ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CSCS
Swiss National Supercomputing Centre

EUREKA - a new data analysis facility at CSCS

Lucerne, May 13th 2011

Thomas Schoenemeyer, Technology Integration, CSCS

Contents

- Motivation
- Analysis Systems
- Underlying File System

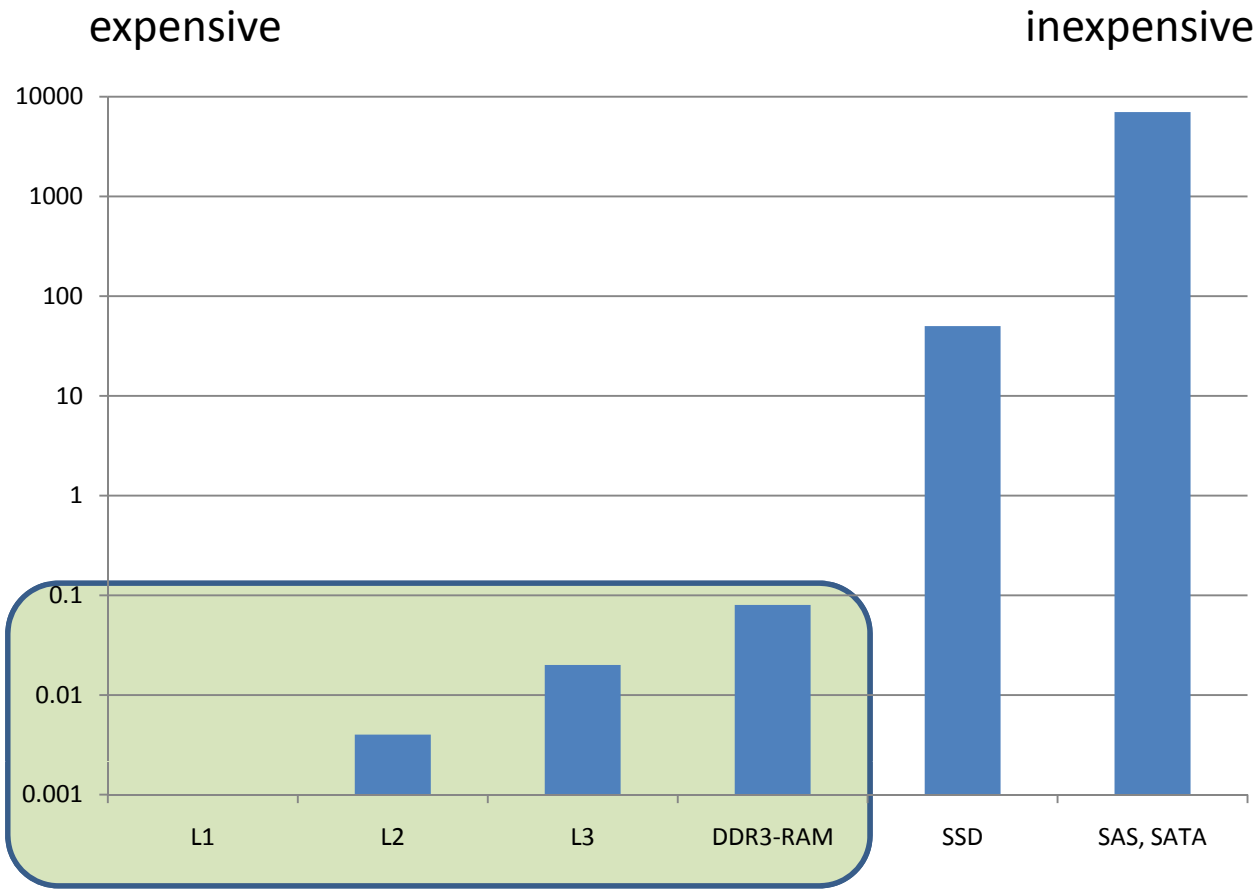
Motivation

- Massive increase of amount of scientific data through simulations and observations
- Trend continues and accelerates
- The next grand challenge is not “simple search of“ but “discovery from“ from big data through complex queries
- Finding useful information that you are not intuitively aware of will be the key of breakthroughs

Observation

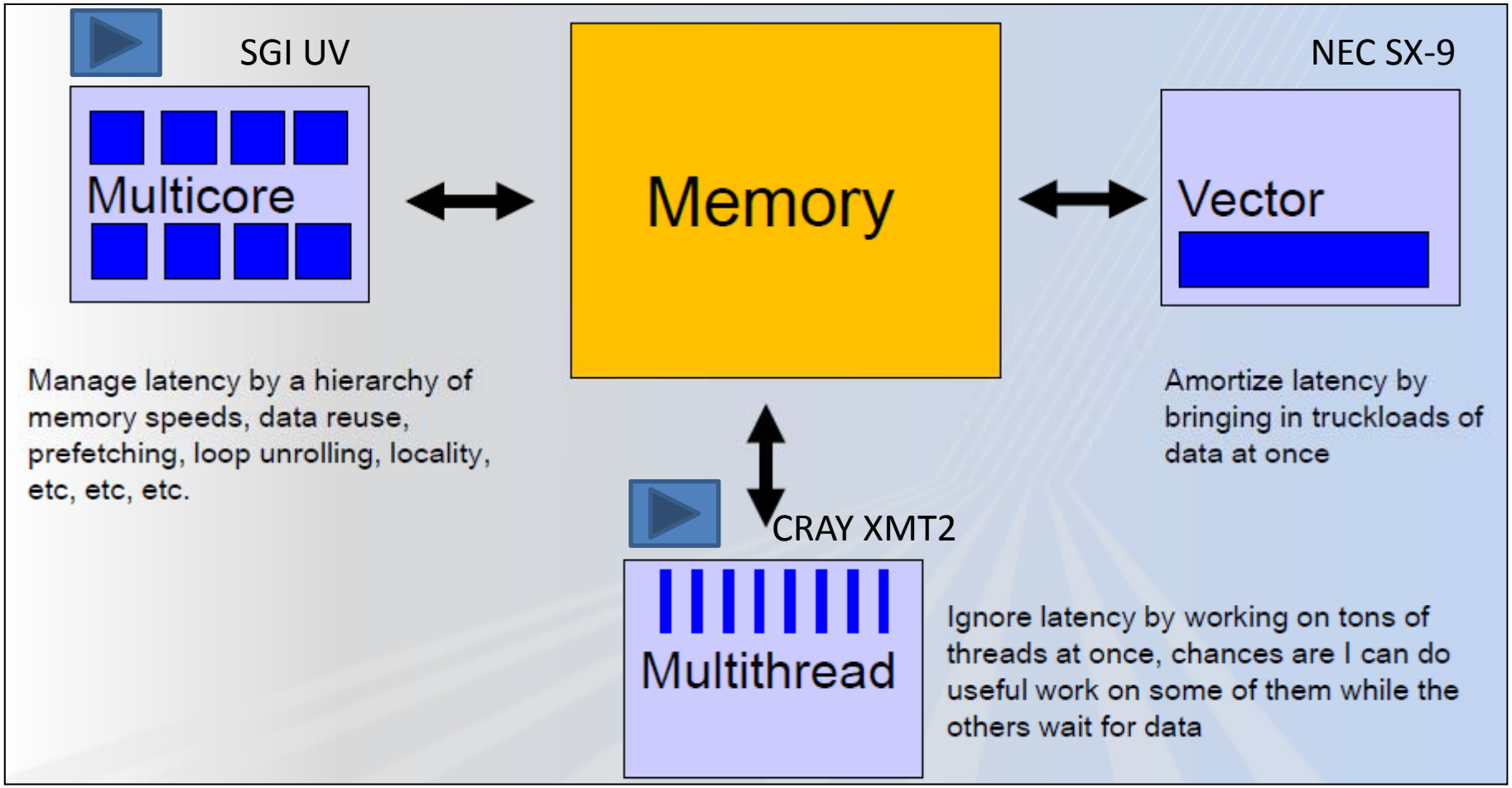
- Impressive trajectory of performance growth
- CPUs, IOPS, parallelism
- 1 CPU Power3 @ 1GF in 1999
- 1 Intel Xeon 6-core @ 70 GF in 2011
- Bandwidth and latency per disk
- Scale bandwidth, but can’t decrease latency with # of disks

Access to Memory



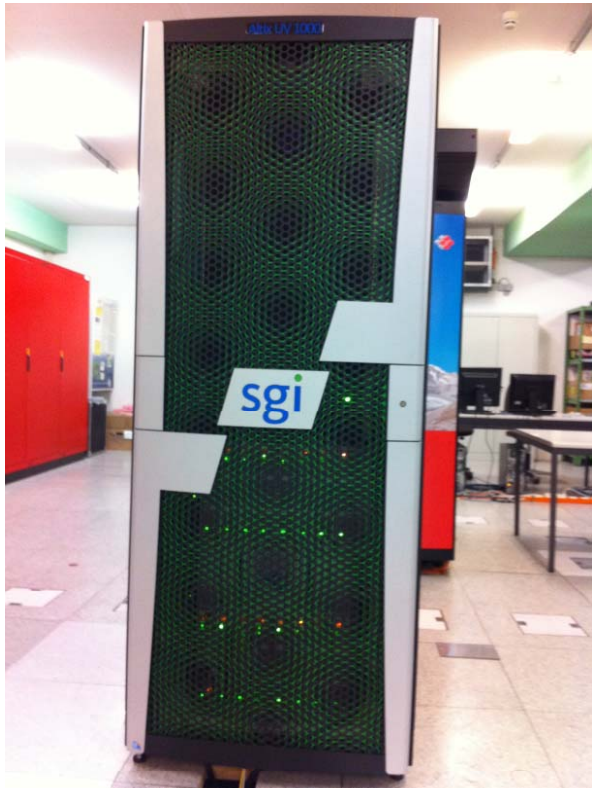
Typical latency in µs for several storage devices

Select the Systems



SGI UV-1000 (Rothorn)

- Architecture
  - 16 dual-socket Blades
  - 128 GB main memory per Blade
  - 32 Intel Xeon E7 (Westmere-EX)
- Westmere-EX
  - 8 cores
  - 2.67 GHz x 4 x 8 = 85 Gflops
  - 24 MB L3-Cache
- NUMAlink 5
  - 15GB/s x 4 = 60 GB/s per blade
  - 2 TB ccNUMA RAM
  - SLES11 and SGI software

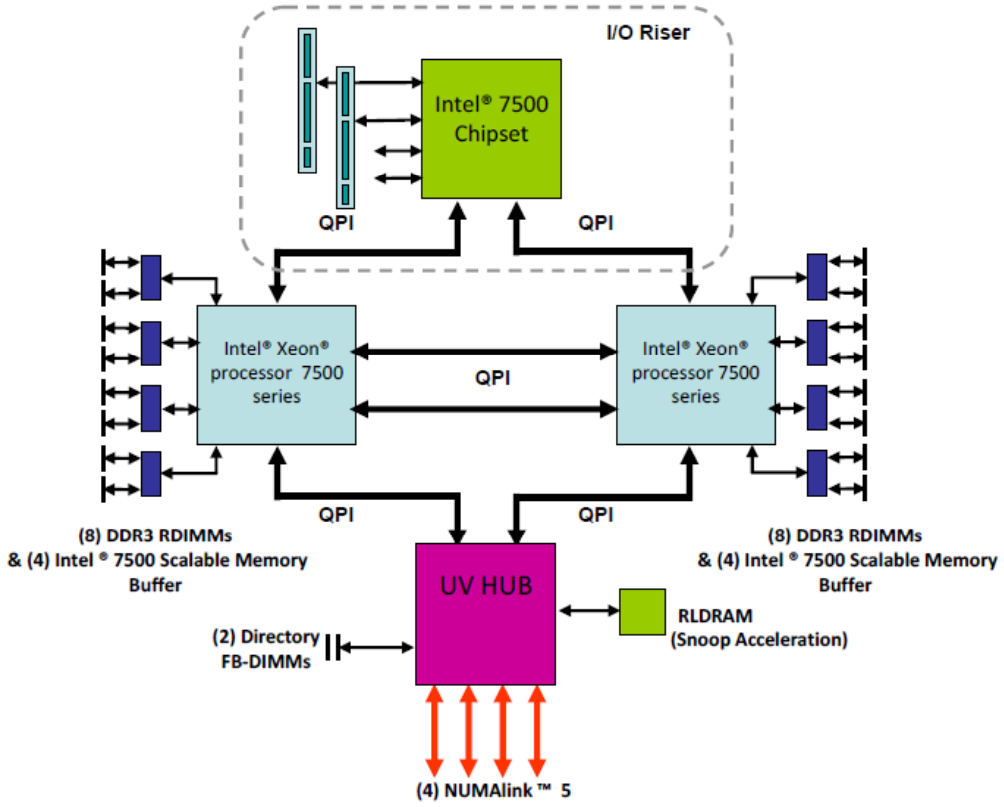


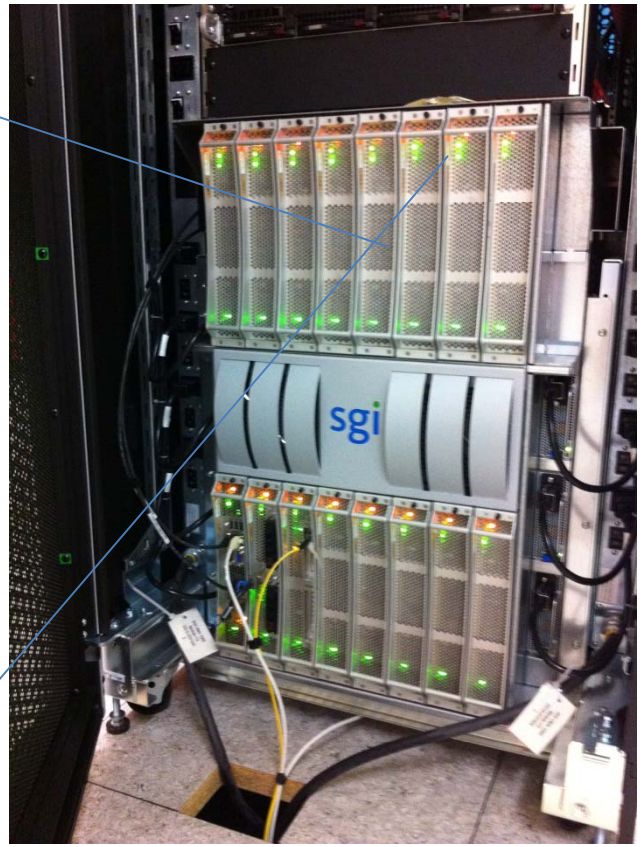
At CSCS since April 29th 2011

Westmere-EX

Processor Number	Processor Frequency	Intel Smart Cache	TDP	Intel Turbo Boost Technology; Intel Hyper-Threading Technology	Cores / Threads	1kU Price
Intel® Xeon Processor E7-8870	2.4 GHz	30M	130W	√	10 / 20	$4,616
Intel® Xeon Processor E7-8860	2.26 GHz	24M	130W	√	10/ 20	$4,061
Intel® Xeon Processor E7-8850	2 GHz	24M	130W	√	10 / 20	$3,059
Intel® Xeon Processor E7-8830	2.13 GHz	24M	105W	√	8/16	$2,280
Intel® Xeon Processor E7-8867L	2.13 GHz	30M	105 W	√	10/20	$4,172
Intel® Xeon Processor E7-8837	2.67 GHz	24M	130 W	Turbo; no HT	8	$2,280
Intel® Xeon Processor E7-4870	2.4 GHz	30M	130W	√	10 / 20	$4,394
Intel® Xeon Processor E7-4860	2.26 GHz	24M	130W	√	10/ 20	$3,838
Intel® Xeon Processor E7-4850	2 GHz	24M	130W	√	10 / 20	$2,837
Intel® Xeon Processor E7-4830	2.13 GHz	24M	105W	√	8/16	$2,059
Intel® Xeon Processor E7-4820	2 GHz	18M	105 W	√	8/16	$1,446
Intel® Xeon Processor E7-4807	1.86 GHz	18M	95 W	No turbo; HT	6/12	$890
Intel® Xeon Processor E7-2870	2.4 GHz	30M	130 W	√	10/20	$4,227

SGI UV 1000 NUMA Blade



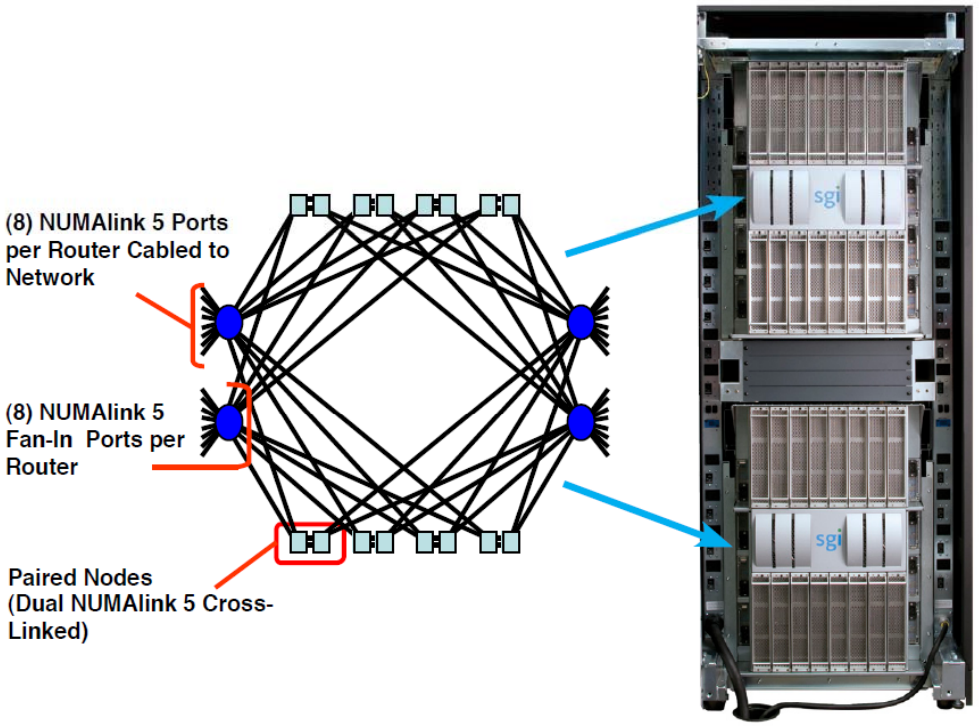


Technical Advances in the SGI® Altix® UV Architecture

Source: *www.**sgi**.com/pdfs/4192.pdf*

4 router Topology

- Each router with 16 NUMAlink ports
- 8 to the network, 8 to the nodes
- Packed into groups of 4 routers
- Linkspeed 7.5 GB/s per direction, 15 GB/s bidir.



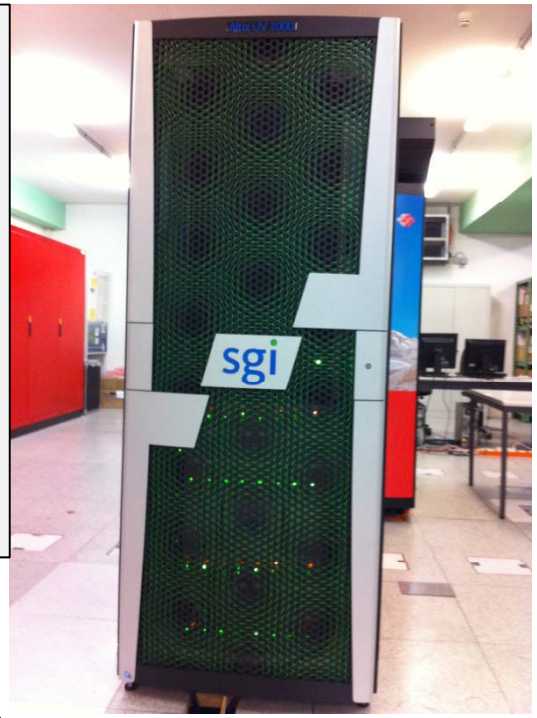
Source: *www.**sgi**.com/pdfs/4192.pdf*

Programming Models

- Sequential codes
- OpenMP
- MPI
- Hybrid OpenMP / MPI
- Pthreads
- SGI UPC compiler
  - conform to Version 1.2 standard
- Java

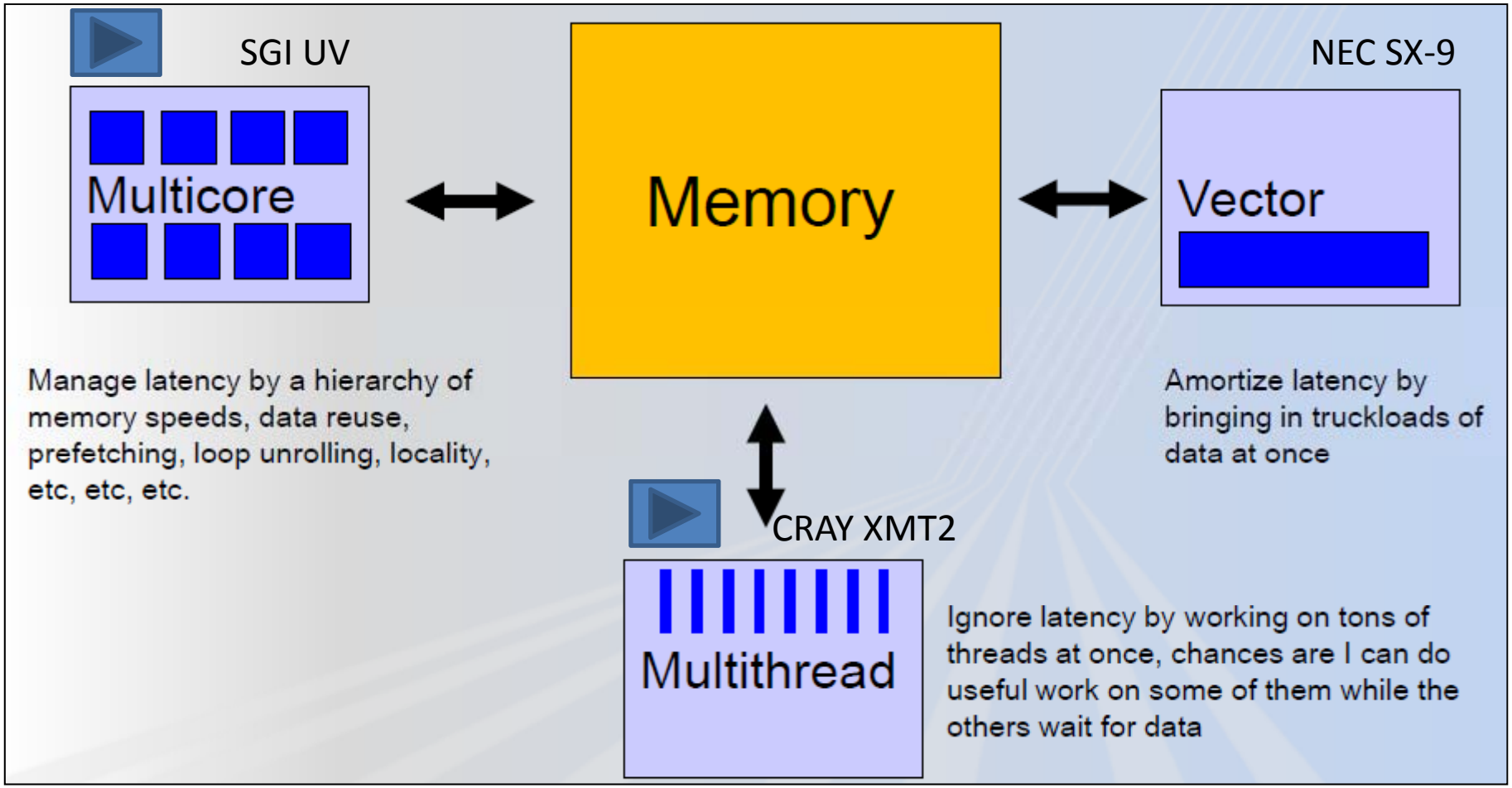
SGI UV-1000

```
total 88
drwxr-sr-x  3 root csstaff 4096 2011-05-05 14:46 cmake
drwxrwsr-x  6 tack csstaff 4096 2011-05-06 10:07 hdf5-1.8.6
lrwxrwxrwx  1 root csstaff   16 2011-05-06 10:08 intel -> ../julier/intel/
drwxrwsr-x 10 tack csstaff 4096 2011-05-06 10:09 modulefiles
drwxr-sr-x  4 tack csstaff 4096 2011-05-04 16:30 mpich2-1.3.2p1
drwx--S---  3 tack csstaff 4096 2011-05-04 09:07 mvapich-1.2rc1
drwx--S---  4 tack csstaff 4096 2011-05-04 16:30 mvapich2-1.6
drwxrwsr-x  5 tack csstaff 4096 2011-05-06 09:00 netcdf-4.1.2
drwxr-sr-x  3 tack csstaff 4096 2011-04-29 13:47 pgi-10.9
drwxrwsr-x  3 tack csstaff 4096 2011-04-29 10:27 pgi-11.4
drwxrwsr-x  3 root csstaff 4096 2011-05-05 09:20 system
thomscho@rothorn:/apps/rothorn> pwd
/apps/rothorn
thomscho@rothorn:/apps/rothorn>
```



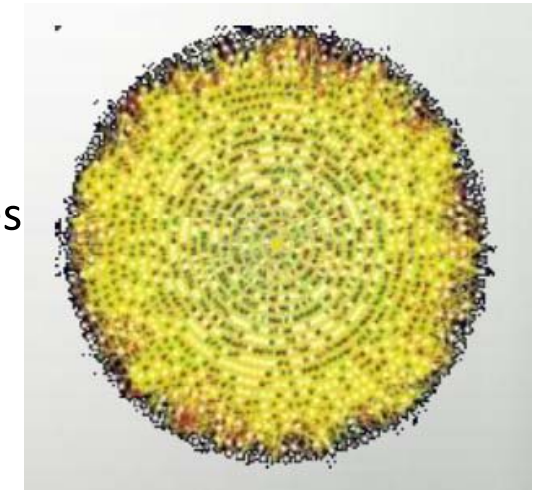
- Visit /apps/rothorn
- Modules Environment
  - Allows to switch between various versions of compiler and libraries
- Batch Scheduler SLURM
  - Start with one demon for all cores
  - One demon for each node

Select the Systems

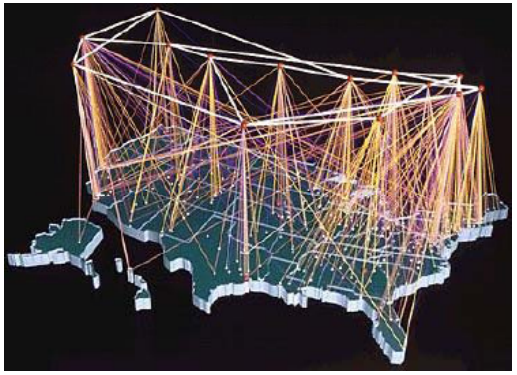


Cray XMT's Applications

- Any application that involves
  - Random or indirect memory accessess
  - Unbalanced subcomputations
  - Unstructured, dynamic, and/or sparse data structures
  - Linked data structures
  - Sorting or searching
  - Huge data sets



- Applications that need to access large amounts of memory and in an unpredictable manner
  - Graph Analysis
  - Data Mining
  - Business intelligence
  - Pattern matching
  - Power grid analysis



CRAY XMT2

- World's **First** Next Generation XMT Supercomputer
- Designed for deep analysis of large datasets
- Special Purpose Processor with 128 Hardware Threads
- Very low power
- Built on the **Cray XT5** Infrastructure
- Highly scalable to 128 TB RAM per System
- Propietary C/C++ Compiler
- Outstanding Performance on Graph Problems, up to 2 or 3 magnitudes over Conventional Clusters

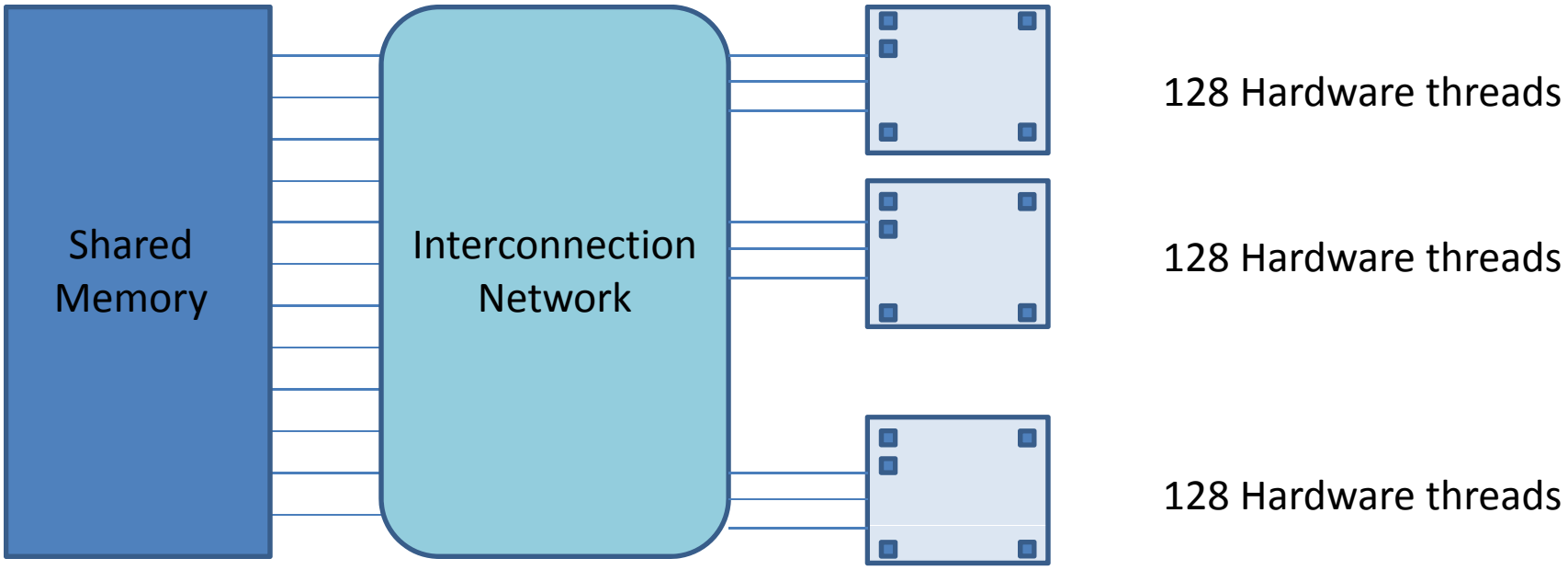
CRAY XMT2 (Matterhorn)

- The Cray XMT provides the programmer an illusion of a globally addressable flat memory hierarchy, thus avoiding the need for load-balanced data partitioning and redistribution among processors

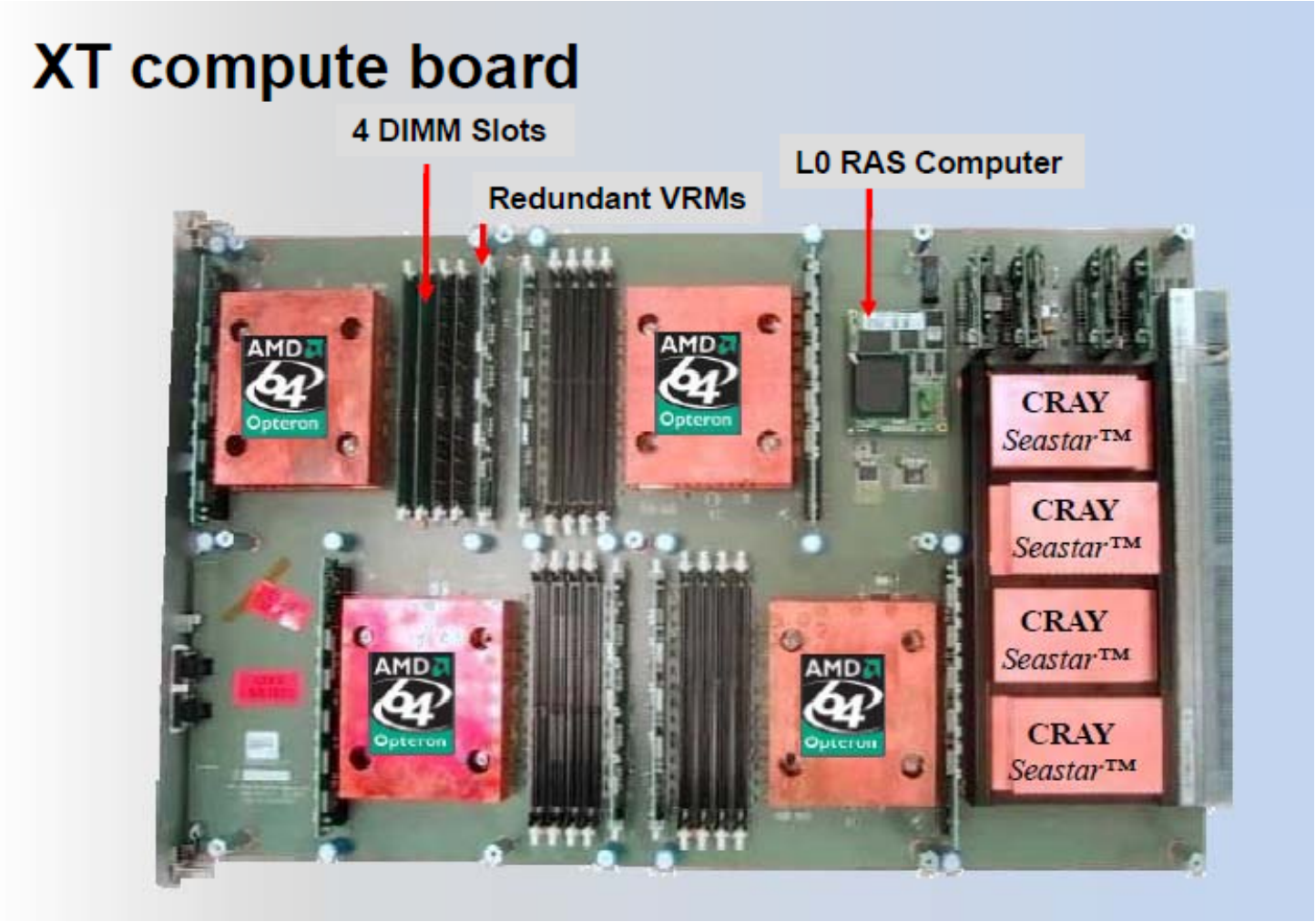
- Architecture
  - 1 Cabinet with 16 quad-socket Compute Blades
  - 64 Cray Threadstorm CPUs with 32 GB per CPU
  - 2 TB of RAM in total (NUMA)
  - 3D Torus Cray Seastar2
  - 8 dual-socket AMD Opteron Service Blades
  - Login Nodes

XMT2 Programming Model

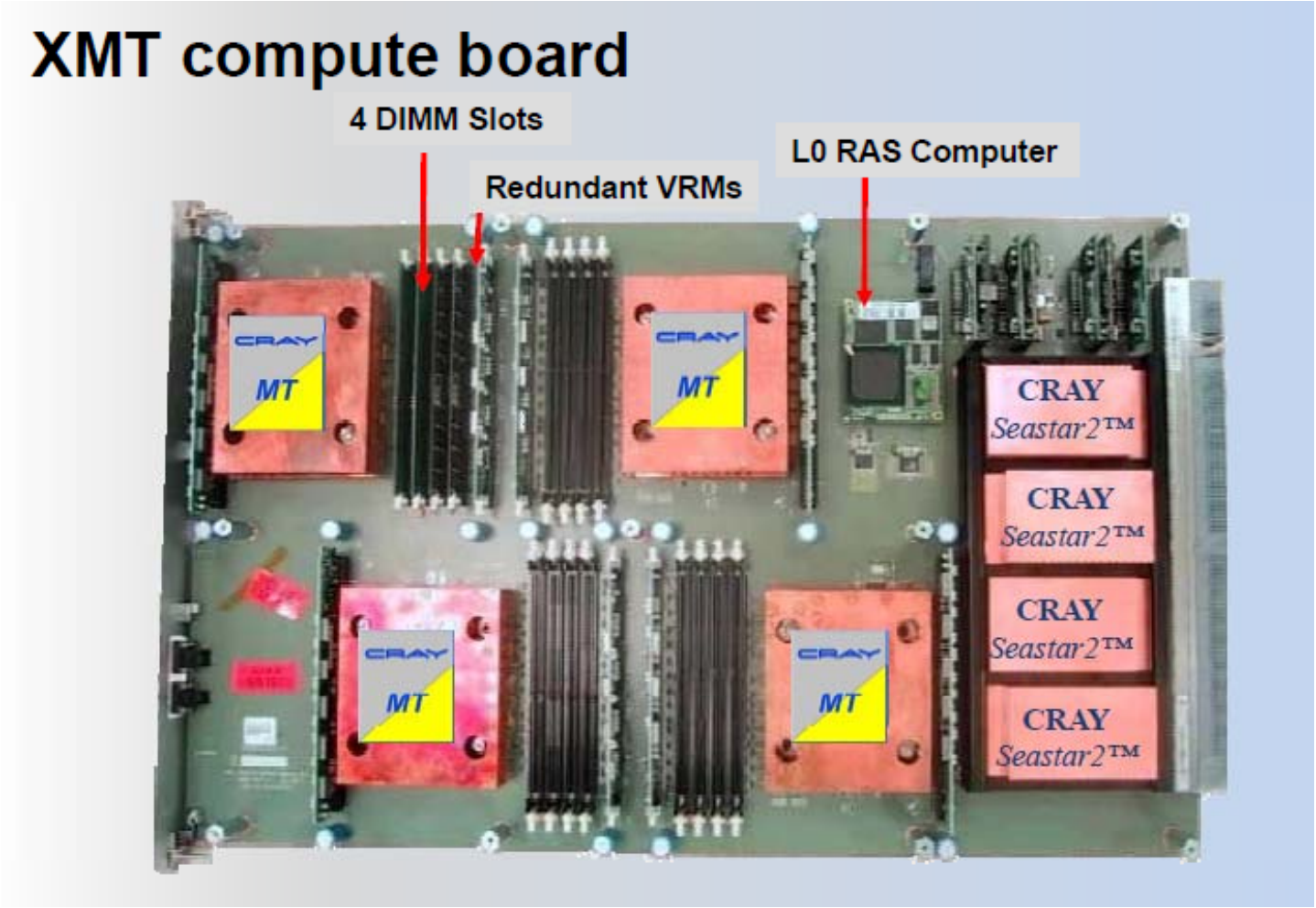
- To the programmer, a XMT processor looks like s single processor, except that the number of threads is increased



Based on XT5


XT compute board
4 DIMM Slots
Redundant VRMs
L0 RAS Computer
AMD 64 Opteron
AMD 64 Opteron
AMD 64 Opteron
AMD 64 Opteron
CRAY Seastar™
CRAY Seastar™
CRAY Seastar™
CRAY Seastar™

Threadstorm CPU socket-compatible



CRAY XMT2 (Matterhorn)

- System not arrived yet, but ready to be shipped
- BM Result was submitted for the next list

The Graph 500 List

Complete Results | Specification | Reference Code

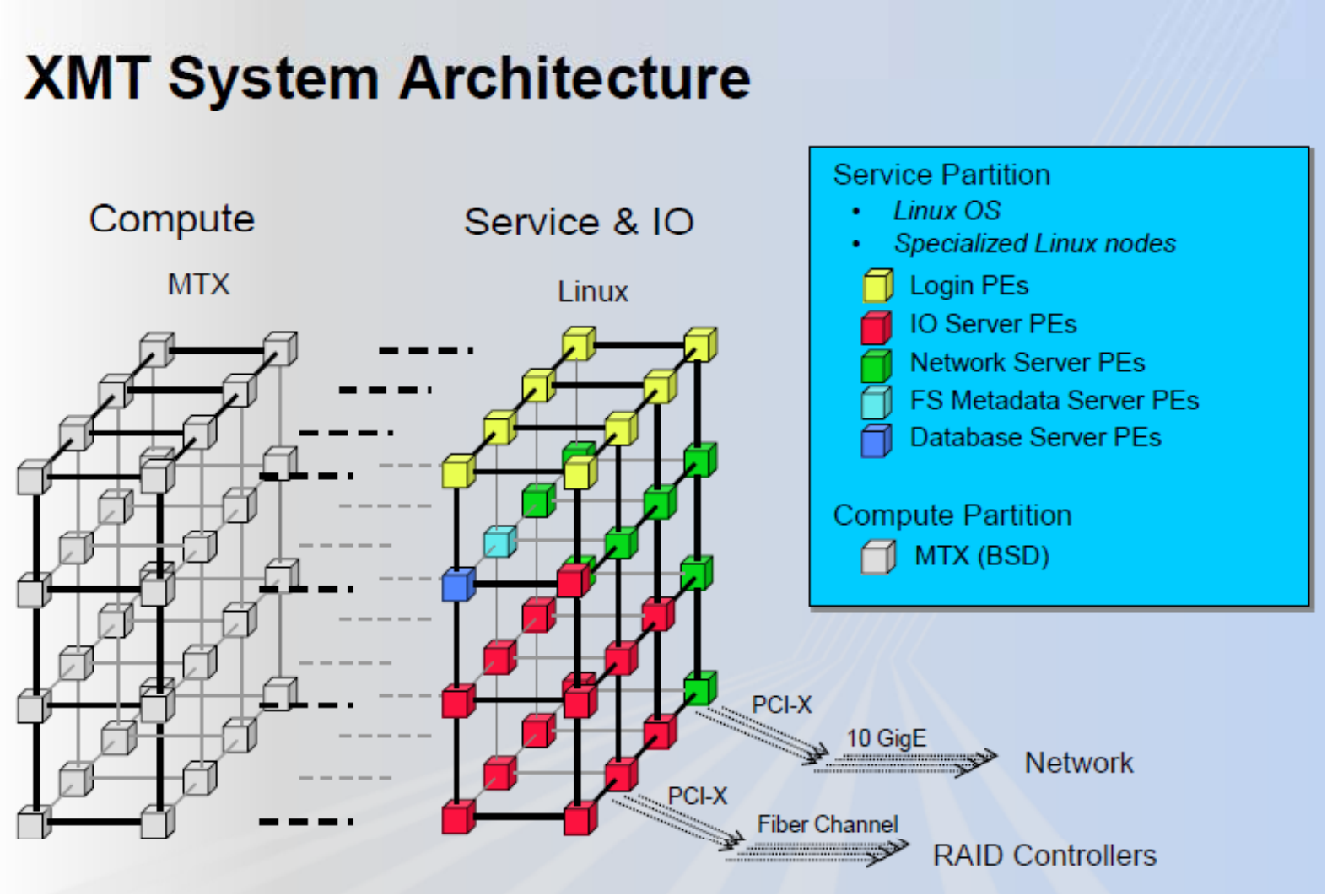
Complete Results November 2010

Rank	Machine	Owner	Problem Size	TEPS	Implementation
1	Intrepid (IBM BlueGene/P, 8192 nodes/32k cores)	Argonne National Laboratory	Scale 36 (Medium)	6.6 GE/s	Optimized
2	Franklin (Cray XT4, 500 of 9544 nodes)	NERSC	Scale 32 (Small)	5.22 GE/s	Optimized
3	cougarxmt (128 node Cray XMT)	Pacific Northwest National Laboratory	Scale 29 (Mini)	1.22 GE/s	Optimized
4	graphstorm (128 node Cray XMT)	Sandia National Laboratories	Scale 29 (Mini)	1.17 GE/s	Optimized
5	Endeavor (256 node, 512 core Westmere X5670 2.93, IB network)	Intel Corporation	Scale 29 (Mini)	533 ME/s	Reference
6	Erdos (64 node Cray XMT)	Oak Ridge National Laboratory	Scale 29 (Mini)	50.5 ME/s	Reference
7	Red Sky (Nehalem X5570 @2.93 GHz, IB Torus, 512 processors)	Sandia National Laboratories	Scale 28 (Toy++)	477.5 ME/s	Reference
8	Jaguar (Cray XT5-HE, 512 node subset)	Oak Ridge National Laboratory	Scale 27 (Toy+)	800 ME/s	Reference
9	Endeavor (128 node, 256 core Westmere X5670 2.93, IB network)	Intel Corporation	Scale 26 (Toy)	615.8 ME/s	Reference

CRAY XMT2 (Matterhorn)

- Software
  - Cray Linux XMT Multithreaded Compute Environment
  - Cray XMT C/C++ Compiler
  - Cray PAT
  - Cray Apprentice
  - Lustre (10 TB)
  - Cray CRMS
  - GPFS client on the service node

CRAY XMT2


XMT System Architecture
Compute
Service & IO
MTX
Linux
Service Partition
Linux OS
Specialized Linux nodes
Login PEs
IO Server PEs
Network Server PEs
FS Metadata Server PEs
Database Server PEs
Compute Partition
MTX (BSD)
PCI-X
10 GigE
Network
PCI-X
Fiber Channel
RAID Controllers

Programming Environment

- C/C++ optimizing compiler
  - Aggressive automatic parallelization capability
  - Support for various hierarchies of parallelization
    - Reductions, linear recurrences
  - Support for atomic memory operations
  - Interprocedural optimization
    - Includes capability to inline library functions
- Incremental recompilation and incremental linking
- Tightly integrated with debugging and performance analysis tools
- Programmer influences the compilers parallelization with pragmas

Resource Management System

- No Batch Scheduler provided
- Slurm to be implemented

Cray XMT Programming Workshop

CRAY XMT Programming Workshop - 16-17 June 2011

CSCS is pleased to announce a 2-day workshop on the Cray XMT Programming. The goal of the course is to familiarize the students with the new Cray XMT architecture and enable them to start programming it for maximum performance.

Registration deadline: June 10, 2011.

Please contact mgg(at)cscs.ch for further technical information and apinna(at)cscs.ch for logistical information.

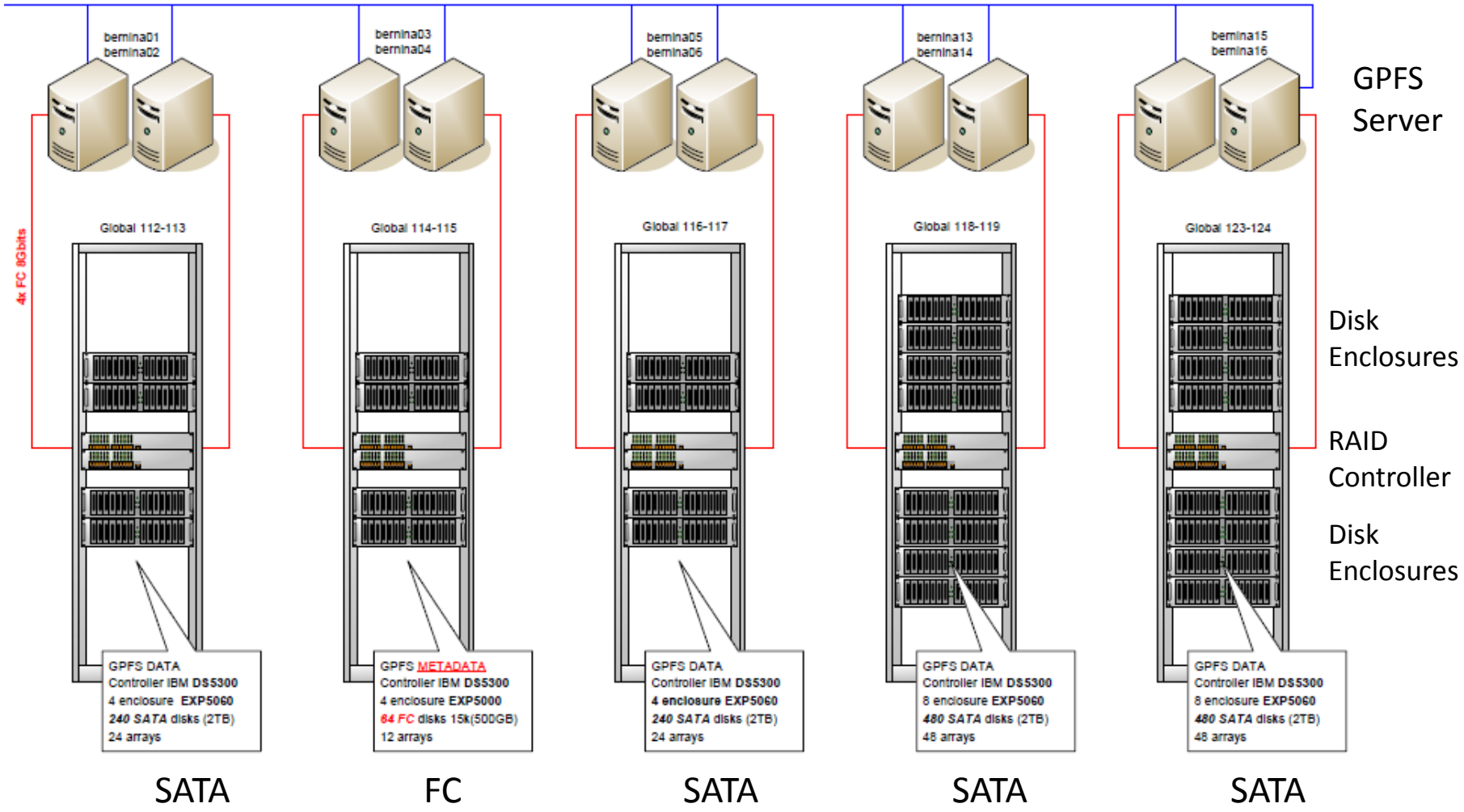
Principal Instructor	***Jim Maltby*** from Cray
Venue	CSCS, Via Cantonale, Galleria 2, 6928 Manno (Please note that on googlemap CSCS is wrongly posted)
Time	9:00 - 17:00 both days
Prerequisites	CSCS next-generation XMT supercomputer "Matterhorn" will be targeted for this workshop.
Maximum number of participants	28
Accommodation	Participants are kindly requested to make their own arrangements for accommodation

Workshop Agenda
The program will cover:

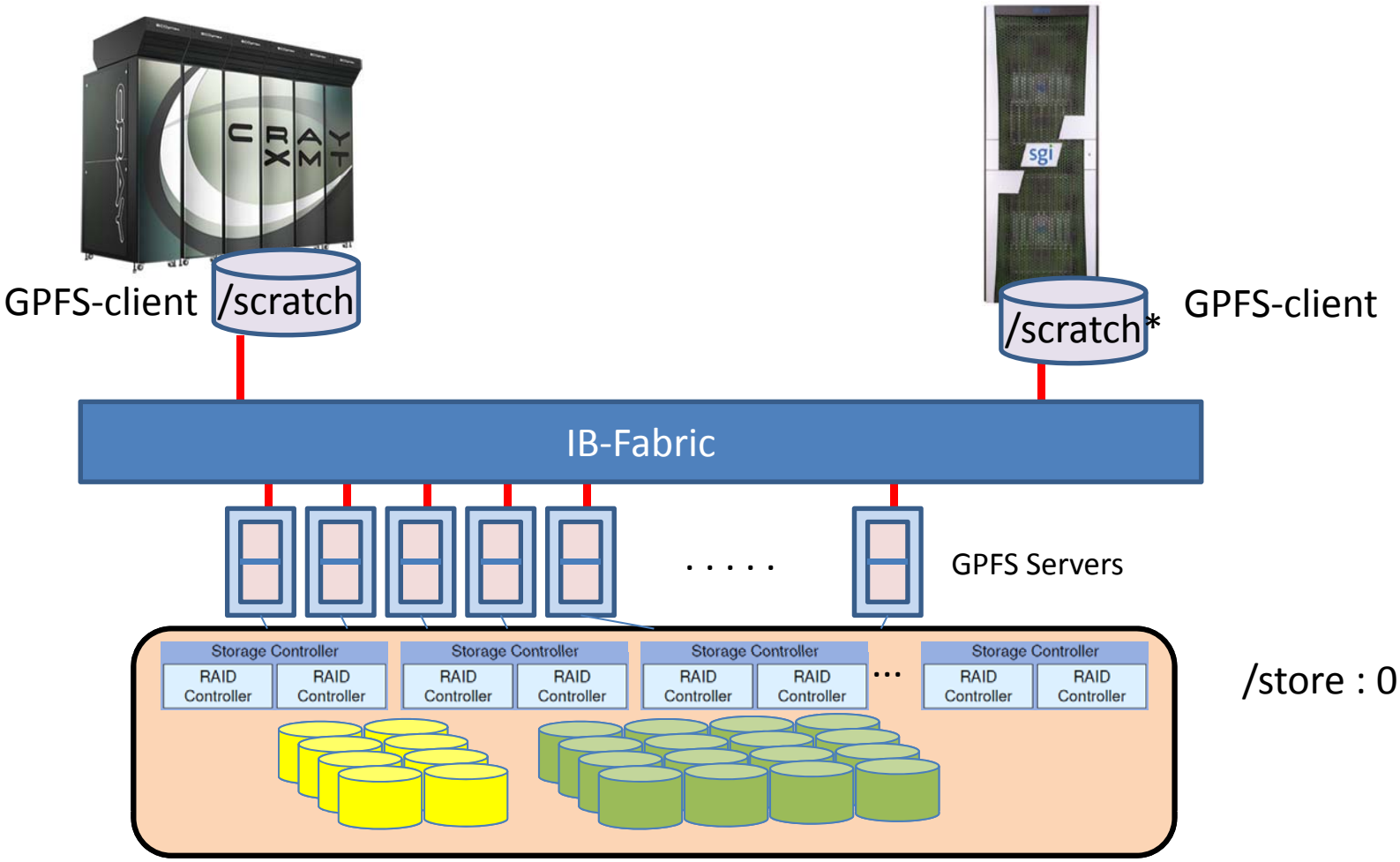
- Overview and history of the Cray XMT architecture.
- Potential application areas for the Cray XMT architecture.
- Programming environment overview.
- Programming for performance.
- I/O programming.
- Using the Cray performance tools and debugger.

In addition a series of hands-on examples will be provided for the students to work on during the course.

CSCS General Parallel File System - 2.1 PB



/store


GPFS-client
/scratch
/scratch*
GPFS-client
IB-Fabric
GPFS Servers
Storage Controller
RAID Controller
RAID Controller
/store : 0.7 PB

/store

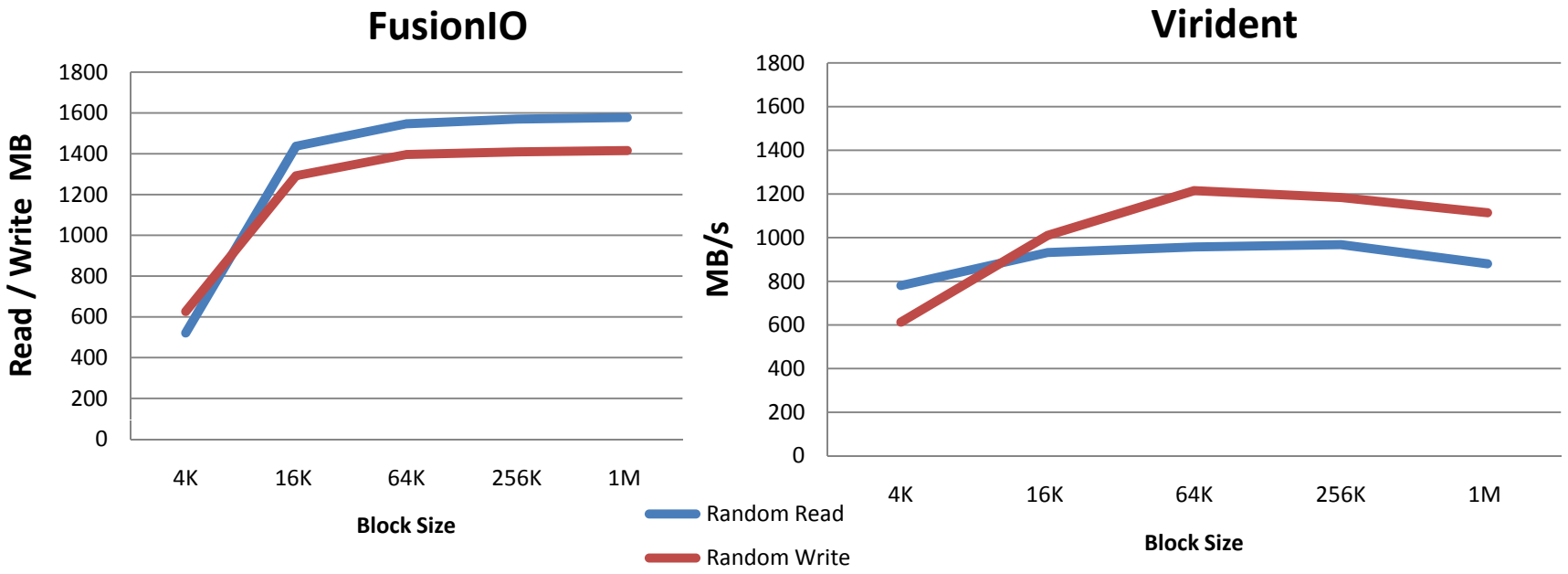
- **Phase 1 (February-May 2011)**
  - 700 TB (as part of Bernina)
  - 10 GB/s
  - FC for Metadata, SATA for user data
  - Integrate Cray XMT and SGI UV in the storage network
- **Phase 2 (October 2011)**
  - 1.5 PB
  - 25+ GB/s
  - Additional GPFS servers
  - GPFS Userdata: No SATA, but SAS NL
  - GPFS Metadata: SSDs
  - Reduce power consumption
  - 60% reduction in foot print

Why SSD`s for Metadata?

- Our current GPFS Server (DS5300) delivers 58,000 IOPs with FC disks
- The new Texas Memory Systems 3U gives 400,000 IOPs with SLC flash drives and max. 10 GB/s bandwidth

SSD demo cards

- Evaluated different SSDs



- GPFS with different type of devices for user data?

User data – store more with less power

LSI Introduces High-Performance, High-Density HPC Storage System

New Engenio 2600-HD storage system enables up to 40GB/s of throughput and 1.8 petabytes of capacity in a single standard rack

SC10, NEW ORLEANS, La., November 15, 2010 – LSI Corporation (NYSE: LSI) today introduced the Engenio® 2600-HD high-density storage system, purpose-built to meet the demanding data requirements of high-performance computing (HPC) file systems. The system offers exceptional performance and a highly scalable, dense architecture that is designed to help HPC organizations maximize productivity and achieve a quicker time to results, while minimizing data center floor space and overall energy expenditures.

The Engenio 2600-HD system consists of two LSI™ 6Gb/s SAS-based controllers integrated into the new Engenio DE6600 high-density SAS drive enclosure. The system is capable of sustaining up to 4GB/s of throughput and housing up to 60 SAS drives in a 4U space.

SUPPORTED DRIVES	
6 Gb/s SAS 15K rpm (3.5-in.)	450 GB, 600 GB
6 Gb/s SAS SLC SSD	150 GB, 300 GB
6 Gb/s SAS 7.2K rpm (3.5-in)	500 GB, 1.0 TB, 2.0 TB, 3.0TB

DenseStak Solution

600 SAS drives housed in just 40Us of rack
Up to 1.8 PB

Summary

- **Two unique and powerful Analysis Systems**
  - Cray XMT2 and SGI UV in the CSCS storage network
  - Both with latest generation CPU technology
  - Will add software as required and available
- **IBM General Parallel File System**
  - Reliable
  - Scalable
  - Latest Generation with TSM/HSM extension
  - /store is integrated in the disaster&recovery concept of CSCS

- For any question, please email me
  - schoenemeyer@cscs.ch