Performance Modeling the Earth Simulator and ASCI Q

Darren J. Kerbyson, Adolfy Hoisie, Harvey J. Wasserman

Performance and Architectures Laboratory (PAL)
Modeling, Algorithms and Informatics Group, CCS-3
Los Alamos National Laboratory
Los Alamos, NM 87545
{djk,hoisie,hjw}@lanl.gov

Abstract

This work gives a detailed analysis of the relative performance between the Earth Simulator and ASCI Q. The Earth Simulator uses vector processing nodes interconnected using a single-stage cross-bar network, whereas ASCI Q is built using superscalar Alpha microprocessors and the Quadrics Elan3 interconnection network. The performance that can be achieved on a system results from an interplay of the system characteristics, application requirements and scalability behavior. Detailed performance models are used here to predict the performance of two codes representative of ASCI computations, namely SAGE and Sweep3D. The performance models encapsulate fully the behavior of these codes and have been previously validated on many large-scale systems. They do not require access to a full sized system but rather rely on characteristics of the system as well as knowledge of the achievable single-processor performance. The models are used to compare the performance of the two machines. One result of this analysis is to size an Alpha System that will have the same performance as the Earth Simulator.

1. Introduction

In this work we compare the performance of the Earth Simulator [10,13], and ASCI Q which will be ranked as the top two systems in the top500 list[1] in June 2003. The Earth Simulator was installed at Yokohama City, Japan in March 2002 and has a peak processing rate of 40Tflops. It had the design goal of giving a 1000-fold increase in the processing capability for atmospheric research, compared to that available when envisioned in 1996, and was initially expected to achieve a sustained performance of 5Tflops (12.5% of system-peak) on atmospheric applications [19]. It has subsequently been demonstrated that 60% of system-peak performance can be achieved on an atmospheric simulation code [14], with other codes achieving up to 40% of peak [20]. ASCI Q was recently installed at Los Alamos National Laboratory (LANL) and has a peak processing rate of 20Tflops.

At present there is much interest in comparing the relative performance between the Earth Simulator and other large-scale systems, in part due to the use of vector processors in comparison to superscalar microprocessors with cache-based memory systems. Most other current Terascale systems are built using more high-volume or COTS (Commodity-Off-The-Shelf) based components. The Earth Simulator is at one end of the spectrum - a system built using low-volume processors and a customized network supplied by a single manufacturer to provide a unique system. At the other end of the spectrum there are high-volume, mass produced, computational nodes which can be obtained from many suppliers and that can use more than one network via standard interfaces. The current highest machines as listed in the top500 fall into different categories in this spectrum. ASCI Q compute nodes are manufactured by a sole supplier (HP), and use the Quadrics network, again a sole supplier. Similarly, ASCI White at Lawrence Livermore National Laboratory (LLNL) uses nodes solely supplied by IBM, but in higher volume than those in the Earth Simulator. MCR, again at LLNL, uses high-volume mass-produced compute nodes which can be obtained from a multitude of suppliers.

It is a complex task to compare the performance of these systems without using simplistic metrics (such as peak flop rating). Thus, comparing the performance of the Earth Simulator with other systems has been restricted so far to a small number of applications that have actually been executed on all systems, or to considering the overall system-peak performance or individual sub-system performance characteristics such as

[1] http://www.top500.org/

achieved MPI intra-node communication bandwidth and latencies [19]. Peak performance nor any of these low-level benchmarks correlates with the time-to-solution for a particular application, the metric of interest in our analysis.

The peak performance of a system results from the underlying hardware architecture including processor design, memory hierarchy, inter-processor communication system, and their interaction. But, the achievable performance is dependent upon the workload that the system is to be used for, and specifically how this workload utilizes the resources within the system.

Performance modeling is a key approach that can provide information on the expected performance of a workload on a given architecture. The modeling approach that we take in this work is application centric. It involves an understanding of the processing flow in the application, the key data structures, and how they use and are mapped to the available resources. From this, a model is constructed that encapsulates key performance characteristics. The aim of the model is to provide insight. By keeping the model general, while not sacrificing accuracy, the achievable performance may be explored for new situations – in terms of both hardware systems and code modifications. This approach has been successfully used on applications that are representative of the ASCI workload including: an adaptive mesh hydro-code [7], structured and unstructured mesh transport codes [4,9], and a Monte-Carlo particle simulation code [11].

We use two existing application performance models in this paper, an S_N transport application on Cartesian grids [4] and a hydro-code [7], in order to compare the relative performance between the Earth Simulator and ASCI Q. The type of computations represented by these codes take a very large fraction of the cycles on all ASCI systems. The models have already been validated with high accuracy on ASCI machines constructed to date. They have also been used to validate performance during the installation of the first phase of ASCI Q [8] and in the initial stages of the procurement of ASCI Purple – a 100Tflop machine to be installed in 2005. The models encapsulate the performance characteristics of the processing nodes, the communication networks, the mapping of the sub-domains to processors, and the processing flow of the application along with their scaling behavior. The performance models are used to predict the time-to-solution of the applications when considering a typical weak-scaling utilization of each system. They are also used to “size” an alphaserver system that will provide the same performance as the Earth Simulator.

In Section 2 an overview of the Earth Simulator and ASCI Q is given along with their salient performance characteristics. In Section 3 we describe the characteristics of the two applications. In Section 4, we compare the performance of the applications on the two systems. This work builds upon an initial study of a performance comparison between the Earth Simulator and AlphaServer systems [6].

2. A Comparison of the Earth Simulator and ASCI Q

The Earth Simulator and ASCI Q are compared below considering some of their main characteristics. This is followed by analysing a representative ASCI workload described in Section 3.

The Earth Simulator

The Earth Simulator consists of 640 nodes inter-connected by a single stage 640x640 crossbar network [17]. Each node is an SMP composed of 8 arithmetic processors, a shared memory of 16GB, a remote access unit (RCU), and an I/O processor (IOP). Each arithmetic processor contains 8 vector units each with 8 vector pipes, a 4-way super-scalar processor and operates at 500MHz. Each is connected to 32 memory units with a bandwidth of 32GB/s (256GB/s aggregate within a node). The peak performance of each arithmetic processor is 8Gflops. The RCU connects a node to the crossbar network. The peak inter-node communication bandwidth is 16GB/s in each direction. Each communication channel is 128bits wide in each direction with a peak transfer rate of 1Gbits per second per wire. The minimum latency for an MPI level communication is quoted as 5.2μsec within a node, and 8.6μsec between nodes [10]. It should also be noted that the Earth Simulator lead to the development of the NEC SX-6 product line [15]. An Earth Simulator node has better memory performance than that of the NEC SX-6, but with a smaller memory capacity.

ASCI Q

ASCI Q, installed at Los Alamos National Laboratory (LANL) [1], is the latest ASCI system with a peak performance of 20Tflop containing 2048 HP AlphaServer ES45 nodes. Each node is a 4-way SMP containing four 21264D EV68 Alpha microprocessors operating at 1.25GHz with 64KB L1 instruction and data cache, 16MB unified L2 cache, and 16GB main memory. A peak memory bandwidth up to 8GB/s is possible within a node using two 256-bit memory buses running at 125MHz. Four independent standard 64-bit PCI buses (running at 66MHz) provide I/O. Nodes are interconnected using two rails of the Quadrics QsNet fat-tree network. The peak bandwidth achievable on an MPI level communication is 300MB/s with a typical latency of 5µsec. The latency increases slightly with the physical distance between nodes. A detailed description of the Quadrics network can be found in [12].

A summary of these two systems is listed in Table 1. Peak performance characteristics as well as configuration details are included. It is clear that the peak performance of the Earth Simulator exceeds that of ASCI Q, and also that the main memory bandwidth per processor is far higher. The inter-node communication performance in terms of latency is worse on the Earth Simulator when compared to ASCI Q (using the Quadrics network) but is better by almost a factor of 40 in terms of bandwidth. Note that the inter-node MPI communication performance listed in Table 1 is based on measured unidirectional inter-node communication performance per network connection reported elsewhere.

Table 1. System Characteristics

	Earth Simulator (NEC)	ASCI Q (HP ES45)
Year of Introduction	2002	2003
Node Architecture	Vector SMP	Microprocessor SMP
System Topology	NEC single-stage Crossbar	Quadrics QsNet Fat-tree
Number of nodes	640	3072 (Total)
Processors - per node - system total	8 5120	4 12288
Processor Speed	500 MHz	1.25 GHz
Peak speed - per processor - per node - system total	8 Gflops 64 Gflops 40 Tflops	2.5 Gflops 10 Gflops 30 Tflops
Memory - per node - per processor - system total	16 GB 2 GB 10.24 TB	16 GB 4 GB 48 TB
Memory Bandwidth (peak) - L1 Cache - L2 Cache - Main (per processor)	 N/A N/A 32 GB/s	 20 GB/s 13 GB/s 2 GB/s
Inter-node MPI - Latency - Bandwidth	8.6 µsec 11.8 GB/s	5 µsec 300 MB/s

3. Representative ASCI Applications

The performance of a system is workload dependent. This is a fundamental observation, as comparing systems based on peak performance, a combination of sub-system performances (such as the IDC so-called balanced ratings), or on LINPACK alone can be at best misleading and at worst lead to incorrect conclusions about the relative achievable performance. In this analysis we compare the performance of systems using two full applications representative of the ASCI workload, namely SAGE and Sweep3D. It is estimated that the type of computations represented by SAGE and Sweep3D represent a high majority of the cycles used on ASCI machines. A brief description of both SAGE and Sweep3D is included in Sections 3.1 and 3.2 below.

In this analysis, measurement of the applications is not currently possible in all the configurations that we are considering in this work. For instance the access to the Earth Simulator is limited. Hence, the approach in this analysis is to use detailed performance models for each of the two codes. An overview of the performance models is given in Section 3.3 along with a quantitative analysis of their accuracy.

3.1. SAGE

SAGE (SAIC's Adaptive Grid Eulerian hydrocode) is a multidimensional (1D, 2D, and 3D), multi-material, Eulerian hydrodynamics code with adaptive mesh refinement (AMR). It comes from the Los Alamos National Laboratory's Crestone project, whose goal is the investigation of continuous adaptive Eulerian techniques to stockpile stewardship problems. SAGE represents a large class of production ASCI applications at Los Alamos that routinely run on 1,000s of processors for months at a time.

SAGE performs hydro-dynamic and heat radiation operations on a structured spatial mesh that is adaptively refined on a cell by cell basis as necessary at the end of each processing cycle. Each cell at the topmost level (level 0) can be considered as the root node of an oct-tree of cells in lower levels. For example, the shock-wave indicated in the 3D spatial domain in Figure 1 by the solid line may cause cells close to it to be split into smaller cells. In this example, a cell at level 0 is considered as not being refined, while a cell at level n represents a physical domain that is 8^n times smaller.

The key characteristics of SAGE are:

Data decomposition – The spatial domain is partitioned across processors in 1D *slab* sub-grids. The problem size grows proportionally with the number of processors – a weak-scaling characteristic.

Processing flow – the processing proceeds in cycles. In each cycle there are a number of stages that involve the three operations of: one (or more) data gathers to obtain a copy of sub-grid boundary data from remote processors, computation on each of the gathered cells, and one (or more) scatter operations to update data on remote processors.

AMR and load-balancing – at the end of each cycle, each cell can either be split into a block of smaller 2x2x2 cells, combined with its neighbors to form a single larger cell, or remain unchanged. A load-balancing operation takes place if any processor contains 10% more cells than the average number of cells across all processors.

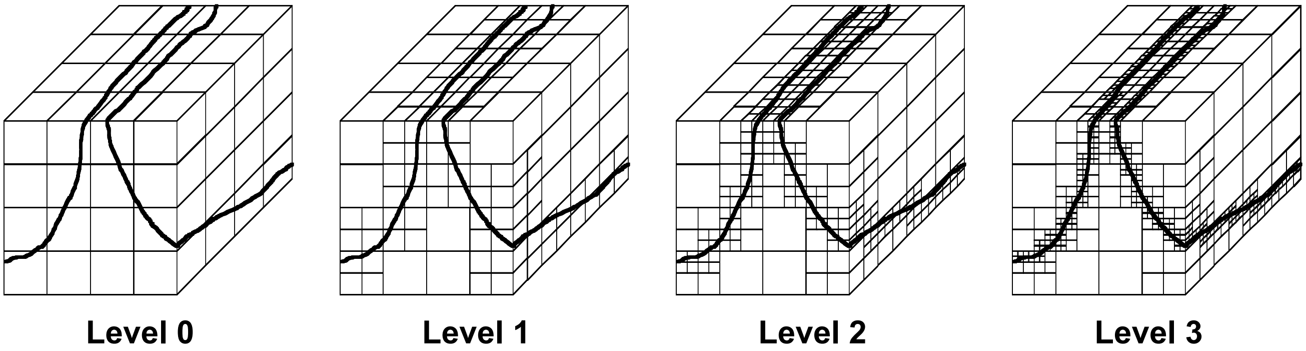


Figure 1. Example of SAGE Adaptive Mesh Refinement at multiple levels

The 1D slab decomposition leads to a number of important factors that influence the achievable performance. For instance the amount of data transfer in gather-scatter communication increases, and also the distance between processors increases as the number of processors increases. Full details on the scaling behavior of SAGE as well as its performance model have been previously described [7].

3.2. Sweep3D

Sweep3D is a time-independent, Cartesian-grid, single-group, "discrete ordinates" deterministic particle transport code. Estimates are that deterministic particle transport accounts for 50-80% of the execution time of many realistic simulations on current ASCI systems; this percentage may expand on future 100Tflops systems. The basis for neutron transport simulation is the time-independent, multigroup, inhomogeneous Boltzmann transport equation [4].

Sweep3D is characteristic of a larger class of algorithms known as wavefront algorithms. These algorithms exhibit a complex interplay between their surface-to-volume and processor utilization. As such an investigation into the performance of Sweep3D cannot be done without a detailed application model. Details of the performance characteristics and performance model of Sweepd3D has been previously published for MPP systems[4], and for SMP clusters [5].

The 3D spatial domain in Sweep3D is mapped to a logically 2D processor array. Wavefronts (or “sweeps”) scan the processor array originating from all corners in a pipelined fashion – an example of which is shown in Figure 2 were wavefronts are traveling from the upper-left corner to the lower-right. The larger the sub-grid size, the more favorable the surface-to-volume is, but results in a corresponding decrease in processor utilization. In order to achieve optimality between the two, Sweep3D uses blocking in one spatial dimension and also in angles. The tuning of the blocking parameters has an important effect on the runtime of the application. Our model captures the effect of the blocking parameters, hence allowing the selection of optimum values that minimize the runtime for a particular system and configuration.

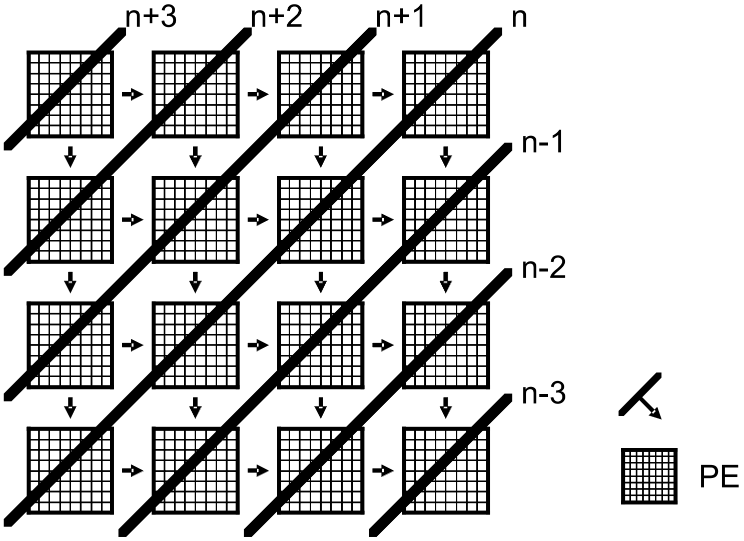


Figure 2. The pipeline processing of sweeps on a 2D processor array. n denotes a sweep that is currently processed on the major diagonal of the processor array. Other wavefronts shown (n-3…n+3) are either ahead or behind sweep n.

3.3. Performance Models

The performance models for both SAGE and Sweep3D incorporate all of the applications key processing characteristics and the way in which they map to and utilize the available resources within a system. These models are analytical, that is they are described through analytical formulations which are parameterized in terms of system characteristics and application characteristics. Within this work we do not attempt to fully describe these models. However, it is important to understand the main principles of the application execution and thus we illustrate below the main aspect of the models along with the key input parameters. The input parameters can be used to explore the performance space and to answer what-if type predictive studies. Full details on the performance model of SAGE can be found in [7] and that for Sweep3D in [4]

SAGE is an iterative code in which each cycle does the same processing but with a possibly changing spatial-grid. The cycle time of SAGE is modeled as a combination of: single-processor computation time, T_{comp}, (based on the number of cells per processor and the time to process a cell), the time spent exchanging boundary data between processors in gather/scatter operations (T_{GS_msg} with a frequency of f_{gs} per cycle, collective allreduce operations, AMR operations of combining and dividing cells ($T_{combine}$ and T_{divide} respectively), and load-balancing (T_{load}):

$$T_{cycle} = T_{comp} + f_{GS}.C.T_{GS\_msg} + f_{allreduce}.T_{allreduce} + T_{divide} + T_{combine} + T_{load} \tag{1}$$

Each term in this model has many subcomponents. Due to the 1D slab decomposition the boundary sizes and also distance between processors (defined as the number and location of the processors involved in one boundary exchange) scale in a very distinctive way (see [7]). Message sizes range from 4 words to half the number of cells in the subgrid owned by each processor. Equation 1 gives a view of the performance from an application perspective and is combined with the processing characteristics of a particular system including: single-processor time to process a single-cell, intra-node and inter-node communication performances, and also the performance of collectives.

In contrast, a single iteration of Sweep3D processes a spatial-grid which does not change, but requires many wavefronts that originate in a defined order from each of the four corners of a logical 2D processor array. For each wavefront, each cell within a block is processed. The cycle time of Sweep3D is modeled as a combination of the single-processor time to process a block of cells, and the time taken to communicate block boundary data in the direction of the wavefront.

$$T_{cycle} = \left(2.P_x + 4.P_y \quad 6\right) T_{comp} + \frac{(1 + P_{SMP})}{CL}.T_{msg} \quad + N_{sweep}. \; T_{comp} + \frac{2.(1 + P_{SMP})}{CL}.T_{msg} \tag{2}$$

Inherent in the performance of Sweep3D is a pipeline due to the wavefront processing and a repetition of processing over the number of blocks on each processor. The first part of equation 2 represents the pipeline cost and the second term the block processing. The pipeline length is in terms of the 2D processor array size (P_x, P_y), and the block processing is in terms of the number of sweeps (N_{sweep}). The block size is a component of the computation time (T_{comp}), the messaging time (T_{msg}), and also N_{sweep}. The communication time is dependent on the number of processors within an SMP (P_{SMP}) sharing the available communication channels (CL). Message sizes depend on blocking parameters but typically are in the range of 100-1000 words.

Table 2. Accuracy of the SAGE performance model.

System	Number of Configurations tested	Maximum Processors tested	Maximum error (%)	Average error (%)
ASCI Blue (SGI O2K)	13	5040	12.6	4.4
ASCI Red (Intel Tflops)	13	3072	10.5	5.4
ASCI White (IBM SP3)	19	4096	11.1	5.1
ASCI Q (HP AlphaServer ES45)	24	3716	9.8	3.4
TC2K (HP AlphaServer ES40)	10	464	11.6	4.7
T3E (Cray)	17	1450	11.9	4.1

The prediction accuracy of the performance models for SAGE and Sweep3D have been validated on numerous systems including all ASCI machines that have been installed to date. A summary of the accuracy of the SAGE performance model is given in Table 2 for six systems. Both the average and maximum prediction errors are shown. It can be seen that the model has a maximum prediction error of 12.6% across all configurations and systems listed. A configuration in this context is a specific processor count. The errors tend to increase with increasing processor count.

4. Predictive performance comparison

It is only through knowledge of the workload the systems are to be used for that a meaningful performance comparison can be made. Thus in this section we compare the performance of the Earthy Simulator and ASCI Q using the two codes representative of the ASCI workload – SAGE and Sweep3D. In order to produce the estimate of the runtime of these applications on the Earth Simulator we have used the performance models as discussed in Section 3. They take into account all architectural details that are parameters to our models as well as low-level performance data found in the Earth Simulator literature, such as the latency and bandwidth for MPI communications [16]. Four sets of analysis are included below:

1) The rate at which each ASCI Q executes both SAGE and Sweep3D are compared.
2) The performance of the Earth Simulator is compared with that of the full sized ASCI Q. This is done separately for SAGE and for Sweep3D.
3) The size of an equivalent performing system to the Earth Simulator is calculated based on scaling of the number of AlphaServer node count.
4) The performance of a combined workload consisting of an assumed 60% Sweep3D and 40% SAGE is considered. Again, an equivalently sized AlphaServer system is calculated.

In these comparisons we use both SAGE and Sweep3D in a weak-scaling mode in which the sub-grid sizes on each processor remain a constant. However, since the main memory per processor varies across machines we use sub-grid sizes per processor which are in proportion to their main memory size. This corresponds to the applications using all the available memory. Both the weak scaling scenario and the use of as much memory as possible are typical of they way in which large-scale ASCI computations are performed. The number of cells per processor for SAGE is set to be 37,500 on ASCI Q, and 17,500 on the Earth Simulator. The sub-grid sizes in Sweep3D are set to be 12x12x280 (40320 cells) on ASCI Q and 8x8x280 (17920 cells) on the Earth Simulator. These sizes correspond approximately to the main memory capacities per processor of 4GB in ASCI Q and 2GB per processor in the Earth Simulator.

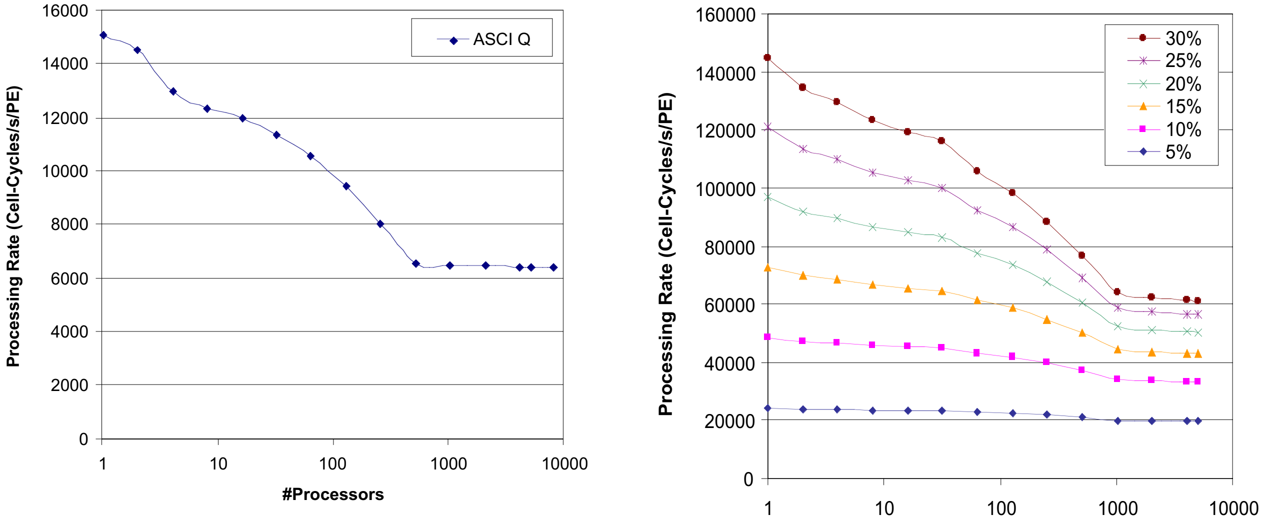
Both performance models are a function of the time to compute a sub-grid of data on a single processor. This time has been measured on the ASCI system but is unknown on the Earth Simulator. To circumvent this, in this analysis we predict the runtime of the applications on the Earth Simulator for a family of curves. Each curve assumes a speed of either: 5%, 10%, 15%, 20%, 25%, or 30% of single processor-peak. This has the advantage in that the analysis will be valid over a period of time: since codes can be optimized over time for a particular system, the percentage of single processor-peak may improve. The performance of both SAGE and Sweep3D has been measured on ASCI Q.

Further parameters which have not been measured are related to the inter-node communication performance. This would require access to the cross-bar network of the Earth Simulator for a higher accuracy. Several architectural aspects of the communication subsystems have been taken into account in order to calculate the best estimate for these parameters. Specific assumptions relate to the serialization of messages when several processors within a node perform simultaneous communications to another node on the Earth Simulator. This results in a multiplicative factor on the time taken to perform a single communication. The analysis below is sensitive to this parameter. In addition, uniform bandwidth and latency is assumed between any two nodes in both systems. This assumption holds true to a high degree for the Quadrics network.

It should also be noted that SAGE currently achieves approximately 10% of single processor-peak performance on microprocessor systems such as the AlphaServer ES45 used in ASCI Q, and Sweep3D achieves approximately 14%. Both codes may need to be modified (at worst re-coded) in order to take advantage of the vector processors in the Earth Simulator. However, none of these codes is particularly tuned for the architectural features of the RISC architectures either, particularly the on-chip parallelism and the memory hierarchy [3]. Low levels of single processor peak performance have currently been observed on the NEC SX-6 for both SAGE and Sweep3D. These are discussed below.

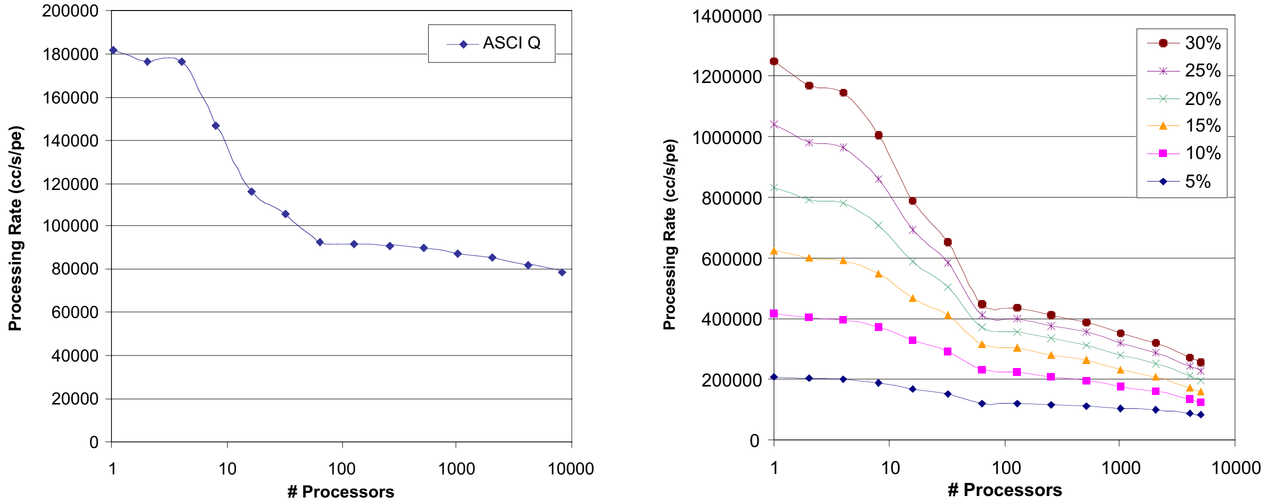
4.1. Performance of SAGE and Sweep3D

The performance of SAGE on ASCI Q is shown in Figure 3a), and for the Earth Simulator in Figure 3b) using the range of assumed percentages of sustained single processor performance. The metric used in this analysis is the cell processing rate – the number of cell-cycles per second per processor (cc/s/pe). Although the performance models are formulated in terms of cycle time, the number of cells per processor between the systems vary due to the differences in main memory capacity and thus the cycle time also varies in accordance to this. It can be seen that the value of cc/s/pe decreases with increasing processor count due to the increase in parallelism costs. A similar comparison is given in Figure 4 for Sweep3D.



(a) ASCI Q **(b) Earth Simulator**

Figure 3. A comparison of the performance of SAGE



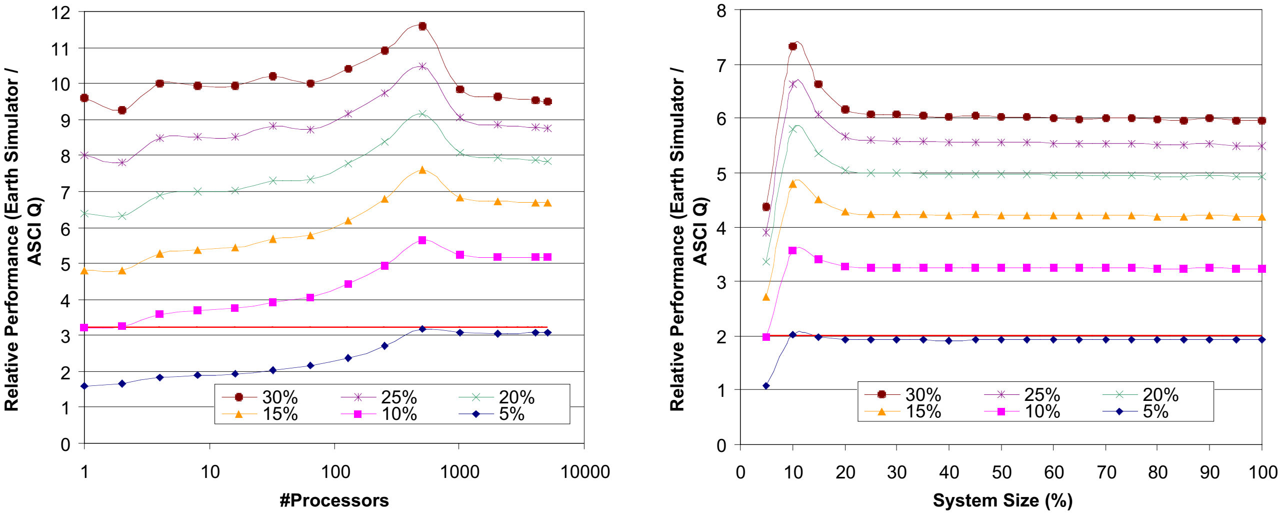
(a) ASCI Q **(b) Earth Simulator**

Figure 4. A comparison of the performance of Sweep3D

4.2 Relative performance between the Earth Simulator and ASCI Q.

SAGE

In Figure 5 the performance of the Earth Simulator is compared with that of ASCI Q for SAGE. Figure 5a) compares the performance on a like-for-like processor count whereas Figure 5b) compares the performance on the percentage of the system used. When comparing performance for a fixed number of processors (Figure 5a) one should remember that one Earth Simulator processor has an 8Gflops peak performance, and each ASCI Q Alpha processor has a 2.5Gflops peak performance. The relative advantage of the Earth Simulator based on peak performance alone is a factor of 3.2 (indicated by a single horizontal line in Figure 5a). A value greater than 3.2 in Figure 5a) indicates a performance advantage of the Earth Simulator compared to ASCI Q over and above the ratio of single processor-peak performance.



(a) equal processor count **(b) percentage of total system**

Figure 5. Relative performance between the Earth Simulator and ASCI Q (SAGE)

It can be seen that the relative performance is better on the Earth Simulator in all cases on a like-for-like processor count. Moreover, if performance of 10% or greater of single processor-peak is assumed, the performance advantage of the Earth Simulator becomes larger than just the ratio of the single-processor peak of 3.2. Depending on the percentage of single processor-peak that will be achieved, the performance advantage of the Earth simulator on the largest configuration is between a factor of 3 and a factor of 9.

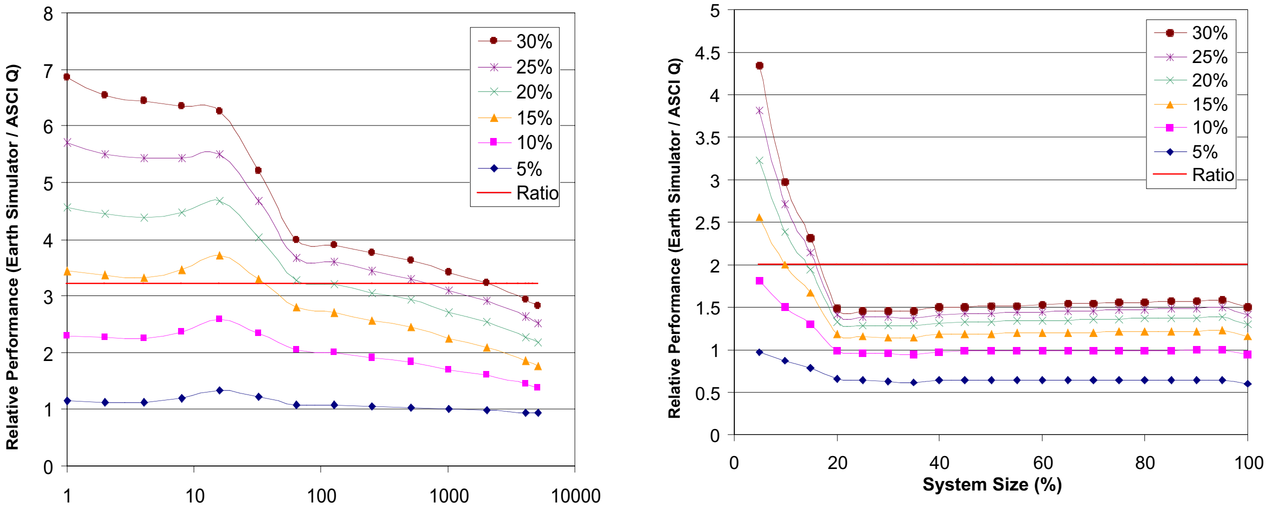
When comparing the percentage of the system recall that the Earth Simulator contains 5120 processors and ASCI Q contains 12288 processors. The relative performance of the systems based on their peak is simply 2 (40Tflops/20Tflops) again indicated by a horizontal line in Figure 5b). It can be seen that the performance advantage of the Earth Simulator when considering the fully sized machines in Figure 5b) is between a factor of 2 and 6.

The shape of the curves in Figure 5 indicates the various surface-to-volume regimes as the machine sizes increase. Clearly the y-axis intercepts for all curves indicate the difference in processing speed for each assumed percentage of single processor-peak achieved on the Earth Simulator. As the machine size increases, the communication starts having a bearing on the surface-to-volume. In particular, for very large configurations, for which the communication requirements become very large, the bandwidth plays an important role in the performance advantage of the Earth Simulator.

Sweep3D

In Figure 6 the performance of the Earth Simulator is compared with that of ASCI Q for Sweep3D. Performance is compared using the same basis as that for SAGE.

It can be seen that the relative performance on the Earth Simulator decreases with the increase in the number of processors used. Depending on the percentage of single-processor peak that will be achieved, the performance advantage of the Earth Simulator on a equal processor count basis on the largest processor count is between a factor of 1 and 3. The Earth Simulator performs worse than the relative single processor-peak speed advantage of 3.2. This is due in part to the communication requirements in this application being largely latency-bound, and also in part due to the difference in the scaling behavior of the different problem sizes as a result of the difference in memory capacities. Hence the lower ratios in these regimes compared to SAGE. Similarly when comparing the fully sized systems in Figure 6b), the Earth Simulator only performs better than ASCI Q if the achieved single processor peak performance on Sweep3D is greater than 10%.



(a) equal processor count **(b) percentage of total system**

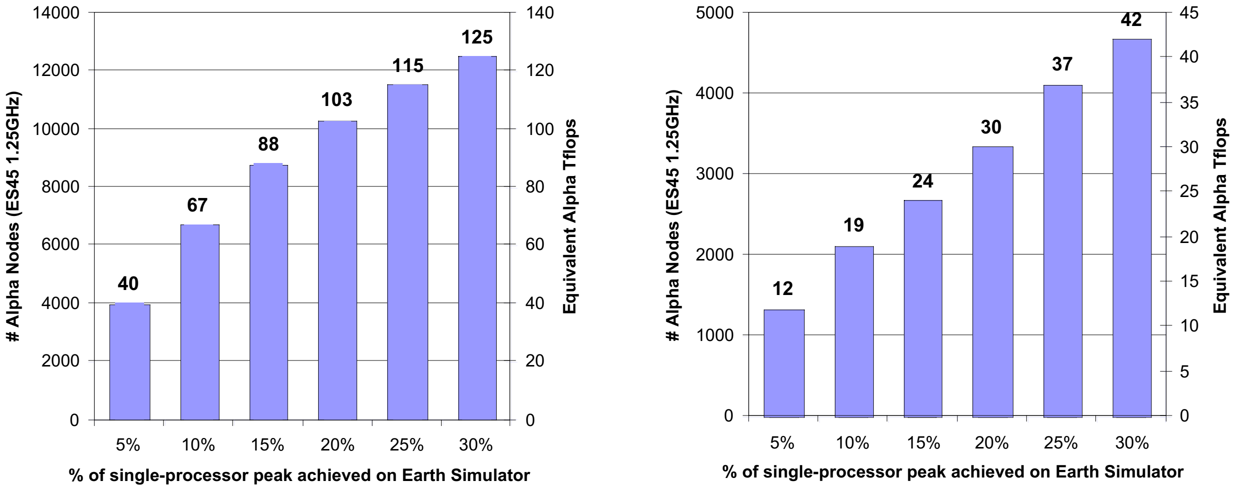
Figure 6. Relative performance between Earth Simulator and ASCI Q (Sweep3D)

4.3 Sizing equivalent performing systems to the Earth Simulator

Here, we calculate the size of system would achieve the same performance as the Earth Simulator. In this comparison we consider scaling up ASCI Q to a point at which it would each achieve the performance of the Earth Simulator for each of SAGE and Sweep3D. We again assume weak-scaling of the applications and utilization of the available memory. We also assume that all other architectural characteristics of the machine (processors per node, network characteristics, etc) remain the same.

Figure 7 shows the size of the systems required, in terms of peak Tflops, for the Alpha EV68 processors in order to achieve the same performance as the Earth Simulator on SAGE and for Sweep3D. The size of an equivalent sized machine to the Earth Simulator grows with the increase of the achieved single processor speed for each application on the Earth Simulator. This is a key parameter in this analysis which has not been directly measured on the Earth Simulator. The performance of SAGE has been measured on a single NEC SX-6 node. Recall that the SX-6 is based on an Earth Simulator node but has a reduced memory performance. The measurement showed that SAGE currently achieves only 5% of single-processor peak. It is expected that over time this value may increase with further code optimization for increased vectorization. The main processing loop inside Sweep3D currently does not vectorize and hence achieves a very low percentage of peak.

Thus, at the present time, an equivalent sized system can be approximately obtained from Figure 7b) for a single processor-peak on the Earth Simulator of between 5 and 10%, and for Sweep at the lowest end of the scale. For example, an equivalent Alpha System would have a peak Tflop rating of approximately 40Tflops for SAGE, and less than 13Tflops for Sweep3D.



(a) SAGE **(b) Sweep3D**

Figure 7. Equivalent system sizes (in peak Tflops) to the Earth Simulator

4.5 Composite Workload

By assuming a hypothetical workload consisting of 40% SAGE and 60% Sweep3D an equivalent sized system to the Earth Simulator can also be calculated. This is shown in Table 3 for an Alpha system. In this analysis we include the same range of single-processor peak percentages as before.

It can be seen in Table 3 that given the current situation of SAGE achieving 5% of single-processor peak on an NEC SX-6, and Sweep3D achieving a very low percentage, that the equivalent sized Alpha system would have a peak of 23Tflops. The full ASCI Q at 20Tflops thus achieves a comparable performance on the two representative ASCI applications at the current time. The information in Table 3 could be coupled with the cost of the two systems to determine which system provides better price-performance for a given cost.

Thus the information presented here for a family of curves, with each curve being based on a different level of achieved single processor performance, provides information that will remain valid over time while the optimization of the codes increases. We re-state here that the current implementations of these codes is not optimized for RISC architectures either, hence the comparison between the two machines using the current implementation is fair.

Table 3. Peak-Tflop rated Alpha system required to achieve the same performance of the Earth Simulator using assumed application weighting (SAGE 40%, Sweep3D 60%). (Numbers in this table represent peak performance in Tflops)

		SAGE % of single-processor peak					
		5%	10%	15%	20%	25%	30%
Sweep3D	5%	23	34	42	48	53	57
	10%	27	38	47	53	57	61
	15%	30	41	50	56	60	64
	20%	34	45	53	59	64	68
	25%	38	49	57	63	68	72
	30%	41	52	60	66	71	75

5. Conclusions

We have compared the performance of the Earth Simulator against ASCI Q for two applications representative of the ASCI workload. The applications considered are Sweep3D, representative of S_N transport computations and SAGE, representative of hydro computations.

All performance data for the applications were generated using highly accurate models for the runtime of these applications developed at Los Alamos and validated on a variety of large-scale parallel systems, including all ASCI machines. In order to bypass the limitations of not having had the opportunity to run these applications on the Earth Simulator to obtain single-processor performance, a family of curves was generated for the Earth Simulator that assume a performance of 5, 10, 15, 20, 25 and 30% of single-processor peak speed.

We have analyzed the performance of the machines as they are. No architectural changes of any kind were considered. That includes the amount of memory with which these systems are equipped, the Earth Simulator \ having 2GB/processor and ASCI Q having 4GB/processor. In this way we have tried to avoid the dangers of analyzing a moving target, choosing instead to compare snapshots of the machines in their current configuration.

Our analysis shows that for all assumed serial performance of the 2 applications on the Earth Simulator, there is a performance advantage of this machine over the ASCI Q on an equivalent processor count. The advantage is more pronounced for SAGE, in which computation and bandwidth are the main contributors to the performance of the code. For SAGE the performance on the Earth Simulator is between a factor of 3-9 larger than on the Alpha system. On Sweep3D, while the Earth Simulator maintains the performance advantage, the relative gain is only between a factor of 1-3.

By considering a number of values of the achieved single-processor performance on the Earth Simulator we think that we also covered the grounds given the distinct possibility that no single value will be generated by various benchmarkers. A multitude of variants of these codes may exist, some of which may vectorize better than others.

An intuitive, but unsubstantiated by a thorough analysis, principle was submitted to the scientific community a decade ago. The argument went as follows: specialized processors (including vector processors) are faster but expensive as they are not backed up by the marketplace, commodity production. Hence, since price-performance is the most important consideration, let’s compensate the performance disadvantage of the COTS microprocessors by building large-scale machines with a greater number of them.

In this paper we quantify this thesis, by showing how large an Alpha-based system would be required in order to achieve an equivalent performance as the Earth Simulator for the representative ASCI workload under consideration. Given the current performance of SAGE and Sweep3D as measured on a single NEC SX-6 node, we estimate that an Alpha system with a peak performance of 23Tflops would equal the performance of the Earth Simulator. Thus, the fully installed ASCI Q with a peak performance of 20Tflops will have a slightly lower performance to that of the Earth Simulator on this workload.

Only through modeling such analysis is possible given the complexity and the non-linearity of the performance of an application and the multi-dimensional performance space that needs to be considered.

Acknowledgements

The authors would like to thank Scott Pakin, Fabrizio Petrini and Kei Davis of Los Alamos and also David Addison of Quadrics for their comments on this work. This work was funded by ASCI Institutes. It was also supported in part by a grant of HPCMP resources from the Arctic Region Supercomputing Center. Los Alamos National Laboratory is operated by the University of California for the National Nuclear Security Administration of the US Department of Energy.

References

[1] ASCI Q, http://www.lanl.gov/asci

[2] ASCI White, http://www.llnl.gov/asci/platforms/white

[3] S. Goedecker, A. Hoisie, Performance Optimization of Numerically Intensive Codes, SIAM Press, March 2001.

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

[5] Adolfy Hoisie, Olaf Lubeck, Harvey Wasserman, Fabrizio Petrini, and Hank Alme, "A General Predictive Performance Model for Wavefront Algorithms on Clusters of SMPs”, in Proc. of ICPP Toronto, 2000.

[6] D.J. Kerbyson, A. Hoisie, H.J. Wasserman, “A Comparison Between the Earth Simulator and AlphaServer Systems using Predictive Application Performance Models”, in Proc. Int. Parallel and Distributed Processing Symposium (IPDPS), Nice, France, April 2003.

[7] D.J. Kerbyson, H.J. Alme, A. Hoisie, F. Petrini, H.J. Wasserman, M. Gittings, “Predictive Performance and Scalability Modeling of a Large-Scale Application”, in Proc. SuperComputing, Denver, November 2001.

[8] D.J. Kerbyson, A. Hoisie, H.J. Wasserman, “Use of Predictive Performance Modeling During Large-Scale System Installation”, to appear in Parallel Processing Letters, World Scientific Publishing, 2003.

[9] D.J. Kerbyson, S.D. Pautz, A. Hoisie, “Predictive Modeling of Parallel Sn Sweeps on Unstructured Meshes”, Los Alamos National Laboratory report LA-UR-02-2662, May 2002.

[10] Kitawaki, M. Yokokawa, “Earth Simulator Running”, Int. Supercomputing Conference, Heidelberg, June 2002.

[11] M. Mathis, D.J. Kerbyson, “Performance Modeling of MCNP on Large-Scale Systems”, in. Proc. Int. Conference on Computational Sciences (ICCS), LNCS, Springer-Verlag, June 2003.

[12] F. Petrini, W.C. Feng, A. Hoisie, S. Coll, E. Frachtenberg, “The Quadrics Network: High-Performance Clustering Technology”, *IEEE Micro*, 22(1), 2002, pp. 46-57.

[13] T. Sato, “Can the Earth Simulator Change the Way Humans Think?”, Keynote address, Int. Conf. Supercomputing, New York, June 2002.

[14] S. Shingu, Y. Tsuda, W. Ohfuchi, K. Otsuka, H. Takahara, T. Hagiwara, S. Habata “A 26.58 Tflops Global Atmospheric Simulation with the Spectral Transform Method on the Earth Simulator”, in Proc. SuperComputing, Baltimore, November 2002.

[15] The NEC SX-6, NEC product description, NEC Corporation, http://www.sw.nec.co.jp/hpc/sx-e

[16] H. Uehara, M. Tamura, M. Yokokawa, “An MPI Benchmark Program Library and Its Application to the Earth Simulator, ISHPC 2002, LNCS Vol. 2327, Springer-Verlag, 2002, pp 219-230.

[17] T. Watanabe, “A New Era in HPC: Single Chip Vector Processing and Beyond”, Proc. NEC Users Group meeting, XIV, May 2002.

[18] P. Worley, “Preliminary SX-6 Evaluation’, http://www.csm.ornl.gov/evaluation/sx6

[19] M. Yokokawa, “Present Status of Development of the Earth Simulator’, in IWIA’01, IEEE Computer Press, 2001, pp. 93-99.

[20] M. Yokokawa, K. Itakura, A. Uno, T. Ishihara, Y. Kaneda, “16.4-Tflops Direct Numerical Simulation of Turbulence by a Fourier Spectral Method on the Earth Simulator”, Proc. SC2002, Baltimore, 2002.