The Red Storm Architecture at a Sustained 100 TeraFlops

Jim Tomkins

SOS8

Charleston, South Carolina

April 12 - 14, 2004

Outline

Application Codes and Code Characteristics

Sustained Performance

The Red Storm Architecture at a Sustained 100 TeraFlops

Some Important Codes

CTH - Hydro / Shock Physics, Structured Grid with AMR

Alegra - Hydro, Shock Physics, Radiation Transport - Unstructured Grid, ALE

Presto - Structural Dynamics - Unstructured Grid, Contact Algorithm

Calore - Heat Transfer - Unstructured Grid

Salinas - Structural Mechanics - Unstructured Grid

ITS - Implicit Monte Carlo Radiation Transport with Electrical Impacts

Sage - Hydro, Block Structured with AMR

Partisn - Radiation Transport, Structured Grid

ALE3D - Hydro - Unstructured Grid, ALE

LAMMPS - Molecular Dynamics

Xyce - Circuit Modeling (Parallel Spice)

Application Code Characteristics

Focus is on Scientific and Engineering Codes - Mostly PDEs

Many Codes are 3-D Meshes

- **Structured Grids**
- **Unstructured Grids - Indirect Addressing**
- **Adaptive Mesh Refinement - Move lots of data around machine**

Sparse Matrices - Low computation to memory access ratio

Complex Equations of State - Lots of wasted cache lines

Solvers

- **Explicit**
- **Implicit**
- **Monte Carlo**
- **Transient and Steady State**

Application Code Characteristics

Memory Access

- Codes go through most of the node memory each time step
- A lot of indirect addressing
- Poor cache reuse for data
- Bandwidth and Latency are extremely important to performance

Node to Node Communication

- Most Codes are tightly synchronized
- Lots of communication
- Latency and Bandwidth are extremely important to scalability

What is Sustained Performance

Peak - Not to be exceeded number

Linpack - 70 - 90^{+}% of peak on a single processor

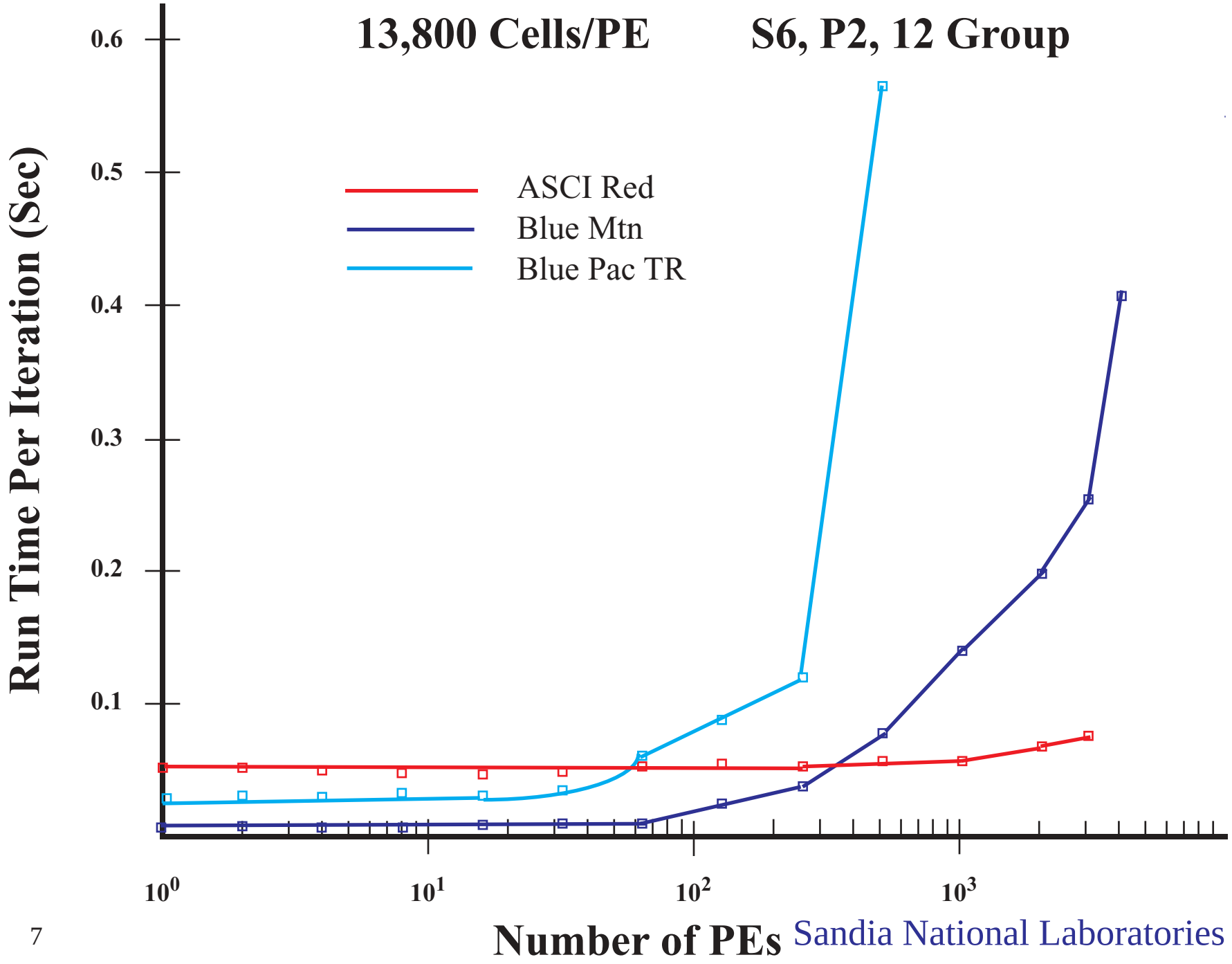
Parallel Scalability

Production Codes

- **Floating Point Operations per second**
- **Integer Operations per second**
- **Total Operations per second**

Problem Run-Time

Parallel S_n Neutronics



Why

Load Imbalance - No

Each processor has the same amount of work to do, 13500 cells per processor. There is no AMR and the grid is structured.

Operating System Noise - Yes

Blue Mtn shows scaling loss inside SMP box.
For both Blue Mtn and Blue Pac OS Noise impacts collective operations and communications in general - messagaes are typically a few hundred bytes

Communication Overhead

Latency - ~15μs on ASCI Red, closer to 100 μs on the Blues
Bandwidth - 2 B/F for ASCI Red, ~0.04 B/F BM, ~0.14 B/F

Parallel Efficiency dominates overall machine efficiency for large problems - 1000s of processors

Red Storm Architecture
at a
Sustained 100 TeraFlops

A Sustained 100 TeraFlops

Linpack at 100 TF - A peak of 150 TF should be enough

Production Code Experience

- **10 - 30% on ASCI Red for a number of Apps**
- **Use 20% as a goal - Implies 500 TF peak to sustain 100 TF**

Design Goals

Architecture - Distributed Memory MIMD MPP, 3-D Mesh

Balanced System Performance - CPU, Memory, Interconnect, and I/O.

Usability - Functionality of hardware and software meets needs of users for Massively Parallel Computing.

Scalability - System Hardware and Software scale, single cabinet system to 65K processor system.

Reliability - 50 hrs MTBI for Applications and 100 hrs for system

Space, Power, Cooling - High density, relatively low power system.

Price/Performance - Excellent performance per dollar, use high volume commodity parts where feasible.

Design Parameters

Time Frame - 2007

Hardware

500.0 TF peak performance
~12,500 nodes / 25,000 processors
~250 TB of memory
~10 PB of disk storage
0.5 TB/s sustained disk bandwidth

System Software

Partitioned OS - LWK for Compute nodes, full Linux for Service and I/O nodes, and streamlined Linux for RAS nodes
Scalable tools and run-time software

Programming Model - Explicit message passing

Red Storm PetaFlop Design Parameters

Topology - 33 X 16 X 24 compute nodes and 2 X 8 X 24 service and I/O nodes (x, y, z).

132 compute node cabinets, 12,672 nodes (25,344 processors)

8 service and I/O node cabinets with 384 service and I/O nodes (768 processors)

Functional Hardware Partitioning - Service and I/O nodes, Compute nodes, and RAS nodes

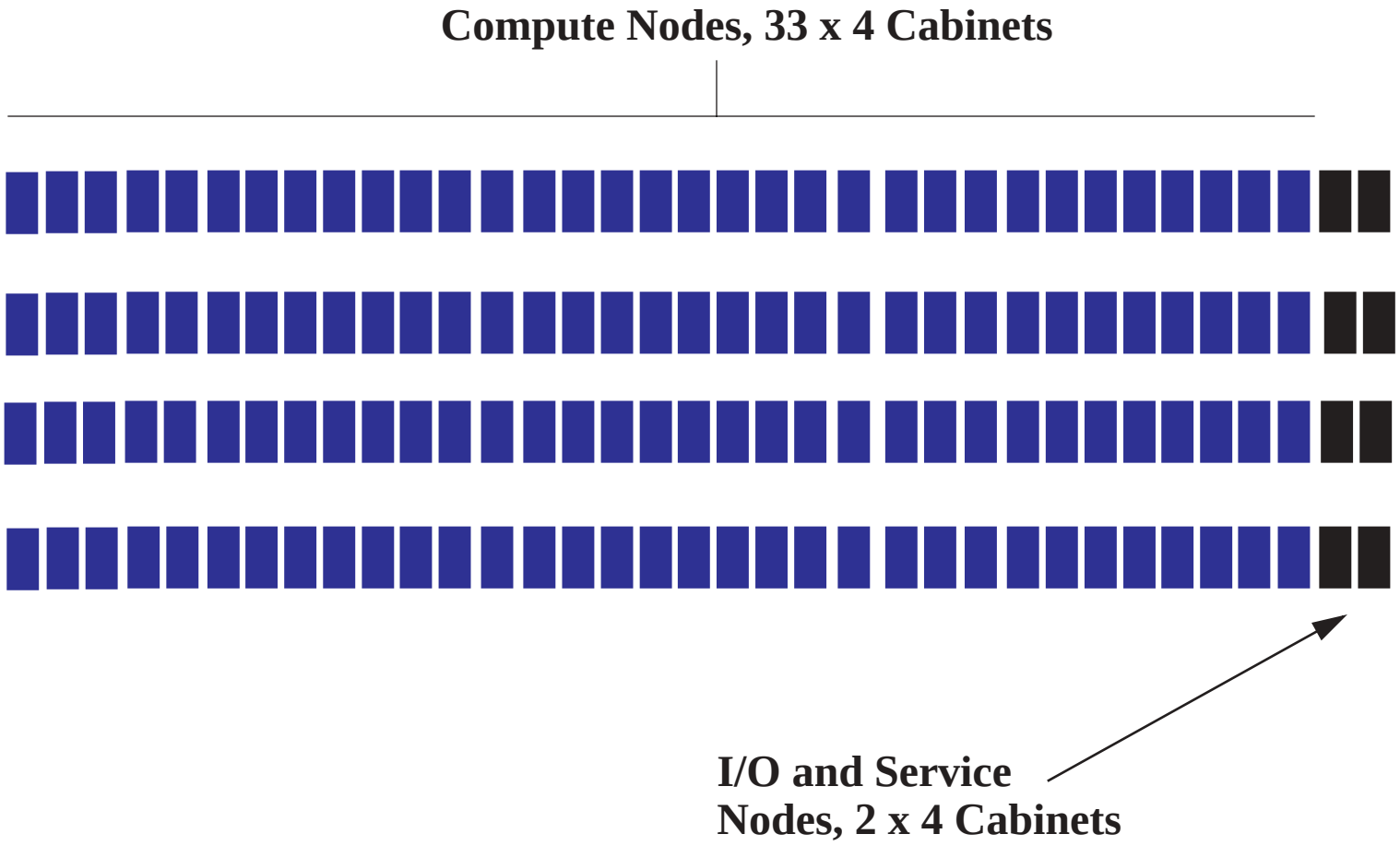
Functional System Software Partitioning - Linux for the Service and I/O nodes, LWK for the compute nodes, and real-time for the RAS nodes.

Power and Cooling - 3-4 MW

Space - ~3000 ft^2

Layout

(33 X 32 X 24 mesh)



Processor and Node Architecture and Performance

20 Gflop per processor

- Commodity micro-processor
- On Chip 2 level cache
- 5 GHz clock rate
- 4 pipelined floating point units per processor

Single Chip, 2 Processor SMP Node

- 40 GFlops per node
- 80 GB/s bandwidth to backplane

Memory System

- Memory controller integrated in processor
- 20 GB of memory per node
- Latency ~250 Cpu clocks (50 nano-seconds)
- Bandwidth of 80 GB/s per node (160 GB/s at 4 B/F)

Nodes per Board - 4 for Compute, 2 for Service and I/O

Interconnect Architecture and Performance

Topology

3-D Mesh - Highly scalable, matches codes, simple cabling
Not a Torus - Requires longer cables and more complex cabling

Performance

MPI Message passing latency - < 500 ns for nearest neighbor
Link bandwidth - 80 GB/s (40 GB/s each direction) per node
Bi-section bandwidth - 30.7 TB/s (Y-Z), 42.2 TB/s (X-Y), 63.4 TB/s (X-Z)

RAS System

Nearly Separate Parallel Computer System

RAS Workstations

- **Separate and redundant RAS workstations**
- **System administration and monitoring interface.**
- **Error logging and monitoring for all major system components including processors, memory, NIC/Router, power supplies, fans, disk controllers, RAS nework, and disk system.**

RAS Network - Dedicated Ethernet network for connecting RAS nodes to RAS workstations.

RAS Nodes - One for each card cage

System Software

Operating Systems

- Compute nodes - LWK (Catamount)
- Service and I/O nodes - Linux
- RAS nodes - Linux
- Single System View

Compilers - Fortran, C, C++

Interactive Parallel Debugger

Performance Monitor - Operations, Interconnect, Memory System

Libraries - MPI, Math, I/O

Comparison of Red Storm and PetaFlops

	ASCI Red	Red Storm	100 TF Sustained (500 TF Peak)
Full System Operational Time Frame	June 1997	August 2004	2007
Theoretical Peak (TF)	3.15	41.47	506.8
MP-Linpack Performance (TF)	2.38	~30	~350
Architecture	Distributed Memory MIMD	Distributed Memory MIMD	Distribute Mem-ory MIMD
Number of Compute Nodes / Pro-cessors	4730 / 9460	10368 / 10368	12672 / 25344
Processor	333 MHz PII (200 MHz PPro)	AMD Opteron @ 2.0 Ghz	?
Total Memory	~1.2 TB	10.4 TB (up to 80 TB)	~250 TB
System Memory B/W	2.53 TB/s	55 TB/s	1000 - 2000 TB/s
Disk Storage	13 TB	240 TB	10,000 TB

	ASCI Red	Red Storm	100 TF Sustained (500 TF Peak)
Parallel File System B/W	1.0 GB/s per color	50.0 GB/s each color	0.5 TB/s
External Network B/W	0.2 GB/s	25 GB/s each color	250 GB/s
Interconnect Topology	3-D Mesh 38 X 32 X 2	3-D Mesh (x, y, z) 27 X 16 X 24	3-D Mesh (x, y, z) 33 X 16 X 24
Interconnect Performance MPI Latency Bi-Directional Link B/W Minimum Bi-section B/W	15 μs 1 hop, 20 μs max 800 MB/s 51.2 GB/s	2.0 μs 1 hop, 5 μs max 7.6 GB/s 2.9 TB/s	~0.5 μs 1 hop, 2.0 μs max 80 GB/s 61.4 TB/s
Full System RAS RAS Network RAS Processors	10 Mb Ethernet 1 for each card cage	100 Mb Ethernet 1 for each 4 CPUs	1 Gbit Ethernet 1 for each card cage

	ASCI Red	Red Storm	100 TF Sustained (500 TF Peak)
Operating System Compute Nodes Service and I/O Nodes RAS Nodes	Cougar TOS (OSF1) VX-Work	Catamount (Cougar) Linux Linux	LWK Linux Linux
Red Black Switching	2260 - 4940 - 2260	2688 - 4992 - 2688	
System Foot Print	~2500 ft^2	~ 3000 sq ft	~3000 sq ft
Power and Cooling Requirement	850 KW	2.0 MW	3 - 4 MW

Expected Application Performance

Parallelism

- ~2.5 times the number of processors as in ASCI Red or Red Storm
- Interconnect latency is increasing relative to CPU clock speed
- Interconnect bandwidth scaling with processor speed and balance is comparable to Red Storm and a little better than for ASCI Red
- Manageable increase in the level of parallelism required for efficient use of machine

Node Performance

- Memory bandwidth balance as good or better than current machines
- Memory latency is increasing in terms of CPU clocks but decreasing in absolute time

Overall application code scaling should be similar to ASCI Red and our expectations for Red Storm

Final Thoughts

This machine can be built in the 2007 time-frame.

The current programming model is unlikely to change from explicit message passing. There is a very large investment in the current message passing codes.