CSC computing resources

Tomasz Malkiewicz
CSC – IT Center for Science Ltd.

Program

- 10-11 CSC presentation
- 11-11:15 TTA presentation
- 11:15-11:30 Round robin
- 11:30-> F2F meetings

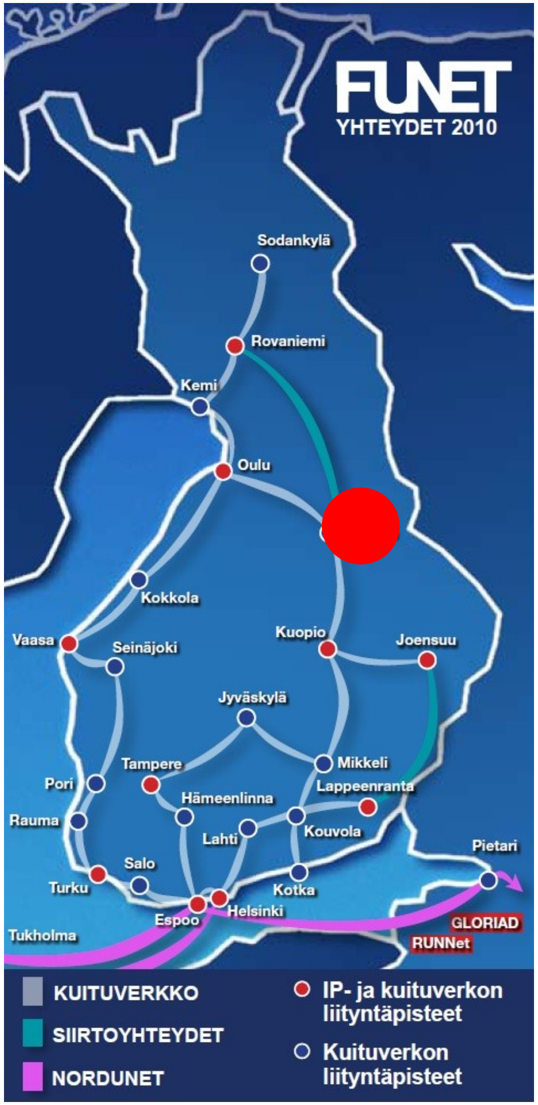
Outline

- CSC at glance
- CSC supercomputers Phase 2
  - *Sisu* (Cray XC30)
  - *Taito* (HP cluster)
- Other resources available for researchers

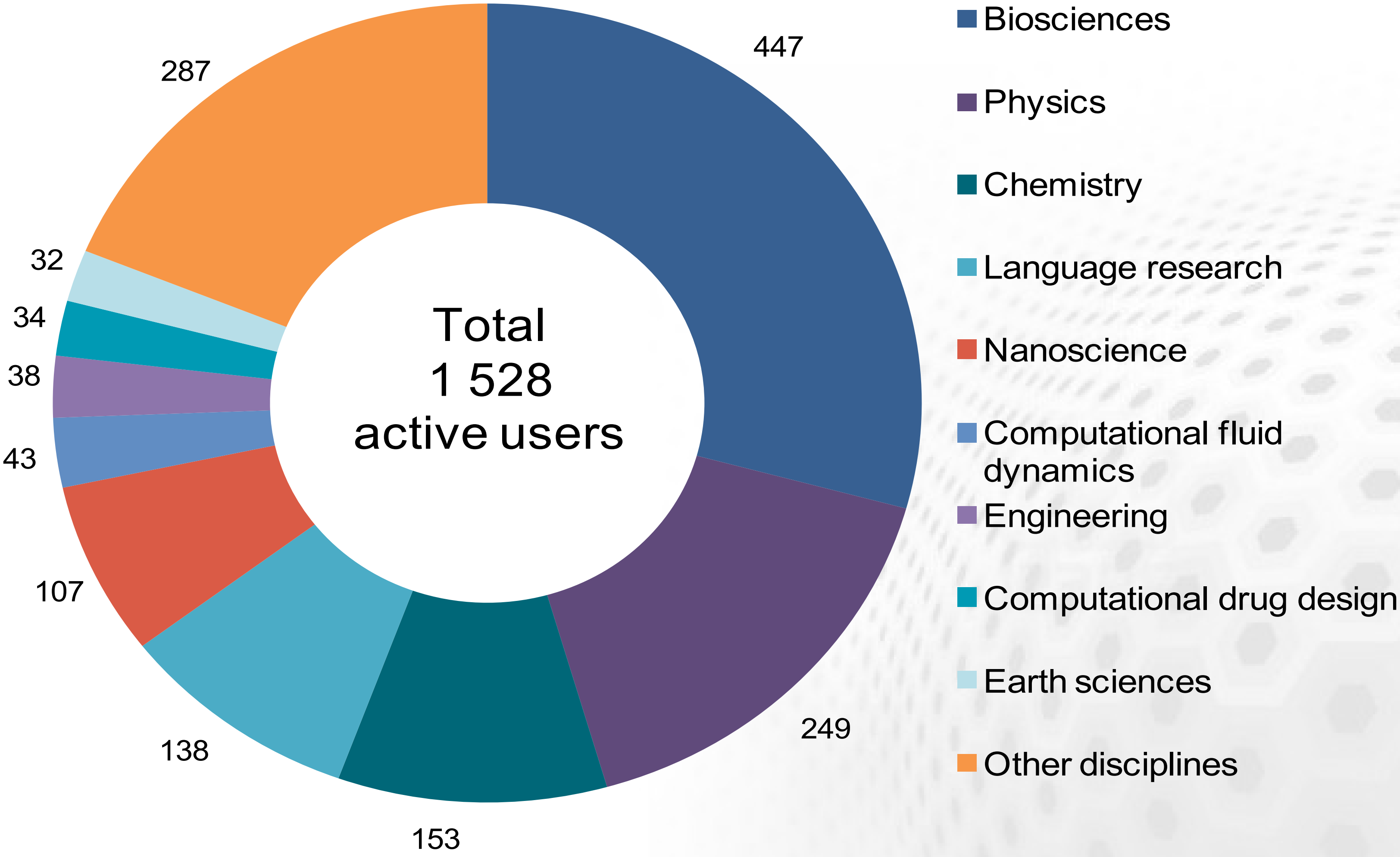
CSC at glance

- Founded in 1971
- Operates on a *non-profit* principle
- Facilities in Espoo and Kajaani
- Staff ~255 people

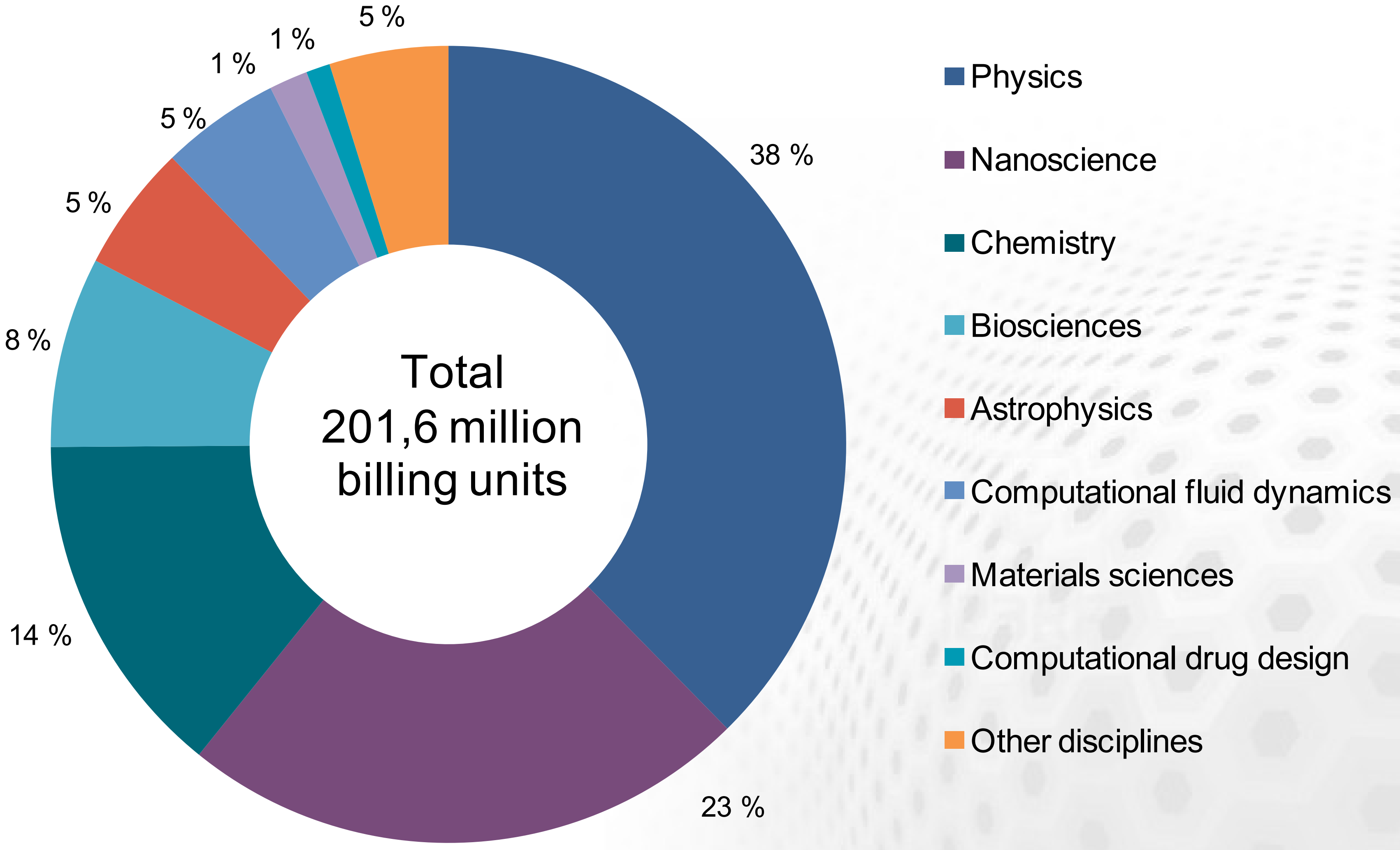
Kajaani modular datacenter



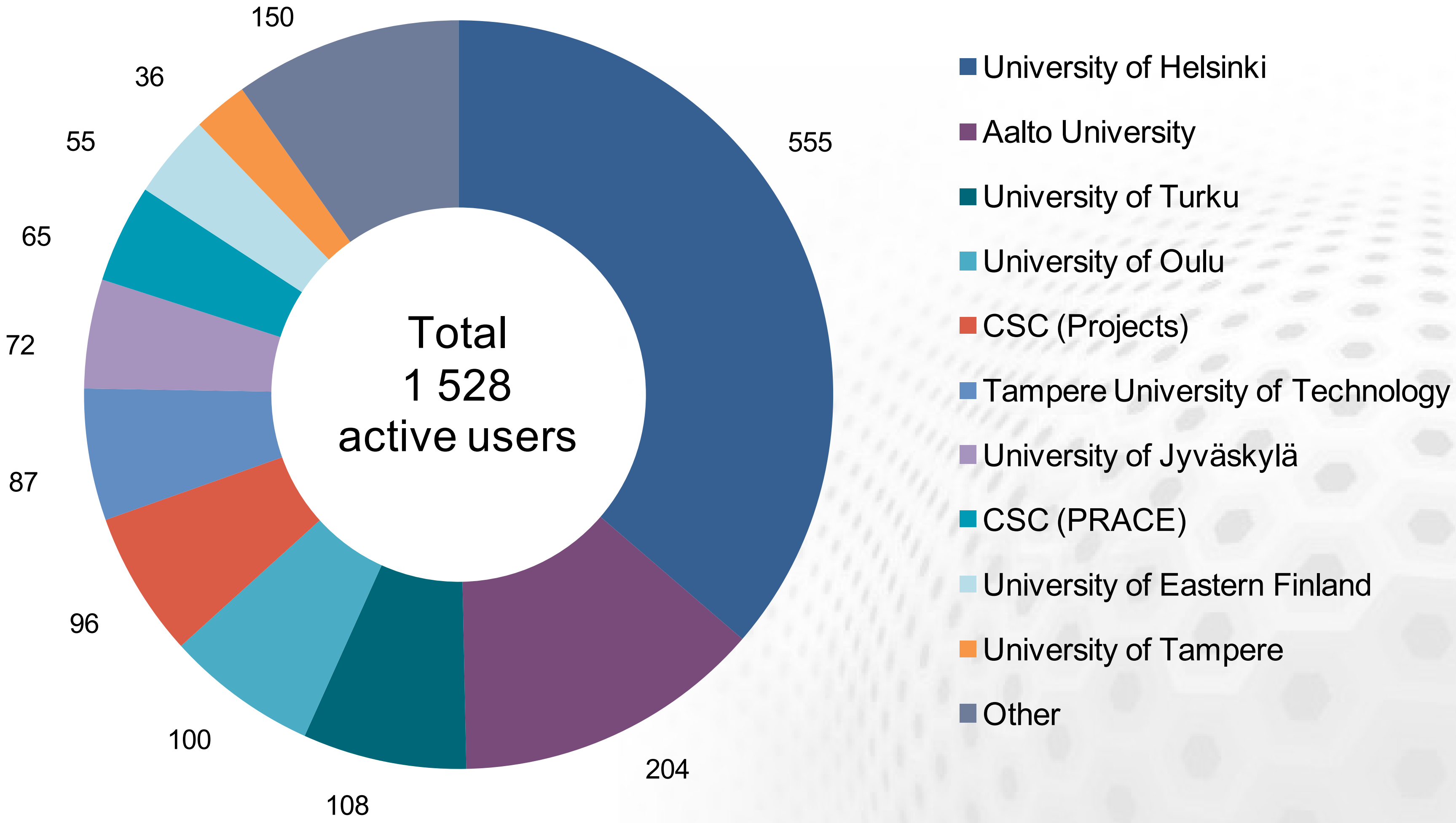
Users of computing resources by discipline 2013



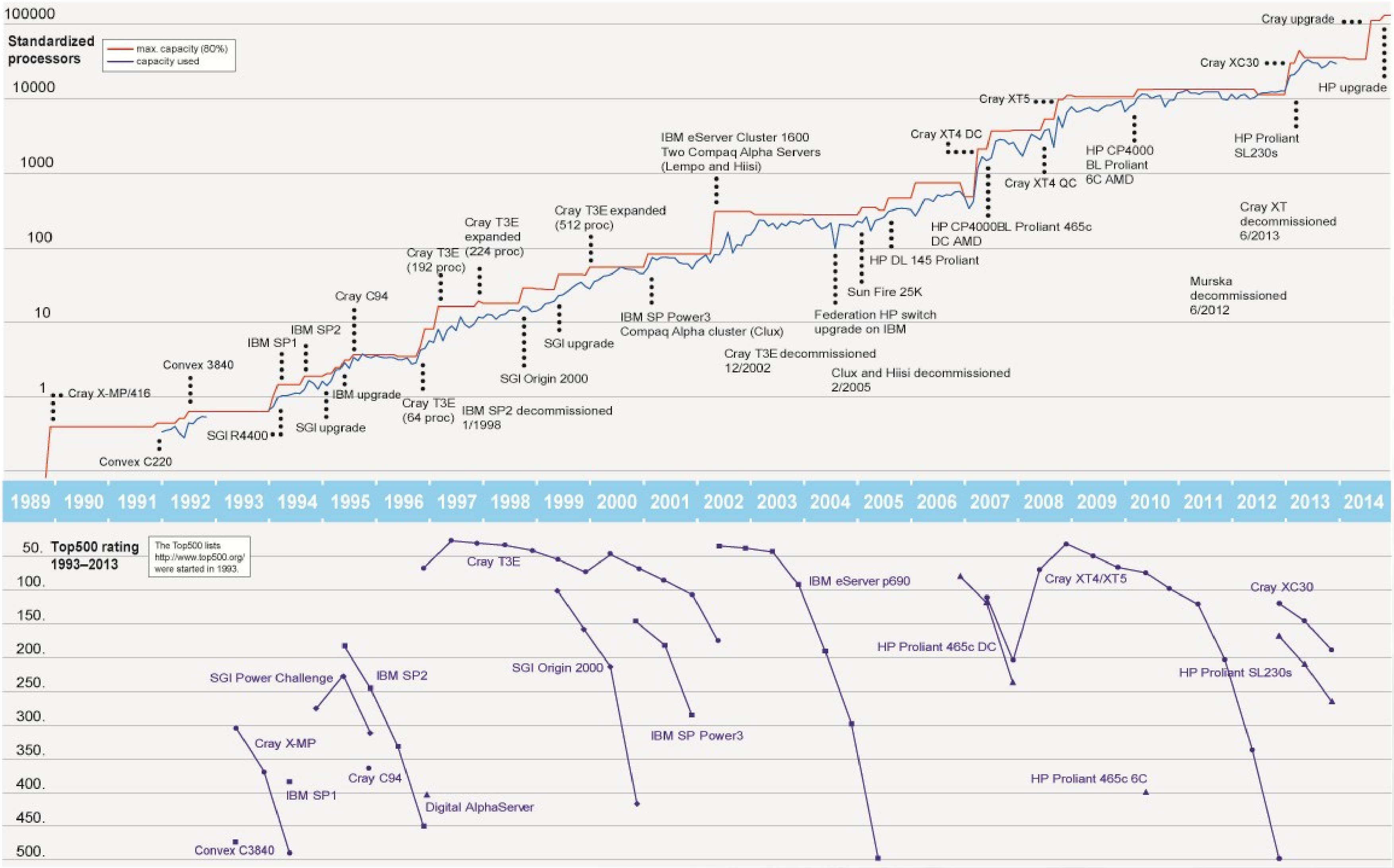
Computing usage by discipline 2013



Users of computing resources by organization 2013



CSC Computing Capacity 1989–2014



PHASE 2 RESOURCES

- SISU
- TAITO
- DDN (PHASE 3)
- BULL

Sisu: Cray Supercomputer

- Future Intel® Xeon® processor E5-2600 v3 product family
- Cray Aries Interconnect
- ~ 40 000 cores
- 64 GB memory per node

How to prepare?

- Old binaries may run off-hand, **CSC advises to recompile the code**
- OS upgrade in June (login nodes)
  - Major upgrade (software on the same level as on Sisu Phase 2)
  - Anything running on ***login nodes*** needs to be recompiled
- Porting strategy
  - Under preparation

Sisu

- AVX-2
  - May need to optimize for wider vectors' size
- DDR4
  - Higher bandwidth, lower power consumption
- Max job size likely to increase
- Native SLURM on the way, unlikely to be available after Sisu July hardware update

Sisu (un)availability in summer 2014

- June 2014 software upgrade break
  - Probably 3 days of downtime
- July 2014 hardware upgrade break
  - At least 2 weeks downtime expected

Taito: HP Supercluster

- Intel® Xeon® processor E5-2600 v2 product family & Future Intel® Xeon® processor E5-2600 v3 family
- FDR InfiniBand interconnect
- ~ 17 000 cores
- Different memory per node sizes: 64, 128, 256 GB and 1.5 TB

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
- Bulk new architecture compute nodes

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

→ reserve only what you need

One SLURM to serve them all…

- Do old applications run on new CPUs
  - May run, CSC **recommends re-compiling**
  - Build your software for both (old and new) architecture
  - Gain depends on architecture
- Batch job scripts need to be updated
  - Number of cores per node may change
  - Memory changes
  - Instructions will be available
- How to submit jobs to either architecture only
  - Specify to which partitions you send your jobs

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-core=something)
  - If you see abuse or think that the setup is unfair, contact *helpdesk@csc.fi*
- SUI has a monitoring tool for your jobs and used resources (*Services -> eServices -> My Project*)

Taito (un)availability in summer 2014

- June 2014 software upgrade break
  - Probably 3 days of downtime

Current Plan for Phase2 Sisu and Taito

- Sisu: planned installation in July-August 2014
  - General availability planned for Q3 2014

- Taito: planned installation in Q4 2014

How to prepare?

- Porting strategy
  - Not much to do at this stage
  - Compilers, libraries, flags, …
  - Preliminary performance data?
  - Add **AVX-2 flag** when compiling your code

Bull

- In pilot/project until end of August 2014
- No guarantee on availability
- 38 NVidia K40 nodes (76 gpus)
  - 12 GB memory per card
- 45 Xeon Phi nodes (90 Xeon Phis)
  - 16 GB memory per card
- Energy efficient CPU’s

How to access (plan)

- Accessing the resources
  - Xeon Phi: ssh taito-mic.csc.fi (TBC)
  - Nvidia K40: ssh taito-gpu.csc.fi (TBC)

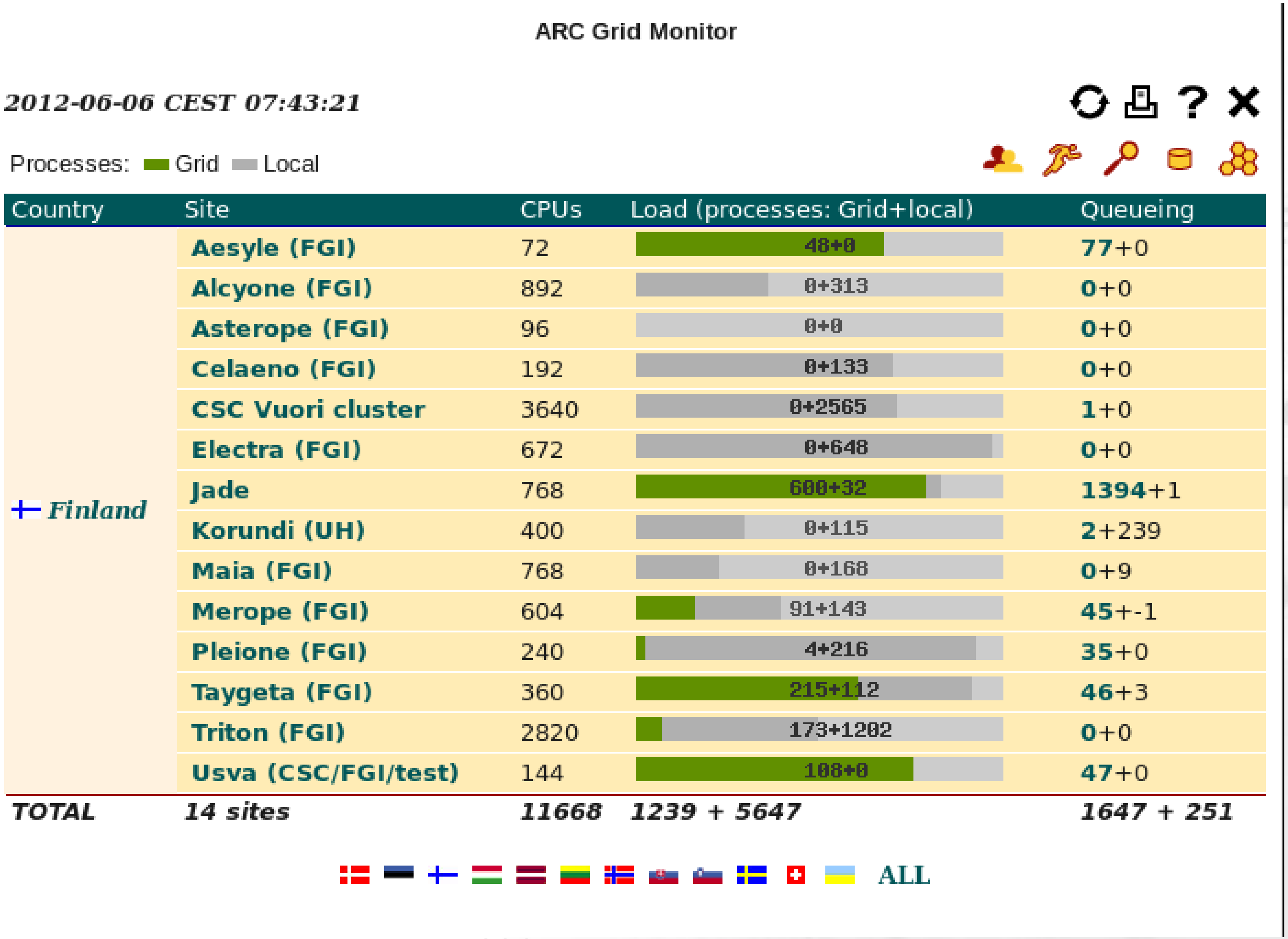
Pettu Phase 3

- System size will increase to ~4 PB
  - About 1.9 PB will added to the current configuration
  - Aggregate bandwidth > 80 GB/s (currently ~48 GB/s)
- Available together with Phase2 supercomputers
- Downtime on all systems (~1 day)

Disks in total

- *4.0 PB on DDN*
  - New $HOME directory (on Lustre)
  - $WRKDIR (*not backed up*), soft quota 5 TB / user
  - Up to 100 TB / project
- *HPC Archive*
  - 2 TB / user, common between Cray and HP
- *3 PB disk space through TTA/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*

Grid computing with Finnish Grid Infrastructutre (FGI)



ARC Grid Monitor

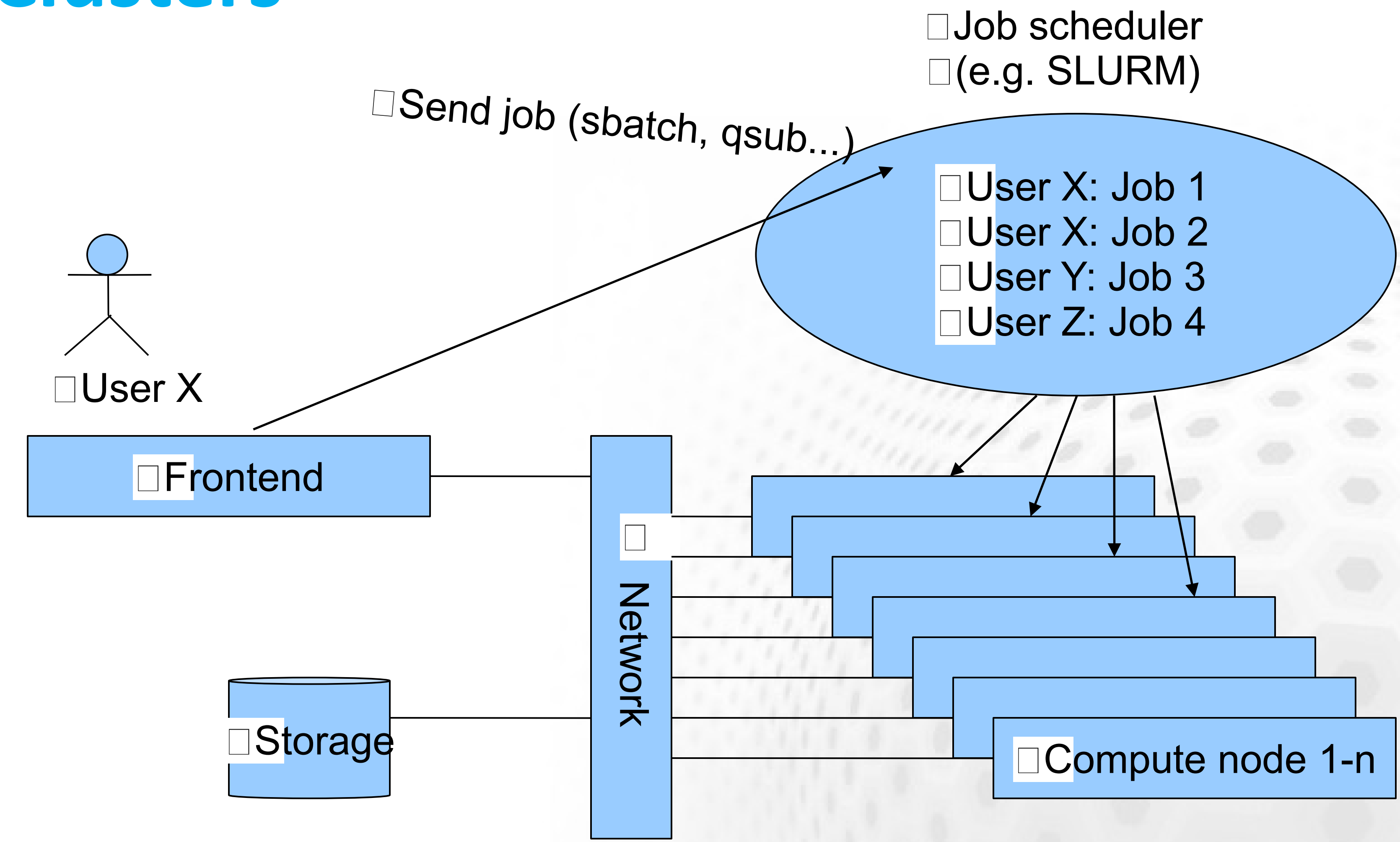
2012-06-06 CEST 07:43:21

Processes: 🟩 Grid ▨ Local

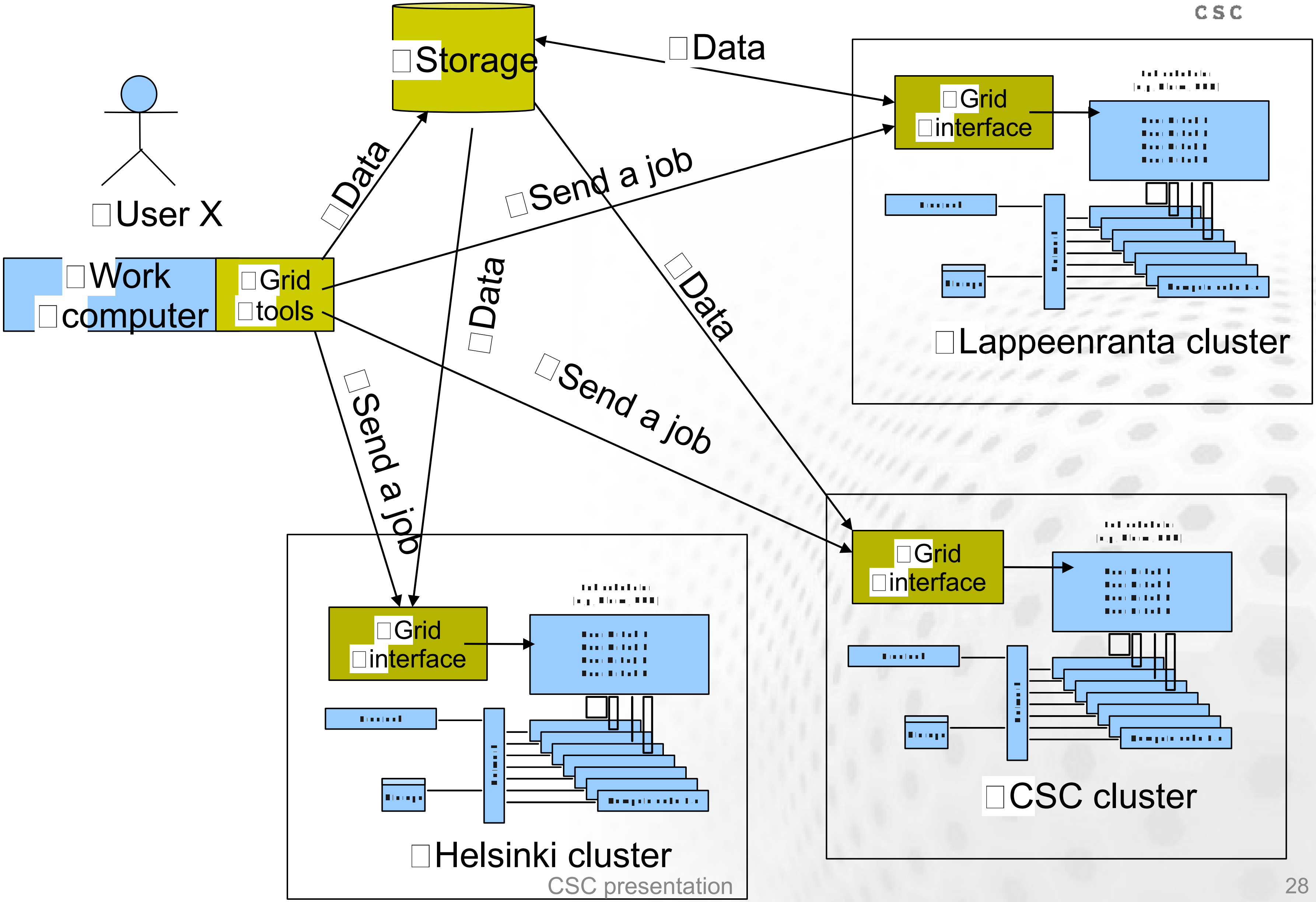
Country	Site	CPUs	Load (processes: Grid+local)	Queueing
Finland	**Aesyle (FGI)**	72	48+0	**77**+0
	Alcyone (FGI)	892	0+313	**0**+0
	Asterope (FGI)	96	0+0	**0**+0
	Celaeno (FGI)	192	0+133	**0**+0
	CSC Vuori cluster	3640	0+2565	**1**+0
	Electra (FGI)	672	0+648	**0**+0
	Jade	768	600+32	**1394**+1
	Korundi (UH)	400	0+115	**2**+239
	Maia (FGI)	768	0+168	**0**+9
	Merope (FGI)	604	91+143	**45**+-1
	Pleione (FGI)	240	4+216	**35**+0
	Taygeta (FGI)	360	215+112	**46**+3
	Triton (FGI)	2820	173+1202	**0**+0
	Usva (CSC/FGI/test)	144	108+0	**47**+0
TOTAL	***14 sites***	***11668***	***1239 + 5647***	***1647 + 251***

ALL

Clusters


Job scheduler
(e.g. SLURM)
Send job (sbatch, qsub...)
User X: Job 1
User X: Job 2
User Y: Job 3
User Z: Job 4
User X
Frontend
Network
Storage
Compute node 1-n

Grids



Storage
Data
User X
Data
Send a job
Work computer
Grid tools
Data
Data
Grid interface
Lappeenranta cluster
Send a job
Send a job
Grid interface
CSC cluster
Grid interface
Helsinki cluster

Getting started with FGI-Grid

1. Apply for a grid certificate from TERENA (a kind of grid passport)
2. Join the FGI VO (Access to the resources)
3. Install the certificate to Scientists' User Interface and Hippu.
4. Install ARC client to your local Mac or Linux machine for local use)
5. Instructions*: http://research.csc.fi/fgi-preparatory-steps*

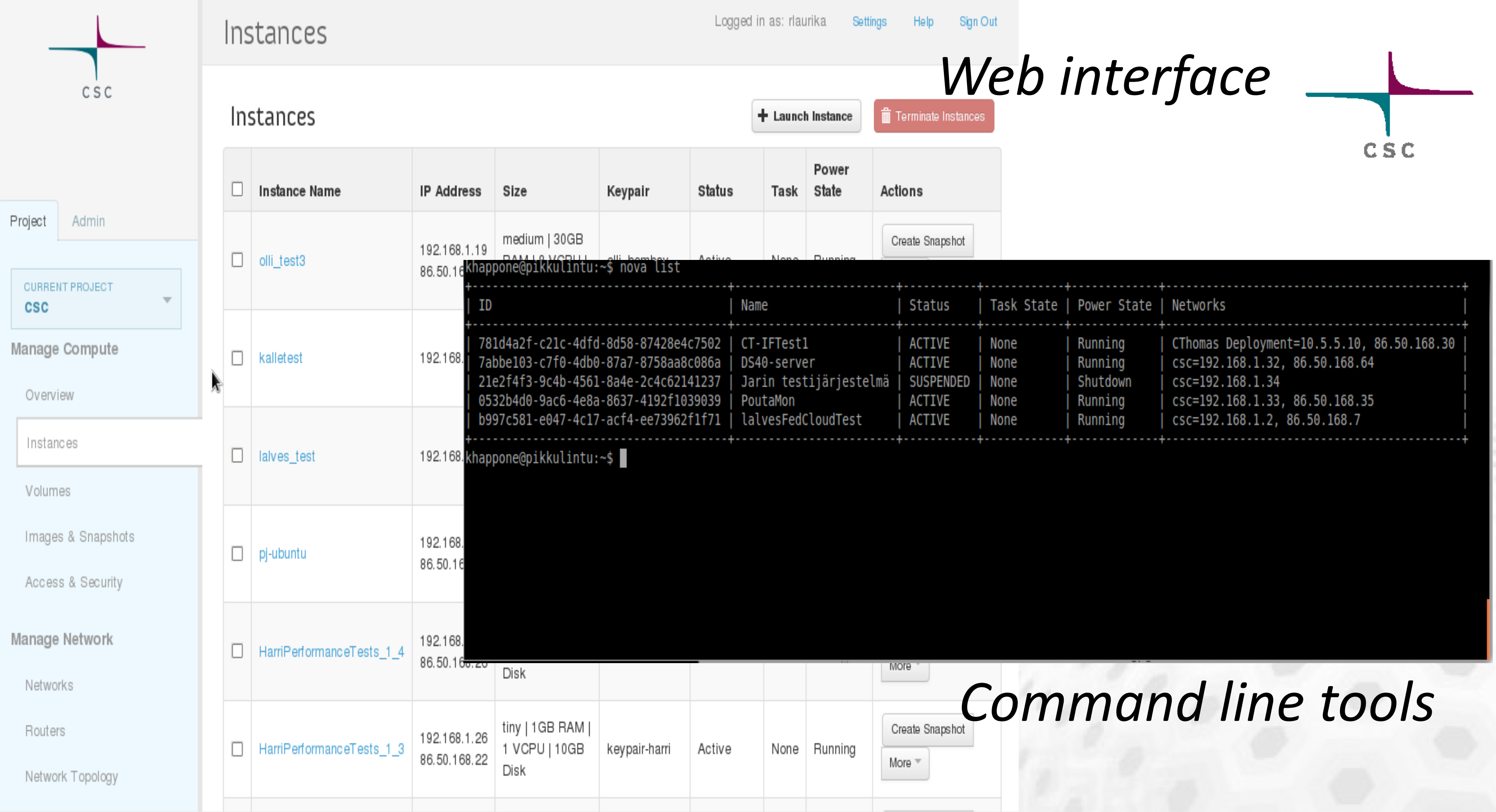
Please ask help to get started: helpdesk@csc.fi

FGI user guide: *http://research.csc.fi/fgi-user-guide*

Pouta – Computing in the Cloud

- Virtual machines on demand
  - Taito hardware
  - Dedicated resources (HPC focus)
- More freedom
- More responsibility
- More work

Web interface



Instances

Instances

☐	Instance Name	IP Address	Size	Keypair	Status	Task	Power State	Actions
☐	olli_test3	192.168.1.19 86.50.16	medium | 30GB					Create Snapshot
☐	kalletest	192.168.						
☐	lalves_test	192.168.						
☐	pj-ubuntu	192.168. 86.50.16						
☐	HarriPerformanceTests_1_4	192.168. 86.50.16	Disk					More
☐	HarriPerformanceTests_1_3	192.168.1.26 86.50.168.22	tiny | 1GB RAM | 1 VCPU | 10GB Disk	keypair-harri	Active	None	Running	Create Snapshot More

```
khappone@pikkulintu:~$ nova list
+--------------------------------------+----------------------+-----------+------------+-------------+------------------------------------------+
| ID                                   | Name                 | Status    | Task State | Power State | Networks                                 |
+--------------------------------------+----------------------+-----------+------------+-------------+------------------------------------------+
| 781d4a2f-c21c-4dfd-8d58-87428e4c7502 | CT-IFTest1           | ACTIVE    | None       | Running     | CThomas Deployment=10.5.5.10, 86.50.168.30 |
| 7abbe103-c7f0-4db0-87a7-8758aa8c086a | DS40-server          | ACTIVE    | None       | Running     | csc=192.168.1.32, 86.50.168.64           |
| 21e2f4f3-9c4b-4561-8a4e-2c4c62141237 | Jarin testijärjestelmä | SUSPENDED | None       | Shutdown    | csc=192.168.1.34                         |
| 0532b4d0-9ac6-4e8a-8637-4192f1039039 | PoutaMon             | ACTIVE    | None       | Running     | csc=192.168.1.33, 86.50.168.35           |
| b997c581-e047-4c17-acf4-ee73962f1f71 | lalvesFedCloudTest   | ACTIVE    | None       | Running     | csc=192.168.1.2, 86.50.168.7             |
+--------------------------------------+----------------------+-----------+------------+-------------+------------------------------------------+
khappone@pikkulintu:~$
```

Command line tools

```
https://pouta.csc.fi:8777/v2/csc/servers/0532b4d0-9ac6-4e8a-8637-4192f1039039
https://pouta.csc.fi:8777/v2/csc/flavors/1a0f1143-47b5-4e8a-abda-eba52ae3c5b9
https://pouta.csc.fi:8777/v2/csc/images/
```

REST API

Pouta audience

- Advanced users – able to manage servers
- Difficult workflows – can’t run on Taito
- Complex software stacks
- Ready made virtual machine images
- Deploying tools with web interfaces
- ”no I really need root access!”

If you can run on Taito – run on Taito
If not – Pouta might be for you

- Pouta user guide: *https://research.csc.fi/pouta-user-guide*

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
  - Deadline: *May 30th, 2014, at 12:00*
- Remember also PRACE/DECI

Courses

- Sisu Phase 2 workshop
  - Possibly Autumn 2014
- Taito Phase 2 workshop
  - Likely in early 2015
- CSC courses: *http://www.csc.fi/courses*
  - CSC HPC Summer School
  - Spring, Autumn, Winter Schools

CSC Phase2 resources' summary

- ***Sisu** supercomputer*
  - Installation planned in *July-August 2014*
  - General availability planned for **Q3 2014**
- ***Taito** supercluster*
  - Installation planned in *Q4 2014*
- ***Bull*** system
  - General availability planned for **Q3 2014**
  - *45 nodes* with *2 Intel Xeon Phi coprocessors* each
  - *38 nodes* with *2 NVIDIA Tesla K40 accelerators* each
- ***DDN*** HPC storage system
  - Added *1.9 PB*, in *Q3 2014* totaling *4 PB of fast parallel storage*
  - Supports Cray and HP systems, aggregate bandwidth > *80 GB/s*

Round robin

Atte Sillanpää
CSC – IT Center for Science Ltd.

Round robin

- 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?
  - 1.5 TB/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?

Feedback form

- *https://www.webropolsurveys.com/S/57665DDA29516729.par*