Implications for Distributed Applications

David X. Wei[1], Pei Cao[2] and Steven H. Low[1]

[1]Division of Engineering and Applied Science
California Institute of Technology
Pasadena, CA, 91125
weixl@cs.caltech.edu, slow@caltech.edu

[2]Department of Computer Science
Stanford University
Stanford, CA, 94305
cao@cs.stanford.edu

Many modern massively distributed systems deploy thousands of nodes to cooperate on a computation task. Network congestions occur in these systems. Most applications rely on congestion control protocols such as TCP to protect the systems from congestion collapse. Most TCP congestion control algorithms use packet loss as signal to detect congestion.

In this paper, we study the packet loss process in sub-round-trip-time (sub-RTT) timescale and its impact on the loss-based congestion control algorithms. Our study suggests that the packet loss in sub-RTT timescale is very bursty. This burstiness leads to two effects. First, the sub-RTT burstiness in packet loss process leads to complicated interactions between different loss-based algorithms. Second, the sub-RTT burstiness in packet loss process makes the latency of data transfers under TCP hard to predict.

Our results suggest that the design of a distributed system has to seriously consider the nature of packet loss process and carefully select the congestion control algorithms best suited for the distributed computation environments.

FixD: Fault Detection, Bug Reporting, and Recoverability for Distributed Applications

Cristian Tapus[1] and David A. Noblet[2]

[1]*Caltech's Center for Advanced Computing Research*
California Institute of Technology
Pasadena, CA, USA
crt@cs.caltech.edu

[2]*Computer Science Department*
California Institute of Technology
Pasadena, CA, USA
dnoblet@cs.caltech.edu

Model checking, logging, debugging, and checkpointing/recovery are great tools to identify bugs in small sequential programs. The direct application of these techniques to the domain of distributed applications, however, has been less effective (mostly owing to the high degree of concurrency in this context).

This paper presents the design of a hybrid tool, FixD, that attempts to address the deficiencies of these tools with respect to their application to distributed systems by using a novel composition of several of these existing techniques. The authors first identify and describe the four abstract components that comprise the FixD tool, then conclude with a proposal for how existing tools can be used to implement these components.

Workshop 5
Int'l Workshop on Java and Components for Parallelism, Distribution and Concurrency JAVAPDC 2007

5 JAVAPDC • Workshop on Java and Components for Parallelism, Distribution and Concurrency

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 and components
- 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;
- Embeded 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:

Jack Dongarra, University of Tenessee
Doug Lea, State University of New York at Oswego
Vladimir Getov, University of Westminster, London, U.K.
George K. Thiruvathukal, Loyola University Chicago
David Walker, Cardiff University, UK

Organization Committee:

Arnaud Casteigts, Université Bordeaux I

Revisiting Deterministic Multithreading Strategies

Jörg Domaschka, Andreas I. Schmied, Hans P. Reiser and Franz J. Hauck

Institute of Distributed Systems
Ulm University
James-Franck-Ring O-27, 89069 Ulm, Germany
{*joerg.domaschka, andreas.schmied, hans.reiser, franz.hauck*}*@uni-ulm.de*

Deterministic behaviour is a prerequisite for most approaches to object replication. In order to avoid the nondeterminism of multithreading, many object replication systems are limited to using sequential method execution. In this paper, we survey existing application-level scheduling algorithms that enable deterministic concurrent execution of object methods. Multithreading leads to a more efficient execution on multiple CPUs and multi-core CPUs, and it enables the object programmer to use condition variables for coordination between multiple invocations. In existing algorithms, a thread may only start or resume if there are no potentially nondeterministic conflicts with other running threads. A decision only based on past actions, without knowledge of future behaviour, must use a pessimistic strategy that can cause unnecessary restrictions to concurrency. Using a priori knowledge about future actions of a thread allows increasing the concurrency. We propose static code analysis as a way for predicting the lock acquisitions of object methods.

Performance and Scalability of a Component-Based Grid Application

Nikos Parlavantzas[1], Matthieu Morel[2], Vladimir Getov[1], Francoise Baude[2] and Denis Caromel[2]

[1]*Harrow School of Computer Science*
University of Westminster
Harrow, United Kingdom
{*N.Parlavantzas, V.S.Getov*}*@westminster.ac.uk*

[2]*INRIA Sophia Antipolis*
Sophia Antipolis, France
{*Matthieu.Morel, Francoise.Baude,*
Denis.Caromel}*@inria.fr*

Component-based software development has emerged as an effective approach to building flexible systems, but there is little experience in applying this approach to Grid programming. This paper presents our initial experience with reengineering a high performance numerical solver of the 3D Maxwell's equations to become a component-based Grid application. The adopted Grid component model is an extension of the generic Fractal model that specifically targets Grid environments. The paper provides qualitative and quantitative evidence that componentisation has improved the modifiability and reusability of the application while not compromising its performance. We also report a set of experimental results about the performance and scalability of our code, which confirm the feasibility of the component-based approach for this class of applications.

Dynamic Load-Balancing and High Performance Communication in Jcluster

Baoyin Zhang[1], Zeyao Mo[1], Guangwen Yang[2] and Weimin Zheng[2]

[1]*Institute of Applied Physics and Computational Mathematics*
Beijing, P.R. China
{*zby, zeyao_mo*}*@iapcm.ac.cn*

[2]*Department of Computer Science and Technology*
Tsinghua University
Beijing, P.R. China
{*ygw, zwm-dcs*}*@tsinghua.edu.cn*

This paper describes the dynamic load-balancing and high performance communication provided in Jcluster, an efficient Java parallel environment. For the efficient load-balancing, we implement a task scheduler based on a Transitive Random Stealing algorithm, which improves the Random Stealing, a well-known load-balancing algorithm. The experiment results show that the scheduler performs efficiently, especially for a large-scale cluster. With the method of asynchronously multithreaded transmission, a high performance PVM-like and MPI-like message passing interface is implemented in pure Java. The evaluation of the communication performance is conducted among Jcluster, LAM-MPI and mpiJava on LAM-MPI based on the Java Grande Forum's pingpong benchmark.

Analysis of Different Future Objects Update Strategies in ProActive

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 large-scale distributed systems, asynchronous communication and future objects are becoming wide spread mechanisms to tolerate high latencies and to improve global performances. Automatic continuation, that is the propagation of a future object outside the activity that has generated it, can be used to further increase concurrency at system level through the anticipation of tasks. An important aspect of automatic continuation, which can cause different performance in different application and deployment scenarios, is the mechanism for updating result values of future objects, when they are ready. In this paper, we analyze the behaviour of the implementation of different updating strategies, by comparing them with the one currently implemented in ProActive. The experimental results show that the lazy home-based strategy behaves better than other strategies in some application scenarios that are very common in distributed applications.

Client-Side Implementation of Dynamic Asynchronous Invocations for Web Services

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

Web Services are becoming more and more fundamental building blocks of Web-based distributed applications and a core technology for Grid systems. Due to their flexibility, Web Services easily combine, in a common and coherent framework, ubiquitous computing with heterogeneous applications composed of different kinds of resources and, typically distributed in many organizations. We expect that this technology will follow the same evolution paths that have characterized other technologies so far, with some specificity due to the openness and size of the application context. In this connection, optimizations tied to invocations and workflows are assuming a primary role in Web Services research. The synchronous request/reply nature of the most diffused underling protocol (HTTP) introduces several restrictions in many application scenarios. On the other hand, asynchronous interactions are allowed by using message oriented middleware platforms, like JMS, which are typically harder to handle than object- and process-oriented middleware. In this paper, we propose a first implementation of a module that allows for dynamic Web Services invocations, which, on the basis of meta-data added to WSDL, is able to select the most appropriate invocation technique for calling a Web Services operation.

Java and asynchronous iterative applications: large scale experiments

Jacques M. Bahi, Raphaël Couturier, David Laiymani and Kamel Mazouzi

Laboratoire d'Informatique de l'université de Franche-Comté (LIFC)
Université de Franche-Comté
IUT de Belfort-Montbéliard, Rue Engel Gros, BP 527, 90016 Belfort cedex, France
{*jacques.bahi, raphael.couturier, david.laiymani, kamel.mazouzi*}*@iut-bm.univ-fcomte.fr*

This paper focuses on large scale experiments with Java and asynchronous iterative applications. In those applications, tasks are dependent and the use of distant clusters may be difficult, for example, because of latencies, heterogeneity, and synchronizations. Experiments have been conducted on the Grid'5000 platform using a new version of the Jace environment. We study the behavior of an application (the Poisson problem) with the following experimentation conditions: one and several sites, large number of processors (from 80 to 500), different communication protocols (RMI, sockets and NIO), synchronous and asynchronous model. The results we obtained, demonstrate both the scalability of the Jace environment and its ability to support wide-area deployments and the robustness of asynchronous iterative algorithms in a large scale context.

Parallel Java: A Unified API for Shared Memory and Cluster Parallel Programming in 100% Java

Alan Kaminsky

Department of Computer Science
Rochester Institute of Technology
Rochester, NY, USA
ark@cs.rit.edu

Parallel Java is a parallel programming API whose goals are (1) to support both shared memory (thread-based) parallel programming and cluster (message-based) parallel programming in a single unified API, allowing one to write parallel programs combining both paradigms; (2) to provide the same capabilities as OpenMP and MPI in an object oriented, 100% Java API; and (3) to be easily deployed and run in a heterogeneous computing environment of single-core CPUs, multi-core CPUs, and clusters thereof. This paper describes Parallel Java's features and architecture; compares and contrasts Parallel Java to other Java-based parallel middleware libraries; and reports performance measurements of Parallel Java programs.

A Survey of Worst-Case Execution Time Analysis for Real-Time Java

Trevor Harmon and Raymond Klefstad

Dept. of Electrical Engineering and Computer Science
University of California, Irvine
Irvine, California, USA
{tharmon, klefstad}@uci.edu

As real-time systems become more prevalent, there is a need to guarantee that these increasingly complex systems perform as designed. One technique involves a static analysis to place an upper bound on worst-case execution time (WCET). Other techniques aim for new architectures and algorithms that reduce the WCET. At the same time, there is a growing interest in using Java for real-time systems. Several WCET analysis prototypes for Java have been created, and more are under development.

This paper provides a comprehensive survey of research that combines WCET analysis with the Java domain. We begin by explaining the importance of WCET analysis and why it is so difficult to perform adequately. We then examine the features that make Java an attractive platform for WCET analysis, as well as the new challenges it brings. Finally, we provide a survey of prior work on this subject, organized as a simple one-level taxonomy.

A Model-Driven Approach to Job/Task Composition in Cluster Computing

Neeraj Mehta, Yogesh Kanitkar, Konstantin Läufer and George K. Thiruvathukal

Emerging Technologies Laboratory, Department of Computer Science
Loyola University Chicago
Chicago, IL 60611, USA
neerajcmehta@gmail.com, yogesh_4u@yahoo.com, {laufer, gkt}@cs.luc.edu

In the general area of high-performance computing, object-oriented methods have gone largely unnoticed. In contrast, the Computational Neighborhood (CN), a framework for parallel and distributed computing with a focus on cluster computing, was designed from ground up to be object-oriented. This paper describes how we have successfully used UML in the following model-driven, generative approach to job/task composition in CN. We model CN jobs using activity diagrams in any modeling tool with support for XMI, an XML-based external representation of UML models. We then export the activity diagrams and use our XSLT-based tool to transform the resulting XMI representation to CN job/task composition descriptors.

High Performance Java Sockets for Parallel Computing on Clusters

Guillermo L. Taboada, Juan Touriño and Ramon Doallo

Dept. of Electronics and Systems
University of A Coruña
A Coruña, Spain
{taboada, juan, doallo}@udc.es

The use of Java for parallel programming on clusters relies on the need of efficient communication middleware and high-speed cluster interconnect support. Nevertheless, currently there are no solutions that fully fulfill these issues. In this paper, a Java sockets library has been tailored to increase the efficiency of Java parallel applications on clusters. This library supports high-speed cluster interconnects and its API has been extended to meet the requirements of a high performance Java RMI implementation and Java parallel applications on clusters. Thus, it provides Java with a more efficient communication middleware on clusters. The performance evaluation of this middleware on a Gigabit Ethernet (GbE) and a Scalable Coherent Interface (SCI) cluster has shown experimental evidence of throughput increase. Moreover, qualitative aspects of the solution such as transparency to the user, interoperability with other systems and no need of source code modifications are decisive to boost the performance of existing Java parallel applications and their developments in high performance Java cluster computing.

Workshop 6
Workshop on Nature Inspired Distributed Computing
NIDISC 2007

• 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

Applying Ant Colony Optimization Metaheuristic to the DAG Layering Problem

Radoslav Andreev, Patrick Healy and Nikola S. Nikolov

Department of Computer Science and Information Systems
University of Limerick
Limerick, IRELAND
{*radoslav.andreev, patrick.healy, nikola.nikolov*}*@ul.ie*

This paper presents the design and implementation of an Ant Colony Optimization based algorithm for solving the DAG Layering Problem. This algorithm produces compact layerings by minimising their width and height. Importantly it takes into account the contribution of dummy vertices to the width of the resulting layering.

A Genetic Approach for Distributing Semantic Databases of Crowd Simulations

Miguel Lozano, Juan Manuel Orduña and Vicente Cavero

Departamento de Informática
Universidad de Valencia
Burjassot, Valencia, Spain
{*Miguel.Lozano, Juan.Orduna, Vicente.Cavero*}*@uv.es*

Last years have witnessed how crowd simulations have become an essential tool for many virtual environment applications. These applications require both rendering visually plausible images and managing the behavior of autonomous agents, and therefore they need a scalable design that allow them to simultaneously tackle these two requirements. One of the main problems in the design of scalable crowd simulations consists of efficiently distributing the semantic database containing the virtual world among different computers.

In this paper, we propose a genetic approach for distributing the semantic database of crowd simulations in such a way that the dependencies among the computers hosting the pieces of the database are minimized. The proposed approach avoids the saturation of these computers by ensuring that the size of the pieces assigned to each computer is properly balanced. The performance evaluation results show that the proposed approach significantly reduces the resulting overhead in regard to other local search methods, regardless of the movement pattern of the agents. Therefore, it allows an effective partition of the semantic database.

Recurrent neural networks towards detection of SQL attacks

Jaroslaw Skaruz[1] and Franciszek Seredynski[1,2,3]

[1] Institute of Computer Science
University of Podlasie
Siedlce, Poland
jaroslaw.skaruz@ap.siedlce.pl, sered@ipipan.waw.pl

[2] Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland

[3] Computer Networks Department
Polish-Japanese Institute of Information Technology
Warsaw, Poland

In the paper we present a new approach based on application of neural networks to detect SQL attacks. SQL attacks are those attacks that take advantage of using SQL statements to be performed. The problem of detection of this class of attacks is transformed to time series prediction problem. SQL queries are used as a source of events in a protected environment. To differentiate between normal SQL queries and those sent by an attacker, we divide SQL statements into tokens and pass them to our detection system, which predicts the next token, taking into account previously seen tokens. In the learning phase tokens are passed to recurrent neural network (RNN) trained by backpropagation through time (BPTT) algorithm. Teaching data are shifted by one token forward in time with relation to input. The purpose of the testing phase is to predict the next token in the sequence. All experiments were conducted on Jordan and Elman networks using data gathered from PHP Nuke portal. Experimental results show that the Jordan network outperforms the Elman network predicting correctly queries of the length up to ten.

An Artificial Immune System for Heterogeneous Multiprocessor Scheduling with Task Duplication

Young Choon Lee and Albert Y. Zomaya

School of Information Technologies/Advanced Networks Research Group
The University of Sydney
Sydney, NSW, Australia
{yclee, zomaya}@it.usyd.edu.au

In this study, we investigate the task scheduling problem in heterogeneous computing environments and propose a novel scheduling algorithm, called the Artificial Immune System with Duplication (AISD) algorithm that efficiently tackles the problem. The AISD algorithm incorporates the clonal selection principle in the immune system and task duplication into the scheduling process. Based on the performance results obtained from extensive experiments conducted with a comprehensive set of both randomly generated and well-known application task graphs and various system configurations, AISD consistently outperformed the two existing algorithms by a noticeable margin, especially when scheduling communication intensive task graphs.

Protein Secondary Structure Prediction using Bayesian Inference method on Decision fusion algorithms

Somasheker Akkaladevi[1] and Ajay K Katangur[2]

[1]*Department of Computer Information Systems*
Virginia State University
Petersburg, VA, USA
sakkaladevi@vsu.edu

[2]*Department of Computing Sciences*
Texas A&M University - Corpus Christi
Corpus Christi, TX, USA
ajay.katangur@tamucc.edu

Prediction of protein secondary structure (alpha-helix, beta-sheet, coil) from primary sequence of amino acids is a very challenging task, and the problem has been approached from several angles. Previously research was performed in this field using several techniques such as neural networks, Simulated annealing (SA) and Genetic algorithms (GA) for improving the protein secondary structure prediction accuracy. Decision fusion methods such as the Committee method and Correlation methods were also used in combination with the profile-based neural networks and AI algorithms for achieving better prediction accuracy. In this research we investigate the Bayesian inference method for predicting the protein secondary structure. The Bayesian inference method proposed in this research uses the results from the committee and correlation methods to achieve better prediction accuracy. Simulations are performed using the RS126 data set. The results show that the protein secondary structure prediction accuracy can be improved by more than 2% using the Bayesian inference method.

Parallel Tabu Search and the Multiobjective Vehicle Routing Problem with Time Windows

Andreas Beham

Institute for Formal Models and Verification
Johannes Kepler University
Linz, Austria
andreas@heuristiclab.com

In this paper the author presents three approaches to parallel Tabu Search, applied to several instances of the Capacitated Vehicle Routing Problem with Time Windows (CVRPTW). Attention in this work was given to keep the parallel implementations simple. The parallel algorithms are of two kinds: Two of them are parallel with respect to functional decomposition and one approach is a collaborative multisearch TS. The implementation builds upon a framework called Distributed metaheuristics or DEME for short. Tests were performed on an SGI Origin 3800 supercomputer at the Johannes Kepler University of Linz, Austria.

A hybrid Evolutionary Algorithm for the Dynamic Resource Constrained Task Scheduling Problem

André Renato Villela Da Silva[1] and Luiz Satoru Ochi[2]

[1]Instituto de Computação
Universidade Federal Fluminense
Niterói, RJ, Brasil
avillela@ic.uff.br

[2]Instituto de Computação
Universidade Federal Fluminense
Niterói, RJ, Brasil
satoru@ic.uff.br

This work presents a new hybrid Evolutionary Algorithm for the Dynamic Resource Constrained Task Scheduling Problem (DRCTSP). The most important differences between the new EA and the previously proposed EAs are an intensification/diversification mechanism that tries to avoid premature convergence in local optimal solutions and a version combining an exact method (CPLEX) with EAs. Some preliminary tests were done and results are very promising.

Parallel Processing for Multi-objective Optimization in Dynamic Environments

Mario Cámara[1], Julio Ortega[1] and Francisco J. Toro[2]

[1]Dept. of Computer Architecture and Technology
University of Granada
Granada, Andalusia, Spain
{mcamara, julio}@atc.ugr.es

[2]Dept. of Signal Theory, Telematics and Communications
University of Granada
Granada, Andalusia, Spain
ftoro@ugr.es

This paper deals with the use of parallel processing for multi-objective optimization in applications in which the objective functions, the restrictions, and hence also the solutions can change over time. These dynamic optimization problems appear in quite different real-world applications with relevant socio-economic impact. The procedure here presented is based on PSFGA, a parallel evolutionary procedure for multi-objective optimization. It uses a master process that distributes the population among the processors in the system (that evolve their corresponding solutions according to an island model), and collects and adjusts the set of local Pareto fronts found by each processor (this way, the master also allows an implicit communication among islands). Moreover, the procedure exclusively uses non-dominated individuals for the selection and variation, and maintains the diversity of the approximation to the Pareto front by using a strategy based on a crowding distance.

Distributed Adaptive Particle Swarm Optimizer in Dynamic Environment

Xiaohui Cui[1] and Thomas E. Potok[2]

[1]*Computational Sciences and Engineering Division*
Oak Ridge National Laboratory
Oak Ridge, TN, USA
cuix@ornl.gov

[2]*Computational Sciences and Engineering Division*
Oak Ridge National Laboratory
Oak Ridge, TN, USA
potokte@ornl.gov

Particle Swarm Optimization (PSO) is a population-based stochastic optimization technique, which can be used to find an optimal, or near optimal, solution to a numerical and qualitative problem. In PSO algorithm, the problem solution emerges from the interactions among many simple individual agents called particles. In the real world, we have to frequently deal with searching and tracking an optimal solution in a dynamical and noisy environment. This demands that the algorithm not only find the optimal solution but also track the trajectory of the non-stationary solution. The traditional PSO algorithm lacks the ability to track the changing optimal solution in a dynamic and noisy environment. In this paper, we present a distributed adaptive PSO (DAPSO) algorithm that can be used to track a non-stationary optimal solution in a dynamically changing and noisy environment.

Evolution of Strategy Driven Behavior in Ad Hoc Networks Using a Genetic Algorithm

Marcin Seredynski[1], Pascal Bouvry[1] and Mieczyslaw A. Klopotek[2]

[1]*Faculty of Sciences, Technology and Communication*
University of Luxembourg
Luxembourg, Luxembourg
{*marcin.seredynski, pascal.bouvry*}*@uni.lu*

[2]*Institute of Computer Science*
Polish Academy of Sciences
Warsaw, Poland
klopotek@ipipan.waw.pl

In this paper we address the problem of selfish behavior in ad hoc networks. We propose a strategy driven approach which aims at enforcing cooperation between network participants. Each node (player) is using a strategy that defines conditions under which packets are being forwarded. Such strategy is based on the notion of trust and activity of the source node of the packet. This way network participants are enforced to forward packets and to reduce the amount of time of being in a sleep mode. To evaluate strategies we use a new game theory based model of an ad hoc network. This model has some similarities with the Iterated Prisoner's Dilemma under the Random Pairing game where randomly chosen players receive payoffs that depend on the way they behave. Our model of the network also includes a simple reputation collection and trust evaluation mechanisms. A genetic algorithm (GA) is applied to find good strategies. Experimental results show that approach can successfully enforce cooperation among ad hoc networks participants.

Time Series Forecasting by means of Evolutionary Algorithms

Cristobal Luque[1], Jose Maria Valls Ferran[2] and Pedro Isasi Viñuela[3]

[1]Departamento de Informática
Carlos III de Madrid
Madrid, Spain
cluque@inf.uc3m.es

[2]Departamento de Informática
Carlos III de Madrid
Madrid, Spain
jvalls@inf.uc3m.es

[3]Departamento de Informática
Carlos III de Madrid
Madrid, Spain
isasi@ia.uc3m.es

Many physical and artificial phenomena can be described by time series. The prediction of such phenomenon could be as complex as interesting. There are many time series forecasting methods, but most of them only look for general rules to predict the whole series. The main problem is that time series usually have local behaviours that don't allow forecasting the time series by general rules. In this paper, a new method for finding local prediction rules is presented. Those local prediction rules can attain a better general prediction accuracy. The method presented in this paper is based on the evolution of a rule system encoded following a Michigan approach. For testing this method, several time series domains have been used: a widely known artificial one, the Mackey-Glass time series, and two real world ones, the Venice Lagon and the sunspot time series.

Efficient Batch Job Scheduling in Grids using Cellular Memetic Algorithms

Fatos Xhafa[1], Enrique Alba[2] and Bernabé Dorronsoro[3]

[1]Department of Languages and Informatic Systems
Polytechnic University of Catalonia
Barcelona, SPAIN
fatos@lsi.upc.edu

[2]Dpto. de Lenguajes y Ciencias de la Computación
E.T.S.I. Informática
Málaga, SPAIN
eat@lcc.uma.es

[3]Dpto. de Lenguajes y Ciencias de la Computación
E.T.S.I. Informática
Málaga, SPAIN
bernabe@lcc.uma.es

Computational Grids are an important emerging paradigm for large-scale distributed computing. As Grid systems become more wide-spread, techniques for efficiently exploiting the large amount of Grid computing resources become increasingly indispensable. A key aspect in order to benefit from these resources is the scheduling of jobs to Grid resources. Due to the complex nature of Grid systems, the design of efficient Grid schedulers becomes challenging since such schedulers have to be able to optimize many conflicting criteria in very short periods of time.

In this work we exploit the capabilities of Cellular Memetic Algorithms (cMAs) for obtaining efficient batch schedulers for Grid systems. A careful design of the cMA methods and operators for the problem yielded to an efficient and robust implementation. Our experimental study, based on a known static benchmark for the problem, shows that this heuristic approach is able to deliver very high quality planning of jobs to Grid nodes and thus it can be used to design efficient dynamic schedulers for real Grid systems. Such dynamic schedulers can be obtained by running the cMA-based scheduler in batch mode for a very short time to schedule jobs arriving to the system since the last activation of the cMA scheduler.

Reconfigurable Architecture for Biological Sequence Comparison in Reduced Memory Space Reduced Memory Space

Azzedine Boukerche[1], Jan M. Correa[2], Alba Cristina M. A. De Melo[2], Ricardo P. Jacobi[2] and Adson F. Rocha[3]

[1]*PARADISE Research Laboratory*
SITE, University of Ottawa
Ottawa, ONT, Canada
boukerch@site.uottawa.ca

[2]*Department of Computer Science*
University of Brasilia
Brasilia, Brazil
{jan, albamn, rjacobi}@cic.unb.br

[3]*Department of Electrical Engineering*
University of Brasilia
Brasilia, Brazil
adson@ene.unb.br

DNA sequence alignment is a very important problem in bioinformatics. The algorithm proposed by Smith-Waterman (SW) is an exact method that obtains optimal local alignments in quadratic space and time. For long sequences, quadratic complexity makes the use of this algorithm impractical. In this scenario, the use of a reconfigurable architecture is a very attractive alternative. This article presents the design and evaluation of an FPGA-based architecture that obtains the similarity score between DNA sequences, as well as its coordinates. The results obtained in a Xilinx xc2vp70 FPGA prototype presented a speedup of 246.9 over the software solution to compare sequences of size 100MBP and 100BP, respectively. Different from others hardware solutions that just calculate alignment scores, our design was able to avoid architectures bottlenecks and accelerate the most computer intensive part of a sequence alignment software algorithm.

A Comparative Study of Parallel Metaheuristics for Protein Structure Prediction on the Computational Grid

Alexandru-Adrian Tantar, Nouredine Melab and El-Ghazali Talbi

LIFL - UMR 8022 USTL/CNRS
USTL - Université des Sciences et Technologies de Lille
Villeneuve d'Ascq, Nord, France
{tantar, melab, talbi}@lifl.fr

A comparative study of parallel metaheuristics executed in grid environments is proposed, having as case study a genetic algorithm, a simulated annealing algorithm and a random search method. The random search method was constructed in order to offer a lower bound for the comparison. Furthermore, a conjugated gradient local search method is employed for each of the algorithms, at different points on the execution path. The algorithms are evaluated using the protein structure prediction problem, the benchmark instances consisting of the *tryptophan-cage* protein (Brookhaven Protein Data Bank ID 1L2Y) and α-*cyclodextrin*. The algorithms are designed to benefit from the grid environment although having no particular optimization for the specified benchmarks. The presented results are obtained by running the algorithms independently and, in a second time, in conjunction with the conjugated gradient search method. Experimentations were performed on a nation-wide grid reuniting five distinct administrative domains and cumulating 400 CPUs. The complexity of the protein structure prediction problem remains prohibitive as far as large proteins are concerned, making the use of parallel computing on the computational grid essential for its efficient resolution.

Workshop 7
Workshop on High Performance Computational Biology
HiCOMB 2007

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:

Srinivas Aluru, Iowa State University, USA

David A. Bader, Georgia Institute of Technology, USA

Program Co-Chairs:

Shankar Subramaniam, University of California at San Diego, USA

Ananth Grama, Purdue University, USA

Program Committee:

Alberto Apostolico, Accademia dei Lincei & Georgia Tech
Joel Bader, Johns Hopkins University
Vineet Bafna, University of California San Diego
Jesus Izaguirre, University of Notre Dame
George Karypis, University of Minnesota
Daisuke Kihara, Purdue University
Vipin Kumar, University of Minnesota
Satoru Miyano, HGC, University of Tokyo
Ben Raphael, Brown University
Naren Ramakrishnan, Virginia Tech
Joel Saltz, Ohio State
Mona Singh, Princeton University
Tandy Warnow, University of Texas, Austin

Keynote – Optical Mapping of the Maize Genome

Michael S. Waterman

Professor of Biological Sciences, Mathematics, Computer Science
University of Southern California
Los Angeles, CA, USA
msw@usc.edu

A new technology, optical mapping, is used to infer the genome map of the location of short sequence patterns called restriction sites. The technology, developed by David Schwartz, allows the visualization of the maps of randomly located single molecules of length from one-half to one million base pairs. The genome map is constructed from overlapping these shorter maps. The substantial mathematical and computational challenges come from modeling the measurement errors and from the process of map assembly. We will report on our experience with assembling the maize genome.

On the Path to Enable Multi-scale Biomolecular Simulations on PetaFLOPS Supercomputer with Multi-core Processors

Sadaf R. Alam and Pratul K. Agarwal

Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN, USA
{*alamsr, agarwalpk*}*@ornl.gov*

Biological processes occurring inside cell involve multiple scales of time and length; many popular theoretical and computational multi-scale techniques utilize biomolecular simulations based on molecular dynamics. Till recently, the computing power required for simulating the relevant scales was even beyond the reach of fastest supercomputers. The availability of petaFLOPS-scale computing power in near future holds great promise. Unfortunately, the bio-simulations software technology has not kept up with the changes in hardware. In particular, with the introduction of multi-core processing technologies in systems with tens of thousands of processing cores, it is unclear whether the existing bio-molecular simulation frameworks will be able to scale and to utilize these resources effectively. While the multi-core processing systems provide higher processing capabilities, their memory and IO subsystems are posing new challenges to application and system software developers. In this preliminary study, we attempt to characterize computation, communication and memory efficiencies of bio-molecular simulations on a Cray XT3 system, which has recently been upgraded to dual-core Opteron processors. We identify that the application efficiencies using the multi-core processors reduce with the increase of the simulated system size. Further, we measure the communication overhead of using both cores in the processor simultaneously and identify that the MPI communication performance can be as low as 50% as compared to the single-core execution times. We conclude that not only the biomolecular simulations need to be aware of the underlying multi-core hardware in order to achieve maximum performance but also the system software needs to provide processor and memory placement features in the high-end systems. Our results on a stand-alone dual-core AMD system confirm that combinations of processor and memory affinity schemes can result in over 12% performance gains.

Analysis of a Computational Biology Simulation Technique on Emerging Processing Architectures

Jeremy S. Meredith, Sadaf R. Alam and Jeffrey S. Vetter

Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN, USA
{*jsmeredith, alamsr, vetter*}*@ornl.gov*

Multi-paradigm, multi-threaded and multi-core computing devices available today provide several orders of magnitude performance improvement over mainstream microprocessors. These devices include the STI Cell Broadband Engine, Graphical Processing Units (GPU) and the Cray massively-multithreaded processors available in desktop computing systems as well as proposed for supercomputing platforms. The main challenge in utilizing these powerful devices is their unique programming paradigms. GPUs and the Cell systems require code developers to manage code and data explicitly, while the Cray multithreaded architecture requires them to generate a very large number of threads or independent tasks concurrently. In this paper, we explain strategies for optimizing a molecular dynamics (MD) calculation that is used in bio-molecular simulations on three devices: Cell, GPU and MTA-2. We show that the Cray MTA-2 system requires minimal code modification and does not outperform the microprocessor runs; but it demonstrates an improved workload scaling behavior over the microprocessor implementation. On the other hand, substantial porting and optimization efforts on the Cell and the GPU systems result in a 5x to 6x improvement, respectively, over a 2.2 GHz Opteron system.

A Graph-Theoretic Analysis of the Human Protein-Interaction Network Using Multicore Parallel Algorithms

David A. Bader and Kamesh Madduri

College of Computing
Georgia Institute of Technology
Atlanta, Georgia, USA
{*bader, kamesh*}*@cc.gatech.edu*

Protein-interaction network (PIN) analysis provides valuable insight into an organism's functional organization and evolutionary behavior. In this paper, we study a PIN formed by high-confidence *human protein interactions* obtained from various public interaction databases. This is the *largest human PIN* studied to date, comprising nearly 18,000 proteins and 44,000 interactions. A novel contribution of this paper is the computation of betweenness centrality, a graph-theoretic metric that is found to be positively correlated with the essentiality and evolutionary age of a protein. We observe that proteins with high betweenness centrality, but low connectivity are abundant in the human PIN. We have designed an efficient and portable parallel implementation for the calculation of this compute-intensive centrality metric. On the Sun Fire T2000 server with the UltraSparc T1 (Niagara) processor, we achieve a relative speedup of about 16 using 32 threads for a typical instance of betweenness centrality, reducing the running time from several minutes to 13 seconds. This work was supported in part by NSF Grants CNS-0614915, CAREER CCF-0611589, DBI-0420513 and ITR EF/BIO 03-31654.

Biomolecular Path Sampling Enabled by Processing in Network Storage

Paul Brenner[1], Justin Wozniak[1], Doug Thain[1], Aaron Striegel[1], Jeff Peng[2] and Jesus Izaguirre[1]

[1]*Computer Science and Engineering*
University of Notre Dame
Notre Dame, IN, USA
{*pbrenne1, jwozniak, dthain, striegel, izaguirr*}*@nd.edu*

[2]*Chemistry and Biochemistry*
University of Notre Dame
Notre Dame, IN, USA
jpeng@nd.edu

Computationally complex and data intensive atomic scale biomolecular simulation is enabled via Processing in Network Storage (PINS): a novel distributed system framework to overcome bandwidth, compute, storage, and security challenges inherent to the wide area computation and storage grid. High throughput data generation requirements for our scientific target are overcome through novel aggregate bandwidth capabilities. Biomolecular simulation methods are correlated with the client tools, hybrid database/file server (GEMS), computation engine (Condor), virtual file system adapter (Parrot), and local file servers (Chirp). PINS performance is reported for the path sampling of a solvated protein domain requiring over 1000 simulations with total output data generation on the order of 1TB.

Preliminary results in accelerating profile HMM search on FPGAs

Arpith C. Jacob, Joseph M. Lancaster, Jeremy D. Buhler and Roger D. Chamberlain

Department of Computer Science and Engineering
Washington University in St. Louis
St. Louis, Missouri, USA
{*jarpith, jmlancas, jbuhler, roger*}*@cse.wustl.edu*

Comparison between biosequences and probabilistic models is an increasingly important part of modern DNA and protein sequence analysis. The large and growing number of such models in today's databases demands computational approaches to searching these databases faster, while maintaining high sensitivity to biologically meaningful similarities. This work describes an FPGA-based accelerator for comparing proteins to Hidden Markov Models of the type used to represent protein motifs in the popular HMMER motif finder. Our engine combines a systolic array design with enhancements to pipeline the complex Viterbi calculation that forms the core of the comparison, and to support coarse-grained parallelism and streaming of multiple sequences within one FPGA. Performance estimates based on a functioning VHDL realisation of our design show a $190\times$ speedup over the same computation in optimised software on a modern general-purpose CPU.

High Performance Database Searching with HMMer on FPGAs

Tim Oliver[1], Leow Yuan Yeow[2] and Bertil Schmidt[3]

[1]Progeniq Pte Ltd
Singapore, Singapore
tim@progeniq.com

[2]Progeniq Pte Ltd
Singapore, Singapore
yuanyeow@progeniq.com

[3]Division of Engineering, Science and Technology
UNSW Asia
Singapore, Singapore
bertil.schmidt@unswasia.edu.sg

Profile Hidden Markov Models (profile HMMs) are used as a popular bioinformatics tool for sensitive database searching, e.g. a set of not annotated protein sequences is compared to a database of profile HMMs to detect functional similarities. HMMer is a commonly used package for profile HMM-based methods. However, searching large databases with HMMer suffers from long runtimes on traditional computer architectures. These runtime requirements are likely to become even more severe due to the rapid growth in size of both sequence and model databases. In this paper, we present a new reconfigurable architecture to accelerate HMMer database searching. It is described how this leads to significant runtime savings on off-the-shelf field-programmable gate arrays (FPGAs).

Exploring the Viability of Cell Broadband Engine for Bioinformatics Applications

Vipin Sachdeva[1], Mike Kistler[1], Evan Speight[1] and Tzy-Hwa Kathy Tzeng[2]

[1]IBM Austin Research Lab
Austin, TX, USA
{vsachde, mkistler, speight}@us.ibm.com

[2]IBM Systems and Technology Group
Poughkeepsie, NY, USA
tzy@us.ibm.com

This paper evaluates the performance of bioinformatics applications on the Cell Broadband Engine recently developed at IBM. In particular we focus on two highly popular bioinformatics applications – Smith-Waterman and ClustalW. The characteristics of these bioinformatics applications, such as small critical time-consuming code size, regular memory accesses, existing vectorized code and embarrassingly parallel computation, make them uniquely suitable for the Cell processing platform. The price and power advantages afforded by the Cell processor also make it an attractive alternative to general purpose processors. We report preliminary performance results for these applications, and contrast these results with the state-of-the-art hardware.

Data-Driven Time-Parallelization in the AFM Simulation of Proteins

L. Ji[1], A. Srinivasan[1], Y. Yu[1] and H. Nymeyer[2]

[1]*Dept. of Computer Science*
Florida State University
Tallahassee, FL, USA
{*ji, asriniva, yu*}*@cs.fsu.edu*

[2]*Dept. of Chemistry and Biochemistry*
Florida State University
Tallahassee, FL, USA
hnymeyer@fsu.edu

Molecular Dynamics is a popular technique to simulate the behavior of physical systems, with resolution at the atomic scale. One of its limitations is that an enormous computational effort is required to simulate to realistic time spans. Conventional parallelization strategies have limited effectiveness in dealing with this difficulty. We recently introduced a more scalable approach to parallelization, where data from prior, related, simulations are used to parallelize a simulation in the time domain. We demonstrated its effectiveness in nano-mechanics simulations. In this paper, we develop our approach so that it can be used in a soft-matter application involving the atomic force microscopy simulation of proteins. We obtain an order of magnitude improvement in performance when we combine time parallelization with conventional parallelization. The significance of this work lies in demonstrating the promise of data-driven time parallelization in soft-matter applications, which are more challenging than the hard-matter applications considered earlier.

RNAVLab: A unified environment for computational RNA structure analysis based on grid computing technology

Michela Taufer[1], Ming-Ying Leung[2,4], Kyle L. Johnson[3] and Abel Licon[1]

[1]*Dept. of Computer Science*
University of Texas at El Paso
El Paso, TX, U.S.A
{*mtaufer, alicon2*}*@utep.edu*

[2]*Dept. of Mathematical Sciences*
University of Texas at El Paso
El Paso, TX, U.S.A
mleung@utep.edu

[3]*Dept. of Biological Sciences*
University of Texas at El Paso
El Paso, TX, U.S.A
kljohnson@utep.edu

[4]*Bioinformatics Program*
University of Texas at El Paso
El Paso, TX, U.S.A

Ribonucleic acid (RNA) molecules play important roles in many biological processes including gene expression and regulation. An RNA molecule is a linear polymer which folds back on itself to form a three dimensional (3D) functional structure. While experimental determination of precise 3D RNA structures is a time consuming and costly process, useful insight into the molecule can be gained from knowing its secondary structure. Structural elements in RNA secondary structures can be separated into two large categories: stem-loops and pseudoknots. The development of mathematical models and computational prediction algorithms for simple stem-loop structures started early in the 1980s. However, building systems that provide the tremendous computer time and memory needed for RNA analysis of both stem-loops and pseudoknots remains a challenge even today. The recently developed grid computing technology can offer a possible solution to this challenge.

In this paper we briefly address mathematical problems associated with the grid computing approach to RNA structure prediction. In particular, we introduce models to partition a large RNA molecule into smaller segments to be assigned to different computers on the grid. Based on these models, we formulate a sampling strategy to select RNA segments for computational prediction to maximize prediction consistency. This strategy is under construction as part of RNAVLab, our unified environment for computational RNA structure analysis, i.e., prediction, alignment, comparison, and classification. A first prototype of RNAVLab is presented and used to investigate the possible association of secondary structure types with RNA functions by analyzing secondary structures for a family of nodavirus genomes.

An Automated Data Processing Pipeline for Virus Structure Determination at High Resolution

Chen Yu[1], Dan C. Marinescu[1], John P. Morrison[2], Brian C. Clayton[2] and David A. Power[2]

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

[2] Computer Science Department
University College Cork
Cork, Cork, Ireland
{j.morrison, b.clayton, d.power}@cs.ucc.ie

The automation of the data processing pipeline for virus structure determination at high resolution is a very challenging problem. The interaction between the data collection process and the theoretical modeling and computer simulation is very complex; routine tasks are mixed with decision making processes and unforeseen conditions. This paper dissects some of the most difficult problems posed by the dynamic coordination of complex computational tasks in a large scale distributed data acquisition and analysis system. A flexible coordination model should be capable of accommodating user actions, handling system related activities such as resource discovery and resource allocation, permitting dynamic process description modification, allowing different level of abstraction, providing some level of fault tolerance and backtracking capability. The Condensed Graphs model of computing developed at University College Cork (UCC) which combines availability-, demand-, and control- driven computation seems to be the most promising for certain classes of problems and complements our previous efforts in developing an intelligent environment for large scale-distributed data acquisition and analysis workflow applications.

Workshop 8
Advances in Parallel and Distributed Computing Models
APDCM 2007

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
Keqiu Li, Dalian Maritime University, China
Weifa Liang, Australian National University, Australia
Ami Marowka Shenkar College of Engineering and Design ,Israel
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

Keynote – Challenges in Large-Scale Internet Search

Tao Yang

Departoment of Computer Science
University of California
Santa Barbara, California, USA
tyang@cs.ucsb.edu

Search engines have become a central mechanism to deal with the exploding volume of information on the Internet. This talk discusses our experiences and computation models involved in developing leading-edge search engine technology at Ask.com. I will present many of challenges faced in processing billions of documents and seeking relevant answers in real time for tens of millions of users everyday using tens of thousands of machines.

Average Execution Time Analysis of a Self-stabilizing Leader Election Algorithm

Juan Paulo Alvarado-Magaña and José Alberto Fernández-Zepeda

Department of Computer Science
CICESE
Ensenada, Baja California, Mexico
{*alvarado, fernan*}*@cicese.mx*

This paper deals with the self-stabilizing leader election algorithm of Xu and Srimani [10] that finds a leader in a tree graph. The worst case execution time for this algorithm is $O(N^4)$, where N is the number of nodes in the tree. We show that the average execution time for this algorithm, assuming two different scenarios, is much lower than $O(N^4)$. In the first scenario, the algorithm assumes a equiprobable daemon and it only privileges a single node at a time. The average execution time for this case is $O(N^2)$. For the second case, the algorithm can privilege multiple nodes at a time. We eliminate the daemon from this algorithm by making random choices to avoid interference between neighbor nodes. The execution time for this case is $O(N)$. We also show that for specific tree graphs, these results reduce even more.

Real-Time Distributed Scheduling of Precedence Graphs on Arbitrary Wide Networks

Franck Butelle, Lucian Finta and Mourad Hakem

LIPN - CNRS UMR 7030
Université Paris Nord
Villetaneuse, France
{*franck.butelle, lucian.finta, mourad.hakem*}*@lipn.univ-paris13.fr*

Previous work on scheduling dynamic competitive jobs is focused on multiprocessors configurations. This paper presents a new distributed dynamic scheduling scheme for sporadic real-time jobs with arbitrary precedence relations on arbitrary wide networks. A job is modeled by a Directed Acyclic Graph (DAG). Jobs arrive on any site at any time and compete for the computational resources of the network. The scheduling algorithm developed in this paper is based upon a new concept of Computing Spheres in order to determine a good neighborhood of sites that may cooperate for the execution of a job if it cannot be guaranteed locally. The salient feature of this new concept is that it allows the algorithm to be performed on arbitrary wide networks since it uses a limited number of sites and communication links.

Novel Broadcast/Multicast Protocols for Dynamic Sensor Networks

Wei Chen[1], A. K. M. Muzahidul Islam[2], Mohan Malkani[1], Amir Shirkhodaie[1], Koichi Wada[2] and Mohamed Zein-Sabatto[1]

[1]*College of Engineerig, Technology and Computer Science*
Tennesee State Unviersity
Nashville, TN, U.S.A
{*wchen, mmalkani, ashirkhodaie, mzein*}*@tnstate.edu*

[2]*Department of Computer and Information Science*
Nagoya Institute of Technology
Nagoya, Aichi, Japan
islam@phaser.elcom.nitech.ac.jp, wada@ nitech.ac.jp

In this paper, we have proposed a time efficient, energy saving and robust broadcast/multicast protocol for reconfigurable cluster-based sensor network. In our broadcast protocol, a broadcast can be executed in $O(hd^2 + D^2)$ rounds and each node needs to be awake in $O(D^2)$ rounds, where D and d are the degrees of G and the sub-network induced by the network backbone, respectively, and h is the height of the backbone. When k channels are available, the broadcast can be executed in $O((hd^2 + D^2)/k)$ rounds and each node needs to be awake in $O(D^2/k)$ rounds. We showed that our broadcast protocol can be readily modified to the one for multicast. The proposed network architecture is self-constructible/reconfigurable with two topological management operations: node-move-in and node-move-out. Details of the protocol along with experimental results are discussed. Simulation results show that the performance of the protocol is much better than that using the theoretical analysis.

Using Coroutines for RPC in Sensor Networks

Marcelo Cohen, Thiago Ponte, Silvana Rossetto and Noemi Rodriguez

Departamento de Informatica
PUC-Rio
Rio de Janeiro, RJ, Brazil
mca@rdc.puc-rio.br, {tponte, silvana, noemi}@inf.puc-rio.br

This paper proposes a concurrency model which integrates the asynchronous and event-driven nature of wireless sensor networks with higher-level abstractions that provide a more familiar programming style for the developer. As a basis for this proposal, we designed and implemented a cooperative multitasking scheduler, based on coroutines, for the TinyOS operating system. We then used this scheduler to implement RPC-like interfaces that capture different communication patterns common in wireless sensor networks. This allows the programmer to work, when appropriate, with a synchronous style, while maintaining an asynchronous model at the message exchange level.

Constant Time Simulation of an R-Mesh on an LR-Mesh

Carlos Alberto Córdova-Flores[1], José Alberto Fernández-Zepeda[1] and Anu G. Bourgeois[2]

[1]*Department of Computer Science/CICESE*
Ensenada, Baja California, Mexico
{cordovaf, fernan}@cicese.mx

[2]*Department of Computer Science*
Georgia State University
Atlanta, Georgia, USA
abourgeois@cs.gsu.edu

Recently, many parallel computing models using dynamically reconfigurable electrical buses have been proposed in the literature. The underlying characteristics are similar among these models, but they do have certain differences that can take form of restrictions on configurations allowed. This paper presents a constant time simulation of an R-Mesh on an LR-Mesh (a restricted model of the R-Mesh), proving that in spite of the differences, the two models possess the same complexity. In other words, the LR-Mesh can simulate a step of the R-Mesh in constant time with a polynomial increase in size. This simulation is based on Reingold's algorithm to solve USTCON in log-space. The simulation is also the first to be executed in constant time.

Scattered Black Hole Search in an Oriented Ring using Tokens

Stefan Dobrev[1], Nicola Santoro[2] and Wei Shi[2]

[1] School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada
sdobrev@site.uottawa.ca

[2] School of Computer Science
Carleton University
Ottawa, Ontario, Canada
{santoro, swei4}@scs.carleton.ca

A *black hole* is a highly harmful host that disposes of visiting agents upon their arrival without any observable trace of the destruction. The problem of locating the black hole in a asynchronous *ring* network is known to be solvable by a team of mobile agents if each node is equipped with a *whiteboard*. A simpler and less expensive inter-communication and synchronization mechanism is provided by *tokens*: each agent has available a bounded number of tokens that can be carried, placed in a node or/and on a port of the node, or removed. All tokens are identical and no other form of communication or coordination is available to the agents.

It is known that locating the black hole in an anonymous ring network using tokens is feasible when the team of agents is initially *co-located* (i.e. they all start from the same host). Recently, the more difficult case when the agents are *scattered* (i.e., when the agents do not start from the same host) has also been examined and solutions requiring only $O(1)$ tokens per agent but using a total of $O(n^2)$ moves have been presented. The number of moves can be reduced to $O(kn + n \log n)$ if the number k of agents is known.

In this paper, we study the impact of orientation and knowledge of team size on the cost of black hole location by scattered agents with tokens. We prove that, in oriented rings, the number of moves can be reduced from $O(n^2)$ to the optimal $\Theta(n \log n)$ using only $O(1)$ tokens per agent, without any knowledge of the team size. This result holds even if both agents and nodes are *anonymous*. Interestingly, the proposed algorithm solves, with the same cost, also the *Leader Election* problem and the *Rendezvous* problem for the scattered agents despite the presence of a black hole.

Cluster-dot Screening by Local Exhaustive Search with Hardware Accelaration

Yasuaki Ito and Koji Nakano

Department of Information Engineering
Hiroshima University
Kagamiyama, Higashi-Hiroshima, JAPAN
{yasuaki, nakano}@cs.hiroshima-u.ac.jp

Screening is an important task to convert a continuous-tone image into a binary image with pure black and white pixels. The main contribution of this paper is to show a new algorithm for cluster-dot screening using the local exhaustive search. Our new algorithm generates 2-cluster, 3-cluster, and 4-cluster binary images, in which all dots have at least 2, 3, and 4 pixels, respectively. The experimental results show that it produces high quality and sharp cluster-dot binary images. We also implemented it on an FPGA to accelerate the computation and achieved a speedup factor of more than 200 over the software implementations.

Implementing Hirschberg's PRAM-Algorithm for Connected Components on a Global Cellular Automaton

Johannes Jendrsczok[1], Rolf Hoffmann[1] and Jörg Keller[2]

[1]*FB Informatik, FG Rechnerarchitektur*
TU Darmstadt
Darmstadt, Germany
{*jendrsczok, hoffmann*}*@ra.informatik.tu-darmstadt.de*

[2]*Fakultät für Mathematik und Informatik*
FernUniversität in Hagen
Hagen, Germany
Joerg.Keller@FernUni-Hagen.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, and it can be implemented on reconfigurable hardware. We discuss the GCA implementation of PRAM algorithms, exemplified by the algorithm of Hirschberg et al., which determines the connected components of a given undirected graph. Insights are that efficient mappings of PRAM algorithms onto GCA exist, and that PRAM and GCA optimality criteria differ because the latter takes memory consumption into account. This makes the GCA a parallel computational model and an implementation platform, thus narrowing the gap between theory and practice.

On Achieving the Shortest-Path Routing in 2-D Meshes

Zhen Jiang[1] and Jie Wu[2]

[1]*Department of Computer Science*
West Chester University
West Chester, Pennsylvania, USA
zjiang@wcupa.edu

[2]*Department of Computer Science and Engineering*
Florida Atlantic University
Boca Raton, Florida, USA
jie@cse.fau.edu

In this paper, we present a fully distributed process to collect and distribute the minimal connected component (MCC) fault information so that the shortest-path between the source and the destination can always be found in the corresponding information-based routing via routing decisions at each intermediate node. Considering the communication cost in the above information distribution, a more practical implementation is provided with only a low number of nodes along the boundary lines involved in the information propagation. The experimental results show the substantial improvement of our approach in terms of the success rate in finding the shortest-path and the average path length.

A Self-Stabilizing Distributed Approximation Algorithm for the Minimum Connected Dominating Set

Sayaka Kamei[1] and Hirotsugu Kakugawa[2]

[1]Dept. of Information Systems
Tottori University of Environmental Studies
Tottori, Tottori, Japan
s-kamei@kankyo-u.ac.jp

[2]Dept. of Computer Science
Osaka University
Toyonaka, Osaka, Japan
kakugawa@ist.osaka-u.ac.jp

Self-stabilization is a theoretical framework of non-masking fault-tolerant distributed algorithms. A self-stabilizing system tolerates any kind and any finite number of transient faults, such as message loss, memory corruption, and topology change. Because such transient faults occur so frequently in mobile ad hoc networks, distributed algorithms on them should tolerate such events. In this paper, we propose a self-stabilizing distributed approximation algorithm for the minimum connected dominating set, which can be used, for example, as a virtual backbone or routing in mobile ad hoc networks. The size of the solution by our algorithm is at most $8|D\_opt| + 1$, where D_opt is a minimum connected dominating set. The time complexity is $O(n^2)$ steps.

A Minimal Access Cost-Based Multimedia Object Replacement Algorithm

Keqiu Li[1], Takashi Nanya[2] and Wenyu Qu[3]

[1]Research Center for Advance Science and Technology
University of Tokyo
Tokyo, Japan
keqiu@hal.rcast.u-tokyo.ac.jp

[2]Research Center for Advance Science and Technology
University of Tokyo
Tokyo, Japan
nanya@hal.rcast.u-tokyo.ac.jp

[3]College of Computer Science and Technology
Dalian Maritime University
Dalian, Liaoning, China
eunice_01@163.com

Multimedia object caching, by which the same multimedia object can be adapted to diverse mobile appliances through the technique of transcoding, is an important technology for improving the scalability of web services, especially in the environment of mobile networks. In this paper, we address the cache replacement problem for multimedia object caching by exploring the aggregate effect of caching multiple versions of the same multimedia object. First, We present an optimal solution for calculating the minimal access cost of caching multiple versions of the same multimedia object. Second, based on this solution, we propose an effective cache replacement algorithm for multimedia object caching. Finally, we evaluate the performance of the proposed solution with a set of simulation experiments for various performance metrics over a wide range of system parameters.

Revisiting Matrix Product on Master-Worker Platforms

Jack Dongarra[1], Jean-François Pineau[2], Yves Robert[2], Zhiao Shi[1] and Frédéric Vivien[2]

[1]*Innovative Computing Laboratory, Department of Computer Science*
University of Tennessee
Knoxville, TN, USA
{*dongarra, shi*}*@cs.utk.edu*

[2]*LIP, CNRS-ENS Lyon-INRIA-UCBL*
École normale supérieure de Lyon
Lyon, France
{*jean-francois.pineau, jean-francois.pineau, frederic.vivien*}*@ens-lyon.fr*

This paper is aimed at designing efficient parallel matrix-product algorithms for homogeneous master-worker platforms. While matrix-product is well-understood for *homogeneous 2D-arrays of processors* (e.g., Cannon algorithm and ScaLAPACK outer product algorithm), there are two key hypotheses that render our work original and innovative:

- Centralized data. We assume that all matrix files originate from, and must be returned to, the master. The master distributes both data and computations to the workers (while in ScaLAPACK, input and output matrices are initially distributed among participating resources). Typically, our approach is useful in the context of speeding up MATLAB or SCILAB clients running on a server (which acts as the master and initial repository of files).

- Limited memory. Because we investigate the parallelization of large problems, we cannot assume that full matrix panels can be stored in the worker memories and re-used for subsequent updates (as in ScaLAPACK). The amount of memory available in each worker is expressed as a given number of buffers, where a buffer can store a square block of matrix elements. These square blocks are chosen so as to harness the power of Level 3 BLAS routines; they are of size 80 or 100 on most platforms.

We have devised efficient algorithms for resource selection (deciding which workers to enroll) and communication ordering (both for input and result messages), and we report a set of MPI experiments conducted on a platform at the University of Tennessee.

A Configuration Control Mechanism Based on Concurrency Level for a Reconfigurable Consistency Algorithm

Christiane V. Pousa[1], Luís F. W. Góes[2] and Carlos A. P. S. Martins[3]

[1]Computational and Digital Systems Group
Pontifical Catholic University of Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
pousa@ieee.org

[2]Computational and Digital Systems Group
Pontifical Catholic University of Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
lfwgoes@yahoo.com.br

[3]Computational and Digital Systems Group
Pontifical Catholic University of Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
capsm@ieee.org

A Reconfigurable Consistency Algorithm (RCA) is an algorithm that guarantees the consistency in Distributed Shared Memory (DSM) Systems. This algorithm modifies its behavior (re-configuration) according to the changes in the workload and DSM system. In a RCA, there is a Configuration Control Layer (CCL) that is responsible for selecting the most suitable RCA configuration (behavior) for a specific workload and DSM system. In previous works, we defined an upper bound performance for RCA based on an ideal CCL, which knows apriori the best configuration for each situation. This ideal CCL is based on a set of workloads characteristics that, in most situations, are difficult to extract from the applications (percentage of shared write and read operations and sharing patterns). In this paper we propose, develop and present a heuristical configuration control mechanism for the CCL implementation. This mechanism is based on an easily obtained applications parameter, the concurrency level. Our results show that this configuration control mechanism improves the RCA performance in 15%, on average, compared to other traditional consistency algorithms. Furthermore, the CCL with this mechanism is independent from the workload and DSM system specific characteristics, like sharing patterns and percentage of writes and reads.

Pipelining Tradeoffs of Massively Parallel SuperCISC Hardware Functions

Colin J. Ihrig, Justin Stander and Alex K. Jones

Electrical and Computer Engineering
University of Pittsburgh
Pittsburgh, PA, USA
{*cihrig, jstander, akjones*}*@engr.pitt.edu*

Parallel processing using multiple processors is a well-established technique to accelerate many different classes of applications. However, as the density of chips increases, another technique to accelerate these applications is the use of application specific hardware processing blocks in parallel within a chip. SuperCISC hardware blocks utilize this method to accelerate scientific, signal, and image processing applications. By applying pipelining methodologies to SuperCISC functions, the effective amount of parallelism already present can be further increased. Automated register placement within a combinational data flow graph (DFG) is governed by the desired maximum operating frequency provided as a parameter to the tool flow, as well as the results of static timing analysis of the circuit. Results presented include the design tradeoffs between increased performance, area, and energy. Additionally, benefits of pipelining compared to hardware replication as a means of achieving further parallelism is studied.

On the Power of the Multiple Associative Computing (MASC) Model Related to That of Reconfigurable Bus-Based Models

Mingxian Jin[1] and Johnnie W. Baker[2]

[1]*Department of Mathematics and Computer Science*
Fayetteville State University
Fayetteville, North Carolina, USA
mjin@uncfsu.edu

[2]*Department of Computer Science*
Kent State University
Kent, Ohio, USA
jbaker@cs.kent.edu

The MASC model is a multi-SIMD model that uses control parallelism to coordinate the interaction of data parallel threads. It supports a generalized associative style of parallel computation. The power of this model has been compared to that of priority CRCW PRAM and enhanced meshes. In this paper, we present the work on simulations between MASC and reconfigurable bus-based models, in particular, different versions of the Reconfigurable Multiple Bus Machine (RMBM). It is shown that MASC and the Basic RMBM (B-RMBM) can simulate each other in constant time if the number of buses on the B-RMBM is Θ(j) where j is the number of MASC instruction streams. Thus, when these two models satisfy the preceding condition, they have the same power. Simulations of other stronger versions of RMBM using MASC are also considered. Since the RMBM model has been shown to be as powerful as a general Reconfigurable Mesh (RM), our simulations can be used to establish a relationship between MASC and RM. As RM has been widely accepted as an extremely powerful model, our work gives a better understanding of the MASC model and provides useful information concerning the power of this model.

Linking Compilation and Visualization for Massively Parallel Programs

Alex K. Jones[1,2], Raymond R. Hoare[3], Joseph St. Onge[3], Joshua M. Lucas[4], Shuyi Shao[2] and Rami Melhem[2]

[1]*Electrical and Computer Engineering*
University of Pittsburgh
Pittsburgh, PA, USA
akjones@ece.pitt.edu

[2]*Computer Science*
University of Pittsburgh
Pittsburgh, PA, USA
{*syshao, melhem*}*@cs.pitt.edu*

[3]*Concurrent EDA, LLC*
Pittsburgh, PA, USA
{*rayhoare, joseph.st.onge*}*@concurrenteda.com*

[4]*Lockheed Martin*
Akron, OH, USA
jmlst79@gmail.com

This paper presents a technique to visualize the communication pattern of a parallel application at different points during its execution. Unlike many existing tools that show the communication pattern for the entire application, our tool breaks this communication pattern down into components to allow the more detailed study of application execution. These patterns are not merely snapshots or windows of the execution but rather are tied to specific code structures comprised of loops in the application. Our technique leverages our compiler, which adds instructions into the code to record where communications and code artifacts occur during execution. This information is stored into a trace format, which is read by our visualization tool. The visualization tool can graphically represent the communication pattern and message volume to allow a user to analyze and optimize the execution. As an example, we show how this information can be used to optimize the execution time and reduce the message delay of applications executed on a system enhanced with optical circuit switch interconnections.

A Prototype Multithreaded Associative SIMD Processor

Kevin Schaffer and Robert A. Walker

Computer Science Department
Kent State University
Kent, OH, USA
{*kschaffe, walker*}*@cs.kent.edu*

The performance of SIMD processors is often limited by the time it takes to transfer data between the central-ized control unit and the parallel processor array. This is especially true of hybrid SIMD models, such as associa-tive computing, that make extensive use of global search operations. Pipelining instruction broadcast can help, but is not enough to solve the problem, especially for mas-sively parallel processors with thousands of processing elements. In this paper, we describe a SIMD processor architecture that combines a fully pipelined broad-cast/reduction network with hardware multithreading to reduce performance degradation as the number of proc-essors is scaled up.

Workshop 9
Communication Architecture for Clusters
CAC 2007

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:

- 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)
- communication and architectural issues related to router/switch organization, flow control, congestion control, routing and deadlock handling, load balancing, reliability, QoS support, topology discovery, dynamic reconfiguration, and clustered storage and fileservers
- novel network-interface and switch architectures for supporting efficient point-to-point and collective communication
- performance measurements, analysis, and early-user experience reports of pre-release and recently released network interfaces, network protocols/emerging standards, and routers/switches

Co-Chairs:

Scott Pakin (LANL)
Craig Stunkel (IBM)
Pankaj Mehra (HP)

Program Committee:

Gheorghe Almasi (IBM)
Angelos Bilas (FORTH & U. Crete)
Ron Brightwell (SNL)
Darius Buntinas (ANL)
Wu-Chun Feng (VT)
José Flich (UPV)
Dave Garcia (HP)
Mitchell Gusat (IBM)
Nectarios Koziris (NTUA)
Ben Lee (OSU)
Andrew Lumsdaine (IU)
Jarek Nieplocha (PNNL)
Greg Pfister (IBM)
Jamie Riotto (Cisco)
Vikram Saletore (Intel)
Cris Simpson (Intel)
Evan Speight (IBM)
Pete Wyckoff (OSC)

Publicity Coordinator

Nectarios Koziris (NTUA)

Efficient Switches with QoS Support for Clusters

Alejandro Martínez[1], Francisco J. Alfaro[1], José L. Sánchez[1] and José Duato[2]

[1]*Computing Systems Department*
Univ. of Castilla-La Mancha
Albacete, Spain
{*alejandro, falfaro, jsanchez*}*@dsi.uclm.es*

[2]*Dept. of Systems Data Processing and Computers*
Tech. Univ. of Valencia
Valencia, Spain
jduato@disca.upv.es

Clusters of PCs provide service to thousands or tens of thousands of concurrent users. Many clusters are used for multimedia applications, which usually present quality of service (QoS) requirements.

Several cluster switches with QoS support have been proposed. All of them incorporate VCs in order to provide QoS support, devoting a different VC to each traffic class. This increases the switch complexity and required silicon area.

In previous work we have proposed a strategy to use just two VCs at each switch port for the provision of QoS, emulating many more VCs. The idea consists in having a strict priority between traffic classes, such as the end-nodes would always inject packets with high priority before packets with low priority. This order of injection could be reused at the switches, producing an efficient design with good performance.

In this paper, we review this proposal and cover its weaknesses. We present a switch design that offers complete QoS support using just two VCs. It does not require strict priority among traffic classes and allows to provide latency guarantees. Moreover, it is based in the table schedulers present both in the specifications of InfiniBand and PCI AS and, thus, can be implemented seamlessly in those network architectures.

Comparing the latency performance of the DTable and DRR schedulers

Raúl Martínez, Francisco J. Alfaro and José L. Sánchez

Computing Systems Department
University of Castilla-La Mancha
Albacete, Spain
{*raulmm, falfaro, jsanchez*}*@dsi.uclm.es*

A key component for networks with Quality of Service (QoS) support is the egress link scheduling algorithm. An ideal scheduling algorithm implemented in a high performance network with QoS support should satisfy two main properties: good end-to-end delay and implementation simplicity. The Deficit Round Robin (DRR) algorithm is known to have a very little implementation complexity. However, depending on the situation, its latency performance can be very bad.

On the other hand, table-based schedulers try to offer a simple implementation and good latency bounds. Some of the latest proposals of network technologies, like Advanced Switching and InfiniBand, include in their specifications one of these schedulers. However, these table-based schedulers do not work properly with variable packet sizes and face the problem of bounding the bandwidth and latency assignments. We have proposed a new table-based scheduler, which we have called Deficit Table (DTable) scheduler, that works properly with variable packet sizes. Moreover, we have proposed a methodology to configure this table-based scheduler to decouple the bounding of bandwidth and latency assignments.

In this paper, we review these proposals and present simulation results that show that the DTable scheduler is able to provide a better latency performance than the DRR scheduler, with only a slightly higher implementaion and computational complexity.

A practically constant-time MPI Broadcast Algorithm for large-scale InfiniBand Clusters with Multicast

Torsten Hoefler[1,2], Christian Siebert[1] and Wolfgang Rehm[1]

[1]*Dept. of Computer Science*
Chemnitz University of Technology
Chemnitz, Saxony, Germany
{*htor, chsi, rehm*}*@cs.tu-chemnitz.de*

[2]*Open Systems Laboratory*
Indiana University
Bloomington, Indiana, USA

An efficient implementation of the MPI_BCAST operation is crucial for many parallel scientific applications. The hardware multicast operation seems to be applicable to switch-based InfiniBand cluster systems. Several approaches have been implemented so far, however there has been no production-ready code available yet. This makes optimal algorithms to a subject of active research. Some problems still need to be solved in order to bridge the semantic gap between the unreliable multicast and MPI_BCAST. The biggest of those problems is to ensure the reliable data transmission in a scalable way. Acknowledgement-based methods that scale logarithmically with the number of participating MPI processes exist, but they do not meet the supernormal demand of high-performance computing. We propose a new algorithm that performs the MPI_BCAST operation in a practically constant time, independent of the communicator size. This method is well suited for large communicators and (especially) small messages due to its good scaling and its ability to prevent parallel process skew. We implemented our algorithm as a collective component for the Open MPI framework using native InfiniBand multicast and we show its scalability on a cluster with 116 compute nodes, where it saves up to 41% MPI_BCAST latency in comparison to the "TUNED" Open MPI collective.

NewMadeleine: a Fast Communication Scheduling Engine for High Performance Networks

Olivier Aumage, Elisabeth Brunet, Nathalie Furmento and Raymond Namyst

Inria, LaBRI
Université Bordeaux 1
Talence, France
{*aumage, brunet, furmento, namyst*}*@labri.fr*

Communication libraries have dramatically made progress over the fifteen years, pushed by the success of cluster architectures as the preferred platform for high performance distributed computing. However, many potential optimizations are left unexplored in the process of mapping application communication requests onto low level network commands. The fundamental cause of this situation is that the design of communication subsystems is mostly focused on reducing the latency by shortening the critical path. In this paper, we present a new communication scheduling engine which dynamically optimizes application requests in accordance with the NICs capabilities and activity. The optimizing code is generic and portable. The database of optimizing strategies may be dynamically extended.

RI2N/UDP: High bandwidth and fault-tolerant network for a PC-cluster based on multi-link Ethernet

Takayuki Okamoto[1], Shin'ichi Miura[2], Taisuke Boku[3], Mitsuhisa Sato[4] and Daisuke Takahashi[5]

[1]*Graduate School of Systems and Information Engineering*
University of Tsukuba
Tsukuba, Ibaraki, Japan
okamoto@hpcs.cs.tsukuba.ac.jp

[2]*Graduate School of Systems and Information Engineering*
University of Tsukuba
Tsukuba, Ibaraki, Japan
miura@hpcs.cs.tsukuba.ac.jp

[3]*Graduate School of Systems and Information Engineering*
University of Tsukuba
Tsukuba, Ibaraki, Japan
taisuke@hpcs.cs.tsukuba.ac.jp

[4]*Graduate School of Systems and Information Engineering*
University of Tsukuba
Tsukuba, Ibaraki, Japan
msato@hpcs.cs.tsukuba.ac.jp

[5]*Graduate School of Systems and Information Engineering*
University of Tsukuba
Tsukuba, Ibaraki, Japan
daisuke@hpcs.cs.tsukuba.ac.jp

PC-clusters with high performance/cost ratio have been one of the typical platforms for high performance computing. To lower costs, Gigabit Ethernet is often used for intercommunication networks. However, the reliability of Ethernet is limited due to hardware failures and tentative errors in the network switches. To solve this problem, we propose an interconnection network system based on multi-link Ethernet named RI2N. In this paper, we developed a user level implementation of RI2N using UDP/IP that is called RI2N/UDP. When this new system was evaluated for performance and fault tolerance, the bandwidth on a 2-link Gigabit Ethernet was 246 MB/s, and the system could remain active during network link failure to provide high system reliability.

Evaluation of Remote Memory Access Communication on the Cray XT3

Vinod Tipparaju[1], Andriy Kot[1], Jarek Nieplocha[1], Monika ten Bruggencate[2] and Nikos Chrisochoides[3]

[1]*Pacific Northwest National Laboratory*
Richland, WA, USA
{*vinod, andriy.kot, jarek.nieplocha*}*@pnl.gov*

[2]*Cray, Inc.*
Albuquerque, NM, USA
monikatb@cray.com

[3]*Computer Science Department*
The College of William and Mary
Williamsburg, VA, USA
nikos@cs.wm.edu

This paper evaluates remote memory access (RMA) communication capabilities and performance on the Cray XT3. We discuss properties of the network hardware and Portals networking software layer and corresponding implementation issues for SHMEM and ARMCI portable RMA interfaces. The performance of these interfaces is studied and compared to MPI performance.

Designing Efficient Asynchronous Memory Operations Using Hardware Copy Engine: A Case Study with I/OAT

Karthikeyan Vaidyanathan, Wei Huang, Lei Chai and Dhabaleswar K. Panda

Department of Computer Science and Engineering
The Ohio State University
Columbus, OH, USA
{*vaidyana, haunwei, chail, panda*}*@cse.ohio-state.edu*

Memory copies for bulk data transport incur large overheads due to CPU stalling, small register-size data movement, etc. Intel's I/O Acceleration Technology offers an asynchronous memory copy engine in kernel space which alleviates such overheads. In this paper, we propose a set of designs for asynchronous memory operations in user space for both single process (as an offloaded *memcpy()*) and IPC using the copy engine. We analyze our design based on overlap efficiency, performance and cache utilization. Our microbenchmark results show that using the copy engine for performing memory copies can achieve close to 87% overlap with computation. Further, the copy engine improves the copy latency of bulk memory data transfers by 50% and avoids cache pollution effects. With the emergence of multi-core architectures, the support for asynchronous memory operations holds a lot of promise in reducing the gap between the memory and processor performance.

10-Gigabit iWARP Ethernet: Comparative Performance Analysis with InfiniBand and Myrinet-10G

Mohammad J. Rashti and Ahmad Afsahi

Department of Electrical and Computer Engineering
Queen's University
Kingston, ON, CANADA
mohammad.rashti@ece.queensu.ca, ahmad.afsahi@queensu.ca

iWARP is a set of standards enabling Remote Direct Memory Access (RDMA) over Ethernet. iWARP supporting RDMA and OS bypass, coupled with TCP/IP Offload Engines, can fully eliminate the host CPU involvement in an Ethernet environment. With the iWARP standard and the introduction of 10-Gigabit Ethernet, there is now an alternative path to the proprietary interconnects for high-performance computing, while maintaining compatibility with existing Ethernet infrastructure and protocols.

Recently, NetEffect Inc. has introduced an iWARP enabled 10-Gigabit Ethernet Channel Adapter. In this paper we assess the potential of such an interconnect for high-performance computing by comparing its performance with two leading cluster interconnects, InfiniBand and Myrinet-10G. The results show that the NetEffect iWARP implementation achieves an unprecedented latency for Ethernet, and saturates 87% of the available bandwidth. It also scales better with multiple connections. At the MPI level, iWARP performs better than InfiniBand in queue usage and buffer re-use.

Implementing the Advanced Switching Fabric Discovery Process

Antonio Robles-Gómez, Aurelio Bermúdez, Rafael Casado and Francisco J. Quiles

Instituto de Investigación en Informática (I3A)
Universidad de Castilla-La Mancha
Albacete, Spain
{*arobles, abermu, rcasado, paco*}*@dsi.uclm.es*

Advanced Switching is a new high-speed industrial standard serial interconnect. It is defined as a switching fabric architecture based on the PCI Express technology. The Advanced Switching specification establishes a management infrastructure which maintains the fabric operation. The topology discovery process is triggered after fabric initialization and every time a topological change is detected. The information gathered by this process is used to build a set of paths between fabric endpoints. This work analyzes the performance of several possible implementations for this management task.

Deterministic versus Adaptive Routing in Fat-Trees

C. Gomez, F. Gilabert, M. E. Gomez, P. Lopez and J. Duato

Grupo de Arquitecturas Paralelas
Universidad Politecnica de Valencia
Valencia, Spain
{*crigore, fragivil*}*@gap.upv.es,* {*megomez, plopez, jduato*}*@disca.upv.es*

Clusters of PCs have become very popular to build high performance computers. These machines use commodity PCs linked by a high speed interconnect. Routing is one of the most important design issues of interconnection networks. Adaptive routing usually better balances network traffic, thus allowing the network to obtain a higher throughput. However, adaptive routing introduces out-of-order packet delivery, which is unacceptable for some applications. Concerning topology, most of the commercially available interconnects are based on fat-tree. Fat-trees offer a rich connectivity among nodes, making possible to obtain paths between all source-destination pairs that do not share any link. We exploit this idea to propose a deterministic routing algorithm for fat-trees, comparing it with adaptive routing in several workloads. The results show that deterministic routing can achieve a similar, and in some scenarios higher, level of performance than adaptive routing, while providing in-order packet delivery.

Workshop 10
NSF Next Generation Software Program
NSFNGS 2007

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

ParalleX: A Study of A New Parallel Computation Model

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
Newark, DE, USA
{ggao, weirong}@capsl.udel.edu

[2]Center for Advanced Computing Research
California Institute of Technology
Pasadena, CA, USA
tron@cct.lsu.edu

[3]Department of Computer Science
Louisiana State University
Baton Rouge, LA, USA

[4]Mathematics and Computer Science Division
Argonne National Laboratory
Argonne, IL, USA
{stevens, hereld}@mcs.anl.gov

This paper proposes the study of a new computation model that attempts to address 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. In this paper, we present the progress of our research on a parallel programming and execution model - mainly, ParalleX. We describe the functional elements of ParalleX, one such model being explored as part of this project. We also report our progress on the development and study of a subset of ParalleX - the LITL-X at University of Delaware. We then present a novel architecture model - Gilgamesh II - as a ParalleX processing architecture. A design point study of Gilgamesh II and the architecture concept strategy are presented.

Improving MPI Independent Write Performance Using A Two-Stage Write-Behind Buffering Method

Wei-Keng Liao[1], Avery Ching[1], Kenin Coloma[1], Alok Choudhary[1] and Mahmut Kandemir[2]

[1]Dept. of Electrical Engineering and Computer Science
Northwestern University
Evanston, Illinois, United States
{wkliao, aching, kcoloma,
choudhar}@ece.northwestern.edu

[2]Dept. of Computer Science and Engineering
Pennsylvania State University
University Park, Pennsylvania, United States
kandemir@cse.psu.edu

Many large-scale production applications often have very long executions times and require periodic data checkpoints in order to save the state of the computation for program restart and/or tracing application progress. These write-only operations often dominate the overall application runtime, which makes them a good optimization target. Existing approaches for write-behind data buffering at the MPI I/O level have been proposed, but challenges still exist for addressing system-level I/O issues. We propose a two-stage write-behind buffering scheme for handing checkpoint operations. The first-stage of buffering accumulates write data for better network utilization and the second-stage of buffering enables the alignment for the write requests to the file stripe boundaries. Aligned I/O requests avoid file lock contention that can seriously degrade I/O performance. We present our performance evaluation using BTIO benchmarks on both GPFS and Lustre file systems. With the two-stage buffering, the performance of BTIO through MPI independent I/O is significantly improved and even surpasses that of collective I/O.

Automatic MPI application transformation with ASPhALT

Anthony Danalis, Lori Pollock and Martin Swany

Computer and Information Sciences
University of Delaware
Newark, DE, USA
{*danalis, pollock, swany*}*@cis.udel.edu*

This paper describes a source to source compilation tool for optimizing MPI-based parallel applications. This tool is able to automatically apply a "prepushing" transformation that causes MPI programs to aggressively send data as soon as it is available, thus improving communication-computation overlap and improving application performance.

In this paper we present *asphalt_transformer*; the Open64-based component of our framework, *ASPhALT*, responsible for automatically performing the prepushing transformation. We also present an extensive study of the performance gains witnessed from automatically transformed codes. In particular, we demonstrate how different levels of aggregation affect the performance of parallel programs executing various computation kernels on different clusters. Furthermore, we discuss the differences in performance improvement between the hand-optimized and automatically optimized codes, as well as the effect of automation on time-to-solution.

Formal Analysis for Debugging and Performance Optimization of MPI

Ganesh L. Gopalakrishnan and Robert M. Kirby

School of Computing
University of Utah
Salt Lake City, UT, USA
{*ganesh, kirby*}*@cs.utah.edu*

High-end computing is universally recognized to be a strategic tool for leadership in science and technology. A significant portion of high-end computing is conducted on clusters running the Message Passing Interface (MPI) library. MPI has become a de facto standard in HPC. MPI programs, as well as MPI library implementations can be buggy, especially when aiming high performance, and running on or porting onto new platforms. Our recent work has addressed the following areas: A TLA+ Formal Semantics of a large subset of MPI-1; A Microsoft Phoenix based Model Extraction and Analysis Framework for MPI programs; Integration into the Visual Studio Environment for error-trace visualization; A new dynamic partial order reduction algorithm (DPOR) tailored to MPI so that the number of interleavings examined during MPI program verification are dramatically reduced; A program called 'inspector' for Analyzing C++ Programs that has found bugs in publicly distributed threaded programs (Inspector automatically instruments PThread programs and searches for races based on a new DPOR); Verified Byte-range Locking Protocols using MPI one-sided Communication - a case study where we found bugs in published byte-range locking protocols, and designed and verified improved versions of these protocols; A New In-situ Model Checker for MPI programs, that traps MPI calls using its profiling interface (PMPI) and orchestrates control to maximize coverage with minimal state saving overhead. The progress made in exploring these directions, our publications, and associated software tools are described, as are our future plans.

Automatic Parallelization of Scripting Languages: Toward Transparent Desktop Parallel Computing

Xiaosong Ma[1,2], Jiangtian Li[1,2] and Nagiza Samatova[1,2]

[1]*Department of Computer Science*
NC State University
Raleigh, NC, USA
{*ma, jli3*}*@csc.ncsu.edu, samatovan@ornl.gov*

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

Desktop computing remains indispensable in scientific exploration, largely because it provides people with devices for human interaction and environments for interactive job execution. However, with today's rapidly growing data volume and task complexity, it is increasingly hard for individual workstations to meet the demands of interactive scientific data processing. The increasing cost of such interactive processing is hindering the productivity of end-to-end scientific computing workflows. While existing distributed computing systems allow people to aggregate desktop workstation resources for parallel computing, the burden of explicit parallel programming and parallel job execution often prohibits scientists to take advantage of such platforms. In this paper, we discuss the need for transparent desktop parallel computing in scientific data processing. As an initial step toward this goal, we present our on-going work on the automatic parallelization of the scripting language R, a popular tool for statistical computing. Our preliminary results suggest that a reasonable speedup can be achieved on real-world sequential R programs without requiring any code modification.

Virtual Execution Environments: Support and Tools

Apala Guha[1], Jason D. Hiser[1], Naveen Kumar[2], Jing Yang[1], Min Zhao[2], Shukang Zhou[1], Bruce R. Childers[2], Jack W. Davidson[1], Kim Hazelwood[1] and Mary Lou Soffa[1]

[1]*Department of Computer Science*
University of Virginia
Charlottesville, VA, USA
{*ag2dx, jdh8d, jy8y, sz9g, jwd, hazelwood, soffa*}*@virginia.edu*

[2]*Department of Computer Science*
University of Pittsburgh
Pittsburgh, PA, USA
{*kumar, zhao, childers*}*@cs.pitt.edu*

In today's dynamic computing environments, the available resources and even underlying computation engine can change during the execution of a program. Additionally, current trends in software development favor the flexibility and cost-effectiveness of dynamically loaded components and libraries. Because of these trends, there has been increased research interest in virtual execution environments (VEEs) for delivering adaptable software suitable for today's rapidly changing, heterogeneous computing environments. In this project, we have been investigating tools and techniques to support implementation of VEEs using software dynamic translation (SDT). This paper highlights some of our recent results. One significant result is that we have developed novel translation techniques that reduce the memory and runtime overhead of SDT to negligible levels. We have also developed innovative debugging and instrumentation tools for SDT-based software environments. Together, these results make SDT-based systems viable for solving a wide range of pressing problems. The paper concludes with a discussion of how SDT may offer a solution to one such problem–inherent process variation in emerging chip multiprocessors.

Intelligent Optimization of Parallel and Distributed Applications

Bhupesh Bansal[2], Jacqueline Chame[1], Ewa Deelman[1], Yolanda Gil[1], Mary Hall[1], Vijay Kumar[3], Kristina Lerman[1], Aiichiro Nakano[2], Yoon-Ju Lee Nelson[1] and Joel Saltz[3]

[1]*Information Sciences Institute*
University of Southern California
Marina del Rey, CA, USA
{*jchame, deelman, gil, mhall, lerman, yoonju*}*@isi.edu*

[2]*Department of Physics and Astronomy*
University of Southern California
Los Angeles, CA, USA
{*bansal, anakano*}*@usc.edu*

[3]*Department of Biomedical Informatics*
The Ohio State University
Columbus, OH, USA
{*vijayskumar, saltz*}*@bmi.osu.edu*

This paper describes a new project that systematically addresses the enormous complexity of mapping applications to current and future parallel platforms. By integrating the system layers – domain-specific environment, application program, compiler, run-time environment, performance models and simulation, and workflow manager – and through a systematic strategy for application mapping, our approach will exploit the vast machine resources available in such parallel platforms to dramatically increase the productivity of application programmers. This project brings together computer scientists in the areas represented by the system layers (i.e., language extensions, compilers, run-time systems, workflows) together with expertise in knowledge representation and machine learning. With expert domain scientists in molecular dynamics (MD) simulation, we are developing our approach in the context of a specific application class which already targets environments consisting of several hundreds of processors. In this way, we gain valuable insight into a generalizable strategy, while simultaneously producing performance benefits for existing and important applications.

Scheduling Issues in Optimistic Parallelization

Milind Kulkarni and Keshav Pingali

Computer Science
University of Texas at Austin
Austin, TX, United States
{*milind, pingali*}*@cs.utexas.edu*

Irregular applications, which rely on pointer-based data structures, are often difficult to parallelize. The input-dependent nature of their execution means that traditional parallelization techniques are unable to exploit any latent parallelism in these algorithms. Instead, we turn to optimistic parallelism, where regions of code are speculatively run in parallel while runtime mechanisms ensure proper execution. The performance of such optimistically parallelized algorithms is often dependent on the schedule for parallel execution; improper choices can prevent successful parallel execution.

We demonstrate this through the motivating example of Delaunay mesh refinement, an irregular algorithm, which we have parallelized optimistically using the Galois system. We apply several scheduling policies to this algorithm and investigate their performance, showing that careful consideration of scheduling is necessary to maximize parallel performance.

New Results on the Performance Effects of Autocorrelated Flows in Systems

Evgenia Smirni[1], Qi Zhang[1], Ningfang Mi[1], Alma Riska[2] and Giuliano Casale[1]

[1]*Department of Computer Science*
College of William and Mary
Williamsburg, VA, USA
{*esmirni, qizhang, ningfang, casale*}*@cs.wm.edu*

[2]*Interfaces and Architecture*
Seagate Research
Pittsburgh, PA, USA
alma.riska@seagate.com

Temporal dependence within the workload of any computing or networking system has been widely recognized as a significant factor affecting performance. More specifically, burstiness, as a form of temporal dependency, is catastrophic for performance. We use the autocorrelation function in a workload flow to formalize burstiness and also to characterize temporal dependence within a flow. We present results from two application areas: load balancing in a homogeneous cluster environment and capacity planning in a multi-tiered e-commerce system. For the load balancing problem, we show that if autocorrelation exists in the arrival stream to the cluster, classic load balancing policies become ineffective and solutions that focus on "unbalancing" the load offer superior performance. For the case of multi-tiered systems, we show that if there is autocorrelation in the flows, we observe the surprising result that in spite of the fact that the bottleneck resource in the system is far from saturation and that the measured throughput and utilizations of other resources are also modest, user response times are very high. For multi-tired systems, this underutilization of resources falsely indicates that the system can sustain higher capacities. We present analysis of the above phenomena that aims at the development of better scheduling policies under autocorrelated flows.

The Adaptive Code Kitchen: Flexible Tools for Dynamic Application Composition

Pilsung Kang[1], Mike Heffner[1], Joy Mukherjee[1], Naren Ramakrishnan[1], Srinidhi Varadarajan[1], Calvin J. Ribbens[1] and Danesh K. Tafti[2]

[1]*Department of Computer Science*
Virginia Tech
Blacksburg, VA, 24061
{*kangp, naren, srinidhi, ribbens*}*@cs.vt.edu,*
mikeh@fesnel.com, jmukherj@vt.edu

[2]*Department of Mechanical Engineering*
Virginia Tech
Blacksburg, VA, 24061
dtafti@vt.edu

Driven by the increasing componentization of scientific codes, the deployment of high-end system infrastructures such as the Grid, and the desire to support high level problem solving primitives, application composition systems have become prevalent in computational science practice. We present the *adaptive code kitchen* which, as the name connotes, is a loose collection of capabilities to help realize complex adaptive composition scenarios. These include function interception, continuation modification, dynamic process checkpointing and rollback, and runtime recommendation. Using these broad primitives, a computational scientist can specify many 'recipes' of adaptivity as complete control systems around native object codes. Runtime systems support then enables loading and linking of native code components, monitoring of performance indicators, consulting a recommender system for algorithmic decisions, and dynamically updating application components in response to the recommendations. We present the architecture of the adaptive code kitchen and the key enabling technologies with brief mention of the applications that will be investigated henceforth during the course of the project.

DOSA: Design Optimizer for Scientific Applications

David A. Bader[1] and Viktor K. Prasanna[2]

[1] College of Computing
Georgia Institute of Technology
Atlanta, GA, USA
bader@cc.gatech.edu

[2] Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
prasanna@ganges.usc.edu

In this work, we propose an application composition system (ACS) that allows design-time exploration and automatic run-time optimizations so that we relieve application programmers and compiler writers from the challenging task of optimizing the computation in order to achieve high performance. Our new framework, called "Design Optimizer for Scientific Applications" (DOSA), allows the programmer or compiler writer to explore alternative designs and optimize for speed (or power) at design-time and use its run-time optimizer as an automatic ACS. The ACS constructs an efficient application that dynamically adapts to changes in the underlying execution environment based on the kernel model, architecture, system features, available resources, and performance feedback. The run-time system is a portable interface that enables dynamic application optimization by interfacing with the output of DOSA. It thus provides an application composition system that determines suitable components and performs continuous performance optimizations. We focus on utilizing advanced architectural features and memory-centric optimizations that reduce the I/O complexity, cache pollution, and processor-memory traffic, in order to achieve high performance. The design-time effort uses a computer-aided design space exploration that provides a user-friendly graphical modeling environment, high-level performance estimation and profiling, and the ability to integrate low-level simulators suitable for HPC architectures.

The TMO Scheme for Wide-Area Distributed Real-Time Computing

K. H. Kim and Stephen F. Jenks

Electrical Engineering and Computer Science
University of California, Irvine
Irvine, CA, USA
{khkim, sjenks}@uci.edu

Some parts of our recent efforts on establishing a technical foundation for wide-area distributed real-time computing (DRC) and distributed time-triggered simulation (DTS) are briefly reviewed. The basic building-block of our technology framework is the Time-triggered Message-triggered Object (TMO) specification and programming scheme. The TMO scheme for local-area DRC has been established in a sound form and its practicality and attractiveness have been extensively demonstrated. However, its extension to fit into wide-area-network based DRC is in an early stage. The distributed time-triggered simulation (DTS) scheme is a new type of an approach to real-time simulation based on parallel distributed computing. The TMO scheme facilitates DTS in efficient forms. Recent developments in TMO-structured wide-area DRC and DTS and the supporting tools are briefly reviewed.

Adaptive Scheduling with Parallelism Feedback

Kunal Agrawal[1], Yuxiong He[2], Wen-Jing Hsu[2] and Charles E. Leiserson[1]

[1]*Massachusetts Institute of Technology*
Cambridge, MA, USA
{*kunal_ag, cel*}*@mit.edu*

[2]*Nanyang Technological University*
Singapore, Singapore
yxhe@mit.edu, hsu@ntu.edu.sg

Multiprocessor scheduling in a shared multiprogramming environment can be structured as two-level scheduling, where a kernel-level job scheduler allots processors to jobs and a user-level thread scheduler schedules the work of a job on the allotted processors. In this context, the number of processors allotted to a particular job may vary during the job's execution, and the thread scheduler must adapt to these changes in processor resources. For overall system efficiency, the thread scheduler should also provide parallelism feedback to the job scheduler to avoid allotting a job more processors than it can use productively.

This paper provides an overview of several adaptive thread schedulers we have developed that provide provably good history-based feedback about the job's parallelism without knowing the future of the job. These thread schedulers complete the job in near-optimal time while guaranteeing low waste. We have analyzed these thread schedulers under stringent adversarial conditions, showing that the thread schedulers are robust to various system environments and allocation policies. To analyze the thread schedulers under this adversarial model, we have developed a new technique, called **trim analysis**, which can be used to show that the thread scheduler provides good behavior on the vast majority of time steps, and performs poorly on only a few. When our thread schedulers are used with dynamic equipartitioning and other related job scheduling algorithms, they are $O(1)$-competitive against an optimal offine scheduling algorithm with respect to both mean response time and makespan for batched jobs and nonbatched jobs, respectively. Our algorithms are the first nonclairvoyant scheduling algorithms to offer such guarantees.

Weaving Atomicity Through Dynamic Dependence Tracking

Suresh Jagannathan

Department of Computer Science
Purdue University
West Lafayette, IN, 47907
suresh@cs.purdue.edu

Programmability is the key hurdle towards effectively utilizing next-generation high-performance computing systems. Current trends in CMP processor design point to the emergence of many-core architectures, in which a single chip will support tens to potentially hundreds of cores. Systems constructed by aggregating these processors can enable *parallel execution of thousands of threads*.

Transactional memory (TM) has been the subject of significant interest in both academia and industry because it offers a compelling alternative to existing concurrency control abstractions, making it especially well-suited for programming applications on scalable multi-core platforms. TM abstractions permit logically concurrent access to shared regions of code, but ensure through some combination of hardware, compiler, and runtime support that such accesses do not violate intended serializability invariants. By doing so, transaction-based abstractions eliminate pernicious errors such as data races that can easily occur using locks, without compromising performance.

While the atomicity and isolation guarantees provided by transactions lead to greater composability and modularity than available using locks, these guarantees may require severe constraints on programmability. In this paper, we describe compiler and runtime techniques that allow structured communication among atomic regions to take place, thus selectively relaxing isolation invariants. Unlike existing proposals, our techniques are completely transparent, and provide a rational semantics for the interplay between transactions, message-passing abstractions, and exceptions.

A Key-based Adaptive Transactional Memory Executor

Tongxin Bai[1], Xipeng Shen[2], Chengliang Zhang[1], William N. Scherer III[3], Chen Ding[1] and Michael L. Scott[1]

[1]Computer Science Department
University of Rochester
Rochester, New York, USA
{bai, zhangchl, cding, scott}@cs.rochester.edu

[2]Computer Science Department
The College of William and Mary
Williamsburg, Virginia, USA
xshen@cs.wm.edu

[3]Computer Science Department
Rice University
Houston, Texas, USA
bill.scherer@cs.rice.edu

Software transactional memory systems enable a programmer to easily write concurrent data structures such as lists, trees, hashtables, and graphs, where non-conflicting operations proceed in parallel. Many of these structures take the abstract form of a *dictionary*, in which each transaction is associated with a search key. By regrouping transactions based on their keys, one may improve locality and reduce conflicts among parallel transactions.

In this paper, we present an executor that partitions transactions among available processors. Our key-based adaptive partitioning monitors incoming transactions, estimates the probability distribution of their keys, and adaptively determines the (usually nonuniform) partitions. By comparing the adaptive partitioning with uniform partitioning and round-robin keyless partitioning on a 16-processor SunFire 6800 machine, we demonstrate that key-based adaptive partitioning significantly improves the throughput of fine-grained parallel operations on concurrent data structures.

Optimizing Sorting with Machine Learning Algorithms

Xiaoming Li[1], Maria Jesus Garzaran[2] and David Padua[2]

[1]Department of Electrical and Computer Engineering
University of Delaware
Newark, Delaware, U.S.A.
xli@ece.udel.edu

[2]Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois, U.S.A.
garzaran@cs.uiuc.edu, padua@uiuc.edu

The growing complexity of modern processors has made the development of highly efficient code increasingly difficult. Manually developing highly efficient code is usually expensive but necessary due to the limitations of today's compilers. A promising automatic code generation strategy, implemented by library generators such as ATLAS, FFTW, and SPIRAL, relies on empirical search to identify, for each target machine, the code characteristics, such as the tile size and instruction schedules, that deliver the best performance. This approach has mainly been applied to scientific codes which can be optimized by identifying code characteristics that depend only on the target machine. In this paper, we study the generation of sorting routines whose performance also depends on the characteristics of the input data.

We present two approaches to generate efficient sorting routines. First, we consider the problem of selecting the best "pure" sorting algorithm as a function of the characteristics of the input data. We show that the relative performance of "pure" sorting algorithms can be encoded as a function of the *entropy* of the input data set. We used machine learning algorithms to compute a function for each target machine that, at runtime, is used to select the best algorithm. Our second approach generalizes the first approach and can build new sorting algorithms from a few primitive operations. We use genetic algorithms and a classifier system to build hierarchically-organized hybrid sorting algorithms capable of adapting to the input data. Our results show that the algorithms generated using this second approach are quite effective and perform significantly better than the many conventional sorting implementations we tested. In particular, the routines generated using the second approach perform better than the most popular libraries available today: IBM ESSL, INTEL MKL and the C++ STL. The best algorithm we have been able to generate is on the average 26% and 62% faster than the IBM ESSL in an IBM Power 3 and IBM Power 4, respectively.

Knowledge and Cache Conscious Algorithm Design and Systems Support for Data Mining Algorithms

Amol Ghoting[1], Gregory Buehrer[1], Matthew Goyder[1], Shirish Tatikonda[1], Xi Zhang[1], Srinivasan Parthasarathy[1], Tahsin Kurc[2] and Joel Saltz[2]

[1]*Department of Computer Science and Engineering*
The Ohio State University
Columbus, OH, USA
{*ghoting, buehrer, goyder, tatikond, srini*}*@cse.ohio-state.edu, xizhang@bmi.osu.edu*

[2]*Department of Biomedical Informatics*
The Ohio State University
Columbus, OH, USA
kurc@bmi.osu.edu, joel.saltz@osumc.edu

The knowledge discovery process is interactive in nature and therefore minimizing query response time is imperative. The compute and memory intensive nature of data mining algorithms makes this task challenging. We propose to improve the performance of data mining algorithms by re-architecting algorithms and designing effective systems support. From the view point of re-architecting algorithms, knowledge-conscious and cache-conscious design strategies are presented. Knowledge-conscious algorithm designs try and re-use repeated computation between iterations and across executions of a data mining algorithm. Cache-conscious algorithm designs on the other hand reduce execution time by maximizing data locality and re-use. The design of systems support that allows a variety of data mining algorithms to leverage knowledge-caching and cache-conscious placement with minimal implementation efforts is also presented.

Memory Optimizations For Fast Power-Aware Sparse Computations

Konrad Malkowski, Padma Raghavan and Mary Jane Irwin

Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA, USA
{*malkowsk, raghavan, mji*}*@cse.psu.edu*

We consider memory subsystem optimizations for improving the performance of sparse scientific computation while reducing the power consumed by the CPU and memory. We first consider a sparse matrix vector multiplication kernel that is at the core of most sparse scientific codes, to evaluate the impact of prefetchers and power-saving modes of the CPU and caches. We show that performance can be improved at significantly lower power levels, leading to over a factor of five improvement in the operations/Joule metric of energy efficiency. We then indicate that these results extend to more complex codes such as a multigrid solver. We also determine a functional representation of the impacts of such optimizations and we indicate how it can be used toward further tuning. Our results thus indicate the potential for cross-layer tuning for multiobjective optimizations by considering both features of the application and the architecture.

A global address space framework for locality aware scheduling

Sriram Krishnamoorthy[1], Umit Catalyurek[2], Jarek Nieplocha[3], Atanas Rountev[1] and P. Sadayappan[1]

[1]*Computer Science & Engineering*
Ohio State University
Columbus, OH, USA
{*krishnsr, rountevr, sadatr*}*@cse.ohio-state.edu*

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

[3]*Computational Science & Mathematics*
Pacific Northwest National Laboratory
Richland, WA, USA
jarek.nieplocha@pnl.gov

In this paper, we present a mechanism for automatic management of the memory hierarchy, including secondary storage, in the context of a global address space parallel programming framework. The programmer specifies the parallelism and locality in the computation. The scheduling of the computation into stages, together with the movement of the associated data between secondary storage and global memory, and between global memory and local memory, is automatically managed. A novel formulation of hypergraph partitioning is used to model the optimization problem of minimizing disk I/O. Experimental evaluation using a sub-computation from the quantum chemistry domain shows a reduction in the disk I/O cost by up to a factor of 11, and a reduction in turnaround time by up to 49%, as compared to alternative approaches used in state-of-the-art quantum chemistry codes.

Speedup using Flowpaths for a Finite Difference Solution of a 3D Parabolic PDE

Darrin M. Hanna[1], Anna M. Spagnuolo[2] and Michael Duchene[3]

[1]*Dept. of Computer Science and Engineering*
Oakland University
Rochester, MI, 48309
dmhanna@oakland.edu

[2]*Dept. of Mathematics and Statistics*
Oakland University
Rochester, MI, 48309
spagnuol@oakland.edu

[3]*Dept. of Computer Science and Engineering*
Oakland University
Rochester, MI, 48309
mjduchen@oakland.edu

Partial differential equations (PDEs) are used to model physical phenomena and then appropriate convergent numerical algorithms are employed to solve them and create computer simulations. In many important applications, such as weather prediction and contaminant transport processes, simulation outputs are required in real time or even faster, yet the spatial component of the problem is very large, thereby increasing the computational time. In addition, often times numerical scientists work in groups to create a large-scale code, but they work individually on PCs to test components of the code, so that speedup of the computational algorithms on PCs is desirable. There is a benefit to creating and using custom hardware to perform the numerical calculations faster than commodity hardware. This work uses a high-level programming language (Java) to behaviorally describe, and then implement, a finite difference solution of a parabolic PDE as a custom hardware circuit targeted to an FPGA. The results show that the circuits can perform the calculations 1 to 2 orders of magnitude faster than commodity hardware.

NGS: Service Adaptation in Open Grid Platforms

Krishnaveni Budati[1], Jinoh Kim[2], Abhishek Chandra[3] and Jon Weissman[4]

[1]*Department of Computer Science*
UMn
Minneapolis, MN, USA
budati@cs.umn.edu

[2]*Department of Computer Science*
UMn
Minneapolis, MN, USA
jkim@cs.umn.edu

[3]*Department of Computer Science*
UMn
Minneapolis, MN, USA
chandra@cs.umn.edu

[4]*Department of Computer Science*
UMn
Minneapolis, MN, USA
jon@cs.umn.edu

Large-scale donation-based distributed infrastructures need to cope with the inherent unreliability of participant nodes. A widely-used work scheduling technique in such environments is to redundantly schedule the outsourced computations to a number of nodes. We present the design and implementation of RIDGE, a reliability-aware system which uses a node's prior performance and behavior to make more effective scheduling decisions. We have implemented RIDGE on top of the BOINC distributed computing infrastructure and have evaluated its performance on a live PlanetLab testbed. Our experimental results show that RIDGE is able to match or surpass the throughput of the best BOINC configuration by automatically adapting to the characteristics of the underlying environment. In addition, RIDGE is able to provide much lower workunit makespans compared to BOINC. RIDGE is also able to produce significantly lower communication makespans for downloading clients. Collectively, the results suggest that RIDGE has great promise for service-oriented environments with time constraints.

Creating a Robust Desktop Grid using Peer-to-Peer Services

Jik-Soo Kim[1], Beomseok Nam[1], Michael Marsh[1], Peter Keleher[1], Bobby Bhattacharjee[1], Derek Richardson[2], Dennis Wellnitz[2] and Alan Sussman[1]

[1]*UMIACS and Department of Computer Science*
University of Maryland
College Park, MD, U.S.A.
{jiksoo, bsnam, mmarsh, keleher, bobby, als}@cs.umd.edu

[2]*Department of Astronomy*
University of Maryland
College Park, MD, U.S.A.
{dcr, wellnitz}@astro.umd.edu

The goal of the work described in this paper is to design and build a scalable infrastructure for executing grid applications on a widely distributed set of resources. Such grid infrastructure must be decentralized, robust, highly available, and scalable, while efficiently mapping application instances to available resources in the system. However, current desktop grid computing platforms are typically based on a client-server architecture, which has inherent shortcomings with respect to robustness, reliability and scalability. Fortunately, these problems can be addressed through the capabilities promised by new techniques and approaches in Peer-to-Peer (P2P) systems. By employing P2P services, our system allows users to submit jobs to be run in the system and to run jobs submitted by other users on any resources available in the system, essentially allowing a group of users to form an ad-hoc set of shared resources. The initial target application areas for the desktop grid system are in astronomy and space science simulation and data analysis.

Locality-aware Buffer Management: Algorithms Design and Systems Implementation for Data Intensive Applications

Xiaodong Zhang

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

The speed gap between data processing in CPU and data accessing in disks has reached to an intolerable level, and will only become worse as time goes by. This bottleneck has seriously hindered the development of large scale computing systems for data-intensive applications that demand fast accesses to a huge amount of data. Viable and cost-effective solutions to address this problem are to build large memory buffers to cache data for reuse by taking advantage of low price and large capacity of DRAM memory, and to prefetch data for predicted future use by taking advantage of high and idle bandwidths of networks. We are working on four different research projects on locality-aware buffer management for data intensive applications. This report briefly presents the background, motivation, and working progress of our work funded by the NSF NGS program.

Designing Efficient Systems Services and Primitives for Next-Generation Data-Centers

Karthikeyan Vaidyanathan, Sundeep Narravula, Pavan Balaji and Dhabaleswar K. Panda

Department of Computer Science and Engineering
The Ohio State University
Columbus, OH, USA
{vaidyana, narravul, balaji, panda}@cse.ohio-state.edu

Current data-centers lack in efficient support for intelligent services, such as requirements for caching documents and cooperation of caching servers, efficiently monitoring and managing the limited physical resources, load-balancing, controlling overload scenarios, 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). In this paper, we extend our previously proposed 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 present the performance results using data-center services such as cooperative caching and active resource monitoring and data-center primitives such as distributed data sharing substrate and distributed lock manager, which demonstrate significant performance benefits achievable by our framework as compared to existing data-centers in several cases.

Supporting Quality of Service in High-Performance Servers

Yan Solihin, Fei Guo, Seongbeom Kim and Fang Liu

Deptartment of Electrical and and Computer Engineering
North Carolina State University
Raleigh, NC, USA
{*solihin, fguo, skim16, fliu3*}*@ece.ncsu.edu*

This paper describes issues that we have analyzed and technology that we have developed in supporting Quality of Service in high-performance servers. More specifically, we target on-chip cache resource allocation and efficiency needed for guaranteeing certain performance levels on Chip Multi-Processor (CMP) architectures. Both prior and ongoing work are summarized in this paper.

Enhancing Energy Efficiency in Multi-tierWeb Server Clusters via Prioritization

Tibor Horvath[1], Kevin Skadron[2] and Tarek Abdelzaher[3]

[1]*Department of Computer Science*
University of Virginia
Charlottesville, VA, USA
tibor@cs.virginia.edu

[2]*Department of Computer Science*
University of Virginia
Charlottesville, VA, USA
skadron@cs.virginia.edu

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

This paper investigates the design issues and energy savings benefits of service prioritization in multi-tier web server clusters. In many services, classes of clients can be naturally assigned different priorities based on their performance requirements. We show that if the whole multitier system is effectively prioritized, additional power and energy savings are realizable while keeping an existing cluster-wide energy management technique, through exploiting the different performance requirements of separate service classes. We find a simple prioritization scheme to be highly effective without requiring intrusive modifications to the system. In order to quantify its benefits, we perform extensive experimental evaluation on a real testbed. It is shown that the scheme significantly improves both total system power savings and energy efficiency, at the same time as improving throughput and enabling the system to meet per-class performance requirements.

Autonomic Power and Performance Management for Large Scale Data Centers

Bithika Khargharia[1], Salim Hariri[1], Ferenc Szidarovszky[2], Hesham El-Rewini[3], Manal Houri[3], Samee Khan[4], Ishfaq Ahmad[4] and Mazin S. Yousif[5]

[1]*Electrical and Computer Engineering*
University of Arizona
Tucson, AZ, USA
bithikak@email.arizona.edu, hariri@ece.arizona.edu

[2]*Systems and Industrial Engineering*
University of Arizona
Tucson, AZ, USA
szidar@sie.arizona.edu

[3]*Department of Computer Science and Engineering*
Southern Methodist University
Dallas, TX, USA
{rewini, mhouri}@engr.smu.edu

[4]*Computer Science and Engineering Department*
The University of Texas at Arlington
Arlington, TX, USA
{sakhan, iahmad}@cse.uta.edu

[5]*Intel*
Hillsboro, OR, USA
mazin.s.yousif@intel.com

With the rapid growth of servers and applications spurred by the Internet, the power consumption of servers has become critically important and must be efficiently managed. High energy consumption also translates into excessive heat dissipation which in turn, increases cooling costs and causes servers to become more prone to failure. This paper presents a theoretical and experimental framework and general methodology for hierarchical autonomic power & performance management in high performance distributed data centers. We optimize for power & performance (performance/watt) at each level of the hierarchy while maintaining scalability. We adopt mathematically-rigorous optimization approach to provide the application with the required amount of memory at runtime. This enables us to transition the unused memory capacity to a low power state. Our experimental results show a maximum performance/watt improvement of 88.48% compared to traditional techniques. We also present preliminary results of using Game Theory to optimize performance/watt at the cluster level of a data center. Our cooperative technique reduces the power consumption by 65% when compared to traditional techniques (min-min heuristic).

Improving Data Access Performance with Server Push Architecture

Xian-He Sun, Surendra Byna and Yong Chen

Department of Computer Science
Illinois Institute of Technology
Chicago, IL, USA
{sun, renbyna, chenyon1}@iit.edu

Data prefetching, where data is fetched before CPU demands for it, has been considered as an effective solution to mask data access latency. However, the current client-initiated prefetching strategies do not work well for applications with complex, non-contiguous data access patterns. While technology advances continue to enlarge the gap between computing and data access performance, trading computing power for data access delay has become a natural choice. We propose a server-based data-push approach. In this server-push architecture, a dedicated server named Data Push Server (DPS) initiates and proactively pushes data closer to the client in time. We present the DPS architecture and study the issues such as what data to fetch, when to fetch, how to push, and data access modeling.

SimX meets SCIRun: A Component-based Implementation of a Computational Study System

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

[1]*Computer Science*
New York University
New York, NY, USA
{*smyau, vijayk, dzorin*}*@cs.nyu.edu*

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

This paper describes the ongoing implementation of the SimX system for multi-experiment computational studies within the SCIRun problem solving environment. The modular, component-based nature of SCIRun enables a natural integration of the SimX runtime modules with the simulation codes that constitute the experiments underlying the study, and provides a rich steering and visualization environment for study interactions. Experience with a computational study involving a SCIRun defibrillator device simulation code (DefibSim) highlights these advantages, and identifies several avenues for future work.

VIProf: Vertically Integrated Full-System Performance Profiler

Hussam Mousa, Chandra Krintz, Lamia Youseff and Rich Wolski

Computer Science Dept
UC Santa Barbara
Santa Barbara, CA, USA
{*husmousa, ckrintz, lyouseff, rich*}*@cs.ucsb.edu*

In this paper, we present VIProf, a full-system, performance sampling system capable of extracting runtime behavior across an entire software stack. Our long-term goal is to employ VIProf profiles to guide online optimization of programs and their execution environments according to the dynamically changing execution behavior and resource availability. VIProf thus, must be transparent while producing accurate and useful performance profiles.

We overview the design and implementation of VIProf and empirically evaluate the system using a popular software stack – one that includes a Linux operating system, a Java Virtual Machine, and a set of applications. This composition is commonly employed and important for high-end systems such as application and web servers as well as Computational Grid services. We show that VIProf introduces little overhead and is able to capture accurate (function-level) full-system performance data that previously required multiple profiles and extensive, manual, and offline post-processing of profile data.

Model Predictive Control for Memory Profiling

Sean Callanan, Radu Grosu, Justin Seyster, Scott A. Smolka and Erez Zadok

Computer Science Department
Stony Brook University
Stony Brook, New York, United States
{*spyffe, grosu, jseyster, sas, ezk*}*@cs.sunysb.edu*

We make two contributions in the area of memory profiling. The first is a *real-time, memory-profiling toolkit* we call *Memcov* that provides both allocation/deallocation and access profiles of a running program. Memcov requires *no recompilation or relinking* and significantly reduces the barrier to entry for new applications of memory profiling by providing a clean, non-invasive way to perform two major functions: processing of the stream of memory-allocation events in real time and monitoring of regions in order to receive notification the next time they are hit.

Our second contribution is an *adaptive memory profiler and leak detector* called *Memcov_MPC*. Built on top of Memcov, Memcov_MPC uses *Model Predictive Control* to derive an optimal control strategy for leak detection that maximizes the number of areas monitored for leaks, while minimizing the associated runtime overhead. When it observes that an area has not been accessed for a user-definable period of time, it reports it as a potential leak. Our approach requires neither mark-and-sweep leak detection nor static analysis, and reports a superset of the memory leaks actually occurring as the program runs. The set of leaks reported by Memcov_MPC can be made to approximate the actual set more closely by lengthening the threshold period.

Understanding Measurement Perturbation in Trace-Based Data

Todd Mytkowicz[1], Amer Diwan[1], Matthias Hauswirth[2] and Peter F. Sweeney[3]

[1]*Computer Science*
University of Colorado
Boulder, CO, USA
{*Todd.Mytkowicz, Amer.Diwan*}*@colorado.edu*

[2]*Informatics*
University of Lugano
Lugano, Switzerland
matthias.hauswirth@unisi.ch

[3]*IBM TJ Watson Research Center*
Hawthorne, NY, USA
pfs@us.ibm.com

Performance analysts commonly use trace-based data containing hardware and software metrics to understand performance. The trace data is generated by instrumenting the code to increment a counter when an event occurs and to collect hardware and software metrics in a trace. Unfortunately, the act of collecting a trace can perturb the behavior that the trace is trying to capture.

In this paper, we gain an understanding of perturbation due to measurement instrumentation of the system. We identify two mechanisms to quantify perturbation: inner and outer perturbation. Using inner perturbation, a performance analyst can determine when a run is perturbed by collecting too much information. Using outer perturbation, the performance analyst can determine if she can use the data from multiple runs as if the data were all from a single run.

Our evaluation of these mechanisms lead to two results. First, we are surprised to find that even with minimal instrumentation overhead, which increased instructions executed by less than 3%, high perturbation resulted, which prevented one from correctly reasoning about metrics within a trace or across traces. Second, the instrumentation of different software metrics interact in subtle, and not always obvious, ways making the impact of instrumentation on perturbation difficult, if not impossible, to predict.

Finally, we outline a methodology for collecting data while avoiding perturbation. When inner perturbation occurs, the performance analyst can spread out the data collection over multiple runs. When outer perturbation occurs, she can try different strategies for spreading out the data collection over multiple runs.

PROTOFLEX: FPGA-accelerated Hybrid Functional Simulator

Eric S. Chung, Eriko Nurvitadhi, James C. Hoe, Babak Falsafi and Ken Mai

Computer Architecture Lab (CALCM)
Carnegie Mellon University
Pittsburgh, PA, USA
{*echung, enurvita, jhoe, babak, kenmai*}*@ece.cmu.edu*

PROTOFLEX is an FPGA-accelerated hybrid simu-lation/emulation platform designed to support large-scale multiprocessor hardware and software research. Unlike prior attempts at FPGA multiprocessor system emulators, PROTOFLEX emulates full-system fidelityi.e., runs stock commercial operating systems with I/O support. This is ac-complished without undue effort by leveraging a hybrid emulation technique called transplanting. Our transplant technology uses FPGAs to accelerate only common-case behaviors while relegating infrequent, complex behaviors (e.g., I/O devices) to software simulation. By working in concert with existing full-system simulators, transplanting avoids the costly and unnecessary construction of the entire target system in FPGA. We report preliminary findings from a working hybrid PROTOFLEX emulator of an UltraSPARC workstation running Solaris 8.

We have also started developing a novel multiprocessor emulation approach that interleaves the execution of many (10s to 100s) processor contexts onto a shared emulation engine. This approach decouples the scale and complexity of the FPGA host from the simulated system size but neverthe-less enables us to scale the desired emulation performance by the number of emulation engines used. Together, the transplant and interleaving techniques will enable us to develop full-system FPGA emulators of up to thousands of processors without an overwhelming development effort.

Models and Heuristics for Robust Resource Allocation in Parallel and Distributed Computing Systems

David L. Janovy[1], Jay Smith[1,3], Howard Jay Siegel[1,2] and Anthony A. Maciejewski[1]

[1]*Electrical and Computer Engineering*
Colorado State University
Fort Collins, CO, USA
{*djanovy, hj, aam*}*@colostate.edu, bigfun@us.ibm.com*

[2]*Computer Science*
Colorado State University
Fort Collins, CO, USA

[3]*IBM*
Boulder, CO, USA

This is an overview of the robust resource allocation research efforts that have been and continue to be conducted by the CSU Robustness in Computer Systems Group.

Parallel and distributed computing systems, consisting of a (usually heterogeneous) set of machines and networks, frequently operate in environments where delivered performance degrades due to unpredictable circumstances. Such unpredictability can be the result of sudden machine failures, increases in system load, or errors caused by inaccurate initial estimation. The research into developing models and heuristics for parallel and distributed computing systems that create robust resource allocations is presented.

Model-Driven Performance Analysis Methodology for Distributed Software Systems

Swapna S. Gokhale[1], Paul J. Vandal[1], Aniruddha S. Gokhale[2], Dimple Kaul[2], Arundhati Kogekar[2], Jeff Gray[3] and Yuehua Lin[3]

[1]*Computer Science and Engineering*
Univ. of Connecticut
Storrs, CT, USA
{*ssg, pvandal*}*@engr.uconn.edu*

[2]*Electrical Engineering and Computer Science*
Vanderbilt Univ.
Nashville, TN, USA
a.gokhale@vanderbilt.edu, dkaul@isis.vanderbilt.edu, akogekar@dre.vanderbilt.edu

[3]*Computer and Information Sciences*
Univ. of Alabama at Birmingham
Birmingham, AL, USA
{*gray, liny*}*@cis.uab.edu*

A key enabler of the recently popularized, assembly-centric development approach for distributed software systems is QoS-enabled middleware, which provides reusable building blocks in the form of design patterns that codify solutions to commonly recurring problems. These patterns can be customized by choosing an appropriate set of configuration parameters. The configuration options of the patterns exert a strong influence on system performance, which is of paramount importance in many distributed software systems. Despite this considerable influence, currently there is a lack of significant research to analyze performance of middleware at design time, where performance issues can be resolved at a much earlier stage of the application life cycle and with substantially less costs. This project seeks to develop a performance analysis methodology for design-time performance analysis for distributed software systems implemented using middleware patterns and their compositions. The methodology is illustrated on a producer/consumer system implemented using the Active Object (AO) pattern in middleware. Finally, broader impacts of the methodology for middleware specialization are also described.

J-Sim: An Integrated Environment for Simulation and Model Checking of Network Protocols

Ahmed Sobeih, Mahesh Viswanathan, Darko Marinov and Jennifer C. Hou

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{*sobeih, vmahesh, marinov, jhou*}*@cs.uiuc.edu*

In this paper, we report our work [24, 26] on extending the J-Sim network simulator [13] to be an integrated environment for both simulation and model checking of network protocols. We also present a case study in which we model-checked AODV in J-Sim.

Early Results with Precision Abstraction: Using Data-flow Analysis to Improve the Scalability of Model Checking

Adam Brown, James C. Browne and Calvin Lin

Department of Computer Sciences
The University of Texas at Austin
Austin, TX, USA
{*abrown, browne, lin*}*@cs.utexas.edu*

This paper presents a new state space reduction technique that applies to model checking of software. The new technique, *precision abstraction*, borrows ideas from data-flow analysis to identify procedures that can be analyzed context-insensitively without affecting the accuracy of the verification of a given property. These context-insensitive procedures can then be represented with fewer states than would be needed context-sensitive analysis. Preliminary results indicate that the number of transitions in the analysis prescribed by our approach is at least 155 times fewer than the exhaustive analysis a model checker would otherwise perform.

Static Verification of Design Constraints and Software Correctness Properties in the Hob System

Patrick Lam[1] and Martin Rinard[2]

[1]*Department of Electrical Engineering and Computer Science*
MIT
Cambridge, MA, USA
plam@mit.edu

[2]*Department of Electrical Engineering and Computer Science*
MIT
Cambridge, MA, USA
rinard@mit.edu

Sets of objects are an intuitive foundation for many object-oriented design formalisms, serving as a key concept for describing elements of the design and promoting communication between members of the development team. It may be natural for the sets of the objects in the design to correspond to the sets of objects in the implementation. In practice, however, the object structure of the implementation is much more complex than that of the design. Moreover, the lack of an enforced connection between the implementation and the design enables the implementation to diverge from the design, rendering the design unreliable as a source of information about the implementation.

Hob allows developers to express and verify the connection between abstract sets of design objects and concrete sets of implementation objects. Abstraction maps define the meaning of the design sets in terms of the objects in the implementation, enabling the elimination of implementation complexity not relevant to the design. An abstract set specification language enables the developer to state important relationships (such as inclusion and disjointness) between abstract sets of objects; our verification system statically checks that the implementation correctly preserves these design-level correctness properties. We have implemented Hob and used it to develop several software systems. Our experience shows that Hob enables the effective expression and verification of precise design constraints that manifest themselves as important correctness properties that the implemented system is guaranteed to preserve.

ExPert: Dynamic Analysis Based Fault Location via Execution Perturbations

Neelam Gupta and Rajiv Gupta

Department of Computer Science
University of Arizona
Tucson, AZ, USA
{*ngupta, gupta*}*@cs.arizona.edu*

We are designing dynamic analysis techniques to identify executed program statements where a fault lies, i.e. the fault candidate set. To narrow the set of statements in the fault candidate set, automated dynamic analyses are being developed which consider not only a failed run of a program but also execution perturbations of the failed run. The goal of this work is to focus the users attention on a small subset of statements in the fault candidate set.

An Analysis of Availability Distributions in Condor

Rich Wolski[1], Daniel Nurmi[1] and John Brevik[2]

[1]*Computer Science*
UCSB
Santa Barbara, CA, USA
rich@cs.ucsb.edu, nurmi@cs.ucsb.edi

[2]*ComputerScience*
UCSB
Santa Barbara, CA, USA
jbrevik@csulb.edu

[3]*Mathematic and Statistics*
Cal State Long Beach
Long Beach, CA, USA

In this paper, we consider the problem of modeling machine availability for the enterprise-area and wide-area distributed system known as Condor. Using availability data gathered from Condor, we detail the suitability of three potential statistical distributions for the data: Weibull, Log-normal, and hyperexponential. In each case, we use software we have developed to determine the necessary parameters automatically from each data collection.

These results indicate that there are two classes of machines in the Condor pool and that different statistical models are best suited to each.

Identifying and Addressing Uncertainty in Architecture-Level Software Reliability Modeling

Leslie Cheung[1], Leana Golubchik[1,2], Nenad Medvidovic[1] and Gaurav Sukhatme[1]

[1]Computer Science Department
University of Southern California
Los Angeles, CA, USA
{lccheung, leana, neno, gaurav}@usc.edu

[2]EE-Systems Dept, IMSC
University of Southern California
Los Angeles, CA, USA

Assessing reliability at early stages of software development, such as at the level of software architecture, is desirable and can provide a cost-effective way of improving a software systems quality. However, predicting a components reliability at the architectural level is challenging because of uncertainties associated with the system and its individual components due to the lack of information. This paper discusses representative uncertainties which we have identified at the level of a systems components, and illustrates how to represent them in our reliability modeling framework. Our preliminary evaluation indicates promising results in our framework's ability to handle such uncertainties.

A Markov Reward Model for Software Reliability

Youngmin Kwon and Gul Agha

Open Systems Laboratory, Department of Computer Science
University of Illinois at Urbana-Champaign
Champaign, IL, USA
{ykwon4, agha}@cs.uiuc.edu

A compositional method for estimating software reliability of many threaded programs is developed. The method uses estimates of the reliability of individual modules and the probability of transitions between the modules to estimate the reliability of a program in terms of its current state. The reliability of a program is expressed using *iLTL*, a probabilistic linear temporal logic whose atomic propositions are linear inequalities about transitions of the probability mass function of a Discrete Time Markov Chain. We then use a Markov reward model to estimate software reliability. The technique is illustrated in terms of an example.

Modeling Modern Micro-architectures using CASL

Edward K. Walters II, J. Eliot B. Moss, Trek Palmer, Timothy Richards and Charles C. Weems

Department of Computer Science
University of Massachusetts Amherst
Amherst, MA, United States
{*ekw,moss,trekp,richards,weems*}*@cs.umass.edu*

We overview CASL, the CoGenT Architecture Specification Language, a mixed behavioral-structure architecture description language designed to facilitate fast prototyping and tool generation for computer architectures with deep pipelines and complicated timing. We show how CASL can describe pipelines, dynamic information contexts, and contention using the DLX/MIPS architecture as an example.

Rethinking Automated Synthesis of MPSoC Architectures

Brett H. Meyer and Donald E. Thomas

Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA, USA
{*bhm, thomas*}*@ece.cmu.edu*

Emerging heterogeneous multiprocessors will have custom memory and bus architectures that must balance resource sharing and system partitioning to meet cost constraints. We propose an augmented simulated annealing synthesis tool that uses system performance and layout evaluation to drive simultaneous data mapping, memory allocation and bus synthesis. A detailed look at the resulting automated design process reveals an approach that, contrary to prior approaches, optimizes bus topology first rather than last, providing design insight for the development of future tools.

A Reconfigurable Chip Multiprocessor Architecture to Accommodate Software Diversity

Engin İpek, Meyrem Kırman, Nevin Kırman and José F. Martínez

Computer Systems Laboratory
Cornell University
Ithaca, NY, USA
{*engin, mkirman, nkirman, martinez*}*@csl.cornell.edu*

We present *core fusion*, a reconfigurable chip multiprocessor (CMP) architecture where groups of fundamentally independent cores can dynamically morph into a larger CPU, or they can be used as distinct processing elements, as needed at run time by applications. Core fusion gracefully accommodates software diversity and incremental parallelization in CMPs. It provides a single execution model across all configurations, requires no additional programming effort or specialized compiler support, maintains ISA compatibility, and leverages mature micro-architecture technology.

Scalable, Dynamic Analysis and Visualization for Genomic Datasets

Grant Wallace[1], Matthew Hibbs[1,2], Maitreya Dunham[2], Rachel Sealfon[1], Olga Troyanskaya[1,2] and Kai Li[1]

[1]*Department of Computer Science*
Princeton University
Princeton, NJ, USA
{*gwallace, mhibbs, ogt, li*}*@cs.princeton.edu,*
sealfon@princeton.edu

[2]*Lewis-Sigler Institute for Integrative Genomics*
Princeton University
Princeton, NJ, USA
maitreya@princeton.edu

A challenge in data analysis and visualization is to build new-generation software tools and systems to truly accelerate scientific discoveries. The recent focus of Princetons next-generation software project is to investigate how to develop new-generation data analysis and visualization capabilities for genomic scientists to analyze high-throughput genomic datasets. This paper describes the software tools we have recently developed to enable dynamic, large-scale data analysis and visualization of multiple datasets on large-scale, high-resolution display wall systems. Our initial experience with the deployed tools at Princetons Lewis-Sigler Institute for Integrative Genomics is very encouraging. Scientists can effectively learn new knowledge from multiple datasets, find new insights, and generate new hypotheses that are not possible with current methods.

Scalable Distributed Execution Environment for Large Data Visualization

Micah Beck, Huadong Liu, Jian Huang and Terry Moore

Dept. of Computer Science
University of Tennessee
Knoxville, TN, USA
{*mbeck, hliu, huangj, tmoore*}*@cs.utk.edu*

To use heterogeneous and geographically distributed resources as a platform for parallel visualization is an intriguing topic of research. This is because of the immense potential impact of the work, and also because of its use of a full range of challenging techniques. In this work, we designed an execution environment for visualization of massive scientific datasets, using network functional units (NFU) for processing power, logistical networking for storage management and visualization cookbook library (vcblib) for visualization operations. This environment is based solely on computers distributed across the Internet that are owned and operated by independent institutions, while being openly shared for free. Those Internet computers are inherently of heterogeneous hardware configuration and running a variety of operating systems. The system is enabled by new techniques Using 100 such processors, we have been able to obtain the same level of performance offered by a 64-node cluster of 2.2 GHz P4 processors, while processing a 75GBs subset of TSI simulation data. Due to its inherently shared nature, this execution environment for data-intensive visualization could provide a viable means of collaboration among geographically separated computational scientists.

Annotation Integration and Trade-off Analysis for Multimedia Applications

Radu Cornea, Alex Nicolau and Nikil Dutt

Donald Bren School of Information and Computer Science
University of California, Irvine
Irvine, CA, 92697
{*radu, nicolau, dutt*}*@ics.uci.edu*

Multimedia applications for mobile devices, such as video/audio streaming, process streams of incoming data in a regular, predictable way. Content-aware optimizations through annotations allow us to highly improve the power savings at the various levels of abstraction: hardware/OS, network, application. However, in a typical system there is a continuous interaction between the components of the system at all levels, which requires a careful analysis of the combined effect of the aforementioned techniques. We investigate such an interaction and we describe metrics for estimating the effect various trade-off have on power and quality. By applying our metrics at the various abstraction levels we show how better energy savings can be achieved with lower quality degradations, through power-quality trade-offs and cross-layer interaction.

Workshop 11
High-Performance, Power-Aware Computing
HPPAC 2007

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:

Kirk W. Cameron, Virginia Polytechnic Institute and State University, USA

Padma Raghavan, Pennsylvania State University, USA

Program Committee:

Kirk W. Cameron, Virginia Polytechnic Institute and State University, USA
Padma Raghavan, Pennsylvania State University, USA
Frank Bellosa, University of Karlsruhe, Germany
Bronis de Supinski, Lawenrce Livermore National Laboratory, USA
Wu-Chun Feng, , Virginia Polytechnic Institute and State University, USA
Xizhou Feng, , Virginia Polytechnic Institute and State University, 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
David K. Lowenthal, University of Georgia, USA

A Power-Aware Prediction-Based Cache Coherence Protocol for Chip Multiprocessors

Ehsan Atoofian and Amirali Baniasadi

Electrical and Computer Engineering
University of Victoria
Victoria, BC, Canada
atoofian@uvic.ca, amirali@ece.uvic.ca

Snoopy cache coherence protocols broadcast requests to all nodes, reducing the latency of cache to cache transfer misses at the expense of increasing interconnect power. We propose speculative supplier identification (SSI) to reduce power dissipation in binary tree interconnects in snoopy cache coherence implementations. In SSI, instead of broadcasting a request to all processors, we send the request to the node more likely to have the missing data. We reduce power as we limit access only to the interconnect components between the requestor and the supplier node. We evaluate SSI using shared memory applications. We show that SSI reduces interconnect power by 23% in a 4-way multiprocessor. This comes with negligible performance cost and hardware overhead. SSI does not change existing coherence protocols and is completely transparent to software and the operating system.

Link Shutdown Opportunities During Collective Communications in 3-D Torus Nets

Sarah Conner, Sayaka Akioka, Mary Jane Irwin and Padma Raghavan

Computer Science and Engineering Department
The Pennsylvania State University
University Park, PA, USA
{sconner, tobita, mji, raghavan}@cse.psu.edu

As modern computing clusters used in scientific computing applications scale to ever-larger sizes and capabilities, their operational energy costs have become prohibitive. While it is an emerging trend in modern cluster design to optimize for low energy consumption in the individual computational nodes, little attention has been paid to reducing the energy used by the communication network that connects the nodes. In this work. we consider a 3-D torus network similar to the one in BlueGene/L to explore opporunities for link shutdown during collective communication operations. For example, we demonstrate that in the case of all-to-one reduce codes, approximately 99% of the total network link time can be spent in a shutoff state on a 64-node toroidal network, thus reducing the overall system energy by approximately 15–28%

A High Performance Cluster System Design by Adaptive Power Control

Masaaki Kondo, Yoshimichi Ikeda and Hiroshi Nakamura

Research Center for Advanced Science and Technology
The University of Tokyo
Tokyo, Japan
{*kondo, ikeda, nakamura*}*@hal.rcast.u-tokyo.ac.jp*

The first order design constraint in dense packaged clusters is power consumption. The currently developed cluster systems are conservatively designed so that the expected peak power does not exceed the power limit. However, practical power consumption seldom reaches the peak power. In this paper, we propose a new approach to design a high performance cluster system by an adaptive power control technique. Our approach is to integrate many computation nodes into a system whose total theoretical peak power exceeds the limit and to control runtime effective power by optimizing the number of working nodes and/or the clock frequency of the processors. We show the algorithm of the adaptive power control and erformance evaluation by using a real cluster system. Evaluation results show that our proposed approach greatly improves performance as large as 46% compared to a conventional cluster system.

Load Miss Prediction - Exploiting Power Performance Trade-offs

Konrad Malkowski, Greg Link, Padma Raghavan and Mary Jane Irwin

Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA, USA
{*malkowsk, link, raghavan, mji*}*@cse.psu.edu*

Modern CPUs operate at GHz frequencies, but the latencies of memory accesses are still relatively large, in the order of hundreds of cycles. Deeper cache hierarchies with larger cache sizes can mask these latencies for codes with good data locality and reuse, such as structured dense matrix computations. However, cache hierarchies do not necessarily benefit sparse scientific computing codes, which tend to have limited data locality and reuse. We therefore propose a new memory architecture with a *Load Miss Predictor (LMP)*, which includes a data bypass cache and a predictor table, to reduce access latencies by determining whether a load should bypass the main cache hierarchy and issue an early load to main memory. Our architecture uses the L2 (and lower caches) as a victim cache for data removed from our bypass cache. We use cycle-accurate simulations, with SimpleScalar and Wattch to show that our LMP improves the performance of sparse codes, our application domain of interest, on average by 14%, with a 13.6% increase in power. When the LMP is used with dynamic voltage and frequency scaling (DVFS), performance can be improved by 8.7% with system power savings of 7.3% and energy reduction of 17.3% at 1800MHz relative to the base system at 2000MHz. Alternatively our LMP can be used to improve the performance of SPEC benchmarks by an average of 2.9% at the cost of 7.1% increase in average power.

Leakage Energy Reduction in Value Predictors through Static Decay

Juan Manuel Cebrián, Juan Luis Aragón and José Manuel García

Department of Computer Engineering
University of Murcia
Murcia, Murcia, Spain
{*jcebrian, jlaragon, jmgarcia*}*@ditec.um.es*

As process technology advances toward deep submicron (below 90nm), static power becomes a new challenge to address for energy-efficient high performance processors, especially for large on-chip array structures such as caches and prediction tables. Value Prediction appeared as an effective way of increasing processor performance by overcoming data dependences, but at the risk of becoming a thermal hot spot due to the additional power dissipation.

This paper proposes the design of low-leakage Value Predictors by applying static decay techniques in order to disable unused entries from the prediction tables. We explore decay strategies suited for the three most common Value Predictors (STP, FCM and DFCM) studying the particular tradeoffs for these prediction structures. Our mechanism reduces VP leakage energy efficiently without compromising VP accuracy nor processor performance. Results show average leakage energy reductions of 52%, 65% and 75% for the STP, DFCM and FCM Value Predictors, respectively.

Determining the Minimum Energy Consumption using Dynamic Voltage and Frequency Scaling

Min Yeol Lim and Vincent W. Freeh

Computer Science
North Carolina State University
Raleigh, NC, USA
{*mlim, vwfreeh*}*@ncsu.edu*

While improving raw performance is of primary interest to most users of high-performance computers, energy consumption also is a critical concern. Some microprocessors allow voltage and frequency scaling, which enables a system to reduce CPU power and performance when the CPU is not on the critical path. When properly directed, such dynamic voltage and frequency scaling can produce significant energy savings with little performance penalty. Various DVFS scaling algorithms have been proposed. However, the benefit is application-dependent. We can not see if they achieve the energy consumption as minimum as possible. So, it is important to establish the baseline of the DVFS scheduling for any application. This paper determines minimum energy consumption in voltage and frequency scaling systems for a given time delay. We assume we have a set of fixed points where scaling can occur. A brute-force solution is intractable even for a moderately sized set (although all programs presented in this paper can be solved with the brute-force). Our algorithm efficiently chooses the exact optimal schedule satisfying the given time constraint by estimation. We evaluate our time and energy estimates in NPB serial benchmark suite. The results show that the running time can be reduced significantly with our algorithm. Besides, our time and energy estimations from the optimal schedule have reasonable accuracy with 1.48% of differences at maximum.

Scaling and Packing on a Chip Multiprocessor

Vincent W. Freeh[1], Tyler K. Bletsch[1] and Freeman L. Rawson, III[2]

[1]Department of Computer Science
North Carolina State University
Raleigh, NC, 27695-8206
vin@csc.ncsu.edu, tkbletsc@ncsu.edu

[2]IBM Austin Research Laboratory
Austin, TX, 78758
frawson@us.ibm.com

Power management is critical in server and high-performance computing environments as well as in mobile computing. Many mechanisms have been developed over recent years to support a wide a variety of power management techniques. In particular, general purpose microprocessors now support dynamically modifying the power-performance state through voltage and frequency changes. This development spawned a very important area of this research in *dynamic voltage and frequency scaling* (DVFS). On the other hand, in a multiprocessor environment one can perform power management by offlining and idling processors when computational demand is low in a technique called *CPU packing*. This paper examines the effect of combining voltage and frequency scaling and CPU packing in a multiprocessor. Furthermore, it examines DVFS on a *chip multiprocessor* in which multiple processor cores are placed on a single die. This paper shows that in general one should use DVFS first, then CPU packing. Furthermore, we find that the effectiveness of CPU packing is application-dependent: commercial workloads (e.g. Apache) with periods of low utilization can reduce power by as much as 19% via packing, while the improvement to HPC workloads ranges from small to negligible.

An Implementation of Page Allocation Shaping for Energy Efficiency

Matthew E. Tolentino[1], Joseph Turner[2] and Kirk W. Cameron[3]

[1]Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA, USA
matthew.e.tolentino@intel.com

[2]Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA, USA
thanksjamesjoyce@hotmail.com

[3]Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA, USA
cameron@cs.vt.edu

Main memory in many tera-scale systems requires tens of kilowatts of power. The resulting energy consumption increases system cost and the heat produced reduces reli-ability. Emergent memory technologies will provide sys-tems the ability to dynamically turn-on (online) and turn-off (offline) memory devices at runtime. This technology, coupled with slack in memory demand, offers the poten-tial for significant energy savings in clusters of servers. However, to realize these energy savings, OS-level mem-ory allocation and management techniques must be modi-fied to minimize the number of active memory devices while satisfying application demands. We propose sev-eral page shaping techniques and structural enhance-ments to proactively and reactively direct allocations to a minimal number of devices. To evaluate these techniques on real systems, we implemented these shaping tech-niques in the Linux kernel. Experiments using our OS extensions coupled with a simple history-based heuristic (to track demand and control state transitions) yield up to 60% energy savings with less than 1% performance loss for various benchmarks including lmbench and SPEC.

Power, Performance, and Thermal Management for High-Performance Systems

Heather Hanson[1], Stephen W. Keckler[2], Karthick Rajamani[3], Soraya Ghiasi[3], Freeman Rawson[3] and Juan Rubio[3]

[1]*Dept. of Electrical & Computer Engineering*
The University of Texas at Austin
Austin, TX, USA
hhanson@mail.utexas.edu

[2]*Dept. of Computer Sciences*
The University of Texas at Austin
Austin, TX, USA
skeckler@cs.utexas.edu

[3]*Austin Research Laboratory*
IBM
Austin, TX, USA
{karthick, sghiasi, frawson, rubioj}@cs.utexas.edu

In future high-performance systems it will be essential to balance often-conflicting objectives of performance, power, energy, and temperature under variable workload and environmental conditions. In this work, we describe a goal-driven approach that conveys multiple expectations to managers that dynamically tune operating states to best meet those demands. We show the benefit of a concise goal specification for complex objectives and the feasibility of managing multiple constraints while maintaining high performance and safe operation. We evaluate key features of our approach with a prototype implementation on a Pentium M platform with Red Hat Enterprise 4 that controls voltage and frequency scaling to achieve the desired performance, power and temperature goals.

Green Supercomputing in a Desktop Box

Wu-Chun Feng[1], Avery Ching[2] and Chung-Hsing Hsu[3]

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

[2]*Electrical & Computer Engineering*
Northwestern University
Evanston, IL, USA
aching@ece.northwestern.edu

[3]*Computer & Computational Sciences*
Los Alamos National Laboratory
Los Alamos, NM, USA
chunghsu@lanl.gov

The advent of the Beowulf cluster in 1994 provided dedicated compute cycles, i.e., supercomputing for the masses, as a cost-effective alternative to large supercomputers, i.e., supercomputing for the few. However, as the cluster movement matured, these clusters became like their large-scale supercomputing brethren: a shared (and power-hungry) datacenter resource that must reside in a actively-cooled machine room in order to operate properly. The above observation, coupled with the increasing performance gap between the PC and supercomputer, provides the motivation for a green supercomputer in a desktop box. Thus, this paper presents and evaluates such an architectural solution: a 12-node personal desktop supercomputer that offers an interactive environment for developing parallel codes and achieves 14 Gflops on Linpack but sips only 185 watts of power at load, all in the approximate form factor of a Sun SPARCstation 1 pizza box.

Workshop 12
High Performance Grid Computing
HPGC 2007

12 HPGC • Workshop on High Performance Grid Computing

Workshop Description:

Computational grids allow the federation of significant computational and storage resources to solve challenging problems in science, engineering, medicine, finance, and entertainment. Involvement of multi-core platforms and wireless communications in the traditional grids comprised of clusters, workstations, and supercomputers pose new challenges to manage the grids and open new opportunities in using them. 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. Solution of large problems on grids.
- 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.
- Multi-core processors as grid components.
- 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
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
Wolfgang Gentzsch, German D-Grid Initiative
Alexey Lastovetsky, University College Dublin, Ireland
Thuy T. Le, San Jose State University, USA
Xiaolin (Andy) Li, Oklahoma State University, USA
Gabriel Mateescu, National Research Council, Ottawa, Canada
Francois Pellegrini, INRIA and LaBRI, Universite Bordeaux, France
Sushil Prasad, Georgia State University, USA
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 T. Yang, St. Francis Xavier University, Canada
Rob F. Van Der Wijngaart, Intel Corporation, USA

Experiments in running a scientific MPI application on Grid5000

Stephane Genaud[1], Marc Grunberg[2] and Catherine Mongenet[1]

[1]*LSIIT-ICPS*
UMR 7005 CNRS-ULP
Strasbourg, France
{*genaud, mongenet*}*@icps.u-strasbg.fr*

[2]*IPGS*
UMR 7516
Strasbourg, France
grunberg@eost.u-strasbg.fr

Over the last couple of years, several dedicated grid platforms have been set up to test applications and middleware for grids. Among these is Grid'5000, a reconfigurable platform gathering resources at nine remote geographical sites in France. This paper presents one of the eight experiments that have tested software scalability at the scale of a thousand processors (i.e. 500-1000) on this grid testbed. The experiment aims at analyzing the behavior of a geophysical application (a seismic ray tracing in a 3D mesh of the Earth). The application is computationnally intensive but requires an all-to-all communication phase during which processors exchange their results, which has shown to be a real bottleneck on many hardware platforms. We analyze various runs and show that this application scales well up to about 500 processors on such a grid.

Cosmological Simulations using Grid Middleware

Yves Caniou[1], Eddy Caron[1], Benjamin Depardon[1], Hélène Courtois[2] and Romain Teyssier[3]

[1]*LIP - GRAAL*
ENS de Lyon
Lyon, France
{*yves.caniou, eddy.caron,*
benjamin.depardon}*@ens-lyon.fr*

[2]*CRAL*
ENS de Lyon
Lyon, France
courtois@IfA.Hawaii.edu

[3]*CEA*
Gif-sur-Yvette, France
romain.teyssier@cea.fr

Large problems ranging from numerical simulation can now be solved through the Internet using grid middleware. This paper describes the different steps involved to make available a service in the DIET grid middleware. The cosmological RAMSES application is taken as an example to detail the implementation. Furthermore, several results are given in order to show the benefits of using DIET, among which the transparent usage of numerous clusters and a low overhead (finding the right resource and submitting the computing task).

A Parallel Hybrid Method of GMRES on GRID System

Ye Zhang, Guy Bergere and Serge Petiton

Laboratoire dInformatique Fondamentale de Lille
Université des Sciences et Technologies de Lille
Villeneuve D'Ascq, Nord, France
{*Ye.Zhang, bergereg, petiton*}*@lifl.fr*

Grid computing focuses on making use of a very large amount of resources from a large-scale computing environment. It intends to deliver high-performance computing over distributed platforms for computation and data-intensive applications. In this paper, we will present an effective parallel hybrid asynchronous method to solve large sparse linear systems by the use of a Grid Computing platform Grid5000. This hybrid method combines a parallel GMRES(m) (Generalized Minimum RESidual) algorithm with the Least Square method that needs some eigenvalues obtained from a parallel Arnoldi algorithm. All of these algorithms run on the different processors of the platform Grid5000. Grid5000, a 5000 CPUs nation-wide infrastructure for research in Grid computing, is designed to provide a scientific tool for computing. We discuss the performances of this hybrid method deployed on Grid5000, and compare these performances with those on the IBM SP series supercomputers.

Experiments with a Software Component Enabling NetSolve with Direct Communications in a Non-Intrusive and Incremental Way

Xin Zuo and Alexey Lastovetsky

School of Computer Science and Informatics
University College Dublin
Dublin, Ireland
{*xin.zuo, Alexey.Lastovetsky*}*@ucd.ie*

The paper presents a software component that enables NetSolve with direct communications between servers in a non-intrusive and incremental way. Non-intrusiveness means that the software component is supplementary, working on top of the original system, which does not change at all. Increment means that the software component does not have to be installed on all computers to enable applications with the new feature. It can be done incrementally, step by step, and the new feature will be enabled in part, with the completeness dependent on how many nodes have been upgraded with the software component. The paper describes the design and implementation of the software component. The paper also reports on experiments with three typical scientific NetSolve applications having different communication structures: (i) protein tertiary structure prediction, (ii) image processing using sequential algorithms, and (iii) the matrix chain product. The presented experimental results show that the performance of these Grid applications can be easily and significantly improved by using the proposed supplementary software component.

Management of Virtual Machines on Globus Grids Using GridWay

Antonio Juan Rubio-Montero[1], Eduardo Huedo[2], Rubén S. Montero[2] and Ignacio M. Llorente[2]

[1]*Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas*
Madrid, Spain
antonio.rubio@ciemat.es

[2]*Facultad de Informática*
Universidad Complutense de Madrid
Madrid, Spain
ehuedo@fdi.ucm.es, {rubensm, llorente}@dacya.ucm.es

Virtual machines are a promising technology to overcome some of the problems found in current Grid infrastructures, like heterogeneity, performance partitioning or application isolation. In this work, we present an straightforward deployment of virtual machines in Globus Grids. This solution is based on standard services and does not require additional middleware to be installed. Also, we assess the suitability of this deployment in the execution of a high throughput scientific application, the XMM-Newton Scientific Analysis System.

Topaz: Extending Firefox to Accommodate the GridFTP Protocol

Richard Zamudio[1], Daniel Catarino[1], Michela Taufe[1], Brent Stearn[2] and Karan Bhatia[2]

[1]*Dept. of Computer Science*
University of Texas at El Paso
El Paso, TX, USA
{rzamudio, dcatarino1, mtaufer}@utep.edu

[2]*San Diego Supercomputer Center*
University of California San Diego
La Jolla, CA, USA
{flujul, karan}@sdsc.edu

As grid infrastructures mature, an increasing challenge is to provide end-user scientists with intuitive interfaces to computational services, data management capabilities, and visualization tools. One novel approach, being successfully applied in the domain of Computational Chemistry, is to leverage the capabilities of the Mozilla framework to provide rich end-user tools that seamlessly integrate with remote resources such as web grid services and data repositories. The Mozilla framework provides much of the infrastructure to build rich end-user applications, but lacks the capability to integrate with Grid protocols and APIs.

In this paper we present the design and evaluation of Topaz, a Mozilla-based component that provides GridFTP functionality to the popular Firefox browser. Topaz provides end-user scientists with a familiar and user-friendly interface with which to access arbitrary GridFTP servers by providing upload and download functionalities as well as by obtaining and managing users' grid certificates.

A Study of Publish/Subscribe Systems for Real-Time Grid Monitoring

Chenxi Huang[1], Peter R. Hobson[2], Gareth A. Taylor[3] and Paul Kyberd[4]

[1]*Electronic and Computer Engineering*
Brunel University
Uxbridge, Middlesex, UK
chenxi.huang@brunel.ac.uk

[2]*Electronic and Computer Engineering*
Brunel University
Uxbridge, Middlesex, UK
peter.hobson@brunel.ac.uk

[3]*Electronic and Computer Engineering*
Brunel University
Uxbridge, Middlesex, UK
gareth.taylor@brunel.ac.uk

[4]*Electronic and Computer Engineering*
Brunel University
Uxbridge, Middlesex, UK
paul.kyberd@brunel.ac.uk

Monitoring and controlling a large number of geographically distributed scientific instruments is a challenging task. Some operations on these instruments require real-time (or quasi real-time) response which make it even more difficult. In this paper, we describe the requirements of distributed monitoring for a possible future Electrical Power Grid based on real-time extensions to Grid computing. We examine several standards and publish/subscribe middleware candidates, some of which were specially designed and developed for Grid monitoring. We analyze their architecture and functionality, and discuss the advantages and disadvantages. We report on a series of tests to measure their real-time per-formance and scalability.

Implementation of Distributed Loop Scheduling Schemes on the TeraGrid

Satish Penmatsa[1], Anthony T. Chronopoulos[1], Nicholas T. Karonis[2,3] and Brian R. Toonen[2]

[1]*Department of Computer Science*
University of Texas at San Antonio
San Antonio, TX, USA
{spenmats, atc}@cs.utsa.edu

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

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

Grid computing can be used for high performance computations. However, a serious difficulty in concurrent programming of such heterogeneous systems is how to deal with scheduling and load balancing of such systems which may consist of heterogeneous computers on different sites. Distributed scheduling schemes suitable for parallel loops with independent iterations on heterogeneous computer clusters have been proposed and analyzed in the past. Here, we implement the previous schemes in MPICH-G2 and MPIg on the TeraGrid. We present performance results for three loop scheduling schemes on single and multi-site TeraGrid clusters.

Implementing OLAP Query Fragment Aggregation and Recombination for the OLAP Enabled Grid

Michael Lawrence[1], Frank Dehne[2] and Andrew Rau-Chaplin[3]

[1]*Dept. of Computer Science*
University of British Columbia
Vancouver, BC, Canada
mklawren@cs.ubc.ca

[2]*School of Computer Science*
Carleton University
Ottawa, ON, Canada
frank@dehne.net

[3]*Faculty of Computer Science*
Dalhousie University
Halifax, NS, Canada
arc@cs.dal.ca

In this paper we propose a new query processing method for the OLAP Enabled Grid, which blends sophisticated cache extraction techniques and data grid scheduling to efficiently satisfy OLAP queries in a distributed fashion. The heart of our approach is our query Fragment Aggregation and Recombination (FAR) strategy that partitions OLAP queries into subqueries which can be effectively answered by retrieving and aggregating multiple fragments of cached data from nearby grid sources, or as a last resort, more remote backend data warehouses. We have implemented and experimentally evaluated our query processing method and found that our strategy reduces query time between 50% and 60% for practical user cache sizes and network parameters.

GridCopy: Moving Data Fast on the Grid

Rajkumar Kettimuthu[1,2], William Allcock[1,2], Lee Liming[1,2], John-Paul Navarro[1,2] and Ian Foster[1,2,3]

[1]*Mathematics and Computer Science Division*
Argonne National Laboratory
Argonne, IL, USA
{*kettimut, allcock, liming, navarro, foster*}*@mcs.anl.gov*

[2]*Computation Institute*
The University of Chicago
Chicago, IL, USA

[3]*Department of Computer Science*
The University of Chicago
Chicago, IL, USA

An important type of communication in grid and distributed computing environments is bulk data transfer. GridFTP has emerged as a de facto standard for secure, reliable, high-performance data transfer across resources on the Grid. GridCopy provides a simple GridFTP client interface to users and extensible configuration that can be changed dynamically by administrators to make efficient data movement in the Grid easier for users.

Online Grid Replication Optimizers to Improve System Reliability

Ming Lei, Susan V. Vrbsky and Zijie Qi

Computer Science
University of Alabama
Tuscaloosa, AL, USA
{*mlei, vrbsky, zqi*}*@cs.ua.edu*

In a data intensive Grid system, many data replica schemes and models have been proposed to improve the system response time or data consistency, but little attention has been paid to the system reliability. In this paper, we investigate how these data replica schemes will impact the system reliability. We use several metrics of system reliability we previously proposed (System Bytes Missing Rate and System File Missing Rate), and we model the system availability problem assuming limited replica storage and different sized files. In order to achieve higher system data availability, in this paper we propose two new replica optimizers, MinDmr-β and MinDmr-γ, to minimize the Data Miss Rate (MinDmr). A comparison of our replica optimizers using a simulation on the OptorSim demonstrates that by utilizing our strategy, we can increase the data availability for files with different sizes.

Workshop 13
Workshop on Parallel and Distributed Scientific and Engineering Computing PDSEC 2007

13 PDSEC • 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.

PDSEC-07 Workshop Chairs:

Thomas Rauber, University of Bayreuth, Germany
Ruppa K. Thulasiram, University of Manitoba, Canada
Gudula Rünger, Chemnitz University of Technology, Germany
Tao Xie, San Diego State University, USA
Laurence T. Yang, St. Francis Xavier University, Canada
Yi Pan, Georgia State University, USA

Steering Committee:

Richard P. Brent, Oxford University, UK
Jack Dongarra, University of Tennessee and Oak Ridge National Laboratory, USA
Gerhard Joubert, Technical University of Clausthal, Germany
John Gustafson, Sun Microsystems Inc, USA
Yi Pan (Chair), Georgia State University, USA
Laurence T. Yang, St. Francis Xavier University, Canada
Xiaodong Zhang, College of William and Mary and National Science Foundation, USA

Technical Program Review Committee:

Eric Aubanel, University of New Brunswick, Canada
Danier Beer, Chemnitz University of Technology, Germany
Ioana Banicescu, Mississippi State University, USA
Angelos Bilas, University of Crete and FORTH, Greece
Anu Bourgeois, Georgia State University, USA
Martin Buecker, Aachen University of Technology, Germany
Jörg Dümmler, Chemnitz University of Technology, Germany
Vincenzo De Florio, University of Antwerp, Belgium
Len Freeman, University of Manchester, UK
George A. Gravvanis, Democritus University of Thrace, Greece
Marco Hobbel, University of Bayreuth, Germany
Ralf Hoffman, University of Bayreuth, Germany
Michael Hoffman, Chemnitz University of Technology, Germany
Sacha Hunold, University of Bayreuth, Germany
Helen Karatza, Aristotle University of Thessaloniki, Greece
Matthias Korch, University of Bayreuth, Germany
Nectarios Koziris, National Technical University of Athens, Greece
Matthias Kuehnemann, Chemnitz University of Technology, Germany
Raphael Kunis, Chemnitz University of Technology, Germany
Raik Nagel, University of Bayreuth, Germany
John O'Donnell, University of Glasgow, UK
Thomas Rauber, University of Bayreuth, Germany
Gudula Rünger, Chemnitz University of Technology, Germany
Carsten Scholtes, University of Bayreuth, Germany
Michael Schwind, Chemnitz University of Technology, Germany
Parimala Thulasiraman, University of Manitoba, Canada
Ruppa K. Thulasiram, University of Manitoba, Canada
Karen Tomko, University of Cincinnati, USA
Lorenzo Verdoscia, ICAR, Italian National Research Council (CNR), Italy
Tao Xie, San Diego State University, USA
Laurence T. Yang, St. Francis Xavier University, Canada
Bing-Bing Zhou, University of Sydney, Aurstralia

Additional Reviewers:

Daniel Beer, Chemnitz University of Technology, Germany
Jörg Dümmler, Chemnitz University of Technology, Germany
Judith Hippold, Chemnitz University of Technology, Germany
Marco Hoebbel, University of Bayreuth, Germany
Ralf Hoffmann, University of Bayreuth, Germany
Sascha Hunold, University of Bayreuth, Germany
Matthias Korch, University of Bayreuth, Germany
Matthias Kühnemann, Chemnitz University of Technology, Germany
Raphael Kunis, Chemnitz University of Technology, Germany
Raik Nagel, University of Bayreuth, Germany
Carsten Scholtes, University of Bayreuth, Germany
Michael Schwind, Chemnitz University of Technology, Germany
Sven Trautmann, Chemnitz University of technology, Germany

Keynote – Petascale Computing for Large-Scale Graph Problems

David A. Bader

College of Computing
Georgia Institute of Technology
Atlanta, GA, USA

Graph theoretic problems are representative of fundamental kernels in traditional and emerging computational sciences such as chemistry, biology, and medicine, as well as applications in national security. Yet they pose serious challenges for parallel machines due to non-contiguous, concurrent accesses to global data structures with low degrees of locality. Few parallel graph algorithms outperform their best sequential implementation due to long memory latencies and high synchronization costs. In this talk, we consider several graph theoretic kernels for connectivity and centrality and discuss how the features of petascale architectures will affect algorithm development, ease of programming, performance, and scalability.

Implementing and Evaluating Automatic Checkpointing

Antonio S. Martins Jr.[1] and Ronaldo A. L. Goncalves[2]

[1]*Antonio S. Martins Jr.*
State University of Maringa
Maringa, Paraná, Brazil
asmartins@uem.br

[2]*Ronaldo A. L. Goncalves*
State University of Maringa
Maringa, Paraná, Brazil
ronaldo@din.uem.br

As the size and popularity of computer clusters go on growing, fault tolerance is becoming a crucial factor to ensure high performance and reliability for applications. To provide this facility, a checkpoint mechanism is used to recover a failed parallel application rolling it back to an execution moment prior to occurrence of the failure. In this work we present a mechanism for managing checkpoint operations during the failures automatically. This mechanism records periodically the applications context, identifies failed nodes and restarts MPI processes on the remaining nodes, allowing the continuity of the application and taking advantage of the computing accomplished previously. We describe a lot of changes inside source of the LAM/MPI. Experiments with an application for recognizing DNA similarity showed that despite the overhead caused by periodic checkpoints, the benefits can reach about 50% on a small cluster.

United-FS: A Logical File System Providing a Single Image of Multiple Physical File Systems on NFS Server

Huan Chen[1,2], Yi Zhao[1,2], Jin Xiong[1], Jie Ma[1] and Ninghui Sun[1]

[1]National Research Center For Intelligent Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
{huanchen, zhaoyi, xj, majie, snh}@ncic.ac.cn

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

NFS is considered to be the bottleneck in cluster computing environment because of its limited resources and centralized data management. With the development of hardware, NFS server has more than one I/O channel, more storage space and more powerful CPU. In this paper, we describe the design and the implementation of a new logical file system called United-FS. It can make storage devices connected to multiple I/O channels work concurrently and cooperatively. It can be exported by NFS server to provide a single file system image to clients by hiding a variety of native file systems built on different type of storage devices. This paper also compares the United-FS with the Software RAID system both from theoretical analysis and experiments. The results show that United-FS is much more flexible and its performance is better than Software RAID in most cases.

An Energy-Efficient Framework for Large-Scale Parallel Storage Systems

Ziliang Zong[1], Matt Briggs[2], Nick O'Connor[3] and Xiao Qin[4]

[1]Computer Science Department
New Mexico Institute of Mining and Technology
Socorro, NM, U.S.A
zzong@nmt.edu

[2]Computer Science Department
New Mexico Institute of Mining and Technology
Socorro, NM, U.S.A
mbriggs@nmt.edu

[3]Electrical Engineering Department
New Mexico Institute of Mining and Technology
Socorro, NM, U.S.A
nick03d@nmt.edu

[4]Computer Science Department
New Mexico Institute of Mining and Technology
Socorro, NM, U.S.A
xqin@cs.nmt.edu

Huge energy consumption has become a critical bottleneck for further applying large-scale cluster systems to build new data centers. Among various components of a data center, storage subsystems are one of the biggest consumers of energy. In this paper, we propose a novel buffer-disk based framework for large-scale and energy-efficient parallel storage systems. To validate the efficiency of the proposed framework, a buffer-disk scheduling algorithm is designed and implemented. Our algorithm can provide more opportunities for underlying disk power management schemes to save energy by keeping a large number of idle data disks in sleeping mode as long as possible. The trace-driven simulation results based on a revised disksim simulator show that this new framework can significantly improves the energy efficiency of large-scale parallel storage systems.

Porting the GROMACS Molecular Dynamics Code to the Cell Processor

Stephen Olivier[1], Jan Prins[1], Jeff Derby[2] and Ken Vu[2]

[1]*Department of Computer Science*
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
{*olivier, prins*}*@cs.unc.edu*

[2]*Systems and Technology Group*
IBM Corporation
Research Triangle Park, NC, USA
{*jhderby, kenvu*}*@us.ibm.com*

The Cell processor offers substantial computational power which can be effectively utilized only if application design and implementation are tuned to the Cell architecture. In this paper, we examine application characteristics which facilitate efficient use of the Cell processor, and those which present obstacles to it. Moreover, we consider possible solutions designed to mitigate inefficiencies. The target application in our study is the GROMACS molecular dynamics package. We have accelerated the most-often used compute-intensive kernel while maintaining the constraints imposed by the structure of the surrounding program. The significant contribution of this paper is the consideration of the kernel in the context of a complex end-to-end application, with irregular data and code patterns, rather than an isolated kernel code. For this challenging scenario, our results show a 2X speedup versus hand-tuned VMX/SSE code running on high-end PowerPC and x86 uniprocessor machines.

Middleware and Performance Issues for Computational Finance Applications on Blue Gene/L

Thomas Phan[1], Ramesh Natarajan[2], Satoki Mitsumori[3] and Hao Yu[2]

[1]*IBM Almaden Research Center*
San Jose, CA, USA
thomascphan@yahoo.com

[2]*IBM T.J. Watson Research Center*
Yorktown Heights, NY, USA
{*nramesh, hyu*}*@us.ibm.com*

[3]*NIWS Co., Ltd.*
Tokyo, Japan
satoki@us.ibm.com

We discuss real-world case studies involving the implementation of a web services middleware tier for the IBM Blue Gene/L supercomputer to support financial business applications. These programs that are representative of a class of modern financial analytics that take part in distributed business workflows and are heavily database-centric with input and output data stored in external SQL data warehouses. We describe the design issues related to the development of our middleware tier that provides a number of core features, including an automated SQL data extraction and staging gateway, a standardized high-level job specification schema, a well-defined web services (SOAP) API for interoperability with other applications, and a secure HTML/JSP web-based interface suitable for general users. Further, we provide observations on performance optimizations to support the relevant data movement requirements.

A Parallel Algorithmic Approach for Microwave Tomography in Breast Cancer Detection

Meilian Xu[1], Abas Sabouni[2], Parimala Thulasiraman[1], Sima Noghanian[2] and Stephen Pistorius[3]

[1]*Computer Science*
Manitoba
Winnipeg, MB, Canada
{*maryx, thulasir*}*@cs.umanitoba.ca*

[2]*Electrical and Computer Engg.*
Manitoba
Winnipeg, MB, Canada
{*sabouni, sima*}*@ee.umanitoba.ca*

[3]*CancerCare Manitoba*
Manitoba
Winnipeg, MB, Canada
Stephen.Pistorius@cancercare.mb.ca

Different technologies have been used for breast cancer detections clinically. But they have weaknesses in terms of sensitivity and specificity. Microwave imaging technique, on the contrary, uses the apparent dielectric property contrasts between different breast tissues at microwave frequencies and is a prospective direction to find small tumor at their early stage. Microwave tomography falls in one category of microwave imaging technique. There are two main components in microwave tomography to detect abnormalities in breasts: Genetic Algorithm (GA) and Finite-Difference Time-Domain (FDTD). Both GA and FDTD are time-consuming, but, they are data-parallel in nature. In this paper, we have designed a parallel framework for microwave tomography: parallel GA combined with parallel FDTD. The algorithms are implemented on distributed memory machines running MPI. The execution time of the sequential algorithm (GA and FDTD combined) is 10,131 seconds. The total execution time obtained on 16 processors which is approximately 2000 seconds surpasses the sequential algorithm.

Performance evaluation of two parallel programming paradigms applied to the symplectic integrator running on COTS PC cluster

Lorena B. C. Passos[1], Gerson H. Pfitscher[1] and Tarcísio M. Rocha Filho[2]

[1]*Department of Computer Science*
University of Brasilia
Brasilia, DF, Brazil
{*lbrasil, gerson*}*@unb.br*

[2]*Institute of Physics*
University of Brasilia
Brasilia, DF, Brazil
marciano@fis.unb.br

There are two popular parallel programming paradigms available to high performance computing users such as engineering and physics professionals: message passing and distributed shared memory. It is interesting to have a comparative evaluation of these paradigms to choose the most adequate one. In this work, we present a performance comparison of these two programming paradigms using a Computational Physics problem as a case study. The self-gravitating ring model (Hamiltonian Mean Field model) for N particles is extensively studied in the literature as a simplified model for long range interacting systems in Physics. We parallelized and evaluated the performance of a simulation that uses the symplectic integrator to model an N particle system. From the obtained results it is possible to observe that message passing implementation of the symplectic integrator presents better results than distributed shared memory implementation.

Parallel Audio Quick Search on Shared-Memory Multiprocessor Systems

Yurong Chen, Wei Wei and Yimin Zhang

Intel China Research Center
Intel Corporation
Beijing, China
{*yurong.chen, yimin.zhang*}*@intel.com, lzuweiwei@yahoo.com.cn*

Audio search plays an important role in analyzing audio data and retrieving useful audio information. In this paper, a Partially Overlapping Block-Parallel Active Search method (POBPAS) is proposed to perform audio quick search on shared-memory multiprocessor systems (SMPs). This method uses a proper data segmentation to achieve parallelism and performs a high level of parallelism with little additional work. Several techniques including I/O optimization, proper data partition and dynamic scheduling are also introduced to maximize its scalability performance. In addition, we conduct a detailed performance characterization analysis of the parallel implementation of the POBPAS for three data sets on two Intel Xeon SMPs. Experimental results indicate that there are no obvious parallel limiting factors in the implementation except memory bandwidth. As a result, it can achieve 11.3X speedup for a larger data set (searching a 15 seconds clip in a 27 hours audio stream) on the 16-way processor system.

iC2mpi: A Platform for Parallel Execution of Graph-Structured Iterative Computations

Harnish Botadra[1], Qiong Cheng[1], Sushil K. Prasad[1], Eric Aubanel[2] and Virendra Bhavsar[2]

[1]*Computer Science Department*
Georgia State University
Atlanta, Georgia, U.S.A.
{*hbotadra1, qcheng1*}*@student.gsu.edu,*
sprasad@gsu.edu

[2]*Faculty of Computer Science*
University of New Brunswick
Fredericton, New Brunswick, Canada
{*aubanel, bhavsar*}*@unb.ca*

Parallelization of sequential programs is often daunting because of the substantial development cost involved. Previous solutions have not always been successful, partly because many try to address all types of applications. We propose a platform for parallelization of a class of applications that have similar computational structure, namely graph-structured iterative applications. iC2mpi is a unique proof-of-concept prototype platform that provides relatively easy parallelization of existing sequential programs and facilitates experimentation with static partitioning and dynamic load balancing schemes. We demonstrate with various generic application graph topologies that our platform can produce good performance with very little effort. The iC2mpi platform has a good potential for further performance improvements and for extensions to related classes of application domains.

Securing Grid Data Transfer Services with Active Network Portals

Onur Demir, Michael R. Head, Kanad Ghose and Madhusudhan Govindaraju

Department of Computer Science
Binghamton University (SUNY)
Binghamton, NY, USA
{*onur, mike, ghose, mgovinda*}*@cs.binghamton.edu*

Widely available and utilized Grid servers are vulnerable to a variety of threats from Denial of Service (DoS) attacks, overloading caused by flash crowds, and compromised client machines. The focus of our paper is the design, implementation and evaluation of a set of admission control policies that permit the server to maintain sustained throughput to legitimate clients even in the face of such overloads and attacks. We propose several schemes to effectively, and importantly in an automated fashion, deal with these attacks and overloads. We discuss how these schemes can be efficiently implemented on an active network adapter based gateway that controls access to a pool of backend data servers. Performance tests conducted on a system based on a dual-ported active NIC demonstrate that efficient optimization schemes can be implemented on such a gateway to minimize the grid service response time and to improve server throughputs under heavy loads and DoS attacks. Our results, using the GridFTP server available with Globus Toolkit 4.0.1, demonstrate that even in adverse load conditions, the response times can be maintained at a level similar to normal, low-load conditions.

Integrating Performance Tools with Large-Scale Scientific Software

Meng-Shiou Wu[1], Jonathan L. Bentz[1], Fang Peng[1], Masha masha Sosonkina[1], Mark S. Gordon[1,2] and Ricky A. Kendall[3]

[1]*Scalable Computing Laboratory*
Ames Laboratory, U.S. DOE
Ames, Iowa, USA
{*mswu, jnbntz, fangp, masha*}*@scl.ameslab.gov,*
mark@si.fi.ameslab.gov

[2]*Department of Chemistry*
Iowa State University
Ames, Iowa, USA

[3]*National Center for Computational Science*
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
kendallra@ornl.gov

Modern performance tools provide methods for easy integration into an application for performance evaluation. For a large-scale scientific software package that has been under development for decades and with developers around the world, several obstacles must be overcome in order to utilize modern performance tools and explore performance bottlenecks. In this paper, we present our experience in integrating performance tools with one popular computational chemistry package. We discuss the difficulties we encountered and the mechanisms developed to integrate performance tools into this code. With performance tools integrated, we show one of the initial performance evaluation results, and discuss what other challenges we are facing to conduct performance evaluation for large-scale scientific packages.

CRAC: a Grid Environment to Solve Scientific Applications with Asynchronous Iterative Algorithms

Raphaël Couturier and Stéphane Domas

Laboratoire d'Informatique de l'Université de Franche-Comté (LIFC)
Université de Franche-Comté
IUT Belfort, rue Engel Gros, BP 527, 90016 Belfort cedex, France
{*raphael.couturier, stephane.domas*}*@iut-bm.univ-fcomte.fr*

This paper presents CRAC, an environment dedicated to design efficient asynchronous iterative algorithms for a grid architecture. Those algorithms are particularly suited for grid architecture since they naturally allow to overlap communications by computations. Each processor computes its iterations freely without any synchronization with its neighbors. All the characteristics of CRAC are described. A real application using four distant clusters, with a total of 120 processors, shows the interest of this environment and of asynchronous algorithms.

`FEMS`: An Adaptive Finite Element Solver

Alberto Bertoldo

Dept. of Information Engineering
University of Padova
Padova, Italy
cyberto@dei.unipd.it

In this paper we investigate how to obtain high-level adaptivity on complex scientific applications such as Finite Element (FE) simulators by building an adaptive version of their computational kernel, which consists of a sparse linear system solver. We present the software architecture of `FEMS`, a parallel multifrontal solver for FE applications whose main feature is an install-time training phase where adaptation to the computing platform takes place. `FEMS` relies on a simple model-driven mesh partitioning strategy, which makes it possible to perform efficient static load balancing on both homogeneous and heterogeneous machines.

Tera-scalable Fourier Spectral Element Code for DNS of Channel Turbulent Flow at High Reynolds Number

Jin Xu

Physics Division
Argonne National Laboratory
Argonne, IL, USA
jin_xu@anl.gov

Due to the extensive requirement of memory and speed for direct numerical simulation (DNS) of channel turbulence, people can only perform DNS at moderate Reynolds number before. With the fast development of supercomputers, it has become more and more approachable for researchers to perform DNS of turbulence at high Reynolds number. This makes it imperative to consider the development of tera-scalable DNS codes that are capable of fully exploiting these massively parallel machines. In order to achieve this, three parallel models (1D, 2D and 3D domain decompositions) have been implemented and benchmarked. All these models have been successfully ported on BlueGene/L. We have benchmarked these models on BG/L at ANL and BGW at IBM Watson center. Details of these models have been described, discussed and presented in this paper. The optimized model can be used to perform DNS at high Reynolds number in the near future.

Coarse-grain Parallel Execution for 2-dimensional PDE Problems

Georgios Goumas, Nikolaos Drosinos, Vasileios Karakasis and Nectarios Koziris

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

This paper presents a new approach for the execution of coarse-grain (tiled) parallel SPMD code for applications derived from the explicit discretization of 2-dimensional PDE problems with finite-differencing schemes. Tiling transformation is an efficient loop transformation to achieve coarse-grain parallelism in such algorithms, while rectangular tile shapes are the only feasible shapes that can be manually applied by program developers. However, rectangular tiling transformations are not always valid due to data dependencies, and thus requiring the application of an appropriate skewing transformation prior to tiling in order to enable rectangular tile shapes. We employ cyclic mapping of tiles to processes and propose a method to determine an efficient rectangular tiling transformation for a fixed number of processes for 2-dimensional, skewed PDE problems. Our experimental results confirm the merit of coarse-grain execution in this family of applications and indicate that the proposed method leads to the selection of highly efficient tiling transformations.

Synchronous Distributed Load Balancing on Totally Dynamic Networks

Jacques M. Bahi[1], Raphaël Couturier[1] and Flavien Vernier[2]

[1]*Laboratoire d'Informatique de l'Université de Franche-Comté (LIFC)*
Université de Franche-Comté
IUT de Belfort-Montbéliard, BP 527, 90016 Belfort cedex, France
{jacques.bahi, raphael.couturier}@iut-bm.univ-fcomte.fr

[2]*LISTIC - Polytech'Savoie*
Université de Savoie
Domaine Universitaire, BP 80439, 74944 Annecy le Vieux cedex, France
flavien.vernier@univ-savoie.fr

In this paper, first order diffusion load balancing algorithms for totally dynamic networks are investigated. Totally dynamic networks are networks in which the topology may change dynamically. Some edges or nodes can appear, disappear or move during the time. In our previous works on dynamic networks, the dynamism was limited to the edges. The main result of this study consists in proving that the load balancing algorithms reduce the unbalance on arbitrary dynamic networks. Notice that the hypotheses of our result are realistic and that for example the network does not have to be maintained connected. To study the behavior of these algorithms, we compare the load evolution by several simulations.

Load Balancing of Parallel Simulated Annealing on a Temporally Heteogeneous Cluster of Workstations

Soo-Young Lee and Sourabh Moharil

Electrical & Computer Engineering
Auburn University
Auburn, AL, U.S.A
leesooy@eng.auburn.edu, ssmoharil@gmail.com

Simulated annealing (SA) is a general-purpose optimization technique widely used in various combinatorial optimization problems. However, the main drawback of this technique is a long computation time required to obtain a good quality of solution. Clusters have emerged as a feasible and popular platform for parallel computing in many applications. Computing nodes on many of the clusters available today are temporally heterogeneous. In this study, multiple Markov chain (MMC) parallel simulated annealing (PSA) algorithms have been implemented on a temporally heterogeneous cluster of workstations to solve the graph partitioning problem and their performance has been analyzed in detail. Temporal heterogeneity of a cluster of workstations is harnessed by employing static and dynamic load balancing techniques to further improve efficiency and scalability of the MMC PSA algorithms.

A Performance Model of Many-to-One Collective Communications for Parallel Computing

Alexey Lastovetsky and Maureen O' Flynn

Computer Science & Informatics
University College Dublin
Dublin, Ireland
{*alexey.lastovetsky, maureen.oflynn*}*@ucd.ie*

This paper presents a performance model of Many-to-One collective communications for MPI platforms on a switched Ethernet network. The model is based on empirical findings from observation of Many-to-One operations over a wide range of message sizes. The model reflects a significant increase in the execution time for medium-sized messages, persistently observed for different parallel platforms and MPI implementations and not reflected in traditional communication performance models. We also demonstrate that the use of the model can significantly improve the performance of parallel applications, intensively using Many-to-One communications.

Adaptive Distributed Database Replication Through Colonies of Pogo Ants

Sarah Abdul-Wahid, Razvan Andonie, Joseph Lemley, James Schwing and Jonathan Widger

Computer Science Department
Central Washington University
Ellensburg, WA, USA
{*abdulwahids, joelemley, jon.widger*}*@gmail.com*, {*andonie, schwing*}*@cwu.edu*

We address the problem of optimizing the distribution of partially replicated databases over a computer network. Replication is used to increase data availability in the presence of site or communication failures and to decrease retrieval costs by local access if possible. We present a new bio-inspired replication management approach which is adaptive, completely decentralized, and based on swarm intelligence. Each node has the autonomy to start at any time, depending on the internal state of its stored data objects, a redistribution process. "Redistribution" means replicate, create, delete, update, or move data objects to other nodes of the network. The redistribution process is a dynamic load-balancing scheme which runs with lower priority in the background. The system is event-driven, but the learning process is not synchronized with the events.

Mobility of Data in Distributed Hybrid Computing Systems

Philippe Faes, Mark Christiaens and Dirk Stroobandt

Dept. of Electronics and Information Systems
Ghent University
Ghent, Belgium
{*pfaes, mchristi, dstrooba*}*@elis.UGent.be*

In distributed hybrid computing systems, traditional sequential processors are loosely coupled with reconfigurable hardware for optimal performance. This loose coupling proves to be a communication challenge; the processor units cannot efficiently share a physical memory. This paper proposes a distributed shared memory architecture and a method for effective data migration within that shared memory. Data is moved using a novel garbage collection scheme, the dual semispace collector. The new garbage collector and the distributed memory prove to be an effective means of data migration in distributed hybrid computing systems.

Incorporating Latency in Heterogeneous Graph Partitioning

Eric Aubanel and Xiaochen Wu

Faculty of Computer Science
University of New Brunswick
Fredericton, New Brunswick, Canada
{*aubanel, Xiaochen.Wu*}*@unb.ca*

Parallel applications based on irregular meshes make use of mesh partitioners for efficient execution. Some mesh partitioners can map a mesh to a heterogeneous computational platform, where processor and network performance may vary. Such partitioners generally model the computational platform as a weighted graph, where the weight of a vertex gives relative processor performance, and the weight of a link indicates the relative transmission rate of the link between two processors. However, the performance of a network link is typically characterized by two parameters, bandwidth and latency, which cannot be captured in a single weight. We show that taking into account the network heterogeneity of a computational resource can significantly improve the quality of a domain decomposition obtained using graph partitioning. Furthermore, we show that taking into account bandwidth and latency of the network links is significantly better than just considering the former. This work is presented as an extension to the PaGrid partitioner, and includes a model for estimated execution time, which is used as a cost function by the partitioner but could also be used for performance prediction by application-oriented schedulers.

Workshop 14
Performance Modelling, Evaluation, and Optimisation of Parallel and Distributed Systems
PMEO-PDS 2007

14 PMEO-PDS • 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 pose 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:

G. Min, University of Bradford, U.K.

M. Ould-Khaoua, University of Glasgow, U.K.

Publicity Co-Chairs:

Xiaolong Jin, University of Bradford, U.K.
Mirela Sechi Moretti Annoni Notare, Barddal University, Brazil

Program Committee:

K. Al-Begain, Univ. of Glamorgan (UK)
A. Al-Dubai, Napier Univ. (UK)
H. R. Arabnia, Univ. of Georgia (USA)
I. Awan, Univ. of Bradford (UK)
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)
K. Djemame, Univ. of Leeds (UK)
T. El-Ghazawi, George Washington University (USA)
R. Fatoohi, San Jose State University (USA)
E. Gelenbe, Imperial College London (UK)
M. Gueroui, University of Cergy-Pontoise (France) (mogue@prism.uvsq.fr)
X. He, Tennessee Technological Univ. (USA)
R. Ibbett, Univ. of Edinburgh (UK)
S. Jarvis, Univ. of Warwick (UK)
X. Jin, Univ. of Bradford (UK)
H. Karatza, Univ. of Thessaloniki (Greece)
A. Katangur, Texas A&M Univ. (USA)
A. Khonsari, IPM (Iran)
W. Knottenbelt, Imperial College London (UK)
K. Li, State Univ. of New York at New Paltz (USA)
H. Liu, Huazhong Univ. of Science and Technology (CHINA)
S. Loucif, Emirates University, (UAE)
L.M. Mackenzie, Univ. of Glasgow (UK)
Y. Pan, Georgia State Univ. (USA)
D. K. Pradhan, Univ. of Bristol (UK)
X. Qin, New Mexico Inst. of Mining & Technology (USA)
H. Sarbazi-Azad, Sharif Univ. & IPM (Iran)
A. Shahrabi, Glasgow Caledonian Univ. (UK)
E. Song, Huazhong Univ. of Science and Technology (CHINA)
X.H. Sun, Illinois Institute of Technology (USA)
N. Thomas, Univ. of Newcastle (UK)
A. Touzene, Sultan Qaboos Univ. (Oman)
X. Wang, Villanova Univ. (USA)
M. Woodward, Univ. of Bradford (UK)
J. Wu, Florida Atlantic Univ. (USA)
L. Xiao, Michigan State Univ. (USA)
T. Xie, San Diego State University (USA)
C.Z. Xu, Wayne State Univ. (USA)
Z. Xu, Suffolk Univ. (USA)
S. Yan, Univ. of Bradford (UK)
Laurence T. Yang, St Francis Xavier Univ. (CANADA)
X. Zhou, University of Colorado at Colorado Springs (USA)
A. Zomaya, Univ. of Sydney (Australia)

Average-Case Performance Analysis of Online Non-clairvoyant Scheduling of Parallel Tasks with Precedence Constraints

Keqin Li

Department of Computer Science
State University of New York
New Paltz, New York 12561, USA
lik@newpaltz.edu

We evaluate the average-case performance of three approximation algorithms for online non-clairvoyant scheduling of parallel tasks with precedence constraints. We show that for a class of wide task graphs, when task sizes are uniformly distributed in the range $[1..C]$, the online non-clairvoyant scheduling algorithm LL-SIMPLE has an asymptotic average-case performance bound of $M/(M-(3-(1+1/C)^{C+1})C-1)$, where M is the number of processors. For arbitrary probability distributions of task sizes, we present numerical and simulation data to demonstrate the accuracy of our general asymptotic average-case performance bound. We also report extensive experimental results on the average-case performance of online non-clairvoyant scheduling algorithms LL-GREEDY and LS. Algorithm LL-GREEDY has better performance than LL-SIMPLE by using an improved algorithm to schedule independent tasks in the same level. Algorithm LS produces even better schedules due to break of boundaries among levels.

A Probabilistic Approach to Measuring Robustness in Computing Systems

Behdis Eslamnour and Shoukat Ali

Dept. of Electrical and Computer Engineering
University of Missouri-Rolla
Rolla, MO, USA
{ben88, shoukat}@umr.edu

System builders are becoming increasingly interested in robust design. We believe that a methodology for generating robustness metrics will help the robust design research efforts and, in general, is an important step in the efforts to create robust computing systems. The purpose of the research in this paper is to quantify the robustness of a resource allocation, with the eventual objective of setting a standard that could easily be instantiated for a particular computing system to generate a robustness metric. We present our theoretical foundation for a robustness metric and give its instantiation for a particular system.

Dynamic Load Balancing of Unbalanced Computations Using Message Passing

James Dinan[1], Stephen Olivier[2], Gerald Sabin[1], Jan Prins[2], P. Sadayappan[1] and Chau-Wen Tseng[3]

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

[2]*Dept. of Computer Science*
Univ. of North Carolina at Chapel Hill
Chapel Hill, NC, USA
{*olivier, prins*}*@cs.unc.edu*

[3]*Dept. of Computer Science*
Univ. of Maryland at College Park
College Park, MD, USA
tseng@cs.umd.edu

This paper examines MPI's ability to support continuous, dynamic load balancing for unbalanced parallel applications. We use an unbalanced tree search benchmark (UTS) to compare two approaches, 1) work sharing using a centralized work queue, and 2) work stealing using explicit polling to handle steal requests. Experiments indicate that in addition to a parameter defining the granularity of load balancing, message-passing paradigms require additional parameters such as polling intervals to manage runtime overhead. Using these additional parameters, we observed an improvement of up to 2X in parallel performance. Overall we found that while work sharing may achieve better peak performance on certain workloads, work stealing achieves comparable if not better performance across a wider range of chunk sizes and workloads.

Software Tools for Performance Modeling of Parallel Programs

Diego R. Martínez[1], Vicente Blanco[2], Marcos Boullón[1], José C. Cabaleiro[1], Casiano Rodríguez[2] and Francisco F. Rivera[1]

[1]*Dept. Electronics and Computer Science*
University of Santiago de Compostela
Santiago de Compostela, Spain
{*diegorm, marcos, caba, fran*}*@dec.usc.es*

[2]*Dept. Statistics and Computer Science*
La Laguna University
La Laguna, Spain
{*vblanco, casiano*}*@ull.es*

This paper presents a framework based on a user driven methodology to obtain analytical models of MPI applications on parallel systems in a systematic and easy to use way. This methodology consists of two stages. In the first one, instrumentation of the source code is performed using CALL, which is a profiling tool for interacting with the code in an easy, simple and direct way. New features are added to CALL to obtain different performace metrics and store the performance information in XML files. Using this information, an analytical model of the performance behavior is obtained in the second stage by means of R, a language and environment for statistical analysis. The structure of the whole framework is detailed in this paper, and some selected examples are used to show its practical use.

Predicting the Effect on Performance of Container-Managed Persistence in a Distributed Enterprise Application

David A. Bacigalupo[1], James W. J. Xue[1], Simon D. Hammond[1], Stephen A. Jarvis[1], Donna N. Dillenberger[2] and Graham R. Nudd[1]

[1]*Department of Computer Science*
University of Warwick
Coventry, England, UK
{*daveb, xuewj2, sdh, saj, grn*}*@dcs.warwick.ac.uk*

[2]*IBM T.J. Watson Research Centre*
Yorktown Heights, NY, USA
engd@us.ibm.com

Container-managed persistence is an essential technology as it dramatically simplifies the implementation of enterprise data access. However it can also impose a significant overhead on the performance of the application at runtime. This paper presents a layered queuing performance model for predicting the effect of adding or removing container-managed persistence to a distributed enterprise application, in terms of response time and throughput performance metrics. Predictions can then be made for new server architectures - that is, server architectures for which only a small number of measurements have been made (e.g. to determine request processing speed). An experimental analysis of the model is conducted on a popular enterprise computing architecture based on IBM Websphere, using Enterprise Java Bean-based container-managed persistence as the middleware functionality. The results provide strong experimental evidence for the effectiveness of the model in terms of the accuracy of predictions, the speed with which predictions can be made and the low overhead at which the model can be rapidly parameterised.

Experimental Evaluation of Emerging Multi-core Architectures

Abdullah Kayi[1], Yiyi Yao[1], Tarek El-Ghazawi[1] and Greg Newby[2]

[1]*Dept. of Electrical and Computer Engineering*
The George Washington University
Washington, DC, USA
{*apokayi, yyy, tarek*}*@gwu.edu*

[2]*Arctic Region Supercomputing Center*
Fairbanks, AK, USA
newby@arsc.edu

The trend of increasing speed and complexity in the single-core processor as stated in the Moores law is facing practical challenges. As a result, the multi-core processor architecture has emerged as the dominant architecture for both desktop and high-performance systems. Multi-core systems introduce many challenges that need to be addressed to achieve the best performance. Therefore, a new set of benchmarking techniques to study the impacts of the multi-core technologies is necessary. In this paper, multi-core specific performance metrics for cache coherency and memory bandwidth/latency/contention are investigated. This study also proposes a new benchmarking suite which includes cases extended from the High Performance Computing Challenge (HPCC) benchmark suite. Performance results are measured on a Sun Fire T1000 server with six cores and an AMD Opteron dual core system. Experimental analysis and observations in this paper provide for a better understanding of the emerging multi-core architectures.

Optimization and evaluation of parallel I/O in BIPS3D parallel irregular application

Rosa Filgueira[1], David E. Singh[2], Florin Isaila[3], Jesus Carretero[4] and Antonio Garcia Loureiro[5]

[1]Departement of Computer Science
University Carlos III
Madrid, Spain
rosaf@arcos.inf.uc3m.es

[2]Departement of Computer Science
University Carlos III
Madrid, Spain
desingh@arcos.inf.uc3m.es

[3]Departement of Computer Science
University Carlos III
Madrid, Spain
florin@arcos.inf.uc3m.es

[4]Departement of Computer Science
University Carlos III
Madrid, Spain
jcarrete@arcos.inf.uc3m.es

[5]Departement of Electronics and Computer Science
University of Santiago de Compostela
Santiago de Compostela, Spain
antonio@dec.usc.es

This paper presents the optimization and evaluation of parallel I/O for the BIPS3D parallel irregular application, a 3-dimensional simulation of BJT and HBT bipolar devices. The parallel version of BIPS3D employs Metis, a library for partitioning graphs, finite element meshes, or sparse matrices. First, we show how the partitioning information provided by Metis can be used in order to improve the performance of parallel I/O. Second, we propose a novel technique, called Interval Data Grouping (IDG), which exploits the data replication of mesh nodes for optimizing the scheduling of the parallel file operations. Finally, we evaluate the parallel I/O version of BIPS3D for various existing parallel I/O techniques and present an in-depth analysis of the IDG performance.

Modeling of NAMD's Network Input/Output on Large PC Clusters

Nancy Tran[1] and Daniel A. Reed[2]

[1]Dept. of Pathology & Laboratory Medicine
University of North Carolina
Chapel Hill, North Carolina, USA
nancytran@unc.edu

[2]Renaissance Computing Institute
University of North Carolina
Chapel Hill, North Carolina, USA
dan_reed@unc.edu

This study examined the interplay among processor speed, cluster interconnect and file I/O, using parallel applications to quantify interactions. We focused on a common case where multiple compute nodes communicate with a single master node for file accesses. We constructed a predictive model that used time characteristics critical for application performance to estimate the number of nodes beyond which further performance improvement became unattainable. Predictions were experimentally validated with NAMD, a representative parallel application designed for molecular dynamics simulation. Such predictions can help guide decision making to improve machine allocations for parallel codes in large clusters.

A Model and Prototype of a Resource-Efficient Storage Server for High-Bitrate Video-on-Demand

Yung Ryn Choe, Chase Douglas and Vijay S. Pai

Electrical and Computer Engineering
Purdue University
West Lafayette, IN, USA
{*yung, cndougla, vpai*}*@purdue.edu*

This paper presents a mathematical model and a prototype of a resource-efficient storage server for high-bitrate video-on-demand (VoD) applications. Rapid exponential growth of disk capacity enables the storage of high-bitrate VoD streams; however, a server system must be carefully designed to allow those streams to be retrieved from disk and delivered to the network efficiently. Additionally, a cost-effective server should be implemented using only commodity components, such as standard PCs, SATA disks and controllers, and Gigabit Ethernet links.

Previous parallel I/O performance models have been either oversimplified theoretical models that ignore hardware and application characteristics or complex hardware models that consider detailed disk behaviors such as inter-track variations. This paper presents a model between these extremes: detailed enough to account for the rate-based nature of streaming video, the buffering time allowed by the application, and average-case disk hardware characteristics while remaining simple enough to use for algorithm and system design. This paper then describes a prototype storage server designed to serve large video files at the specified bitrates and finds its performance to agree closely with the model (with an average discrepancy of 11% for high-bitrate streams). The system uses up to 8 SATA-300 disks and can simultaneously serve 290 distinct DVD-quality (6 Mbps) streams or 74 distinct HDTV-quality (25 Mbps) streams from disk, achieving an aggregate network throughput of 1.85 Gbps.

Loss Probability of LRD and SRD Traffic in Generalized Processor Sharing Systems

Xiaolong Jin and Geyong Min

Department of Computing
University of Bradford
Bradford, UK
{*x.jin, g.min*}*@brad.ac.uk*

Generalized Processor Sharing (GPS) is an efficient and flexible scheduling mechanism for sharing server capacity and providing differentiated Quality-of-Service (QoS) owing to its appealing properties of fairness, traffic isolation, and work conservation. This paper analytically investigates the loss probabilities of individual traffic flows in GPS systems subject to heterogeneous Long-Range Dependent (LRD) and Short-Range Dependent (SRD) traffic, which have not been studied in the open literature. We derive and validate the closed-form expressions of the loss probabilities of both traffic flows. We then evaluate the effects of Hurst parameter of LRD traffic on the performance of GPS systems in terms of traffic loss probability.

An Adaptive Fault Identification Protocol for an Emergency/Rescue-Based Wireless and Mobile Ad-Hoc Network

Mourad Elhadef, Azzedine Boukerche and Hicham Elkadiki

School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada
{*elhadef, boukerch, elkadiki*}*@site.uottawa.ca*

In this paper, we consider the fault diagnosis problem in MANETs, i.e. the problem of identifying faulty hosts by fault-free ones. The diagnosis scheme that we consider is that based on the comparison approach, where hosts transmit test tasks to their neighbors and the outcomes are compared. By comparing the received outcomes fault-free hosts are able to diagnose the fault status of the network. We propose an adaptive distributed diagnosis algorithm that uses an adaptable spanning tree to disseminate the local diagnosis views throughout the ad-hoc network. The protocol allows all fault-free hosts to correctly identify all faulty ones, and it constitutes a viable addition to existing self-diagnosis protocols.

Distributed Broadcast Scheduling in Mobile Ad Hoc Networks with Unknown Topologies

Guang Tan, Stephen A. Jarvis, James W. J. Xue and Simon D. Hammond

Computer Science
The University of Warwick
Coventry, UK
{*gtan, saj, xuewj2, sdh*}*@dcs.warwick.ac.uk*

Broadcasting is a fundamental communication task in mobile ad hoc networks, and minimizing broadcasting time (or latency) is crucial to the performance of many applications. Extensive studies have been conducted on the minimization of broadcasting time in the context of *radio networks*, which are usually modeled as general graphs. In this paper, we consider how to achieve this goal with distributed algorithms based on a more realistic (and restricted) network model. We propose a randomized algorithm that completes broadcasting in $O(D\log(n/D)+\log^2 n)$ time, where n is the number of nodes in the network and D the eccentricity (maximum distance from the source node to any other node). Compared with a previous optimal algorithm that achieves the same result for general networks, our algorithm obviates the need to know the network eccentricity D beforehand. We also propose a deterministic broadcasting algorithm that works in $O(n)$ time, which is in contrast with the best known result of $O(n\log^2 D)$ for general networks.

A Design and Analysis of a Hybrid Multicast Transport Protocol for the Haptic Virtual Reality Tracheotomy Tele-Surgery Application

Azzedine Boukerche, Haifa Maamar and Abuhossain

PARADISE Research Laboratory
University of Ottawa
Ottawa, ONT, Canada
{*boukerch, abuhoss*}*@site.uottawa.ca, haifaraja@yahoo.fr*

Nowadays, distributed collaborative virtual environments are used in many scenarios such as tele-surgery, gaming, and industrial training, However several challenging issues remain to be resolved before haptic virtual reality based class of applications become a common place. In this paper, we focus upon a tracheotomy tele-surgery application that is based on closely coupled and highly synchronized haptic tasks that require a high- level of coordination among the participants. We also propose a hybrid protocol that is able to satisfy all the collaborative and haptic virtual environment requirements in general and tracheotomy tele-surgery in particular. We discuss our C-HAVE tracheotomy tele-surgery framework and report on the performance results we have obtained to evaluate our protocol using an extensive set of experiments. .

Low-Overhead LogGP Parameter Assessment for Modern Interconnection Networks

Torsten Hoefler[1,2], Andre Lichei[1] and Wolfgang Rehm[1]

[1]*Dept. of Computer Science*
Technical University of Chemnitz
Chemnitz, Saxony, Germany
{*htor, lica, rehm*}*@cs.tu-chemnitz.de*

[2]*Open Systems Laboratory*
Indiana University
Bloomington, Indiana, USA

Network performance measurement and prediction is very important to predict the running time of high performance computing applications. The LogP model family has been proven to be a viable tool to assess the communication performance of parallel architectures. However, non-intrusive LogP parameter assessment is still a very difficult task. We compare well known measurement methods for Log(G)P parameters and discuss their accuracy and network contention. Based on this, a new theoretically exact measurement method that does not saturate the network is derived and explained in detail. Our method only uses benchmarked values instead of computed parameters to compute other parameters to avoid propagation of first-order errors. A methodology to detect protocol changes in the underlying communication subsystem is also proposed. The applicability of our method and the protocol change detection is shown for the low-level API as well as MPI implementations of different modern high performance interconnection networks. The whole method is implemented in the tool Netgauge and it is available as open source to the public.

Performance Modelling of Necklace Hypercubes

Sina Meraji[1,2], Hamid Sarbazi-Azad[1,2] and Ahmad Patooghy[1,2]

[1] IPM School of Computer Science
Tehran, Iran
{meraji, patooghy}@ce.sharif.edu, azad@ipm,ir

[2] Computer Engineering Department
Sharif University of Technology
Tehran, Iran

The necklace hypercube has recently been introduced as an attractive alternative to the well-known hypercube. Previous research on this network topology has mainly focused on topological properties, VLSI and algorithmic aspects of this network. Several analytical models have been proposed in the literature for different interconnection networks, as the most cost-effective tools to evaluate the performance merits of such systems. This paper proposes an analytical perform-ance model to predict message latency in wormhole-switched necklace hypercube interconnection networks with fully adaptive routing. The analysis focuses on a fully adaptive routing algorithm which has been shown to be the most effec-tive for necklace hypercube networks. The results obtained from simulation experiments confirm that the proposed model exhibits a good accuracy under different operating conditions.

Performance Evaluation of A Load Self-Balancing Method for Heterogeneous Metadata Server Cluster Using Trace-Driven and Synthetic Workload Simulation

Bin Cai[1], Changsheng Xie[1] and Guangxi Zhu[2]

[1] Department of Computer Science and Technology,
Wuhan National Laboratory for Optoelectronics,
Huazhong University of Science and Technology.
Wuhan, Hubei, China.
hust_caibin@sohu.com, csxie@263.net

[2] Department of Electronics and Information Engineering,
Wuhan National Laboratory for Optoelectronics,
Huazhong University of Science and Technology.
Wuhan, Hubei, China.
gxzhu@mail.hust.edu.cn

In cluster-based storage systems, the metadata server cluster must be able to adaptively distribute responsibility for metadata to maintain high system performance and long-term load balance, due to workload skew and metadata servers heterogeneity. In this paper, we describe a simple and adaptive metadata load management scheme, called Self-Balancing Uniform (SBU) randomization, to efficiently and continually adapt the metadata distribution to current demands in heterogeneous metadata server cluster. We implement our system within a discrete event driven simulation environment, along with two other systems, simple randomization (SR) and performance aware distribution (PAD) to serve as points of comparison, and evaluate the performance of our SBU algorithms against SR and PAD algorithms by both a trace workload and a synthetic workload. Simulation results verify that our SBU algorithm achieves load self-balance, provides consistent response latencies and resource utilization. Simulation results also indicate that SR cannot cope with skew and heterogeneity and PAD requires a larger shared state to achieve optimal performance.

Evaluating the Performance of Adaptive Fault-Tolerant Routing

Farshad Safaei[1,3], Ahmad Khonsari[2,1], Mahmood Fathy[3], Amir Hossein Shantia[4] and Mohamed Ould-Khaoua[5]

[1]*IPM School of Computer Science*
IUST
Tehran, Tehran, Iran
safaei@ipm.ir

[2]*Dept. of Electrical and Computer Engineering,*
University of Tehran
Tehran
Tehran, Tehran, Iran
ak@ipm.ir

[3]*Dept. of Computer Engineering, Iran University of Science and Technology*
IUST
Tehran, Tehran, Iran
mahfathy@iust.ac.ir

[4]*Islamic Azad University, North Branch*
Azad
Tehran, Tehran, Iran
shantia_amirhosein@computer.org

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

One of the fundamental problems in parallel computing is how to efficiently perform routing in a faulty network each component of which fails with some probability. This paper presents a comparative performance study of ten prominent adaptive fault-tolerant routing algorithms in wormhole-switched 2-D mesh interconnect networks. These networks carry a routing scheme suggested by Boppana and Chalasani as an instance of a fault-tolerant method. The suggested scheme is widely used in the literature to achieve high adaptivity and support inter-processor communications in parallel computer systems due to its ability to preserve both communication performance and fault-tolerant demands in these networks. The performance measures studied are the throughput, average message latency and average usage of virtual channels per node. Results obtained through simulation suggest two classes of presented routing schemes as high performance candidate in most faulty networks.

Message Routing and Scheduling in Optical Multistage Networks using Bayesian Inference method on AI algorithms

Ajay K Katangur[1] and Somasheker Akaladevi[2]

[1]*Department of Computing Sciences*
Texas A&M University - Corpus Christi
Corpus Christi, TX, USA
ajay.katangur@tamucc.edu

[2]*Department of Computer Information Systems*
Virginia State University
Petersburg, VA, USA
sakkaladevi@vsu.edu

Optical Multistage Interconnection Networks (MINs) suffer from optical-loss during switching and crosstalk problem in the switches. The crosstalk problem is solved by routing messages using time division multiplexing (TDM) approach. This paper focuses on minimizing the number of groups (time slots) required to realize a permutation. Many researchers concentrated on this NP-hard problem and concluded that AI algorithms perform better than the heuristic algorithms. They also showed that majority of the times the performance of Genetic Algorithm (GA) was better than Simulated Annealing Algorithm (SAA). In this research, we implement a new approach to minimize the number of passes required for scheduling a given permutation. A combinational method is developed which comprises the use of Bayesian inference method on GA and SAA to always guarantee the best solution, instead of only using either GA or SAA. Simulations are performed in java using multiple threads to run SA and GAA in parallel and to evaluate the performance of the new method. The results are then compared to those obtained from GA and SAA.

Workshop 15
Dependable Parallel, Distributed and Network-Centric Systems
DPDNS 2007

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:

Bruno Ciciani, University of Roma, Italy

Steering Committee:

D. Avresky (Chair)
E. Maehle

Program Committee:

M. Atighetchi, BBN Technologies, Cambridge, USA
G. Deconinck, University of Leuven, Belgium
A. Doering, IBM Research Zurich, Switzerland
S. Geoghegan, University of Arkansas at Little Rock, USA
K.-E. Grosspietsch, Fraunhofer AIS, Germany
K. Kanoun, LAAS-CNRS, France
T. Kikuno, Osaka University, Japan
M. Malek, Humboldt University, Germany
N. Neogi, University of Illinois at UC, USA
E. Nett, University of Magdeburg, Germany
D. Nikolos, University of Patras, Greece
F. Quaglia, University of Roma, Italy
P. Romano, University of Roma, Italy
M. Roy, LAAS-CNRS, France
J. G. Silva, University of Coimbra, Portugal
P. Sobe, University of Luebeck, Germany
C. Trinitis, TU Munich, Germany
M. Vouk, NC State University, USA

Recent Advances in Trusted Grids and Peer-to-Peer Computing Systems

Kai Hwang

Dept of EE System
University of Southern California
Los Angeles, CA, USA
kaihwang@usc.edu

Computational Grids and peer-to-peer (P2P) are emerging as two of the most promising distributed computing technologies that may change the world in the next decade. In this talk, Dr. Hwang presents recent advances in network security technologies, cyber trust systems, and integrated solutions for trusted computing over the Internet. The talk covers the integration of web services with P2P Grid computing, new cybertrust models, Internet worm containment, P2P reputation systems, and hybrid defense systems to protect distributed resources from network worms, DDoS attacks or peer intrusions or collusions. Research findings and benchmark results from the USC GridSec project will be reported for automated trust management to facilitate security binding and defense against worms and DDoS attacks in Grids, P2P systems, and web services. He will assess frontier research topics on fast reputation aggregation for trusted P2P file sharing, security-aware Grid job scheduling, game-theoretic modeling of non-cooperative Grids, new performance metrics, and DETER experiments for cybertrust development. The fortified Grids, P2P systems, and Internet resources will benefit many security-sensitive applications in digital government, e-commerce, distance learning, distributed supercomputing, etc.

A Framework for Experimental Validation and Performance Evaluation in Fault Tolerant Distributed System

Hein Meling

Dept. of Electrical Engineering and Computer Science
University of Stavanger
Stavanger, Norway
hein.meling@uis.no

Performing experimental evaluation of fault tolerant distributed systems is a complex and tedious task, and automating as much as possible of the execution and evaluation of experiments is often necessary to test a broad spectrum of possible executions of the system to obtain good coverage. The confidence of the results obtained from an experimental evaluation depends on the degree of control over the environment in which experiments are being executed. Typically, an uncontrolled environment is exposed to numerous sources of external influence that can affect the obtained results. Automated and repeated executions can be used to reduce the impact of such influences.

In this paper, a framework for experimental validation and performance evaluation of fault management in a fault tolerant distributed system is presented. The framework provides a facility to execute experiments in a configured target system. It is based on injecting faults or other events needed to test the fault handling capability of the system. Relevant events are logged and collected for post-processing and analysis, e.g. to construct a single global timeline of events occurring at different nodes in the target system. This timeline of events can then be used to validate the behavior a system, and to evaluate its performance.

Dependability Modeling and Analysis in Dynamic Systems

Salvatore Distefano and Antonio Puliafito

Department of Mathematics
University of Messina
Messina, Italy
salvatdi@ingegneria.unime.it, apuliafito@unime.it

Dependability evaluation is an important, often indispensable, step in (critical) systems design and analysis processes. The introduction of control and/or computing systems to automate processes increases the overall system complexity and therefore has an impact in terms of dependability. When a system grows, dynamic effects, not present or manifested before, could arise or become significant in terms of reliability/availability: the system could be affected by common cause failures, the system components could interfere, effects due to load sharing arise and therefore should be considered. Moreover it is of interest to evaluate redundancy and maintenance policies. In those cases it is not possible to recur to notations as reliability block diagrams (RBD), fault trees (FT) or reliability graphs (RG) to represent the system, since the statistical independence assumption is not satisfied. Also more enhanced formalisms as dynamic FT (DFT) could result not adequate to the goal. To overcome those problems we developed a new formalism derived from RBD: the dynamic RBD (DRBD). In this paper we explain how to use the DRBD notation in system modeling and analysis, coming inside a methodology that, starting from the system structure, drives to the overall system availability evaluation following modeling and analysis phases. To do this we use an example drawn from literature consisting of a multiprocessor distributed computing system, also comparing our approach with the DFT one.

A Combinatorial Analysis of Distance Reliability in Star Network

Xiaolong Wu[1], Shahram Latifi[2] and Yingtao Jiang[3]

[1]*Department of Electrical and Computer Engineering*
University of Nevada, Las Vegas
Las Vegas, NV, USA
xiaolong@egr.unlv.edu

[2]*Department of Electrical and Computer Engineering*
University of Nevada, Las Vegas
Las Vegas, NV, USA
latifi@egr.unlv.edu

[3]*Department of Electrical and Computer Engineering*
University of Nevada, Las Vegas
Las Vegas, NV, USA
yingtao@egr.unlv.edu

This paper addresses a constrained two-terminal reliability measure referred to as Distance Reliability (DR) between the source node u and the destination node I with the shortest distance, in an n-dimensional star network, Sn. The shortest distance restriction guarantees the optimal communication delay between processors and high link/node utilization across the network. This paper uses a combinatorial approach by limiting the number of node, link and node/link failures. For each failure model, two different cases depending on the relative positions of u and I, are analyzed to compute DR. Furthermore, DR for the antipodal communication, where every node must communicate with its antipode, is investigated as a special case. For this case, lower bound on DR of those disjoint paths is also derived.

ABARIS: An Adaptable Fault Detection/Recovery Component Framework for MPIs

Hideyuki Jitsumoto[1], Toshio Endo[2] and Satoshi Matsuoka[2,,3]

[1]*Matematical and Computing Science*
Tokyo Institute of Technology
Meguro-ku, Tokyo, JAPAN
jitsumo0@is.titech.ac.jp

[2]*Global Scientific Information and Computing Center*
Tokyo Institute of Technology
Meguro-ku, Tokyo, JAPAN
endo@gsic.titech.ac.jp, matsu@is.titech.ac.jp

[3]*National Institute of Informatics*
Chiyoda-ku, Tokyo, JAPAN

Long-running MPI applications on clusters and grids that are prone to node and network failures, motivates the use of fault tolerant MPI implementations. However, previous fault tolerant MPIs lack the ability to allow the user to easily choose appropriate fault recovery strategies according to the execution environment, independent of the application codes—rather, the user often had to hard-code restoration strateties in accordance to diverse sets of fault patterns, which could be numerous: for instance, if the fault is transient to a particular process, we merely have to restart the process on the same computing node; on the other hand, if the fault is due to repetitive hardware unreliability, we must migrate the process to a new node in its recovery. ABARIS is our new Fault/Recovery model aware component framework for MPI, where users can customize MPI fault detection and recovery algorithms according to their application and execution environmental requirements by merely selecting appropriate fault/recovery components, independent of the application code. Currently, the ARABIS framework prototype is implemented on top of MPICH-P4MPD. Preliminary evaluation of the prototype using NPB on our MPI fault simulator demonstrates that overhead compared to the original MPICH-P4MPD is almost negligible (less than 1%) under normal execution, and when faults occur, appropriate selections and pairings of fault model and recovery method components for corresponding to the execution environment is significant to the overall execution time.

Self Adaptive Application Level Fault Tolerance for Parallel and Distributed Computing

Zizhong Chen[1], Ming Yang[1], Guillermo Francia[1] and Jack Dongarra[2]

[1]*MCIS Department*
Jacksonville State University
Jacksonville, AL, USA
{zchen, myang, gfrancia}@jsu.edu

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

Most application level fault tolerance schemes in literature are non-adaptive in the sense that the fault tolerance schemes incorporated in applications are usually designed without incorporating information from system environments such as the amount of available memory and the local or network I/O bandwidth. However, from an application point of view, it is often desirable for fault tolerant high performance applications to be able to achieve high performance under whatever system environment it executes with as low fault tolerance overhead as possibile.

In this paper, we demonstrate that, in order to achieve high reliability with as low performance penalty as possible, fault tolerant schemes in applications need to be able to adapt themselves to different system environments. We propose a framework under which different fault tolerant schemes can be incorporated in applications using an adaptive method. Under this framework, applications are able to choose near optimal fault tolerance schemes at run time according to the specific characteristics of the platform on which the application is executing.

The Design and Implementation of Checkpoint/Restart Process Fault Tolerance for Open MPI

Joshua Hursey[1], Jeffrey M. Squyres[2], Timothy I. Mattox[1] and Andrew Lumsdaine[1]

[1]*Open Systems Laboratory*
Indiana University
Bloomington, IN, USA
{*jjhursey, timattox, lums*}*@osl.iu.edu*

[2]*Server Virtualization Business Unit*
Cisco Systems, Inc.
San Jose, CA, USA
jsquyres@cisco.com

To be able to fully exploit ever larger computing platforms, modern HPC applications and system software must be able to tolerate inevitable faults. Historically, MPI implementations that incorporated fault tolerance capabilities have been limited by lack of modularity, scalability and usability. This paper presents the design and implementation of an infrastructure to support checkpoint/restart fault tolerance in the Open MPI project. We identify the general capabilities required for distributed checkpoint/restart and realize these capabilities as extensible frameworks within Open MPI's modular component architecture. Our design features an abstract interface for providing and accessing fault tolerance services without sacrificing performance, robustness, or flexibility. Although our implementation includes support for some initial checkpoint/restart mechanisms, the framework is meant to be extensible and to encourage experimentation of alternative techniques within a production quality MPI implementation.

Intelligent Dynamic Network Reconfiguration

Juan Ramon Acosta and Dimiter Avresky

Dept. of Electrical and Computer Engineering
Northeastern University
Boston, MA, USA
{*jracosta, avresky*}*@ece.neu.edu*

Dynamic network reconfiguration is a technique in which the routing tables of the nodes in the vicinity of a failure are updated in real-time. The technique has been proved effective only if no failures occur after the reconfiguration process has started.

This paper, presents enhancements to Agent NetReconf to allow it tolerate new failures if the reconfiguration was already started for a different failure. Agent NetReconf is an intelligent dynamic network reconfiguration algorithm. The improvements were made on the following three phases: Restoration Tree Construction (Phase 1), Multiple Failures synchronization (Phase 2) and Routing Information Update (Phase 3). The proposed strategy consists of: 1) Activate Agent NetReconf recursively, if a new node/link failure occurs and the reconfiguration of a different failure was started, 2) Use a pair of gateway nodes to help the restoration leaders, to reach consensus and to define the order in which each leader will execute the reconfiguration. The complexity, in terms of the number agents created, is analyzed for all phases. Termination is also proved for all phases.

Distributed Interval Voting with Node Failures of Various Types

Behrooz Parhami

Department of Electrical and Computer Engineering
University of California
Santa Barbara, CA, USA
parhami@ece.ucsb.edu

Intervals constitute one of the most important tools for dealing with uncertainty in computations. Researchers in the fields of interval arithmetic and constraint propagation have devised elaborate methods for computing with interval variables. In this interpretation, an interval represents the proposition: I dont know what the correct value is, but it cannot be outside this range. However, intervals also have another use, which is captured in the statement: Any value in this range would be fine with me. In devising voting schemes for data fusion and fault-tolerant distributed computation, these two meanings, and a number of other lesser known variations, must be completely understood in order to design and implement meaningful voting strategies. Irregularities and paradoxes in voting schemes, extensively studied by mathematicians and social scientists, must also be taken into account to avoid serious pitfalls. In this paper, we discuss the two interpretations of interval voting, along with their practical implications, and show how voting strategies differ in their time and communication complexities, performance, and resilience according to the meaning intended and the types of failure assumed.

Fault-Tolerant Earliest-Deadline-First Scheduling Algorithm in Uniprocessor Embedded Systems

Hakem Beitollahi[1], Seyed Ghassem Miremadi[2] and Geert Deconinck[3]

[1]*Electrical Engineering*
Katholieke Universiteit Leuven
Leuven, Leuven, Belgium
Hakem.Beitollahi@esat.kuleuven.be

[2]*Computer Engineering*
Sharif University
Tehran, Tehran, Iran
Miremadi@sharif.edu

[3]*Electrical Engineering*
Katholieke Universiteit Leuven
Leuven, Leuven, Belgium
Geert.Deconinck@esat.kuleuven.be

The general approach to fault tolerance in uniprocessor systems is to maintain enough time redundancy in the schedule so that any task instance can be re-executed in presence of faults during the execution. In this paper a scheme is presented to add enough and efficient time redundancy to the Earliest-Deadline-First (EDF) scheduling policy for periodic real-time tasks. This scheme can be used to tolerate transient faults during the execution of tasks. We describe a recovery scheme which can be used to re-execute tasks in the event of transient faults and discuss conditions that must be met by any such recovery scheme. For performance evaluation of this idea a tool is developed.

IntraCache: An Interest group-based P2P Web Caching System

Huifang Cheng, Zhimin Gu and Junchang Ma

Department of Computer Science and Engineering
Beijing Institute of Technology
Beijing, China
chenghuifang@126.com

An interest group-based P2P browser cache collaborative system, named IntraCache, is proposed in the paper. IntraCache is scalable, resilient to node failures and easy to manage nodes. In the system, the peers with similar interest are organized into autonomous group by PB grouping method and documents are located by similarity based search method. Trace-driven experiments show that PB-Grouping method can utilize local browser cache more efficiently than previous grouping methods. Even if using small cache size, PB grouping method can get preferable hit ratio. Moreover, the interest group-based search method can more efficiently prune the P2P search space and reduce the latency than previous search methods.

Availability/Consistency Balancing Replication Model

Johannes Osrael, Lorenz Froihofer and Karl M. Goeschka

Institute of Information Systems
Vienna University of Technology
Vienna, Austria
{*johannes.osrael, lorenz.froihofer, karl.goeschka*}*@tuwien.ac.at*

Replication combined with explicit management of data integrity constraints can be used to enhance availability of object-oriented, data-centric distributed systems when node and link failures occur. Our approach enhances availability by temporarily relaxing non-critical data integrity constraints during degraded situations. This requires new kinds of optimistic replication protocols that support the configuration of this trade-off. The contribution of this paper is a replication model called Availability/Consistency Balancing Replication Model that allows replicas to diverge in degraded situations if data integrity can be temporarily relaxed and re-establishes both replica consistency and data integrity during repair time. The Primary-per-Partition-Protocol and Adaptive Voting are two concrete protocols following our model. The feasibility of our approach has been shown by several prototype implementations.

Combining Compression, Encryption and Fault-tolerant Coding for Distributed Storage

Peter Sobe[1] and Kathrin Peter[2]

[1]*Institute of Computer Engineering*
University of Luebeck
Luebeck, Germany
sobe@iti.uni-luebeck.de

[2]*Computer Science Research*
Zuse Institute Berlin
Berlin, Germany
kathrin.peter@zib.de

Storing data in distributed systems aims to offer higher bandwidth and scalability than storing locally. But, a couple of disadvantageous issues must be taken into account such as unreliability caused by faults, temporal downtimes and malicious attacks. To improve dependability, redundancy codes like parity can be used as well as more sophisticated codes such as Reed/Solomon. Another issue - security requirements - arise when data is kept in untrusted units in a network. To encrypt data, it is common to use security algorithms like AES. For efficient transfer and storage, the amount of data can be reduced by compression algorithms. All these techniques - data distribution, fault-tolerant coding, encryption and compression - can be employed together using independent algorithms, but in a proper combination. A superposition of these techniques exploiting synergies is still an issue for research. Thus, in this paper we study proper technique combinations applied to distributed storage. The combinations are classified and examined with respect to their potential benefit and limitations. For our model, performance parameters from the distributed storage system NetRAID are used.

Workshop 16
International Workshop on Security in Systems and Networks
SSN 2007

• 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 trust 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:

Li Xiao, Michigan State University, USA

Program Committee:

Kevin Butler, Pennsylvania State University, USA
David Chadwick, University of Salford, UK
Songqing Chen, George Mason University, USA
Wenliang Du, Syracuse University, USA
Yong Guan, Iowa State University, USA
Minaxi Gupta, Indiana University, USA
Anca Ivan, IBM T. J. Watson Research Center, USA
James B. D. Joshi, University of Pittsburgh, USA
Alex X. Liu, Michigan State University, USA
Donggang Liu, University of Texas at Arlington, USA
Geyong Min, University of Bradford, UK
Vassilis Prevelakis, Drexel University, USA
Jian Ren, Michigan State University, USA
Chik How Tan, Gjovik University College, Norway
Wietse Venema, IBM T.J. Watson Research Center, USA
Haining Wang, College of William and Mary, USA
Jianbin Wei, South Dakota School of Mines, USA
Greg B. White, The University of Texas at San Antonio, USA
S. Felix Wu, University of California at Davis, USA
Ye Xia, University of Florida, USA
Bin Xiao, Hong Kong Polytechnic University
Dongyan Xu, Purdue University, USA
Dong Xuan,The Ohio State University, USA
Ossama Younis, University of Arizona, USA
Sheng Zhong, State University of New York at Buffalo, USA
Sencun Zhu, Pennsylvania State University, 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

Transaction Based Authentication Scheme for Mobile Communication: A Cognitive Agent Based Approach

B. Sathish Babu and Pallapa Venkataram

Electrical Communication Engineering
Indian Institute of Science
Bangalore, Karnataka, India
{*bsb, pallapa*}*@ece.iisc.ernet.in*

The vulnerable air interface, device level constraints, and insecure encryption techniques of wireless networks have naturally increased the chance of attacker obtaining users information fraudulently. Most of the existing authentication systems for mobile communication principally depends on the strength of authenticating identifiers. Once the client who may be genuine or an attacker, successfully proves the possession of the identifiers the system accepts all the transactions of a session under single risk level, which is the most important point of vulnerability. We propose a novel Transaction Based Authentication Scheme(TBAS) for mobile communication using cognitive agents. The proposed approach intensifies the procedure of authentication by deploying authentication scheme based on the transaction sensitivity and client's transaction time behaviors. The TBAS provides effective authentication solution, by relieving the conventional authentication systems, from being dependent on only the strength of authentication identifiers. Additionally the transaction time behavior analysis by cognitive agents provides rational approach towards establishing the legitimacy or illegitimacy of the mobile client. The method has been simulated with different applications over in-house established wired and wireless networks. The Agent Factory framework is used for cognitive agents generation and communication. The simulation results are quite encouraging.

An Approach to Detect Executable Content for Anomaly Based Network Intrusion Detection

Like Zhang[1] and Gregory B. White[2]

[1]*Computer Science*
University of Texas at San Antonio
San Antonio, TX, U.S.A.
lzhang@cs.utsa.edu

[2]*Computer Science*
University of Texas at San Antonio
San Antonio, TX, U.S.A.
greg.white@utsa.edu

Since current internet threats contain not only malicious codes like Trojan or worms, but also spyware and adware which do not have explicit illegal content, it is necessary to have a mechanism to prevent hidden executable files downloading in the network traffic. In this paper, we present a new solution to identify executable content for anomaly based network intrusion detection system (NIDS) based on file byte frequency distribution. First, a brief introduction to application level anomaly detection is given, as well as some typical examples of compromising user computers by recent attacks. In addition to a review of the related research on malicious code identification and file type detection in section 2, we will also discuss the drawback when applying them for NIDS. After that, the background information of our approach is presented with examples, in which the details of how we create the profile and how to perform the detection are thoroughly discussed. The experiment results are crucial in our research because they provide the essential support for the implementing. In the final experiment simulating the situation of uploading executable files to a FTP server, our approach demonstrates great performance on the accuracy and stability.

Security Threat Prediction in a Local Area Network Using Statistical Model

Somak Bhattacharya and S K Ghosh

School of Information Technology
Indian Institute of Technology, Kharagpur
Kharagpur, West Bengal, India
somakb@sit.iitkgp.ernet.in, skg@iitkgp.ac.in

In todays large and complex network scenario vulnerability scanners play a major role from security perspective by proactively identifying the known security problems or vulnerabilities that exists across a typical organizational network. Identifying vulnerabilities before they can be exploited by malicious user often helps to test, maintain, and assess the risk of the existing network. Still there are many problems with currently available state of the art vulnerability scanners like hampering system resource. One possible solution to this problem might be reducing the number of vulnerability scans, along with the quantitative approach towards different vulnerability category in order to identify which class of vulnerability should enjoy preference in the risk mitigation procedure. This paper introduces a model that predicts vulnerabilities that will occur in near future on a Local Area Network (LAN) by using statistical measures and vulnerability history data. Two case studies have also been presented to validate the model.

Distributed IDS using Reconfigurable Hardware

Ashok Kumar Tummala[1] and Parimal Patel[2]

[1]*Department of Electrical Engineering*
University of Texas at San Antonio
San Antonio, TX, USA
ashoktummala@yahoo.com

[2]*Department of Electrical Engineering*
University of Texas at San Antonio
San Antonio, TX, USA
parimal.patel@utsa.edu

With the rapid growth of computer networks and network infrastructures and increased dependency on the internet to carry out day-to-day activities, it is imperative that the components of the system are secured. In the last few years a number of Intrusion Detection Systems (IDS) have been developed as network security tools, both in commercial and academic sectors. While considerable progress has been made in the areas of string matching, header processing and detecting DoS attacks at network level, complete systems have not yet been demonstrated that provide all of the functionality necessary to perform intrusion detection at each host system there by securing the entire network. In this paper we are proposing the architecture of a Distributed Intrusion Detection System (DIDS) for use in high-speed networks. The proposed DIDS has Host IDS component at each host that combines the above-mentioned functionalities along with the capability of collecting the events at the application level to look for any signs of intrusion at the network level. DIDS consists of Central IDS component which performs sophisticated processing to detect any signs of distributed attacks on the entire network and update rules in each host system.

For high speed networks it can be difficult to keep up with intrusion detection using purely software approach without affecting performance of the system intended for designed application. It is essential to use hardware systems or software with hardware accelerators. The proposed DIDS is a custom hardware implemented on Field Programmable Gate Arrays (FPGAs). This move to customized hardware-based systems allows the introduction of higher degree of parallelism than might be possible in software at a reasonable cost. The key aspects of this system are flexibility and partial reconfigurability. The nature of future attacks to the Internets infrastructure is difficult to predict, and partial reconfigurability feature of FPGA will allow the system to be adapted to a constant change allowing the system to adapt to new threats.

ESSTCP: Enhanced Spread-Spectrum tcp

Amir R. Khakpour and Hakima Chaouchi

LOR Department
Institut National des Telecommunications
Evry, France
amir.khakpour@int-evry.fr

Having stealth and lightweight authentication methods is empowering network administrators to shelter critical services from adversaries. Spread-Spectrum TCP (SSTCP) is one of these methods by which the client sends an authentic sequence of SYN packets to the server for authentication. Since SSTCP have some certain drawbacks and security flaws, we propose an enhanced version of SSTCP (ESSTCP) which modifies the original algorithm to reduce the computational cost and cover its vulnerabilities from denial of service and replay attacks. Some performance problems like time synchronization are also resolved. We finally try to extend the functionality of this method for different applications and numbers of users by which ESSTCP can be performed as a secure Remote Procedure Call (RPC).

On the Security of Ultrasound as Out-of-band Channel

Rene Mayrhofer and Hans Gellersen

Computing Department
Lancaster University
Lancaster, UK
{*rene, hwg*}*@comp.lancs.ac.uk*

Ultrasound has been proposed as out-of-band channel for authentication of peer devices in wireless ad hoc networks. Ultrasound can implicitly contribute to secure communication based on inherent limitations in signal propagation, and can additionally be used explicitly by peers to measure and verify their relative positions. In this paper we analyse potential attacks on an ultrasonic communication channel and peer-to-peer ultrasonic sensing, and investigate how potential attacks translate to application-level threats for peers seeking to establish a secure wireless link. Based on our analysis we propose a novel method for authentic communication of short messages over an ultrasonic channel.

PCPP: On Remote Host Assessment via Naïve Bayesian Classification

Thomas H. Morris[1] and V. S. S. Nair[2]

[1]*High Assurance Computing and Networking Lab (HACNet)*
Southern Methodist University
Dallas, TX, USA
tmorris@engr.smu.edu

[2]*High Assurance Computing and Networking Lab (HACNet)*
Southern Methodist University
Dallas, TX, USA
nair@engr.smu.edu

Private Computing on a Public Platform (PCPP) is a new paradigm in public computing in which an application executes on a previously unknown remote system securely and privately. The first step in the PCPP process is remote assessment of a prospective remote host to determine whether it is capable of executing the PCPP application and to classify the host as a potential threat or non-threat. This paper explores the use of a Naive Bayesian classifier to classify prospective remote hosts. We show that the Naïve Bayesian classifier learns to recognize subtle patterns in historical host measurements and performs the classification task accurately and with minimal negative performance implications.

A Scenario-Based Protocol Checker for Public-Key Authentication Scheme

Takamichi Saito

Computer Science
Meiji University
Kawasaki, Kanagawa, Japan
saito@cs.meiji.ac.jp

Communication security depends on security protocol such like Secure SHell or Secure Socket Layer. One of the important features is authentication. Its correctness is strongly related with the whole of communication security. In this paper, we introduce three types of attack-models that can be actualized as their attack-scenarios, and provide an authentication protocol checker to apply the three types of the attack-scenarios. We also show some problems in security protocols.

A Global Security Architecture for Intrusion Detection on Computer Networks

Abdoul Karim Ganame, Julien Bourgeois, Renaud Bidou and Francois Spies

LIFC
University of Franche-Comte
Montbeliard, France
{*ganame, bourgeois, bidou, spies*}*@lifc.univ-fcomte.fr*

Detecting all kinds of intrusions efficiently requires a global view of the monitored network. Built to increase the security of computer networks, traditional IDS are unfortunately unable to give a global view of the security of a network. To overcome this situation, we are developing a distributed SOC (Security Operation Center) which is able to detect attacks occurring simultaneously on several sites in a network and to give a global view of the security of that network.

In this article, we present the global architecture of our system, called DSOC as well as several methods used to test its accuracy and performance.

Improving Secure Communication Policy Agreements by Building Coalitions

Srilaxmi Malladi[1], Sushil K. Prasad[1] and Shamkant B. Navathe[2]

[1]*Department of Computer Science*
Georgia State University
Atlanta, GA, U.S.A.
{*cscsrmx, sprasad*}*@cs.gsu.edu*

[2]*College of Computing*
Georgia Institute of Technology
Atlanta, GA, U.S.A.
sham@cc.gatech.edu

In collaborative applications, participants agree on certain level of secure communication based on communication policy specifications. Given secure communication policy specifications of various group members at design time, the minimum set of resources for a pair, called Resolved Policy Level Agreement (RPLA) is translated into appropriate security service implementations, for the pair-wise communication to take place. We propose a novel idea that the members may extend pair-wise communication quality through other trusted nodes whose communication resources offer more security. We propose a heuristic algorithm which finds the best quality of protection (QoP), a measure of the resistance to an attack, path through coalition of trusted nodes. The results from our experiments indicate a significant improvement in QoP in the range of 13% to 48% over pair-wise communications.

Workshop 17
Workshop on System Management Techniques, Processes, and Services SMTPS 2007

17 SMTPS • Workshop on System Management Techniques, Processes and Services

Workshop Description:

In today's high-performance computing, system management and related services play a key role. With service business accounting for more than half of the U.S. economy, in our third year of SMTPS we would like to broaden the scope of our workshop to cover all aspects of system management, going beyond scientific computing. In order to satisfy the systems needs of both commercial and scientific applications, the focus on system management now includes not only the tools and user interfaces, but also other aspects such as services, processes and system control. Businesses in the IT area are focusing more and more on innovative techniques, processes and methods to manage commercial or scientific systems remotely, in order to optimize the use of resources by minimizing system down time. As a result, there are requirements not only to revisit some of the traditional methods used to develop system management tools for today's servers but also to evaluate the implications and benefits the new programming models can provide for parallel and distributed systems in terms of system services performance and utilization. This workshop is intended to bring together researchers and practitioners to identify the new challenges imposed by this trend and investigating efficient software tools, techniques and service processes to improve the performance, reliability and operation of enterprise servers including parallel and distributed 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 check-pointing tools
- Self-healing and self-management tools
- Power management for enterprise servers leading to efficient systems management
- 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
- Optimization techniques for services management
- Services engineering and utility computing techniques
- Web services and Services oriented architecture and implications to system management aspects

General Chair:

Ramendra Sahoo, IBM Research, USA

Program Co-Chairs:

Kyung Dong Ryu, IBM Research
Fabrizio Petrini, Pacific Northwest National Lab
Yanyong Zhang, Rutgers University, USA

Program Committee:

Ricardo Bianchini, Rutgers
Henri Casanova, Hawaii
I-hsin Chung, IBM Research
Dick Epema, Delft
Dror Feitelson, Hebrew University
John Janakiraman, HP
Joefon Jann, IBM Research
Jose E. Moreira, IBM
Manish Parashar, Rutgers
Rolf Riesen, Sandia
Anand Sivasubramaniam, Penn State
Rajeev Thakur, Argonne
Andy Yoo, LLNL

Keynote: Five Years with the High Productivity Computing Systems Program — A Perspective

Elmootazbellah Elnozahy

IBM Research
Austin, TX, USA
mootaz@us.ibm.com

For the past five years, I had the very enviable task of leading IBM's effort in DARPA's High Productivity Computing Systems (HPCS) program. IBM competed successfully with other contestants in and survived two down-selects, producing along the way ground-breaking research for peta-scale systems aimed at changing the status quo in high end computing. The HPCS program is unique in that it states productivity as a broader definition of the system value than just performance. Commercial viability is another goal, meant to add realism and produce usable systems at the end of the program with productivity and performance goals that well exceed the projected improvements using today's technology. This unprecedented mix adds interesting and challenging constraints on the research program, and the traditional ways of approaching the problem do not apply. This talk will give an overview of the challenges of running projects of this kind, and gives a forward looking statement about the future of the program and its projected impact on the industry and the academic communities.

Detecting Runtime Environment Interference with Parallel Application Behavior

Rashawn L. Knapp[1], Karen L. Karavanic[1] and Douglas M. Pase[2]

[1]*Computer Science*
Portland State University
Portland, OR, USA
{knappr, karavan}@cs.pdx.edu

[2]*IBM System x HPC Portfolio Development*
IBM Corporation
Research Triangle Park, NC, USA
pase@us.ibm.com

Many performance problems observed in high end systems are actually caused by the runtime system and not the application code. Detecting these cases will require parallel performance tools to incorporate information about the runtime system; however many current tools do not. We present a test suite for evaluating the ability of performance tools to reach a correct diagnosis in cases where a problem is caused by the runtime environment. We include a set of results for one of the tests, which measures application performance as NFS server load is increased. We also present a model for performance diagnosis that combines system and application level information.

Automatic Path Migration Over InfiniBand: Early Experiences

Abhinav Vishnu, Amith R. Mamidala, Sundeep Narravula and Dhabaleswar K. Panda

Computer Science and Engineering
The Ohio State University
Columbus, OH, USA
{*vishnu, mamidala, narravul, panda*}*@cse.ohio-state.edu*

High computational power of commodity PCs combined with the emergence of low latency and high bandwidth interconnects has escalated the trends of cluster computing. Clusters with InfiniBand are being deployed, as reflected in the TOP 500 Supercomputer rankings. However, increasing scale of these clusters has reduced the Mean Time Between Failures (MTBF) of components. Network component is one such component of clusters, where failure of Network Interface Cards (NICs), cables and/or switches breaks existing path(s) of communication. InfiniBand provides a hardware mechanism, Automatic Path Migration (APM), which allows user transparent detection and recovery from network fault(s), without application restart. In this paper, we design modules; which work together for providing network fault tolerance for user level applications leveraging the APM feature. Our performance evaluation at the MPI Layer shows that APM incurs negligible overhead in the absence of faults in the system. In the presence of network faults, APM incurs negligible overhead for reasonably long running applications.

A Selective Profiling Tool: Towards Automatic Performance Tuning

Abhinav Bhatele[1] and Guojing Cong[2]

[1]*Computer science*
UIUC
Urbana, IL, USA
bhatele2@uiuc.edu

[2]*IBM TJ Watson Research Center*
Yorktown heights, NY, USA
gcong@us.ibm.com

We present some preliminary results of selective profiling in our efforts towards automatic performance tuning for scientific codes. Performance analysis and tuning are becoming very important with the increasing complexity and speed of high performance systems. Great efforts are necessary to tune applications for optimal performance on such systems.

In our efforts to automate most, if not all, of the performance tuning process, we developed a flexible profiling tool that can quickly pinpoint the performance bottleneck and further refine the problem area. This is an important first step in our open framework with a rule-based approach in our on-going PERCS project.

A Flexible Resource Management Architecture for the Blue Gene/P Supercomputer

Sam Miller, Mark Megerian, Paul Allen and Tom Budnik

Systems & Technology Group
IBM
Rochester, MN, USA
{*samjmill, megerian, pvallen, tbudnik*}*@us.ibm.com*

Blue Gene®/P is a massively parallel supercomputer intended as the successor to Blue Gene/L. It leverages much of the existing architecture of its predecessor to provide scalability up to a petaflop of peak computing power. The resource management software for such a large parallel system faces several challenges, including system fragmentation due to partitioning, presenting resource usage information using a polling or event driven model, and acting as a barrier between external resource management systems and the Blue Gene/P core.

This paper describes how the Blue Gene/P resource management architecture is extremely flexible by providing multiple methodologies for obtaining resource usage information to make scheduling decisions. Three distinctly separate resource management services will be described. First, the Bridge API, a full-featured API suitable for fine tuning scheduling and allocation decisions. Second, a light-weight Allocator API for allocating resources without substantial development costs. And lastly, configuring the system into static partitions. Job scheduling strategies utilizing each of the methods will be discussed.

Encompass: Managing Functionality

Oleg Goldshmidt, Benny Rochwerger, Alex Glikson, Inbar Shapira and Tamar Domany

IBM Haifa Research Lab
Haifa, Israel
{*olegg, benny, glikson, inbar_shapira, tamar*}*@il.ibm.com*

Today's system management tools focus on a computer as a visible enclosure of both computational resources (CPU and memory) and the functionality and data that reside in the storage subsystem. Recent technological trends, such as shared SAN or NAS storage and virtualization, have the potential to break this tight association between functionality and machines.

We describe the design and implementation of Encompass - an image management system centered around a shared storage repository of "master system images", each representing different functionality. The functionality is provisioned by "cloning" master images, associating the resulting "clone images" with specified physical and/or virtual resources ("machines"), customizing the clone images for the specific environment and circumstances, and automatically performing the necessary operations to activate the clones. "Machines" - physical or virtual - are merely computational resources that do not have any permanent association with functionality.

Encompass supports the complete lifecycle of a system image, including reallocation and re-targeting of resources, maintenance, updates, etc. It separates image creation from image management from resource allocation policies - an emerging trend that is manifested in particular by proliferation of turn-key "virtual appliances".

Base Operating System Provisioning and Bringup for a Commercial Supercomputer

David Daly, Jong Hyuk Choi, Jose E. Moreira and Amos Waterland

IBM T.J. Watson Research Center
Yorktown Heights, NY, USA
{*dmdaly, jongchoi, jmoreira, apw*}*@us.ibm.com*

Commercial Scale-Out is a new research project at IBM Research. Its main goal is to investigate and develop technologies for the use of large scale parallelism in commercial applications, eventually leading to a commercial supercomputer. The project leverages and explores the features of IBM's BladeCenter family of products. A significant challenge in using a large cluster of servers is the installation and provisioning of the base operating system in those servers. Compounding this problem is the issue of maintenance of the software image in each server after its provisioning. This paper describes the system we developed to manage the installation, provisioning, and maintenance process for a cluster of blades, providing a base level of functionality to be used by higher level management tools. The system leverages the management facilitation features of BladeCenter, and exploits the network and storage architecture of the Commercial Scale-Out prototype cluster. It uses a single shared root filesystem image to reduce management complexity, and completely automates the process of bringing a new blade into the cluster upon its insertion into a BladeCenter chassis.

Scale-up x Scale-out: A Case Study using Nutch/Lucene

Maged Michael, Jose E. Moreira, Doron Shiloach and Robert W. Wisniewski

IBM T.J. Watson Research Center
Yorktown Heights, NY, USA
{*magedm, jmoreira, doron, bobww*}*@us.ibm.com*

Scale-up solutions in the form of large SMPs have represented the mainstream of commercial computing for the past several years. The major server vendors continue to provide increasingly larger and more powerful machines. More recently, scale-out solutions, in the form of clusters of smaller machines, have gained increased acceptance for commercial computing. Scale-out solutions are particularly effective in high-throughput web-centric applications. In this paper, we investigate the behavior of two competing approaches to parallelism, scale-up and scale-out, in an emerging search application. Our conclusions show that a scale-out strategy can be the key to good performance even on a scale-up machine. Furthermore, scale-out solutions offer better price/performance, although at an increase in management complexity.

Performance Studies of a WebSphere Application, Trade, in Scale-out and Scale-up Environments

Hao Yu, Jose E. Moreira, Parijat Dube, I-Hsin Chung and Li Zhang

IBM Thomas J. Watson Research Center
Yorktown Heights, NY, USA
{*yuh, jmoreira, pdube, ihchung, zhangli*}*@us.ibm.com*

Scale-out approach, in contrast to scale-up approach (exploring increasing performance by utilizing more powerful shared-memory servers), refers to deployment of applications on a large number of small, inexpensive, but tightly packaged and tightly interconnected servers. Recently, there has been an increasing interest in scale-out approach. The purpose of this study is to discover advantages or disadvantages of scale-out systems with a typical enterprise workload, IBM Trade Performance Benchmark Sample for WebSphere Application Server (a.k.a. Trade6). In this work, through cross system performance comparison, we show that for such workload, scale-out approach has better performance/cost effect. In term of scalability, we show that WebSphere Application Server packages for distributed environment scale well while the possible bottleneck of the application deployment is the database tier. We present preliminary results to show that both database partitioning feature (DPF) and federated database server approaches are not exactly suitable for providing scale-out solution for the database tier of workloads similar to Trade (small tables and short transactions). In addition, we discuss our on-going effort on further performance study: (1) studies of performance/scalability for larger deployments by adopting the IBM AMBIENCE queuing network modeling tool, (2) performance breakdowns utilizing IBM ACTC hardware counter library.

Storage Optimization for Large-Scale Distributed Stream Processing Systems

Kirsten Hildrum[1], Fred Douglis[1], Joel Wolf[1], Philip Yu[1], Lisa Fleischer[2] and Akshay Katta[3]

[1]*IBM T.J. Watson Research Center*
Hawthorne, NY, USA
{*hildrum, fdouglis, jlwolf, psyu*}*@us.ibm.com*

[2]*Computer Science*
Dartmouth
Hanover, NY, USA
lkf@us.ibm.com

[3]*Amazon Corportation*
Seattle, WA, USA
ark2001@cs.columbia.edu

We consider storage in an extremely large-scale distributed computer system designed for stream processing applications. In such systems, incoming data and intermediate results may need to be stored to enable future analyses. The quantity of such data would dominate even the largest storage system. Thus, a mechanism is needed to keep the most useful data. One recently introduced approach is to employ retention value functions, which effectively assign each data object a value that changes over time. Storage space is then reclaimed automatically by deleting data of lowest current value. In such large systems, there will naturally be multiple file systems available, each with different properties. Choosing the right file system for a given incoming data stream presents a challenge. In this paper we provide a novel and effective scheme for optimizing the placement of data within a distributed storage subsystem employing retention value functions. The goal is to keep the data of highest overall value, while simultaneously balancing the read load to the file system.

Peak-Performance DFA-based String Matching on the Cell Processor

Daniele Paolo Scarpazza[1], Oreste Villa[2] and Fabrizio Petrini[1]

[1]Computational & Information Sciences Division
Pacific Northwest National Laboratory
Richland, WA, United States of America
{daniele.scarpazza, fabrizio.petrini}@pnl.gov

[2]Dipartimento di Elettronica e Informazione
Politecnico di Milano
Milano, Italy
ovilla@elet.polimi.it

The security of your data and of your network is in the hands of intrusion detection systems, virus scanners and spam filters, which are all critically based on string matching. But network links are getting faster and faster, and string matching is getting more and more difficult to perform in real time. Traditional processors are not keeping up with the performance demands, whereas specialized hardware will never be able to compete with commodity hardware in terms of cost effectiveness, reusability and ease of programming. Advanced multi-core architectures like the IBM Cell Broadband Engine promise unprecedented performance at a low cost, thanks to their popularity and production volume. Nevertheless, the suitability of the Cell processor to string matching has not been investigated so far.

In this paper we investigate the performance attainable by the Cell processor when employed for string matching algorithms based on Deterministic Finite-state Automata (DFA). Our findings show that the Cell is an ideal candidate to tackle modern security needs: two processing elements alone, out of the eight available on one Cell processor provide sufficient computational power to filter a network link with bit rates in excess of 10 Gbps.

An Adaptive Semantic Filter for Blue Gene/L Failure Log Analysis

Yinglung Liang[1], Hui Xiong[2], Yanyong Zhang[1] and Ramendra Sahoo[3]

[1]ECE Department
Rutgers University
Piscataway, NJ, 08854
ylliang@eden.rutgers.edu, yyzhang@ece.rutgers.edu

[2]MSIS Department
Rutgers University
Newark, NJ, 07102
hui@rbs.rutgers.edu

[3]System and Network Services Department
IBM T. J. Watson Research Center
Yorktown Heights, NY, 10598
rsahoo@us.ibm.com

Frequent failure occurrences are becoming a serious concern to the community of high-end computing, especially when the applications and the underlying systems rapidly grow in size and complexity. In order to better understand the failure behavior of such systems and further develop effective fault-tolerant strategies, we have collected detailed event logs from IBM Blue Gene/L, which has as many as 128K processors, and is currently the fastest supercomputer in the world. Due to the scale of such machines and the granularity of the logging mechanisms, the logs can get voluminous and usually contain records which may not all be distinct. Consequently, it is crucial to filter these logs towards isolating the specific failures, which can then be useful for subsequent analysis. However, existing filtering methods either require too much domain expertise, or produce erroneous results. This paper thus fills this crucial void by designing and developing an Adaptive Semantic Filtering (ASF) method, which is accurate, light-weight, and more importantly, easy to automate. Specifically, ASF exploits the semantic correlation between two events, and dynamically adapts the correlation threshold based on the temporal gap between the events. We have validated the ASF method using the failure logs collected from Blue Gene/L over a period of 98 days. Our experimental results show that ASF can effectively remove redundant entries in the logs, and the filtering results can serve as a good base for future failure analysis studies.

Workshop 18
Workshop on Performance Optimization for High-Level Languages and Libraries
POHLL 2007

• 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

Atanas (Nasko) Rountev, The Ohio 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
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, The University of Texas at Austin
Marcus Pueschel, Carnegie Mellon University
J. (Ram) Ramanujam, Louisiana State University
Atanas (Nasko) Rountev, The Ohio 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

POET: Parameterized Optimizations for Empirical Tuning

Qing Yi[1], Keith Seymour[2], Haihang You[2], Richard Vuduc[3] and Dan Quinlan[3]

[1]*Computer Science*
University of Texas at San Antonio
San Antonio, TX, USA
qingyi@cs.utsa.edu

[2]*Computer Science*
University of Tennessee at Knoxville
Knoxville, TN, USA
{seymour, you}@cs.utk.edu

[3]*CASC*
Lawrence Livermore National Laboratory
Livermore, CA, USA
{vuduc2, dquinlan}@llnl.gov

The excessive complexity of both machine architectures and applications have made it difficult for compilers to statically model and predict application behavior. This observation motivates the recent interest in performance tuning using empirical techniques. We present a new embedded scripting language, POET (Parameterized Optimization for Empirical Tuning), for parameterizing complex code transformations so that they can be empirically tuned. The POET language aims to significantly improve the generality, flexibility, and efficiency of existing empirical tuning systems. We have used the language to parameterize and to empirically tune three loop optimizations—interchange, blocking, and unrolling—for two linear algebra kernels. We show experimentally that the time required to tune these optimizations using POET, which does not require any program analysis, is significantly shorter than that when using a full compiler-based source-code optimizer which performs sophisticated program analysis and optimizations.

Experience of Optimizing FFT on Intel Architectures

Daniel Orozco, Liping Xue, Murat Bolat, Xiaoming Li and Guang R. Gao

Electrical and Computer Engineering
University of Delaware
Newark, DE, USA
orozco@eecis.udel.edu, {xue, ggao}@capsl.udel.edu, murat@udel.edu, xli@ece.udel.edu

Automatic library generators, such as ATLAS [**?**], Spiral [**?**] and FFTW [**?**], are promising technologies to generate efficient code for different computer architectures. The library generators usually tune programs using two layers of optimizations: the search at the algorithm level, and the optimization for micro kernels. The micro optimizations are important for the performance of library, because the optimized micro kernels are the bases of algorithm level search, and have a great impact on the overall performance of the generated libraries. A successfully optimized micro kernel requires thorough understanding of the interaction between architectural features and highly optimized code. However, literature on library generators focus more on the algorithm level optimization, and usually give only simple discussion of how kernel codes are generated and tuned. As a result, the optimization of micro kernels is still an art that depends on individual expertise, and is insufficiently documented. In this paper, we study the problem of how to generate efficient FFT kernels. We apply a series of micro optimizations, for example, memory hierarchy locality enhancements, to several FFT routines, and use hardware counters to observe the interactions between those optimizations with Intel Pentium 4 and the latest Intel Core 2 architecture. We achieve good speedups, and more importantly, we present methods that can be used to generate high-performance FFT kernels on different architectures.

Optimizing the Fast Fourier Transform on a Multi-core Architecture

Long Chen[1], Ziang Hu[1], Junmin Lin[2] and Guang R. Gao[1]

[1]CAPSL, ECE Dept.
University of Delaware
Newark, Delaware, USA
{lochen, ggao}@capsl.udel.edu, hu@ee,udel.edu

[2]Dept. of Computer Technology
Tsinghua University
Beijing, China
linjunmin@tsinghua.org.cn

The rapid revolution in microprocessor chip architecture due to multicore technology is presenting unprecedented challenges to the application developers as well as system software designers: how to best exploit the parallelism potential due to such multi-core architectures ? In this paper, we report an in-depth study on such challenges based on our experience of optimizing the Fast Fourier Transform (FFT) on the IBM Cyclops-64 chip architecture - a large-scale multi-core chip architecture consisting 160 thread units, associated memory banks and an interconnection network that connect them together in a shared memory organization.

We demonstrate how multi-core architectures like the C64 could be used to achieve a high performance implementation of FFT both in 1D and 2D cases. We analyze the optimization challenges and opportunities including problem decomposition, load balancing, work distribution, and data-reuse, together with the exploiting of the C64 architecture features such as the multi-level of memory hierarchy and large register files.

Furthermore, the experience learned during the hand-tuned optimization process have provided valuable guidance in our compiler optimization design and implementation.

The main contributions of this paper include: 1) our study demonstrates that successful optimization for C64-like large-scale multi-core architectures requires a careful analysis that can identify certain domain-specific features of a target application (e.g. FFT) and match them well with some key multi-core architecture features; 2) Our optimization, assisted with hand-tunned process, provided quantitative evidence on the importance of each optimization identified in 1) ; 3) Automatic optimization by our compiler, the design and implementation of which is guided by the feedbacks from 1) and 2), shows excellent results that are often comparable to the results derived from our time-consuming hand-tunned code.

Performance Analysis of a Family of WHT Algorithms

Michael Andrews and Jeremy Johnson

Department of Computer Science
Drexel University
Philadelphia, PA, USA
mjand@drexel.edu, jjohnson@cs.drexel.edu

This paper explores the correlation of instruction counts and cache misses to runtime performance for a large family of divide and conquer algorithms to compute the Walsh–Hadamard transform (WHT). Previous work showed how to compute instruction counts and cache misses from a high–level description of the algorithm and proved theoretical results about their minimum, maximum, mean, and distribution. While the models themselves do not accurately predict performance, it is shown that they are statistically correlated to performance and thus can be used to prune the search space for fast implementations. When the size of the transform fits in cache the instruction count itself is used; however, when the transform no longer fits in cache, a linear combination of instruction counts and cache misses is used. Thus for small transforms it is safe to ignore algorithms which have a high instruction count and for large transforms it is safe to ignore algorithms with a high value in the combined instruction count/cache miss model. Since the models can be computed from a high–level description of the algorithms, they can be obtained without runtime measurement and the previous theoretical results on the models can be applied to limit empirical search.

Model-Guided Empirical Optimization for Multimedia Extension Architectures: A Case Study

Chun Chen[1], Jaewook Shin[2], Shiva Kintali[3], Jacqueline Chame[1] and Mary Hall[1]

[1]*Information Sciences Institute*
University of Southern California
Marina del Rey, CA, USA
{*chunchen, jchame, mhall*}*@isi.edu*

[2]*MCS Division*
Argonne National Laboratory
Argonne, IL, USA
jaewook@mcs.anl.gov

[3]*College of Computing*
Georgia Institute of Technology
Atlanta, GA, USA
kintali@cc.gatech.edu

Compiler technology for multimedia extensions must effectively utilize not only the SIMD compute engines but also the various levels of the memory hierarchy: superword registers, multi-level caches and TLB. In this paper, we describe a compiler that combines optimization across all levels of the memory hierarchy with automatic generation of SIMD code for multimedia extensions. At the high-level, model-guided empirical optimization is used to transform code to optimize for all levels of the memory hierarchy. This compiler interacts with a backend compiler exploiting superword-level parallelism that takes sequential code as input and produces SIMD code. This paper discusses how we have combined these technologies into a single framework. Through a case study with matrix multiply, we observe performance results that outperform the hand-tuned Intel MKL library, and achieve performance that is within 4% of the ATLAS self-tuning library with architectural defaults and more than 4X faster than the native Intel compiler.

Automatic Program Segment Similarity Detection in Targeted Program Performance Improvement

Haiping Wu[1], Eunjung Park[1], Mihailo Kaplarevic[1], Yingping Zhang[2], Murat Bolat[1], Xiaoming Li[1] and Guang R. Gao[1]

[1]*Dept. of Electrical and Computer Engineering*
University of Delaware
Newark, DE, USA
hwu@capsl.udel.edu, {*epark, murat*}*@udel.edu,* {*kaplar, xli, ggao*}*@ece.udel.edu*

[2]*Digital Enterprise Group*
Intel Cooperation
Chandler, AZ, USA
ying.m.zhang@intel.com

Targeted optimization of program segments can provide an additional program speedup over the highest default optimization level, such as -O3 in GCC. The key challenge is how to automatically search for performance sensitive program segments in a given code, to which a customized set of optimization compiler options could be applied.

In this paper we propose a method for automatic detection of performance sensitive program segments based on program segment similarity. First we create a proxy segment template database trained over a set of random input programs. The compiler identifies program segments by correlating them to the pre-build proxy segment templates using the syntax structure and architecture-dependent behavior similarity. We argue that the identified program segments can be custom optimized to improve the overall program performance.

The method is evaluated on the Intel XScale PXA255 platform using randomly selected benchmarks. The experimental results show that our method can provide additional speedups over the highest optimization level in GCC 3.3 (-O3) for an arbitrary set of applications.

From Hardware to Software Synthesis of Linear Feedback Shift Registers

Lauradoux Cédric

INRIA
Team CODES
Le Chesnay, Yvelines, France
cedric.lauradoux@inria.fr

Linear Feedback Shift Registers (LFSRs) have always received considerable attention in computer science especially in coding theory and in cryptography. The scope of applications of LFSRs is wide: data scrambling, spread spectrum, build in self tests (BISTs)... They have to be implemented either in hardware or in software. Unlike hardware, software applications have not been very popular. The main reason is that, even if the LFSR synthesis in software is very similar to the LFSR synthesis on Xilinx FPGA, the overall processing is parallel in hardware while it is almost sequential in software, leading to low throughput implementations. If the naive LFSR implementation is in favor of hardware, increasing the number of LFSR steps computed at the same time can considerably improve software implementation. For instance, we obtain a 103 speedup factor for a 128-bit LFSR on 64-bit processors. Unfortunately, this cannot be obtain for all LFSRs. We here describe how LFSR parameters must be chosen to obtain an ef cient implementation.

Code Generation: On the Scheduling of DAGs Using Worm-Partition

Hatem M. El-Boghdadi[1] and Mohamed Bohalfaeh[2]

[1]*Computer Engineering Dept.*
Cairo University
Giza, EGYPT
helboghdadi@eng.cu.edu.eg

[2]*Computer Science Dept.*
Cairo University
Giza, EGYPT
bohalfaehm@yahoo.co.uk

Code generation consists of three main stages, instruction selection, scheduling and register allocation. The scheduling stage is very important because it affects the execution time of resulting code as well as the associated memory space needed to store the program. This paper deals with scheduling directed acyclic graphs (DAGs) using worm-partition. First, we develop a new algorithm to partition DAGs into a collection of worms while ensuring that the worm-partition is legal. Although the algorithm does not guarantee the minimum number of worms, it runs in optimal $O(|V| + |E|)$ time. This is in contrast to the known method for producing the minimum number of worms that runs in $O(|V|^2 + |V||E|)$. We apply the algorithm to benchmark real problems and show its comparable results to the previous method. Then we study some DAG properties that are related to worm partitioning. We derive a necessary condition for the minimum number of worms in a given DAG. In other words, a lower bound for the number of worms. Then we identify two important classes of DAGs, for which this necessary condition is sufficient as well; i.e. we show that the lower bound is a tight one. Finally, we show that our algorithm generates the minimum number of worms for theses classes of DAGs.

Library Function Selection in Compiling Octave

Daniel Mcfarlin and Arun Chauhan

Computer Science
Indiana University
Bloomington, Indiana, USA
{*dmcfarli, achauhan*}*@cs.indiana.edu*

One way to address the continuing performance problem of high-level domain-specific languages, such as Octave or MATLAB, is to compile them to a relatively lower level language for which good compilers are available. As a first step in this direction, specializing the high-level operations in the source, based on operand types, leads to significant gains. However, simple translation of the high-level operations to the underlying libraries can often miss important opportunities to improve performance. This paper presents a global algorithm to select functions from a target library, utilizing the semantics of the operations as well as the platform-specific performance characteristics of the library. Making use of the library properties, the simple and easy-to-implement selection algorithm is able to achieve as much as three times performance improvement for certain linear algebra kernels, over a straight mapping of operations, which are compiled to the vendor-tuned BLAS.

A Portable Framework for High-Speed Parallel Producer/Consumers on Real CMP, SMT and SMP Architectures

Richard T. Saunders[1], Clinton L. Jeffery[2] and Derek T. Jones[1]

[1]*Rincon Research Corporation*
Tucson, AZ, USA
{*rts, dtj*}*@rincon.com*

[2]*Department of Computer Science*
University of Idaho
Moscow, ID, USA
jeffery@cs.uidaho.edu

This paper explores generating efficient, portable High-Speed Producer Consumer (HSPC) code on current shared memory architectures: Chip Multi-Processors (CMP), Simultaneous Multi-Threading processors (SMT) and Shared Memory Processors (SMP). To build an HSPC, we use a code generation approach in two stages.

Stage One generates data structures to eliminate memory interference. This is done by adjusting and timing cache/buffer/stack placements and lengths for an idealized producer/consumer. Perfect load-balancing is achievable for CMP and SMP, but not for SMT due to simultaneous-execution interference.

In Stage Two, the codebase is refined inside its target application: profiling events sent from Python to a consumer that computes profiling information. Stage two further tests the impact of altering event sizes, synchronization primitives, container libraries, and processor affinity. Stage two achieves near perfect balancing for CMP and SMP architectures, but SMT still performs poorly.

libDMC: a Library to Operate Efficient Distributed Model Checking

Alexandre Hamez, Fabrice Kordon and Yann Thierry-Mieg

LIP6/MoVe
Pierre and Marie Curie
Paris, France
{*alexandre.hamez, fabrice.kordon, yann.thierry-mieg*}*@lip6.fr*

Model checking is a formal verification technique that allows to automatically prove that a system's behavior is correct. However it is often prohibitively expensive in time and memory complexity, due to the so-called state space explosion problem. We present a generic multi-threaded and distributed infrastructure library designed to allow distribution of the model checking procedure over a cluster of machines. This library is generic, and is designed to allow encapsulation of any model checker in order to make it distributed. Performance evaluations are reported and clearly show the advantages of multi-threading to occupy processors while waiting for the network, with linear speedup over the number of processors.

Workshop 19
International Workshop on Hot Topics in Peer-to-Peer Systems
HOTP2P 2007

19 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), whose first edition has been held in Volendam, the Netherlands, on Oct. 8th, 2004, second edition has been held in San Diego, California, on Jul. 21st, 2005, and third edition has been held in Rhodes Island, Greece, on 29th April 2006, 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 Co-Chairs:

Giovanni Chiola, Universita' di Genova (Italy)
Franck Cappello, INRIA/Universitè Paris Sud (France)

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)
Michele Colajanni, Universita' di Modena e Reggio Emilia (Italy)
Antonio Corradi, Universita' di Bologna (Italy)
Paul Ezhilchelvan, University of Newcastle (UK)
Thomas Fuhrmann, Universität Karlsruhe (TH)
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)
Manish Parashar, Rutgers University, New Jersey (USA)
Giancarlo Ruffo, Universita' di Torino (Italy)
Sanjeev Setia, George Mason University (USA)
Rich Wolski, University of California, Santa Barbara (USA)

Towards threat-adaptive dynamic fragment replication in large scale distributed systems

Roberto Di Pietro[1], Luigi V. Mancini[2] and Alessandro Mei[2]

[1]*Department of Mathematics*
University of Rome 3
Rome, Italy
dipietro@mat.uniroma3.it

[2]*Department of Computer Science*
University of Rome
Rome, Italy
{mancini, mei}@di.uniroma1.it

In this paper, we consider new issues in building secure p2p file sharing systems. In particular, we define a powerful adversary model and consequently present the requirements to address when implementing a threat-adaptive secure file sharing system. We describe the main components of such a system: An early warning mechanism to perform pre-emptive actions against new vulnerabilities; a mechanism to sanitize corrupted nodes; a protocol to securely "migrate" data from non-safe nodes; and an efficient dynamic secret sharing mechanism.

Effects of Replica Placement Algorithms on Performance of structured Overlay Networks

Bassam A. Alqaralleh[1], Chen Wang[2], Bing Bing Zhou[3] and Albert Zomaya[4]

[1]*School of Information Technologies*
University of Sydney
Sydney, NSW, Australia
bassam@it.usyd.edu.au

[2]*School of Information Technologies*
University of Sydney
Sydney, NSW, Australia
cwang@it.usyd.edu.au

[3]*School of Information Technologies*
University of Sydney
Sydney, NSW, Australia
bbz@it.usyd.edu.au

[4]*School of Information Technologies*
University of Sydney
Sydney, NSW, Australia
zomaya@it.usyd.edu.au

In DHT-based P2P systems, Replication-based content distribution and load balancing strategies consists of such decisions as which files should be replicated, how many replicas should be created and where to replicate them in order increase the system performance in the presence of non-uniform data and access distribution. There are many works on replica placement policies; however, the impact of system workload on different replica placement strategies is not well studied. We investigate this problem under the context of content addressable overlay networks. We compare a trace based replica placement algorithm with two of its variations, namely random placement and priority based placement under different workloads. Our experimental results show that the effect of replica placement policy is highly affected by the workload of the system, which indicates that an adaptive replica placement strategy is desirable for content distribution in an overlay network.

A Resource Allocation Problem in Replicated Peer-to-peer Storage Systems

Sriram Ramabhadran and Joseph Pasquale

Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA, USA
{*sriram, pasquale*}*@cs.ucsd.edu*

This paper focuses on peer-to-peer storage systems that achieve availability through replication. We study the problem of resource allocation when the system must replicate multiple files using a fixed amount of resource. We characterize the optimal allocation that maximizes the average availability of the files in the system, and also study two simple, decentralized allocation schemes, viz., uniform allocation, where each file is allocated equal shares of the resource, and proportional allocation, where each file is allocated a share of the resource proportional to its size. We show that while uniform allocation is fair in terms of allocating resources, it may be arbitrarily sub-optimal. On the other hand, proportional allocation, though unfair in resource allocation, is competitive with the optimal allocation.

P2PADM: An In-kernel Gateway Architecture for Managing P2P Traffic

Ying-Dar Lin[1], Po-Ching Lin[1], Meng-Fu Tsai[1], Tsao-Jiang Chang[1] and Yuan-Cheng Lai[2]

[1]*Department of Computer Science*
National Chiao Tung University
Hsinchu, Taiwan
{*ydlin, pclin, mftsai, tjchang*}*@cis.nctu.edu.tw*

[2]*Department of Information Management*
National Taiwan University of Science and Technology
Taipei, Taiwan
laiyc@cs.ntust.edu.tw

This work presents an in-kernel gateway architecture on Linux, namely *kP2PADM*, for managing P2P traffic on dynamic ports. This design can effectively eliminate redundant data passing between the kernel space and the user space. The management functions include (1) classifying and filtering P2P traffic, (2) scanning viruses on shared files, (3) auditing chatting messages and transferred files, and (4) bandwidth control. Practical implementation issues and techniques in the system design are discussed herein. This design proposes a dual-queue architecture to handle packet reassembly and resolve head-of-line blocking. A connection cache accelerates handling the reconnection requests from the peers. The throughput can achieve up to 185.73 Mbps even with content filtering, and remains around 79.09 Mbps when virus scanning is enabled. The impacts of each management function and out-of-order packets on performance are also analyzed through the internal benchmarks.

A Peer-to-Peer Infrastructure for Autonomous Grid Monitoring

Laurent Baduel and Satoshi Matsuoka

Tokyo Institute of Technology
Tokyo, Japan
baduel@smg.is.titech.ac.jp, matsu@is.titech.ac.jp

Modern grids have become very complex by their size and their heterogeneity. It makes the deployment and maintenance of systems a difficult task requiring lots of efforts from administrators and programmers. Our goal is to investigate the concepts that underlie autonomic computing systems, especially for grid environment. We believe that peer-to-peer overlay networks are a valuable basis to support some of the main issues of autonomic computing in the particular case of grids.

This article presents the construction of an autonomous, decentralized, scalable, and efficient grid monitoring system. The components of this application negotiate through a peer-to-peer network in order to provide autonomic behaviors and exchange data. We present a solution based on a gossip broadcast protocol upon a hierarchical, directed, and acyclic graph to rapidly diffuse information in the system while limiting the number of messages. The software architecture is detailed, and then the first results of its performance are presented and analyzed.

A Pretty Flexible API for Generic Peer-to-Peer Programming

Giuseppe Ciaccio

DISI
Universita' di Genova
Genova, Italy
ciaccio@disi.unige.it

We propose and motivate an API for programming distributed applications using a structured overlay network of peers as infrastructure. The API offers simple primitives and powerful mechanisms, in a way that is independent from the underlying overlay.

The dynamic set of participants is abstracted by providing a flat space of keys, transparently scattered across all participants in the overlay. The API primitives allow application instances to send messages towards individual keys. Two different kinds of messages can be exchanged, namely, unidirectional and request-response; the latter takes place in a split-phase non-blocking way, so that the application can be made latency-tolerant and thus more performing. The request-response pattern is also shown to be crucial for those applications demanding a degree of user anonymity.

The semantics of messages is not defined by the API itself. Rather, the API offers a mechanism to allow the application to set up handlers, which are upcalls to run upon message arrivals at each peer. The overall behaviour of the application is thus shaped by the handlers.

The API also allows to define application-level handlers for other two typical tasks of any dynamic peer-to-peer system, namely, the migration of keys across peers after new peer arrivals, and the regeneration of missing keys after peer departures.

Spinneret: A Log Random Substrate for P2P Networks

Jeffrey Rose, Cyrus Hall and Antonio Carzaniga

Department of Informatics
University of Lugano
Lugano, Switzerland
{*jeffrey.rose, cyrus.hall*}*@lu.unisi.ch, antonio.carzaniga@unisi.ch*

Until now, structured and unstructured networks have been considered in absentia of each other. We believe that next-generation P2P services will require both structured and unstructured algorithms, and that it therefore makes sense to consider a unified substrate that provides good service for both. In this paper we argue for the creation of a semi-structured overlay substrate, called Spinneret, which can serve as the base layer for a variety of structured and unstructured search algorithms. In order to validate that this structure forms a good foundation for various services, we present two algorithms simulated on top of the Spinneret substrate: an unstructured k-walker random walk search as well as a logarithmic DHT search. Further, we argue that such a substrate strikes a balance between the resilience and reliability of unstructured networks and the efficiency of structured networks.

Using Linearization for Global Consistency in SSR

Kendy Kutzner[1] and Thomas Fuhrmann[2]

[1]*Computer Science Department*
University of Karlsruhe
Karlsruhe, Germany
kutzner@ira.uka.de

[2]*Computer Science Department*
Technical University of Munich
Munich, Germany
fuhrmann@net.in.tum.de

Novel routing algorithms such as *scalable source routing* (SSR) and *virtual ring routing* (VRR) need to set up and maintain a virtual ring structure among all the nodes in the network. The *iterative successor pointer rewiring protocol* (ISPRP) is one way to bootstrap such a network. Like its VRR-analogon, ISPRP requires one of the nodes to flood the network to guarantee consistency.

Recent results on self-stabilizing algorithms now suggest a new approach to bootstrap the virtual rings of SSR and VRR. This so-called linearization method does not require any flooding at all. Moreover, it has been shown that linearization with shortcut neighbors has on average polylogarithmic convergence time, only.

Performance Modelling of Peer-to-Peer Routing

Idris A. Rai, Andrew Brampton, Andrew Macquire and Laurent Mathy

Computing Department
Lancaster University
Lancaster, United Kingdom
{*rai, brampton, macquire, laurent*}*@comp.lancs.ac.uk*

We propose several models based on discrete-time Markov chains for the analysis of Distributed Hash Tables (DHTs). Specifically, we examine the Pastry routing protocol, as well as a Stealth DHT adaptation of Pastry to compute their exact expressions for average number of lookup hops. We show that our analytical models match with the protocols' simulation results almost perfectly, making them ideal for rapid evaluation.

Reliable Routing of Event Notifications over P2P Overlay Routing Substrate in Event Based Middleware

Shruti P. Mahambre[1] and Umesh Bellur[2]

[1]*School of Information Technology*
Indian Institute of Technology Bombay
Mumbai, Maharashtra, India
shruti@it.iitb.ac.in

[2]*School of Information Technology*
Indian Institute of Technology Bombay
Mumbai, Maharashtra, India
umesh@it.iitb.ac.in

Event Broker Networks (EBN) are a scalable incarnation of the publish subscribe paradigm for building asynchronous systems. These take the form of overlays of broker nodes and several routing schemes exist that deliver events from publishers to subscribers efficiently on different overlay structures. However quality of service based routing schemes are rare and our work addresses this gap. Specifically we look into the prospect of routing events based on reliability requirements of subscribers for an event type being delivered via the EBN. In this paper, we formally define reliability and propose a multiplicative model which calculates reliability of the P2P overlay routing substrate and an algorithm based on this model, to deliver event notifications to the client. We employ a technique called 'pruning' by which we restrict flooding the entire overlay routing substrate, when finding a reliable path. The complexity analysis of our algorithm shows that it finds a reliable path with a lower message complexity, as compared to the flooding approach. Our algorithm also determines a path with higher reliability than the path established by Hermes. We present initial simulation results, using the Hermes middleware simulator.

PON: Exploiting Proximity on Overlay Networks

Gennaro Cordasco, Alberto Negro, Alessandra Sala and Vittorio Scarano

Dipartimento di Informatca e Applicazioni
Università di Salerno
Fisciano, Italy
{*cordasco, alberto, sala, vitsca*}*@dia.unisa.it*

We define a proximity overlay network (PON) which allow to realize DHT systems whose aim is to combine routing efficiency – i.e. an optimal degree/diameter tradeoff – and proximity awareness.

The proposed systems is parameterized with a positive integer s which measures the amount of flexibility offered by the network. Varying the value of s the system goes from a quite rigid network ($s = 2$) which offer an optimal degree/diameter tradeoff. Increasing s to relatively low values allows to increase the flexibility of the network and consequently improves the stretch, that is, the ratio between the latency of two nodes on the overlay network and the unicast latency between those nodes.

We are able to reconcile the conflict between the load balancing and proximity relationship by proving the efficiency of the main performance metrics. In particular we analytically prove that our system can result in lookup latencies proportional to the maximum latency of the underlying physical network, provided that the physical network has a power law latency expansion.

Proximity-Aware Collaborative Multicast for Small P2P Communities

Francisco De Asís López-Fuentes and Eckehard Steinbach

Institute of Communication Networks Media Technology Group
Technische Universität München
Munich, Germany
{*fcoasis, Eckehard.Steinbach*}*@tum.de*

In this paper we describe a novel solution for delay sensitive one-to-many content distribution in P2P networks based on cooperative m-ary trees. Our scheme maximizes the overall throughput while minimizing end-to-end delay by exploiting the full upload capacities of the participating peers and their proximity relationship. Our delivery scheme is based on cooperation between the source, the content-requesting peers and the helper peers. In our solution, the source splits the content into several blocks and feeds them into multiple m-ary trees rooted at the source. Every peer contributes its upload capacity by being a forwarding peer in at least one of the m-ary trees. Our performance evaluation shows that our proposal achieves similar throughput as the best known solution in the literature (Mutualcast) while at the same time reducing content delivery delay.

Shrack: Description and Performance Evaluation of a Peer-to-Peer System for Document Sharing and Tracking using Pull-Only Information Dissemination

Hathai Tanta-Ngai, Vlado Keselj and Evangelos E. Milios

Computer Science
Dalhousie
Halifax, NS, Canada
{*hathai, vlado, eem*}*@cs.dal.ca*

Shrack is a peer-to-peer framework for document sharing and tracking. Shrack peers provide support to researchers in forming direct collaboration in autonomous sharing and keeping track of newly published documents based on their interests. We propose a pull-only information dissemination protocol for peers to distribute information about new documents among peers with similar interests. Each peer can use the disseminated information to build a local view of semantic overlay of peer interests in the network. Each peer can later use the semantic overlay to find new contact information about other peers with a particular interests, as well as search for documents archived by other peers. After presenting an overview architecture of the system and the dissemination protocol, we present the evaluation results of the system performance, based on a simulated environment. The results indicate that the Shrack protocol is scalable and reliable as the network size increases.

Workshop 20
Workshop on Large-Scale and Volatile Desktop Grids
PCGRID 2007

20 PCGRID • Workshop on Large-Scale, Volatile Desktop Grids

Workshop Description:

Desktop grids utilize the free resources available in Intranet and Internet environments for supporting large-scale computation and storage. For over a decade, desktop grids have been one of the largest and most powerful distributed computing systems in the world, offering a high return on investment for applications from a wide range of scientific domains (including computational biology, climate prediction, and high-energy physics). While desktop grids sustain up to Teraflops/second of computing power from hundreds of thousands to millions of resources, fully leveraging the platform's computational power is still a major challenge because of the immense scale, high volatility, and extreme heterogeneity of such systems. The workshop seeks to bring desktop grid researchers together from theoretical, system, and application areas to identify plausible approaches for supporting applications with a range of complexity and requirements on desktop environments. Moreover, the purpose of the workshop is to provide a forum for discussing recent advances and identifying open issues for the development of scalable, fault-tolerant, and secure desktop grid systems.

Topics of interest include, but are not limited to:

- desktop grid middleware and software infrastructure (including management)
- incorporation of desktop grid systems with Grid infrastructures
- desktop grid programming environments and models
- modeling, simulation, and emulation of large-scale, volatile environments
- resource management and scheduling
- resource measurement and characterization
- novel desktop grid applications
- data management (strategies, protocols, storage)
- security on desktop grids (reputation systems, result verification)
- fault-tolerance on shared, volatile resources
- peer-to-peer (P2P) algorithms or systems applied to desktop grids

General Chairs:

Derrick Kondo, INRIA Futurs, France
Franck Cappello, INRIA Futurs, France

Program Chair:

Gilles Fedak, INRIA Futurs, France

Program Committee:

David Anderson, University of California at Berkeley, USA
Artur Andrzejak, Zuse Institute of Berlin, Germany
MaengSoon Baik, Samsung Research, Korea
Henri Bal, Vrije Universiteit, The Netherlands
Zoltan Balaton, SZTAKI, Hungary
James C. Browne, University of Texas at Austin, USA
Denis Caromel, INRIA, France
Abhishek Chandra, University of Minnesota, USA
Rudolf Eigenmann, Purdue University, USA
JoonMin Gil, Catholic University of Daegu, Korea
Renato Figueiredo, University of Florida, USA
Fabrice Huet, University of Nice Sophia Antipolis, France
Adriana Iamnitchi, University of South Florida, USA
Mario Lauria, Ohio State University, USA
Virginia Lo, University of Oregon, USA
Grzegorz Malewicz, Google Inc., USA
Fernando Pedone, University of Lugano, Switzerland
Serge Petiton, University of Lille, France
Olivier Richard, ID-IMAG, France
Arnold L. Rosenberg, University of Massachusetts Amherst, USA
Mitsuhisa Sato, University of Tsukuba, Japan
Luis Silva, University of Coimbra, Portugal
Alan Sussman, University of Maryland, USA
Michela Taufer, University of Texas at El Paso, USA
Douglas Thain, University of Notre Dame, USA
Bernard Traversat, SUN, USA
Jon Weissman, University of Minnesota, USA
Rich Wolski, University of California at Santa Barbara, USA

Open Internet-based Sharing for Desktop Grids in iShare

Xiaojuan Ren[1], Ayong Basumallik[1], Zhelong Pan[2] and Rudolf Eigenmann[1]

[1]*School of ECE*
Purdue University
West Lafayette, IN, USA
{*xren, basumall, eigenman*}*@purdue.edu*

[2]*Performance Group*
VMWare Inc.
Palo Alto, CA, USA
zpan@vmware.com

This paper presents iShare, a distributed peer-to-peer Internet-sharing system, that facilitates the sharing of diverse resources located in different administrative domains over the Internet. iShare addresses the challenges of resource management in desktop grids, and integrates these resources with production grids. In this paper, we present a brief overview of the iShare system and describe how iShare leverages existing standards to provide novel solutions to the problems of resource dissemination, resource allocation and trust in desktop grids. We also discuss how iShare integrates production grid systems, such as the Teragrid, with desktop resources and compare the iShare approach with web-based user portals for production grids. To quantitatively evaluate our techniques, we measured the efficiency of resource allocation in iShare and the overheads associated with establishing trust and providing the iShare user interface for production grids. The evaluation results demonstrate that iShare enables open Internet sharing with efficiency, reliability, and security.

Decentralized Dynamic Host Configuration in Wide-Area Overlays of Virtual Workstations

Arijit Ganguly, David Wolinsky, P. Oscar Boykin and Renato Figueiredo

Advanced Computing and Information Systems Laboratory
University of Florida
Gainesville, FL 32611, USA
{*aganguly, davidiw, boykin, renato*}*@acis.ufl.edu*

Wide-Area Overlays of Virtual Workstations (WOWs) have been shown to provide excellent infrastructure for deploying high throughput computing environments on commodity desktop machines by (1) offering scalability to a large number of nodes, (2) facilitating addition of new nodes even if they are behind NATs/Firewalls and (3) supporting unmodified applications and middleware. However, deployment of WOWs from scratch still requires setting up a bootstrapping network and managing centralized DHCP servers for IP address management. In this paper we describe novel techniques that allow multiple users to create independent, isolated virtual IP namespaces for their WOWs without requiring a dedicated bootstrapping infrastructure, and to provision dynamic host configuration (e.g. IP addresses) to unmodified DHCP clients without requiring the setup and management of a central DHCP server. We give qualitative and quantitative arguments to establish the feasibility of our approach.

SZTAKI Desktop Grid: a Modular and Scalable Way of Building Large Computing Grids

Zoltán Balaton[1], Gábor Gombás[1], Péter Kacsuk[1], Ádám Kornafeld[1], József Kovács[1], Attila Csaba Marosi[1], Gábor Vida[1], Norbert Podhorszki[2] and Tamás Kiss[3]

[1]*MTA SZTAKI Computer and Automation Research Institute of the Hungarian Academy of Sciences*
Budapest, Hungary
{*balaton, gombasg, kacsuk, kadam, smith, atisu, vida*}*@sztaki.hu*

[2]*Computer Science Department*
University of California, Davis
Davis, CA, United States of America
pnorbert@cs.ucdavis.edu

[3]*Cavendish School of Computer Science*
University of Westminster
London, United Kingdom
T.Kiss@westminster.ac.uk

So far BOINC based desktop Grid systems have been applied at the global computing level. This paper describes an extended version of BOINC called SZTAKI Desktop Grid (SZDG) that aims at using Desktop Grids (DGs) at local (enterprise/institution) level. The novelty of SZDG is that it enables the hierarchical organisation of local DGs, i.e., clients of a DG can be DGs at a lower level that can take work units from their higher level DG server. More than that, even clusters can be connected at the client level and hence work units can contain complete MPI programs to be run on the client clusters. In order to easily create Master/Worker type DG applications a new API, called as the DC-API has been developed. SZDG and DC-API has been successfully applied both at the global and local level, both in academic institutions and in companies to solve problems requiring large computing power.

Direct Execution of Linux Binary on Windows for Grid RPC Workers

Yoshifumi Uemura[1], Yoshihiro Nakajima[2] and Mitsuhisa Sato[3]

[1]*Graduate School of Systems and Information Engineering*
University of Tsukuba
Tsukuba, Japan
uemura@hpcs.cs.tsukuba.ac.jp

[2]*Graduate School of Systems and Information Engineering*
University of Tsukuba
Tsukuba, Japan
ynaka@hpcs.cs.tsukuba.ac.jp

[3]*Graduate School of Systems and Information Engineering*
University of Tsukuba
Tsukuba, Japan
msato@hpcs.cs.tsukuba.ac.jp

Local area or campus-type networks consist of PCs using different operating systems such as Windows and Linux. These PCs are expected to have enormous potential computing power for grid computing. The majority of PCs in this type of environment run on Windows, while grid applications and middleware are often developed on Linux. The challenge is to absorb the heterogeneity of operating systems. Grid RPC is a promising programming model for the development of grid applications. We have designed and implemented an agent called BEE, which enables direct execution of Linux binary programs on Windows for a Grid RPC worker. We have integrated the BEE agent into an OmniRPC system in order to make use of Windows PCs as computing resources in a hybrid grid environment combining Windows PCs into grid computing resources. The BEE agent allows Linux binaries of the program of the OmniRPC worker to be exported and run under Windows without any modification of its Linux binaries. The results of our experiments show that the performance of a worker program using BEE is almost the same as that of Windows native binary and Cygwin and is better that that of using VMware. We have demonstrated a hybrid grid environment combining Window PCs in a conventional grid of Linux nodes.

Local Scheduling for Volunteer Computing

David Anderson[1] and John Mcleod VII[2]

[1]Space Sciences Lab
U.C. Berkeley
Berkeley, CA, USA
davea@ssl.berkeley.edu

[2]652 Crescent Ridge Trail
Mableton, GA, USA
jm7@acm.org

BOINC, a middleware system for volunteer computing, involves projects, which distribute jobs, and hosts, which execute jobs. The local (host-level) scheduler addresses two issues: when to fetch new jobs from a project and, of the currently runnable jobs, which to execute. It seeks to simultaneously satisfy a number of constraints such as maintaining given long-term ratios of work between projects, meeting deadlines for job reporting, and providing variety to the volunteer using uncertain, dynamic information about resources and jobs. We describe these goals and factors, and discuss BOINC's local scheduling policies.

Moving Volunteer Computing towards Knowledge-Constructed, Dynamically-Adaptive Modeling and Scheduling

Michela Taufer[1], Andre Kerstens[1], Trilce Estrada[1], David A. Flores[1], Richard Zamudio[1], Patricia J. Teller[1], Roger Armen[2] and Charles L. Brooks[2]

[1]Dept. of Computer Science
University of Texas at El Paso
El Paso, TX, U.S.A
{mtaufer, akerstens, tpestrada, daflores, rzamudio, pteller}@utep.edu

[2]Dept. of Molecular Biology
The Scripps Research Institute
La Jolla, CA, U.S.A.
{rarmen, brooks}@scripps.edu

Volunteer computing projects supported by BOINC have been exploring new research directions. For example, mature projects like Folding@home are moving towards the use of a broader range of architectures and computers. Other projects such as Docking@Home are exploring multi-scale, resource-driven and application-driven adaptations of the volunteer system.

This paper presents results that enforce the need for knowledge-constructed capabilities in volunteer computing projects, i.e., the capability to drive simulations based on application-results and resource-status. The Docking@Home project, which uses volunteer resources to study putative drugs by computationally simulating the behavior of small molecules (ligands) when docking to a protein, serves as a case study to positively assess two key hypotheses. The first hypothesis claims that the adaptive selection of computational models for docking simulations based on the features of the protein and ligand can positively affect the final accuracy of the prediction. The second hypothesis claims that the adaptive selection of volunteer resources can ultimately improve project throughput.

Towards Deployment Contracts in Large Scale Clusters & Desktop Grids

Francoise Baude, Denis Caromel, Alexandre Di Costanzo, Christian Delbe and Mario Leyton

INRIA Sophia - I3S - CNRS
Universite de Nice Sophia Antipolis
Sophia Antipolis, Alpes Maritimes, France
{*Francoise.Baude, Denis.Caromel, Alexandre.Di_Costanzo, Christian.Delbe, Mario.Leyton*}*@sophia.inria.fr*

While many dream and talk about Service Level Agreement (SLA) and Quality of Service (QoS) for Service Oriented Architectures (SOA), the practical reality of Grid computing is still far from providing effective techniques enabling such contractual agreements.

Towards this goal, this paper provides an overview of the techniques offered by ProActive to set and use contractual agreements. Based on the identification of roles, *application developer, infrastructure manager, application user*, the actors of a Grid environment can specify what is required or what is provided at various levels. The results are both flexibility and adaptability, matching the application constraints and the environment characteristics with various techniques.

Proxy-based Grid Information Dissemination

Deger Cenk Erdil, Michael J. Lewis and Nael B. Abu-Ghazaleh

Department of Computer Science
State University of New York (SUNY) at Binghamton
Binghamton, NY, USA
{*erdil, mlewis, nael*}*@cs.binghamton.edu*

Resource scheduling in large-scale, volatile desktop grids is challenging because resource state is both dynamic and eclectic. Matching available resources with requests is not always possible with existing approaches. Partial dissemination protocols, such as gossiping, may provide efficient schedules when resource requesters are located near providers that can meet their needs. However, when requesters are distant from available resources, regular information dissemination techniques can waste communication bandwidth with futile messages. Thus, it may be advantageous to attempt to advertise to select remote regions of the grid, without necessarily also going through all intermediate nodes. This paper proposes dissemination proxies to increase coverage footprints and reduce dissemination overhead. We incorporate selecting and adjusting the amount of proxy nodes into an adaptive dissemination algorithm, and show that dissemination proxies are able to reduce dissemination overhead, and handle available resource distribution scenarios where regular information dissemination approaches may not produce efficient protocols. We also report initial results that indicate that randomly selecting nodes to serve as proxies can perform as well as strategies that select seemingly better-qualified proxies.

Challenges in Executing Data Intensive Biometric Workloads on a Desktop Grid

Christopher Moretti, Timothy C. Faltemier, Douglas Thain and Patrick J. Flynn

Computer Science and Engineering Department
University of Notre Dame
Notre Dame, IN, USA
{*cmoretti, tfaltemi, dthain, flynn*}*@cse.nd.edu*

Desktop grids have traditionally focused on executing computation intensive workloads. Can they also be used to execute data-intensive workloads? To answer this question, we present a case study of a data intensive biometric application which is infeasible to process on a single machine. We evaluate the capacity of a desktop grid to store and deliver the data need to execute the workload, and compare several general techniques for data deployment. Selecting the most scalable technique, we execute and evaluate five large production workloads on a 350-CPU desktop grid. We observe that this technique is sensitive to many parameters, and propose that an ideal system should be responsible for choosing the proper decomposition of a workload.

Storage@home: Petascale Distributed Storage

Adam L. Beberg[1] and Vijay S. Pande[2]

[1]*Computer Science Department*
Stanford University
Stanford, CA, 94305
beberg@cs.stanford.edu

[2]*Chemistry Department*
Stanford University
Stanford, CA, 94305
pande@stanford.edu

Storage@home is a distributed storage infrastructure developed to solve the problem of backing up and sharing petabytes of scientific results using a distributed model of volunteer managed hosts. Data is maintained by a mixture of replication and monitoring, with repairs done as needed. By the time of publication, the system should be out of testing, in use, and available for volunteer participation.

Applying IC-Scheduling Theory to Familiar Classes of Computations

Gennaro Codasco[1], Grzegorz Malewicz[2] and Arnold L. Rosenberg[3]

[1]*Dipt. di Informatica e Applicazioni*
Università di Salerno
Baronissi, SA, Italy
cordasco@dia.unisa.it

[2]*Dept. of Engineering*
Google Inc.
Mountain View, CA, USA
malewicz@google.com

[3]*Dept. of Computer Science*
University of Massachusetts
Amherst, MA, USA
rsnbrg@cs.umass.edu

Earlier work has developed the underpinnings of *IC-Scheduling theory*, an algorithmic framework for scheduling computations having intertask dependencies for Internet-based computing (IC). The Theory aims to produce schedules that render tasks eligible for execution at the maximum possible rate, so as to: (a) utilize remote clients' computational resources well, by always having work available for allocation; (b) lessen the likelihood that a computation will stall for lack of tasks that are eligible for execution. The current paper reconnects the Theory, which models computations abstractly, with a variety of significant *real* computations and computational paradigms, by illustrating how to schedule these computations optimally.

A combinatorial model for self-organizing networks

Yuri Dimitrov[1], Carlo Giovine[2], Gennaro Mango[2] and Mario Lauria[3,4]

[1]*Dept. of Mathematics*
The Ohio State University
Columbus, OH, USA
yuri@math.ohio-state.edu

[2]*Dip. di Informatica e Sistemistica*
Universita' di Napoli
Naples, Italy
{*carlo.giovine, gennaro.mango*}*@unina.it*

[3]*Dept. of Computer Science and Engineering*
The Ohio State University
Columbus, OH, USA
lauria@cse.ohio-state.edu

[4]*Telethon Institute of Genetics and Medicine*
(TIGEM)
Naples, Italy

In a previous work we have proposed to use of self-organization based on emergent design as a model for the programming of very large aggregates of heterogeneous computing resources. In our approach, a large scale computation is divided in small independent units of computation, each provided with its own uniform, autonomous behavior; only local information is used by each unit of computation to take all the decisions needed to carry out the computation.

One of the challenges of this novel approach is to provide some theoretical foundation that can assist in the rational design of new systems. The purpose of this paper is to demonstrate the use of combinatorial techniques for obtaining quantitative analytical models of the organization pattern emerging from a specific type of self-organizing computation.

Specifically, in this paper we derive an analytical expression describing how nodes distribute themselves over a tree overlay network based on their performance. This result represents a first instance of a theoretical tool that can be used to predict global patterns emerging from a self-organizing computation, and that can be used to establish a direct connection between global features and local behavior parameters.

Workshop 21
Workshop on Multi-Threaded Architectures and Applications
MTAAP 2007

21 MTAAP • Workshop on Multi-Threaded Architectures and Applications

Workshop Description:

Multithreading (MT) programming and execution models are starting to permeate the high-end and mainstream computing scene. This trend is driven by the need to increase processor utilization and deal with the memory-processor speed gap. Recent and upcoming examples of architectures that fit this profile are Cray's Eldorado, IBM Cyclops, and several SMT processors from Sun (UltraSparc T1), IBM (Power5+, Power6), Intel (Xeon with hyperthreading). The underlying rationale to increase processor utilization is a varying mix of new metrics that take performance improvements as well as better power and cost budgeting into account. Yet, it remains a challenge to identify and productively program applications for these architectures with a resulting substantial performance improvement. This workshop intends to identify applications that are amenable to MT and the MT programming and execution models as well as the underlying architectures on which they can thrive. The workshop seeks to explore programming frameworks in the form of MT languages and libraries, compilers, analysis and debugging tools to increase the programming productivity.

Topics of interest include, but are not limited to:

- Multithreaded Performance Metrics and Evaluations
- Multithreaded Architectures
- Multithreaded Programming Framework
- Innovative applications for MT architectures
- Compilation and Optimization for MT architectures
- Multithreaded Libraries and runtime systems
- Multithreaded Performance Analysis and Debugging Tools

Workshop Co-Chairs:

Luiz DeRose (Cray)
Jarek Nieplocha (PNNL)

Program Committee:

David Bader (Georgia Tech)
Frank Baetke (HP)
Calin Cascaval (IBM)
Barbara Chapman (U. Houston)
John Feo (Cray)
Guang Gao (U. Delaware)
Bruce Hendrickson (Sandia National Laboratory)
Andres Marquez (Pacific Northwest National Laboratory)
Michael Merrill (DoD)
Jose Moreira (IBM)
Walid Najjar (U. California Riverside)
Fabrizio Petrini (Pacific Northwest National Laboratory)
Josep Torellas (University of Illinois)
Mateo Valero (Universitat Politècnica de Catalunya)
Jeff Vetter (Oak Ridge National Laboratory, Georgia Tech)

A Heterogeneous Lightweight Multithreaded Architecture

Sheng Li[1], Amit Kashyap[1], Shannon Kuntz[1], Jay Brockman[1], Peter Kogge[1], Paul Springer[2] and Gary Block[2]

[1]*Computer Science and Engineering*
Notre Dame
Notre Dame, IN, USA
{*sli2, akashyap, skuntz, jbb, kogge*}*@nd.edu*

[2]*NASA Jet Propulsion Laboratory*
Pasadena, CA, USA
{*pls, Gary.L.Block*}*@jpl.nasa.gov*

Programs with irregular patterns of dynamic data structures and/or those with complicated control structures such as recursion are notoriously difficult to parallelize efficiently. For some highly-irregular applications, such as a SAT solver, it has been nearly impossible to obtain significant parallel speedups on conventional SMP systems over serial implementations. Lightweight multithreading, as found in the Cray MTA and the upcoming XMT (Eldorado), has been demonstrated as an effective approach to attacking these problems. In this paper, we describe a heterogeneous lightweight multithreading that extends ideas found in the Cray machines to support larger numbers of threads while reducing the cost of thread management and synchronization.

Exploring a Multithreaded Methodology to Implement a Network Communication Protocol on the Cyclops-64 Multithreaded Architecture

Ge Gan, Ziang Hu, Juan Cuvillo and Guang R. Gao

Electrical and Computer Engineering
University of Delaware
Newark, DE, U.S.A
{*gan, hu, jcuvillo, ggao*}*@capsl.udel.edu*

The IBM Cyclops-64 (C64) chip employs a multithreaded architecture that integrates a large number of hardware thread units on a single chip. A cellular supercomputer is being developed based on a 3D-mesh connection of the C64 chips. This paper introduces the Cyclops Datagram Protocol (CDP) developed for the C64 supercomputer system. CDP is inspired by the TCP/IP protocol, yet simpler and more compact. The implementation of CDP leverages the abundant hardware thread-level parallelism provided by the C64 multithreaded architecture.

The main contributions of this paper are: (1) We have completed a design and implementation of CDP that is used as the fundamental communication infrastructure for the C64 supercomputer system. (2) CDP successfully exploits the massive thread-level parallelism provided on the C64 hardware, achieving good performance scalability; (3) CDP is quite efficient. Its peak throughput reaches 884Mbps on the Gigabit Ethernet, even it is running at the user-level on a single-processor Linux machine; (4) Extensive application test cases are passed and no reliability problems have been reported.

OS Mechanism for Continuation-based Fine-grained Threads on Dedicated and Commodity Processors

Shigeru Kusakabe[1], Satoshi Yamada[1], Mitsuhiro Aono[1], Masaaki Izumi[1], Satoshi Amamiya[1], Yoshinari Nomura[2], Hideo Taniguchi[2] and Makoto Amamiya[1]

[1]*Grad. Sch. of Information Sci. & Electrical Eng.*
Kyushu University
Fukuoka, Japan
kusakabe@csce.kyushu-u.ac.jp, {satoshi, aono}@ale.csce.kyushu-u.ac.jp, {masaaki, rogeri}@al.is.kyushu-u.ac.jp, amamiyai@is.kyushu-u.ac.jp

[2]*The Grad. School of Natural Sci. & Tech.*
Okayama University
Okayama, Japan
{nom, tani}@cs.okayama-u.ac.jp

Fine-grained multithreading based on a natural model, such as dataflow model, is promising in achieving high efficiency and high programming productivity. In this paper, we discuss operating system issues for fine-grained multithread programs. We are developing an operating system called CEFOS based on a dataflow based computation model. A program on CEFOS consists of zero-wait threads which run to completion without suspension once started. Firing control among such threads is performed in a dataflow manner along with continuation relations in the program. Target platforms include Fuce processor, which is dedicated to fine-grained multithreading, and commodity processors such as Intel x86. In this paper, after introducing our basic model and our operating system model, we discuss implementation issues on Fuce and commodity platforms. The evaluation results indicate that our approach on commodity platforms is effective in reducing overheads while our approach on a special architecture naturally exploits parallelism even in I/O handling.

On the Role of Deterministic Fine-Grain Data Synchronization for Scientific Applications: A Revisit in the Emerging Many-Core Era

Weirong Zhu, Ziang Hu and Guang R. Gao

Department of Electrical and Computer Engineering
University of Delaware
Newark, DE, USA
{weirong, hu, ggao}@capsl.udel.edu

The design of microprocessor chip for high-end computing systems is moving towards many-core architectures with 10s or 100+ processing units. An important class of the target applications for such architectures are scientific numerical computations, many of which are intrinsically deterministic - that is for a given input a fixed output(result) should be produced no matter how the program is parallelized. It is critical that the read-after-write data dependencies in such programs should be implemented correctly and efficiently via fine-grain data synchronization. In this paper, we investigate the parallelization of three representative scientific computation kernels using fine-grain data synchronization supported by an recently proposed architectural mechanism for many-core chips, called *Synchronization State Buffer (SSB)*. Using detailed simulation on a simulator for the IBM 160-core Cyclops-64 chip architecture with the SSB extension, our experiments demonstrate significant performance advantage of using fine-grain data synchronization based parallelization schemes for scientific workloads.

SWARM: A Parallel Programming Framework for Multicore Processors

David A. Bader, Varun Kanade and Kamesh Madduri

College of Computing
Georgia Institute of Technology
Atlanta, GA, USA
{*bader, varunk, kamesh*}*@cc.gatech.edu*

Due to fundamental physical limitations and power constraints, we are witnessing a radical change in commodity microprocessor architectures to multicore designs. Continued performance on multicore processors now requires the exploitation of concurrency at the algorithmic level. In this paper, we identify key issues in algorithm design for multicore processors and propose a computational model for these systems. We introduce SWARM (**S**oft**W**are and **A**lgorithms for **R**unning on **M**ulti-core), a portable open-source parallel library of basic primitives that fully exploit multicore processors. Using this framework, we have implemented efficient parallel algorithms for important primitive operations such as prefix-sums, pointer-jumping, symmetry breaking, and list ranking; for combinatorial problems such as sorting and selection; for parallel graph theoretic algorithms such as spanning tree, minimum spanning tree, graph decomposition, and tree contraction; and for computational genomics applications such as maximum parsimony. The main contributions of this paper are the design of the SWARM multicore framework, the presentation of a multicore algorithmic model, and validation results for this model. SWARM is freely available as open-source from http://multicore-swarm.sourceforge.net
This work was supported in part by NSF Grants CNS-0614915, CAREER CCF-0611589, DBI-0420513 and ITR EF/BIO 03-31654.

A Comprehensive Analysis of OpenMP Applications on Dual-Core Intel Xeon SMPs

Ryan E. Grant and Ahmad Afsahi

Department of Electrical and Computer Engineering
Queen's University
Kingston, ON, CANADA
ryan.grant@ece.queensu.ca, ahmad.afsahi@queensu.ca

Hybrid chip multithreaded SMPs present new challenges as well as new opportunities to maximize performance. Our intention is to discover the optimal operating configuration of such systems for scientific applications and to identify the shared resources that might become a bottleneck to performance under the different hardware configurations. This knowledge will be useful to the research community in developing software techniques to improve the performance of shared memory programs on modern multi-core multiprocessors.

In this paper, we study a two-way dual-core Hyper-Threaded (HT) Intel Xeon SMP server under single program and multi-program multithreaded workloads using the NAS OpenMP benchmark suite. Our performance results indicate that in the single-program case, the CMP-based SMP and CMT-based SMP configurations have the highest average speedup across all of the applications. The most efficient architecture is a single HT-enabled dual-core processor that is almost comparable to the performance of a 2-way dual-core HT-disabled system.

Improving Scalability of OpenMP Applications on Multi-core Systems Using Large Page Support

Ranjit Noronha and Dhabaleswar Panda

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

Modern multi-core architectures have become popular because of the limitations of deep pipelines and heating and power concerns. Some of these multi-core architectures such as the Intel Xeon have the ability to run several threads on a single core. The OpenMP standard for compiler directive based shared memory programming allows the developer an easy path to writing multi-threaded programs and is a natural fit for multi-core architectures. The OpenMP standard uses loop parallelism as a basis for work division among multiple threads. These loops usually use arrays in their computation with different data distributions and access patterns. The performance of accesses to these arrays may be impacted by the underlying page size depending on the frequency and strides of these accesses. In this paper, we discuss the issues and potential benefits from using large pages for OpenMP applications. We design an OpenMP implementation capable of using large pages and evaluate the impact of using large page support available in most modern processors on the performance and scalability of parallel OpenMP applications. Results show an improvement in performance of up to 25% for some applications. It also helps improve the scalability of these applications.

STAMP: A Universal Algorithmic Model for Next-Generation Multithreaded Machines and Systems

Michel Dubois[1], Hyunyoung Lee[2] and Lan Lin[2]

[1]*Dept. of Electrical Engineering*
University of Southern California
Los Angeles, CA, USA
dubois@paris.usc.edu

[2]*Dept. of Computer Science*
University of Denver
Denver, CO, USA
{*hlee, llin*}*@cs.du.edu*

We propose a generic algorithmic model called STAMP (Synchronous, Transactional, and Asynchronous Multi-Processing) as a universal performance and power complexity model for multithreaded algorithms and systems. We provide examples to illustrate how to design and analyze algorithms using STAMP and how to apply the complexity estimates to better utilize CMP(Chip MultiProcessor)-based machines within given constraints such as power.

Software and Algorithms for Graph Queries on Multithreaded Architectures

Jonathan W. Berry[1], Bruce Hendrickson[1], Simon Kahan[2] and Petr Konecny[3]

[1]*Discrete Math and Algorithms*
Sandia National Labs.
Albuquerque, NM, 87112
{*jberry, bahendr*}*@sandia.gov*

[2]*Research*
Google, Inc.
Seattle, WA, 98033
simon.kahan@gmail.com

[3]*Eldorado Software*
Cray, Inc.
Seattle, WA, 98101
pekon@cray.com

Search-based graph queries, such as finding short paths and isomorphic subgraphs, are dominated by memory latency. If input graphs can be partitioned appropriately, large cluster-based computing platforms can run these queries. However, the lack of compute-bound processing at each vertex of the input graph and the constant need to retrieve neighbors implies low processor utilization. Furthermore, graph classes such as scale-free social networks lack the locality to make partitioning clearly effective.

Massive multithreading is an alternative architectural paradigm, in which a large shared memory is combined with processors that have extra hardware to support many thread contexts. The processor speed is typically slower than normal, and there is no data cache. Rather than mitigating memory latency, multithreaded machines tolerate it. This paradigm is well aligned with the problem of graph search, as the high ratio of memory requests to computation can be tolerated via multithreading.

In this paper, we introduce the MultiThreaded Graph Library (MTGL), generic graph query software for processing semantic graphs on multithreaded computers. This library currently runs on serial machines and the Cray MTA-2, but Sandia is developing a run-time system that will make it possible to run MTGL-based code on Symmetric MultiProcessors. We also introduce a multithreaded algorithm for connected components and a new heuristic for inexact subgraph isomophism. We explore the performance of these and other basic graph algorithms on large scale-free graphs. We conclude with a performance comparison between the Cray MTA-2 and Blue Gene/Light for s-t connectivity.

Analyzing the Scalability of Graph Algorithms on Eldorado

Keith D. Underwood[1], Jonathan Berry[1], Bruce A. Hendrickson[1] and Megan Vance[2]

[1]*Scalable Computing Systems*
Sandia National Laboratories
Albuquerque, NM, USA
{*kdunder, jberry, bahendr*}*@sandia.gov*

[2]*Computer Science and Engineering Department*
University of Notre Dame
Notre Dame, IN, USA
mvance@cse.nd.edu

The Cray MTA-2 system provides exceptional performance on a variety of sparse graph algorithms. Unfortunately, it was an extremely expensive platform. Cray is preparing an Eldorado platform that leverages the Cray XT3 network and system infrastructure while integrating a new revision of the MTA-2 processors that is pin compatible with the AMD Opteron socket. Unlike the MTA-2, this platform will have a more constrained network bisection bandwidth and will pay a high penalty for random memory accesses. This work assesses the hardware level scalability of the Eldorado platform on several graph algorithms.

Advanced Shortest Paths Algorithms on a Massively-Multithreaded Architecture

Joseph R. Crobak[1], Jonathan W. Berry[2], Kamesh Madduri[3] and David A. Bader[3]

[1]Dept. of Computer Science
Rutgers University
Piscataway, NJ, USA
crobakj@cs.rutgers.edu

[2]Sandia National Laboratories
Albuquerque, NM, USA
jberry@sandia.gov

[3]College of Computing
Georgia Institute of Technology
Atlanta, GA, USA
{kamesh, bader}@cc.gatech.edu

We present a study of multithreaded implementations of Thorup's algorithm for solving the Single Source Shortest Path (SSSP) problem for undirected graphs. Our implementations leverage the fledgling MultiThreaded Graph Library (MTGL) to perform operations such as finding connected components and extracting induced subgraphs. To achieve good parallel performance from this algorithm, we deviate from several theoretically optimal algorithmic steps. In this paper, we present simplifications that perform better in practice, and we describe details of the multithreaded implementation that were necessary for scalability.

We study synthetic graphs that model unstructured networks, such as social networks and economic transaction networks. Most of the recent progress in shortest path algorithms relies on structure that these networks do not have. In this work, we take a step back and explore the synergy between an elegant theoretical algorithm and an elegant computer architecture. Finally, we conclude with a prediction that this work will become relevant to shortest path computation on structured networks.

Probability Convergence in a Multithreaded Counting Application

Chad Scherrer[1], Nathaniel Beagley[1], Jarek Nieplocha[1], Andres Marquez[1], John Feo[2] and Daniel Chavarria-Miranda[1]

[1]Pacific Northwest National Laboratory
Richland, WA, USA
{chad.scherrer, nathaniel.beagley, jarek.nieplocha, andres.marquez, daniel.chavarria}@pnl.gov

[2]Cray, Inc.
Seattle, WA, USA
feo@cray.com

The problem of counting specified combinations of a given set of variables arises in many statistical and data mining applications. To solve this problem, we introduce the *PDtree* data structure, which avoids exponential time and space complexity associated with prior work by allowing user specification of the tree structure. A straightforward parallelization approach using a Cray MTA-2 provides a speedup that is linear in the number of processors, but introduces nondeterminism into probability estimates. We prove a general convergence result that bounds the nondeterministic deviation of probability estimates relative to a sequential implementation. Beyond PDtrees, this convergence result applies to any counting application that takes a multithreaded streaming approach.

Index