New CSC computing resources

Olli-Pekka Lehto and Tomasz Malkiewicz
CSC – IT Center for Science Ltd.

Program

- 10-10:45 Overview of new resources
- 10:45-11 First experience from Sisu and Taito
- 11-11:30 Round robin
- 11:30-> F2F meetings

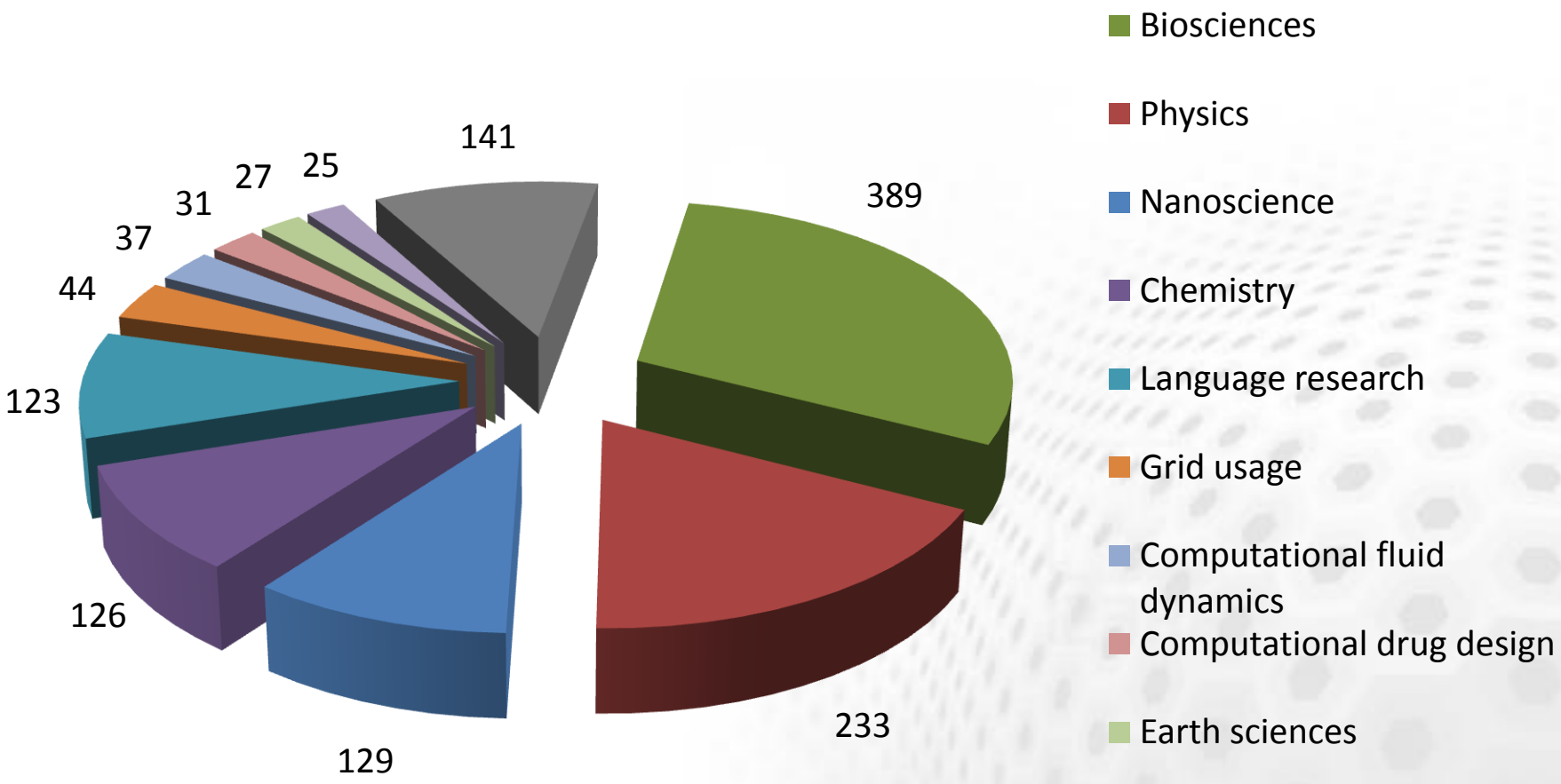
Overview of new resources: outline

- CSC at glance
  - New Kajaani Data Centre
- Finland's new supercomputers
  - *Sisu* (Cray XC30)
  - *Taito* (HP cluster)
  - Other resources available for researchers

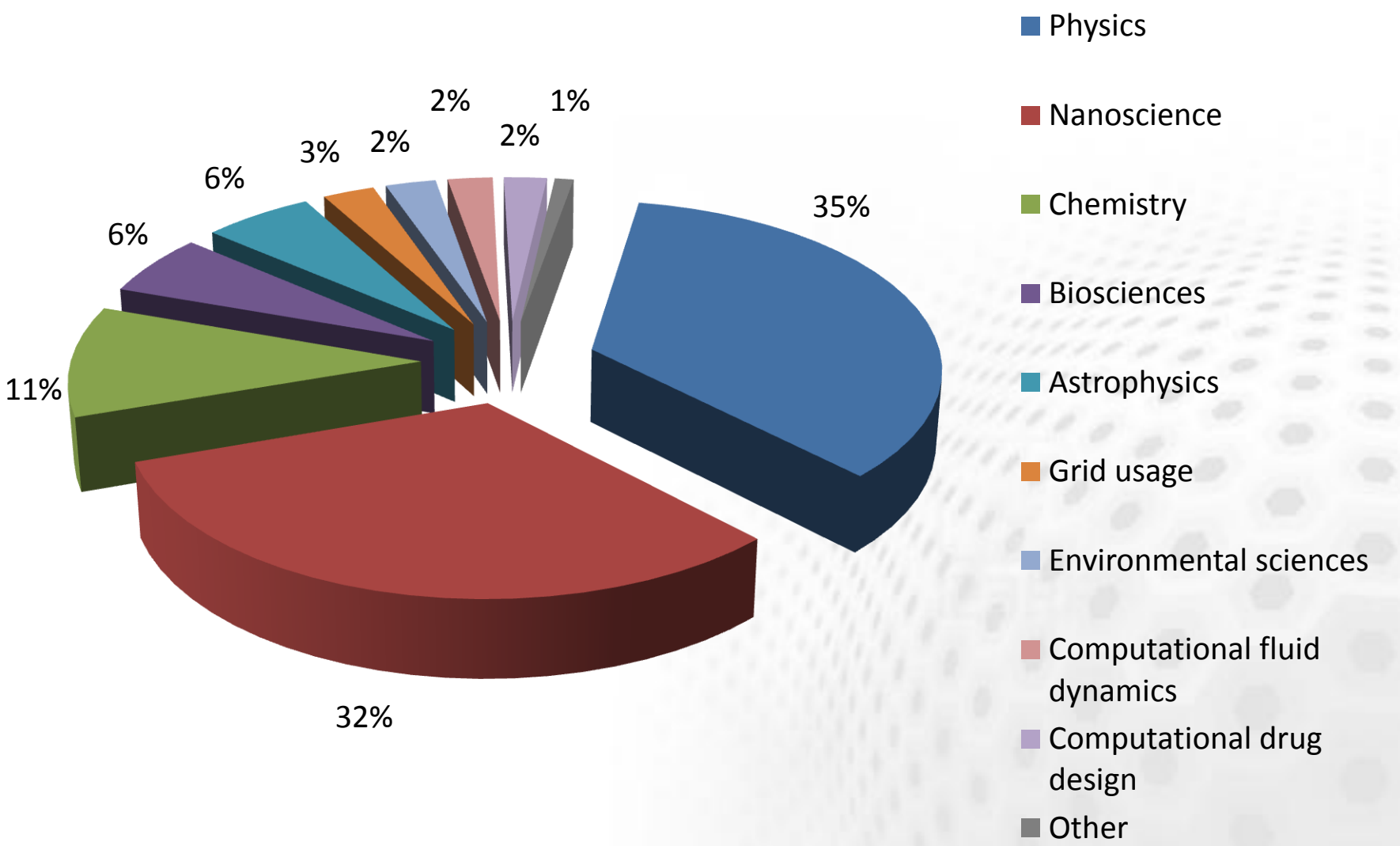
CSC at glance

- Founded in 1971
- Operates on a *non-profit* principle
- Facilities in Espoo and Kajaani
- Staff ~250 people

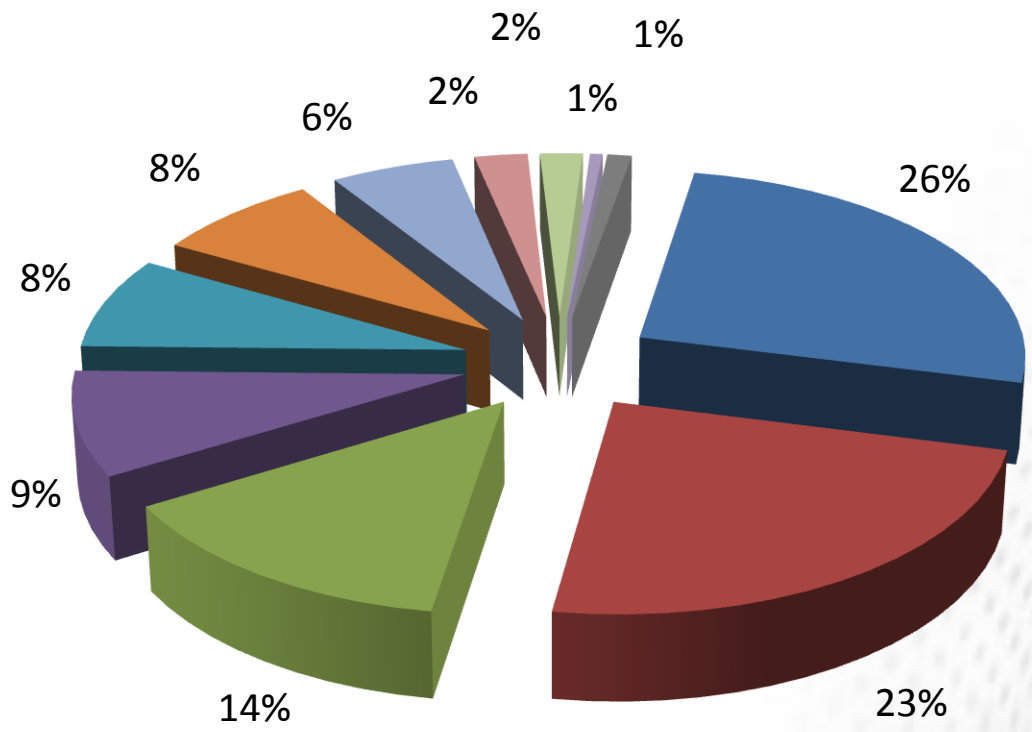
Users of computing resources by discipline 3Q/2012 (total 1305 users)



Usage of processor time by discipline 3Q/2012 (total 2012 96 million cpuh)



Usage of processor time by organization 3Q/2012 (total 2012 96 million cpuh)

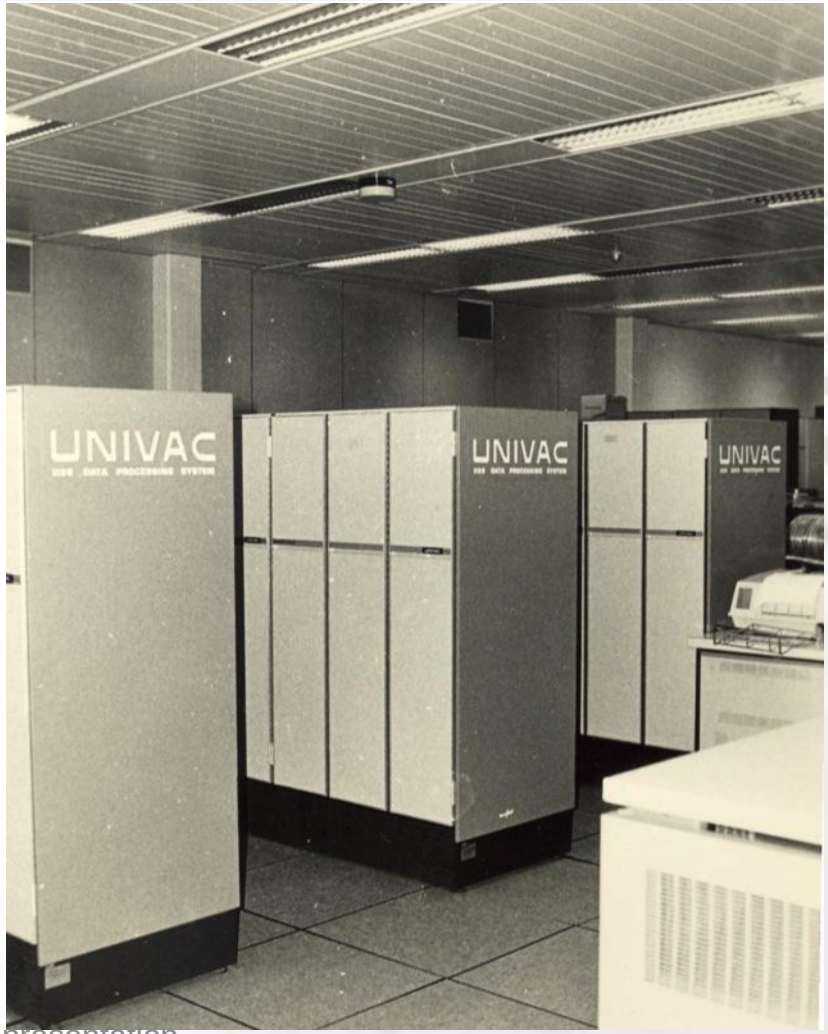


- University of Helsinki
- Aalto University
- University of Jyväskylä
- CSC (Grand Challenge)
- Tampere University of Technology
- Lappeenranta University of Technology
- CSC (PRACE)
- University of Oulu
- University of Turku
- CSC (HPC-Europa2)
- Other

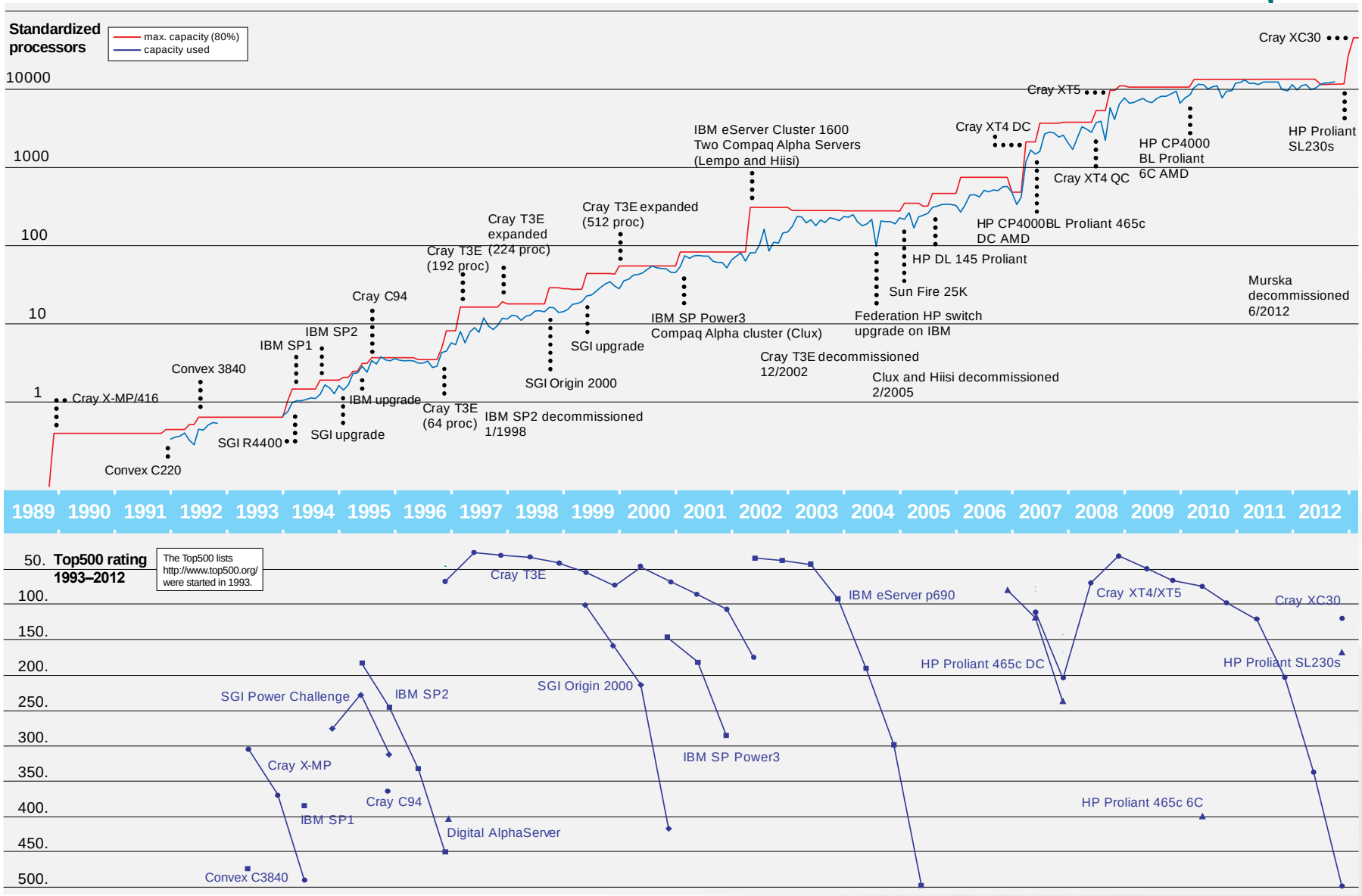
FUNET and Data services

- ***FUNET***
  - Connections to all higher education institutions in Finland
  - Haka-identity Management
  - Campus Support
  - The NORDUnet network
- ***Data services***
  - Digital Preservation and Data for Research
    - Data for Research (TTA), National Digital Library (KDK)
  - Database and information services
    - Nic.funet.fi – freely distributable files with FTP since 1990
  - Memory organizations (Finnish university and polytechnics libraries, Finnish National Audiovisual Archive, Finnish National Archives, Finnish National Gallery)

CSC and High Performance Computing

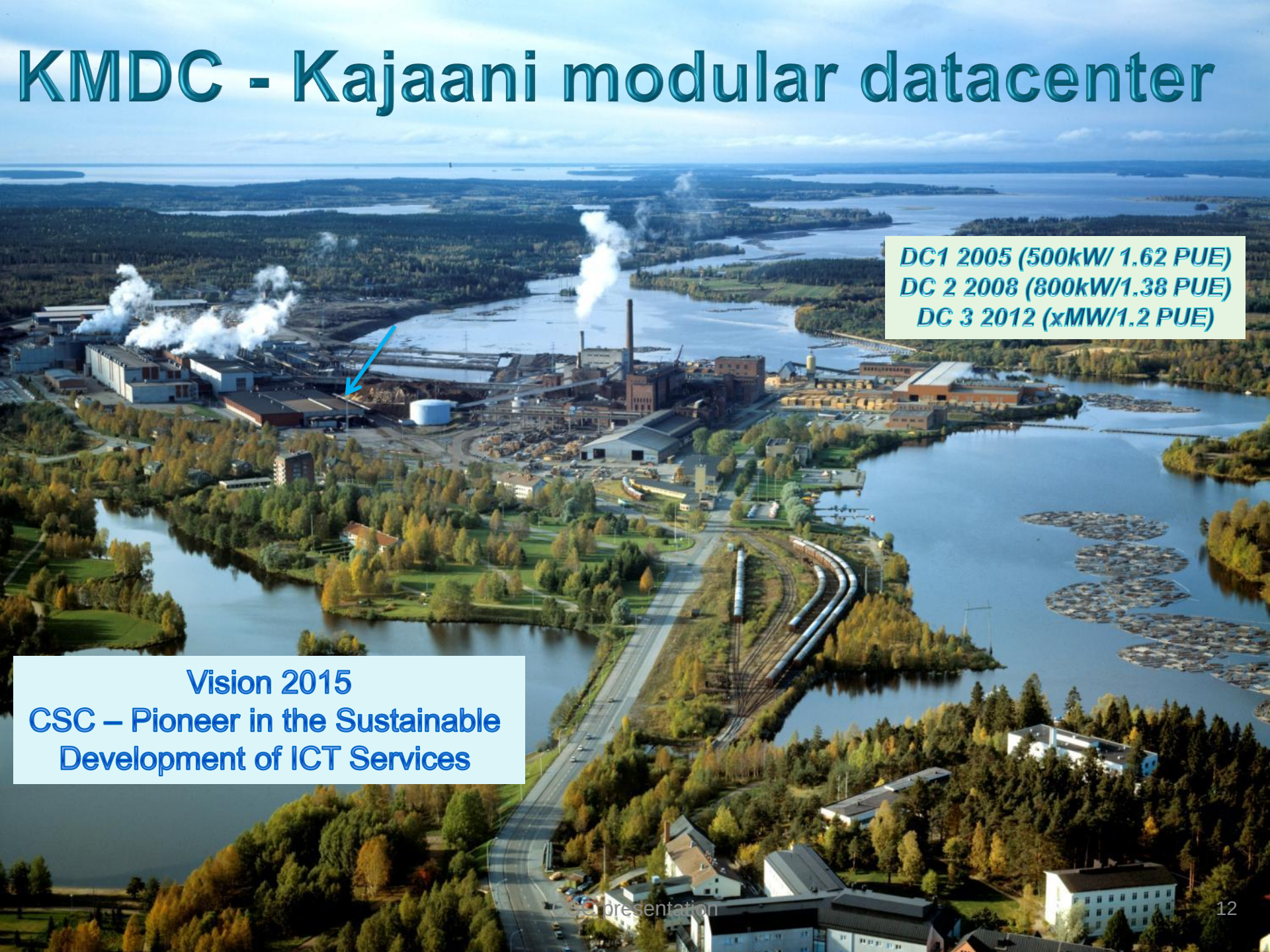


CSC Computing Capacity 1989–2012



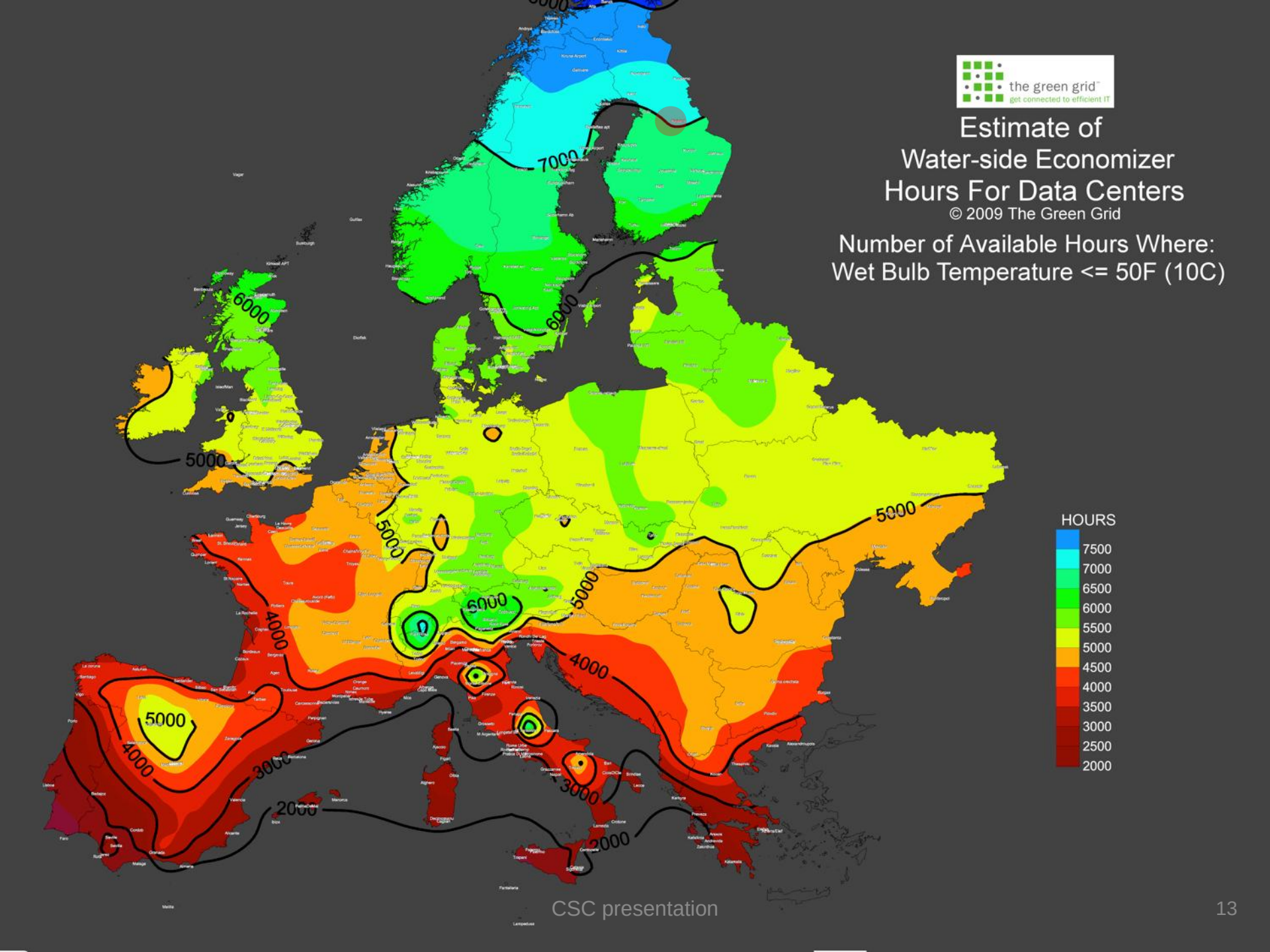
THE NEW DATACENTER

KMDC - Kajaani modular datacenter

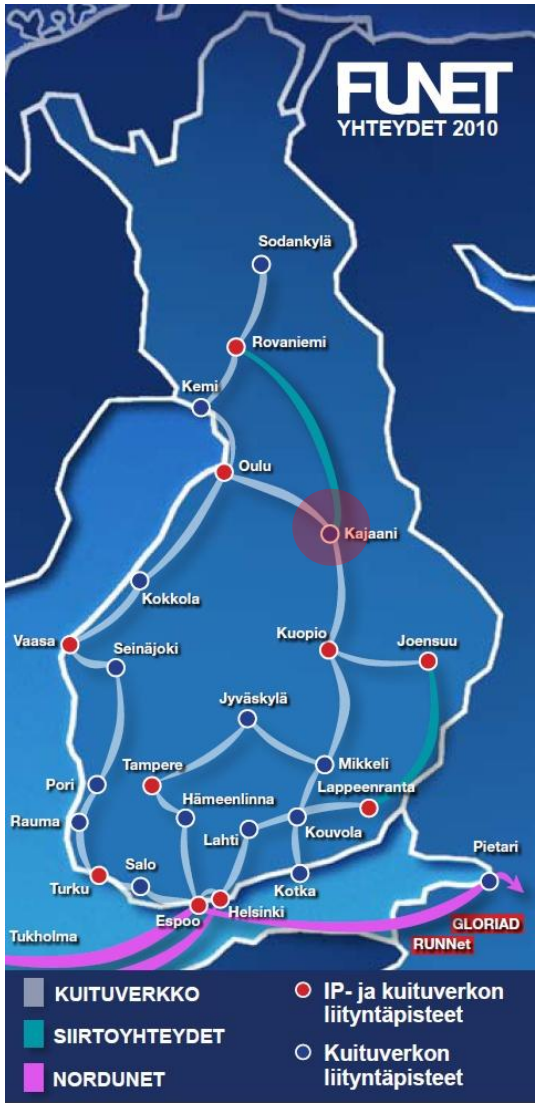


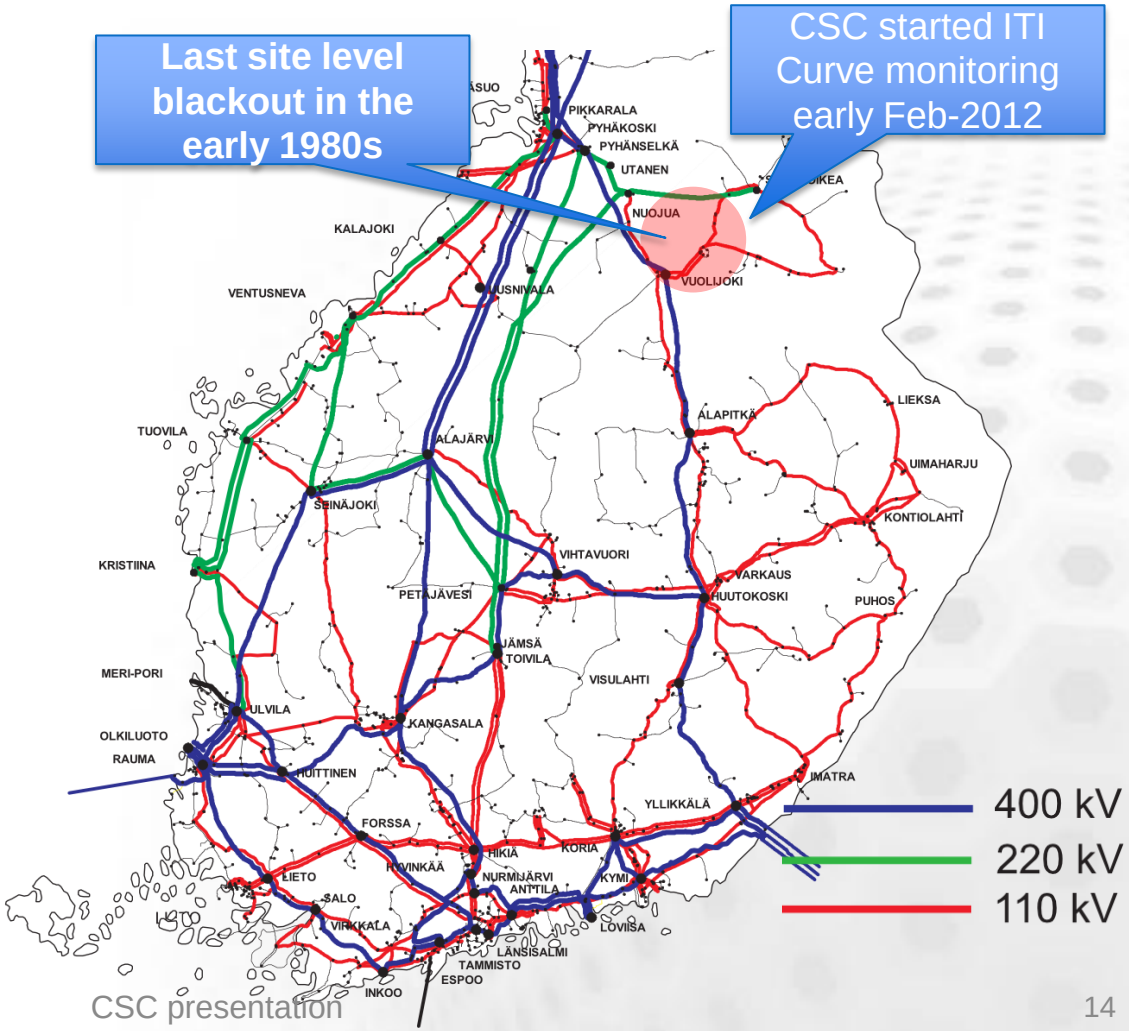
DC1 2005 (500kW/ 1.62 PUE)
DC 2 2008 (800kW/1.38 PUE)
DC 3 2012 (xMW/1.2 PUE)

Vision 2015
CSC – Pioneer in the Sustainable Development of ICT Services


the green grid
get connected to efficient IT
Estimate of
Water-side Economizer
Hours For Data Centers
© 2009 The Green Grid
Number of Available Hours Where:
Wet Bulb Temperature <= 50F (10C)
HOURS
7500
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
7000
6000
5000
4000
3000
2000

Power distribution (FinGrid)



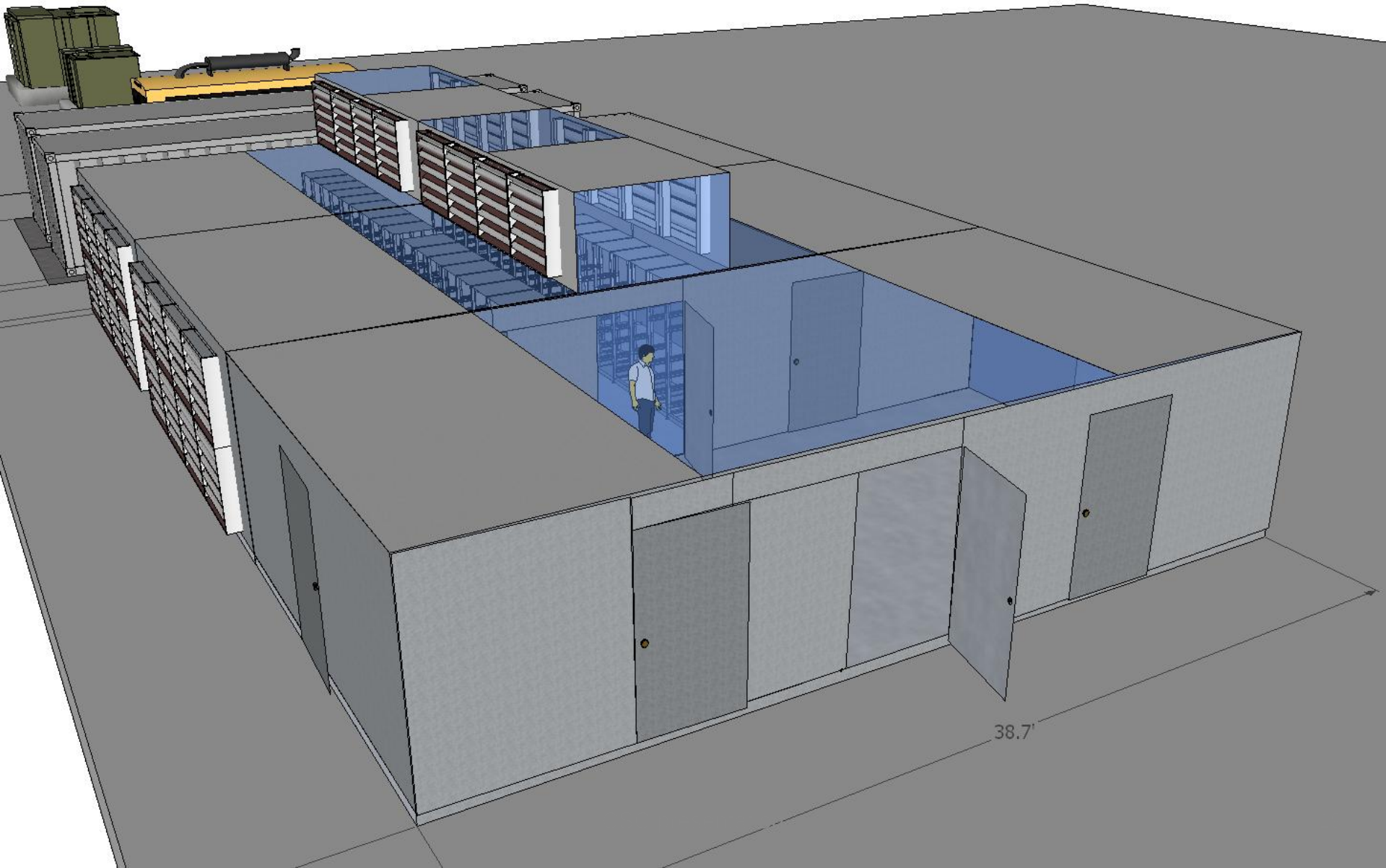


The machine hall

Sisu (Cray) supercomputer housing

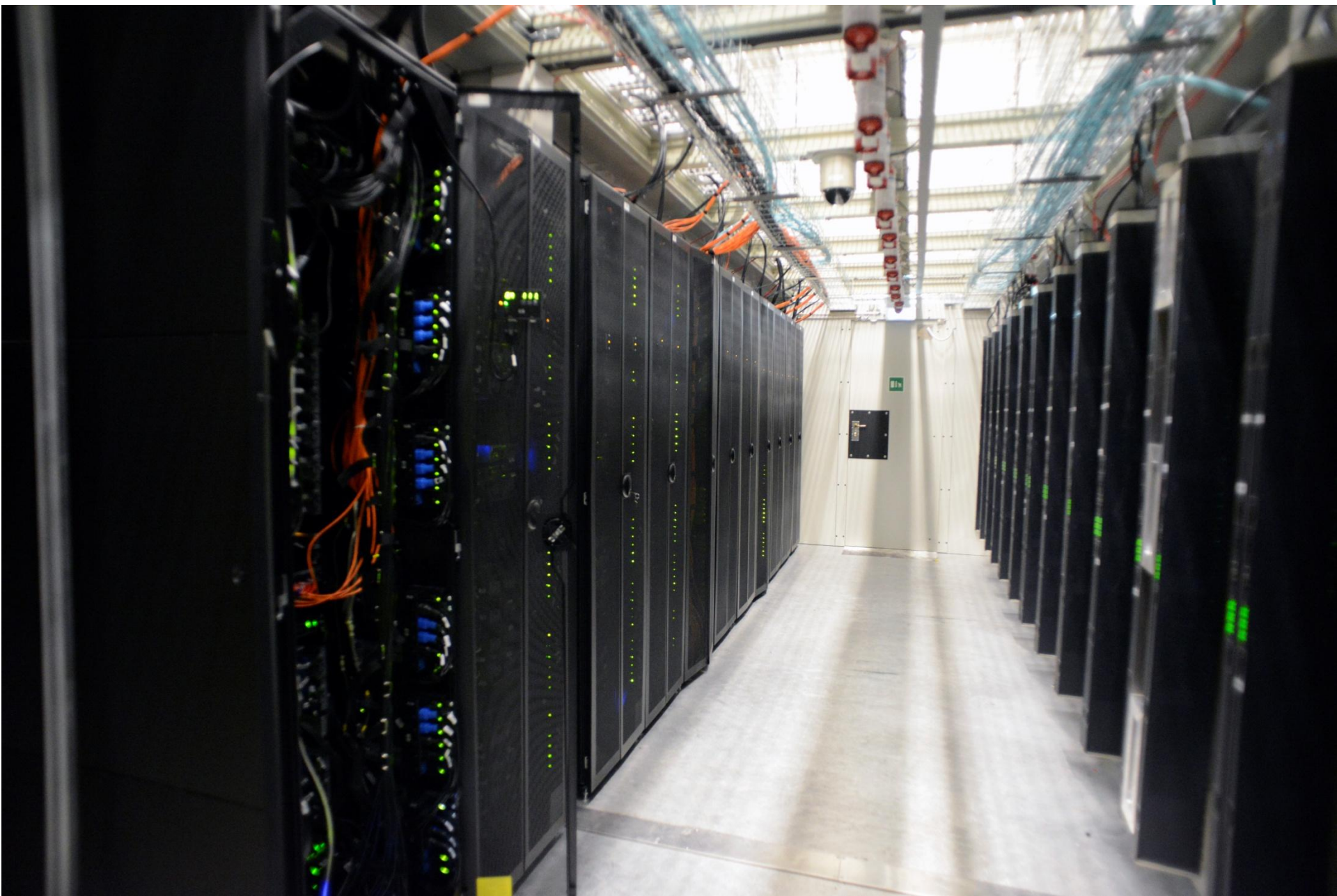
Sisu

SGI Ice Cube R80, hosting Taito (HP)



SGI Ice Cube R80

Taito



Data center specification

- 2.4 MW combined hybrid capacity
- 1.4 MW modular free air cooled datacenter
  - Upgradable in 700 kW factory built modules
  - Order to acceptance in 5 months
  - 35 kW per extra tall racks – 12 kW common in industry
  - PUE forecast < 1.08 ($pPUE_{L2,YC}$)
- 1MW HPC datacenter
  - Optimised for Cray super & T-Platforms prototype
  - 90% Water cooling

CSC NEW SUPERCOMPUTERS

Overview of New Systems

	Phase 1		Phase 2	
	Cray	**HP**	**Cray**	**HP**
Deployment	Done	Done	Probably 2014	
CPU	Intel Sandy Bridge 16 cores @ 2.6 GHz		Next generation processors	
Interconnect	Aries	FDR InfiniBand (56 Gbps)	Aries	EDR InfiniBand (100 Gbps)
Cores	11776	9216	~40000	~17000
Tflops	244 (2x Louhi)	180 (5x Vuori)	1700 (16x Louhi)	515 (15x Vuori)
Tflops total	424 (3.6x Louhi)		2215 (20.7x Louhi)	

IT summary

- Cray XC30 supercomputer (Sisu)
  - Fastest computer in Finland
  - Phase 1: 385 kW, 244 Tflop/s, 16 x 2 GB cores per computing node, 4 x 256 GB login nodes
  - Phase 2: ~1700 Tflop/s
  - Very high density, large racks

IT summary cont.

- HP (Taito)
  - 1152 Intel CPUs
    - 16 x 4 GB cores per node
    - 16 fat nodes with 16 x16 GB cores per node
    - 6 x 64 GB login nodes
  - 180 TFlop/s
  - 30 kW 47 U racks

DataDirect™ NETWORKS

- HPC storage
  - 1 + 1.4 + 1.4 PB of fast parallel storage
  - Supports Cray and HP systems

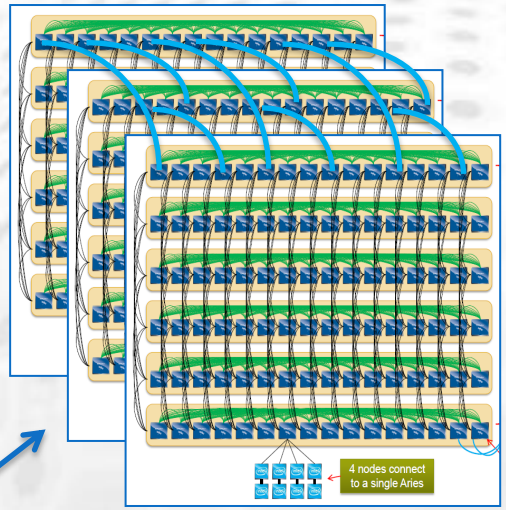
Features

- Cray XC30
  - Completely new system design
    - Departure from the XT* design (2004)
  - First Cray with Intel CPUs
  - High-density water-cooled chassis
    - ~1200 cores/chassis
  - New ”Aries” interconnect

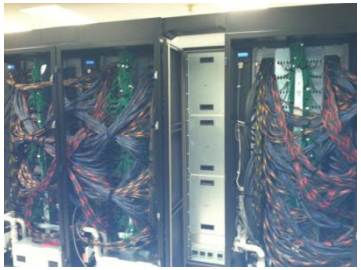
- HP Cluster
  - Modular SL-series systems
  - Mellanox FDR (56 Gbps) Interconnect

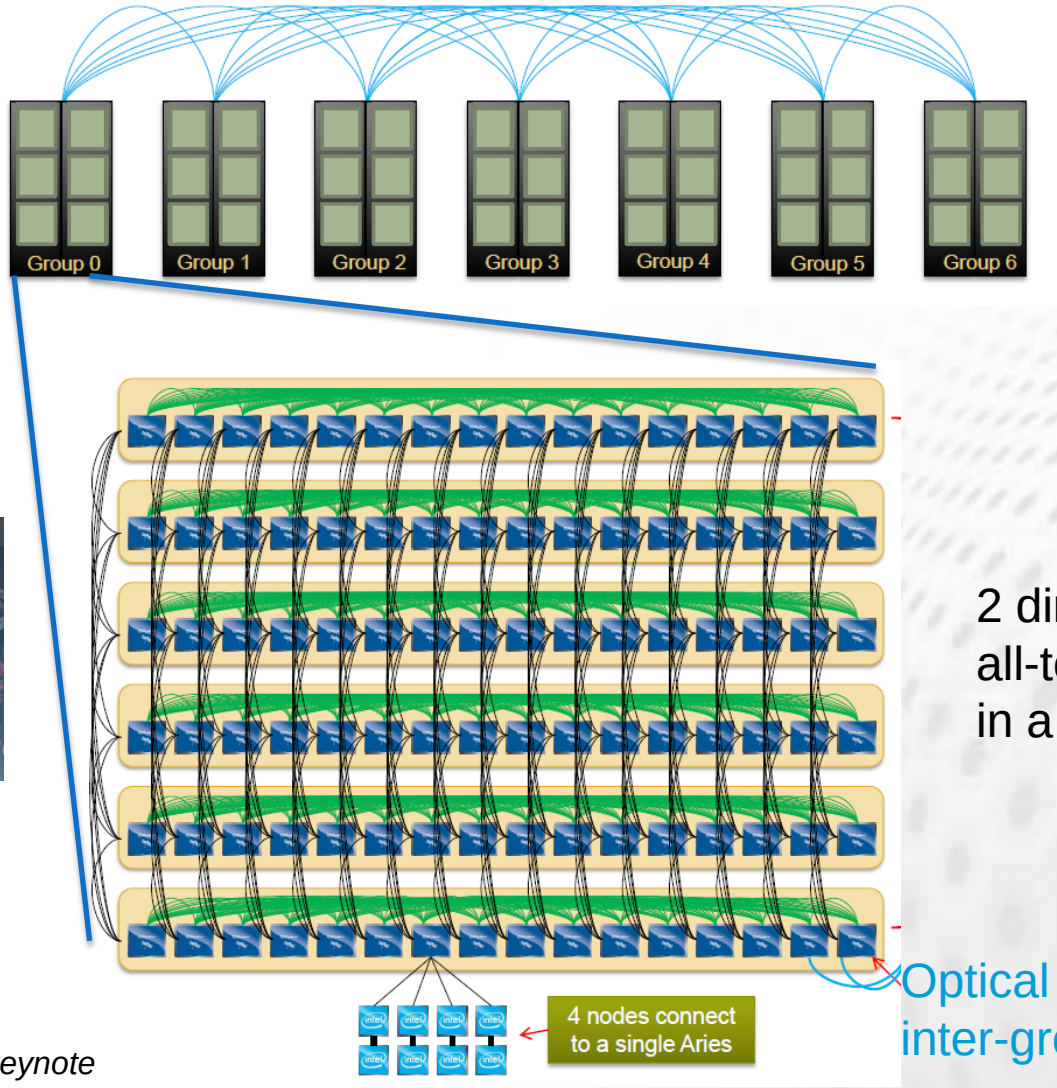
CSC new systems: What's new?

- Sandy Bridge CPUs
  - 4->8 cores/socket
  - ~2.3x Louhi flops/socket
    - 256-bit SIMD instructions (AVX)
- Interconnects
  - Performance improvements
    - Latency, bandwidth, collectives
    - One-sided communication
  - New topologies
    - Cray: "Dragonfly": Islands of 2D Meshes
    - HP: Islands of fat trees



Cray Dragonfly Topology





All-to-all network between groups

2 dimensional all-to-all network in a group

Optical uplinks to inter-group net

Source:
Robert Alverson, Cray
Hot Interconnects 2012 keynote

Cray environment

- Typical Cray environment
- Compilers: **Cray**, Intel and GNU
- Cray mpi, Cray tuned versions of all usual libraries
- SLURM
- Module system similar to Louhi
- Default shell: ***bash*** (previously tcsh)
- Character encoding: UTF-8
  - Latin-15 *alias* iso8859-15 currently on Louhi, Vuori and Hippu will be kept as is

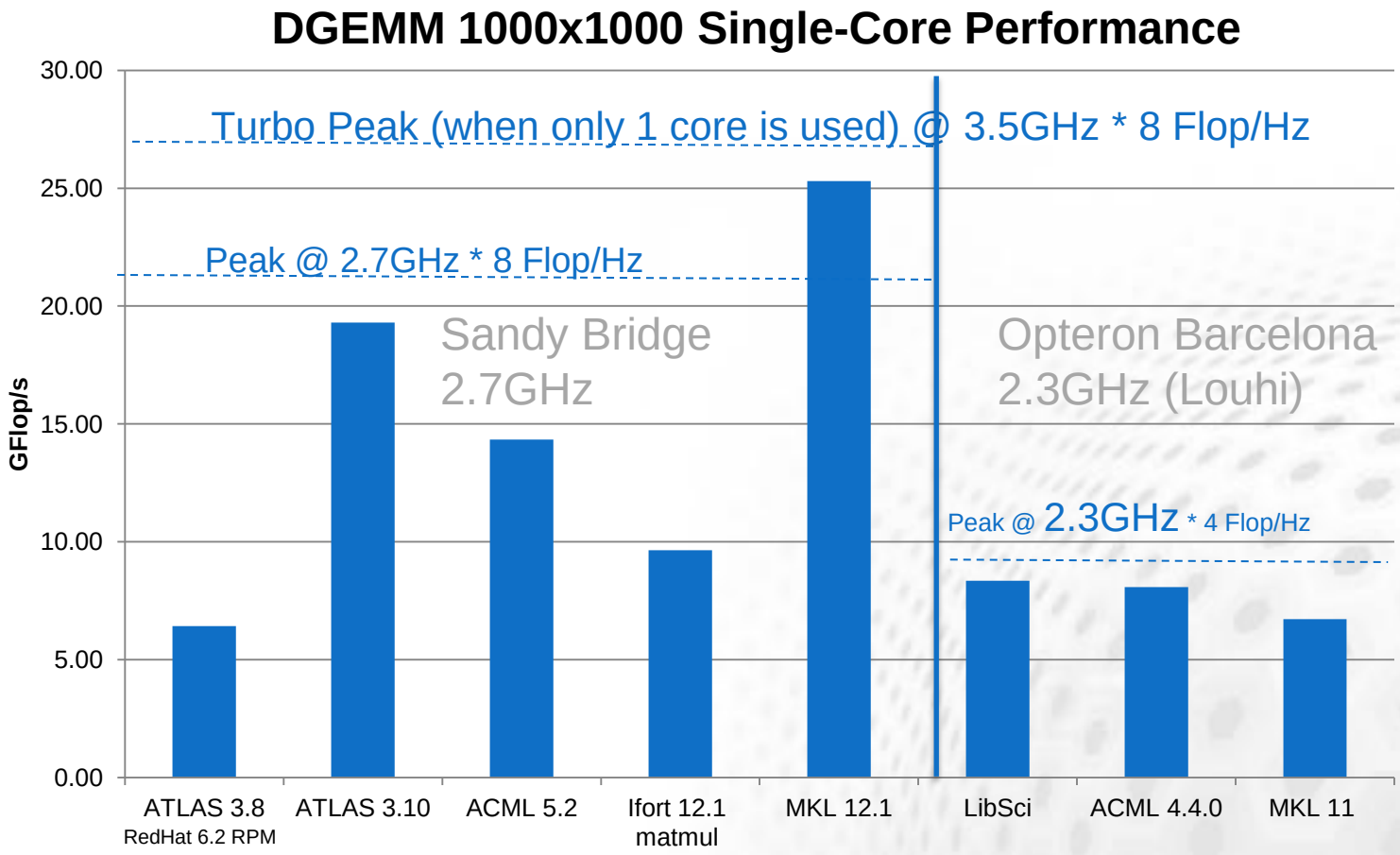
HP Environment

- Compilers: **Intel**, GNU
- MPI libraries: Intel, mvapich2, OpenMPI
- Batch queue: SLURM
- New more robust module system
  - Only compatible modules shown with *module avail*
  - Use *module spider* to see all
- Disk system changes
- Default shell: ***bash*** (used to be tcsh)
- Character encoding: UTF-8

Core development tools

- Intel XE Development Tools
  - Compilers
    - C/C++ (icc), Fortran (ifort), Cilk+
  - Profilers and trace utilities
    - Vtune, Thread checker, MPI checker
  - MKL numerical library
  - Intel MPI library (only on HP)
- Cray Application Development Environment
- GNU Compiler Collection
- Tokens shared between HP and Cray
- TotalView debugger

Performance of numerical libraries



MKL the best choice on Sandy Bridge, for now.

(On Cray, LibSci will likely be a good alternative)

Compilers, programming

- Intel
  - Intel Cluster Studio XE 2013
  - http://software.intel.com/en-us/intel-cluster-studio-xe
- GNU
  - GNU-compilers, e.g. GCC 4.7.2.
  - http://gcc.gnu.org/
- Intel can be used together with GNU
  - E.g. gcc or gfortran + MKL + IntelMPI
- mvapich2 MPI-library also supported
  - It can be used that Intel or GNU

Available applications

- Biggest ready
  - Taito: Gromacs, NAMD, Gaussian, Turbomole, Amber, CP2K, Elmer, VASP
  - Sisu: Gromacs, GPAW, Elmer, VASP
- CSC offers ~240 scientific applications
  - Porting them all is a big task
  - Most if not all (from Vuori) should be available
    - Some installations upon request
  - Do you have priorities?

Porting strategy

- At least recompile
  - Legacy binaries may run, but not optimally
  - Intel compilers preferred for performance
  - Use Intel MKL or Cray LibSci (not ACML!)
    - ***http://software.intel.com/sites/products/mkl/***
  - Use compiler flags (i.e. -xhost -O2 (includes `-xAVX`))
- Explore optimal thread/task placement
  - Intra-node and internode
- Refactor the code if necessary
  - OpenMP/MPI workload balance
  - Rewrite any SSE assembler or intrinsics
- HPC Advisory Council has best practices for many codes
  - *http://www.hpcadvisorycouncil.com/subgroups_hpc_works.php*
- During (and after) pilot usage, share your makefiles and optimization experiences in the wiki

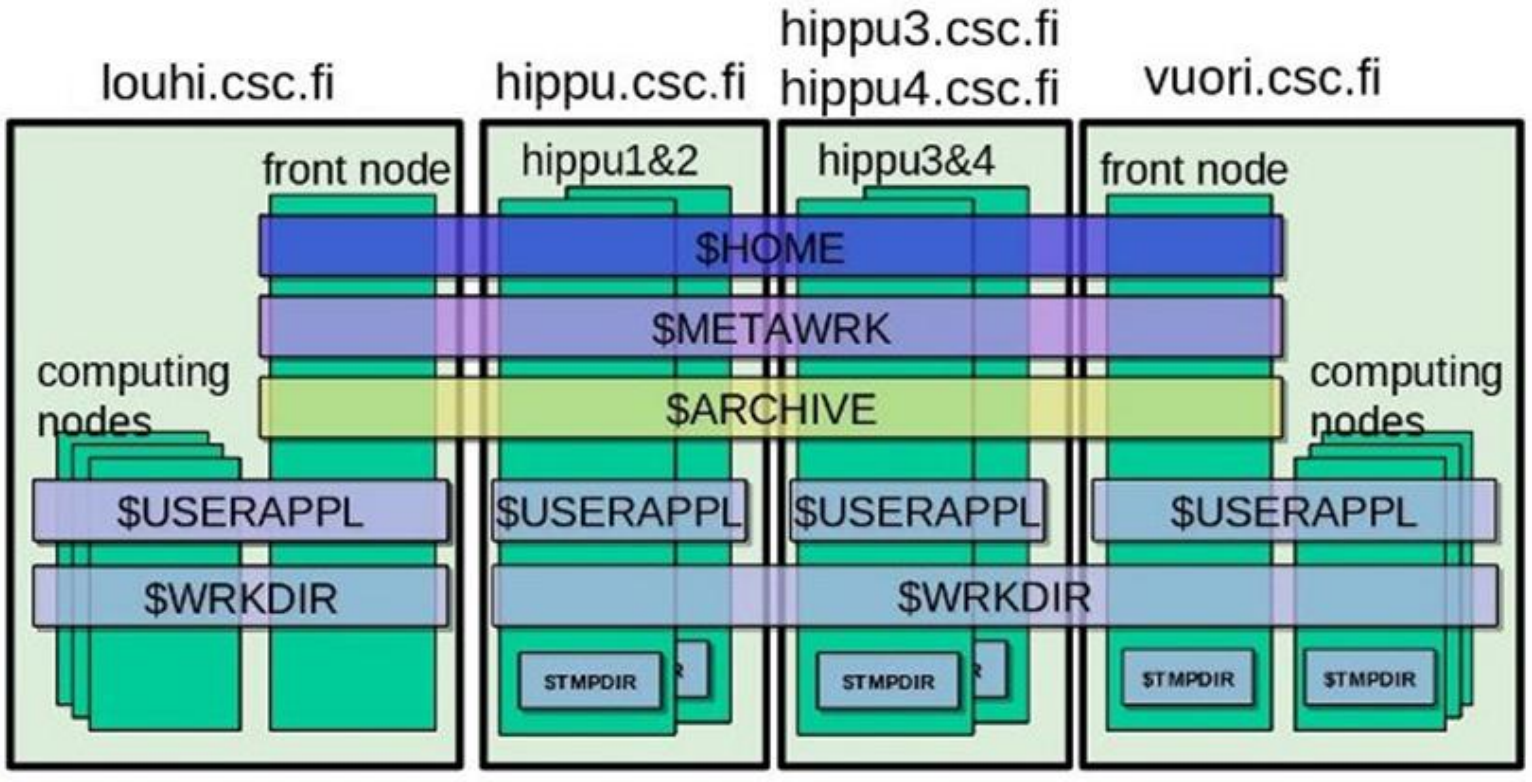
Modules

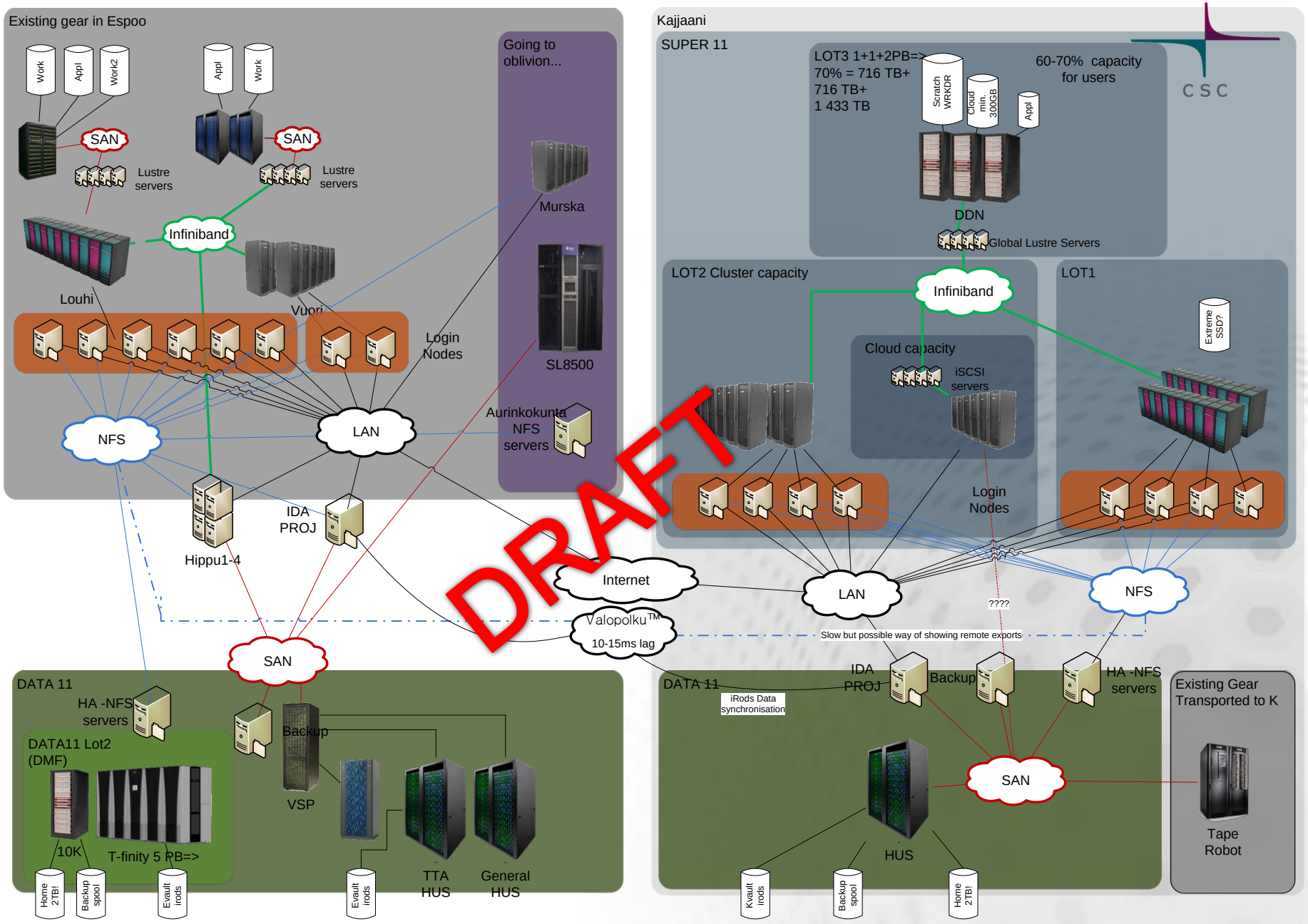
- Some software installations are conflicting with each other
  - For example different versions of programs and libraries
- Modules facilitate the installation of conflicting packages to a single system
  - User can select the desired environment and tools using module commands
  - Can also be done "on-the-fly"

Key differences (Taito vs. Vuori)

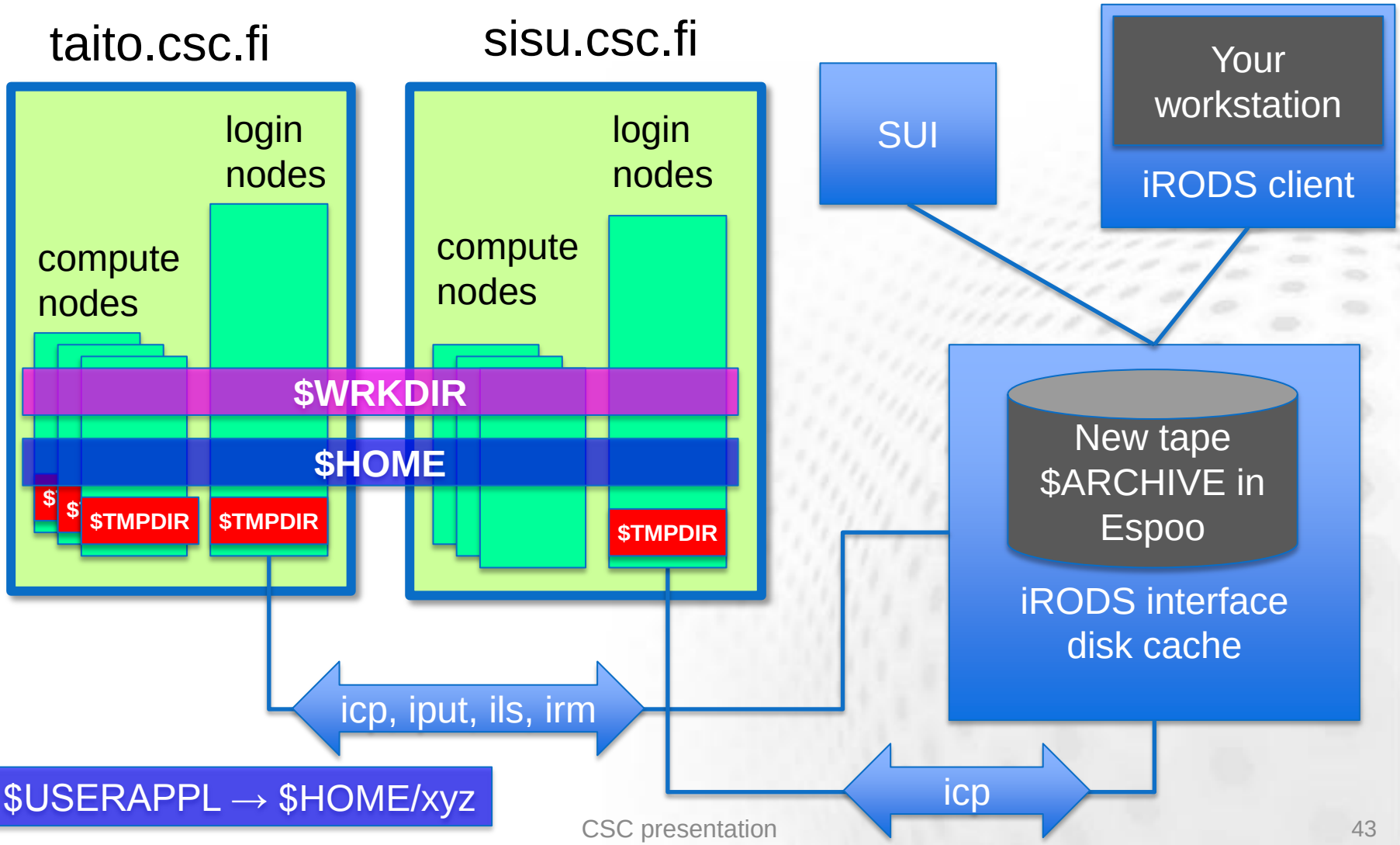
- **`module avail`** shows only those modules that can be loaded to current setup (no conflicts or extra dependencies)
  - Use **`module spider`** to list all installed modules and solve the conflicts/dependencies
- No PrgEnv- modules *(on Taito)*
  - Changing the compiler module switches also MPI and other compiler specific modules

Disks at Espoo




Existing gear in Espoo
Work
Appl
Work2
SAN
Lustre servers
Appl
Work
SAN
Lustre servers
Infiniband
Louhi
Vuori
Login Nodes
NFS
LAN
Going to oblivion...
Murska
SL8500
Aurinkokunta NFS servers
IDA PROJ
Hippu1-4
Kajjaani
SUPER 11
LOT3 1+1+2PB=>
70% = 716 TB+
716 TB+
1 433 TB
Scratch WRKDR
Cloud min. 300GB
Appl
60-70% capacity for users
DDN
Global Lustre Servers
LOT2 Cluster capacity
Infiniband
Cloud capacity
iSCSI servers
LOT1
Extreme SSD?
Login Nodes
Internet
LAN
NFS
????
Slow but possible way of showing remote exports
Valopolku™
10-15ms lag
SAN
DATA 11
HA -NFS servers
DATA11 Lot2 (DMF)
10K
T-finity 5 PB=>
Backup
VSP
Home 2TB!
Backup spool
Evault irods
Evault irods
TTA HUS
General HUS
DATA 11
iRods Data synchronisation
IDA PROJ
Backup
HA -NFS servers
SAN
HUS
Kvault irods
Backup spool
Home 2TB!
Existing Gear Transported to K
Tape Robot
DRAFT

Disks at Kajaani



Disks

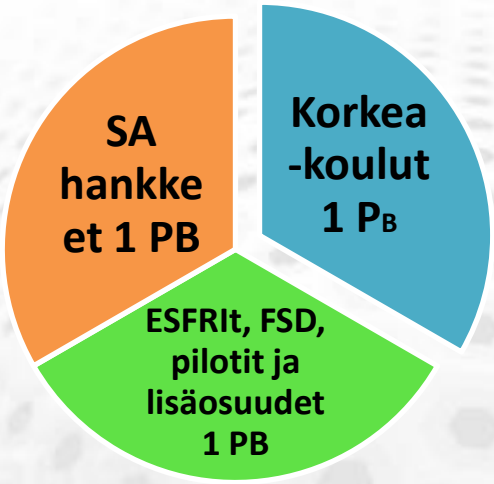
- 3.8 PB on DDN
  - New $HOME directory (on Lustre)
  - $WRKDIR (*not backed up*), soft quota ~ 5 TB
- $ARCHIVE ~1 - 5 TB / user, common between Cray and HP
- Disk space through IDA
  - 1 PB for Univerisities
  - 1 PB for Finnish Academy (SA)
  - 1 PB to be shared between SA and ESFRI
  - more could be requested
- ***/tmp*** (around 1.8 TB) to be used for *compiling codes*

IDA service promise

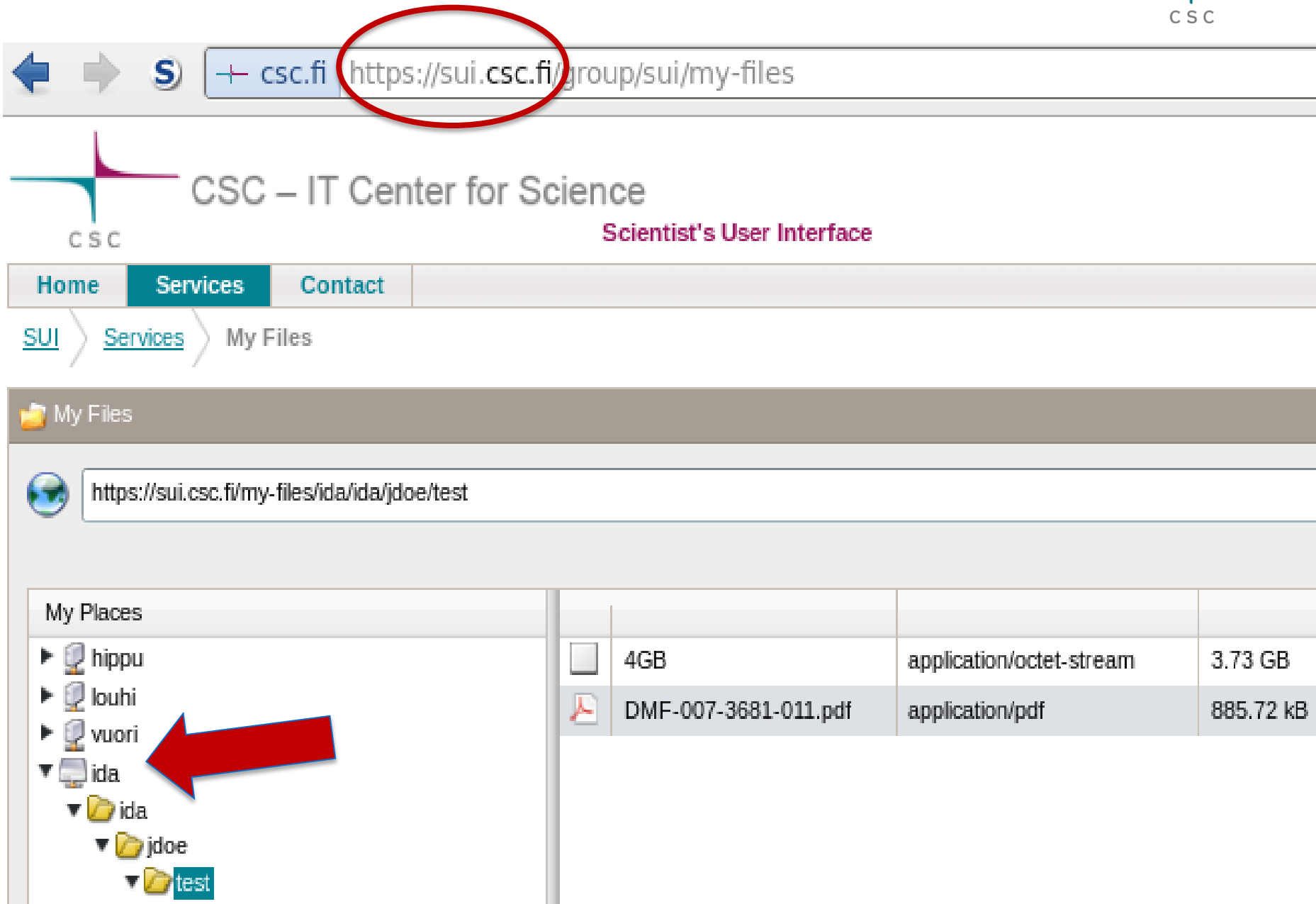
- The IDA service guarantees data storage till at least end of year 2017. By that timeline it will be decided if IDA or some other service solution (for example Long Term Storage solution) continues to serve as data storage. If some other solution is available, the data will be automatically transferred to this service, no interaction needed from users.
- **The IDA service guarantees at least 3 petabytes storage capacity**.
- When transferring data to IDA, some metadata will automatically be attached to the data.
- During this time period, there are no costs to the user within agreed service policy.
- Owners of data can decide on data use, openness and data policy.
- TTA strongly recommends clear ownership and IPR management for all datasets
- After 2017, metadata attached to datasets has to be more extensive than minimum metadata.

Datasets served by TTA

- Projects funded by Finnish Academy (akatemiahankkeet, huippuyksiköt, tutkimusohjelmat and tutkimusinfrastruktuurit)
  - 1 PT capacity
- Universities and Polytechnics
  - 1 PT capacity
- ESFRI-projects (tex. BBMRI, CLARIN)
- Other important research projects via special application process



IDA Interface



Moving files, best practices

- tar & bzip first
- rsync, not scp
  - `rsync -P username@hippu1.csc.fi:/tmp/huge.tar.gz .`
- Blowfish may be faster than AES (CPU bottleneck)
- Funet FileSender (max 50 GB)
  - *https://filesender.funet.fi*
  - Files can be downloaded also with wget

- Consider: SUI, IDA, iRODS, batch-like process, staging
- CSC can help to tune e.g. TCP/IP parameters
  - *http://www.csc.fi/english/institutions/funet/networkservices/pert*
- FUNET backbone 10 Gbit/s

ARCHIVE, dos and don'ts

- Don't put small files in $ARCHIVE
  - Small files waste capacity
  - Less than 10 MB is small
  - Keep the number of files small
  - Tar and bzip files

- Don't use $ARCHIVE for incremental backup (store, delete/overwrite, store, …)
  - Space on tape is not freed up until months or years!
- Maximum file size 300GB
- Default quota 2 TB per user, new likely up to 5 TB
- New ARCHIVE being installed, consider if you really need all your old files. *Transfer from old to new needed*.

Use profiles

- Taito (HP)
- Serial and parallel upto about 256 cores (TBD)
- Sisu (Cray XE30)
- Parallel up to thousands of cores
- Scaling tests

Queue/server policies

- Longrun queue has drawbacks
  - Shorter jobs can be chained
- Apps that can’t restart/write checkpoint?
  - Code you use to run very long jobs?
- Large memory jobs to Hippu/HP big memory nodes
  - Think about memory consumption
- Minimum job size in Cray

Documentation and support

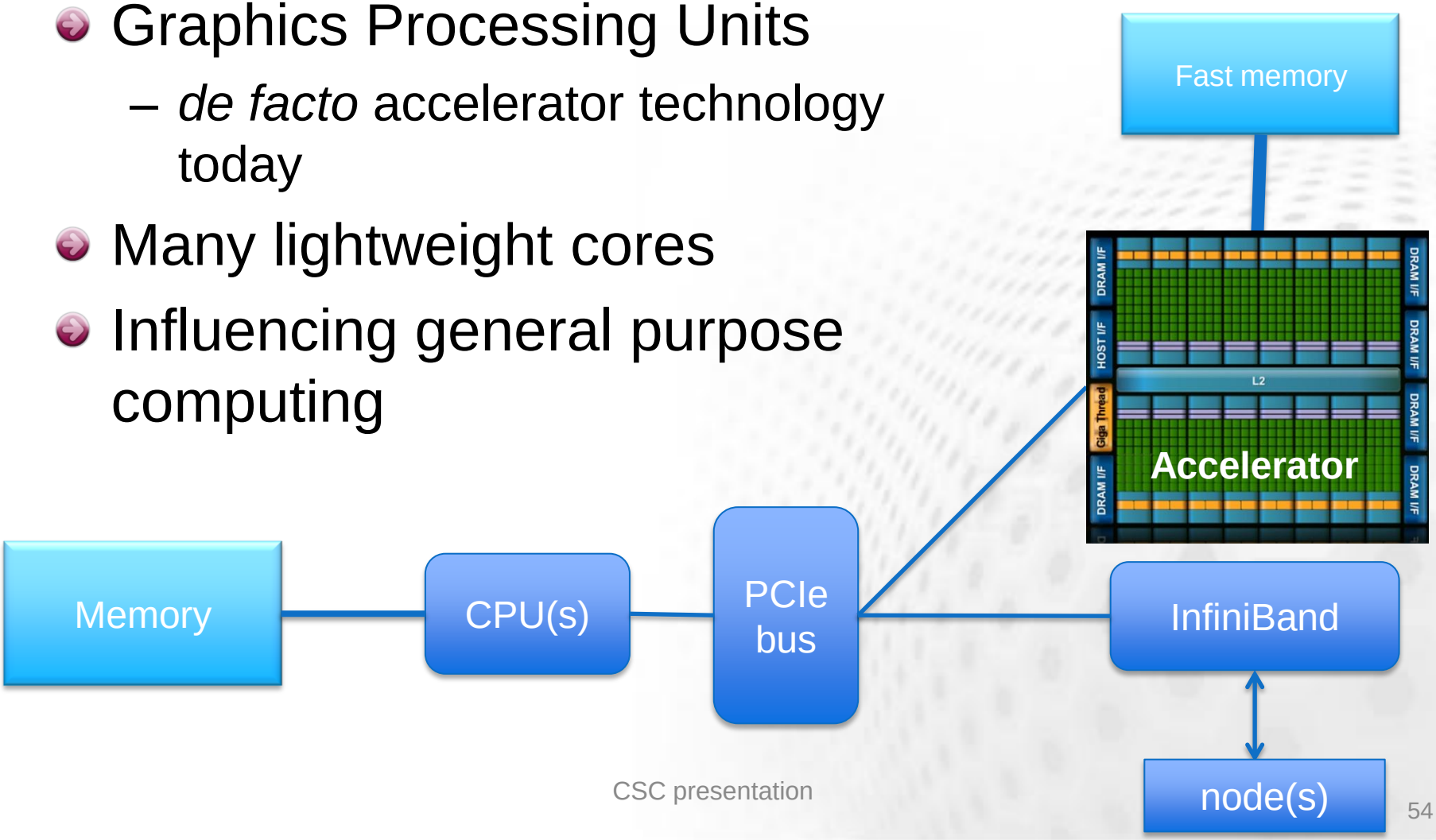
- User manual being built, FAQ here:
  - *https://datakeskus.csc.fi/en/web/guest/faq-knowledge-base*
  - Pilot usage during acceptance tests
- User documetation's link collection
  - *http://www.csc.fi/english/research/sciences/chemistry/intro*
- Porting project
  - All code needs to be recompiled
  - Help available for porting your code
- List of first codes, others added later, some upon request
- User accounts
  - ***HP: recent Vuori users moved automatically***
  - ***Cray: recent Louhi users moved automatically***
  - Others: request from **usermgr@csc.fi** with *current* contact information

Grand Challenges

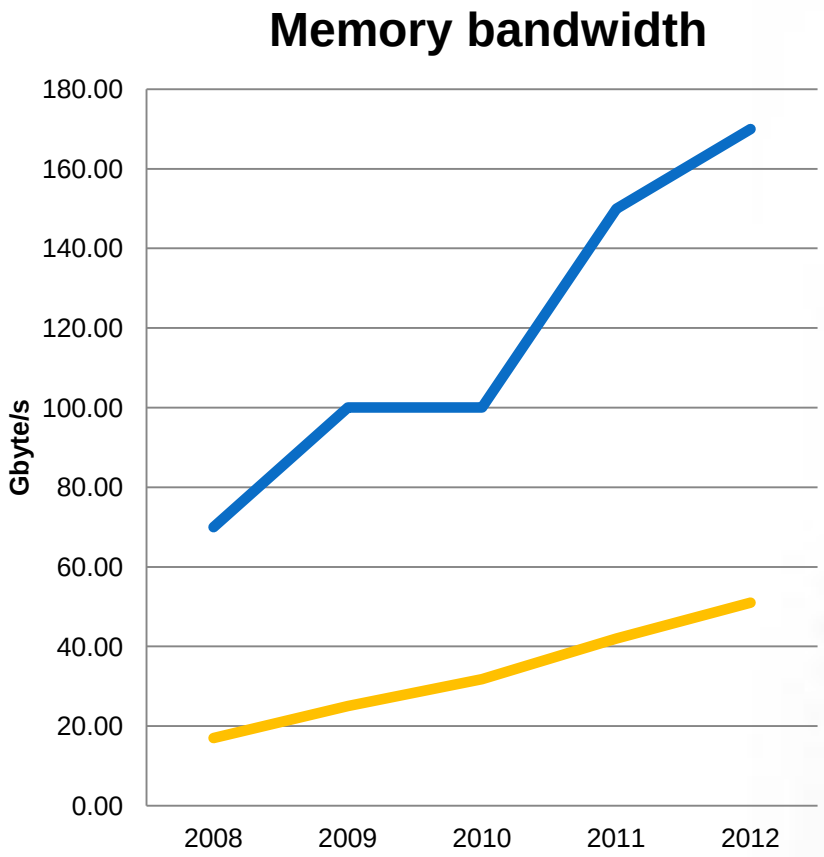
- Normal GC *(in half a year / year)*
  - new CSC resources available for a year
  - no bottom limit for number of cores
- Special GC call (mainly for Cray) *(based on your needs)*
  - possibility for short (day or less) runs with the whole Cray
- Remember also PRACE/DECI
  - http://www.csc.fi/english/csc/news/customerinfo/DECI10callopen

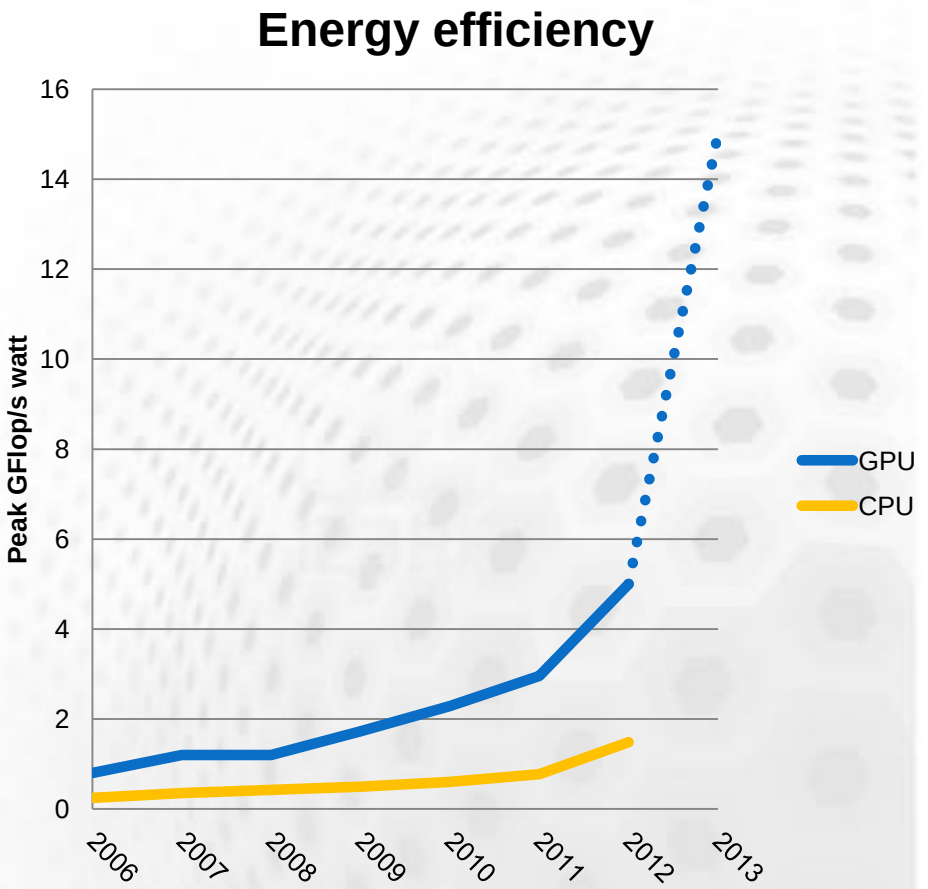
Accelerators

- Add-on processors
- Graphics Processing Units
  - *de facto* accelerator technology today
- Many lightweight cores
- Influencing general purpose computing



Evolution of CPU and GPU performance





Future directions in parallel programming

- **MPI-3** standard being finalized
  - Asynchronous collective communication etc.
- Partitioned Global Address Space (PGAS)
  - Data sharing via global arrays
  - Finally starting to see decent performance
  - Most mature: **Unified Parallel C**, **Co-Array Fortran** (in Fortran 2008), **OpenSHMEM**
- Task Dataflow -based parallel models
  - Splits work into a graph (DAG) of tasks
  - **SmpSs**, **DAGUE**, **StarPU**

CSC RESOURCES AVAILABLE FOR RESEARCHERS

Currently available computing resources

- Massive computational challenges: Sisu (Louhi being decommisioned)
  - > 10 000 cores, >23TB memory
  - Theoretical peak performance > 240 Tflop/s
- HP-cluster Taito (Vuori by the end 2013)
  - Small and medium-sized tasks
  - Theoretical peak performance 180 Tflop/s (40)
- Application server Hippu
  - Interactive usage, without job scheduler
  - Postprocessing, e.g. vizualization
- FGI

Novel resources at CSC

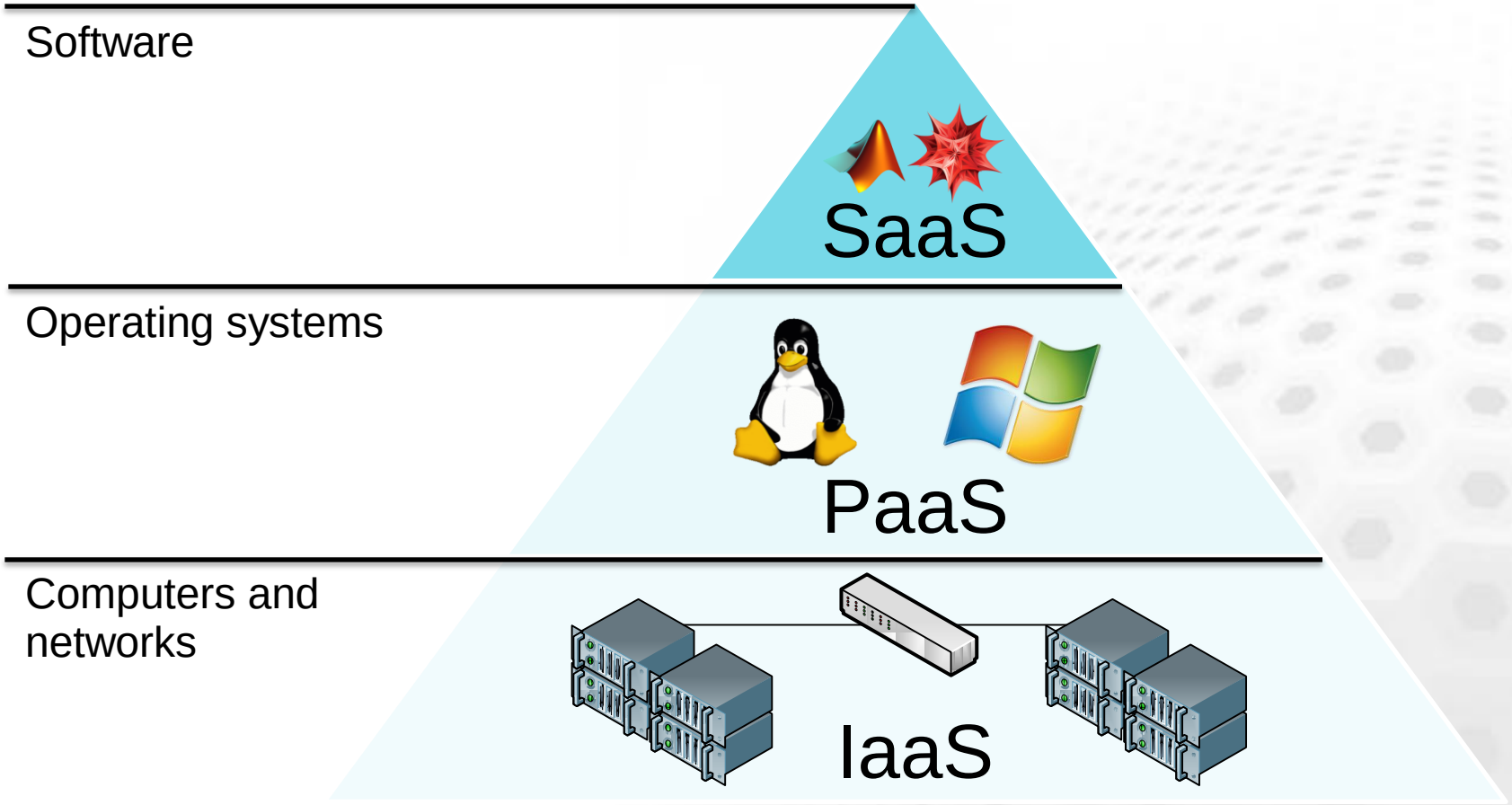
- Production (available for all Finnish researchers)
  - **Vuori**: 8 Tesla GPU nodes
  - **FGI**: 88 GPUs (44 Tesla 2050 + 44 Tesla 2090)
    - GPU nodes located at **HY**, **Aalto**, **ÅA**, **TTY**
- Testing
  - **Mictest**: Intel MIC nodes
    - Several cards
  - **Tunturi:** Sandy Bridge node, cluster (*for CSC experts*)
    - Porting to AVX instruction set

Old capacity decommissions

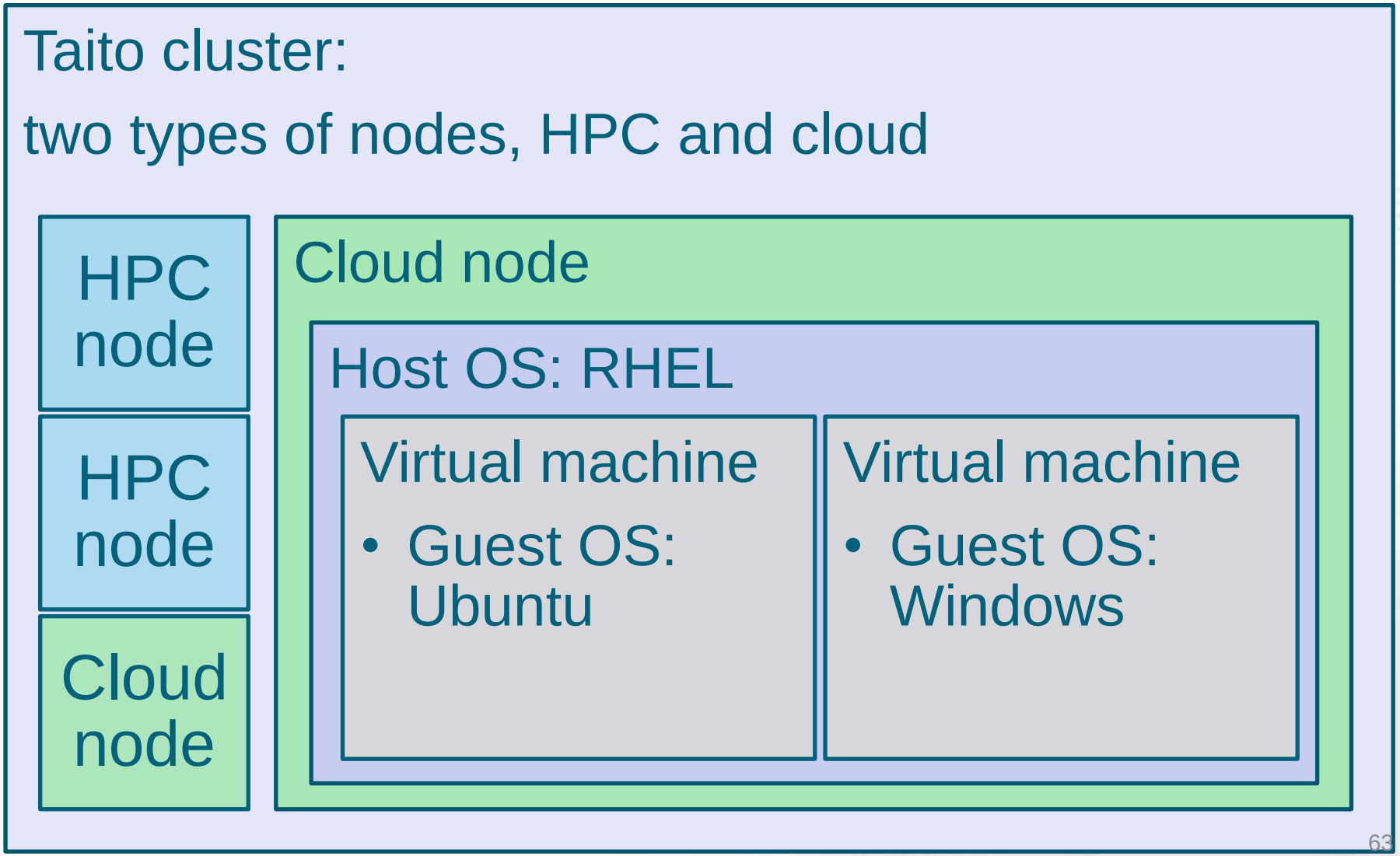
- Louhi to be decommissioned
  - guaranteed access likely by May 1, 2013
  - *communicated via Louhi's MOTD (message of the day), louhi-users mailing list, customer bulletin*
- Vuori decommission by the end of 2013

CSC cloud services

Three service models of cloud computing



Example: Virtualization in Taito



Traditional HPC vs. IaaS

	Traditional HPC environment	Cloud environment Virtual Machine
Operating system	Same for all: CSC's cluster OS	Chosen by the user
Software installation	Done by cluster administrators Customers can only install software to their own directories, no administrative rights	Installed by the user The user has admin rights
User accounts	Managed by CSC's user administrator	Managed by the user
Security e.g. software patches	CSC administrators manage the common software and the OS	User has more responsibility: e.g. patching of running machines
Running jobs	Jobs need to be sent via the cluster's Batch Scheduling System (BSS)	The user is free to use or not use a BSS
Environment changes	Changes to SW (libraries, compilers) happen.	The user can decide on versions.
Snapshot of the environment	Not possible	Can save as a Virtual Machine image
Performance	Performs well for a variety of tasks	Very small virtualization overhead for most tasks, heavily I/O bound and MPI tasks affected more

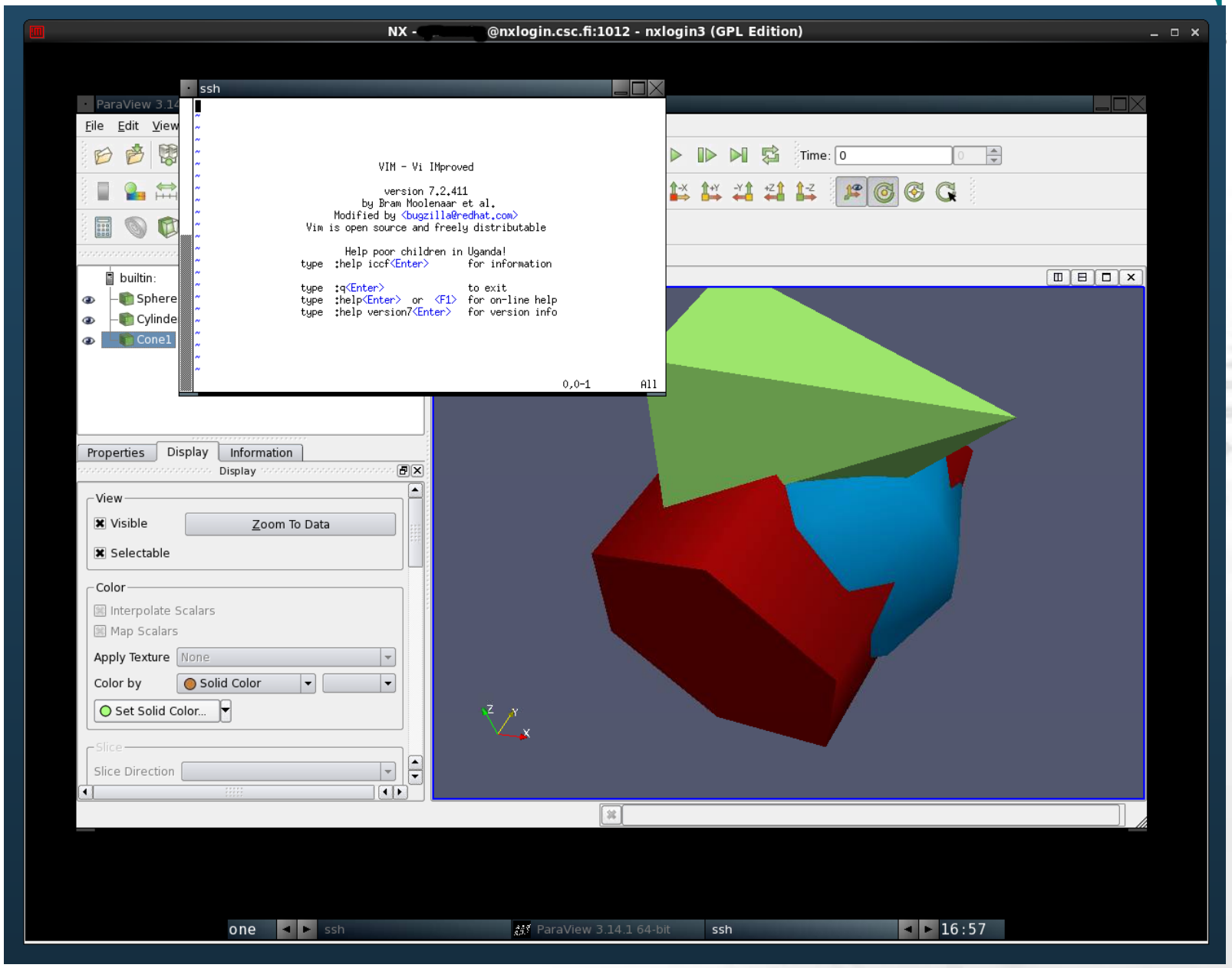
Cloud: Biomedical pilot cases

- Several pilots (~15)
- Users from several institutions, e.g. University of Helsinki, Finnish Institute for Molecular Medicine and Technical University of Munich
- Many different usage models, e.g.:
  - Extending existing cluster
  - Services run on CSC IaaS by university IT department for end users (SaaS for end users)

NX for remote access

- Optimized remote desktop access
  - Near local speed application responsiveness over high latency, low bandwidth links
- Customized launch menus offer direct access CSC supported applications
- Working session can saved and restored at the next login
- Further information: *http://www.csc.fi/english/research/software/freenx*

NX screenshot



Customer training

- Taito (HP)
  - ongoing: Taito workshop
- Sisu (Cray)
  - May 14 - 17 (for all users, a PATC course, i.e. expecting participants from other countries too)
  - December 2013 (likely) MIC programming
- CSC courses: *http://www.csc.fi/courses*
- CSC HPC Summer School; Spring, Winter Schools

Engineering internship, location St. Paul (USA); Duration: 3-12 months

- Cray (in collaboration with CSC) is offering an internship position at Cray for a Finnish student
  - https://skills.csc.fi/rekry3/jsp/joblist/ja_job_list.faces
- The deadline for the applications is May 1, 2013.

How to prepare for new systems

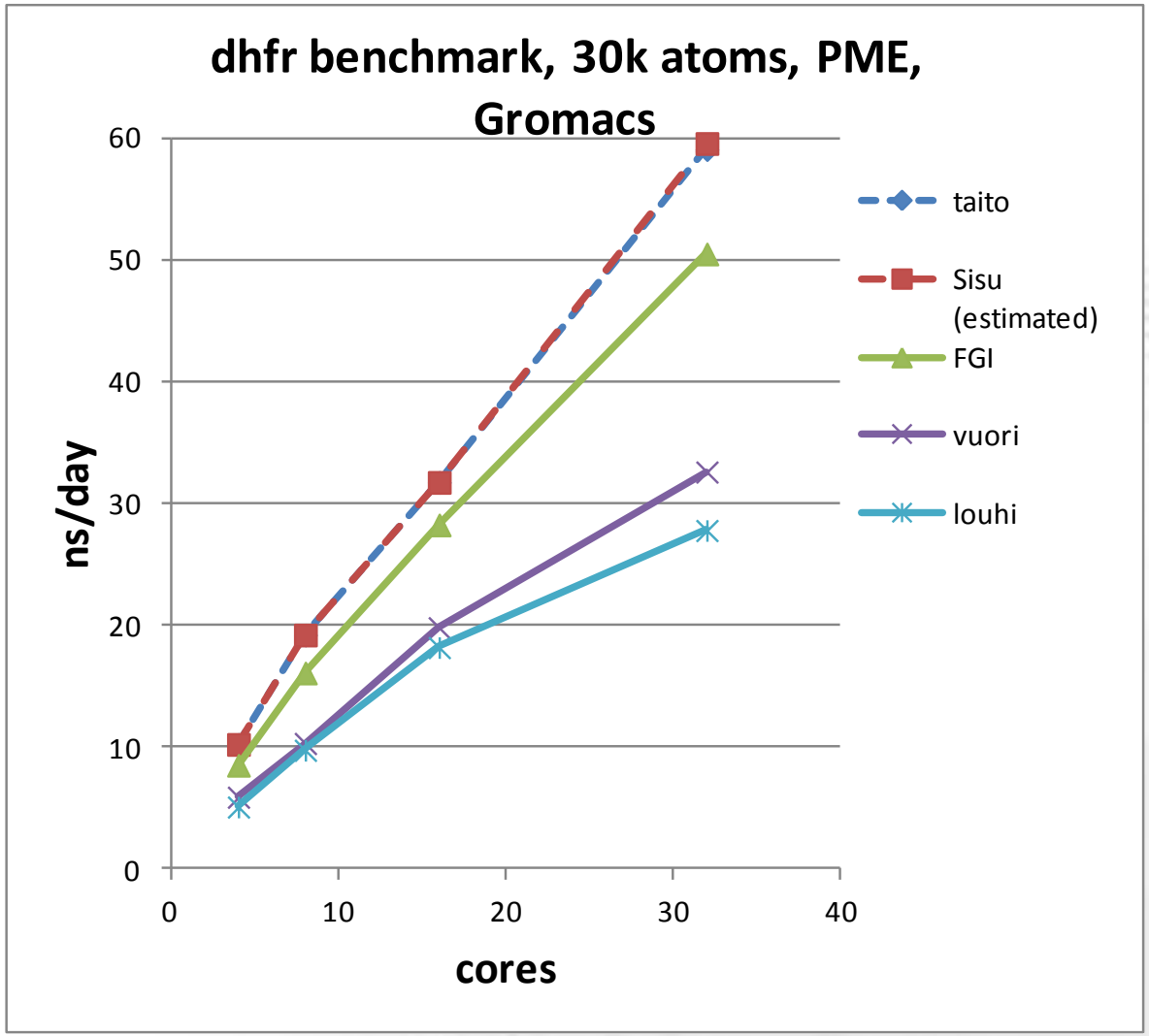
- Participate in system workshops
- Try Intel/GNU compiler in advance, PGI upon request
- Check if your scripts/aliases need fixing (***bash***)
- A lot of resources available in the beginning: prepare ahead what to run!
- The traditional wisdom about good application performance will still hold
  - Experiment with all compilers and pay attention on finding good compiler optimization flags
  - Employ tuned numerical libraries wherever possible
  - Experiment with settings of environment variable that control the MPI library
  - Mind the I/O: minimize output, checkpoint seldom

Sisu&Taito vs. Louhi&Vuori vs. FGI vs. Local Cluster

	Sisu&Taito (Phase 1)	Louhi&Vuori	FGI	Merope
Availability	Available	Available	Available	Available
CPU	Intel Sandy Bridge, 2 x 8 cores, 2.6 GHz, Xeon E5-2670	AMD Opteron 2.3 GHz Barcelona and 2.7 GHz Shanghai / 2.6 GHz AMD Opteron and Intel Xeon	Intel Xeon, 2 x 6 cores, 2.7 GHZ, X5650	
Interconnect	Aries / FDR IB	SeaStar2 / QDR IB	QDR IB	
Cores	11776 / 9216	10864 / 3648	7308	748
RAM/core	2 / 4 GB 16x 256GB/node	1 / 2 / 8 GB	2 / 4 / 8 GB	4 / 8 GB
Tflops	244 / 180	102 / 33	95	8
GPU nodes	in Phase2	- / 8	88	6
Disc space	2.4 PB	110 / 145 TB	1+ PB	100 TB

Conclusions

- Performance comparison
  - Per core performance
    ~2 x compared
    to Vuori/Louhi
  - Better interconnects enhance scaling
  - Larger memory
  - Smartest collective communications
- The most powerful computer(s) in Finland



Round robin

Round robin

- HIP and CSC
- What are your research interest?
  - How CSC can help?
  - Special libraries/tools?
- Queue length: 3 days enough?
  - Codes that can't checkpoint?
- Is memory an issue for you?
  - 256 GB/nodes usage policy?
- Applying for Grand Challenge?
  - Special Grand Challenge?
- Need to move a lot of files? (from where?)
- Interested in GPGPU/MICs? Which code?