Pathway to Petaflops
A vendor contribution

Philippe Trautmann
Business Development Manager HPC & Grid
Global Education, Government & Healthcare

A Petaflop is :

1.000.000.000.000.000 (10^{15}) floating point operations per second!

Is there a limit to compute power requirements ?

My own take is that the first PetaFlop system will be installed in 2007, latest 2008

Sun HPC Vision

To Create the Best Standards based Technologies, Products and Services Needed For High Performance Computing

Major Technology Trends

1. Multicore CPUs are driving larger node sizes in terms of memory, I/O and power per node
2. Virtualization is increasing application footprint and therefore memory per node
3. InfiniBand and 10 GigE are offering low latency and increasingly cost-effective Fabrics
4. Power and cooling bills are sometimes more expensive than the system purchase cost
5. User/Researcher efficiency requires new programming tools and models

Average compute node size is increasing in CPU cores, memory, I/O and power

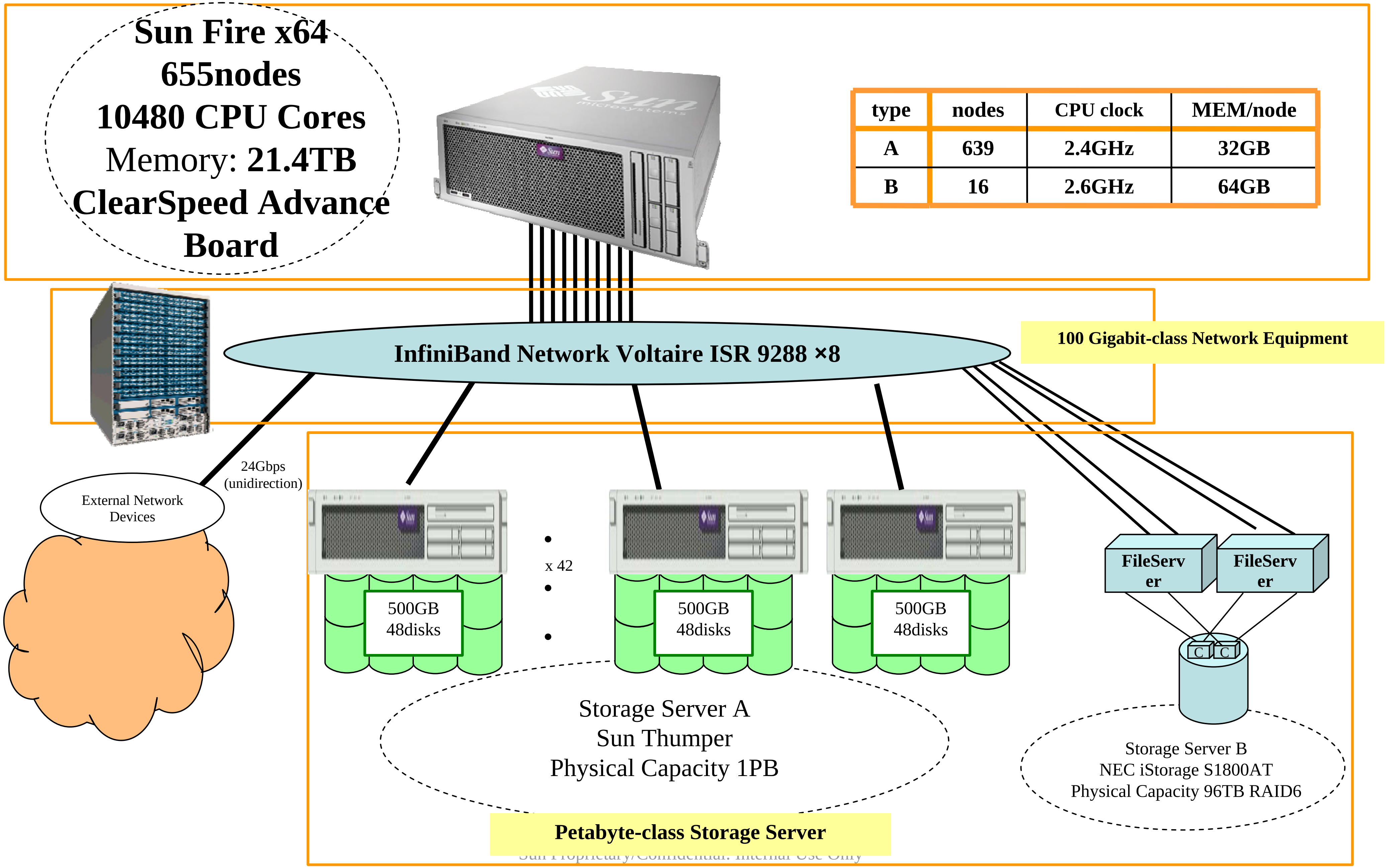
Top500 HPC Trends

- Clusters: > 72.8% of HPC Architecture
- Standards Based Interconnects: InfiniBand and Gigabit Ethernet
  - > 80% of HPC Clusters
- Storage Capacity and Bandwidth
  - Multi-TeraBytes to Multi-PetaBytes
  - Parallel Filesystem
- Compute Nodes
  - Multi-Core CPUs
  - Moving toward Clusters of SMPs

Tokyo Institute of Technology

- World’s Largest x64 Cluster
- Over 10,000 x64 processor cores, 20 TB main memory
- Over 1 PB Sun Storage with Lustre Parallel File System
- World’s Fastest Infiniband Network from Voltaire
- 6 weeks to deploy!
- Using N1SM and N1GE
- 38.16TFlops (Top500 #7), and expect further improvement at SC06

Titech Top Level System Diagram



Sun Fire x64
655nodes
10480 CPU Cores
Memory: 21.4TB
ClearSpeed Advance Board

type	nodes	CPU clock	MEM/node
A	639	2.4GHz	32GB
B	16	2.6GHz	64GB

InfiniBand Network Voltaire ISR 9288 ×8

100 Gigabit-class Network Equipment

24Gbps (unidirection)

External Network Devices

500GB 48disks

x 42

500GB 48disks

500GB 48disks

FileServer

FileServer

Storage Server A
Sun Thumper
Physical Capacity 1PB

Storage Server B
NEC iStorage S1800AT
Physical Capacity 96TB RAID6

Petabyte-class Storage Server

Sun Fire X4500 “Storver” Selected by IN2P3 to run 400TB of data for the LHC-CERN

Base Storage Configuration
48 x4500 data servers
Operating Systems: Solaris/ZFS

High Performance Computing Data Server

Compute

- 2 x Dual Core Opteron processors
- 16GB Memory

Storage

- 48 Serial ATA disks
- Up to 24TB raw capacity

I/O

- Very high throughput
- 2x PCI-X slots
- 4 GigE

Availability

- Hot-swap power, fans, disks

Management

- Same management as other Galaxy servers

Solaris(TM) ZFS

- Ground-breaking file system performance

World's Storage Leader

- 67% of mainframe attached storag
- 37% of world's data
- Delivering NAS performance
- #1 WW Unix platform storage leader by PB delivered
- 8 out of 11 LHC Tier1 sites run StorageTek backup

High Productivity Computing Systems

Providing a new generation of economically viable high productivity computing systems for the national security and industrial user community (2009-2010)

DARPA's “Ultrascale” Goals

- ✓ 2+ PetaFlops on Linpack Top500 benchmark
  - - Solves large linear system using Gaussian Elimination
- ✓ 6.5 PetaBytes/sec data streams bandwidth
  - - Copy, add, scale large vectors locally
- ✓ 3.2 PetaBytes/sec bisection bandwidth
  - - Transpose a large matrix
- ✓ 64,000 GigaUpdate /sec
  - - Random updates to memory

**“May require up to 100,000 processor cores,
but must be easier to use than current 1,000 processor systems.”**

Sun and US Government Partnership Basic Computer Systems Research

2002	1 Year	I: Industry Concept Study	$ ~3M	IBM sgi CRAY hp Sun microsystems
	3 Years	II: Research	$ ~50M (July 2003)	IBM CRAY Sun microsystems
2010	4 Years	III: Product Development	$~250M (End 2006)	??? 1 or 2 to Be Selected

Sun's Qualification of Scientific Linux

- Scientific Linux 4.5 installed and ran successfully on Sun x64 systems
  - > X2200, 2-socket, 4-core AMD x64 systems
- Scientific Linux 4.5 installed and ran successfully on Sun x64 Data Server at CERN
  - > X4500, 2-socket, 4-core AMD x64 system with 48 SATA Disk Drives
  - > Next test will be with Solaris and ZFS
  - > http://pcitapi34.cern.ch/%7Efrederik/X4500.pdf

Sun and DESY Collaboration

- dCache
  - > Used World Wide in most of CERN Tier Sites
  - > A distributed disk caching system as a front-end for Mass Storage System
  - > Written in Java and until recently, only ran on Linux
- Sun worked with DESY to port dCache to Solaris x64 and ZFS
  - > dCache ran on a x4500, half configured with 24 SATA Drives and 8GB of memory
- With Solaris and ZFS
  - > No silent data corruption
  - > Self-healing
  - > Dynamic
  - > High Performing

Summary of Sun's Value to the HEP community

- Sun is Committed to CERN WW for Today and the Future
- We are demonstrating CERN Software stack on Sun Hardware and Solaris (Qualified Scientific Linux)
- Many of the HEP compute sites run Sun to manage Data through Sun StorageTek Technology
- Sun invests in collaborations with the community (Darpa, NSF, CERN LHC, European R&D projects, etc.)
- Collaborations allow creation of new tools and models that will allow researchers to run the PetaFlop systems

Thank you for your attention

Philippe Trautmann
Business Development Manager HPC & Grid
Global Education, Government & Healthcare