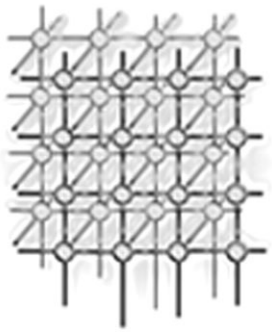
CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2005; **17**:1219–1238
Published online 22 March 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cpe.891

A performance comparison between the Earth Simulator and other terascale systems on a characteristic ASCI workload‡



Darren J. Kerbyson*,†, Adolfy Hoisie and Harvey J. Wasserman

Performance and Architectures Laboratory (PAL), Modeling, Algorithms and Informatics Group, CCS-3, Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.

SUMMARY

This work gives a detailed analysis of the relative performance of the recently installed Earth Simulator and the next top four systems in the Top500 list using predictive performance models. The Earth Simulator uses vector processing nodes interconnected using a single-stage, cross-bar network, whereas the next top four systems are built using commodity based superscalar microprocessors and interconnection networks. The performance that can be achieved results from an interplay of system characteristics, application requirements and scalability behavior. Detailed performance models are used here to predict the performance of two codes representative of the ASCI workload, namely SAGE and Sweep3D. The performance models encapsulate fully the behavior of these codes and have been previously validated on many large-scale systems. One result of this analysis is to size systems, built from the same nodes and networks as those in the top five, that will have the same performance as the Earth Simulator. In particular, the largest ASCI machine, ASCI Q, is expected to achieve a similar performance to the Earth Simulator on the representative workload. Published in 2005 by John Wiley & Sons, Ltd.

KEY WORDS: performance modeling; performance analysis; performance prediction; large-scale systems; application analysis; extreme-scale parallel systems

1. INTRODUCTION

In this work we compare the performance of the Earth Simulator [1,2] and the current top four U.S. Department of Energy (DOE) systems, ranked as numbers 2–5 in the list of the top 500 machines as of June 2003 [3]. The Earth Simulator was installed at Yokohama City, Japan in March 2002. It had the design goal of giving a 1000-fold increase in the processing capability for atmospheric research, compared with that available when envisioned in 1996, and was initially expected to

*Correspondence to: Darren J. Kerbyson, Performance and Architectures Laboratory (PAL), Modeling, Algorithms and Informatics Group, CCS-3, Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.
†E-mail: djk@lanl.gov

Published in 2005 by John Wiley & Sons, Ltd.

Received 14 February 2003
Revised 16 September 2003
Accepted 18 September 2003

achieve a sustained performance of 5 Tflops (12.5% of system-peak) on atmospheric applications [4]. It has subsequently been demonstrated that 60% of system-peak performance can be achieved on an atmospheric simulation code [5], with other codes achieving up to 40% of peak [6]. Several of the DOE systems are part of the Accelerated Strategic Computing Initiative (ASCI).

At present there is much interest in comparing the relative performance between the Earth Simulator and other large-scale teraflop 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 top five machines fall into different categories in this spectrum. ASCI Q (currently listed as number 2 in the Top500) compute nodes are manufactured by a sole supplier (HP), and use the Quadrics network, again a sole supplier. Similarly, ASCI White's (number 4) components and NERSC's system components (number 5) are solely built by IBM, but in a higher volume than those in the Earth Simulator. MCR (number 3) uses high-volume, mass-produced compute nodes which can be obtained from a multitude of suppliers.

There is a large difference in the processor-peak performance between the top five systems. An Earth Simulator node contains eight processors each with a peak processing rate of 8 Gflops. In contrast, an AlphaServer ES45 as used in ASCI Q contains four processors each with a peak of 2.5 Gflops. The communication networks are also quite different—the Earth Simulator has a cross-bar interconnect with a 16 GB s^{-1} bandwidth between any two nodes in comparison to the Quadrics network which has a fat-tree topology interconnecting ES45 nodes with a 300 MB s^{-1} bandwidth. These differences affect the achievable performance of applications, changing both their surface-to-volume ratios as well as their parallel overheads [7].

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 achieved MPI intra-node communication bandwidth and latencies. Neither 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, the inter-processor communication system, and the interaction of these items. 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 [8], structured and unstructured mesh transport codes [9,10], and a Monte Carlo particle simulation code [11].

We use two existing application performance models in this paper, models for an S_N transport application on Cartesian grids [9] and for an adaptive mesh hydro-code [8], in order to compare the relative performance between the Earth Simulator and the other top five machines from the Top500 list. 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 ASCI Q [12], to investigate improvements in performance due to sub-system improvements [13], and in the initial stages of the procurement of ASCI Purple—expected to be a 100 Tflop 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.

The performance models are used to compare the performance of the Earth Simulator with the other top terascale systems and also to 'size' systems architecturally similar to those in the top five that will provide the same performance as the Earth Simulator.

In Section 2 an overview of the current top five systems is given along with their salient performance characteristics. In Section 3 we describe the characteristics of two applications. In Section 4, we compare the performance of the top five systems. This work builds upon an initial study of a performance comparison between the Earth Simulator and AlphaServer systems [14].

2. COMPARISON OF SYSTEMS

In order to compare the performance of the Earth Simulator with other machines, we first compare the main characteristics of the machines in the top five of the Top500 list as published in June 2003 [3]. The Earth Simulator is ranked number 1, having a system-peak performance of 40 Tflops, followed by four DOE machines—ASCI Q at Los Alamos National Laboratory (LANL), MCR at Lawrence Livermore National Laboratory (LLNL), ASCI White also at Livermore, and the IBM machine at the National Energy Research Scientific Computing Center (NERSC). It is interesting to note the LINPACK performance values as used to rank the machines in the Top500. These are listed in Table I, along with the system-peak performances for each.

Several things can be noted from Table I. First, the system-peak performance of the Earth Simulator is almost the summation of the peak performance of ASCI Q, MCR and ASCI White. The LINPACK performance of the Earth Simulator is almost identical to the summation of all the other top five machines. Note also that the LINPACK performance on the NERSC machine was obtained using a newer LINPACK implementation than on ASCI White [3], resulting in an identical performance even though the two systems differ in size (ASCI White contains 8192 processors and the NERC machine contains 6656 processors). The new LINPACK implementation has yet to be run on ASCI White and may result in a higher performance. In order to provide a more meaningful performance comparison we first detail the individual systems and some of their characteristics below. This is followed by examination of a representative ASCI workload described in Section 3.

Table I. The top five systems, June 2003.

Top500 ranking	System	System-peak performance (Gflops)	LINPACK performance (Gflops)
1	Earth Simulator	40 960	35 860
2	ASCI Q	20 480	13 880
3	MCR	11 060	7634
4	ASCI White	12 288	7304
5	NERSC	9984	7304

2.1. The Earth Simulator

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

2.2. ASCI Q

ASCI Q is the largest ASCI system that has been installed at Los Alamos National Laboratory (LANL) [18]. It contains 2048 HP AlphaServer ES45 nodes and has a system-peak performance of 20 Tflops. Each node is a four-way SMP containing four 21264D EV68 Alpha microprocessors operating at 1.25 GHz with 64 KB L1 instruction and data cache, 16 MB unified L2 cache, and 16 GB of main memory. Each processor has a peak performance of 2.5 Gflops. A peak memory bandwidth up to 8 GB s^{-1} is possible within a node using two 256-bit memory buses running at 125 MHz. Four independent standard 64-bit PCI buses (running at 66 MHz) 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 300 MB s^{-1}, with a typical latency of 5 μs. The latency increases slightly with the physical distance between nodes. A detailed description of the Quadrics network can be found in [19].

2.3. MCR

The Multiprogrammatic Capability Cluster (MCR) was recently installed at LLNL and entered service in late 2002. It is implemented using more commodity-like processing nodes. MCR consists of

1152 dual Xeon Intel Pentium IV processing nodes. Each processor operates at 2.4 GHz, with an 8 KB L1 cache, 512 KB L2 cache, and has a peak performance of 4.8 Gflops. Each node contains 4 GB of main memory and has a shared memory bus with a peak bandwidth of 3.2 GB s^{-1}. The nodes are interconnected using the Quadrics QsNet fat-tree network—the same as that used in ASCI Q, but with only a single rail. The MPI level inter-node bandwidth has a peak of 315 MB s^{-1} and a latency of 4.75 μs—this is slightly better performance than on ASCI Q due to a different PCI bus implementation in the nodes.

2.4. ASCI White

ASCI White was installed at Lawrence Livermore National Laboratory (LLNL) in 2000 [20]. It consists of 512 16-way IBM RS/6000 NightHawk-II nodes. The system thus contains a total of 8192 processors and its peak performance is 12 Tflops. Each node contains four processor cards each with four Power3-II processors and 16 GB main memory. Each processor has 32 KB + 64 KB L1 instruction and data cache and an 8 MB unified L2 cache. The processors operate at 375 MHz and have a peak performance of 1.5 Gflops. Nodes are interconnected using a single IBM proprietary SP omega network. Each node is connected to this network via two network cards. This network has a peak inter-node bandwidth of 2 GB s^{-1} (bi-directional) and an MPI latency of 18 μs. The system is typically configured with 460 compute nodes, 16 file-server nodes, 32 visualization nodes, and four interactive I/O nodes.

2.5. NERSC

The NERSC machine is architecturally the same as ASCI White. It contains 16-way IBM RS/6000 NightHawk-II nodes interconnected using a single IBM proprietary omega network. It is however smaller than ASCI White, consisting of 416 nodes for a total of 6656 processors. The amount of memory per node varies between 16 and 64 GB—312 nodes contain 16 GB of memory. This system is typically configured with 380 compute nodes, 20 file-server nodes, six interactive nodes, two network nodes, and eight service nodes.

A summary of these systems is listed in Table II. Peak performance characteristics as well as configuration details are included. The two IBM systems, ASCI White and the NERSC machine, are considered together in Table II as they differ in their node/processor counts and year of introduction. It is clear that the peak performance of the Earth Simulator exceeds that of any of the other systems, and also that the main memory bandwidth per processor is far above that of any other processor. The inter-node communication performance in terms of latency is worse on the Earth Simulator when compared with ASCI Q and MCR (both using the Quadrics network) but is better by almost a factor of 40 in terms of bandwidth. Note that the MPI performance listed in Table II is based on measured unidirectional inter-node communication performance per network connection, and the memory performance is based on measured memory read bandwidths reported elsewhere.

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

Table II. System characteristics.

	Earth Simulator (NEC)	ASCI Q (HP ES45)	MCR (Duel Xeon)	ASCI White and NERSC (IBM SP3)
Year of introduction	2002	2003	2002	2000 (2002)
Node architecture	Vector SMP	Microprocessor SMP	Microprocessor SMP	Microprocessor SMP
System topology	NEC single-stage Crossbar	Quadrics QsNet Fat-tree	Quadrics QsNet Fat-tree	IBM Omega network
Number of nodes	640	2048	1152	512 (416)
Processors				
per node	8	4	2	16
system total	5120	8192	2304	8192 (6656)
Processor speed	500 MHz	1.25 GHz	2.4 GHz	375 MHz
Peak speed				
per processor (Gflops)	8	2.5	4.8	1.5
per node (Gflops)	64	10	9.6	24
system total (Tflops)	40	20	10.8	12 (9.8)
Memory				
per node (GB)	16	16	4	16
per processor (GB)	2	4	2	1
system total (TB)	10.24	48	4.6	8 (7)
Memory bandwidth (GB s^{-1})				
L1 cache	N/A	12	18	6
L2 cache	N/A	8	16	2
Main (per processor)	32	2	2	1
Inter-node MPI				
latency (μs)	8.6	5	4.75	18
bandwidth	11.8 GB s^{-1}	300 MB s^{-1}	315 MB s^{-1}	500 MB s^{-1}

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. These two applications are implemented using MPI and their performances are considered for the codes as they are currently implemented.

In this analysis, measurement of the applications is not currently possible on all systems 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.

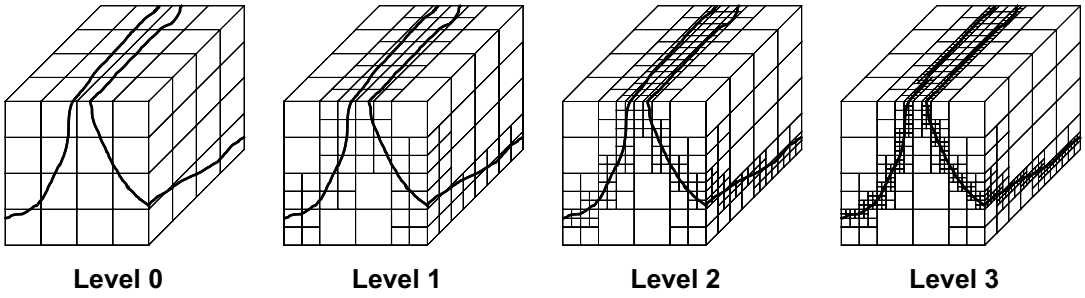


Figure 1. Example of SAGE AMR at multiple levels.

3.1. SAGE

SAGE (SAIC's Adaptive Grid Eulerian hydrocode) is a multi-dimensional (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 thousands 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 not to be refined, while a cell at level n represents a physical domain that is 8^n times smaller.

The key characteristics of SAGE are listed below.

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

Processing flow. The processing proceeds in cycles. In each cycle there are a number of stages that involve three operations: one (or more) data gather(s) to obtain a copy of sub-grid boundary data from remote processors, computation on each of the gathered cells, and one (or more) scatter operation(s) 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 $2 \times 2 \times 2$ 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 cells across all processors.

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

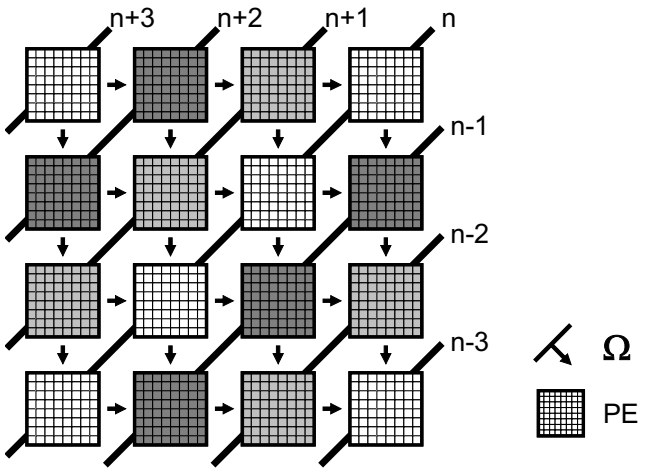


Figure 2. The pipeline processing of sweeps on a 2D processor array. n denotes a sweep that is currently processed on the major diagonal of the processor array. Other wavefronts shown $(n-3 \ldots n+3)$ are either ahead or behind sweep n.

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 [8].

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 100 Tflops systems. The basis for neutron transport simulation is the time-independent, multigroup, inhomogeneous Boltzmann transport equation [9].

Sweep3D is characteristic of a larger class of algorithms known as wavefront algorithms. These algorithms exhibit a complex interplay between their surface-to-volume ratio 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 Sweep3D have previously been published for MPP systems [9], and for SMP clusters [21].

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, where 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 ratio is, but this 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.

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 aspects 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 [8] and that for Sweep3D in [9].

SAGE is an iterative code in which the same processing is done in each cycle, but with a possibly changing spatial grid due to refinement. 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 ($T_{\text{GS\_msg}}$ 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}):

$$T_{\text{cycle}} = T_{\text{comp}} + f_{\text{GS}} \cdot C \cdot T_{\text{GS\_msg}} + f_{\text{allreduce}} \cdot T_{\text{allreduce}} + T_{\text{divide}} + T_{\text{combine}} + T_{\text{load}} \quad (1)$$

Each term in this model has many subcomponents. Due to the 1D slab decomposition, the boundary sizes and also the distance between processors (defined as the number and location of the processors involved in one boundary exchange) scale in a very distinctive way (see [8]). Message sizes range from four words to half the number of cells in the sub-grid 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.

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

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

Table III. 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

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.

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 III 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

Performance is entirely workload dependent, hence performance analysis in abstract of a workload is meaningless. Thus in this section we compare the performance of the current top five systems by 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. The models take into account the characteristics of the applications as they are currently implemented, architectural details 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 analyses are included below.

1. The rates at which each of the top five systems process both SAGE and Sweep3D are compared. Included here is a breakdown of where time is spent in the applications.
2. The performance of the Earth Simulator is compared with that of 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 the scaling of the node counts of each of ASCI Q, ASCI White, the NERSC machine, and MCR.
4. The performance of a combined workload consisting of an assumed 60% Sweep3D and 40% SAGE is considered. In this analysis 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 the 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), 17 500 (on the Earth Simulator and MCR), and 8750 (IBM SP3 machines). The sub-grid sizes in Sweep3D are set to be $12 \times 12 \times 280$ (40 320 on ASCI Q), $8 \times 8 \times 280$ (17 920 on the Earth Simulator and MCR), and $6 \times 6 \times 280$ (10 080 on the IBM SP3 machines). These sizes correspond approximately to the main memory capacities per processor of 4 GB (ASCI Q), 2 GB (Earth Simulator and MCR), and 1 GB (IBM SP3 machines). Note that both ASCI White and the NERSC machine are considered together as they both consist of IBM SP3 differing only in size (number of nodes in the system).

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 systems 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 an achieved application performance 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 and the percentage of single-processor peak may improve. The performance of both SAGE and Sweep3D has been measured on ASCI White and ASCI Q. The performance predictions for MCR are based on measurements made on a small dual Xeon system and on detailed knowledge of the Quadrics network.

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 due to contention. 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 microprocessor architectures with deep memory hierarchies, particularly the on-chip parallelism and the memory hierarchy [7]. 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 the top five systems on SAGE and Sweep3D

A comparison of the performance of the DOE machines is shown in Figure 3(a), and for the Earth Simulator in Figure 3(b) for SAGE 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

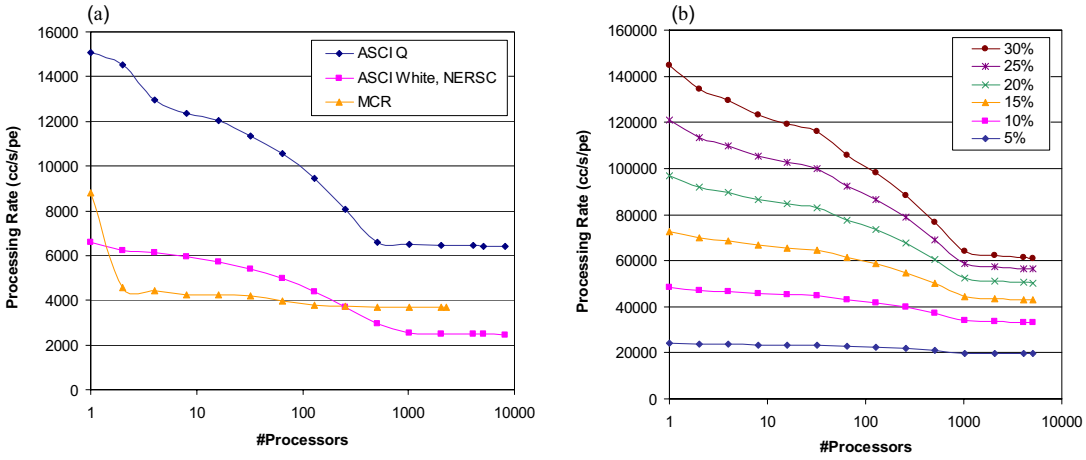


Figure 3. A comparison of the performance of SAGE: (a) DOE machines; (b) Earth Simulator.

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 varies due to the differences in main memory capacity and thus the cycle time also varies in accordance with this. It can be seen that the value of cc/s/pe decreases with increasing processor count due to the increase in parallelism costs. ASCI Q outperforms ASCI White, the NERSC machine, and MCR. Note that MCR has a good single-processor performance but exhibits contention when both processors within a node are utilized. The performance on MCR does however scale better than that on ASCI White and the NERSC machine. A similar comparison is given in Figure 4 for Sweep3D.

The performance breakdown is shown in Figure 5 for both ASCI Q and the Earth Simulator for SAGE. For each system, the percentage of time spent either processing (compute/memory), or in communication (latency or bandwidth) is shown. Figure 5(b) assumes a single-processor peak of 10% for SAGE on the Earth Simulator. It can be seen that on larger processor counts, a greater percentage of time is taken by the bandwidth component on the ASCI Q, compared with the Earth Simulator on which the bandwidth related component of the runtime is much lower. This difference is simply due to the Earth Simulator's much larger available bandwidth. SAGE has an increasing communication requirement with increasing processor count, resulting from the 1D slab decomposition as discussed in Section 3. This decomposition increases sub-grid surfaces as the processor count increases, hence increasing message sizes between processors (see [8] for a full description).

The importance of latency can be seen in the component times for Sweep3D in the comparison between Figure 6(a) and 6(b). Figure 6(b) assumes a single-processor peak of 5% for Sweep3D on the Earth Simulator. The Earth Simulator has a much larger latency component than the Alpha system. The wavefront algorithms require blocking of the sub-grids on each processor to achieve high multi-processor utilization, which results in small messages between processors. Thus communications are dominated by the communication latency. The difference between the pipeline processing and block

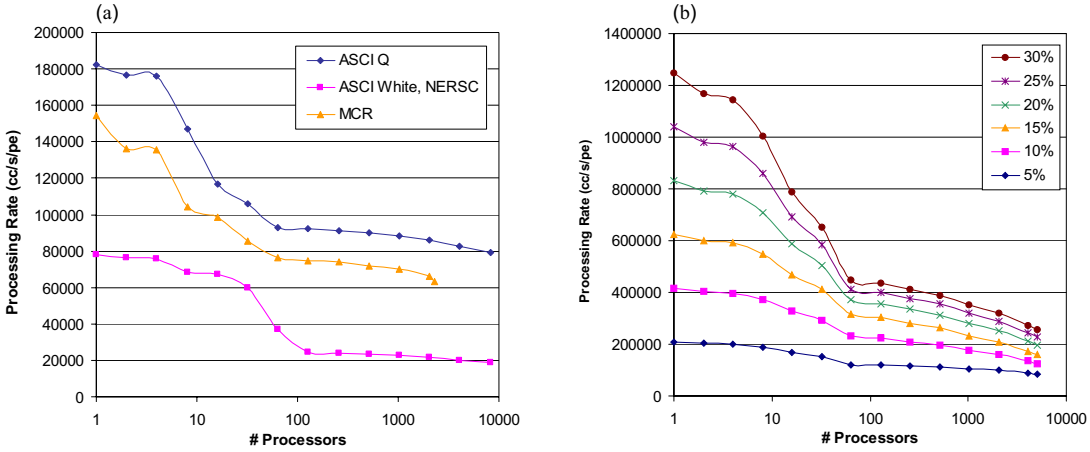


Figure 4. A comparison of the performance of Sweep3D: (a) DOE machines; (b) Earth Simulator.

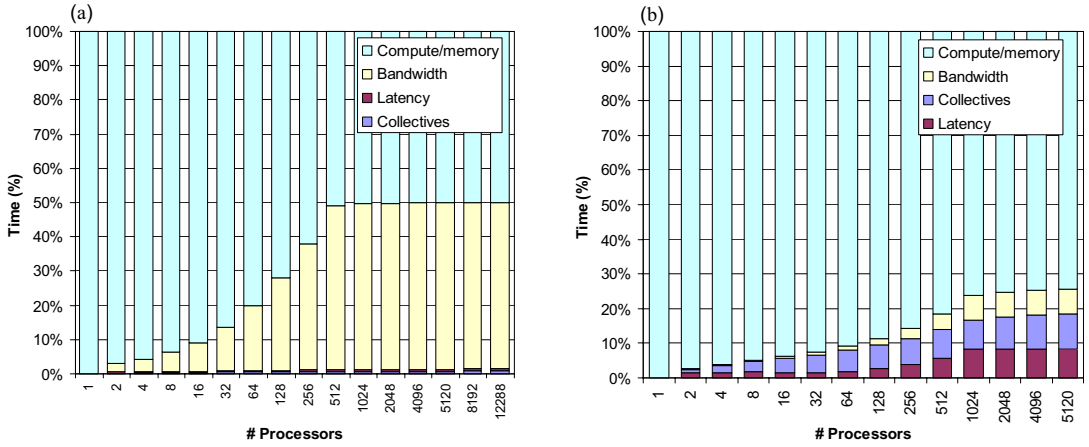


Figure 5. Time spent in processing and communication in SAGE: (a) ASCI Q; (b) Earth Simulator.

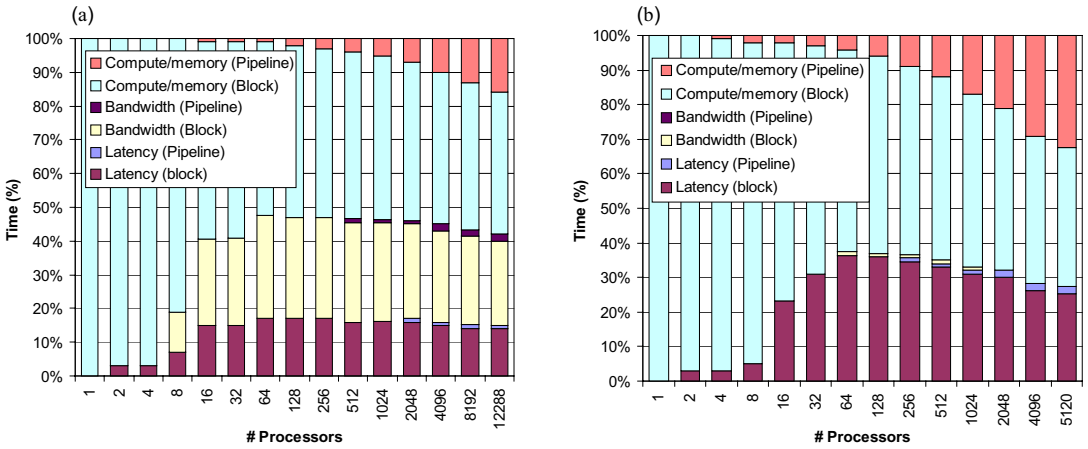


Figure 6. Time spent in processing and communication in Sweep3D: (a) ASCI Q; (b) Earth Simulator.

processing from Equation (2) can also be seen in Figure 6. The pipeline processing increases with the increase in processor count.

4.2. Relative performance of the Earth Simulator and ASCI Q

The relative performance of the Earth Simulator and ASCI Q is considered for both SAGE and Sweep3D below. Using the performance data for each system, as presented in Figures 3 and 4, the relative performance is calculated in two ways: first considering the processor count to be equal on both systems; second considering a percentage of the total system.

4.2.1. SAGE

In Figure 7 the performance of the Earth Simulator is compared with that of ASCI Q for SAGE. Figure 7(a) compares the performance on a like-for-like processor count, whereas Figure 7(b) compares the performance based on the percentage of the system used. When comparing performance for a fixed number of processors (Figure 7(a)) one should remember that one Earth Simulator processor has an 8 Gflops peak performance, and each ASCI Q Alpha processor has a 2.5 Gflops 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 7(a)). A value greater than 3.2 in Figure 7(a) indicates a performance advantage of the Earth Simulator compared with ASCI Q over and above the ratio of single-processor peak performance.

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 a performance of 10% or greater of the single-processor peak is assumed for SAGE on the Earth Simulator, the performance advantage is larger than just the ratio of

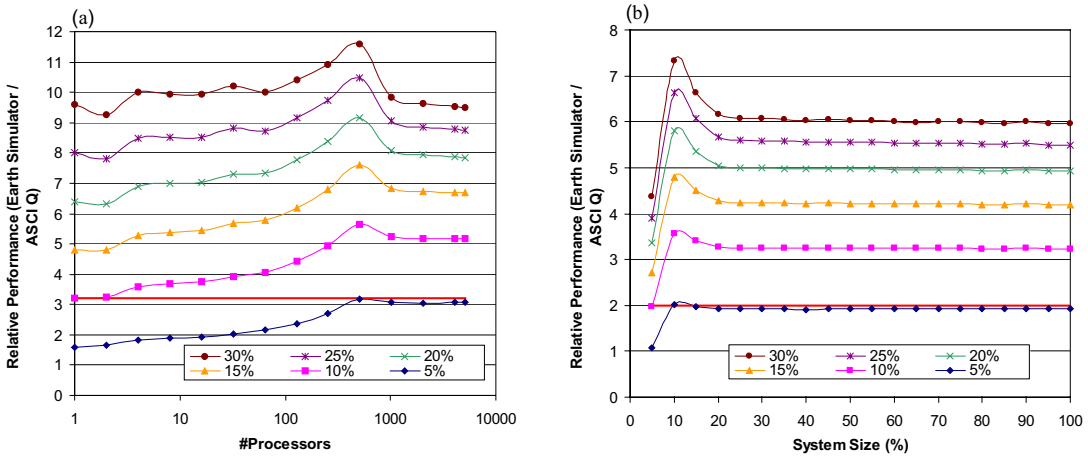


Figure 7. Relative performance between the Earth Simulator and ASCI Q (SAGE): (a) equal processor count; (b) percentage of total system.

the single-processor peak of 3.2. Depending on the percentage of the single-processor peak that will be achieved, the performance advantage of the Earth Simulator over ASCI Q on the largest configuration is between a factor of 3 and 9.

When comparing the performance in terms of the percentage of the system used, recall that the Earth Simulator contains 5120 processors and ASCI Q contains 8192 processors. The relative performance of the systems based on their peak is 2 (40 Tflops/20 Tflops), again indicated by a horizontal line in Figure 7(b). It can be seen that the performance advantage of the Earth Simulator when considering the fully sized machines in Figure 7(b) is between a factor of 2 and 6.

The shape of the curves in Figure 7 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 regime. 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.

4.2.2. Sweep3D

In Figure 8 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 the single-processor peak that will be achieved, the performance advantage of the Earth Simulator on an equal processor count basis on the largest processor count is between a factor of 1 and 3. The Earth Simulator performs worse than

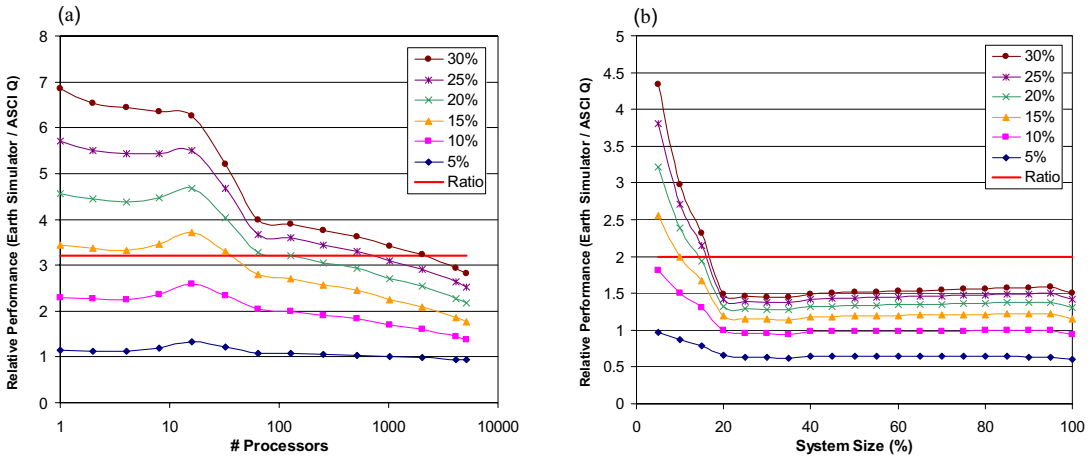


Figure 8. Relative performance between Earth Simulator and ASCI Q (Sweep3D): (a) equal processor count; (b) percentage of total system.

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 various problem sizes as a result of the difference in memory capacities. Hence the lower ratios in these regimes compared with SAGE. Similarly, when comparing the fully sized systems in Figure 8(b), the Earth Simulator only performs better than ASCI Q if the achieved single-processor peak performance on Sweep3D is greater than 20%.

4.3. Sizing equivalent performing systems to the Earth Simulator

Here, we calculate the size of the system that would achieve the same performance as the Earth Simulator on the workload considered. In this comparison we consider scaling up ASCI Q, ASCI White (or the NERSC machine), and MCR to a point at which they 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 9 shows the size of the systems required, in terms of peak Tflops, for each of the Alpha EV68, Power3, and Dual Xeon 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 in 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 the single-processor peak. It is expected that over time this

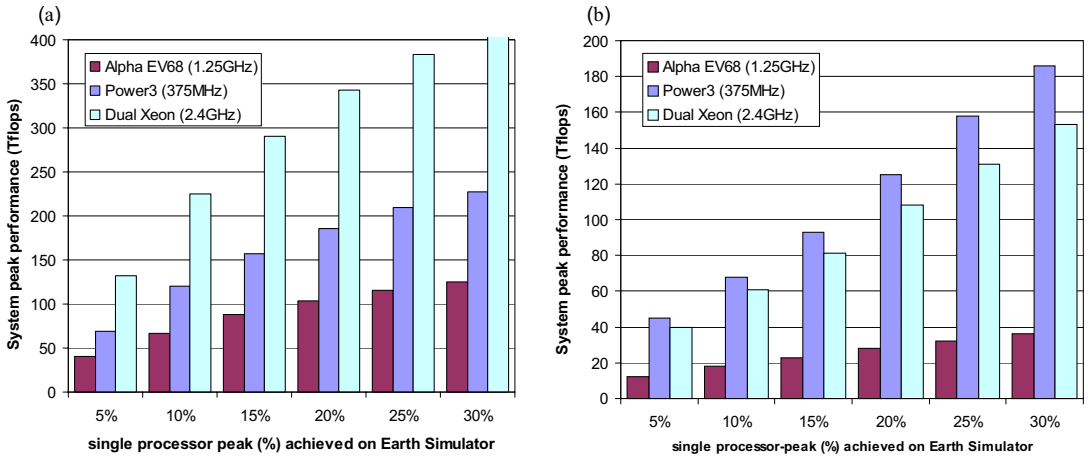


Figure 9. Equivalent system sizes (in peak Tflops) to the Earth Simulator: (a) SAGE; (b) Sweep3D.

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 9(a) for a single value of the processor peak on the Earth Simulator of between 5 and 10% for SAGE, and at the lowest end of the scale for Sweep3D. For example, an equivalent Alpha System would have a peak Tflop rating of approximately 40 Tflops for SAGE, and less than 13 Tflops for Sweep3D.

4.4. 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 IV for an Alpha EV68 system. In this analysis we include the same range of single-processor peak percentages as before.

It can be seen from Table IV that, given the current situation of SAGE achieving 5% of the single-processor peak on an NEC SX-6 and Sweep3D achieving a very low percentage, the equivalent sized Alpha system would have a peak of 23 Tflops, slightly higher than the current ASCI Q. The information in Table IV could be coupled with the cost of the two systems to determine which system provides the better price performance for a given cost.

The information presented here for a family of curves, with each curve being based on a different level of achieved single-processor performance, 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 microprocessor architectures, hence the comparison between the two machines using the current implementation is fair.

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

SWEEP3D % of single-processor peak	SAGE % of single-processor peak					
	5%	10%	15%	20%	25%	30%
5%	23	34	42	48	53	57
10%	27	38	46	52	57	61
15%	30	41	49	55	60	64
20%	33	44	52	58	62	67
25%	35	46	54	60	65	69
30%	38	48	57	63	68	72

5. CONCLUSIONS

We have compared the performance of the Earth Simulator with the current top five systems for two applications representative of the ASCI workload. The applications considered were Sweep3D, representative of S_N transport computations, and SAGE, representative of hydro computations.

All performance data 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, which includes the amount of memory with which these systems were equipped: the Earth Simulator and MCR having 2 GB per processor, ASCI Q having 4 GB per processor, and both ASCI White and the NERSC machine having 1 GB per 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 configurations.

Our analysis shows that for the assumed serial performance of the two applications on the Earth Simulator, there is a performance advantage of this machine over 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 and 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 and 3.

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

An intuitive, but unsubstantiated 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. Hence, since price performance is the most important consideration, let us compensate the performance disadvantage of the COTS microprocessors by building large-scale machines on a larger scale.

In this paper we quantify this by showing how large the Alpha-based system would have to be 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 EV68 system with a peak performance of 23 Tflops would equal the performance of the Earth Simulator. Thus ASCI Q with a peak performance of 20 Tflops is expected to achieve a similar performance to the Earth Simulator on this workload.

Only through modeling is such an analysis 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 supported by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36.

REFERENCES

1. Kitawaki S, Yokokawa M. Earth simulator running. *Proceedings of the International Supercomputing Conference (ISC)*, Heidelberg, Germany, 19–22 June, 2002.
2. Sato T. Can the Earth Simulator change the way humans think? Keynote address. *International Conference on Supercomputing*, New York, June 2002. ACM Press: New York, 2002.
3. Top500 list of machines, June 2003. http://www.top500.org.
4. Yokokawa M. Present status of development of the Earth Simulator. *Innovative Architecture for Future Generation High-Performance Processors and Systems*. IEEE Computer Society Press: Los Alamitos, CA, 2001; 93–99.
5. Shingu S, Tsuda Y, Ohfuchi W, Otsuka K, Takahara H, Hagiwara T, Habata S. A 26.58 Tflops global atmospheric simulation with the spectral transform method on the Earth Simulator. *Proceedings of IEEE/ACM Supercomputing'02*, Baltimore, MD, November 2002. IEEE/ACM Press: New York, 2002.
6. Yokokawa M, Itakura K, Uno A, Ishihara T, Kaneda Y. 16.4-Tflops direct numerical simulation of turbulence by a Fourier spectral method on the Earth Simulator. *Proceedings of IEEE/ACM Supercomputing'02*, Baltimore, MD, 2002. IEEE/ACM Press: New York, 2002.
7. Goedecker S, Hoisie A. *Performance Optimization of Numerically Intensive Codes*. SIAM Press: Philadelphia, PA, 2001.
8. Kerbyson DJ, Alme HJ, Hoisie A, Petrini F, Wasserman HJ, Gittings M. Predictive performance and scalability modeling of a large-scale application. *Proceedings of IEEE/ACM Supercomputing'01*, Denver, CO, 10–16 November 2001. IEEE/ACM Press: New York, 2001.
9. Hoisie A, Lubeck O, Wasserman HJ. Performance and scalability analysis of Teraflop-scale parallel architectures using multidimensional wavefront applications. *International Journal of High Performance Computing Applications* 2000; **14**(4):330–346.
10. Kerbyson DJ, Pautz SD, Hoisie A. Performance modeling of deterministic transport computations. *Performance Analysis and Grid Computing*. Kluwer: Dordrecht, 2003.
11. Mathis M, Kerbyson DJ, Hoisie A. Performance modeling of MCNP on large-scale systems. *Proceedings of the International Conference on Computational Science (ICCS)* (*Lecture Notes in Computer Science*, vol. 2659). Springer: Berlin, 2003; 905–915.

12. Petrini F, Kerbyson DJ, Pakin S. The case of the missing supercomputer performance: Achieving optimal performance on the 8,192 processors of ASCI Q. *Proceedings of IEEE/ACM Supercomputing'03*, Phoenix, AZ, 15–21 November 2003. IEEE/ACM Press: New York, 2003.
13. Kerbyson DJ, Wasserman HJ, Hoisie A. Exploring advanced architectures using performance prediction. *Innovative Architecture for Future Generation High-Performance Processors and Systems*. IEEE Computer Society: Los Alamitos, CA, 2002; 27–37.
14. Kerbyson DJ, Hoisie A, Wasserman HJ. A comparison between the Earth Simulator and AlphaServer systems using predictive application performance models. *Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS)*, Nice, France, 21–25 April 2003. IEEE Press: Piscataway, NJ, 2003.
15. Watanabe T. A new era in HPC: Single chip vector processing and beyond. *Proceedings of the NEC Users Group Meeting, XIV*, Yokohama, Japan, May 2002.
16. Uehara H, Tamura M, Yokokawa M. An MPI benchmark program library and its application to the Earth Simulator. *Proceedings of ISHPC 2002* (*Lecture Notes in Computer Science*, vol. 2327). Springer: Berlin, 2002; 219–230.
17. The NEC SX-6, NEC product description, NEC Corporation. http://www.sw.nec.co.jp/hpc/sx-e [June 2002].
18. ASCI Q. http://www.lanl.gov/asci [June 2002].
19. Petrini F, Feng WC, Hoisie A, Coll S, Frachtenberg E. The Quadrics Network: High-performance clustering technology. *IEEE Micro* 2002; **22**(1):46–57.
20. ASCI White. http://www.llnl.gov/asci/platforms/white [June 2002].
21. Hoisie A, Lubeck O, Wasserman HJ, Petrini F, Alme H. A general predictive performance model for wavefront algorithms on clusters of SMPs. *Proceedings of the International Conference on Parallel Processing (ICPP)*, Toronto, 2000. IEEE Press: Piscataway, NJ, 2000.