CSC resources intro:

most powerful computational resources in Finland

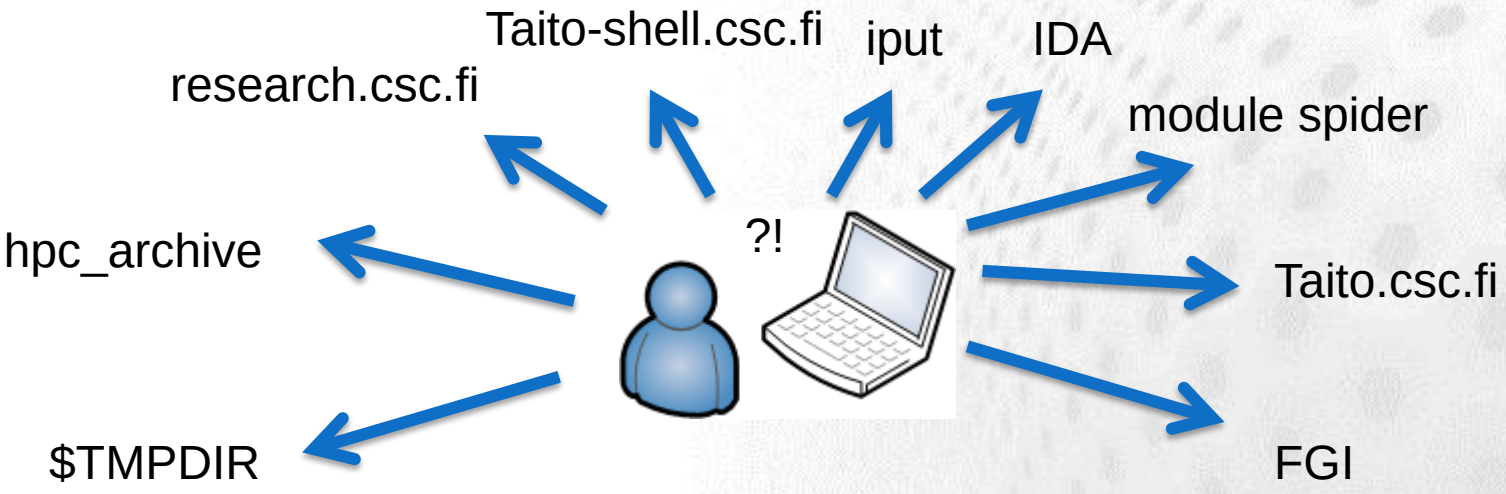
Tomasz Malkiewicz
CSC – IT Center for Science Ltd.

Outline

- Intro
  - Why supercomputers?
  - CSC at glance
  - Kajaani Datacenter
- Finland's supercomputers
  - *Sisu* (Cray XC30)
  - *Taito* (HP cluster)
- ***Live demos/hands-on 1 (Taito)***
- CSC resources available for researchers
- ***Live demos/hands-on 2 (Taito)***

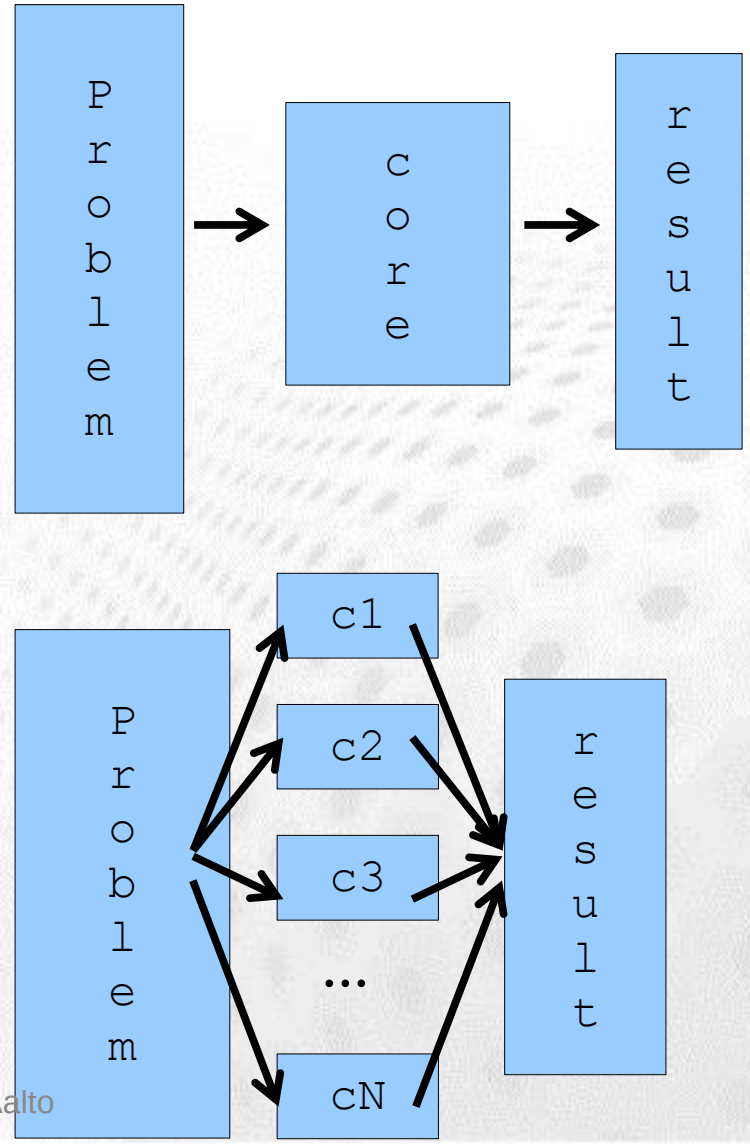
Learning targets

- Be able to use Taito supercluster
  - Create a batch job script, submit a job and view the results
- Get acquainted with what is available
  - Know how to choose right server (resource)



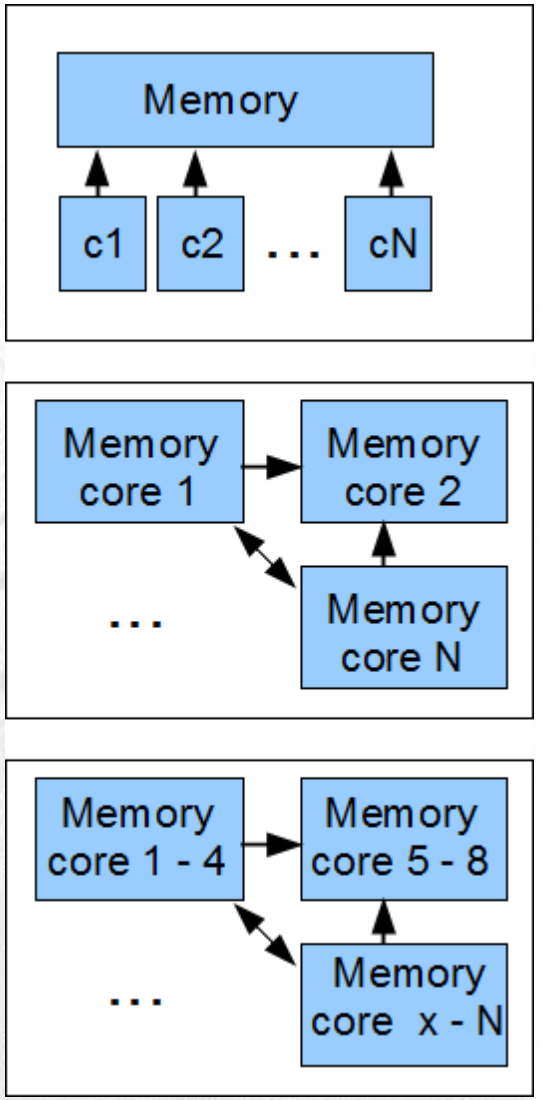
Supercomputing: serial and parallel processing

- Serial computing
  - single processing unit (core) is used for solving a problem
  - single task performed at once
- Parallel computing
  - multiple cores are used for solving a problem
  - problem is split into smaller subtasks
  - multiple subtasks are performed simultaneously



Types of parallel computers

- Shared memory
  - all the cores can access the whole memory
- Distributed memory
  - all the cores have their own memory
  - communication is needed in order to access the memory of other cores
- Current supercomputers combine the distributed memory and shared memory approaches

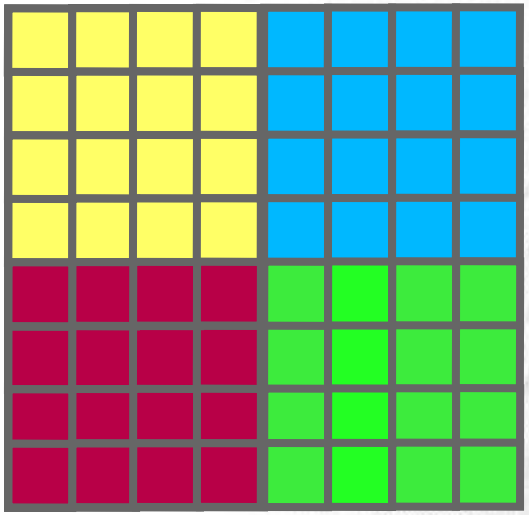


Data parallelism

- Data is distributed to processor cores
- Each core performs (nearly) identical tasks with different data
- Example: summing the elements of a 2D array

core 1: ∑=

core 2: ∑=



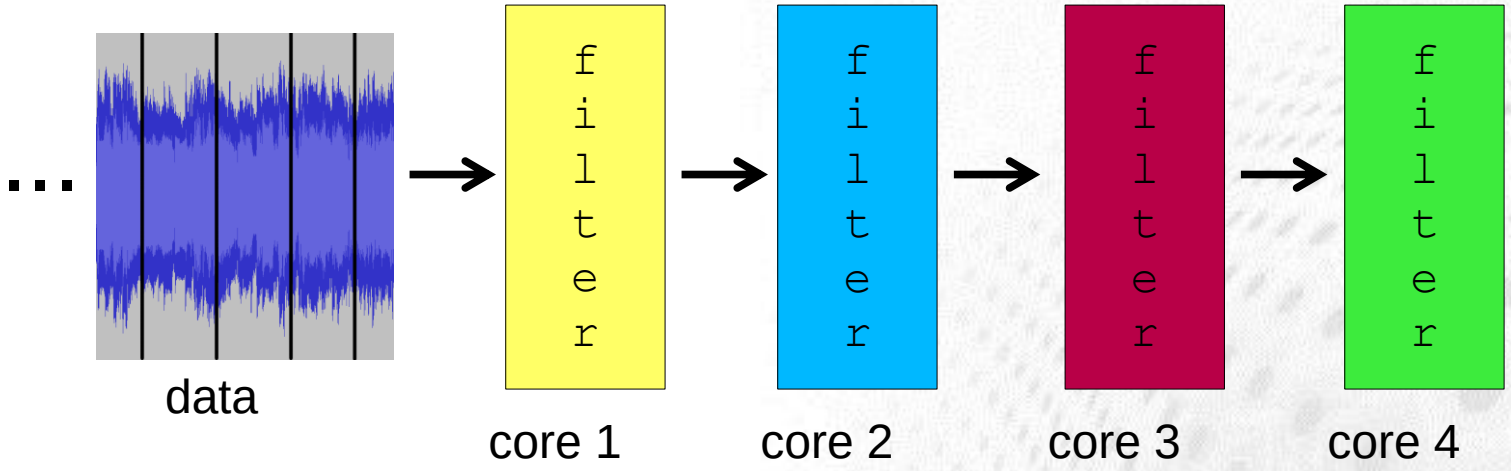
core 3: ∑=

core 4: ∑=

- Each core sums it's part of the array
- The individual sums have to be combined in the end

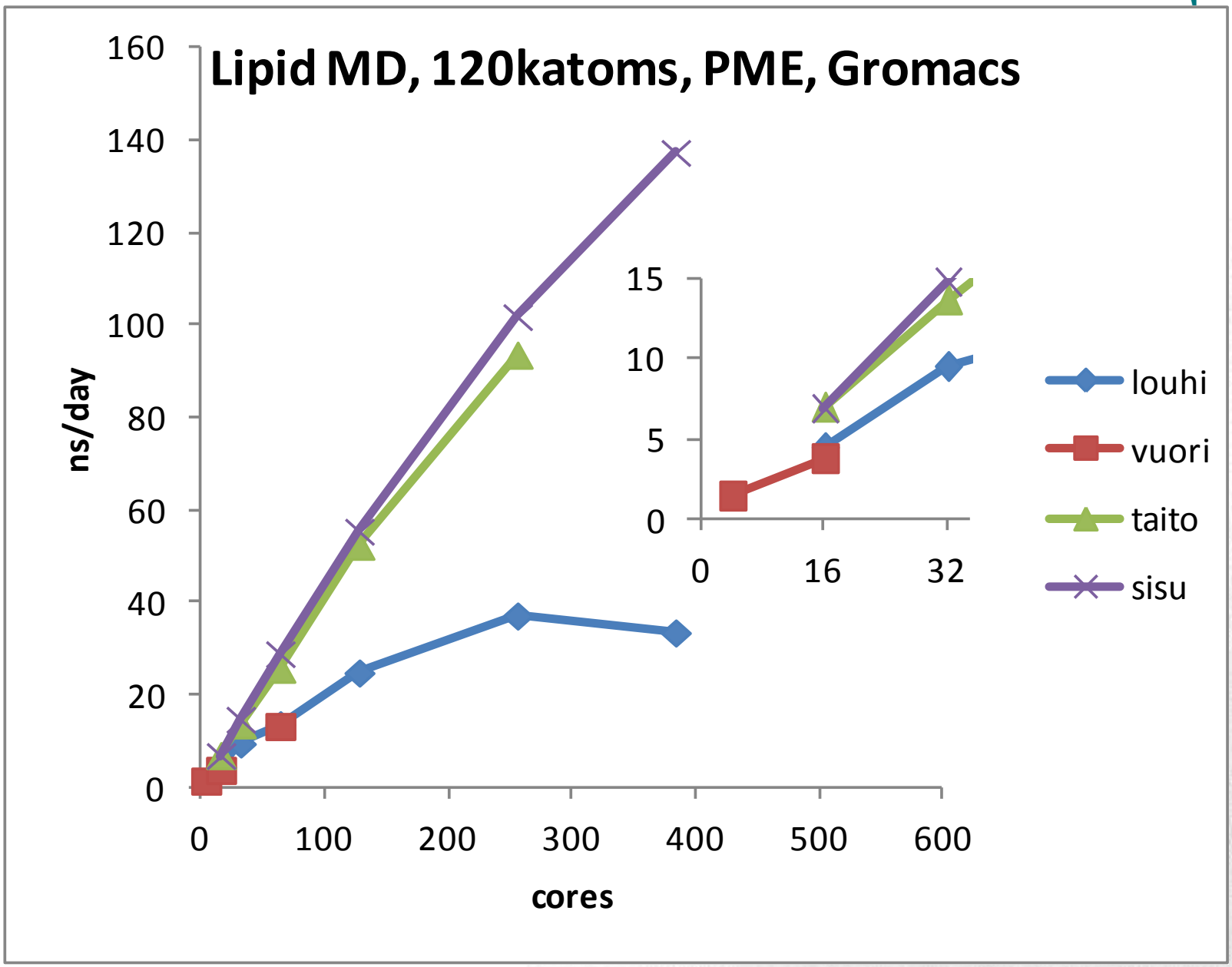
Task parallelism

- Different cores perform different tasks with the same or different data
- Example: signal processing, four filters as separate tasks

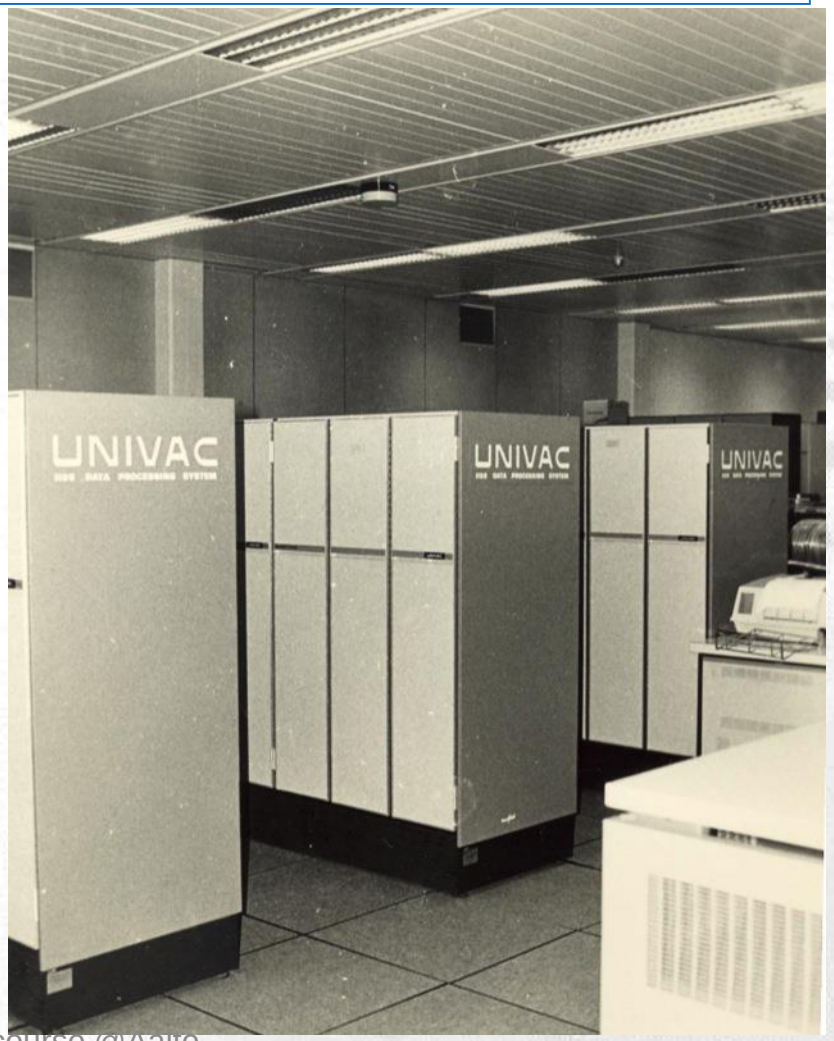


- Data is processed as segments
- Core 2 obtains a segment after core 1 has processed it; core 1 starts to process a new segment
- When the first segment gets to core 4, all cores are busy

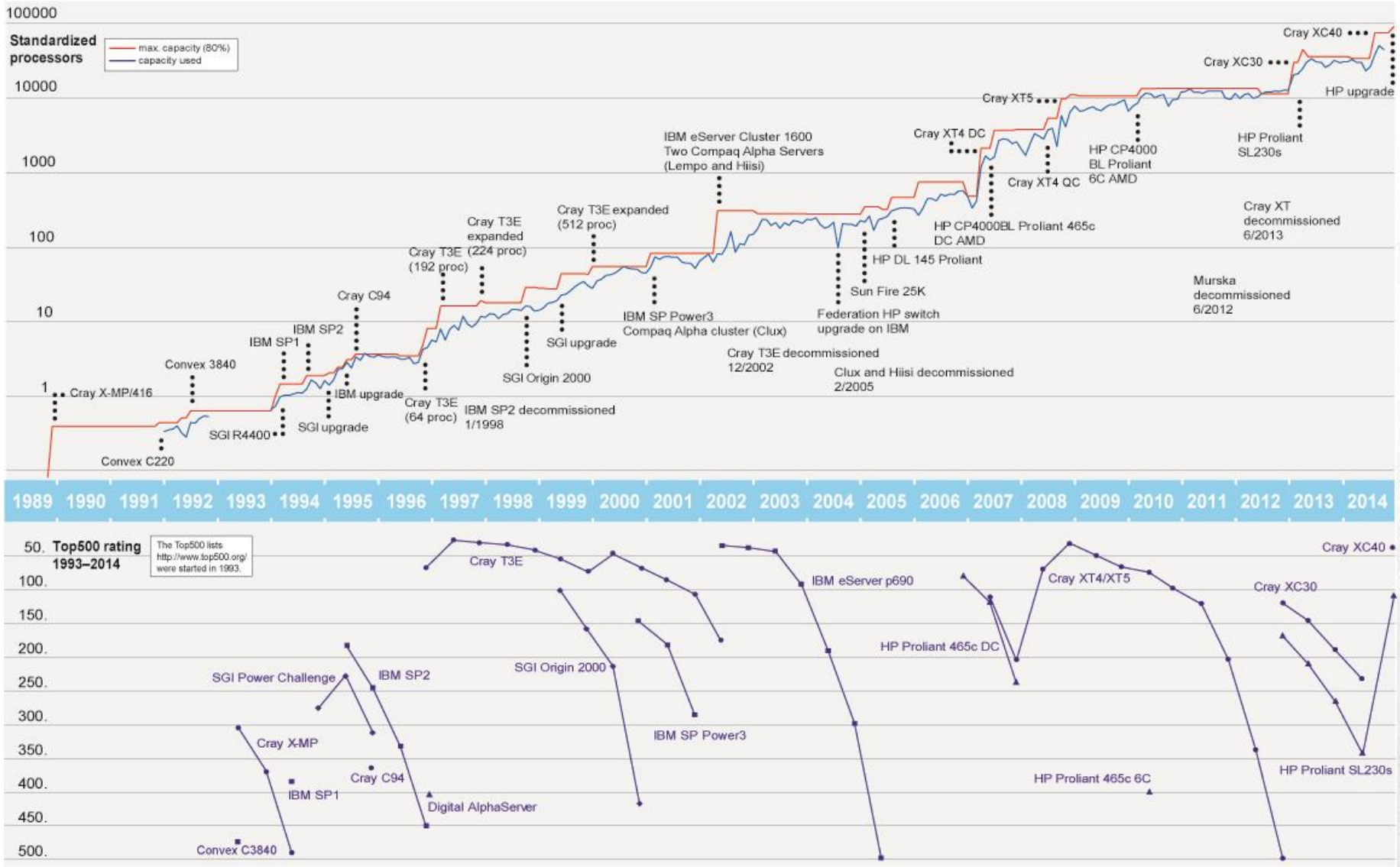
Why supercomputers?



CSC and High Performance Computing



CSC Computing Capacity 1989–2014

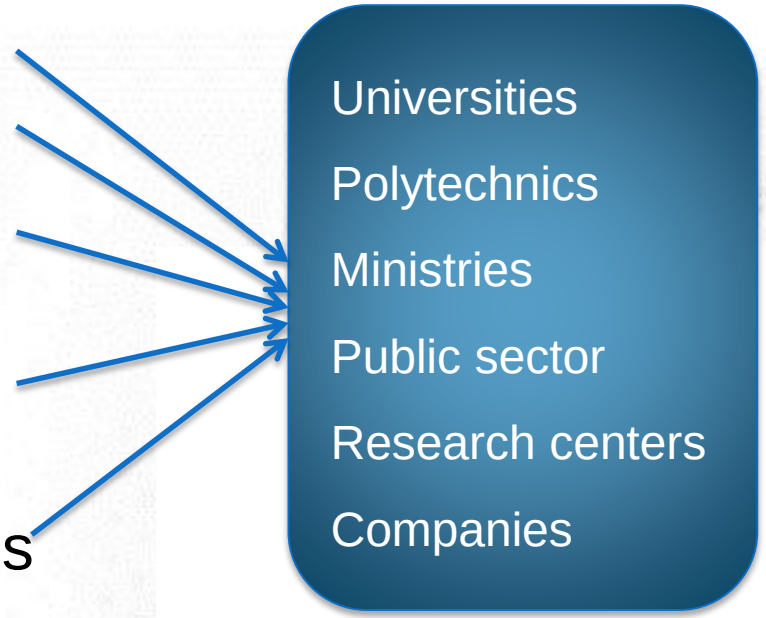


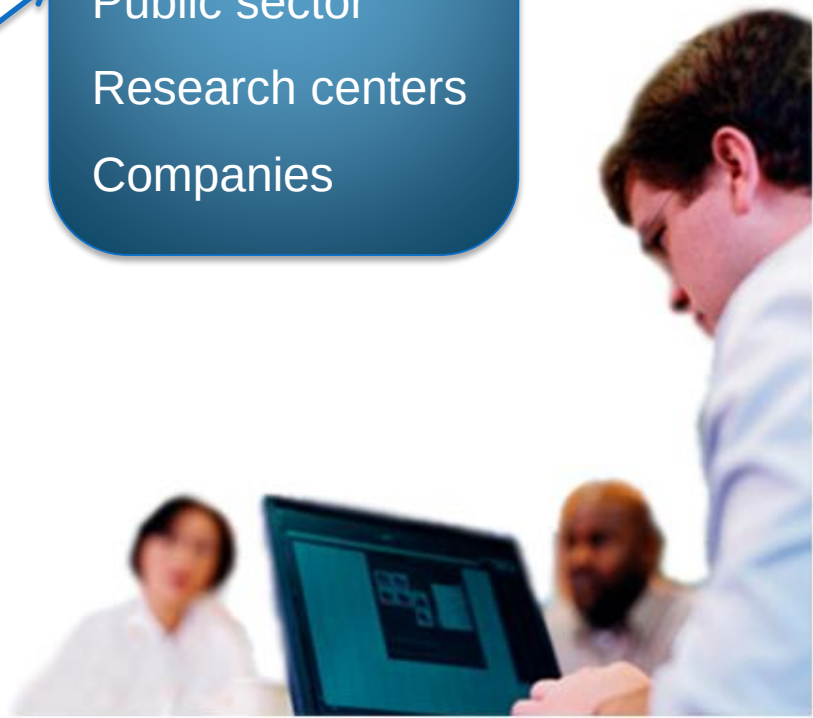
CSC at glance

- Founded in 1971
  - technical support unit for Univac 1108
- Connected Finland to Internet in 1988
- Reorganized as a company, CSC – Scientific Computing Ltd. in 1993
- All shares to the Ministry of Education and Culture of Finland in 1997
- Operates on a *non-profit* principle
- Facilities in Espoo and Kajaani
- Staff ~270 people

CSC's Services

- FUNET Services
- Computing Services
- Application Services
- Data Services for Science and Culture
- Information Management Services





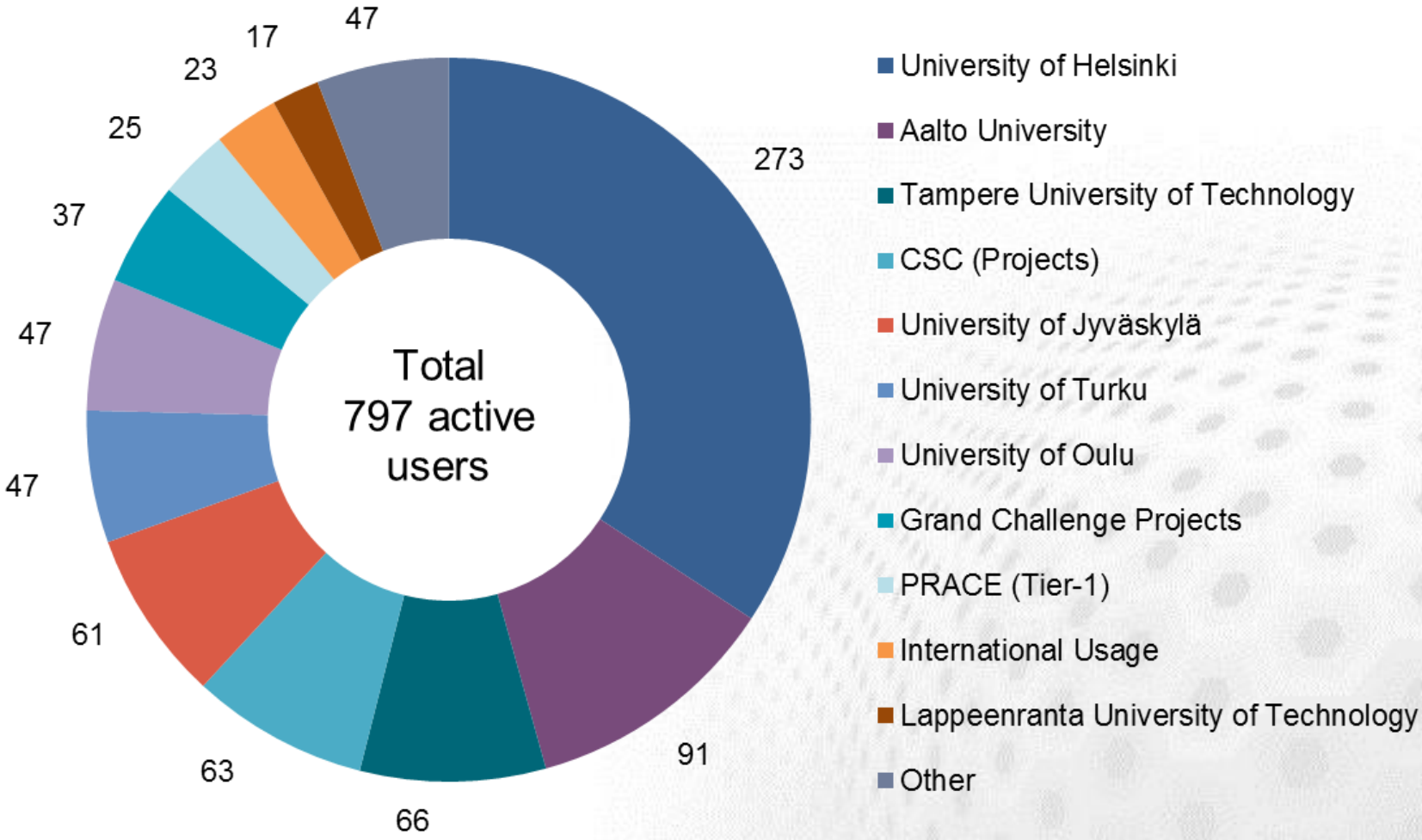
FUNET and Data services

- ***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)

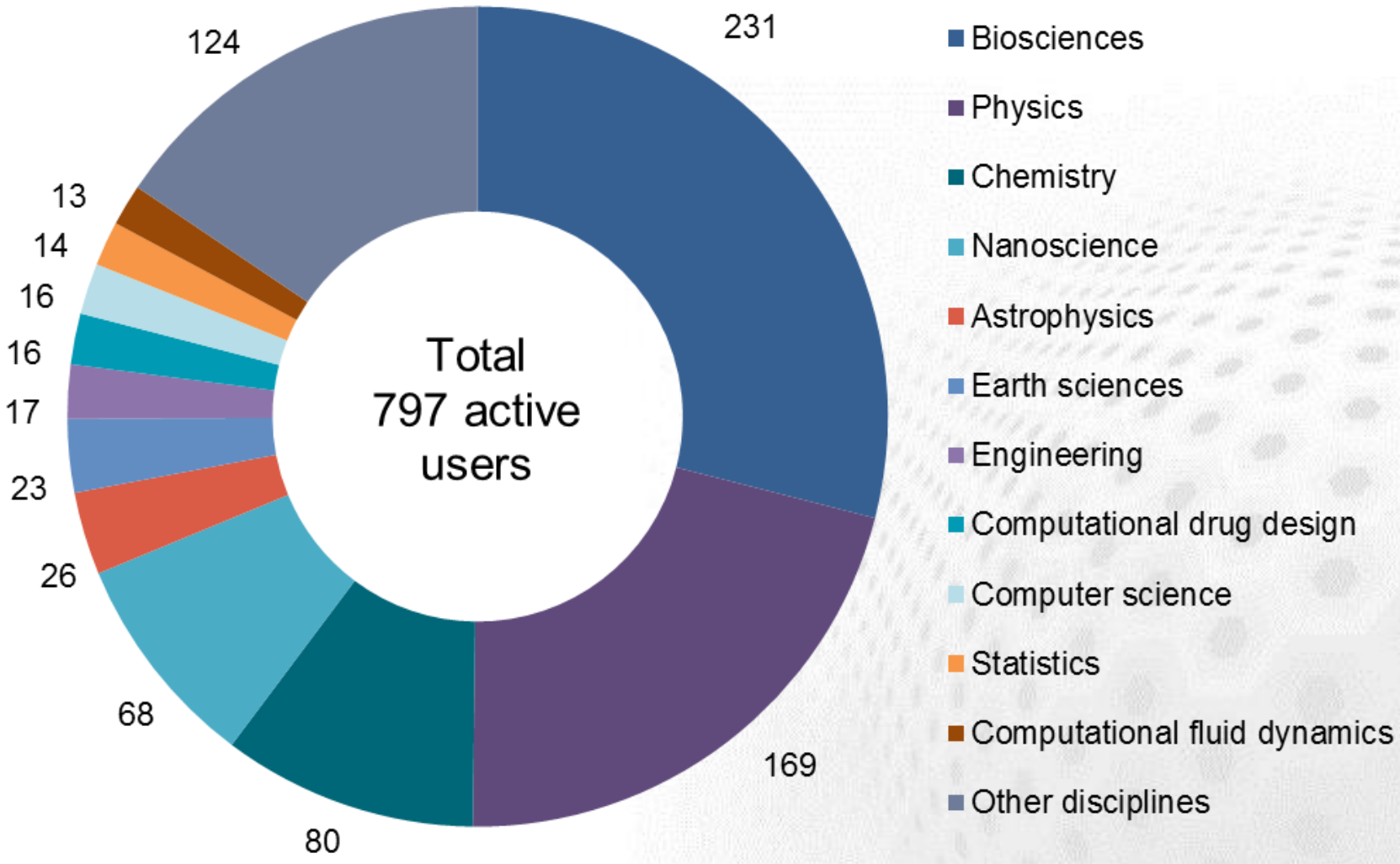
Users

- About 700 active computing projects
  - 3000 researchers use CSC's computing capacity
  - 4250 registered customers
- Haka-identity federation covers all universities and higher education institutes (287 000 users)
- Funet - Finnish research and education network
  - Total of 370 000 end users

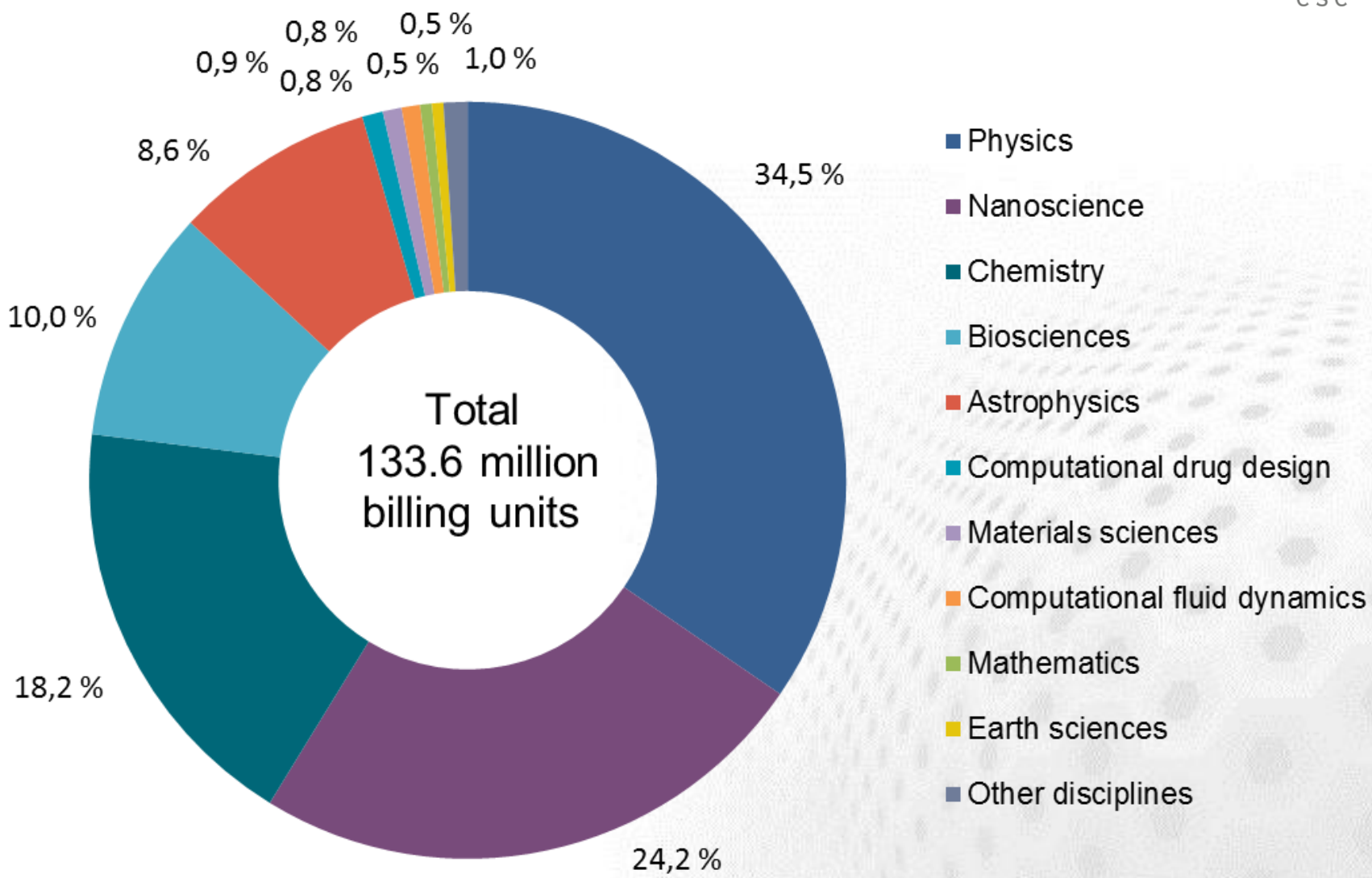
Users of computing resources by organization 1Q/2015



Users of computing resources by discipline 1Q/2015

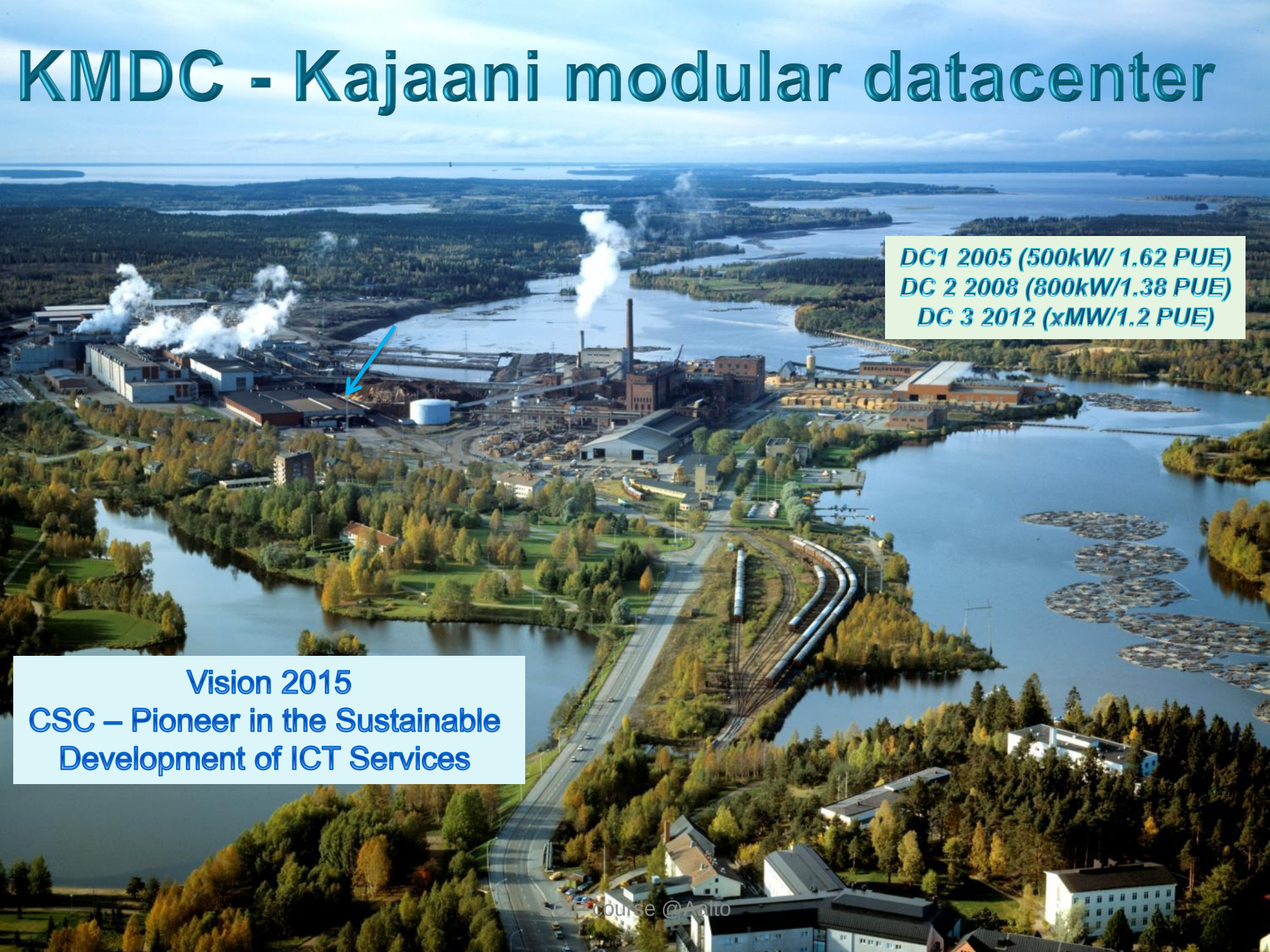


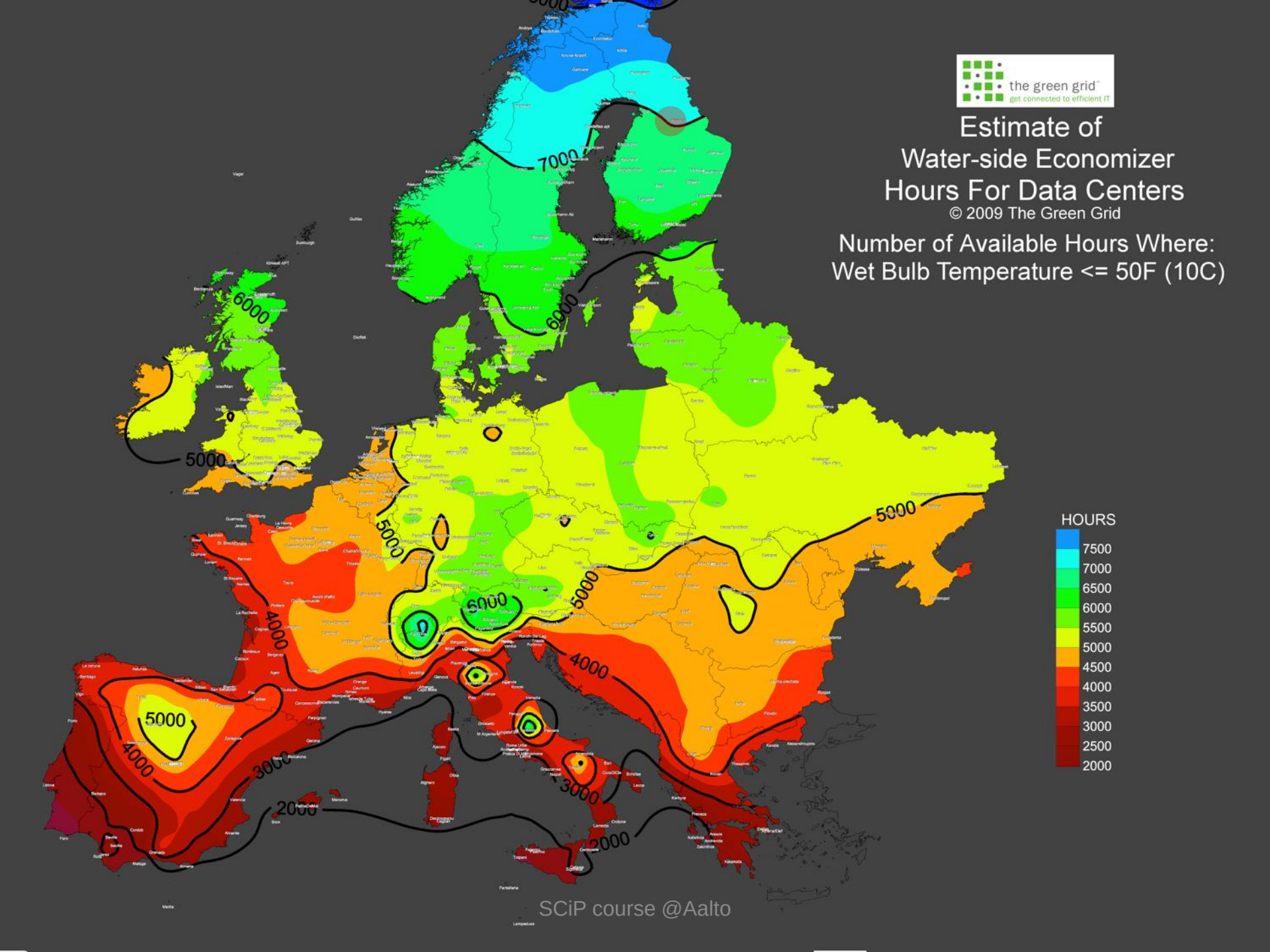
Computing usage by discipline 1Q/2015



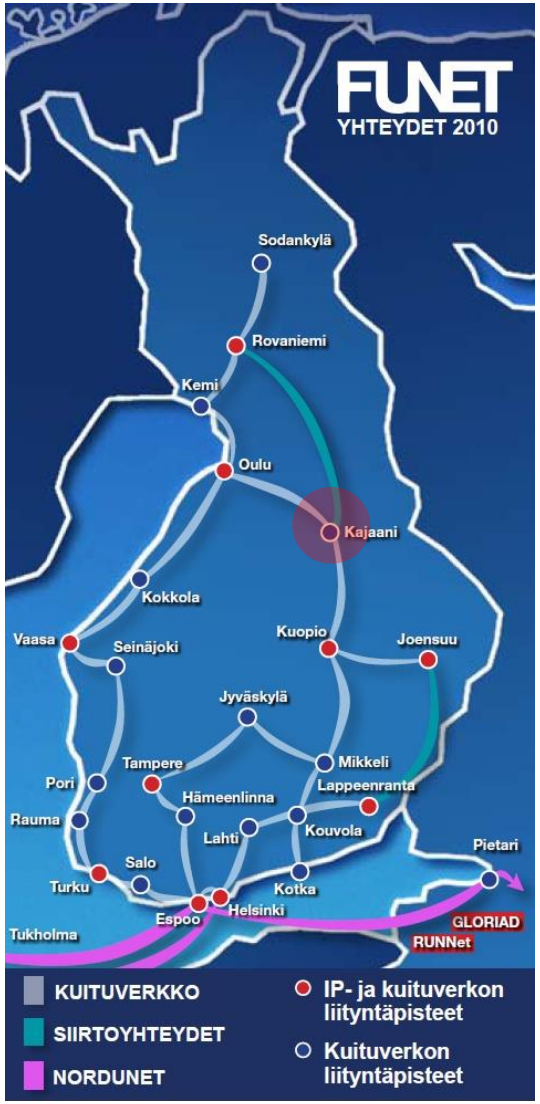
THE KAJAANI DATACENTER

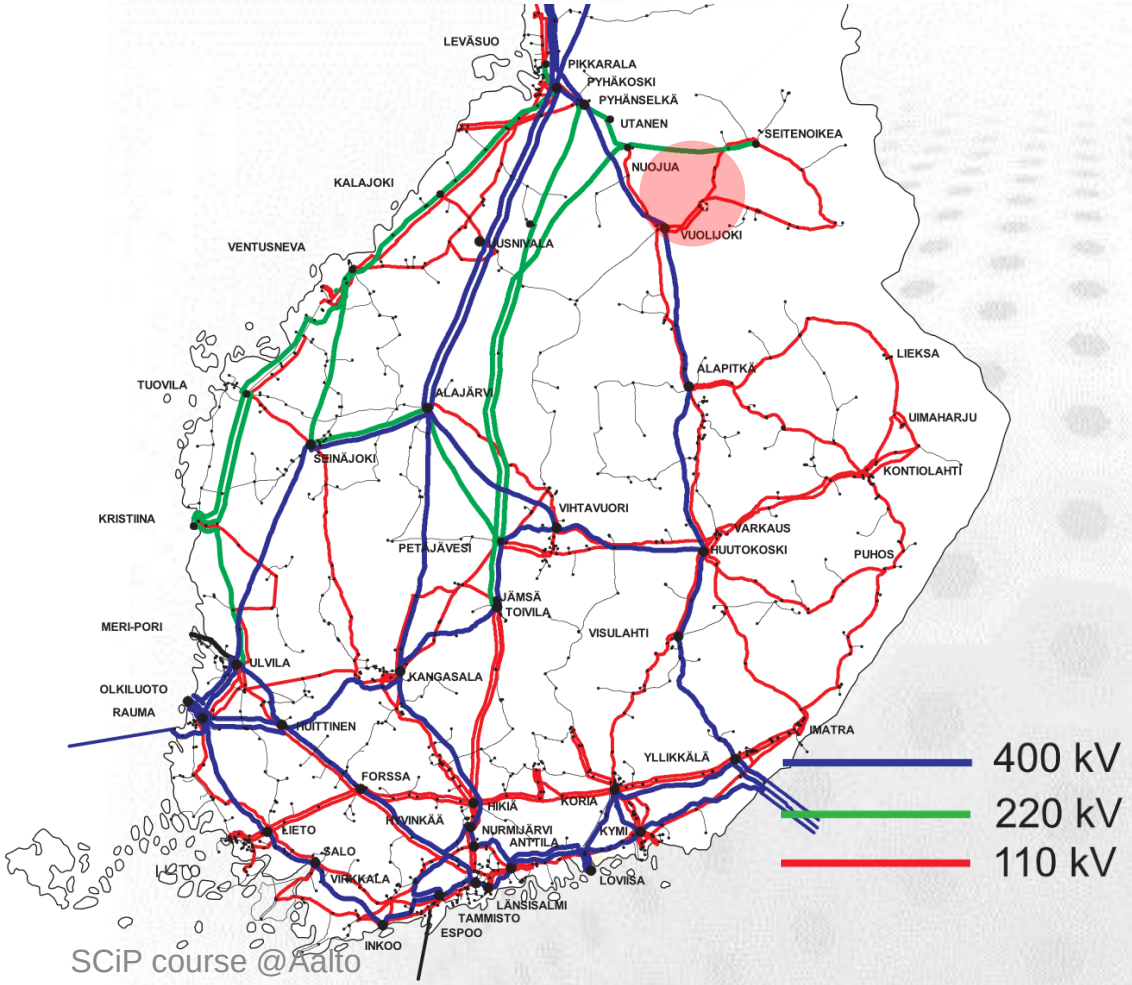
KMDC - Kajaani modular datacenter




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)



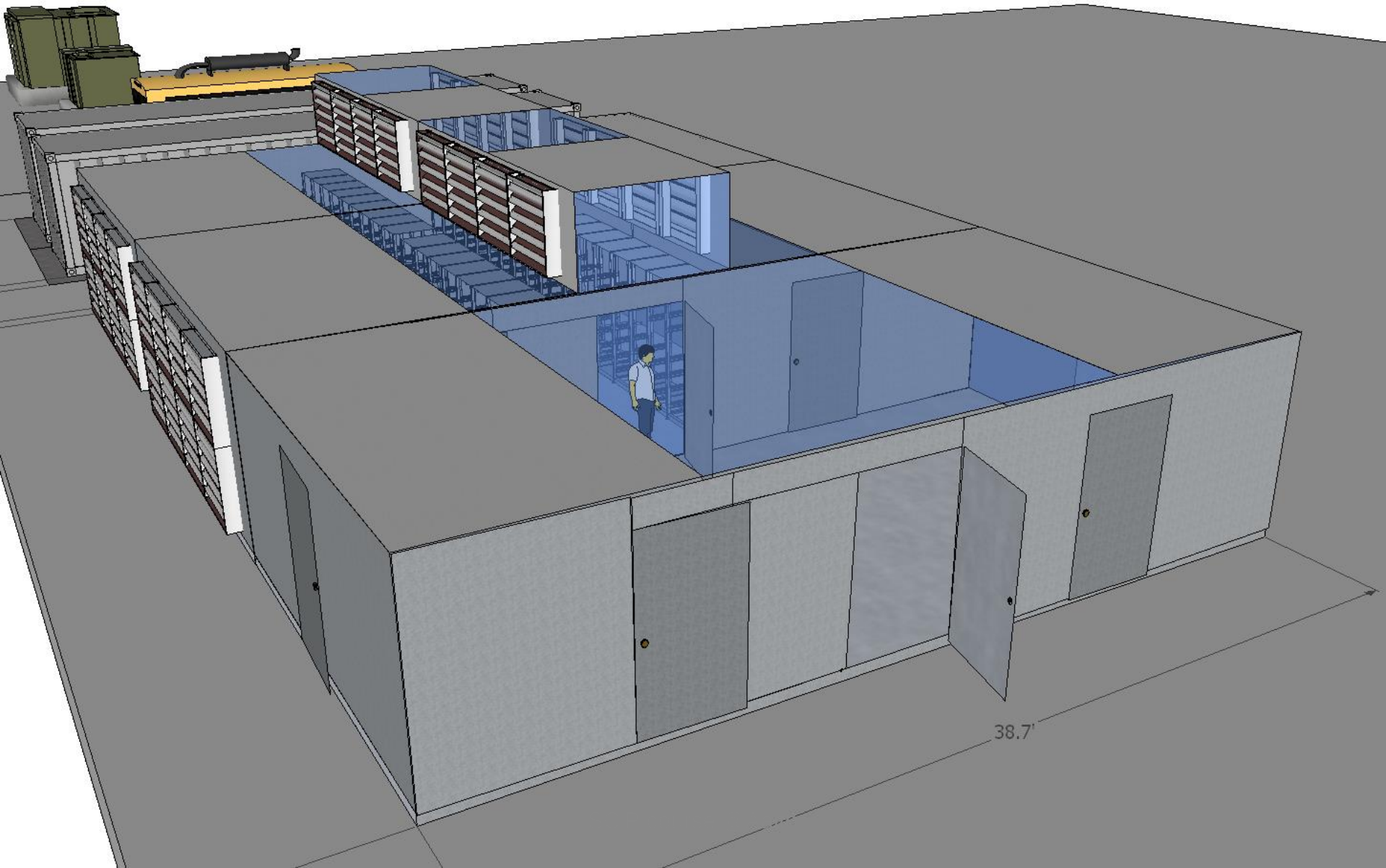


Kajaani site

Sisu

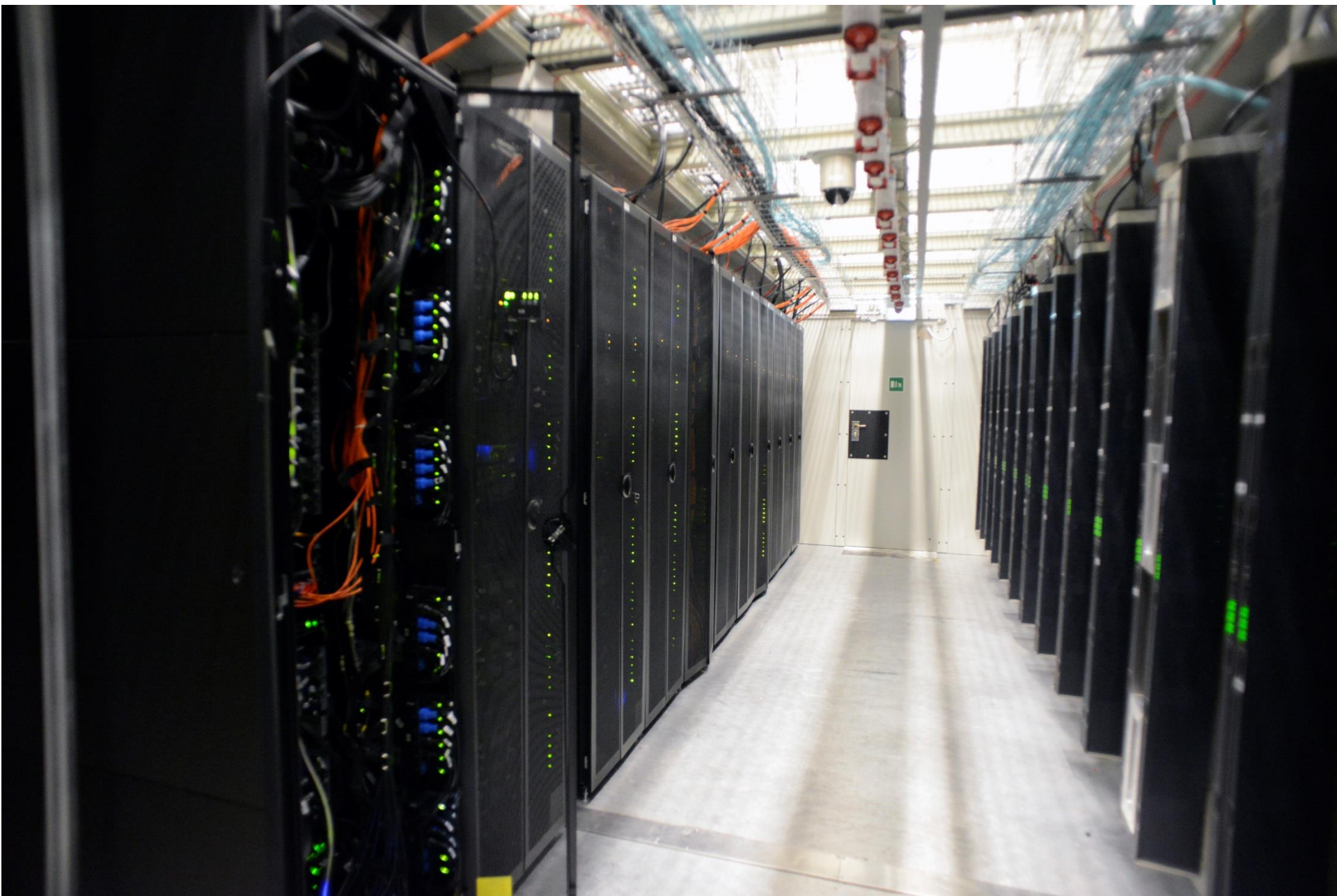
Sisu rear view

Taito (HP) hosted in SGI Ice Cube R80



SGI Ice Cube R80

Taito now



Data center specification

- 2.4 MW combined hybrid capacity
- 1.4 MW modular free air cooled datacenter
  - Hosting e.g. Taito
  - 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 < 1.08 ($pPUE_{L2,YC}$)
- 1MW HPC datacenter
  - Optimised for Cray super & T-Platforms prototype
  - 90% Water cooling

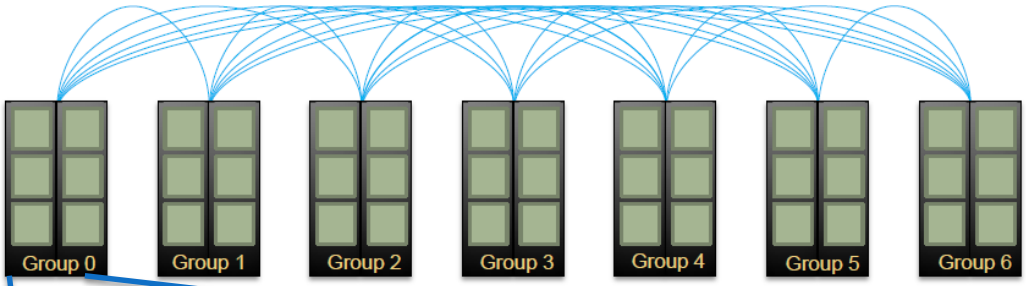
CSC SUPERCOMPUTERS

Main Computing capacity: Sisu,Taito, FGI

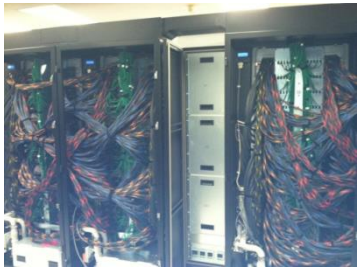
	Sisu (Phase 2)	Taito (Phase 2)	FGI	Taygeta
Availability	2014-	2015-	2012-	2012-
CPU	Intel Haswell and Sandy Bridge, 2 x 12 and 2 x 8 cores, 2.6 GHz, Xeon E5-2690v3 and E5-2670		Intel Xeon, 2 x 6 cores, 2.7 GHZ, X5650	
Interconnect	Aries	FDR IB	QDR IB	
Cores	40512	9768+9216	7308	360
RAM/node	64 GB	64/128/256/ 1536** GB	24/48/ 96 GB	48 GB
Tflops	1688	515	95	4
GPU nodes	-	38	88	-
Disc space	4 PB	4 PB	1+ PB	0.8 TB

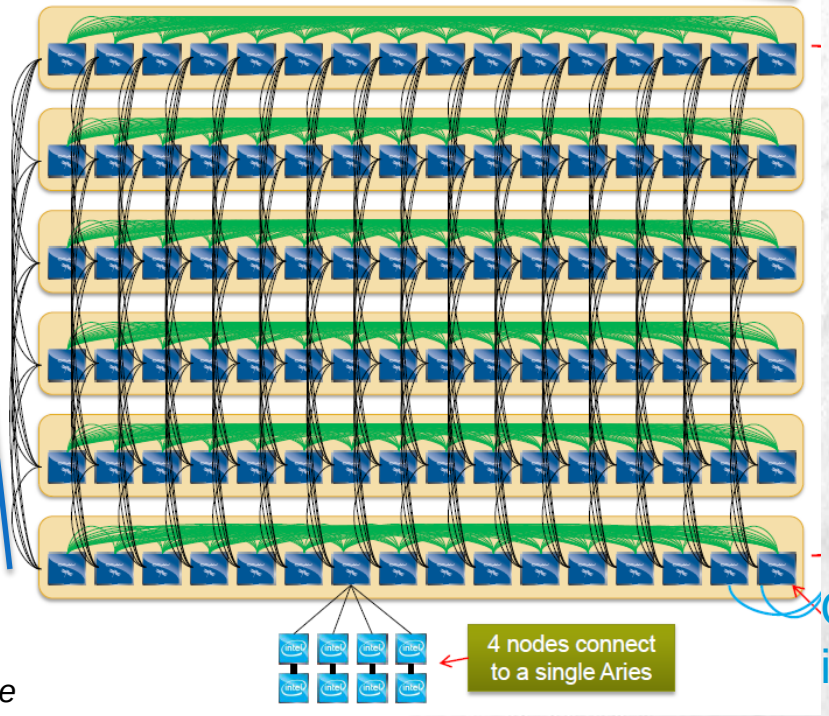
*) 2 nodes a 32 cores with 1,5 TB RAM/node (hugemem-queue)

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 (Sisu)

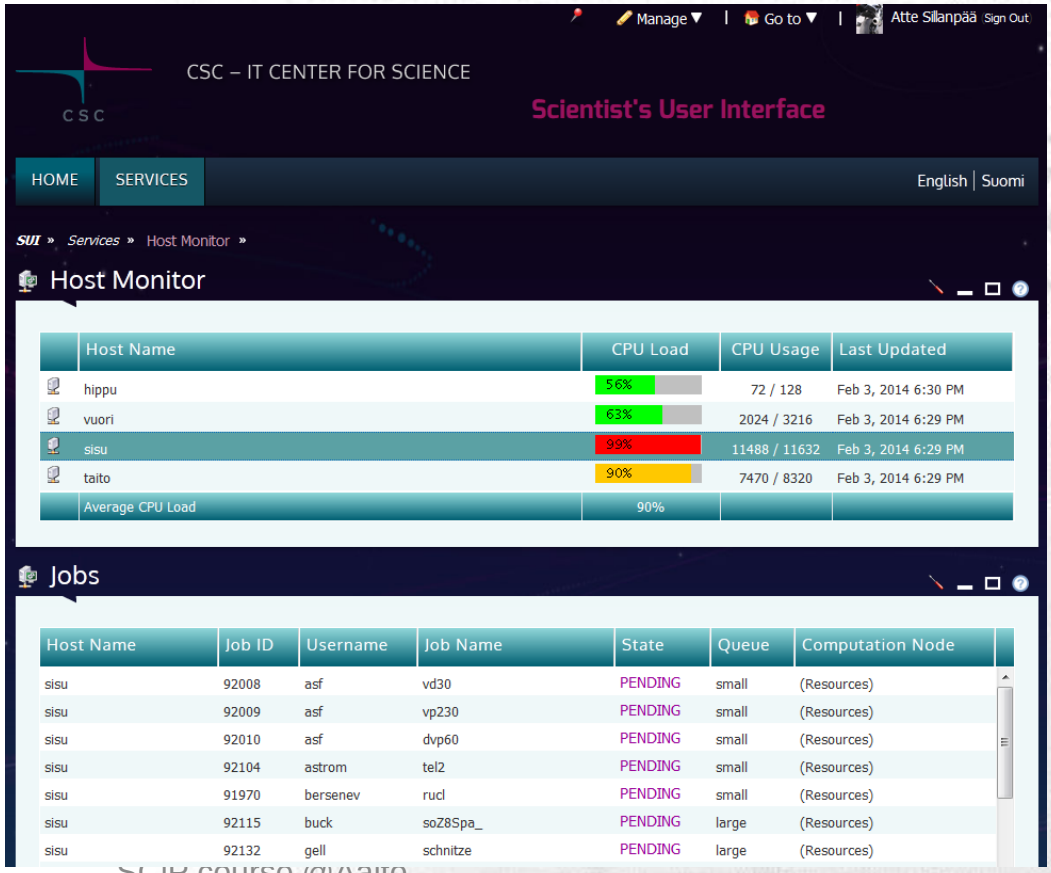
- Typical Cray environment
- Compilers: **Cray**, Intel and GNU
- Cray mpi, Cray tuned versions of all usual libraries
- SLURM
- Default shell: ***bash*** (previously tcsh)
- Character encoding: UTF-8

HP Environment (Taito)

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

How to get access to CSC supercomputers?

- sui.csc.fi (HAKA authentication)
  - Sing up



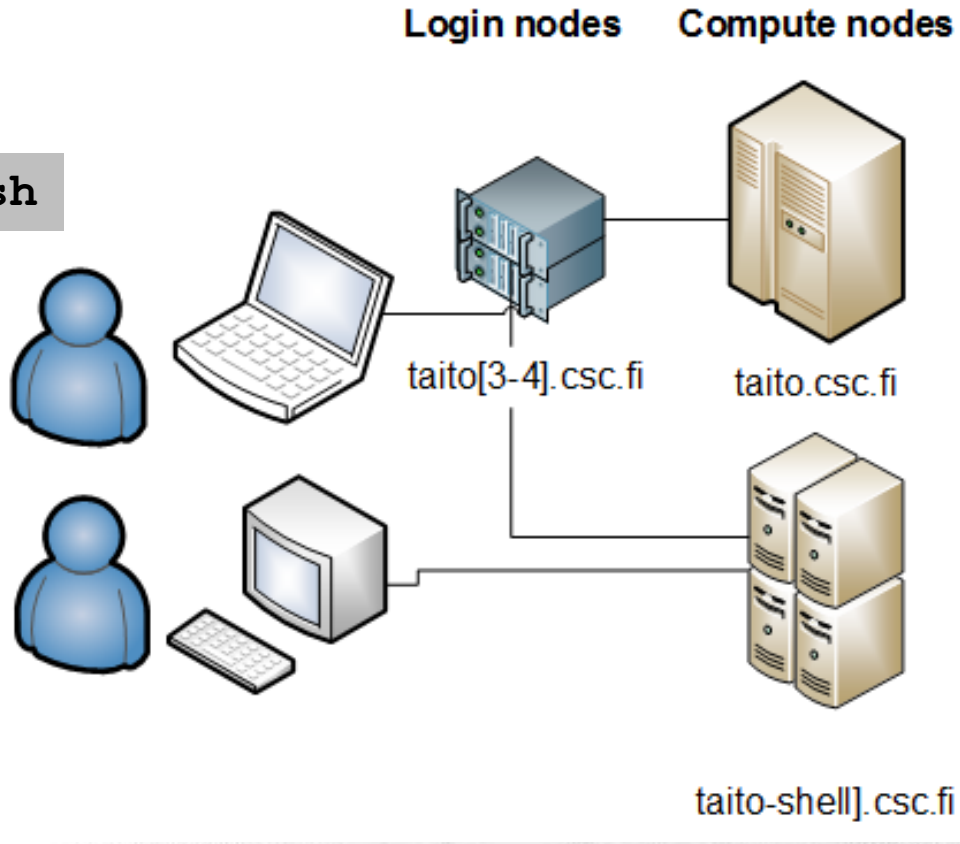
Preparing for demo/hands-on

- Prerequisite:
  - Batch queueing system
  - Module system

Compute nodes are used via queueing system

```
$ sbatch job_script.sh
```

```
$ ./my_prog &
```



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"

Taito module system

- ***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

Live demo/hands-on (Taito)

- ***ssh*** trng86 - trng135 ***@taito.csc.fi***

  (e.g. ssh trng100@taito.csc.fi)

- To be on the safe side:

  mkdir own_username

  cd own_username

- ***module avail***

Live demo/hands-on (1) cont.

- ***nano test_hostname.sh***

```
#!/bin/bash -l
#SBATCH -J print_hostname
#SBATCH -o output.txt
#SBATCH -e errors.t
#SBATCH -t 00:01:00
#SBATCH -p test
#
echo "This job runs on the host:"; hostname
```

CTRL+O; CTRL+X to exit

- ***sbatch test_hostname.sh***

Live demo/hands-on (1) cont.

- Check out the output:
  - ***less output.txt*** (type ***q*** to quit)
  - ***less errors.t*** (type ***q*** to quit)

- More examples:

  https://research.csc.fi/taito-constructing-a-batch-job-file

Taito is a heterogeneous cluster

- Different jobs need different resources
  - Bulk Sandy Bridge compute nodes
  - Largemem Sandy Bridge compute nodes
  - Hugemem Sandy Bridge compute nodes (***1.5 TB*** memory)
  - Bulk Haswell compute nodes
  - Taito-shell (Sandy Bridge) for interactive work

- Choosing between Haswell and Sandy Bridge nodes
  - --constraint=snb or =hsw
    - snb: **16**, hsw: **24**
  - Sandy Bridge: 64 GB memory nodes, Haswell: 128 GB memory in most nodes + 10 with 256 GB
  - More details:
    - *https://research.csc.fi/taito-constructing-a-batch-job-file#3.1.4*

- Local */tmp* disk 2 TB on each node

→ reserve only what you need

Preinstalled modules: optimizations

What architectures a code has been optimized for

- (h) only Haswell → runs only on Haswell
- (sh) Haswell and Sandy Bridge → runs optimally on both
- (g) GPGPU aware → needs to be run on taito-gpu.csc.fi
- No entry → optimized for Sandy Bridge: should run on both SB/H, but not optimally on Haswell (or even SB)

```
[GPU-Env ~]$ module avail

------------------------------ /appl/gpu_modulefiles/Compiler/gcc/4.8.2 ------------
   intelmpi/4.1.3        mkl/11.1.1              openblas/0.2.8
   magma/1.4.1    (g)    mvapich2/2.0-GDR (g)    openmpi/1.8.1  (g)

------------------------------ /homeappl/appl_taito/gpu_modulefiles/Core -------------
   StdEnv      cuda/6.0        gcc/4.8.2 (D)    intel/14.0.1    pgi/14.4
   cuda/5.5    cuda/6.5 (D)    gcc/4.9.1        pgi/14.3        pgi/14.7 (D)

------------------------------ /homeappl/appl_taito/gpu_modulefiles/Linux --------------
   amber-env/14-cuda      (g)    namd-env/2.10b1-cuda (g)        totalview/8.13.0-0
   git/1.9.2                     python-env/2.7.6         (D)    totalview/8.14.0-21  (D)
   gromacs-env/4.6.6-cuda (g)    python-env/3.4.1

  Where:
   (g):  built for GPU
   (D):  Default Module
```

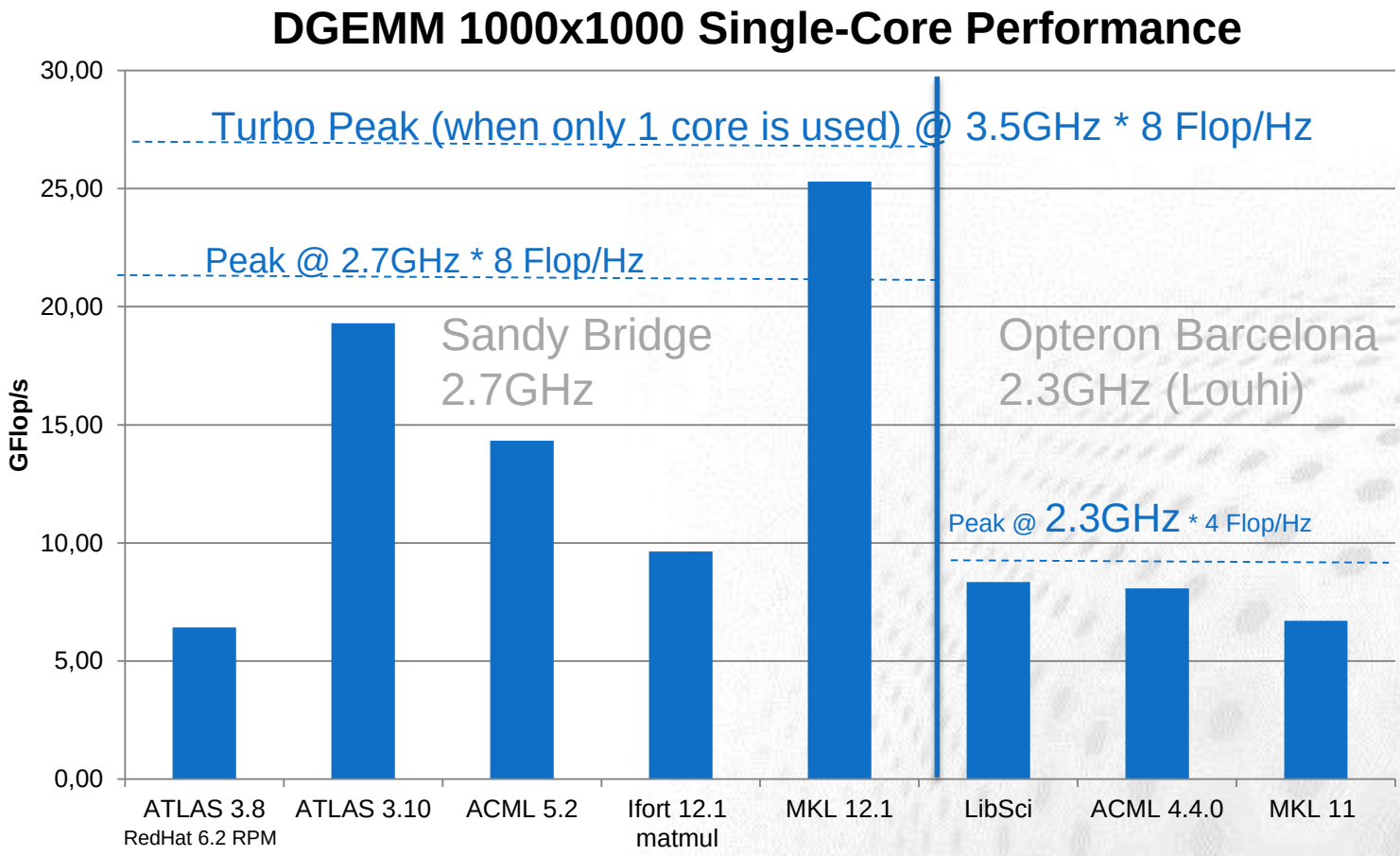
SLURM configuration: Fair usage

- SLURM uses fair share: the highest priority jobs go into execution next
  - Priority is decreased by the total amount of resources used in last 2 weeks per user
  - Priority is increased by time spent queueing
  - Backfiller will try to put small jobs into gaps due to current available resources and highest priority job
  - Jobs labeled ”Association limit” are not eligible to run (due to too many jobs in queue by the user)
- *Due to abuse, a maximum limit of jobs in queue now enforced*
- Chain jobs (--dependency –flag for SLURM) if you need long running time
- Don’t overallocate memory (add this command to your batch script `used_slurm_resources.bash` will print requests vs. used at stdout)
  - If you request a full node (-N 1), use –mem=55000 instead of –mem-per-cpu=something)
  - If you see abuse or think that the setup is unfair, contact *servicedesk@csc.fi*
- SUI has a monitoring tool for your jobs and used resources (*Services -> eServices -> My Project*)

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 a good alternative)

CSC RESOURCES AVAILABLE FOR RESEARCHERS

Server use profiles

- Taito (HP)
- Serial and parallel upto 448/672 cores
- Huge memory jobs
- Lots of preinstalled software

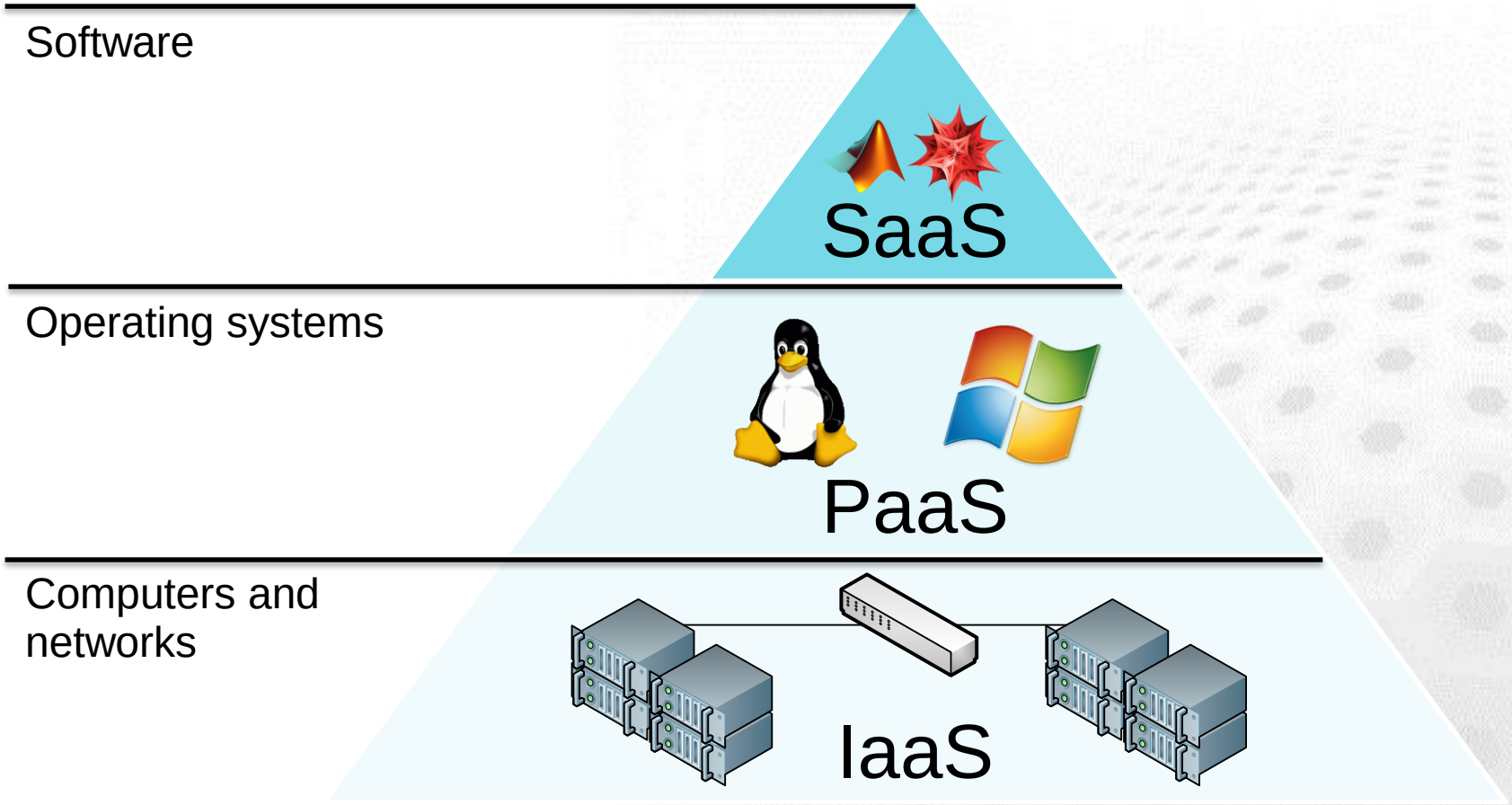
- Taito-shell (HP)
- Interactive jobs
- Very long jobs
- Automatic queue, shared resources

- Sisu (Cray XC40)
- Parallel from 72 up to thousands of cores
- Scaling tests 1008+

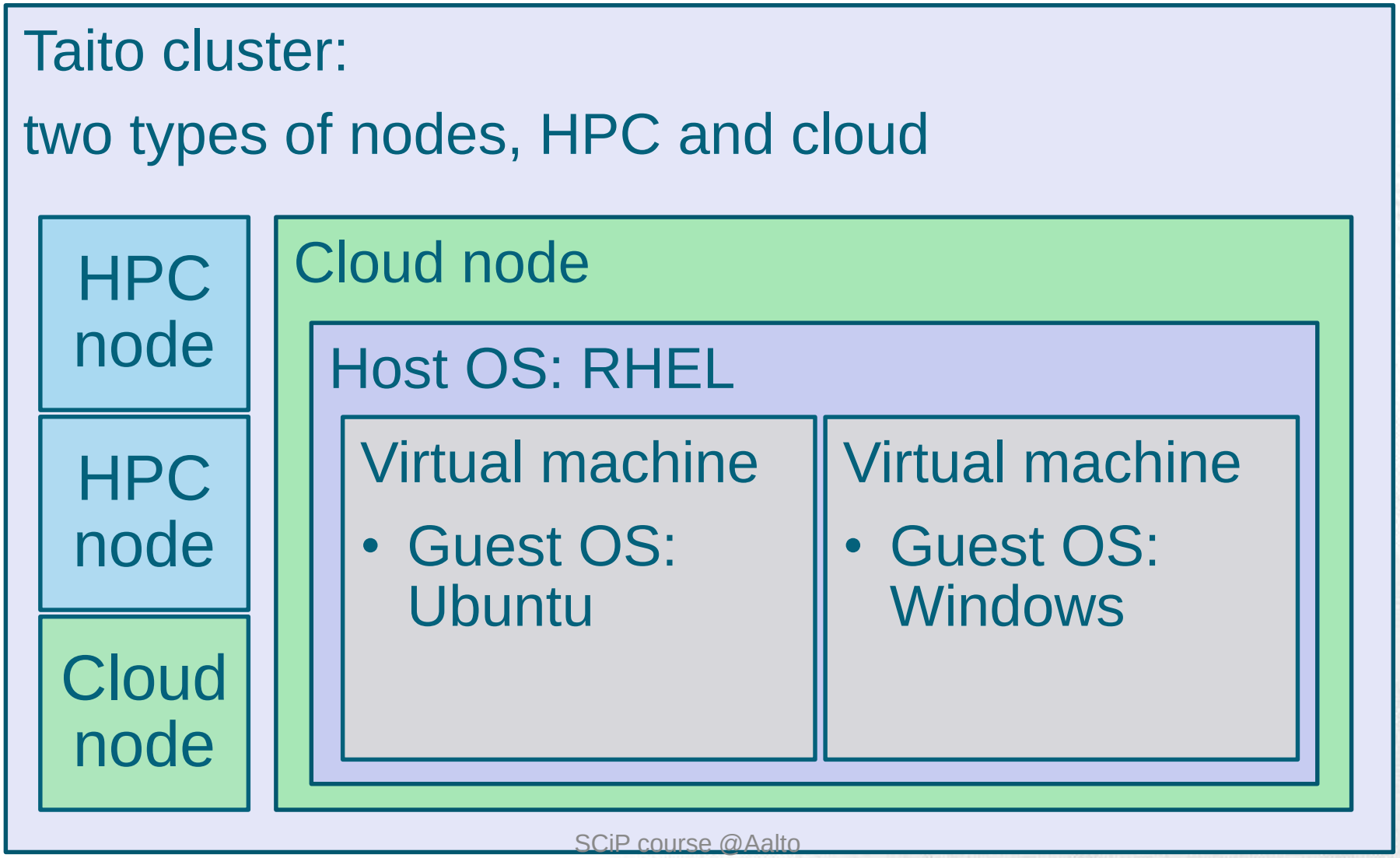
- cPouta (HP) Cloud
- Serial and parallel upto 16 cores

- FGI (HP)
- Serial and parallel (16)

Three service models of cloud computing



Example: Virtualization in Taito



Bull

- 38 NVIDIA K40 nodes (76 gpus)
  - 12 GB memory per card
- 45 Intel Xeon Phi nodes (90 Xeon Phis)
  - 16 GB memory per card
- Energy efficient CPU’s

How to access Bull

- **Accessing the resources**
  - **Intel Xeon Phi:**
    ssh taito-mic.csc.fi
    (*from taito.csc.fi*)
  - **NVIDIA K40:**
    ssh taito-gpu.csc.fi

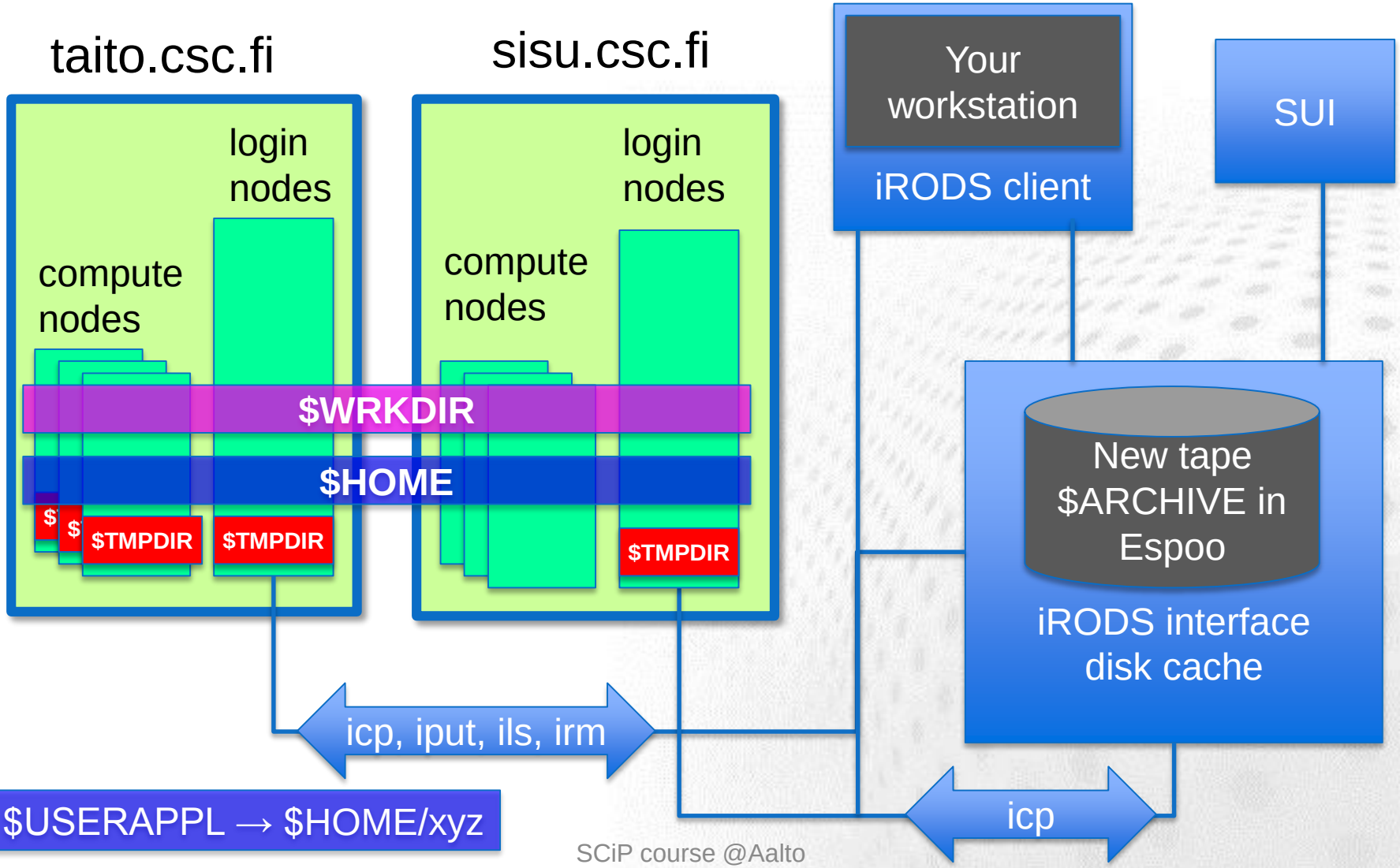
IDA storage service

- Intended for storing research data, the ultimate goal being to facilitate the exploitation of electronic data in research.
- Secure and user-friendly storage service for data and the associated metadata.
- The integrity of the data to be stored is secured by managing copies and their integrity.

Becoming an IDA user

- Universities: Please contact your local IDA contact person (*http://www.tdata.fi/en/idan-kayttajaksi*)
- Universities of applied science: Please contact contact@csc.fi
- Academy of Finland: please contact contact@csc.fi

Disks space


taito.csc.fi
sisu.csc.fi
Your workstation
iRODS client
SUI
login nodes
compute nodes
login nodes
compute nodes
$WRKDIR
$HOME
$TMPDIR
$TMPDIR
$TMPDIR
New tape $ARCHIVE in Espoo
iRODS interface disk cache
icp, iput, ils, irm
icp
$USERAPPL → $HOME/xyz

Disks space cont.

- 4.8 PB on DDN
  - New $HOME directory (on Lustre)
  - $WRKDIR (*not backed up*), soft quota: 5 TB
- HPC_ARCHIVE
  - 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

Directories at CSC Environment (1)

Directory or storage area	Intended use	Default quota/user	Storage time	Backup
$HOME [1]	Initialization scripts, source codes, small data files. Not for running programs or research data.	50 GB	Permanent	Yes
$USERAPPL [1]	Users' own application software.	50 GB	Permanent	Yes
$WRKDIR [1]	Temporary data storage.	5 TB	Until further notice.	No
$TMPDIR [3]	Temporary users' files.	-	~2 days	No
Project [1]	Common storage for project members. A project can consist of one or more user accounts.	On request.	Permanent	No
HPC Archive [2]	Long term storage.	2 TB	Permanent	Yes
IDA [2]	Sharing and long term storage	several TB	At least -2017	Yes

[1]: Lustre parallel ([3]:local) file system in Kajaani [2]: iRODS storage system in Espoo

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 10-100 Gbit/s

ARCHIVE, dos and don'ts

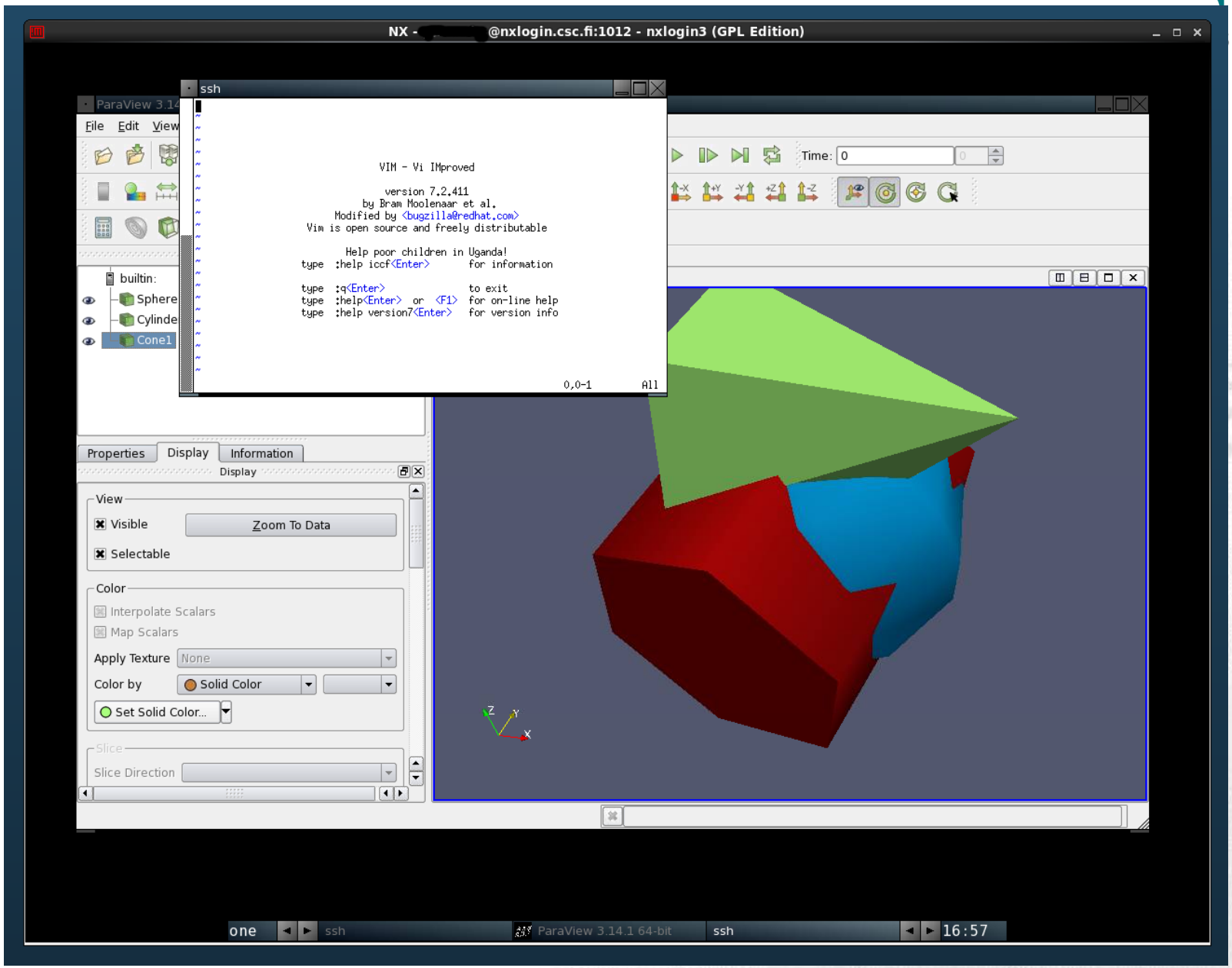
- Don't put small files in HPC ARCHIVE
  - Small files waste capacity
  - Less than 10 MB is small
  - Keep the number of files small
  - Tar and bzip files

- 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 300 GB
- Default quota 5 TB per user

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) *(when needed)*
  - possibility for short (day or less) runs with the whole Cray
- Remember also PRACE/DECI
  - *www.prace-ri.eu/DECI-11-Call*

NX screenshot



Demo / hands on (2)

- Transpose a matrix, e.g. 3X3, by running a batch job of *Matlab* on Taito

Demo / hands on (2) cont.

- Create a matlab file, called e.g. *transpos.m*, which may look like this:

a=[1 2 3;4 5 6;7 8 9];

% non-conjugate transpose of matrix a

b2=transpose(a)

Demo / hands on (2) cont.

- Create a batch job script (e.g. *matlab_transpos.sh*), which will submit it to the queue in a row (non-interactive) mode, e.g. like this (everything in one line, without the "\" character)

srun -v matlab -r "transpos;exit" -no display \ -nosplash -nodesktop –nojvm

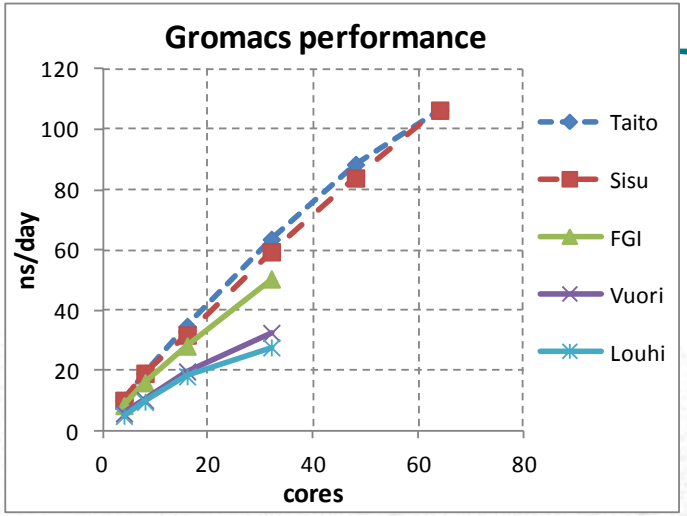
- Remember to load Matlab module before the srun command in the batch script:

module load matlab

Courses at CSC

- **CSC courses**: *http://www.csc.fi/courses*
  - Introduction to Linux and Using CSC Environment Efficiently
  - Pouta training
  - CSC HPC Summer School
  - Spring, Autumn, Winter Schools
  - Parallel Programming
  - Some courses have possibility for remote participation
  - Course materials often available from event website for self study

Summary



- ***Sisu*** *supercomputer*
  - *Fastest in Finland*
- ***Taito*** *supercluster*
  - *Serial & parallel jobs + large memory*
- ***Taito-shell*** *environment*
- ***Pouta Cloud***
  - *cPouta*
  - *ePouta*
- ***Bull*** system
- ***DDN*** HPC storage system