Deploying the Sun Magnum System: The Beginning of NSF Petascale Computing

Jay Boisseau, Director
Texas Advanced Computing Center
The University of Texas at Austin

October 24, 2006

First, Some “Corrections”

- There are **no ClearSpeed** (or any other) accelerators in this system!

First, Some “Corrections”

- There are **no ClearSpeed** (or any other) accelerators in this system
- This is a **capability** system, not only a capacity system: balanced & tightly coupled

First, Some “Corrections”

- There are **no ClearSpeed** (or any other) accelerators in this system
- This is a **capability** system, not only a capacity system: balanced & tightly coupled
- Jobs will run on the **entire system**; it will not be partitioned into smaller systems

First, Some “Corrections”

- There are **no ClearSpeed** (or any other) accelerators in this system
- This is a **capability** system, not only a capacity system: balanced & tightly coupled
- Jobs will run on the **entire system**; it will not be partitioned into smaller systems
- There was **no unique deal** from AMD (but we’ll take one if they offer!)

And Some Presentation Caveats

- The system does not exist yet
  - It is not yet doing science or even drawing power!
  - Key components are to be delivered in 2007
- Some system details are still non-disclosure
  - Ask Giri Chukkapalli
- Our experience with 50K general-purpose CPUs is zero; even with 5K, it’s only 3 weeks
  - This is new territory—none of us are experts *yet*

NSF HPC Vision & Strategy

- Context: NSF Cyberinfrastructure Strategic Plan
- NSF now investing in world-class HPC systems
  - Annual track 2 solicitations ($30M)
  - Single five-year Track1 solicitation ($200M)
- Complementary solicitations forthcoming
  - petascale applications development solicitations
  - Software Development for CI has an HPC component
  - Etc.

Sun System Configuration

[Some data non-disclosure]

- Compute power
  - 13152 Opteron “Deerhound” processors
    - Quad-core, four flops/cycle (dual pipelines)
    - Initial deployment with SantaRosa processors
  - 421 teraflops aggregate peak (at least)
- Memory
  - 2GB/core
  - 105 TB total

Sun System Configuration

[Some data non-disclosure]

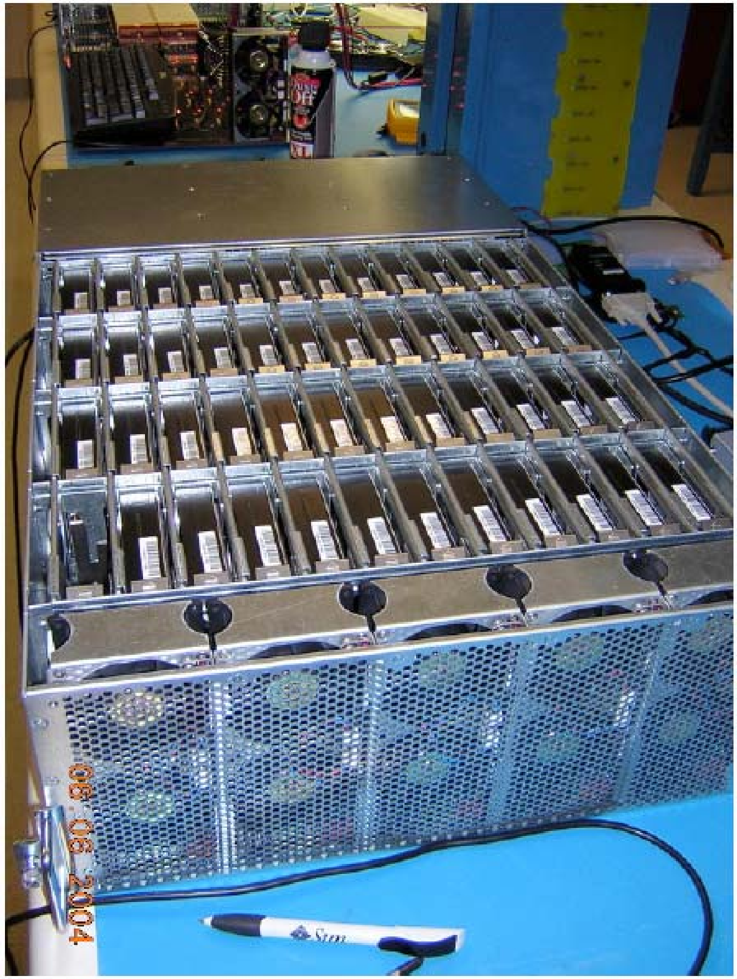
- Interconnect
  - – Sun proprietary switch based on IB
    - Minimum cabling: robustness and simplicity!
  - – PathScale adapters (PCI-Express)
  - – MPI latency: 1.6-1.8 microsec
  - – Peak bi-directional b/w: ~ 1 GB/sec
  - – Total backplane b/w: 13.8 TB/sec

Sun System Configuration

[Some data non-disclosure]

- File system
  - 72 Sun X4500s (“Thumper”)
    - 48 500GB disks per 4U!
  - 1.7 PB total disk
    - 1 PB in largest /work file system
  - Lustre file system
  - Aggregate b/w: 40 GB/s

Thumper Photos

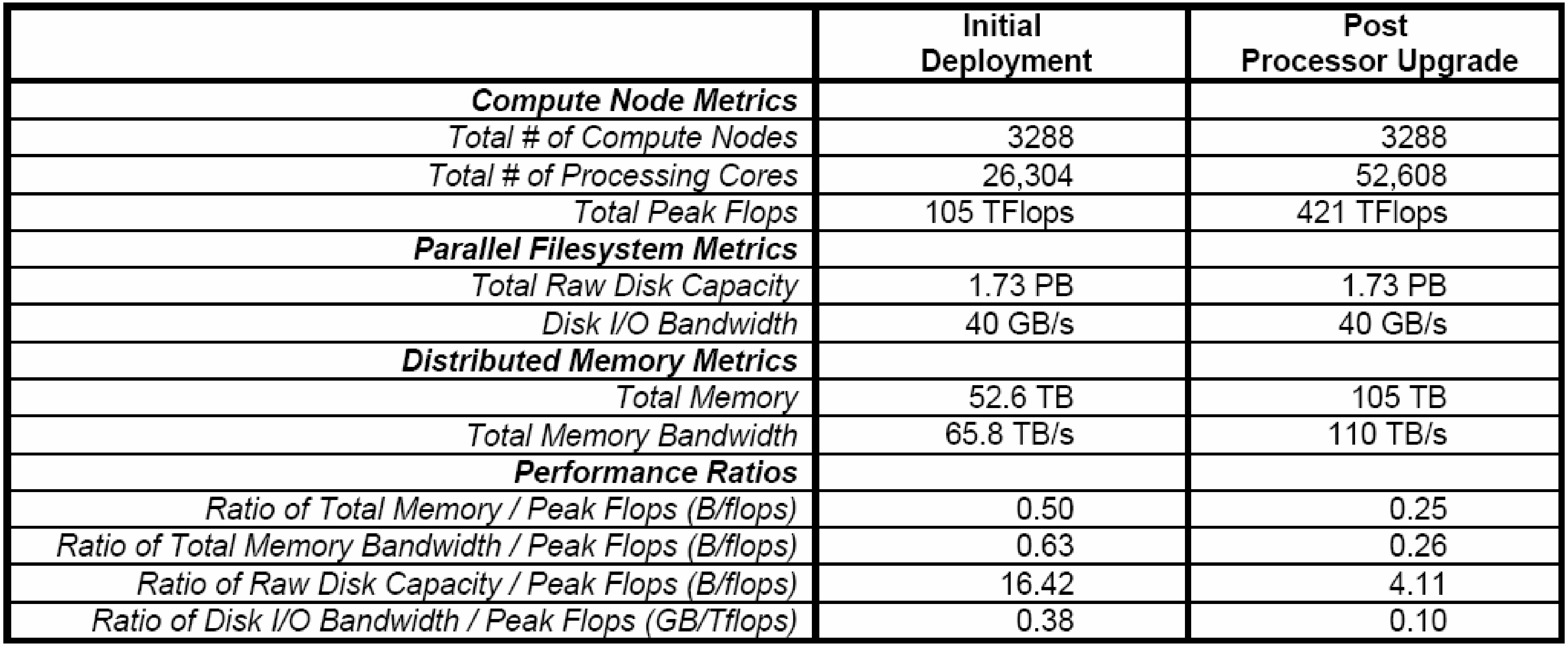


Sun System Configuration

[Some data non-disclosure]

- System Management
  - ROCKS (customized) Cluster Kit
    - *perfctr* patch, etc.
  - Sun N1SM for lights-out management
  - Sun N1GE for job submission
    - Backfill, fairshare, reservations, etc.

Speeds & Feeds



	Initial Deployment	Post Processor Upgrade
Compute Node Metrics		
Total # of Compute Nodes	3288	3288
Total # of Processing Cores	26,304	52,608
Total Peak Flops	105 TFlops	421 TFlops
Parallel Filesystem Metrics		
Total Raw Disk Capacity	1.73 PB	1.73 PB
Disk I/O Bandwidth	40 GB/s	40 GB/s
Distributed Memory Metrics		
Total Memory	52.6 TB	105 TB
Total Memory Bandwidth	65.8 TB/s	110 TB/s
Performance Ratios		
Ratio of Total Memory / Peak Flops (B/flops)	0.50	0.25
Ratio of Total Memory Bandwidth / Peak Flops (B/flops)	0.63	0.26
Ratio of Raw Disk Capacity / Peak Flops (B/flops)	16.42	4.11
Ratio of Disk I/O Bandwidth / Peak Flops (GB/Tflops)	0.38	0.10

Space & Watts

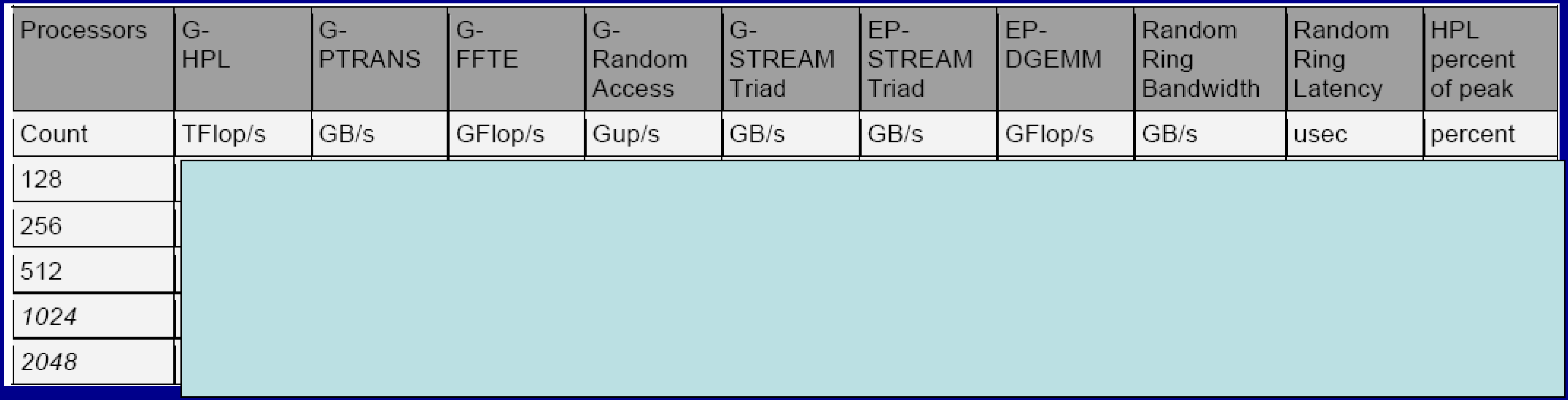
- System power: 2.162 MW
- System space
  - ~80 racks
  - ~1500 sqft for system racks and in-row cooling equipment
  - ~3000 sqft total
- Cooling:
  - In-row units and chillers
  - ~0.6 MW
- Observations:
  - space less an issue than power
  - power distribution less an issue than generation!

Applications Performance Notes

- Obviously, no data for final system
  - Switch doesn’t exist yet
  - Processors don’t exist yet
- Performance predictions can be made from previous & pre-production versions
- Applications performance projections for NSF benchmarks look very promising (MPI only)

Applications Performance Notes

- *Hope to be able to reveal projections at SC06*



Processors	G-HPL	G-PTRANS	G-FFTE	G-Random Access	G-STREAM Triad	EP-STREAM Triad	EP-DGEMM	Random Ring Bandwidth	Random Ring Latency	HPL percent of peak
Count	TFlop/s	GB/s	GFlop/s	Gup/s	GB/s	GB/s	GFlop/s	GB/s	usec	percent
128										
256										
512										
1024										
2048										

- Applications: *WRF, OOCORE, MILC, GAMESS, HOMME…*

User Support Challenges

- NO systems like this exist yet!
  - Will be the first general-purpose system at ½ Pflop
  - Quad-core, massive memory/disk, etc.
- New opportunities, new apps challenges
  - Multi-core optimization
  - Extreme scalability
  - Fault tolerance in apps
  - Petascale data analysis
- Initially, the only such NSF system: demand?

User Support Plans

- User support: the “usual” +
  - User Committee dedicated to this system
  - Applications Engineering
    - algorithmic consulting
    - technology selection
    - performance/scalability optimization
    - data analysis
  - Applications Collaborations
    - Partnership with petascale apps developers and software developers

User Support Plans

- Also
  - Strong support of ‘professionally optimized’ software
    - Community apps
    - Frameworks
    - Libraries
  - *Extensive* Training
    - On-site at TACC, partners, and major user sites, and at workshops/conferences
    - Advanced topics in multi-core, scalability, etc
    - Virtual workshops
  - Increased contact with users in TACC User Group

Technology Insertion

- Again, NO systems like this exist yet!
  - Workshops like this are excellent to start thinking, planning
- System will stimulate new R&D
- System will operate for four years
- Technology identification, tracking, evaluation and insertion will be important!
- Chief Technologist will work with team to:
  - identify new apps, libs, tools, etc.
  - improve perf, ease-of-use, reliability, security

Access & Allocations

- System is primarily NSF funded
  - 90% allocable via the TeraGrid
- System hosted by UT, supported by TX $:
  - 5% for Texas institutions, from R1s to JuCos
    - Excellent EOT opportunities
- System will foster academic/industry collaborations and tech transfer
  - 5% for industrial partners
    - Work with Council on Competitiveness
    - Learn from INCITE

Impact in TeraGrid

- 400M CPU hours to TeraGrid: more than double current total of all TG HPC systems
- 421 Tflops peak
  - 3x total perf of all TG HPC systems
  - 10x top TG HPC system in perf, memory, disk
- Reestablish NSF as a leader in HPC
- Jumpstarts progress to petascale for entire US academic research community

Practice Makes Perfect

- **Lonestar** is the fastest US academic supercomputer in operation
  - – 1300 Dell PowerEdge 1955 blade servers
  - – 2600 Intel Xeon dual-core processors

    5200 cores at 2.66 GHz each
  - – Cisco InfiniBand interconnect

    10 gigabit/sec bandwidth, < 5 microsec latency

THE UNIVERSITY OF

Texas Advanced Computing

Project Timeline

Sep06	award, press, relief, beers
Jan07	equipment begins arriving
Mar07	facilities upgrades complete
May07	allocations requests due (TeraGrid)
May07	very friendly users on <100 Tflops system (dual-core, 2 flop/cycle Santa Rosa procs)
Jun07	friendly users on 100 Tflops system
Jul07	full operations, relief, beers
3Q07	processor upgrade to Deerhound (quad-core, 4 flops/cycle)

Team Partners & Roles

- TACC / UT Austin: project leadership, system hosting & ops, user support, apps collaborations, tech evaluation & dev
- ICES / UT Austin: apps collaborations, algorithm/technique dev & transfer
- Cornell Theory Center: large-scale data management & analysis, user support
- Arizona State HPCI: tech evaluation & dev, user support

Team Partners & Roles

- Project Director: Jay Boisseau (TACC)
- Project Manager: Tommy Minyard (TACC)
- Chief Applications Scientists: Omar Ghattas (ICES / UT Austin), Giri Chukkapalli (Sun)
- Chief Technologists: Jim Browne (ICES / UT Austin)
- Many other TACC, ICES, CTC, ASU staff playing important roles (~25 FTEs)
- Strengthening relationships with other petascale centers

Summary

- Track2 Sun system will be one of most powerful general-purpose open computing system in the world in Oct07
- System will be *balanced, capability* system for scalable numerically- and data-intensive apps
- System will present tremendous opps for applications developers, and s/w developers
- Allocations 1 July 2007, apply by 1 May 2007

Ideas, Suggestions, and Users Welcome!

Jay Boisseau: boisseau@tacc.utexas.edu

Tommy Minyard: minyard@tacc.utexas.edu

Giri Chukkappalli: giridhar.chukkapalli@sun.com