A Conservative Path to PetaFlop Computing
The Red Storm Architecture
Scaled to a PetaFlop and Beyond

Jim Tomkins

The 4th Annual Workshop on Linux Clusters for Supercomputing

National Supercomputer Center (NSC)

Linkoping University, Sweden

October 22 - 24, 2003

Outline

Application Code Characteristics

PetaFlop Assumptions

Key Issues in Scaling to a PetaFlop and Beyond

- **Application Code Scaling**
- **Single Processor Performance**
- **Parallel Efficiency - Communication Overhead**
- **Parallel I/O**
- **Power, Cooling, and Packaging**
- **Reliability**
- **Price/Performance**

The Red Storm Architecture scaled to a Petaflop

- **Design Goals**
- **PetaFlop Point Design**

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

Parallel efficiency dominates application code performance when thousands of processors are involved.

PetaFlop Assumptions

Time Frame - 2010

Moore's Law Continues

Device density doubles every 2 years - Factor of 8 by 2010
Line width <50 nano-meters in 2010

Peak Processor Performance Continues to Increase

Clock rate increases to 10 - 12 GHz by 2010
4 or more floating point operations per processor clock

Processor ASICs will have at least 4 (maybe more) cores per chip in 2010.

Power

Processor ASIC will be <150 W
Memory chips will be < 1 W each at 4 Gbit per chip

Application Code Scaling Issues

Applications Scaling - Codes need to scale to 25000 or more processors at 50% or greater parallel efficiency.

Serial Fraction - at 25000 processors the serial fraction must be less than 1/25000 to achieve 50% parallel efficiency.

Load Imbalance

- **Many processors waiting on one or a few**
- **A factor of 2 imbalance - as much as a 50% loss in parallel efficiency.**

Communication Overhead

- **Tends to grow with the number of processors involved**
- **50% overhead - as much as a 50% loss in parallel efficiency**

Single Node Performance Issues

Major Issues are memory latency and bandwidth.

Goal - at least maintain current memory bandwidth (B/F ratio)

- **AMD Opteron at 2 GHz - 5.3 GB/s (1.33 B/F)**
- **Prefer B/F ratio of 4 - Dependent on commodity path**
- **Bandwidth per pin must increase significantly - > 20 Gbits/s per pin**

Goal - Improve memory latency in absolute terms

- **AMD Opteron is ~80 ns for page miss**
- **Should get to at least 40 ns by 2010 - 3X Cpu clocks**

Operating System Impacts

- **Page Size - Minimize page misses**
- **Contiguous Memory**

Applications will have to become more latency tolerant as processors clock rates continue to increase.

Question - Is 64 bit arithmetic sufficient?

Parallel Efficiency - Communication Overhead Issues

Communication Bandwidth Needed to get Scalability

Peak Bandwidth > 2 B/F per link
Sustained Bandwidth (includes system software overhead) - >1.5 B/F
Minimum Bi-section Bandwidth

Communication Latency

Message startup time
Needs to be as low as possible - Speed of light is an issue
Software protocols add overhead
How to get MPI < 200 ns

Operating System Overhead

Impact of OS Daemons
Use LWK to minimize OS Impacts

Parallel I/O Issues

Sustained Read/Write Performance

1 GB/s per TFlop - For a PetaFlop this is 1 TB/s
Single file - 25000 processors to one file
25000 or more files - one or more files per processor
Mixed single file and many files simultaneously

Meta Data Service

Accesses per second - Need to handle at least 25000 requests
Multiple Servers
Data Integrity

Current File System Development

Lustre
Panassas
PVFS2

Power, Cooling, and Packaging

Power and Cooling

Air Cooled
Vertical Air Flow - bottom to top
~20 KW per cabinet

Packaging - Space

1 U Blades
2 Processor Chips and Memory per Blade
Backplane Based Interconnect

Big Issue - How much power will processors and memory chips require in 2010.

Reliability Issues

Part Count - PetaFlop System will have a large number of Parts. However, the total will be less than Red Storm.

- **Level of integration decreases number of parts**
- **Parts are more complicated**
- **Lower voltages make soft errors more likely**
- **Higher power density impacts reliability**

Availability - Not a Key indicator of Reliability

Mean Time Between Interrupt (MTBI) is Key Indicator

- **Application code interrupt - 50 hrs**
- **System interrupt - 100 hrs**

How to get Reliability

- **Build in redundancy and error correction - Performance**
- **System monitoring**
- **Keep it simple - Hardware and System Software**

Price/Performance

Performance on Real Applications

- Scalability - Dominates overall system performance
- Reliability - Cost of repeating work

Use Commodity Parts where Feasible

Total Costs are Important

- System Costs
- Operations, Support, and Maintenance
- Power and Cooling
- Building Space

Red Storm Architecture Scaled to a PetaFlop

Design Goals

Operational in 2010

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 <u>Massively Parallel Computing</u>.

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.

Red Storm PetaFlop Design Parameters

Time Frame - ~2010

Hardware

- **1.0 Petaflop peak performance**
- **~25 thousand processors**
- **~500 TB of memory**
- **~20 PB of disk storage**
- **1.0 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 32 X 24 compute processors and 2 X 16 X 24 service and I/O processors (x, y, z). (33 X 8 X 24 compute nodes and 2 X 4 X 24 service and I/O nodes.)

132 compute node cabinets with 6336 compute nodes (25344 processors)

8 service and I/O node cabinets with 192 service and I/O nodes (768 processors)

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

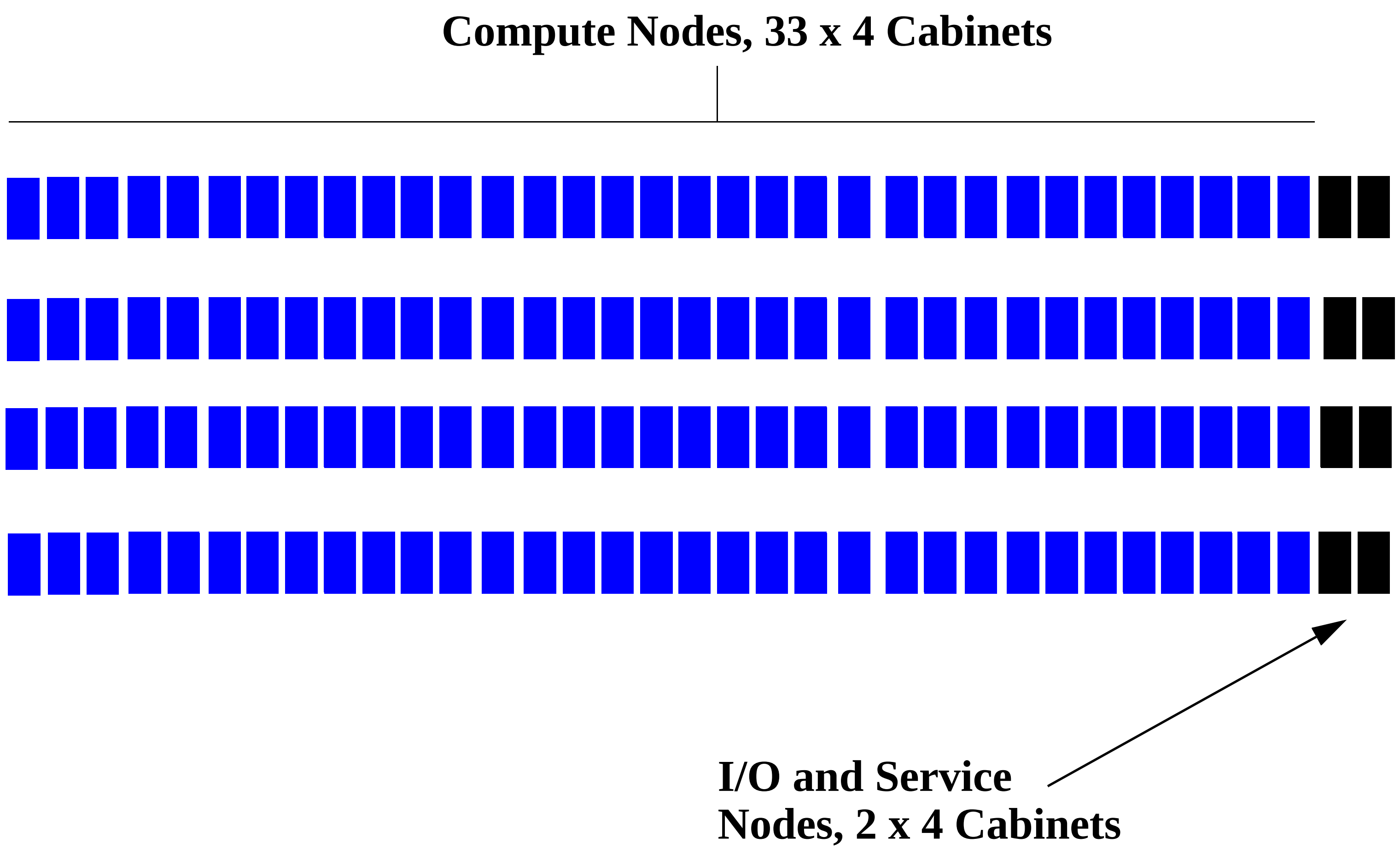
Functional System Software Partitioning - Linux for the Service and I/O nodes, LWK for the compute nodes, and real-time for the RAS nodes.

Power and Cooling - 3-4 MW

Space - ~300 m^2

Red Storm Layout

(33 X 32 X 24 mesh)



Processor and Node Architecture and Performance

40 Gflop per processor

- Commodity micro-processor
- Large 2 level cache
- 10 GHz clock rate
- 4 pipelined floating point units per processor

Single Chip, 4 Processor SMP Node

- 160 GFlop per node
- 320 GB/s bandwidth to backplane

Memory System

- Memory controller integrated in processor
- 96 GB of memory per node
- Latency ~500 Cpu clocks (40 nano-seconds)
- Bandwidth of 213 GB/s per node (640 GB/s at 4 B/F)

Nodes per Board - 2 for Compute, 1 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

Message passing latency - < 500 ns
Link bandwidth - 80 GB/s (40 GB/s each direction) per processor
Bi-section bandwidth - 61.5 TB/s

Red Storm RAS System

Nearly Separate Parallel Computer System

RAS Workstations

- Separate and redundant RAS workstations for Red and Black ends of machine.
- System administration and monitoring interface.
- Error logging and monitoring for major system components including processors, memory, NIC/Router, power supplies, fans, disk controllers, and disks.

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

RAS Nodes - One for each card cage

PetaFlop 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

Libraries - MPI, Math, I/O

Comparison of Red Storm and PetaFlops

	Red Storm	PetaFlop
Full System Operational Time Frame	August 2004	2010
Theoretical Peak (TF)	41.47	1013.8
MP-Limped Performance (TF)	~30 (est)	~700
Architecture	Distributed Memory MIMD	Distribute Memory MIMD
Number of Compute Node Proces-sors	10368	25344
Processor	AMD Patroon @ 2.0 Gaze	?
Total Memory	10.4 TB (up to 80 TB)	~600 TB
System Memory B/W	55 TB/s	2000 - 4000 TB/s
Disk Storage	240 TB	20000 TB
Parallel File System B/W	50.0 GB/s each color	1 TB/s
External Network B/W	25 GB/s each color	500 GB/s

	Red Storm	PetaFlop
Interconnect Topology	3-D Mesh (x, y, z) 27 X 16 X 24	3-D Mesh (x, y, z) 33 X 32 X 24
Interconnect Performance MPI Latency I-Directional Link B/W Minimum I-section B/W	 2.0 µs 1 hop, 5 µs max 6.0 GB/s 2.3 TB/s	 ~0.2 µs 1 hop, 1.5 µs max 80 GB/s 61.4 TB/s
Full System RAS RAS Network RAS Processors	 100 Bit Ethernet 1 for each 4 CPOS	 1 Bit Ethernet 1 for each card cage
Operating System Compute Nodes Service and I/O Nodes RAS Nodes	 Catamount (Cougar) Linux Linux	 LWK Linux Linux
Red Black Switching	2688 - 4992 - 2688	
System Foot Print	~ 3000 sq ft	~3000 sq ft
Power and Cooling Requirement	2.0 MW	3 - 4 MW

Expected Application Performance

Parallelism

- ~2.5 times the number of processors
- Interconnect latency is increasing relative to CPU clock speed
- Interconnect bandwidth scaling with processor speed and balance is comparable to Red Storm
- 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 Red Storm

Final Thoughts

A balanced Petaflop computer system can be built by the end of the decade without a major change in the technology path we are currently on. It is only necessary to continue with Moore's Law scaling.

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

Any proposed new architecture that requires a programming model change and doesn't provide an easy migration path for the current application codes will have a very tough time in the market place.