COMPUTER & COMPUTATIONAL SCIENCES

Performance Modeling the Earth Simulator and ASCI Q

Darren J. Kerbyson

Adolfy Hoisie, Harvey J. Wasserman

Performance and Architectures Laboratory (PAL)
Los Alamos National Laboratory

April 2003

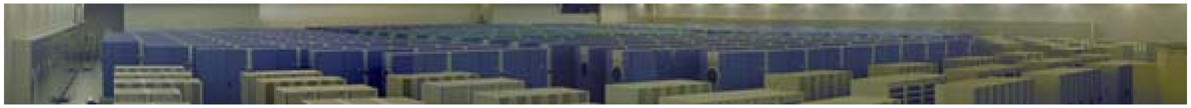
1943 - 2003 Los Alamos NATIONAL LABORATORY

Ideas That Change the World

"26.58Tflops on AFES … 64.9% of peak (640nodes)"

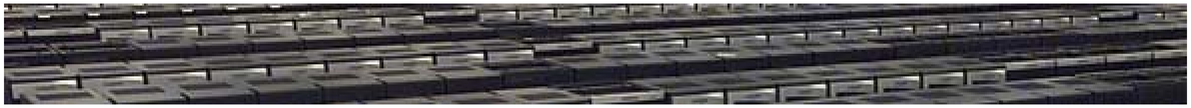
"14.9Tflops on Impact-3D …. 45% of peak (512nodes)"

"10.5Tflops on PFES … 44% of peak (376nodes)"



40Tflops

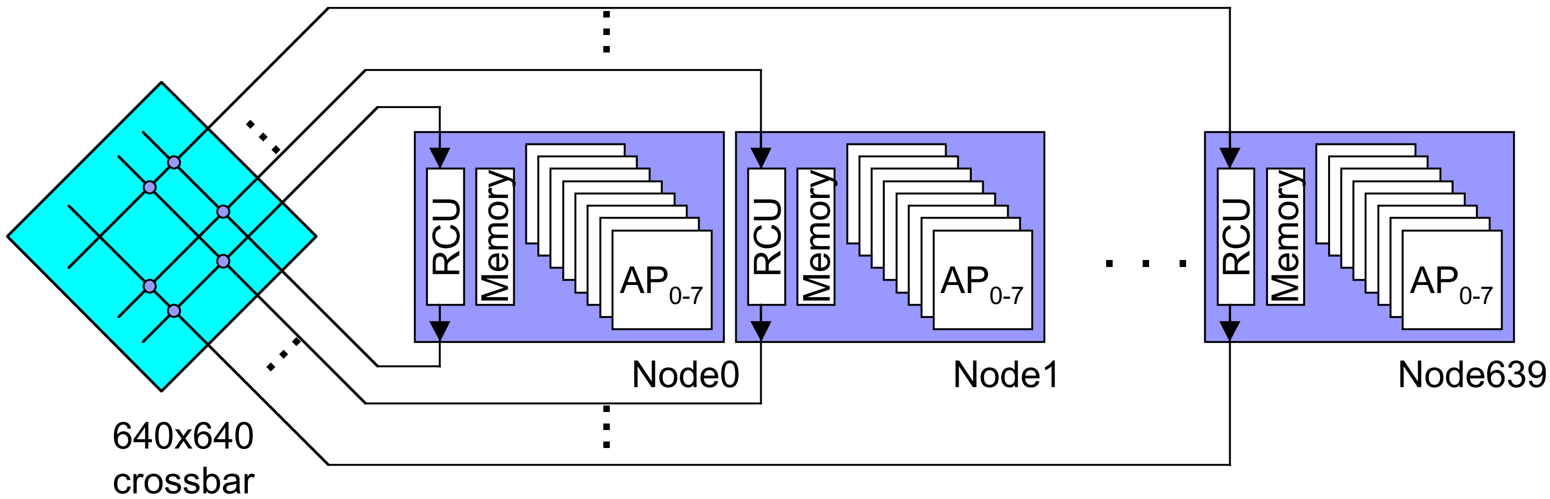
20Tflops



Talk Overview

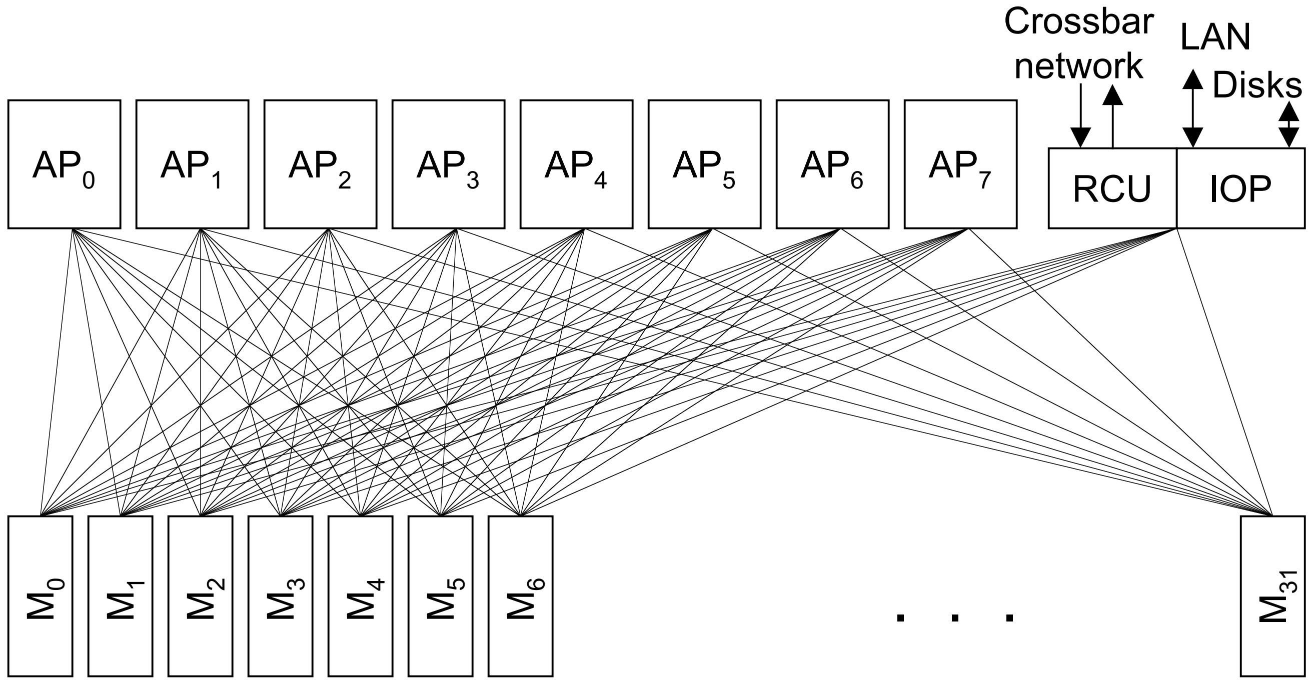
- **Overview of the Earth Simulator**
  - A (quick) view of the architecture of the Earth Simulator (and Q)
  - A look at its performance characteristics
- **Application Centric Performance Models**
  - Method of comparing performance is to use trusted models of applications that we are interested in, e.g. SAGE and Sweep3D.
  - Analytical / Parameterized in system & application characteristics
- **Models can be used to provide:**
  - Predicted performance prior to availability (hardware or software)
  - Insight into performance
  - Performance Comparison (which is better?)
- **System Performance Comparison (Earth Simulator vs ASCI Q)**

Earth Simulator: Overview



- **640 Nodes (Vector Processors)**
- **interconnected by a single stage cross-bar**
  - Copper interconnect (~3,000Km wire)
- **NEC markets a product – SX-6**
  - Not the same as an Earth Simulator Node - similar but different memory sub-system

Earth Simulator Node



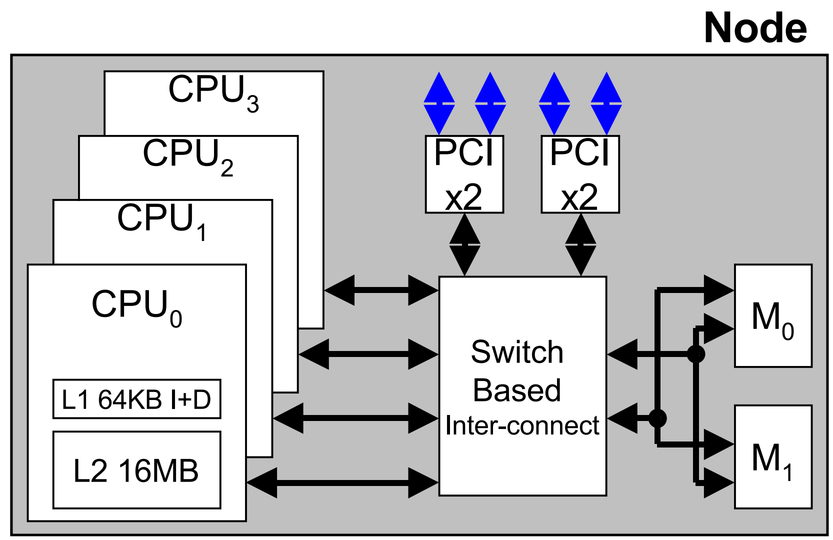
Node contains:

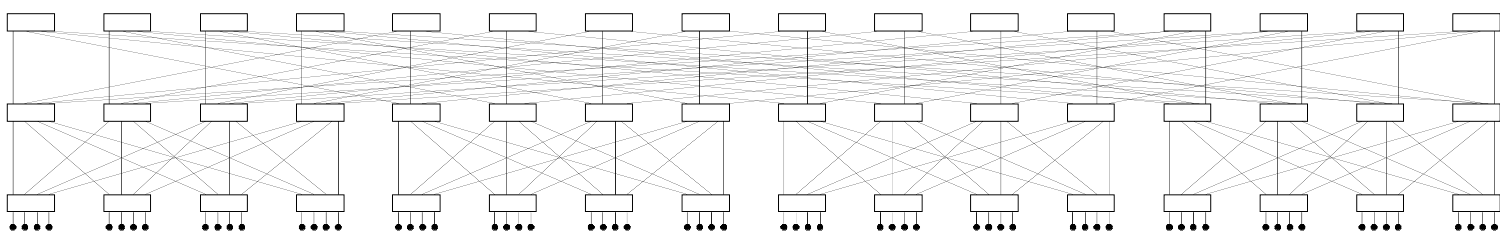
8 vector processors (AP)
16GByte memory
Remote Control Unit (RCU)
I/O Processor (IOP)

- **Each AP has:**
  - A vector unit (8 sets of 6 different vector pipelines), and a super-scalar unit
  - 5,000 pins
  - AP operates mostly at 500MHz:
  - Processor Peak Performance = 500 x 8 (pipes) x 2 (float-point) = 8,000 Mflop/s
  - Memory bandwidth = 32 GB/s per Processor (256GB/s per node)

ASCI Q

- **2048 Nodes**
- **Fat-tree network (Quadrics)**
- **Each node is an HP ES45**
  - 4 x Alpha EV68 micro-processors
  - 1.25GHz
  - Memory Hierarchy:
    - » 64KB I+D L1
    - » 16MB L2 cache
    - » 16 GB memory per node (typical)





	Earth Simulator (similar to NEC SX-6)	ASCI Q (HP ES45)
Node Architecture	Vector SMP	Microprocessor SMP
System Topology	Crossbar (single-stage)	Fat-tree
Number of nodes	640	2048
Processors - per node - system total	8 5120	4 8192
Processor Speed	500 MHz	1.25 GHz
Peak speed - per processor - per node	8 Gflops 64 Gflops	2.5 Gflops 10 Gflops
Memory - per node - per processor	16 GB 2 GB	16 GB (max 32 GB) 4 GB (max 8 GB)
Memory Bandwidth (peak) - L1 Cache - L2 Cache - Main memory (per PE)	 N/A N/A 32 GB/s	 20 GB/s 13 GB/s 2 GB/s
Inter-node MPI communication - Latency - Bandwidth	 8.6 sec 11.8 GB/s	 5 sec 300 MB/s

A quotation from a scientific explorer?

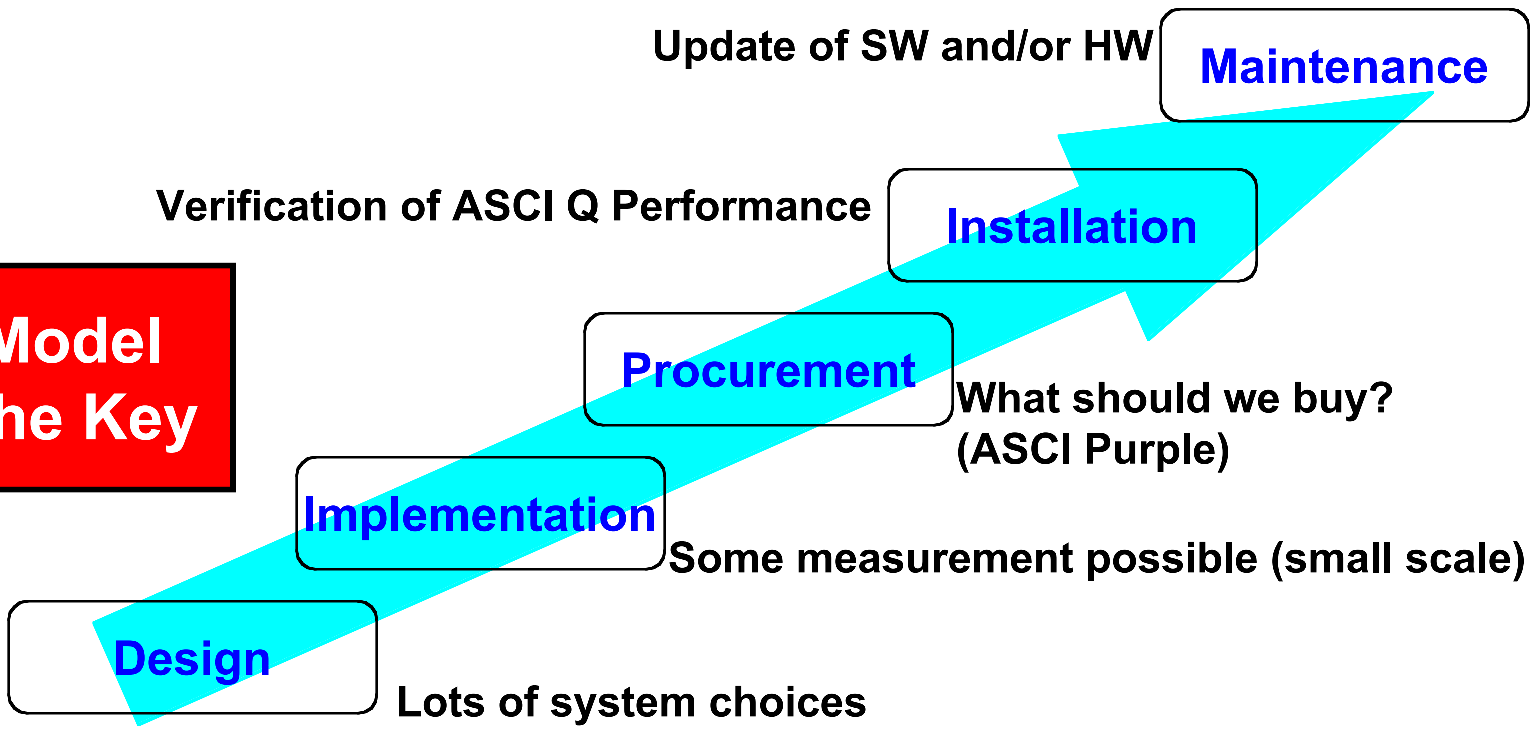
"Its life Jim,
but not as we know it"

"Its performance Jim,
but not as we expected it"

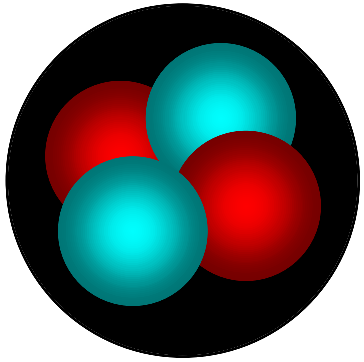
Need to have an expectation

- **Complex machines and software**
  - single processors, interactions within nodes, interaction between nodes (communication networks), I/O
- **Large cost for development, deployment and maintenance**
- **Need to know in advance what performance will be.**

A Model is the Key



How many



—

?

(How many Alphas are equivalent to the Earth Simulator ?)

Answer:

Earth Simulator = 640 x 8 x 8 = 40-Tflops

Nodes PEs/node Gflops/PE

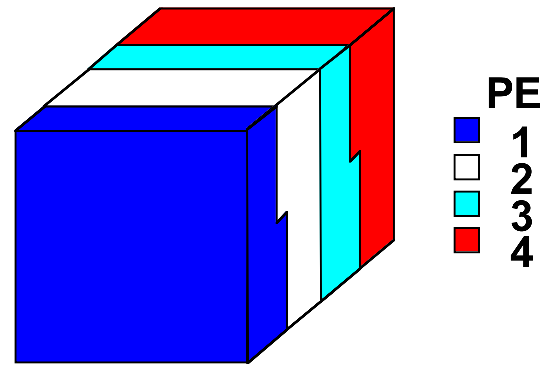
Alpha ES45 System = ? x 4 x 2.5 = 40-Tflops

-> ? = 4,096 nodes

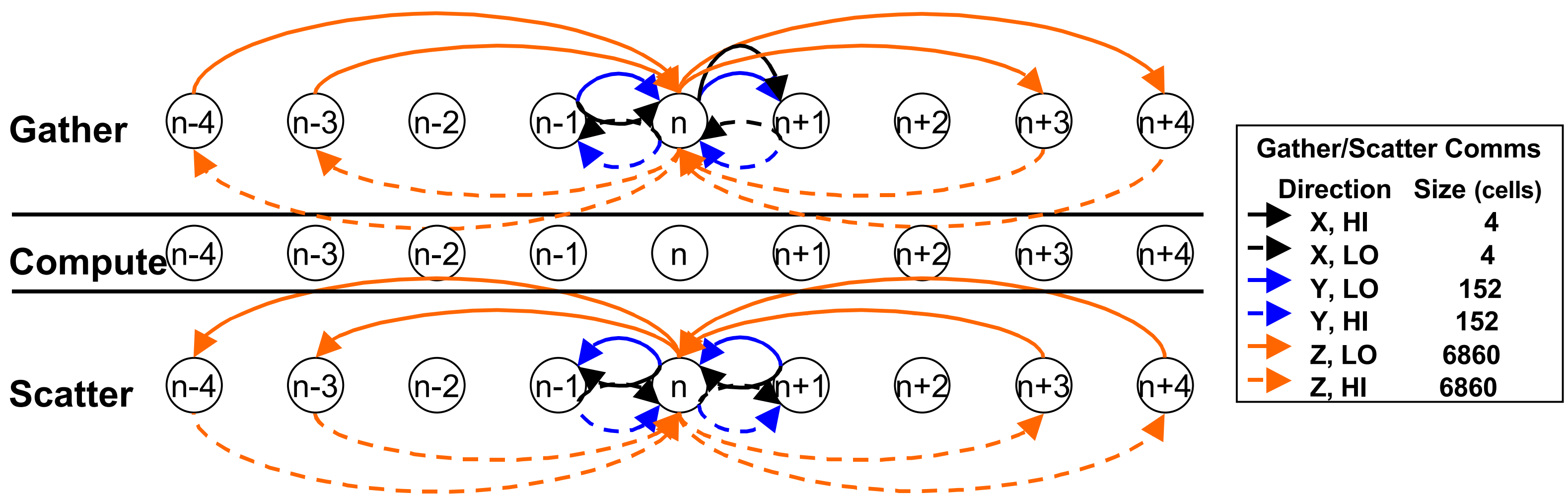
Another Answer: <u>It depends on what is going to run</u>

Performance Model for SAGE

- **Identify and understand the key characteristics of code**
- **Main Data structure decomposition: Slab**
  - communications scale as number of PEs (^2/3)
  - communication distance (PEs) increases
  - Effect of network topology



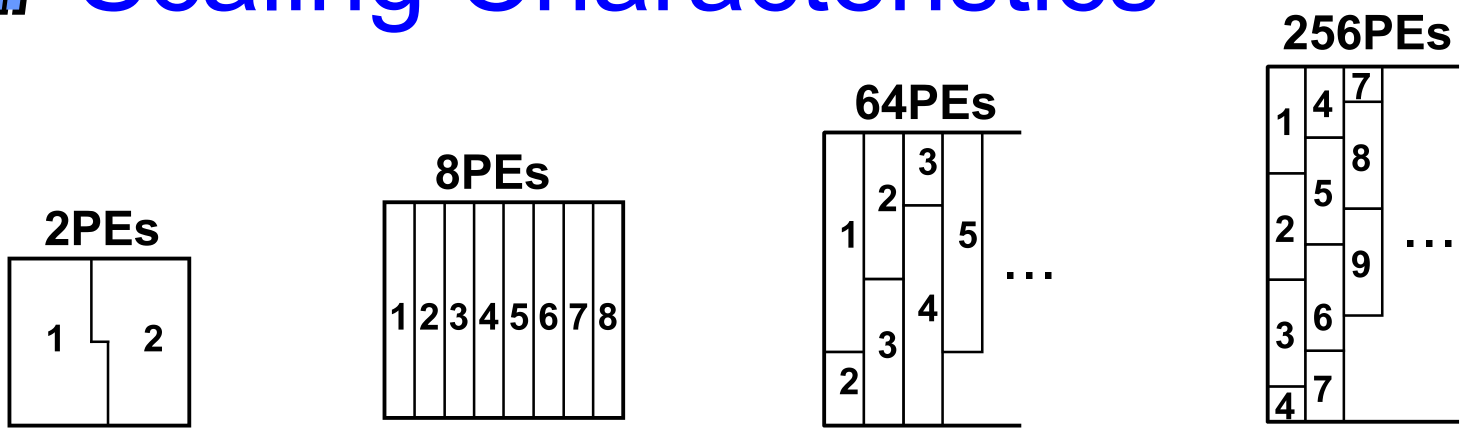
- **Processing Stages**
  - Gather data, Computation, Scatter Data

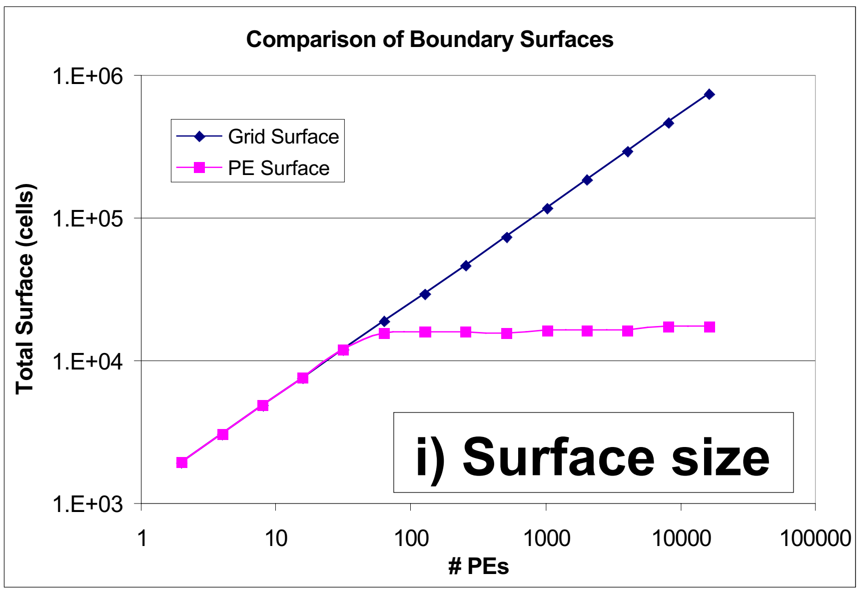


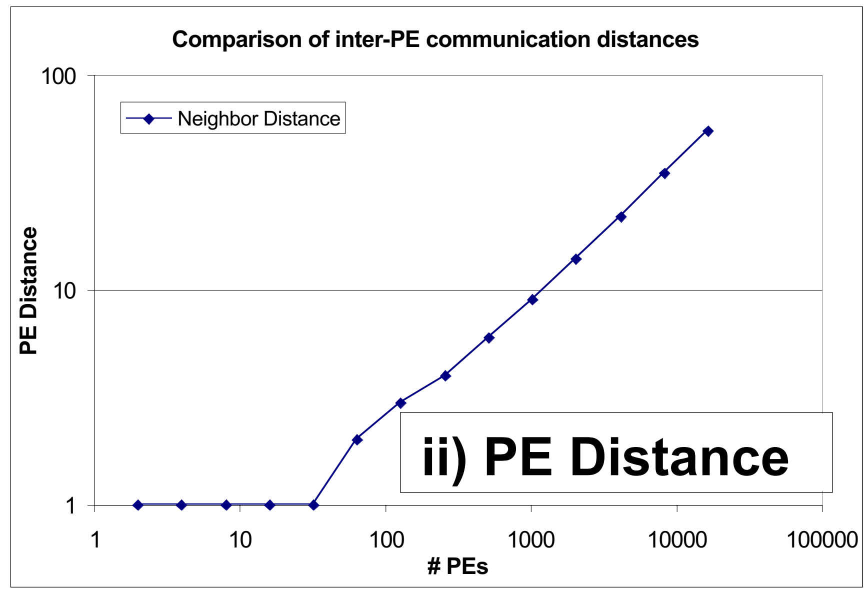
PAL

Scaling Characteristics

CCS-3







Surface split across PEs: $P > \div(E/8)$

PE distance = $(8P^2/E)^{1/3}$

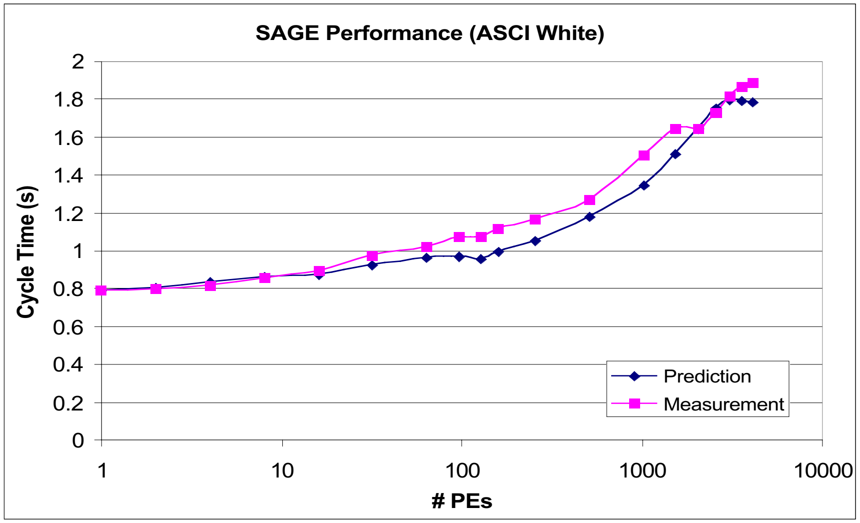
(for cells/PE = 13,500 $P > 41$**)**

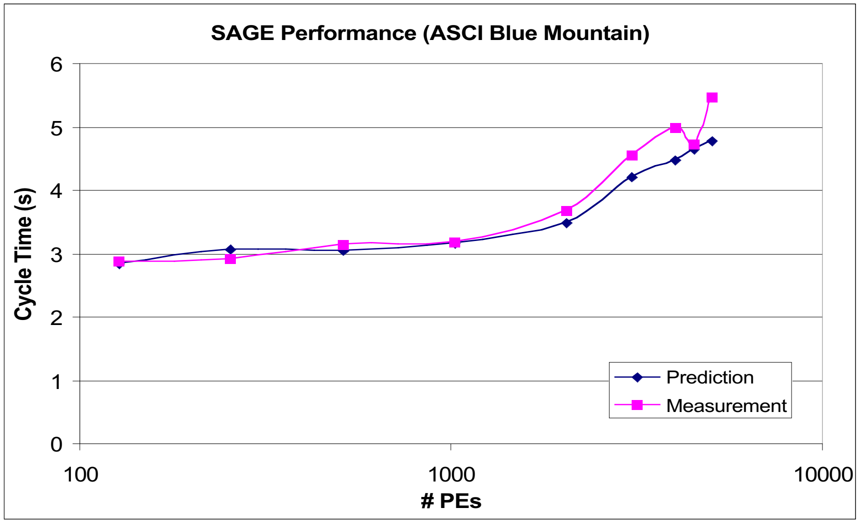
Los Alamos

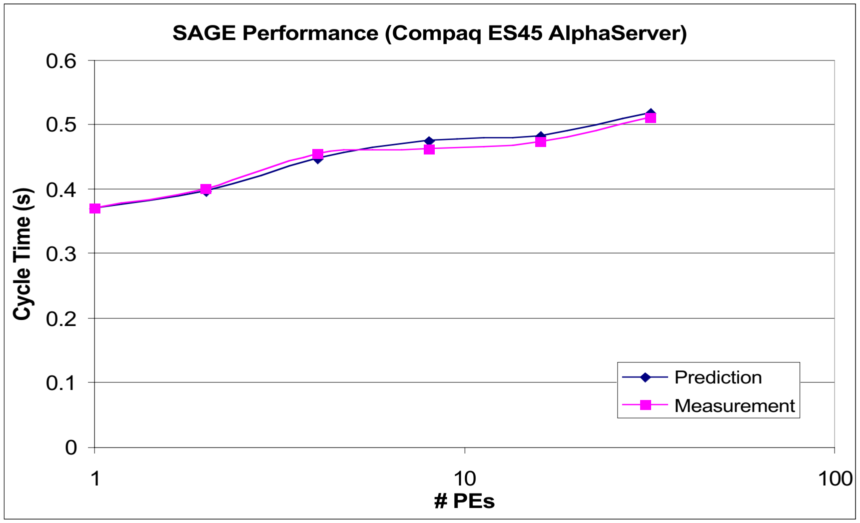
Validation

CCS-3

- **Validated on large-scale platforms:**
  - ASCI Blue Mountain (SGI Origin 2000)
  - CRAY T3E
  - ASCI Red (intel)
  - ASCI White (IBM SP3)
  - Compaq Alphaserver SMP clusters
- **Model is highly accurate (< 10% error)**





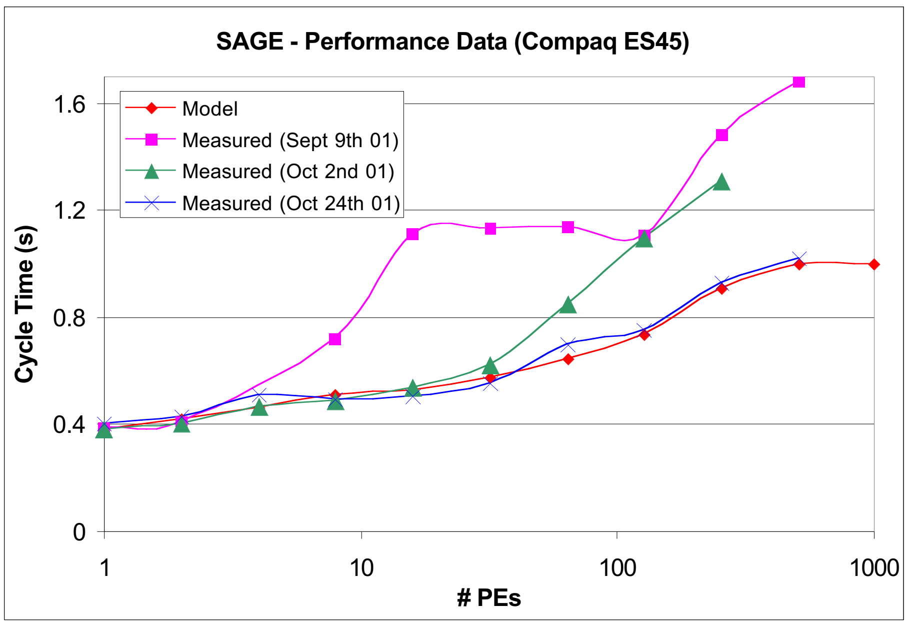


Experience: Compaq Installation

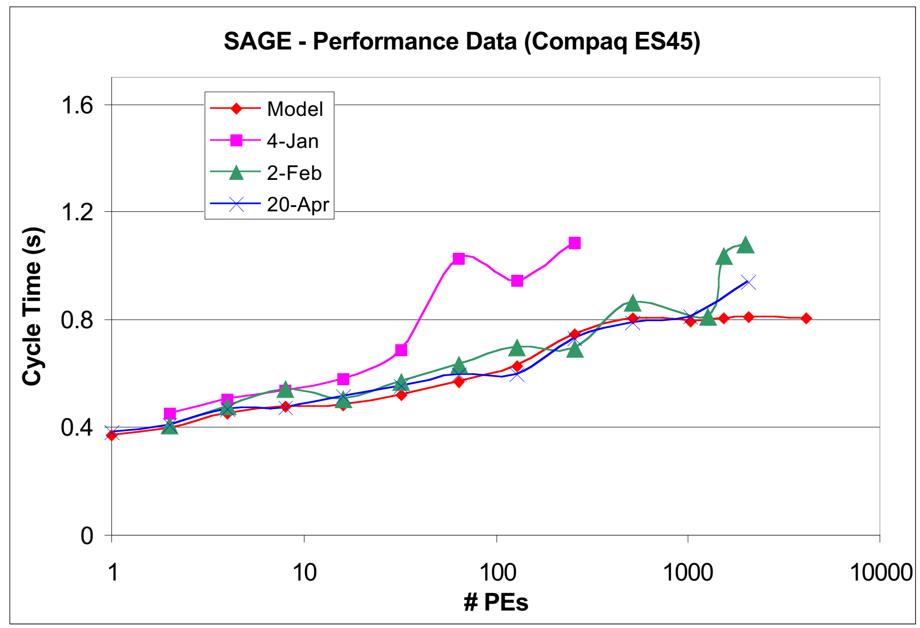
CCS-3

- Model provided expected performance
- Installation performed in stages

Late 2001:



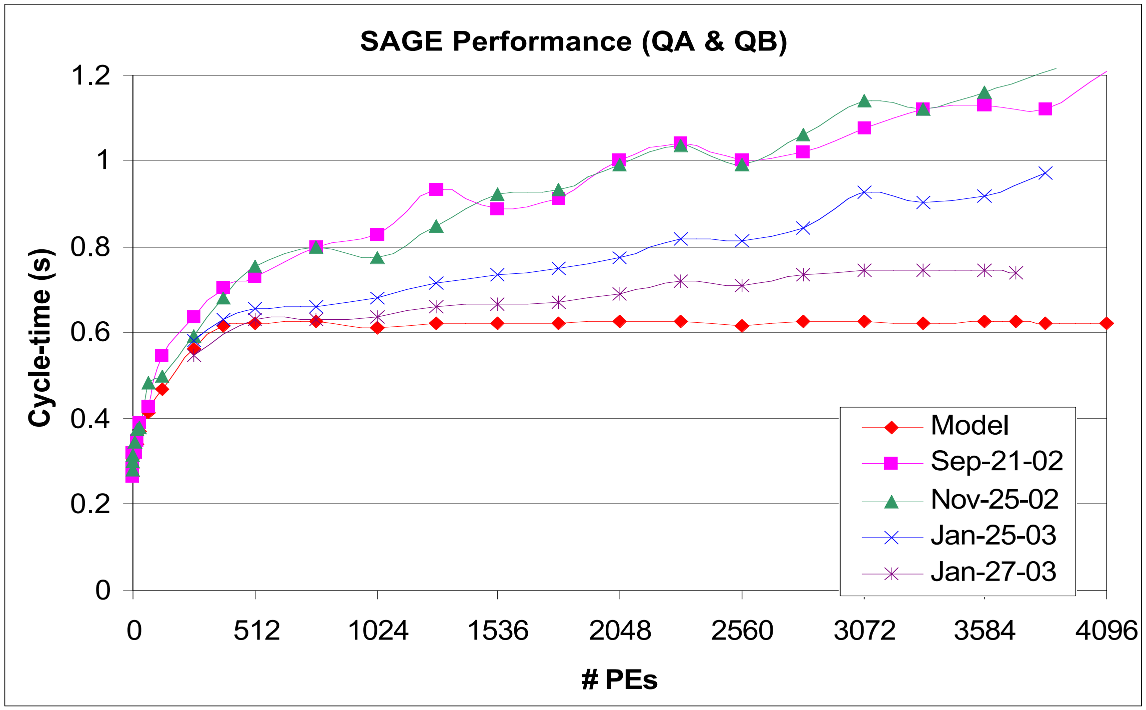
Early 2002: upgraded PCI



-> Model used to validate measurements!

Other factors: accurately predicted when 2-rails improves the performance (P>41)

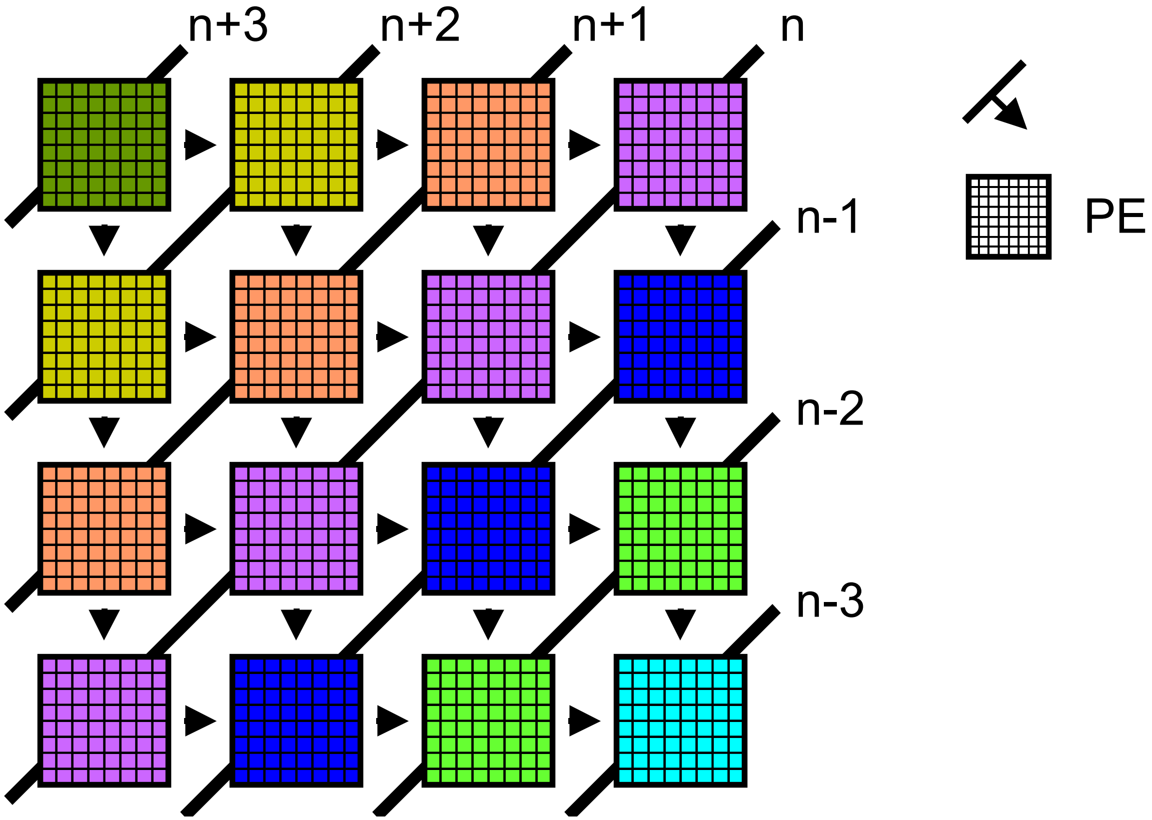
More recently: (Jan 2003)



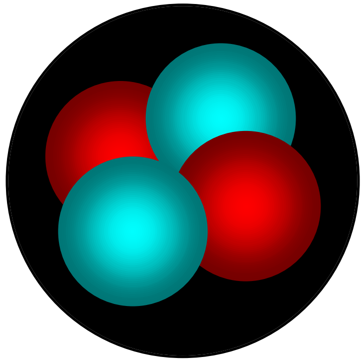
- **Performance of ASCI Q is now with ~10% of our expectation**
- **Without a model we would not have identified (and solved) the poor performance!**

CCS-3

SWEEP3D Particle Transport: 2-D Pipeline Parallelism



- **S_N transport using discrete ordinates method**
- ***wavefronts* of computation propagate across grid**
- **Model for Sweep3D accurately captures the performance**

How many  –  ?

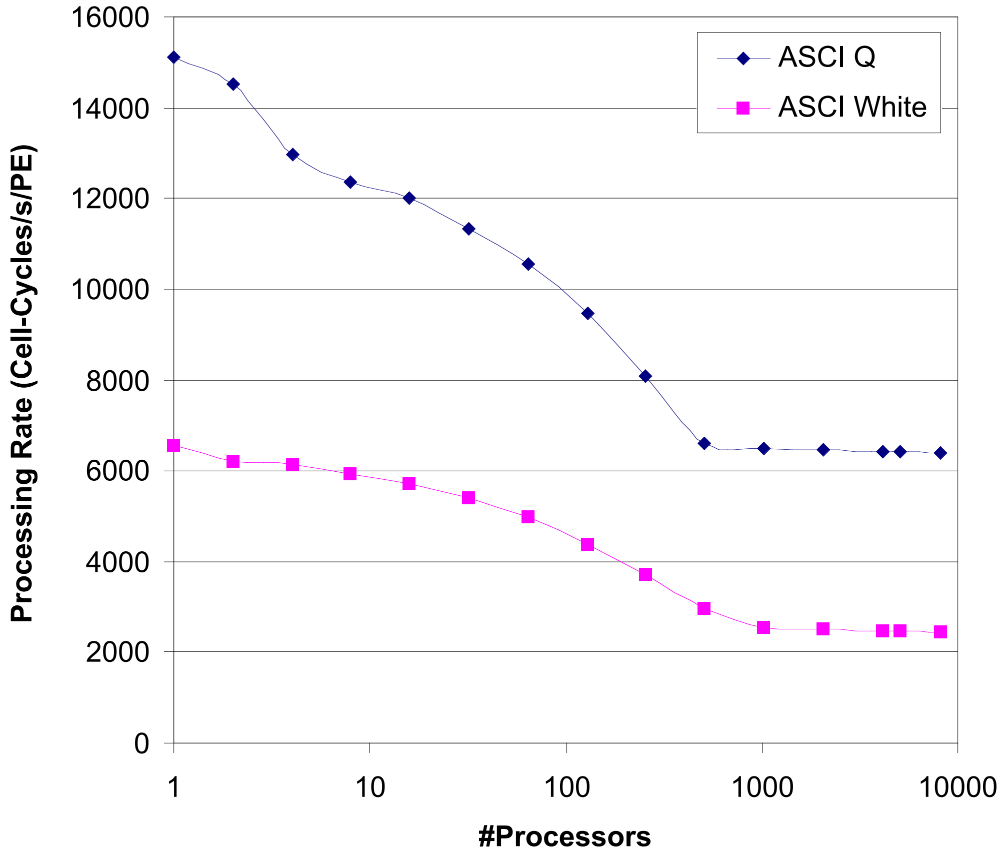
- **Basis for Performance Comparison**
  - Use validated performance models (trusted) to predict performance
  - Large problem size (Weak scaling) to fill available memory
    -> Problem size on Q is double that on the Earth Simulator
- **Compare processing rate:**
  - Equal processor count basis (1 to 5120)
  - % of system used (10 to 100%)
- **But have an unknown:**
  - performance of codes on single Earth Simulator processor

  -> Consider range of values

Model Parameters: SAGE

Parameter		Alpha ES45	Earth Simulator
P_{SMP}	**Processors per node**	**4**	**8**
CL	**Communication Links per Node**	**1**	**1**
E	**Number of level 0 cells**	**35,000**	**17,500**
f_{GS_r} f_{GS_I}	**Frequency of real and integer gather-scatters per cycle**	**377** **22**	**377** **22**
$T_{comp}(E)$	**Sequential Computation time per cell**	**68.6 *s***	42.9 *s* (%5 *of peak*) 21.4 *s* (%10 *of peak*) 14.3 *s* (%15 *of peak*) 10.7 *s* (%20 *of peak*) 8.6 *s* (%25 *of peak*) 7.1 *s* (%30 *of peak*)
$L_c(S,P)$	**Bi-directional MPI communication Latency**	6.10 *s* $S < 64$ 6.44 *s* 64 *S* 512 13.8 *s* $S > 512$	**8 *s***
$B_c(S,P)$	**Bi-directional MPI communication Bandwidth (per direction)**	0.0 $S < 64$ 78*MB*/*s* 64 *S* 512 120*MB*/*s* $S > 512$	**10*GB/s***

SAGE: Processing Rate

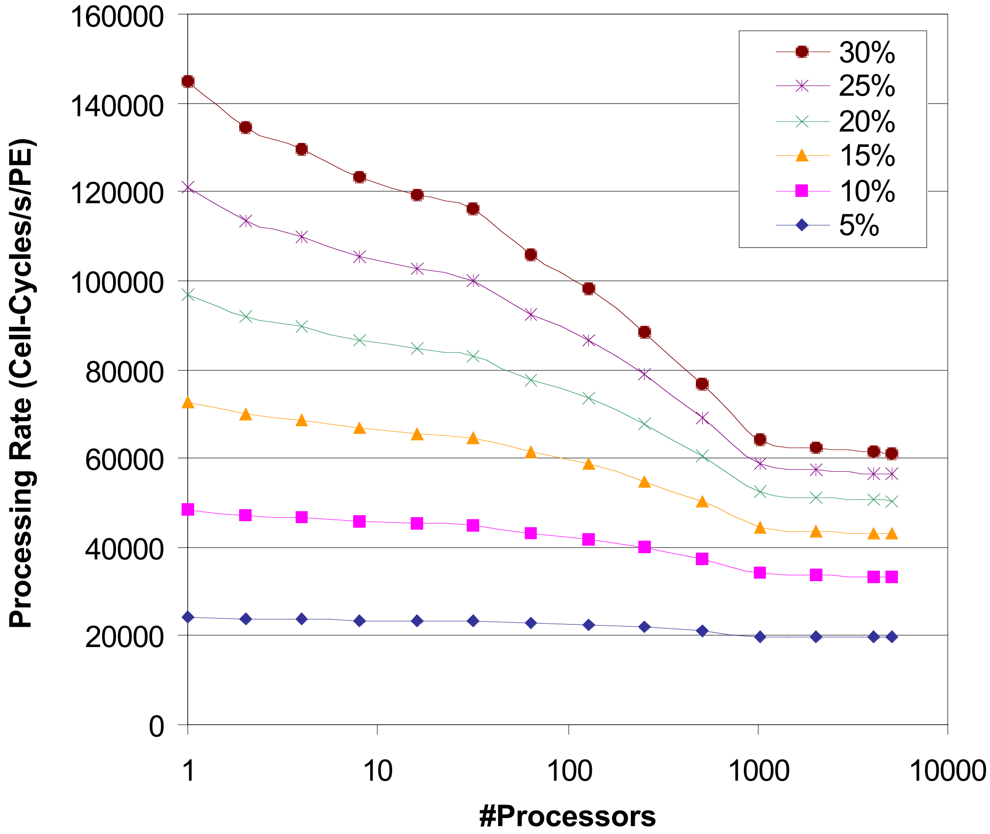


Higher is better

- **Weak-scaling**
- **ASCI Q between 2.2 and 2.8 times faster than ASCI White**
- **Data from Model (validated to 4096 PEs on Q, 8192 PEs on White)**

CCS-3

SAGE: Processing Rate



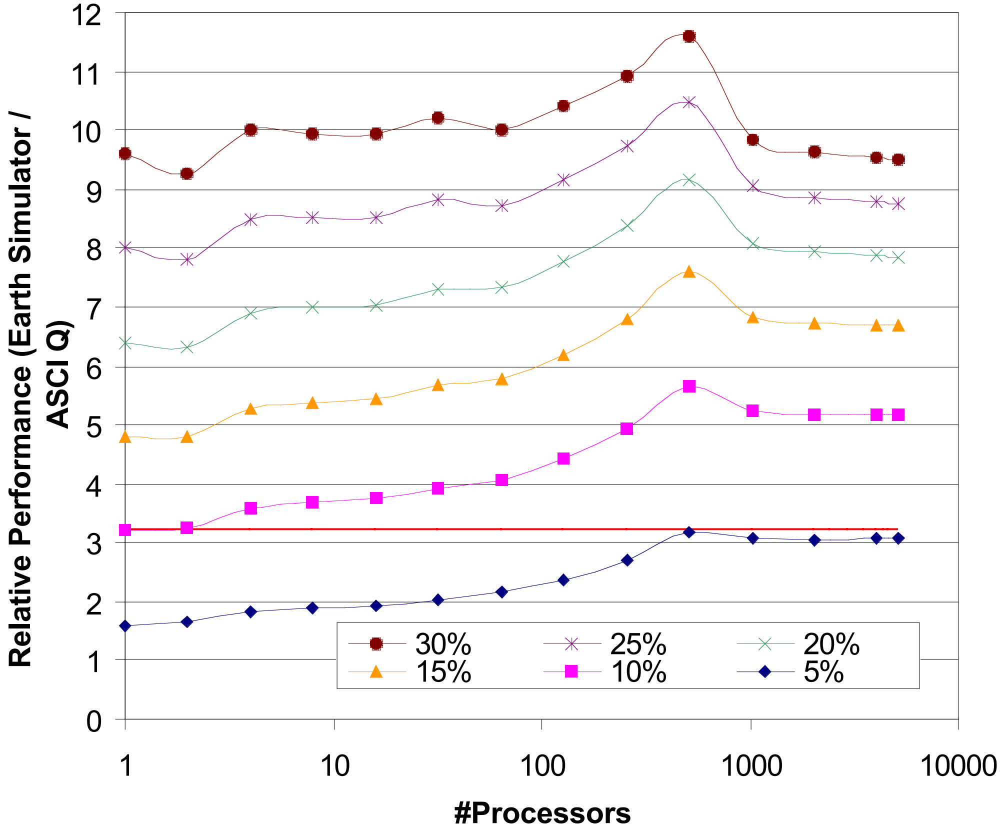
Earth Simulator

Higher is better

- **Single Processor time is unknown – modeled using range of values**

SAGE: Earth Simulator vs Q (Equal processor count)

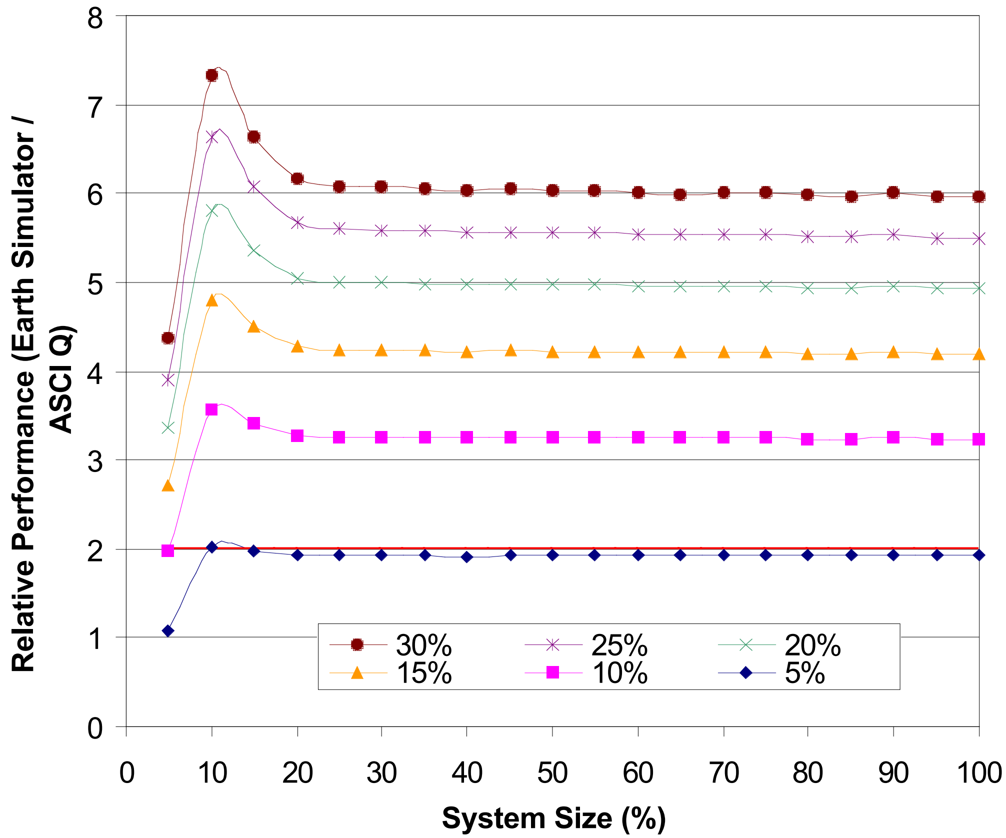
CCS-3



Higher = Earth Simulator Faster

- **Ratio of processor peak-performance is 3.2 (red line)**
- **In nearly all cases: Earth Simulator performance is better than ratio of processor peak-performance**

SAGE: Earth Simulator vs Q (% of system utilized)

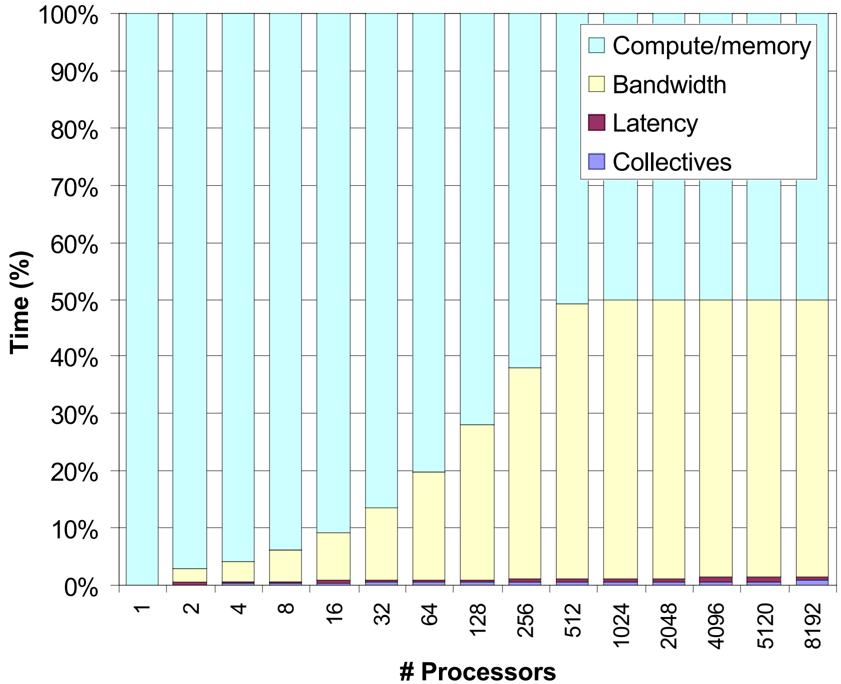


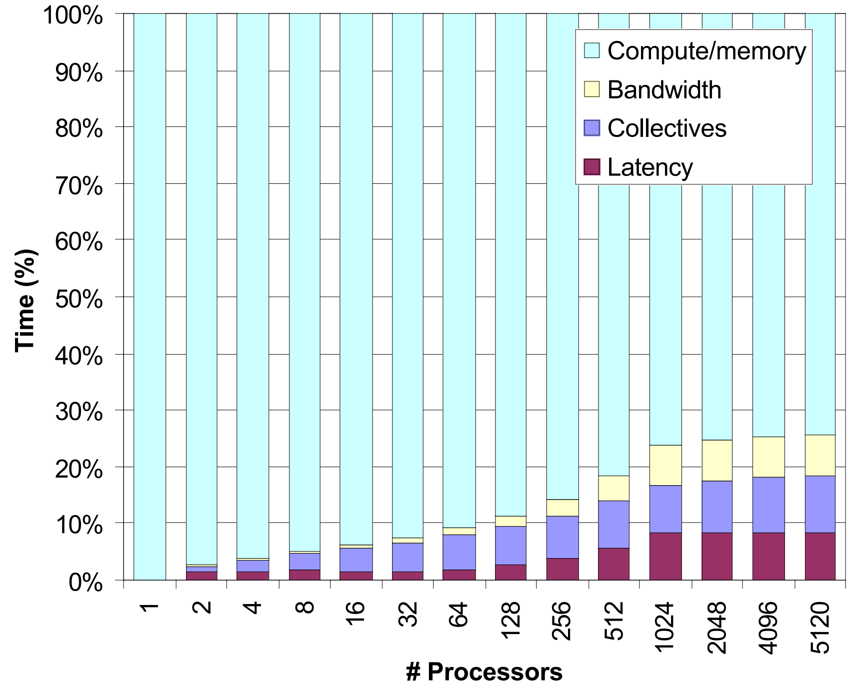
Higher = Earth Simulator Faster

- **Ratio of system peak-performance is 2 (red line)**
- **In nearly all cases: Earth Simulator better than ratio of system peak-performance**

SAGE: Component Time Predictions

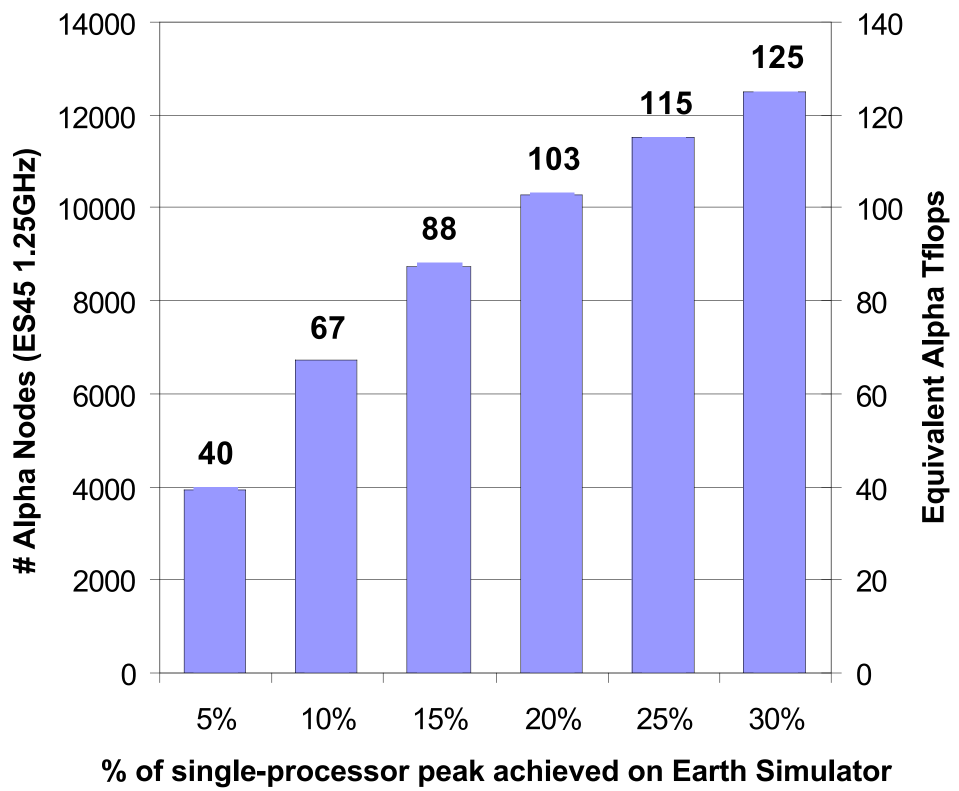
CCS-3





- **10% case considered on Earth Simulator**
- **Higher bandwidth component on Alpha**
- **Latency cost on Earth Simulator is more visible**
  - reduced computation (x3.2), increased bandwidth (x40), latency ~same

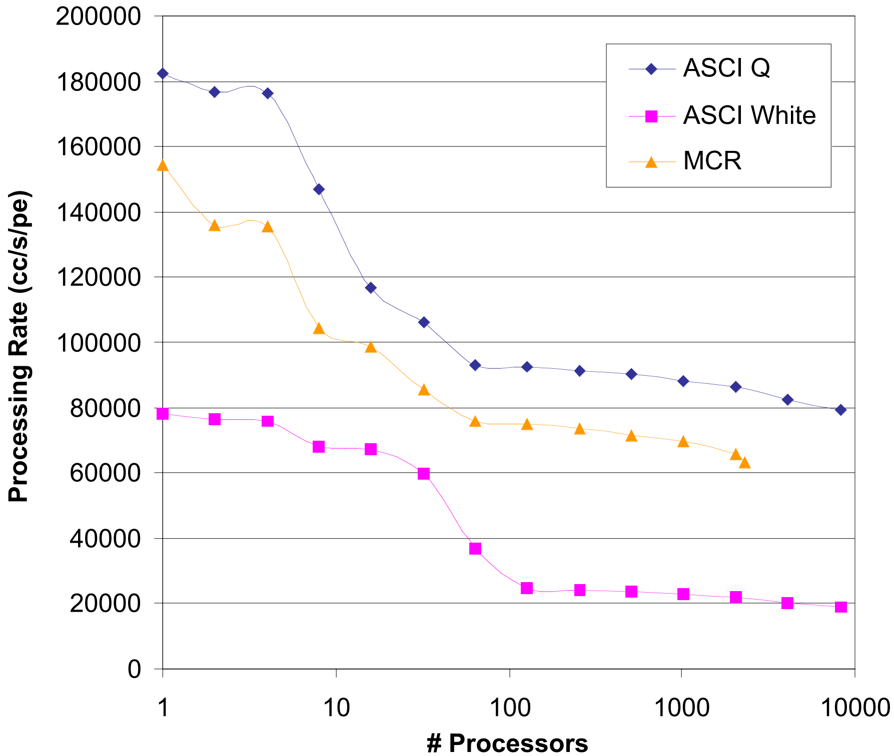
How many Alpha nodes _ Earth Simulator for SAGE?

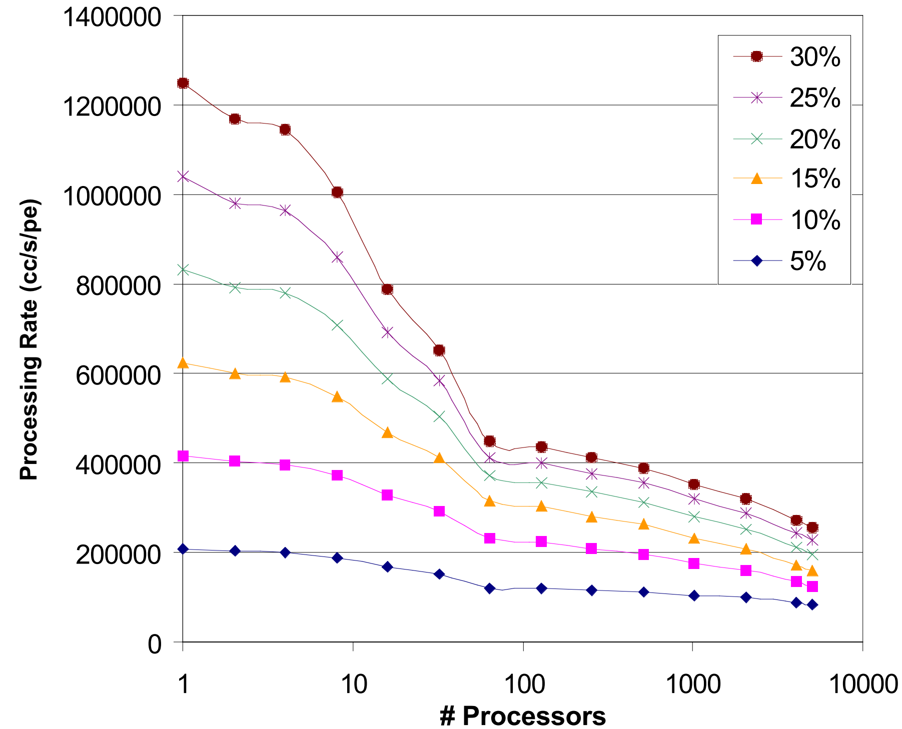


- **1 ES45 1.25GHz = 10GFlops, or 1,000 nodes = 10Tflops**
- **If assume a 10% of peak on ES -> need 67Tflops of EV68 Alphas**

Sweep3D: Processing Rate

CCS-3

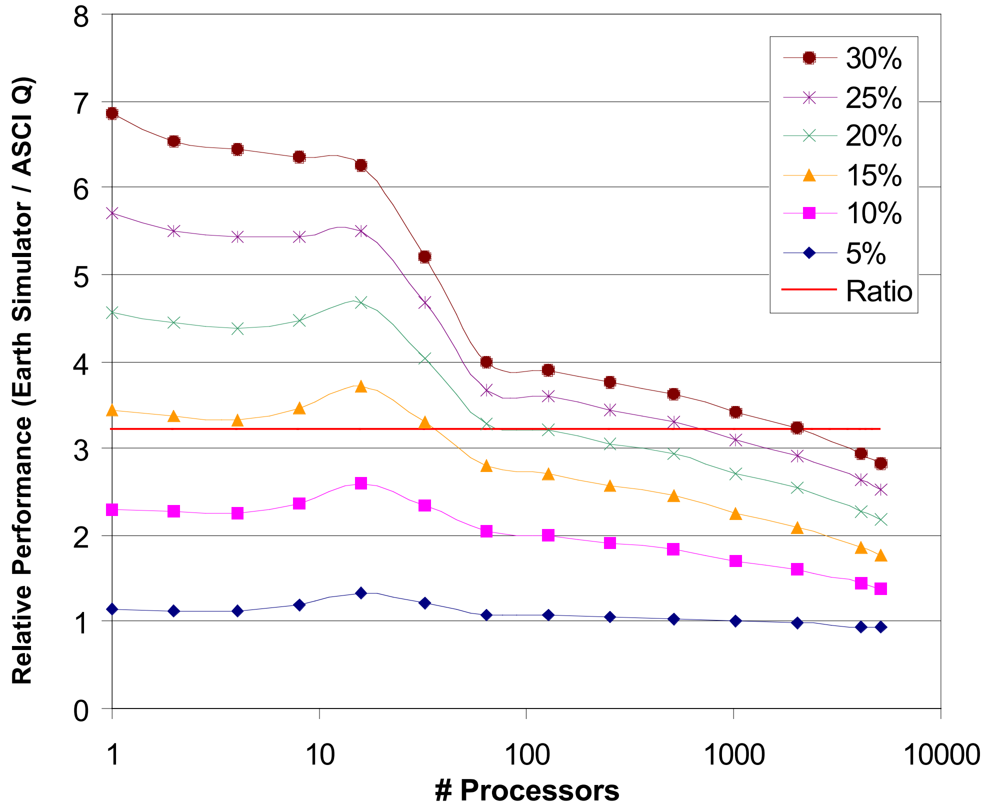




Higher is better

- **Again, range of values for Earth Simulator unknown single processor time**

Sweep3D: Earth Simulator vs Q (Equal processor count)

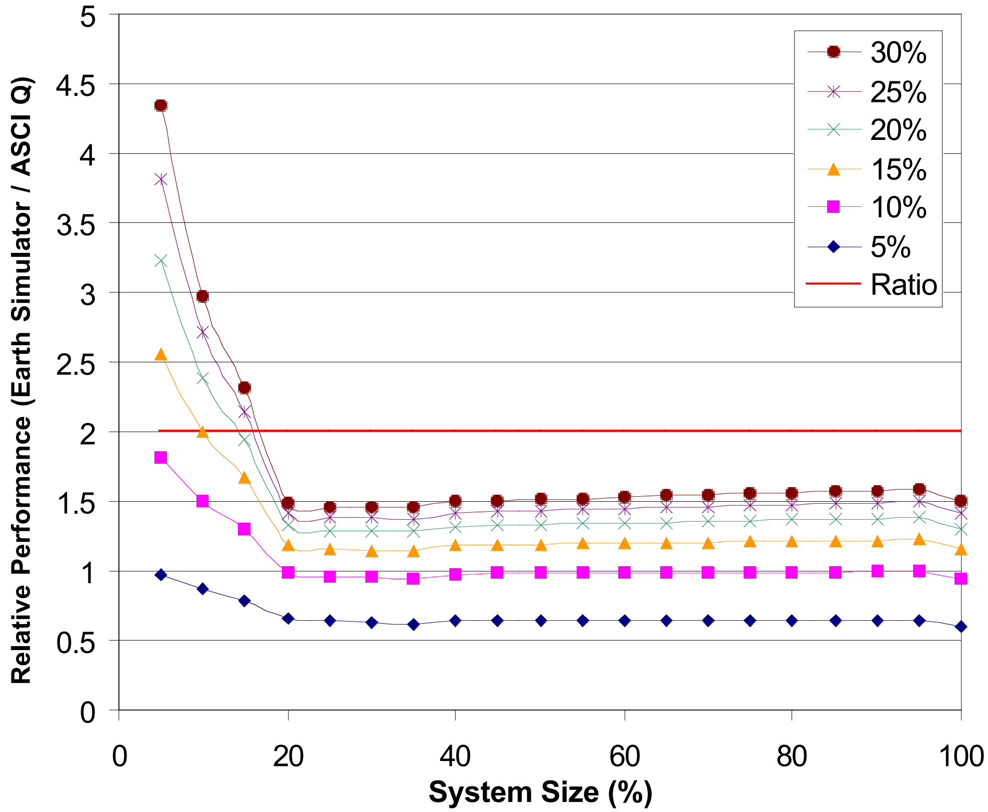


Higher = Earth Simulator Faster

- **Ratio of processor peak-performance is 3.2 (red line)**
- **At large PE counts: Earth Simulator performance is worse than ratio of processor peak-performance**

Sweep3D: Earth Simulator vs Q (% of system utilized)

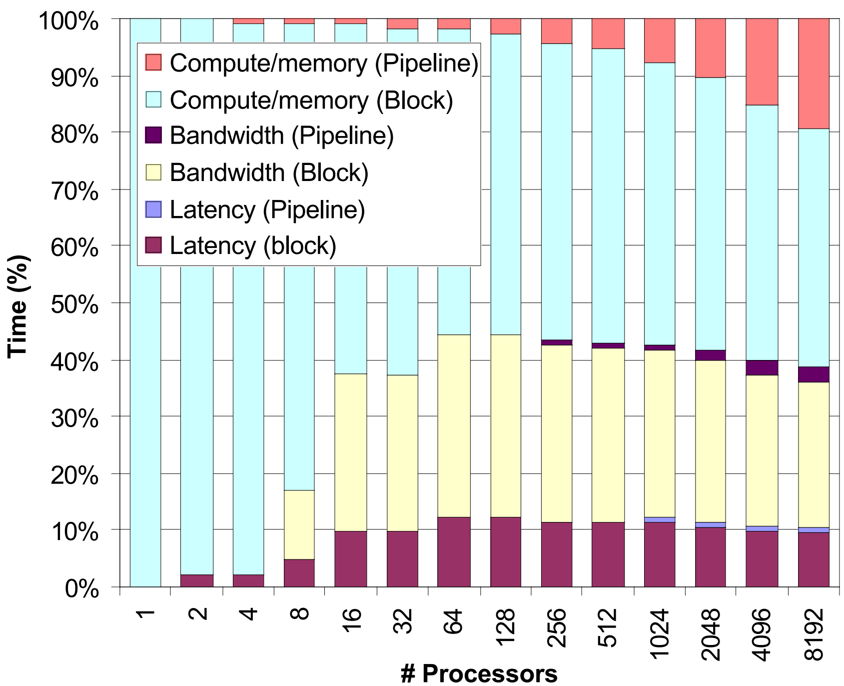
CS-3

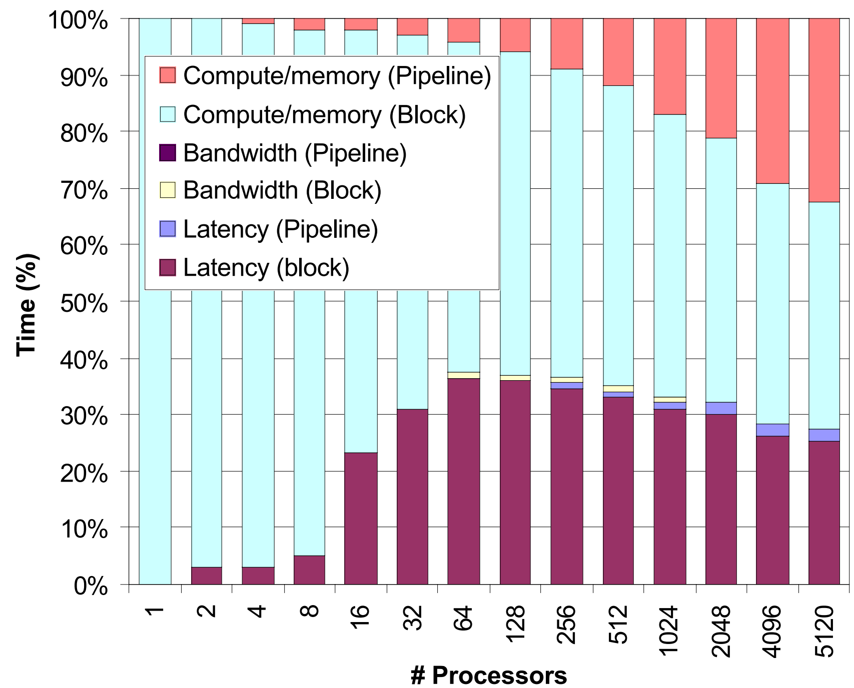


Higher = Earth Simulator Faster

- **Ratio of system peak-performance is 2 (red line)**
- **Using more than 15% of system: Earth Simulator performance is worse than ratio of system peak-performance**

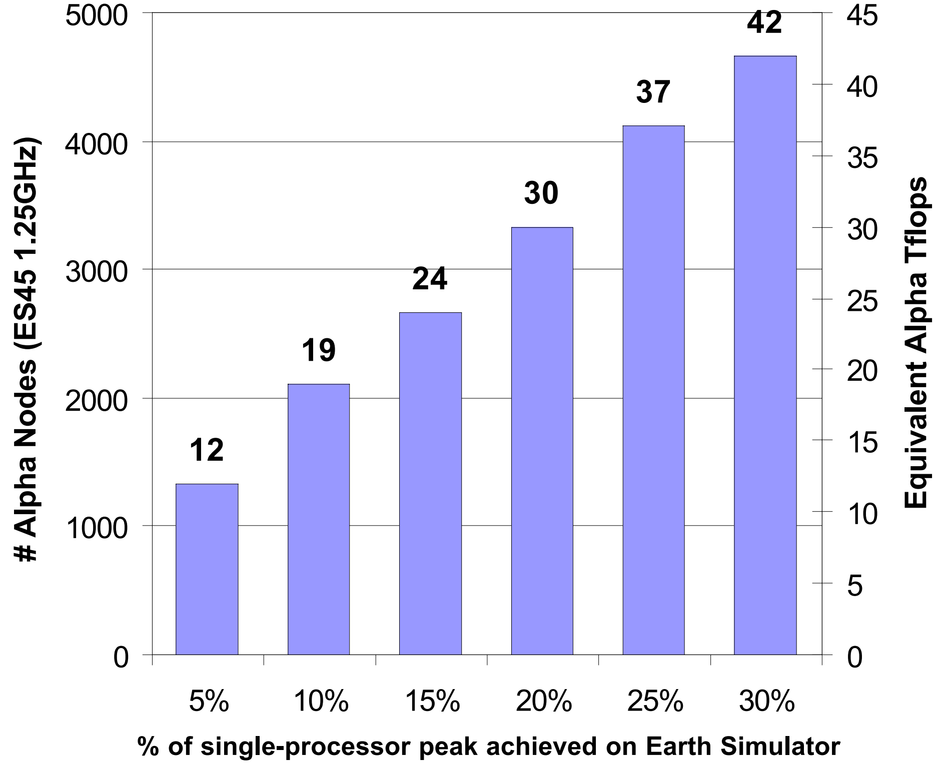
PAL Sweep3D: Component Time Predictions





- **5% case considered for the Earth Simulator**
- **Higher Latency component than SAGE on both systems**
  - Significant
- **Pipeline effect can be seen on higher PE counts**

How many Alpha nodes _ Earth Simulator for Sweep3D?



- **1 ES45 1.25GHz = 10GFlops, or 1,000 nodes = 10Tflops**
- **If assume a 5% of peak on ES -> need 12Tflops of EV68 Alphas**

PAL

Hypothetical Workload:
Assume 40% SAGE and 60% Sweep

		SAGE % of singleprocessor 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

- **Numbers in table indicate a peak Tflop rated Alpha ES45 system that would achieve the same performance as the Earth Simulator**
- **Currently: SAGE on NEC SX-6 achieved 5% on first run (Sweep3D expected to be less). This may improve over time.**

PAL

Summary

- **Models have provided quantitative information on the long-debated efficiency of large, microprocessor-based systems for HPC (instead of smaller, more-powerful but special-purpose vector systems)**
- **Comparison of the Earth Simulator and ASCI Q performance is heavily dependent on the workload**
  - At present, an Alpha system of approx. 23Tflops peak would achieve the same level of performance as the Earth Simulator on the workload considered here (60% Sweep3D, and 40% SAGE).
- **Results gives a 'reference' comparison**
  - The achievable performance on the NEC SX-6 (or Earth Simulator node) may change over time. This analysis will stay valid for the systems as they presently are.
- **Models have also been used for:**
  - Design studies (architecture and software, e.g IBM PERCS HPCS)
  - During ASCI Q installation (to validate observed performance)
  - In the procurement of ASCI Purple
  - Performance comparison of systems