CCS-3

Performance Modeling of Extreme-Scale Systems and Applications

Adolfy Hoisie

work with: **Kevin Barker, Kei Davis, Darren Kerbyson, Greg Johnson, Mike Lang, Scott Pakin, Fabrizio Petrini**

www.c3.lanl.gov/par_arch

Performance and Architecture Lab (PAL)
Computer and Computational Sciences Division
Los Alamos National Laboratory

Work funded by ASC, Office of Science, DARPA

CESC, April 2005, Washington DC

Los Alamos NATIONAL LABORATORY EST. 1943

Outline

CCS-3

- **Performance modeling**
- **Comparison of systems**
- **Performance prediction for advanced architectures**
- **Summary/Tentative conclusions**

Performance and Architecture Lab (PAL)

CCS-3

- **Performance/Architecture team at Los Alamos**
  - Modeling
  - Measurement, system software R&D, simulation
- **Application-centric performance modeling**
- **Analyze existing systems (or near-to-market systems)**
  - e.g. Cray X1, Earth Simulator, BlueGene/L …
- **Examine possible future systems**
  - e.g. IBM PERCS (DARPA HPCS), BlueGene/P, … ?
- **Recent work includes:**
  - Modeling and optimization of ASCI Q
  - Comparison of: Cray X1, Earth Simulator, ASCI Q
  - Scalability analysis of BlueGene/L
  - Initial analysis of BlueGene/P

Why Performance Modeling?

CCS-3

- **Other performance analysis methods fall short in either accuracy or practicality:**
  - » Simulation
    - Greatest architectural flexibility but takes too long for real applications
  - » Trace-driven experiments
    - Results often lack generality
  - » Stochastic
    - System-wide/throughput analysis
  - » Benchmarking
    - Limited to current implementation of the code
    - Limited to currently-available architectures
    - Difficult to distinguish between real performance and machine idiosyncrasies

Performance Modeling Process

CCS-3

- **Basic approach:**

"Fundamental Equation of Modeling"

$$T_{run} = T_{computation} + T_{communication} - T_{overlap}$$

$$T_{run} = f\,(T_{1\text{-}CPU}\,,\ \textit{Scalability})$$

where T1-CPU is the single processor time

- **We are not using first principles to model single-processor computation time.**
  - Rely on measurements for $T_{1\text{-}CPU}$. May be:
    - » **time per subgrid,**
    - » **time per cell,**
    - » **calculated using measured rate and # of FLOPS per subgrid**

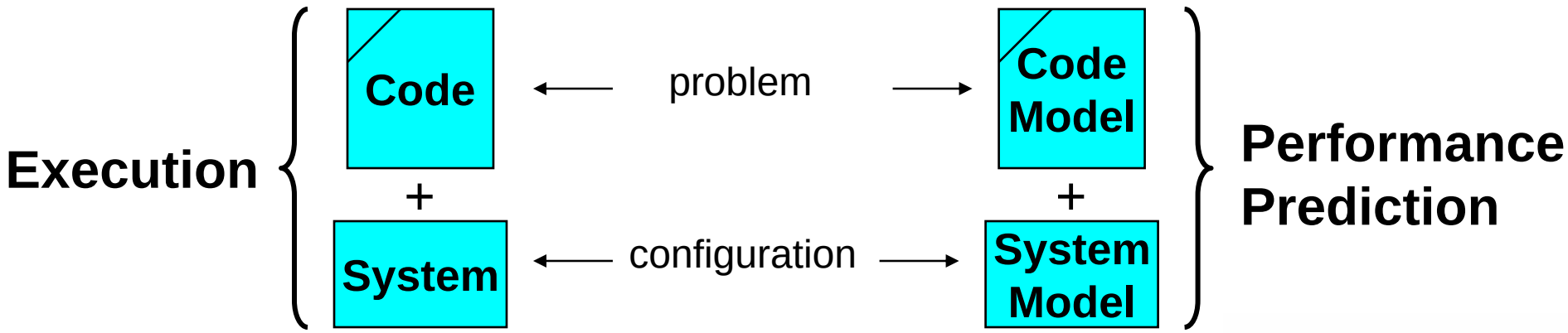
Los Alamos NATIONAL LABORATORY EST. 1943

What is a performance model?

CCS-3

- **A model encapsulates the understanding of:**
  - What resources an application uses during execution

  (i.e. a functional description)
- **This is coupled with a system model:**
  - how long it takes to satisfy various resource requirements
- **Parameterized:**



Los Alamos NATIONAL LABORATORY EST. 1943

Modeling Successes

CCS-3

- **Codes:**
  - SWEEP3D,
  - SAGE,
  - Partisn,
  - LBMHD,
  - HYCOM
  - MCNP,
  - POP,
  - TYCHO,
  - UMT2K,
  - RAGE (in progress),
  - RF-CTH (in progress),
  - CICE (in progress)

- **Machines:**
  - ASCI Q,
  - ASCI BlueMountain,
  - ASCI White,
  - ASCI Red,
  - CRAY T3E,
  - Earth Simulator,
  - Various clusters (AMD, IA-64, Alphas, Xeon-64, etc),
  - BlueGene/L,
  - BlueGene/P,
  - CRAY X-1,
  - CRAY XD-1,
  - IBM PERCS (early designs)
  - Red Storm

Outline

CCS-3

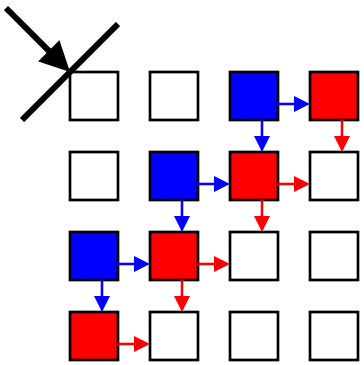
- Performance modeling
- **Comparison of systems**
- An experiment-ahead for advanced architectures
- Tentative conclusions

Application Performance
1) Sweep3D

CCS-3

- **3-D spatial grid, partitioned in 2-D**
- **Pipelined wavefront processing**
  - Dependency in ‘sweep’ direction

2-D example:



- **Parallel Characteristics:**
  - logical neighbors in X and Y
  - Small message sizes: 100’s bytes (typical)
  - Number of processors determines pipeline length ($P_X + P_Y$)

Sweep3D model: Int. Journal High Performance Computing Applications, 2000

Application Performance
2) SAGE

CCS-3

- **Adaptive Mesh Refinement (AMR) hydro-code**
- **3 repeated phases**
  - Gather data (including processor boundary data)
  - Compute
  - Scatter data (send back results)
- **3-D spatial grid, partitioned in 1-D**
- **Parallel characteristics**
  - Message sizes vary, typically 10 - 100's Kbytes
  - Distance between neighbors increases with scale

SAGE model: **Supercomputing 2001, Denver**

Los Alamos NATIONAL LABORATORY EST. 1943

BlueGene/Light Overview

CCS-3

- **Node**
  - Dual Core Embedded PowerPC 440
  - 256MB or 512MB memory
- **700MHz (500MHz prototype)**
- **Network**
  - 3-D torus (point-to-point) & Tree network (broadcast, …)
- **4 floating-point per cycle**
  - 2.8 GFlops per processor core
- **Use either 1 PE or 2 PEs per node**
- **Largest system - Lawrence Livermore, 2005 (ASC)**
  - 32 x 32 x 64 nodes (64K nodes, 128K processor cores)
  - Peak performance: 360 Tflops
- **Small physical footprint**
  - 2 nodes per compute card, 16 cards per board, 32 boards per rack

Los Alamos NATIONAL LABORATORY EST. 1943

RedStorm Overview

CCS-3

- **Node**
  - Single AMD Opteron
  - 1-8 GB main memory
- **2.0GHz, 2.4GHz**
- **Network**
  - 3-D torus (in two dimensions at Sandia)
  - SeaStar Network processor (uses embedded PowerPC)
- **2 floating-point per cycle**
  - 4 GFlops per processor
- **Largest system - Sandia, 2005 (ASC)**
  - 27 x 16 x 24 nodes (10,368 nodes)
  - Peak performance: 41.5 Tflops

System Comparison

CCS-3

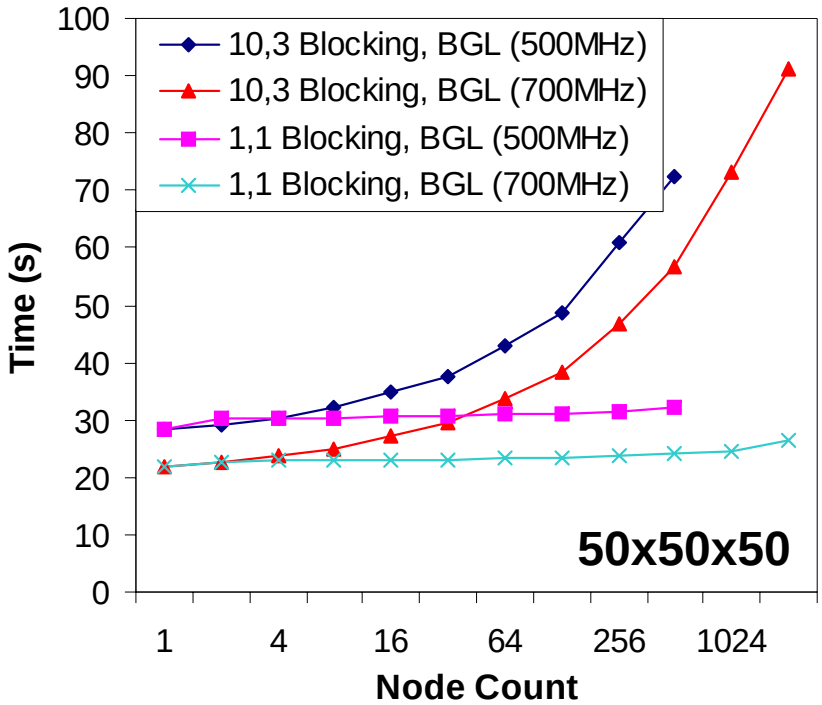
	BlueGene/L	RedStorm	ASCI Q
System Peak	360 Tflops	41.5 Tflops	20 Tflops
Processor Type	PowerPC 440	AMD Opteron	Alpha EV68
Node Count	65,536	10,368	2,048
Processor Count	131,072	10,368	8,192
Processor Speed	0.7GHz	2.0GHz	1.25GHz
Processor Peak	2.8 GF/s	4.0 GF/s	2.5 GF/s
Topology	3D-torus	3D-torus	Fat-tree
MPI network Latency	3µs	2-3µs	6µs
MPI network Bandwidth	160MB/s	2.0GB/s	300MB/s

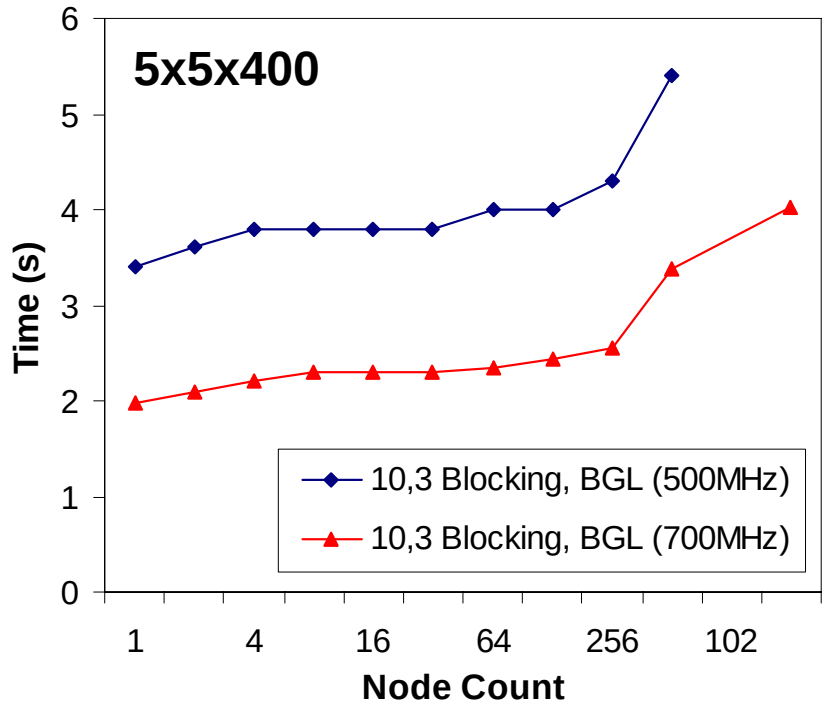
Notes: BlueGene/L based on Lawrence Livermore installation
RedStorm based on Sandia installation
ASCI Q production machine at Los Alamos (2002)

Sweep3D (measured on BG/L)

CCS-3

- **3 sets of measurements:**
  - 50x50x50 grid per PE (blocked in two different ways)
  - 5x5x400 grid per PE (one blocking regime)



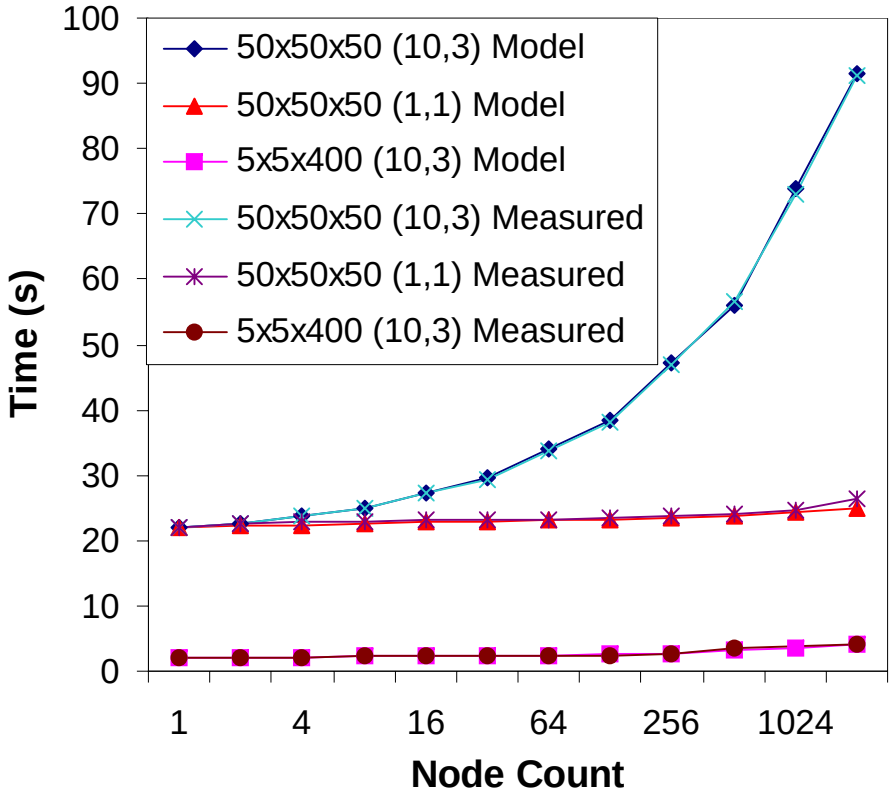


- **Overall: good scaling (using best blocking)**

Los Alamos NATIONAL LABORATORY EST. 1943

Sweep3D Model Validation (BG/L)

CCS-3



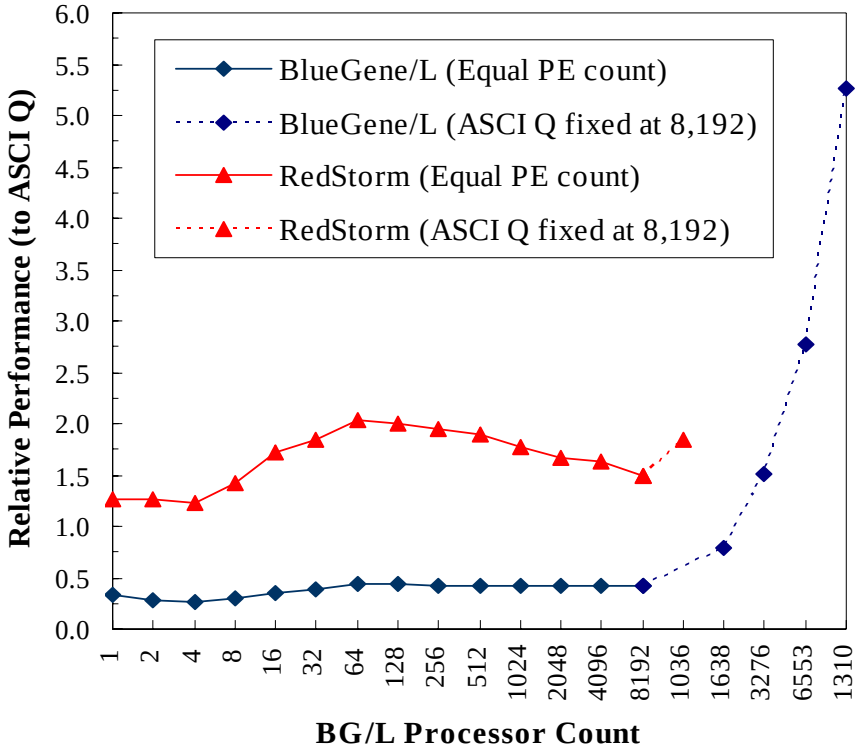
- **700MHz**
- **Model prediction error:**
  - 7.2% (maximum)
  - 1.7% (average)

- **NB: VNM vs. COP mode (using 2 vs 1 PEs per node) :**
  - Factor of ~1.9x higher performance

Relative performance
BG/L and RedStorm to ASCI Q

CCS-3

- **2 regions in graph:**
  - equal processor count (up to 8,192 processors)
  - ASCI Q fixed size (above 8,192 processors)



Equal PE count:

BG/L is **~0.42x** speed of Q

RedStorm is **1.25x - 2x** faster than Q

Full-sized system:

BG/L is **~5.5x** faster than Q

RedStorm is **~1.85x** faster than Q

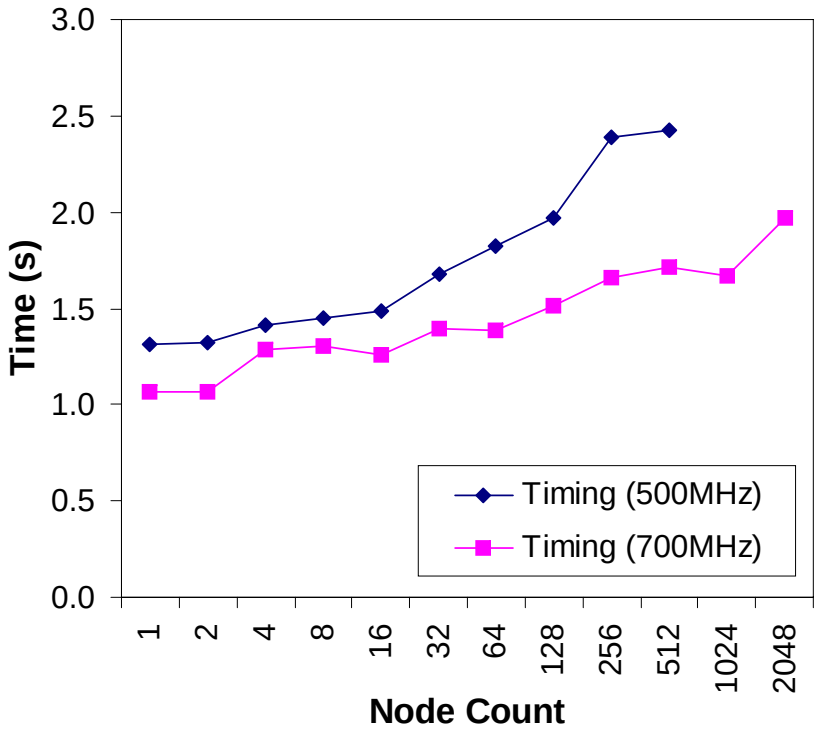
(5x5x400 sub-grids with best blocking)

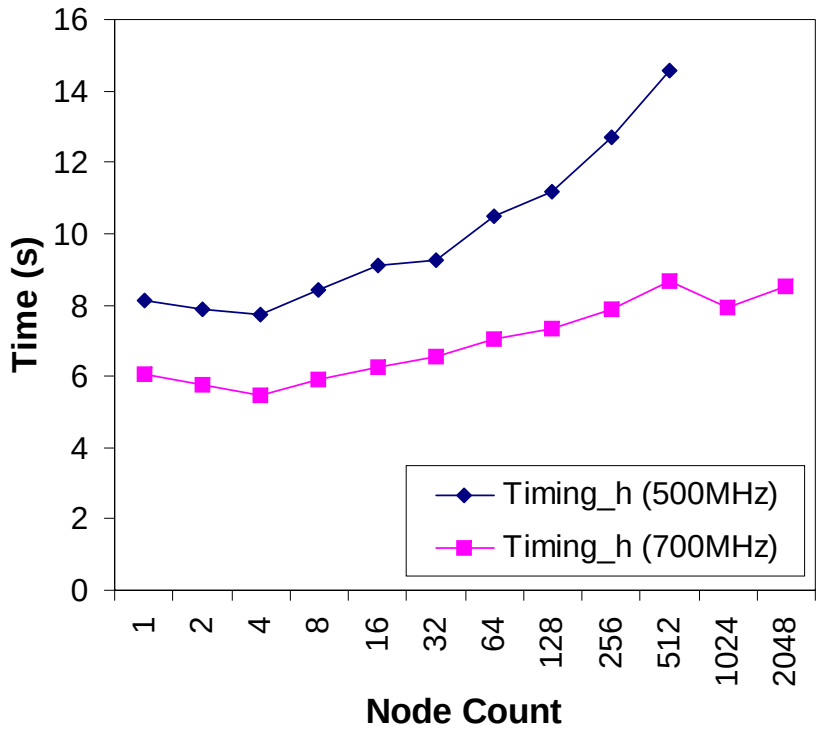
Los Alamos NATIONAL LABORATORY EST. 1943

SAGE (measured performance BG/L)

CCS-3

- **Two inputs:**
  - timing: 13,500 cells per PE
  - timing_h: 35,000 cells per PE

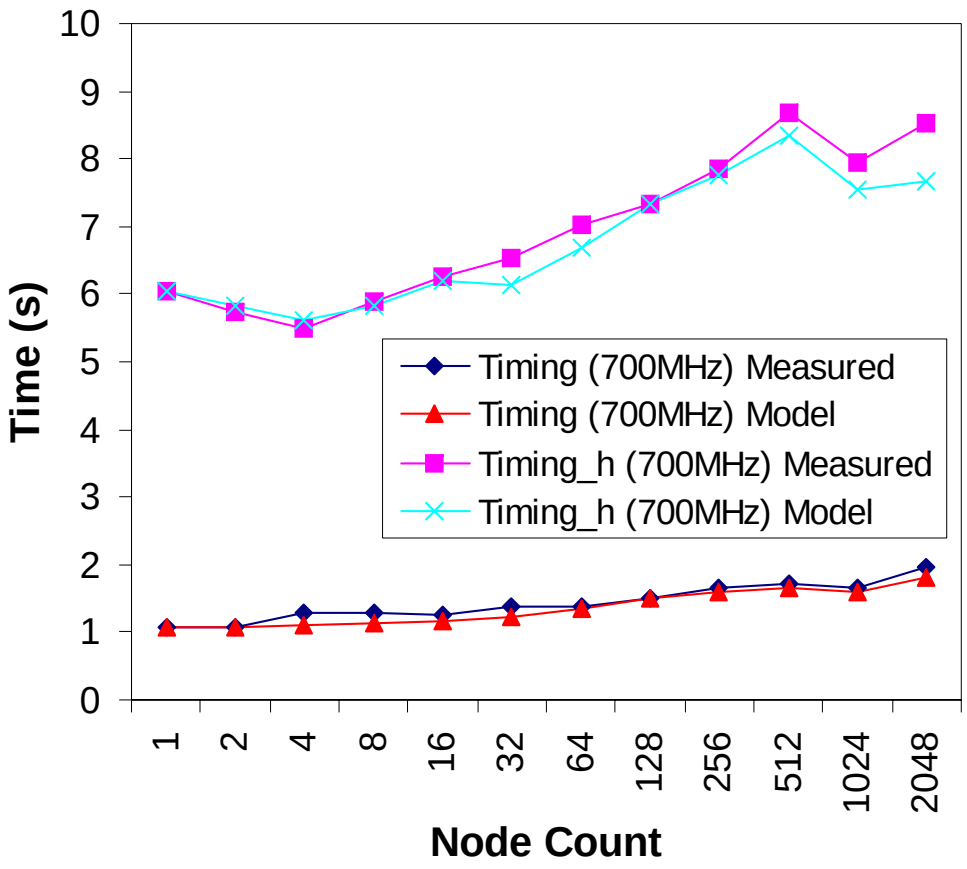




Los Alamos NATIONAL LABORATORY EST. 1943

SAGE model validation BG/L

CCS-3



- **700MHz**
- **Model prediction error:**
  - 10.1% (maximum)
  - 4.1% (average)

- **NB: VNM vs. COP mode (using 2 vs 1 PEs per node) :**
  - Factor of ~1.1x higher performance

Outline

CCS-3

- Performance modeling
- Applications of modeling: rational system integration, large-scale system comparison, performance prediction, system design
- **Performance predictions for advanced architectures**
- Tentative conclusions

An Experiment-Ahead for Petaflop Architectures – The "Racehorse"

CCS-3

- Processor roadmap based on Moore's law
- Network bandwidth based on projected Infiniband for 2008
- Network latency based on current lowest value, with a small assumed improvement over time

	2004	2006	2008
Processor Peak (Gflop/s)	**8**	**20**	**50**
Network Bandwidth (GB/s)	**1**	**4**	**16**
Network Latency (μs)	**4**	**3**	**1**

Los Alamos NATIONAL LABORATORY EST. 1943

An Experiment-Ahead for Petaflop Architectures

CCS-3

- What would the Racehorse look like?

	2004	2006	2008
System Peak (Pflops)	**0.1**	**0.4**	**1.0**
Processor Count	**12,500**	**20,000**	**20,000**

- What would the input deck for the app look like?

	2004	2006	2008
Total problem size	**$2x10^8$**	**$5x10^8$**	**$1x10^9$**
cells size per processor	**16,000**	**25,000**	**50,000**

Los Alamos NATIONAL LABORATORY EST. 1943

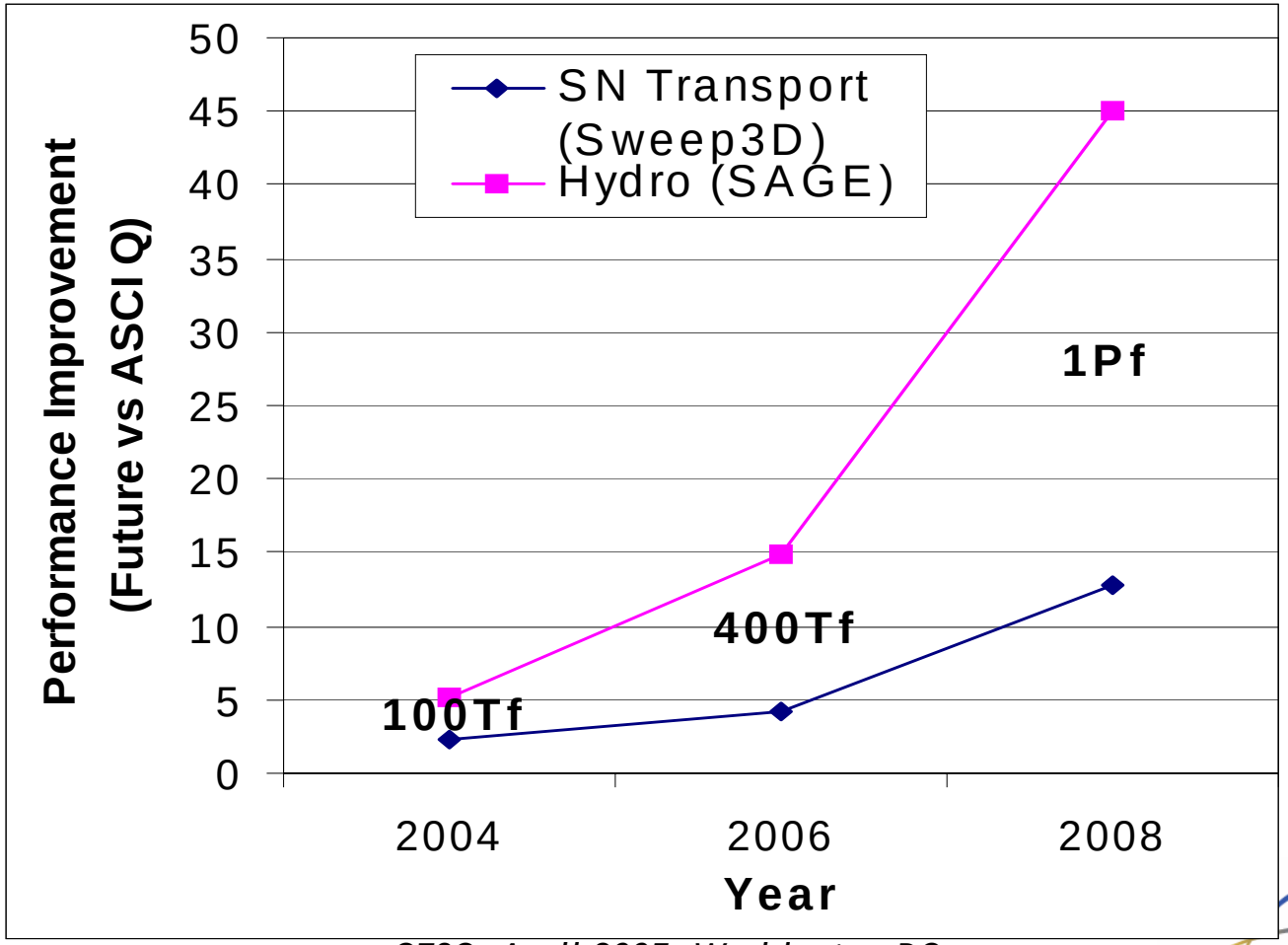
An Experiment-Ahead for Petaflop Architectures

CCS-3

- Use expected system characteristics and problem sizes
- Use PAL's performance models
- Provide expected performance improvements over time
- Assume we well get same % of peak on a single CPU

Los Alamos NATIONAL LABORATORY EST. 1943

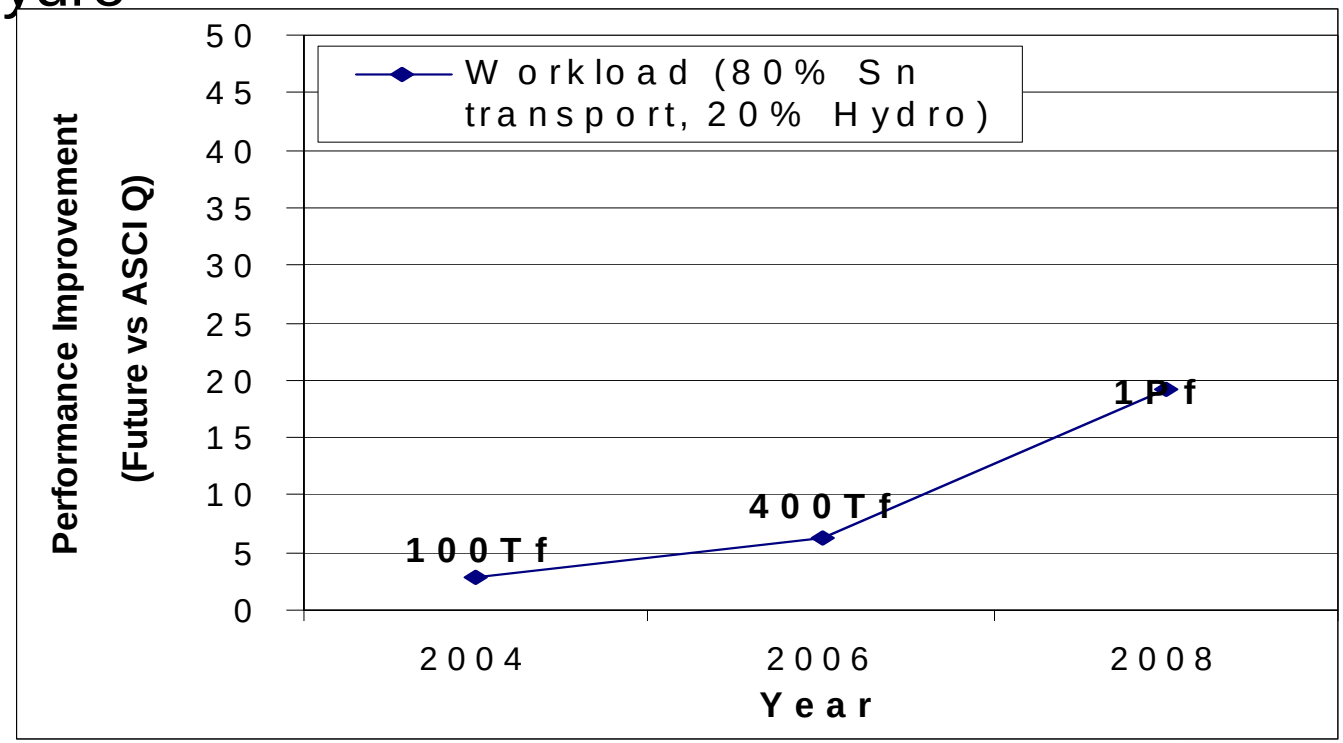
Performance improvements relative to ASCI Q



Performance improvements relative to ASCI Q

CCS-3

- Workload assuming 80% Sn transport and 20% Hydro



Los Alamos NATIONAL LABORATORY EST. 1943

A specific petaflop architecture

		ASCI Q	BlueGene/L	BlueGene/P*
System	System Peak Performance	20 Tflops	360 Tflops	1 Pflops
	Processor Type	Alpha EV68-CD	PowerPC 440	PowerPC 440
	Node Count	2,048	65,536	65,536
	Processor Count	8,192	131,072	262,144
	Processor Speed	1.25GHz	0.7GHz	1.0GHz
Nodes	Processors per Node	4	2	4
	Peak Floating-Point rate	10 GF/s/node	5.6 GF/s/node	16 GF/s/node
	Cache	8MB (L2 / PE)	4MB (L3 / node)	8MB (L3 / node)
	Main Memory (node)	8-16 GB	256-512MB	0.5-4 GB
Network	Topology	Fat-tree	3D-torus	3D-torus
	MPI network Latency	6µs	3µs	1µs
	MPI network Bandwidth	300MB/s	160MB/s	500MB/s

** BG/P parameters shown with permission from IBM. The system characteristics are likely to evolve.*

BG/P Performance characteristics (assumed)

CCS-3

- **No hardware available for measurement**
  - Use assumed hardware performance characteristics

network latency: **1.0 µs**

network bandwidth: **500 MB/s**

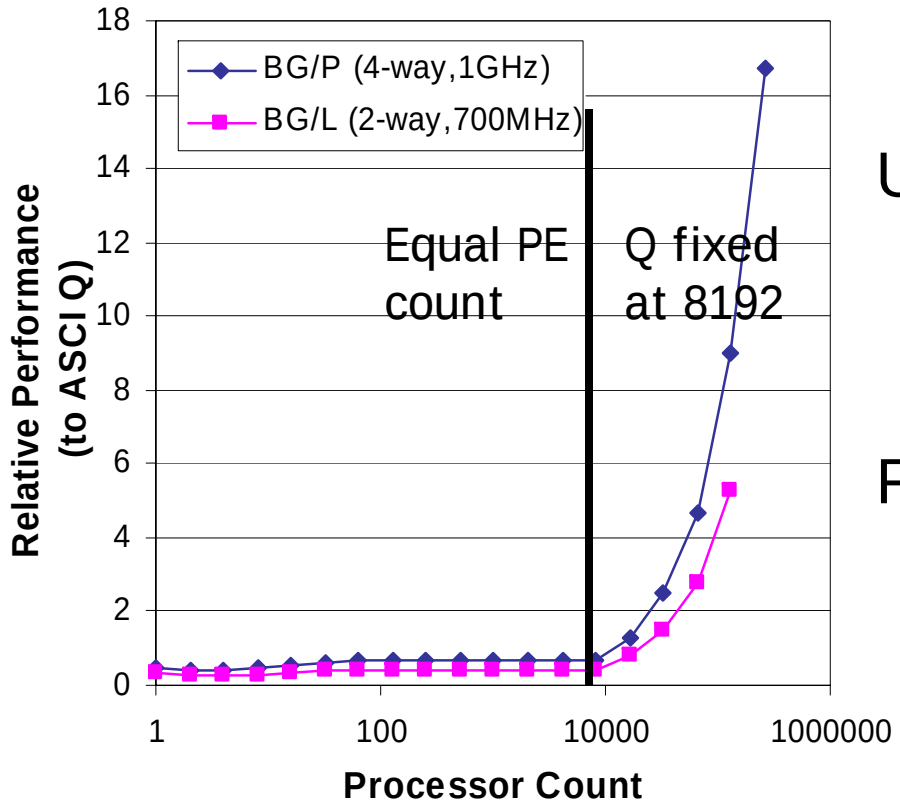
achievable processor speed: **3.75% of peak** **(single node)**

- **Optimistic view**
- **2 types of analysis:**
  - 1) Sensitivity analysis
  - 2) Expected performance
- **Sweep3D : 10K cells per processor** **(5x5x400 sub-grid size)**

Expected Performance improvement Sweep3D

CCS-3

- **Comparison with ASCI Q - 2 regions in graph**
  - Equal processor count (up to 8,192 processors)
  - ASCI Q fixed size (above 8192 processors)



Up to 8192 PEs:

BG/L is **~0.42x** speed of Q

BG/P is **~0.62x** speed of Q

Full-sized system:

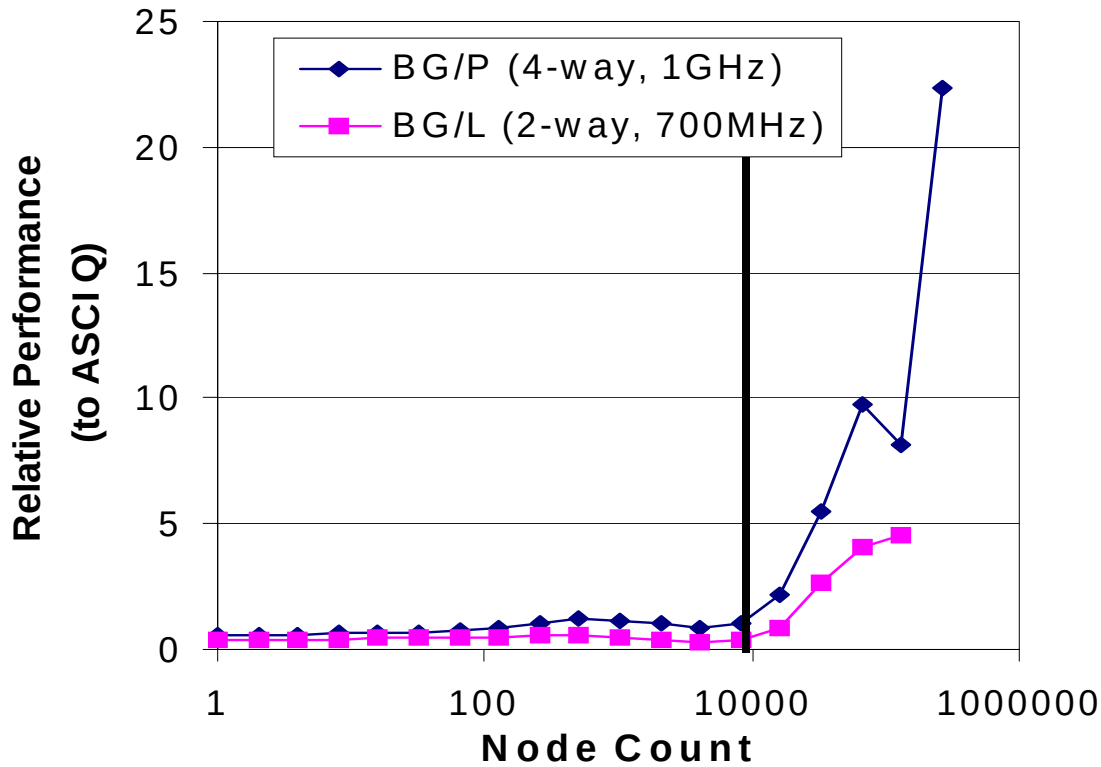
BG/L is **5.3x** faster than Q

BG/P is **16.7x** faster than Q

Los Alamos NATIONAL LABORATORY EST. 1943

Expected Performance improvement SAGE

CCS-3



- Performance depends on the problem, distance between communicating neighbors, and size of each dimension on the 3-D torus
- Mapping assumed is a default one (may be pessimistic)
- Kinks in the curve are a result of this mapping with changing PE count

Los Alamos NATIONAL LABORATORY EST. 1943

Tentative conclusions

CCS-3

- **80% Sweep + 20% Sage: BG/P would be faster by a factor of 18 on the full BG/P configuration – the Racehorse was 19 X…**
- **We ended up in the same place!**
- **Is it a new law of nature?**
- **We are chasing the technology (the Racehorse) or we are throwing more (less powerful) resources at the problem.**
- **The 2 approaches are very similar in nature – and the interplay between surface-to-volume and resources just happened to end up at the same point**
- **We are falling behind in efficiency**
- **Did not consider important issues such as system software, power, footprint, etc**

Summary

CCS-3

- **Accurate performance models of entire applications developed by PAL based on novel methodology**
- **Multiple uses of modeling in the system and application areas**
- **Models can be used to predict performance on future architectures, and conversely, design future systems for a workload of interest**
- **Performance analysis of BG/L and RedStorm, compared against Q**
- **Point design studies for a hypothetical and a specific petaflop architecture**
- **Disclaimer: other applications will result in different performance!**

Acknowledgements

CCS-3

- **Thanks to:**
  - IBM BlueGene Team, especially:
    - » **George Almasi, Manish Gupta, Jose Moreira, Bob Walkup**
  - Sandia RedStorm Team, especially:
    - » **Jim Tomkins, Sue Kelly**
  - US Department of Energy through Los Alamos National Laboratory contract W-7405-ENG-36
  - Work funded by the ASC program

Los Alamos NATIONAL LABORATORY EST. 1943

The End

CCS-3