



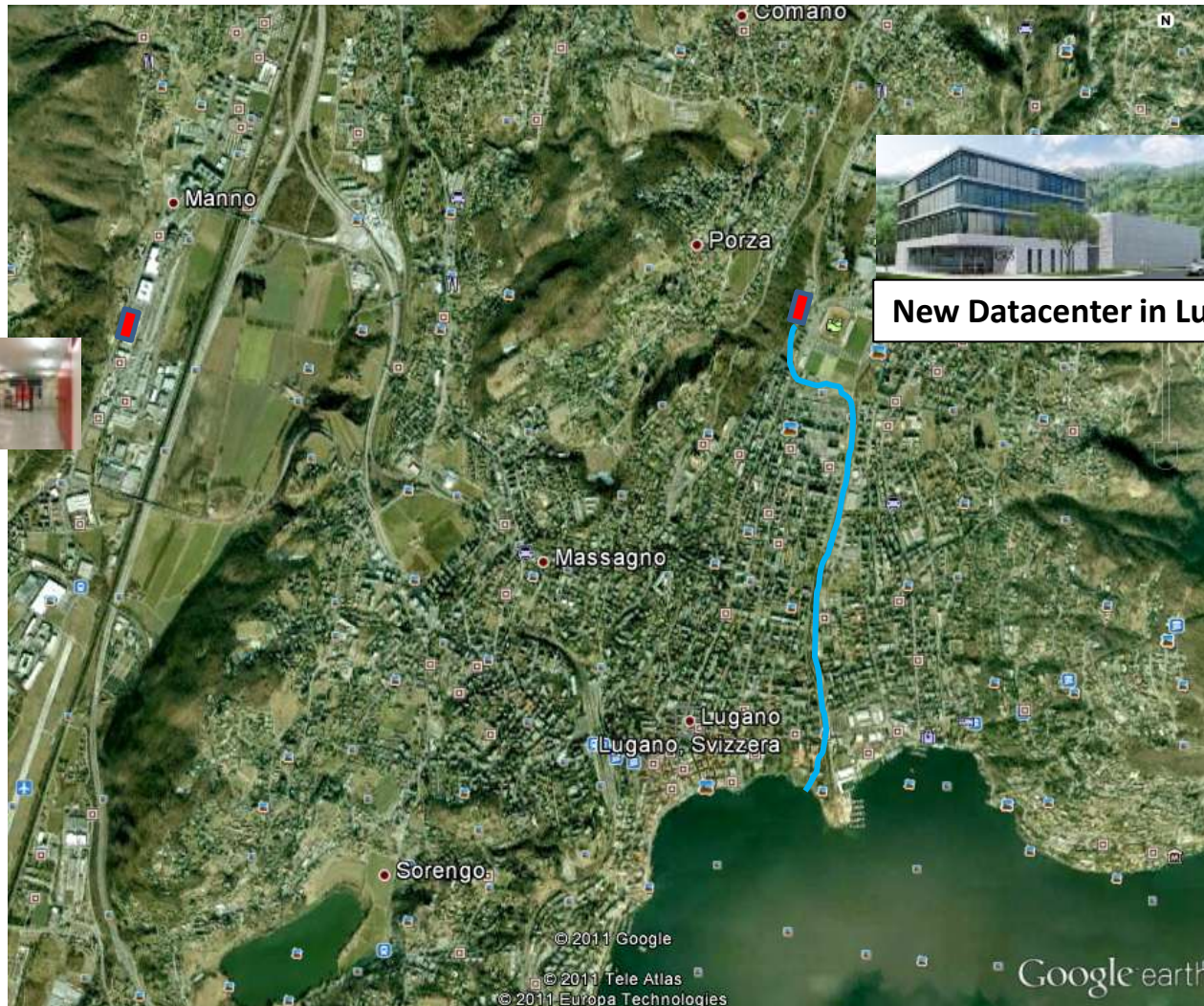
HPC Technology Update – Challenges or Chances?

Swiss Distributed Computing Day

Thomas Schoenemeyer, Technology Integration, CSCS



Move in Feb-April 2012



1500m²
16 MW
Lake-water cooling
PUE 1.2

New Datacenter in Lugano, Cornaredo



Manno

TOP500 (Nov. 2011) – Leaders

- **Japan**

- K-computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect (**10.5 PF**)

- **China**

- Tianhe-1A - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050, (**2.56 PF**)

- **US**

- Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz (**1.76 PF**)

- **EU**

- Tera-100 - Bull bullx super-node S6010/S6030 (**1.05 PF**)

- **CH**

- Monte Rosa – Cray XE6- Opteron 16-core 2.1 GHz (**0.32 PF**)

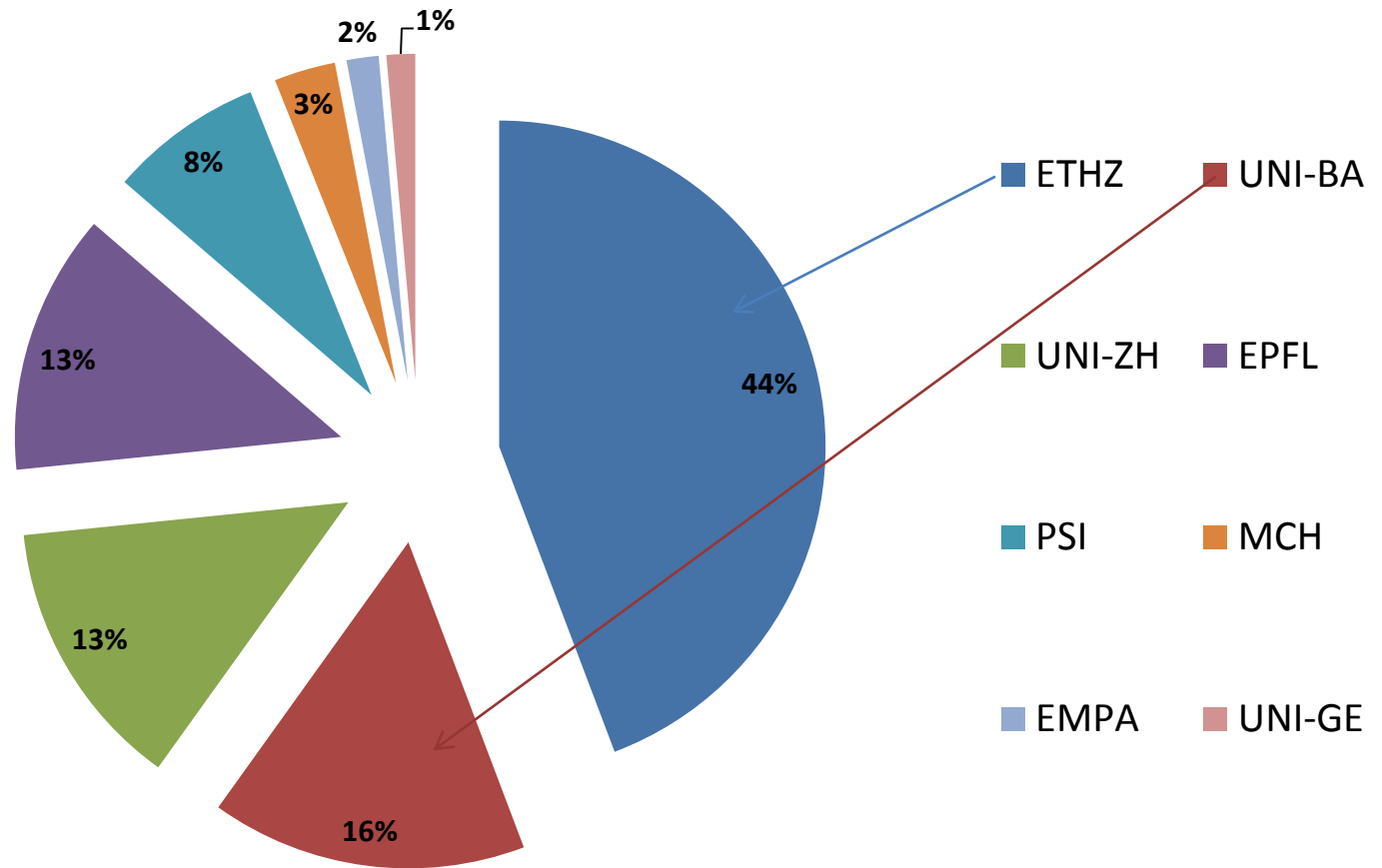
(Rmax score in the HPL benchmark)

TOP500 Power

- **Japan**
 - K-computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect **(10.5 PF) 12.7 MW**
- **China**
 - Tianhe-1A - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050, **(2.56 PF) 4.4 MW**
- **US**
 - Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz **(1.76 PF) 7.0 MW**
- **EU**
 - Tera-100 - Bull bullx super-node S6010/S6030 **(1.05 PF) 4.6 MW**
- **CH**
 - Monte Rosa – Cray XE6- Opteron 16-core 2.1 GHz **(0.30 PF) 0.75 MW**

(Rmax score in the HPL benchmark)

CSCS Resources – Usage by Institution



January 2011 – October 2011: Total 151 Mio CPU hours



Systems at CSCS – Nov. 2011

Our largest systems

- Monte Rosa, Cray XE6 (P)
- Tödi, Cray XK6 (R&D)
- Matterhorn, Cray Next Generation XMT (R&D, P in 1/2012)
- Rothorn, SGI Altix UV 1000 (R&D, P in 1/2012)
- Castor&Pollux, IBM iDataPlex GPU Prototype (R&D)
- Julier, IBM x3850, Postprocessing System (P)
- Purpose built systems for CHIPP and MeteoSwiss (P)

Eiger, Dalco
Visualization System
(R&D and P)

Vis

TSM

Backup



Platform 1: Monte Rosa



- Cray XE6 - Flagship System for National HPC Service
- Upgrade finished on Nov 21th 2011 to AMD Bulldozer (16c) - 1496 dual-socket nodes
- Gemini™ Interconnect with 3D-Torus topology
- 47872 cores - 402 Tflops Peak, 48 TB Memory – 32 GB each node (**316 PFs HPL**)
- 300 TB /scratch – based on Lustre parallel file system
- Slurm workload manager

Platform 2: Tödi



- Cray XK6 – Research & Development System
- 176 nodes in two cabinets, 1st of its kind
- Gemini™ Interconnect with 3D-Torus topology
- Each node one 16-core AMD Interlagos CPU and one **Nvidia X2090 Tesla GPU**
- 140 Tflops hybrid peak performance
- 5.6 TB Memory – 32 GB each node
- A few hundred TBs of /scratch – based on Lustre parallel file system
- SLURM workload manager
- Cray compiler with directives
- 290 Gflops versus 665 Gflops per ‘socket’

Tödi Compute Blade



Platform 3: CSCS Data Analysis Facility – Nov. 2011

Cray XMT

Installed May 2011
64 Threadstorm CPUs
2 TB Memory

SGI UV 1000

Installed March 2011
16 blades
2 TB Memory

GPFS Storage

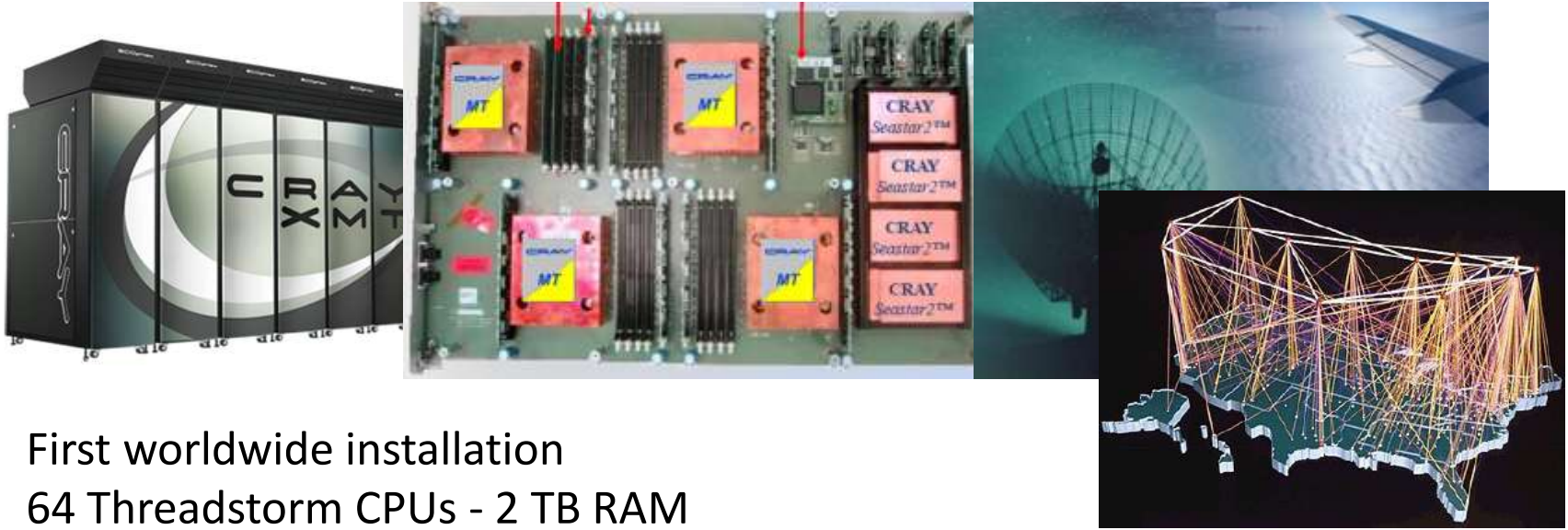
Final Phase installed
November 2011
2.2 PB SAS NL usable
8 TB SSD Storage

GPFS client
Lustre /scratch

GPFS client
Lustre /scratch

DataDirect
Networks

Platform 3 a: Cray Next Generation XMT –Matterhorn

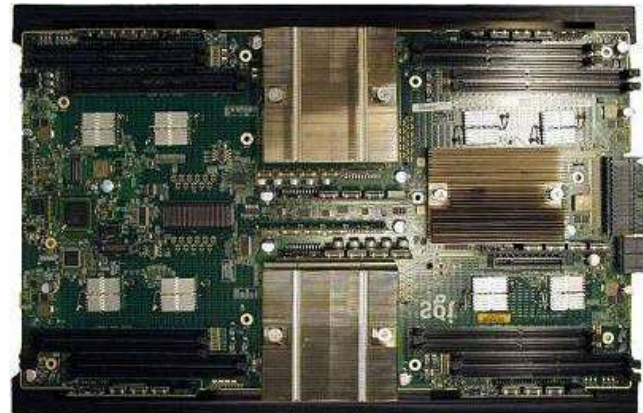


- First worldwide installation
- 64 Threadstorm CPUs - 2 TB RAM
- Proprietary Interconnect
- Designed for Graph Analysis, Data Mining, Pattern Matching
- Unstructured, dynamic and sparse data structures
- Smart power grids, bioinformatics
- not “simple search of” but “discovery from” from big data through complex queries

Platform 3 b: SGI UV 1000 - Rothorn



1 cabinet



1 dual-socket blade

- 2 TB RAM ccNUMA with 16 dual-socket blades with Intel Xeon Westmere EX (8 cores), 16 cores each blade
- Extendable to 16TB Global Shared Memory
- Wide range of applications from in-memory databases to complex data analytics
- Healthcare, Digital Content Management, Financial Services



Platform 4: IBM iDataPlex – Castor&Pollux

- 32 IBM dx360 M3 dual-socket nodes
- 2xM2090 Nvidia GPUs for each node = 46 Tflops hybrid Peak
- Two Mellanox QDR InfiniBand Fabrics
- Research Prototype for GPU Virtualization
- Research on Topologies and Scheduling and GPU-virtualization
- Two partitions with 16 nodes and 32 Nvidia Tesla M2090



Dedicated Visualization System Eiger

- 19 dual-socket nodes for different types of visualization
- Various Nvidia generations C10xx and C20xx and ATI Radeon 6990
- QDR InfiniBand fabric
- 66TB /scratch space
- Nodes manufactured by SMC, system integration by Dalco
- 10 GbE LAN ensures access to /home, /project and /store



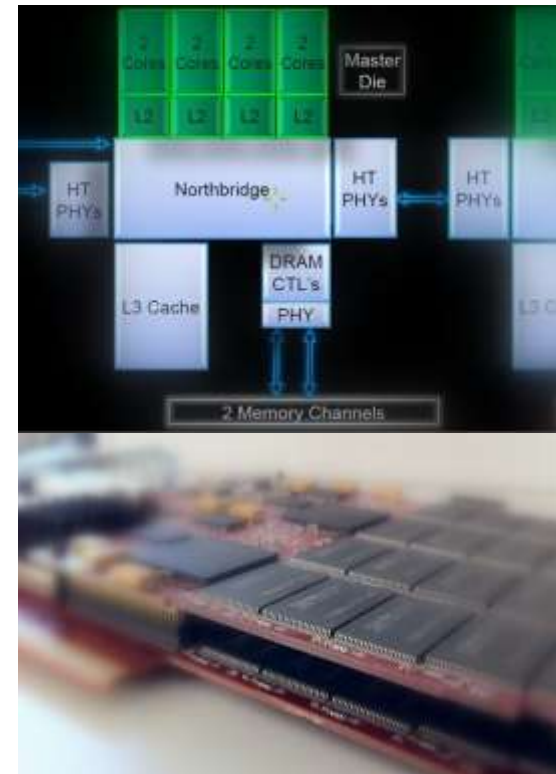
Central Facility: Data Safety

- 5700 cartridges with 8 PB
- 24 LTO5 Drives
- Tivoli Storage Manager
- GPFS-Policies
- Disaster&Recovery
 - Daily changes of ~ 3 TB
 - GPFS metadata recoverable within 2 hours
 - Critical files recoverable within 2 days (10% to 20% of total file system)
 - Non critical files recoverable within 2 to 3 weeks
- InfiniBand Backbone



Challenges

- Exploit new architectures and programming models, with dramatical increases in productivity
- Highly parallel software for heterogenous SC
- I/O Problem can be solved, but needs deep understanding of hardware and the underlying software
- ISV codes may not follow the performance path
- Affordable power envelope will drive the datacenter design
- Operational costs over 4 years exceed the investment costs



Technology Integration at CSCS

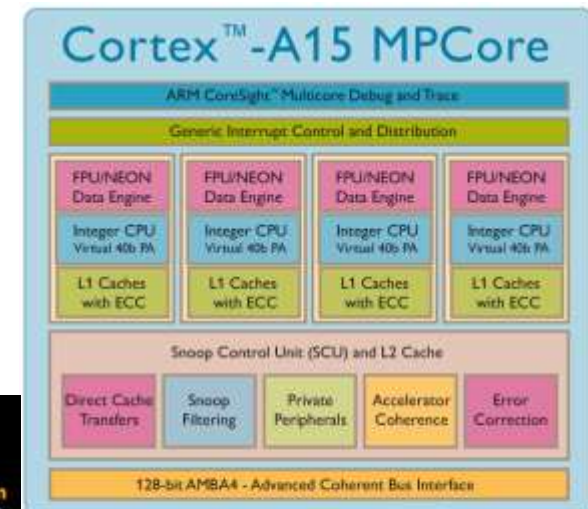
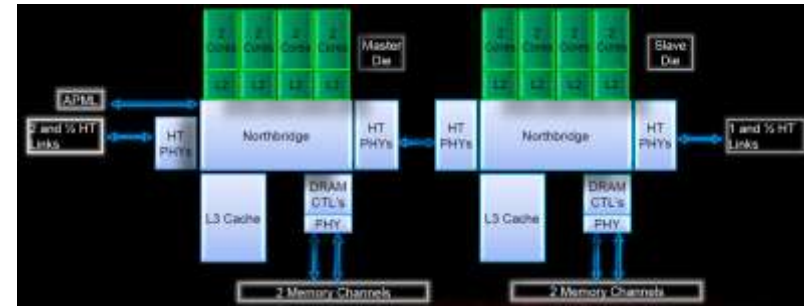
- Future
 - Preselection with other units at CSCS
 - Testing and learning
 - Final Design
 - Integration into production

3 Examples

- Processor: Accelerator
- Storage: SSDs
- Network: Infiniband Range Extender

Processor Architectures for future HPC

- AMD Opteron
- Intel Xeon
- Intel MIC
- ARM
- Nvidia Tesla
- Fujitsu SPARC
- FPGA coprocessors
- BlueGene/Q



"Knights Ferry" Development Platform

Software Development Platform

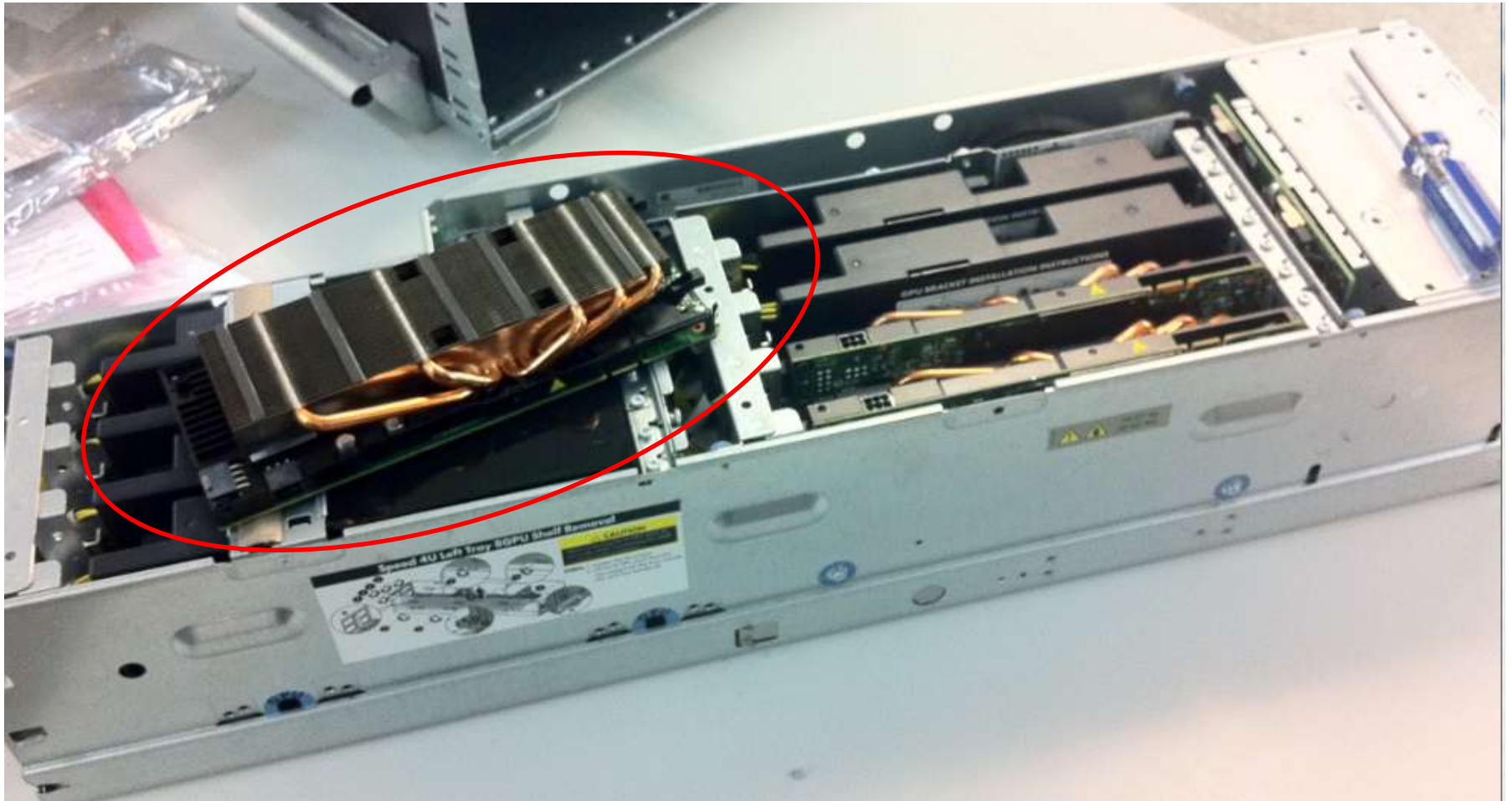
Growing availability through 2011

- Up to 32 cores, up to 1.2 GHz
- Up to 128 threads at 4 threads / core
- Up to 8MB shared coherent cache

TFLOPS Performance

- Up to 2 GB GDDR5 shared memory
- PCIe Card (within 300W envelope)
- Bundled with Intel HPC SW tools

Nvidia M2090 – 512 cores – 665 Gflops DP



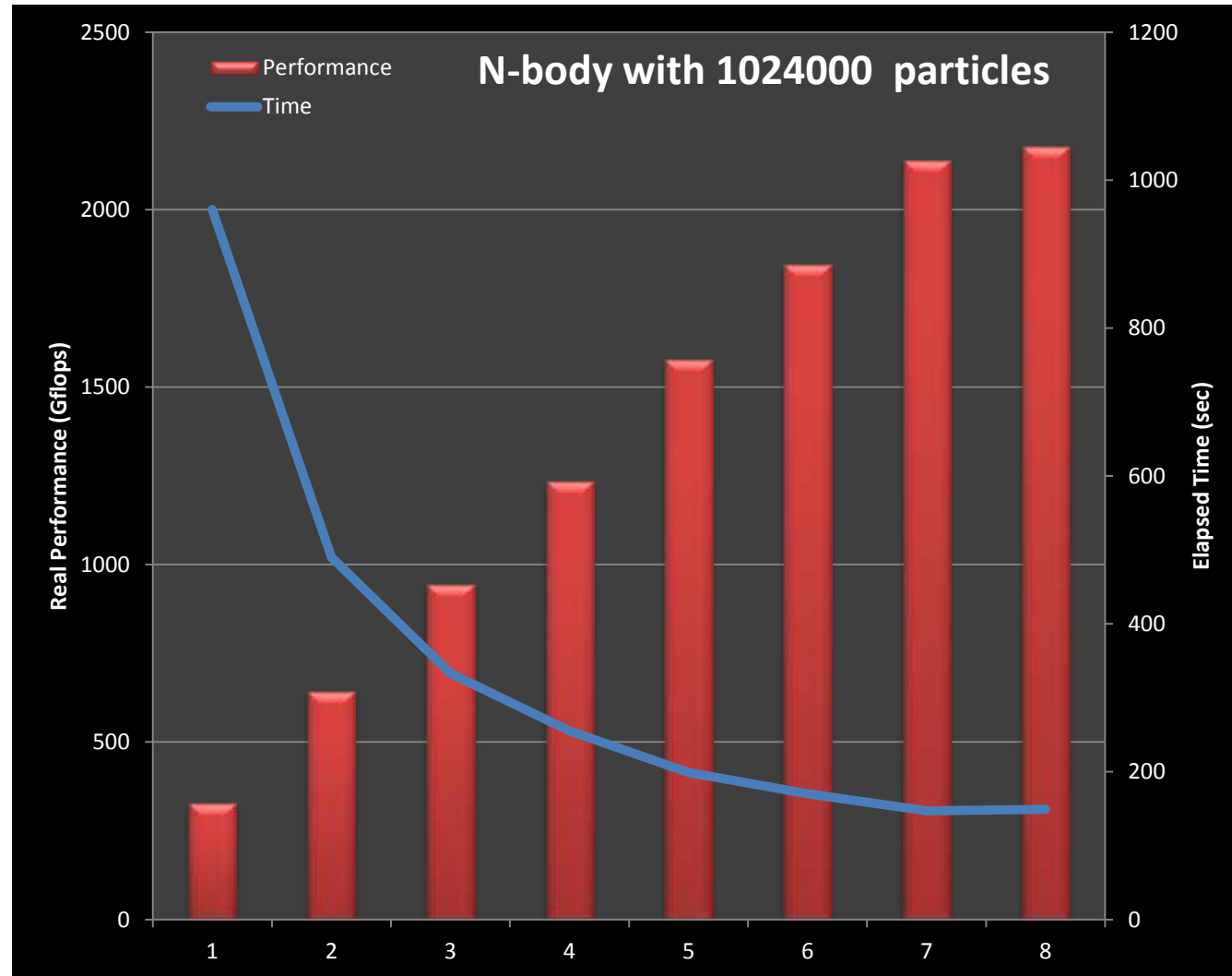
N-Body

- This sample demonstrates efficient all-pairs simulation of a gravitational n-body simulation in CUDA.
- Simple all pair simulation of particles with gravitational forces
- Highly data parallel algorithm to simulate N^2 particle-particle interactions
- E.g. an astrophysical system where a particle represents a galaxy or an individual star
- 1. Example with 64000 particles on Intel Xeon 1 core takes 750 sec, but only 4 secs on a single Nvidia M2090
- Larger cases

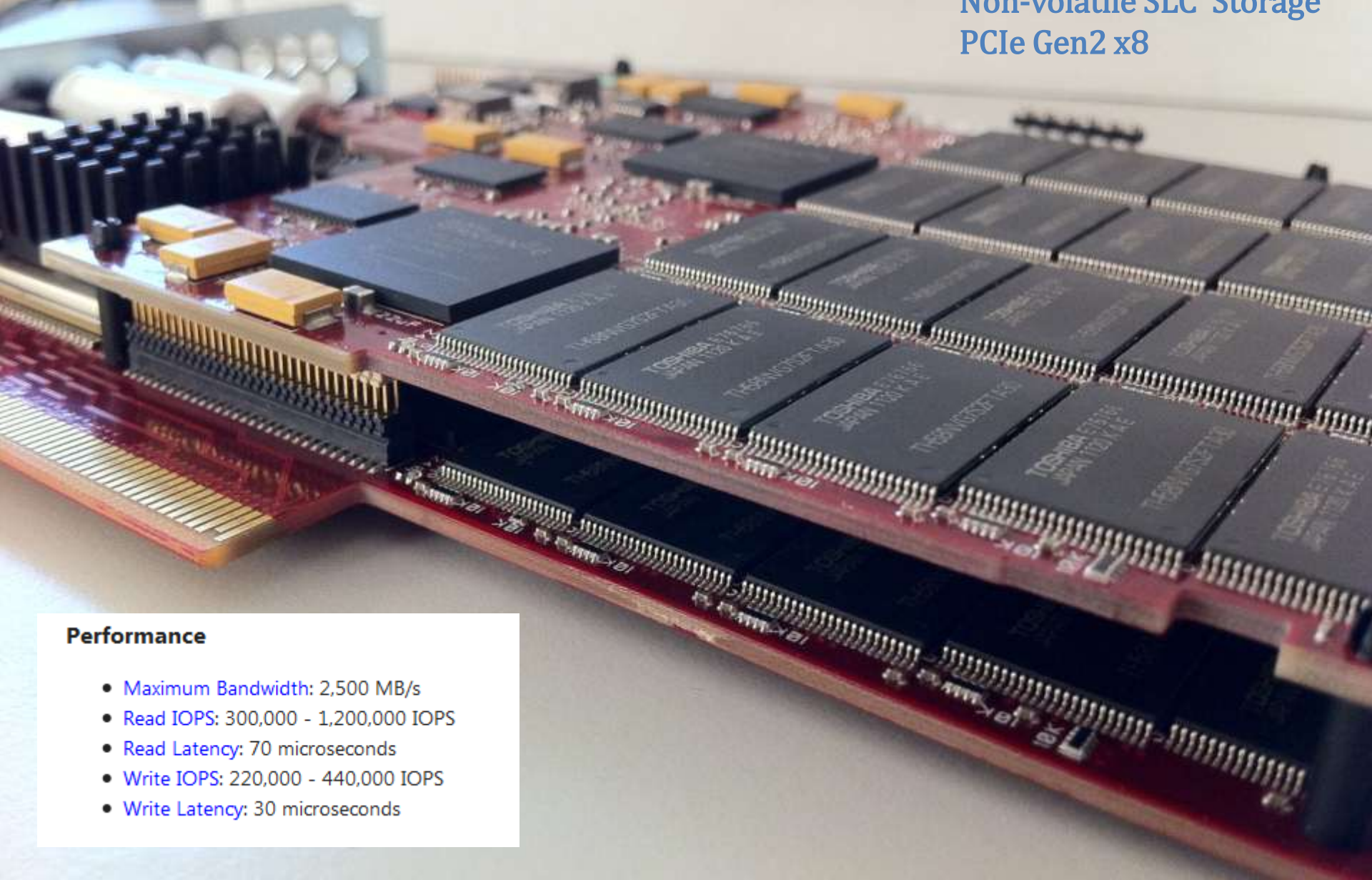
N-body Simulation

HP SL390
Dual-socket Intel Xeon
(12 cores)
8 x Nvidia M2090
(8 x 512 cores)

N-body is a simple highly data parallel algorithm to simulate N^2 particle-particle interactions



Texas Memory Systems
RamSan-70 900 GB
Non-volatile SLC Storage
PCIe Gen2 x8

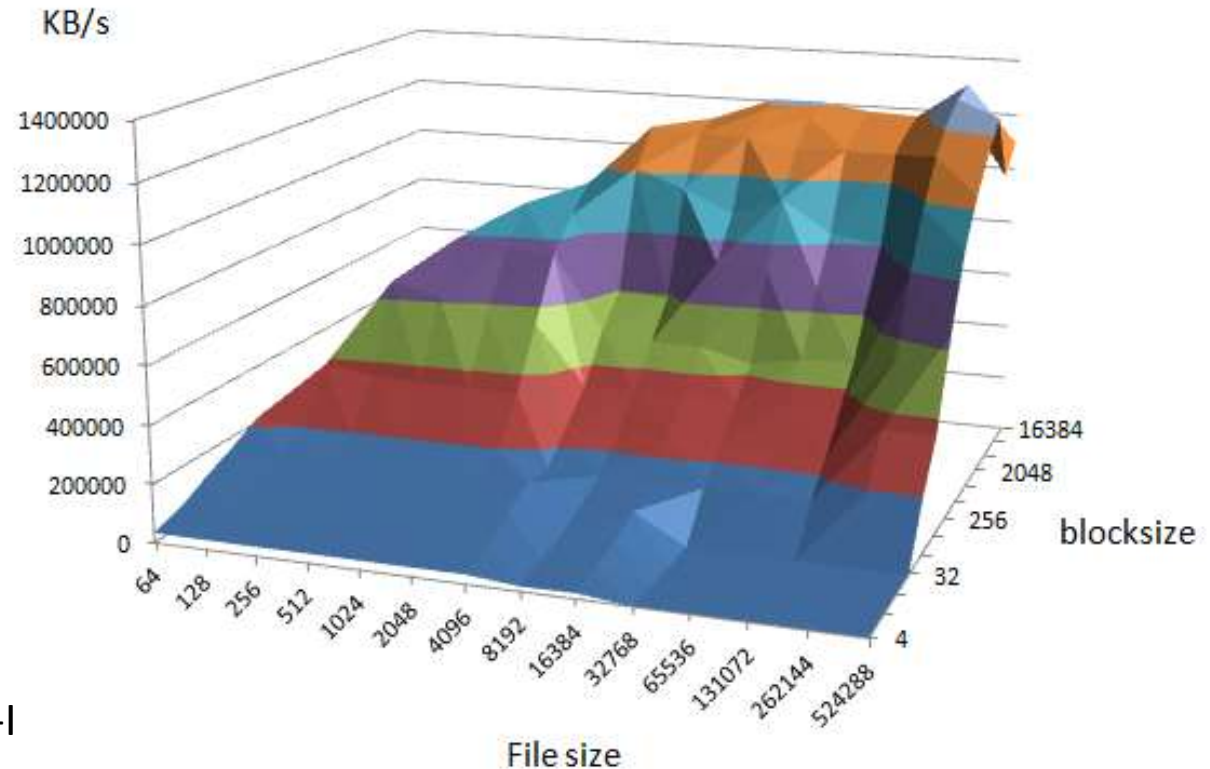


Performance

- **Maximum Bandwidth:** 2,500 MB/s
- **Read IOPS:** 300,000 - 1,200,000 IOPS
- **Read Latency:** 70 microseconds
- **Write IOPS:** 220,000 - 440,000 IOPS
- **Write Latency:** 30 microseconds

Real Measurements

- Virident N800 was measured with 1.3 MIOPS at 512B
- Iozone for ssd and local disk (XFS)
- XFS Performance
- GPFS and Lustre Integration is the challenge



`./iozone -Ra -g 524288 -i 0 -i 1 -l`
one thread (core)

InfiniBand Range Extender

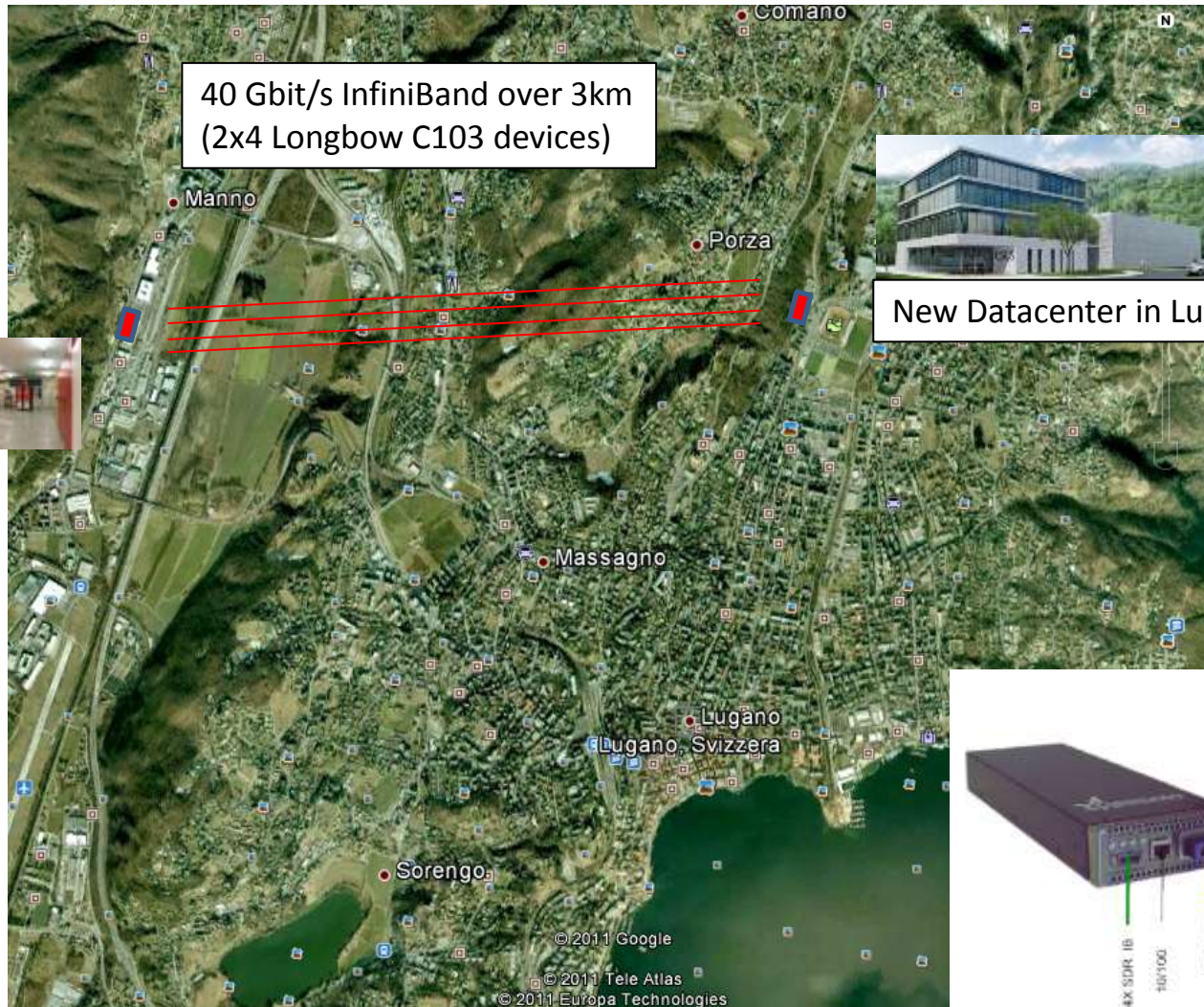
- Create a single File System over two locations
- Aggregate InfiniBand Clusters
- Remote Visualization

Extend the InfiniBand Backbone of CSCS

Providers

- Net.com
- Obsidian
- Bay Microsystems

Obsidian InfiniBand Range Extender



New Datacenter in Lugano, Cornaredo



Manno



Obsidian Longbow C103

- Simulation with 10km IB cable
- *Gpfsperf create* = max. 1560 MB/s
- *Mmdelnsd*: Migration of 6 NSDs (6 LUNs) with 2 TB of data and > 5000 files in < 90 min
- Intensive failover testing
- The technology provided by Obsidian is a very cost-efficient and reliable solution to complete the relocation until June 2012



Full Report on our website – section technical reports



Newsroom

[CSCS on CSCS](#)

[Science](#)

[Publications](#)

▪ [Presentations](#)

▪ [Technical Documents](#)

[Archive](#)

[National Supercomputing Service](#)

[HPC Co-Location Services](#)

[Scientific Computing Research](#)

[Technology Integration](#)

[Service & Compute Resources](#)

[Events](#)

[Working at CSCS](#)

[About us](#)

Technical Documents

CSCS User Information

Reporting

CSCS User Day contributions. Users of CSCS platforms and services are expected to present their work in form of posters at the annual User Day. On this occasion, selected users may be invited to give a presentation.

Acknowledgements

Users are required to acknowledge the use of CSCS resources and services in their publications, posters and scientific presentations. Such recognition is needed as a justification of the funds CSCS receives from the Swiss Confederation to enable research in the field of computational sciences.

Users are obliged to quote the use of CSCS resources as follows:

"This work was supported by a grant from the Swiss National Supercomputing Centre-CSCS under project ID ###".

CSCS Technical Reports

Technology Integration Report:

[Detailed Performance Analysis of Solid State Disk »](#)

[Performance Analysis of TMS SSD »](#)

[Evaluation report for Obsidian »](#)

[Comparing Qlogic and Mellanox »](#)

Summary & Outlook

- Challenges
 - Energy, Energy, Energy
 - Difficult to predict the trends in HPC and adoption of new technologies
 - Codes have to adapt extremely fast in order to take advantage
- Chances
 - Solve scientific problems not been possible before
- Our near term challenge: successful relocation to the new data center in Lugano
- Continue to extend of existing systems according to the needs of the HP2C project



Newsroom

[CSCS on CSCS](#)
[Science](#)
[Publications](#)
[Archive](#)
[National Supercomputing Service](#)
[HPC Co-Location Services](#)
[Scientific Computing Research](#)
[Technology Integration](#)
[Service & Compute Resources](#)
[Events](#)
[Working at CSCS](#)
[About us](#)

hpc-ch News

- [hpc-ch Forum on GPU – Video on Cray XK6 Overview](#) (November 23, 2011)
- [hpc-ch Forum on GPU – Video on Scheduling GRES resources With SLURM](#) (November 23, 2011)
- [hpc-ch Forum on GPU – Video on Installation and Operational Needs of Multi-purpose GPU Clusters](#) (November 22, 2011)
- [hpc-ch Forum on GPU – Video on Multi-resolution flow](#)

Submissions & Events

Project Proposal Submission

Submit Now

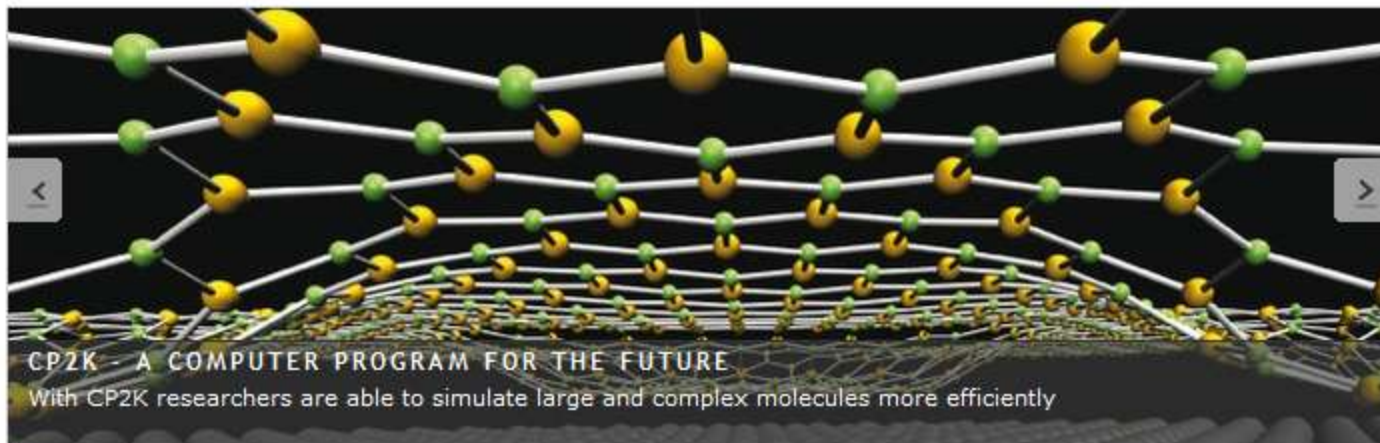
**Deadline
May 11, 2012**

2012 HPC Advisory Council Switzerland Workshop

Register Now

March 13-15, 2012

Latest News



Lake water to cool supercomputers

November 3, 2011

High-performance computing centres use a great deal of electricity. In order to run its new computer centre in Lugano-Cornaredo as energy-efficiently and cost-effectively as possible, CSCS is using the natural resource of Lake Lugano to cool its supercomputers and the new building.

