A Look at Application Performance Sensitivity to the Bandwidth and Latency of Infiniband Networks.

Darren J. Kerbyson

Performance and Architecture Lab (PAL)
Los Alamos National Laboratory
NM 87544 USA
djk@lanl.gov

Abstract

This work explores the expected performance of three applications on a High Performance Computing cluster interconnected using Infiniband. In particular, the expected performance across a range of configurations is analyzed notably Infiniband 4x, 8x and 12x representing link-speeds of 10Gb/s, 20Gb/s, and 30Gb/s respectively as well as near-neighbor MPI message latencies of 4µs and 1.5µs. In addition we also consider the impact of node size, from one to eight processors that share a single network connection. The performance analysis is based on the use of detailed performance models of the three applications developed at Los Alamos. The results of the analysis show that the application performance can range by as much as 60% from best to worst. The relative importance of bandwidth, latency and node size differs between the applications.

1. Introduction

The importance of analyzing, and understanding performance increases as both systems and applications grow in size and in complexity. The performance of a system results from an interplay between the hardware architecture, the communication system, and the applied workload. Knowledge of the processor design, memory hierarchy, inter-processor and network system, and workload arrangement is necessary in order to understand the factors that impact the achievable performance.

At Los Alamos National Laboratory (LANL), we have been developing accurate analytical performance models for some time. The approach that we take is application centric. Constructing a performance model requires an in-depth analysis and understanding of both application and system aspects. In this way we can investigate the performance impact that both system changes and code changes will have. The performance models have been previously validated across many systems and used in the comparison of large-scale systems such as several terascale systems compared to the Earth Simulator [6], the optimization of ASCI Q during installation [11], and in the exploration of possible future systems [5].

In this work we investigate the impact on application performance of Infiniband networks, with different bandwidth and latency characteristics using our performance models. The Infiniband architecture is an industry standard that offers low latency and high bandwidth network communications. It is increasingly becoming popular for building high-performance-clusters. Current configurations of Infiniband operate at either 4x or 8x. The peak data rate of is 10Gb/s and 20Gb/s respectively. However, studies on the performance of MPI level communications has shown that a peak of approximately 900MB/s (unidirectional) for 4x, and approximately 1.6GB/s for 8x are possible [9]. Typical latencies are in the range of 4 to 6µs.

The performance of three applications of interest to Los Alamos is studied in this work on a multitude of potential Infiniband configurations. In particular bandwidths of 4x, 8x and 12x are considered as well as possible near-neighbor MPI latencies of 4µs and 1.5µs. By considering such performance ranges, we can examine the sensitivity of application performance using our models in-advance of procurement for instance. In addition we also consider the impact of node-size (number of processors in a node) on application performance. The node size can lead to increased communication contention on the local NIC if all processors compete for the inter-node communication links at the same time.

In Section 2 we detail our analysis approach using the application performance models. In Section 3 we present the results of this analysis for a multitude of configurations (bandwidths, latencies, and node-sizes). The sensitivity of application performance to the network characteristics is discussed in Section 4. A summary of this work is given in Section 5.

1-4244-0054-6/06/$20.00

2. Approach

To consider the performance that may be achievable from a cluster interconnected with Infiniband of various performance characteristics we utilize detailed performance models of three applications of interest. The performance models of these codes have been previously validated on numerous systems including large-scale ASC machines (ASCI Red, Bluemountain, White, Q, Redstorm, BlueGene), with high accuracy (typically to within 10% of measurements).

An overview of these applications is given in Section 2.1 below along with an empirical analysis of their typical processing and communication characteristics. In Section 2.2 we detail the range of cluster configurations considered in this analysis. In particular we detail our assumptions on the possible performance of the Infiniband interconnect.

2.1. Applications

Three applications were used in this analysis that have their origins in the Department of Energy Accelerated Strategic Computing program - formally known as ASCI. It is not the intention of this work to provide details on the applications themselves or how they are utilized, but rather to examine the impact of Infiniband network configurations on their performance.

Partisn – This application is an implementation of S_N transport, the solution of the Boltzmann equation using the discrete ordinates method, on structured meshes. Some details on Partisn can be found in [1].

SAGE – This is an adaptive mesh hydrodynamics application that is used for the simulation of shock-waves. A detailed description of RAGE, a derivative of SAGE, can be found in [2]. The performance characteristics of SAGE have been studied and modeled in some detail [4].

Sweep3D – This is a kernel application which implements the main processing involved in S_N transport calculations. A description of the algorithm is given in [7]. The performance of Sweep3D has been measured and modeled on many systems e.g. [3].

The characteristics of these applications are listed in Table 1. All three of the applications are typically executed in a weak-scaling mode – that is the sub-grid size per processor remains a constant and increased parallelism is used to increase the global grid size. The sub-grid sizes as specified by an example input deck for each application are listed in Table 1. The applications are iterative and hence the total application run-time is a multiple of the iteration-time. In addition it is also assumed that the applications are run separately, utilizing all of a system or a contiguous part of it.

Table 1. Application characteristics.

	Partisn	**SAGE**	**Sweep3D**
Input Deck	Rep1-5x5	TimingH	standard
Scaling	Weak	Weak	Weak
Sub-grid size	400x5x5	35K cells	5x5x400
Logical topology	2-D mesh	1-D (+)	2-D mesh
Iteration time (s)	6.0	3.6	0.7
Message count /PE/iteration	100K	4K	2K
Typical Message sizes (bytes)	8 400 20,000	8 8,000 150,000	1,200
Collectives	Allreduce Broadcast	Allgather Allreduce Alltoall Reduction	Allreduce Broadcast

Communication characteristics including typical message sizes and the number of such messages that occur per iteration are also listed in Table 1. It can be seen that Sweep3D typically has a large number of small messages whereas SAGE typically has larger sized messages. The message sizes in Partisn are both small and large.

The message sizes mostly represent the surfaces that are communicated between processors when exchanging sub-grid boundary data. They arise from the methods used to parallelize the global grids as well as the way in which sub-grids are processed within each application. The logical topology of Sweep3D and Partisn is 2-D resulting in each processor having at most four neighbors. The sub-grids are processed in smaller blocks in both Partisn and Sweep3D resulting in small message sizes (typically 400 and 1200 bytes respectively). Partisn also exchanges full sub-grid boundaries leading to the additional larger message sizes. SAGE uses a 1-D partitioning that result with communications to only two neighbors on a small scale system. However, as shown in [4], the distance between logical neighbors gradually increases with the processor count, and the number of neighbors can also double. In addition the message sizes can also increase with processor count.

2.2. Cluster configurations

The performance of each application is examined for a cluster consisting of Opteron processors interconnected using Infiniband. Three cluster configuration parameters are varied in this analysis:

(1) Infiniband large-message bandwidth: 4x (10Gb/s), 8x (20Gb/s), and 12x (30Gb/s),
(2) Infiniband small message near-neighbor latency of 4µs or 1.5µs,
(3) Node size in terms of the number of processors in a node sharing a single NIC.

The achievable communication performance is often limited by the speed of the Peripheral Component Interconnect (PCI) bus. PCI-X can support an aggregate of 1GB/s limiting the performance of 4x Infiniband bi-directional communications whereas the newer PCI Express can achieve 2GB/s in each direction matching the Infiniband 8x performance [9].

The small-message latency of 4µs is in the range that can be currently achieved with PCI based NICs. Pathscale has recently introduced a NIC that connects directly to Hypertransport in Opteron based nodes [10]. This has a best quoted latency of 1.29µs. We do not expect latency in the future to decrease much below this figure.

Note that the number of processors here could be considered equal to the number of cores if multi-core processor chips are utilized. For instance considering a node with two dual-core processors to be similar to a node with four single-core processors. This simplification results in a similar use of the network but does not consider the possible contention on the memory buses within the node. From our experience the memory contention within an Opteron dual-core based node is significant whereas the contention within a single-core based node is not.

The peak link bandwidths along with the assumed achievable MPI bandwidths as well as near-neighbor MPI latencies are summarized in Table 2. The listed latency and bandwidth values are used as input to the application performance models in Section 3. The node size is varied between one and eight processors. Any contention that results from using all processors within a node is not included in this analysis.

In addition, NIC contention is fully considered in the analysis but congestion within the network (due to messages colliding within the switch fabric and causing delays) is assumed not significant. This can arise when adaptive routing is used in network.

Table 2. Assumed performance characteristics of the different Infiniband configurations.

Infiniband	Link-speed	Assumed MPI	
		Near-neighbor Latency	Bandwidth
4x	10Gb/s	4µs or 1.5µs	0.9GB/s
8x	20Gb/s	4µs or 1.5µs	1.6GB/s
12x	30Gb/s	4µs or 1.5µs	2.4GB/s

Note that the MPI bandwidths are based on measurements on current systems for the 4x and 8x cases [8,9], and assumed for the 12x case. The network latency is dependent on the number of switch levels a message traverses between the source and destination nodes. In this analysis the switch latency is assumed to be 200ns with each switch having 24 ports (12 down and 12 up). Thus the message latency increases with the distance between source and destination nodes.

3. Predictive Results

The application performance models were used with the system and network configurations parameters as detailed previously. For each application we consider varying our three main variables of network bandwidth, near-neighbor network latency, and node size independently of each other. This ultimately results in quantifying the application performance sensitivity to each of the parameters.

In order to limit the results presented we first consider the impact on communication performance while varying the network bandwidth in Section 3.1. We then consider the impact of overall application execution time when considering the full range of configurations for three system sizes – containing 256, 512 and 1024 processors in Section 3.2.

3.1. Communication component

Sweep3D

The performance of Sweep3D is considered using sub-grids of size 5x5400 cells per processor in a weak scaling mode. The application contains two blocking parameters which were fixed at 10 k-planes and 3 angles per block. In Figure 1 (first column) the cost of communications per Sweep3D iteration is plotted for 4x, 8x, and 12x Infiniband bandwidths, and for a node-size of between one and eight processors (the rows in Figure 1). The near-neighbor MPI latency was 4µs, and the switch latency per hop was 200ns in all cases.

It can be seen in Figure 1 that the communication costs in Sweep3D increase with scale. The first part of the curves (up to 16 processors) results from the gradual increase in number of logically neighboring sub-grids in a 2-D parallelization scheme – for instance on eight processors there are at most 3 neighbors, where as on 16 or more processors there are 4 neighbors. Above 16 processors the communication costs gradually increase with processor count – this is a result of an application pipeline effect (see [3]).

At processor counts of 16 and above the increased bandwidth of a 12x Infiniband network when compared with a 4x network results in a reduction in communication

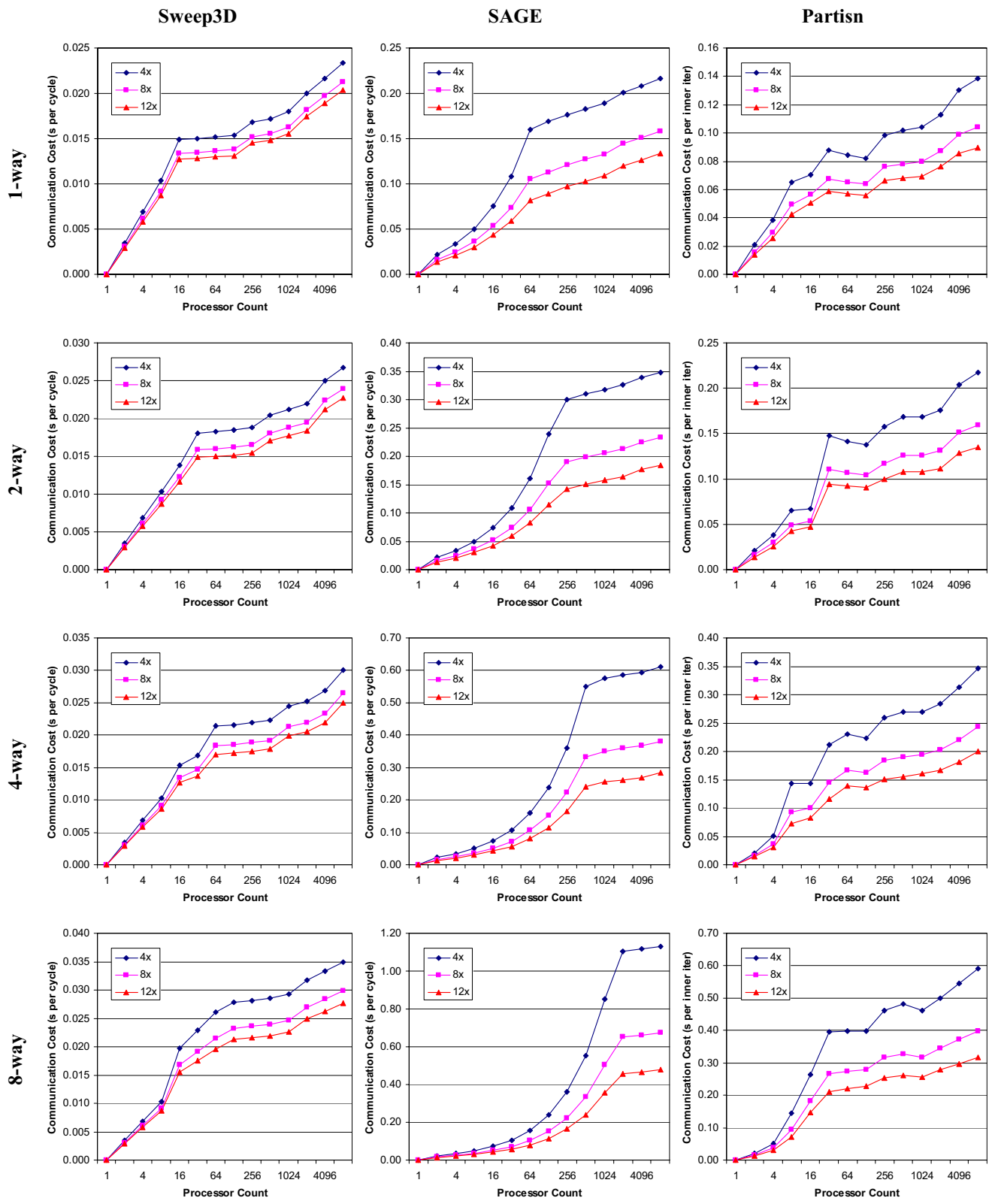

Figure 1. Expected application communication costs for Infiniband 4x, 8x, and 12x with an MPI latency of 4µs.

cost by only 15-20%. Increasing the node size, from one to eight processors, results in the communication cost increasing by approximately 50%.

Note that the communication costs decreases on average by 45% when reducing the MPI latency to 1.5µs. i.e. the communication performance increases on average by approximately a factor of 2. The message sizes in Sweep3D are small and hence messaging costs are more sensitive to latency than to bandwidth. For brevity, this data was not included in this work.

SAGE

The performance of SAGE is considered using an input deck which assigns 35,000 cells to each processor in a weak-scaling mode. The cost of communications per SAGE iteration is shown for 4x, 8x, and 12x Infiniband bandwidths in the second column in Figure 1, and for a node-size of between one and eight processors. The near-neighbor MPI latency was again 4µs, and the switch latency per hop was 200ns in all cases.

The communication costs vary to a greater extent than with Sweep3D. For instance at a processor count of 8,192 the communication performance when using a 12x network is approximately twice that of a 4x network. Also when increasing the node size from one to eight processors, the communication cost increases by over a factor of four.

The interesting shape of the curves, an almost exponential increase followed by a gradual increase, is a characteristic of SAGE resulting from the 1-D parallelization of the spatial grid. The knee in the curve occurs at an increasing processor count as the node size increases. The knee actually corresponds to the scale when all processors within a node undertake inter-node communications on a boundary exchange. Below this knee, there is a mix of intra- and inter-node communications whereas above this knee, all communications are inter-node [4].

The communication costs decreases on average by 15% when reducing the MPI latency to 1.5µs. The message sizes in SAGE are typically 150KB and hence the communication performance is more sensitive to bandwidth than to latency.

Partisn

The performance of Partisn is considered using the an input deck that assigns a sub-grid of 400x5x5 sub-grid (10,000 cells) to each processor in a weak scaling mode. In Figure 1 (third column) the cost of communications per Partisn iteration is shown for 4x, 8x, and 12x Infiniband bandwidths, and for a node-size of between one and eight processors. The MPI latency was again 4µs, and the switch latency per hop was 200ns in all cases.

The communication costs increase reasonably linearly with the processor count. This is an algorithmic effect rather than a parallelization effect. A conjugant-gradient solver is used as part of the processing in Partisn whose number of iterations required for convergence increases with processor count (grid size). At 8,192 processors a 12x network has a 45% higher performance than a 4x network. Also the increase in node size, from one to eight processors, results in the communication cost increasing by between a factor of three to four.

The communication costs decreases on average by 25% when reducing the MPI latency to 1.5µs. The message sizes in Partisn are a mixture of small and large resulting from two distinct computational phases. The transport phase (similar to Sweep3D) has small messages whereas the diffusion phase has larger messages. Thus the communication cost is sensitive to both latency and to bandwidth.

3.2 Overall run-time

In order to compare the performance of the applications across the network performance and node size ranges the computational cost must also be considered. The communication costs for the three applications as shown in Figure 1 by itself does not show depict the overall impact on application performance.

The single processor performance is an input to each of the application performance models. In this analysis we use the single processor performance obtained from a 2GHz AMD Opteron processor. We compare the expected performance of a cluster with the different Infiniband configurations and cluster node-sizes for a system containing 256, 512, and 1024 processors.

The relative performance of different system configurations to a *baseline* configuration is considered. The baseline is taken to consist of Infiniband 4x with 4 processors per node. The relative performance has a positive value if the configuration has a higher performance than that of the baseline, and a negative value if it has a lower performance.

The relative performance for the three applications is plotted in Figures 2a) to 2c) in which the cluster size varies between 256 processors and 1024 processors. In these graphs, the relative performance is indicated for a particular node size (from one to eight processors), and a particular application by a vertical bar. The bottom and top of each vertical bar indicates the relative performance (compared with the baseline system) for the 4x and 12x bandwidths respectively while the middle white line indicates the 8x relative performance. The solid vertical bars are for the case of a 4µs near-neighbor latency, and the shaded bars are for a 1.5µs near-neighbor latency.

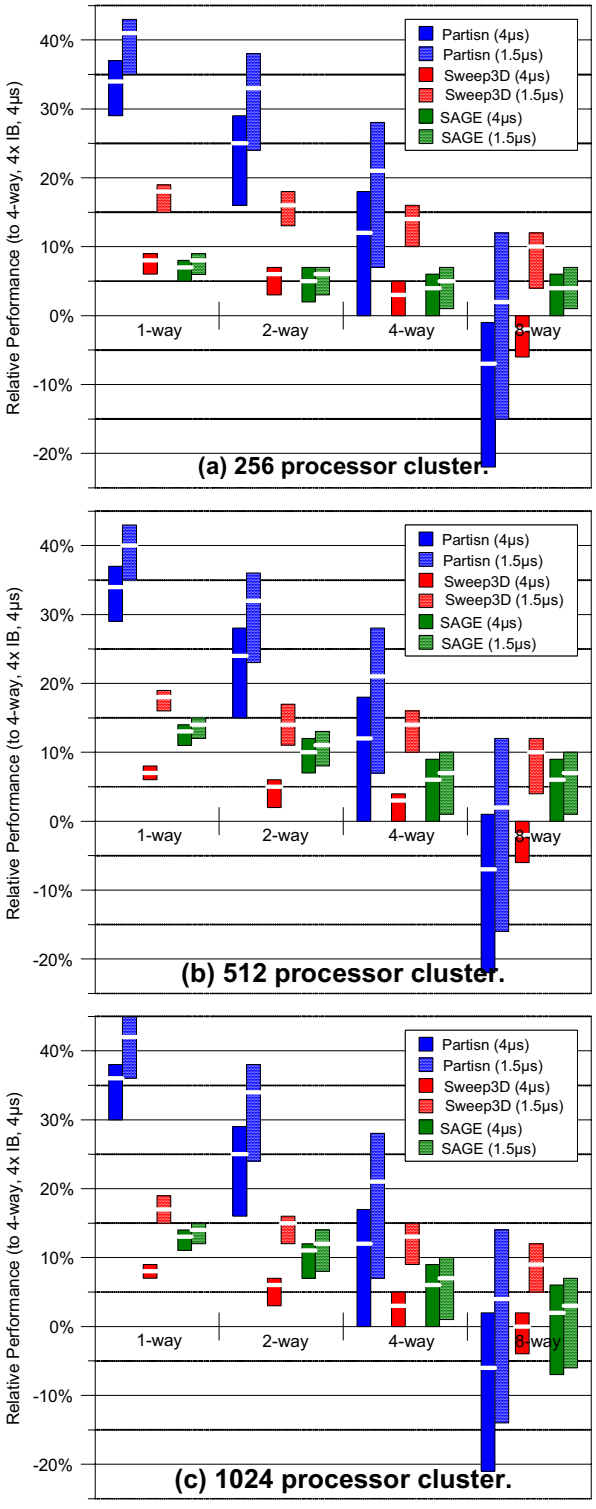

Figure 2. Performance relative to a baseline configuration (4-way nodes with 4µs latency).

4. Discussion

A lot of information is contained in the graphs of Figure 2. It should be first noted that the height of the vertical bars gives an indication of the sensitivity of application performance to the bandwidth of the network. Secondly the difference in vertical position between a solid bar (4µs latency) and the associated shaded bar (1.5µs latency) gives an indication of latency sensitivity. Thirdly the sensitivity to node size can be seen by comparing like-bars across the range in the X-axis (between one and eight processors per node). The worst performance is always that of an 8-way node with 4x bandwidth. The best performance is always with a 1-way node and 12x bandwidth. Also note that the performance of the baseline configuration (4-way nodes with Infiniband 4x and MPI near-neighbor latency of 4µs) relative to itself is always zero.

From this analysis it can be seen that the performance of Partisn is sensitive to both to the bandwidth (the large height of the vertical bars) and the latency (difference in position between the solid and associated shaded bar) of the network. The performance of Sweep3D has a high sensitivity to the network latency. The performance of SAGE has a high sensitivity to the network bandwidth and is impacted very little by the network latency.

Figure 2 considers the impact on application performance when one, two or all three parameters in the study are varied. A summary of the change of application performance resulting from independently changing only one parameter at a time is given in Table 3 for the three processor counts. The change in performance is considered when changing the near-neighbor MPI network latency (from 4µs to 1.5µs), when changing network bandwidth (from 4x to 8x), and when changing node size (from 4-way to 2-way). In all three cases the performance of each characteristic is improving by approximately a factor of two.

Independently changing the network latency, network bandwidth and node size gives a more concise view of their relative impact on application performance. The performance of Sweep3D would benefit the most from an improvement in network latency. The performance of SAGE would benefit the most from an improvement in network bandwidth at the smallest sized system (256 processors) and the most from a reduction in node size for the larger two sized systems. The performance of Partisn would also benefit the most from a reduction in node size.

It is also interesting to note that Partisn has the highest potential improvement across the three applications. This is a reflection of the higher communication to computation ratio exhibited by the application as indicated by the message count and message sizes per iteration indicated in Table 1.

Table 3. Performance change when changing network bandwidth (4x to 8x), latency (4µs to 1.5µs), or node-size (4-way to 2-way).

PE count		Sweep3D (%)	SAGE (%)	Partisn (%)
	Latency	9.9	0.8	7.3
256	Bandwidth	3.4	4.0	11.8
	Node size	3.4	1.7	16.4
	Latency	9.7	0.8	7.1
512	Bandwidth	3.3	6.2	11.9
	Node size	2.0	6.8	15.4
	Latency	9.2	0.9	7.4
1024	Bandwidth	3.1	6.3	11.8
	Node size	3.2	7.3	16.4

Although the improvements as indicated in Table 3 are possible by changing any or all of the configuration parameters, it does not say anything about cost. For instance, halving the node-size for a system with constant processor count would require a relative increase in the number of NICs and switches in the network. Also the cost of the network generally increases with the network bandwidth and reduction in network latency. Such a cost-performance analysis is beyond the scope of this work and is very much a moving target. This information could be used in procurement activities – answering the questions on what improvement in performance would be achievable if a particular configuration was available.

5. Summary

We have analyzed the impact on application performance of a multitude of potential Infiniband network configurations using detailed application performance models. The effect of reducing network latency, increasing network bandwidth, and reducing the node size (processors per node) have all been analyzed on three applications of interest to Los Alamos.

The analysis has once again shown that the possible improvement in performance is workload dependent and varies from application to application. For instance the performance of Sweep3D is most impacted by network latency, whereas the performance of Partisn is most impacted by node size.

When taking the best performance to worst performance across the range in configurations we note that the range in application performance is 68% for Partisn, 15% for SAGE and 24% for Sweep3D.

The performance information contained in this work could be used in either a procurement process or in aiding in the design of future systems. However, in specifying a particular cluster configuration the price of individual performance characteristics need also to be considered if options are available.

Acknowledgements

This work was funded in part by the Accelerated Strategic Computing (ASC) program of the Department of Energy, and by the DARPA High Productivity Computing Systems program in collaboration with the IBM PERCS project. Los Alamos National Laboratory is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36.

References

[1] R.S. Baker. A Block Adaptive Mesh Refinement Algorithm for the Neutral Particle Transport Equation. *Nuclear Science & Engineering*, 141(1):1-12, 2002.

[2] M. Gittings, R. Weaver, M. Clover, T. Betlach, N. Byrne, R. Stefan, D. Ranta. The RAGE Radiation-hydrodynamic code. To appear in *J. Computational Physics*, 2006.

[3] A. Hoisie, O. Lubeck, H.J. Wasserman. Performance and Scalability Analysis of Teraflop-Scale Parallel Architectures using Multidimensional Wavefront Applications. *Int. J. of High Performance Applications*, 14(4):330-346, 2000.

[4] D.J. Kerbyson, H.J. Alme, A. Hoisie, F. Petrini, H.J. Wasserman, M.L. Gittings. Predictive Performance and Scalability Modeling of a Large-scale Application. In *IEEE/ACM Supercomputing (SC'01)*, Nov. 2001.

[5] D.J. Kerbyson, A. Hoisie, H. Wasserman. Exploring Advanced Architectures using Performance Prediction. In *Innovative Architecture for Future Generation High Performance Processors and Systems*, A. Veidenbaum and K. Joe (Eds), IEEE Computer Society, pp. 27-40, 2002.

[6] D.J. Kerbyson, A. Hoisie, and H.J. Wasserman. A Performance Comparison between the Earth Simulator and other Top 5 Terascale Systems on a Characteristics ASCI Workload. *Concurrency and Computation: Practice and Experience*, 17(10):1219-1238, Aug. 2005.

[7] K.R. Koch, R.S. Baker, R.E. Alcouffe. Solution of the First-Order Form of the 3-D Discrete Ordinates Equation on a Massively Parallel Processor. *Trans. of the American Nuclear Soc.*, 65:198-199, 1992.

[8] J. Liu, B. Chandrasekaran, W. Yu, J. Wu, D. Buntinas, S. Kini, P. Wyckoff, D.K. Panda. Micro-Benchmark Performance Comparison of High-Speed Cluster Interconnects, *IEEE Micro*, 14(1):42-51, 2004.

[9] J. Liu, A. Mamidala, A. Vishnu, D.K. Panda. Performance Evaluation of InfiniBand with PCI Express. *IEEE Micro*, 25(1):20-29, 2005.

[10] Pathscale Infinipath HTX Adapter: Low-Latency Cluster Interconnect for Infiniband. Available from http://www.pathscale.com

[11] F. Petrini, D.J. Kerbyson, S. Pakin. The Case of the Missing Supercomputer Performance: Achieving Optimal Performance on the 8,192 Processors of ASCI Q. *IEEE/ACM Supercomputing (SC'03)*, 2003.