

# How CSC is supporting science

**Introduction to data science for researchers**

2015-05-06

Atte Sillanpää

# Outline

- Support for research
  - Stats, www, preinstalled software, Servicedesk, training (old materials, webinars)
- Hardware, access: lots of capacity
  - Batch jobs, interactive taito-shell, NoMachine remote desktop
- Data storage & sharing
  - Directories, best practices, IDA

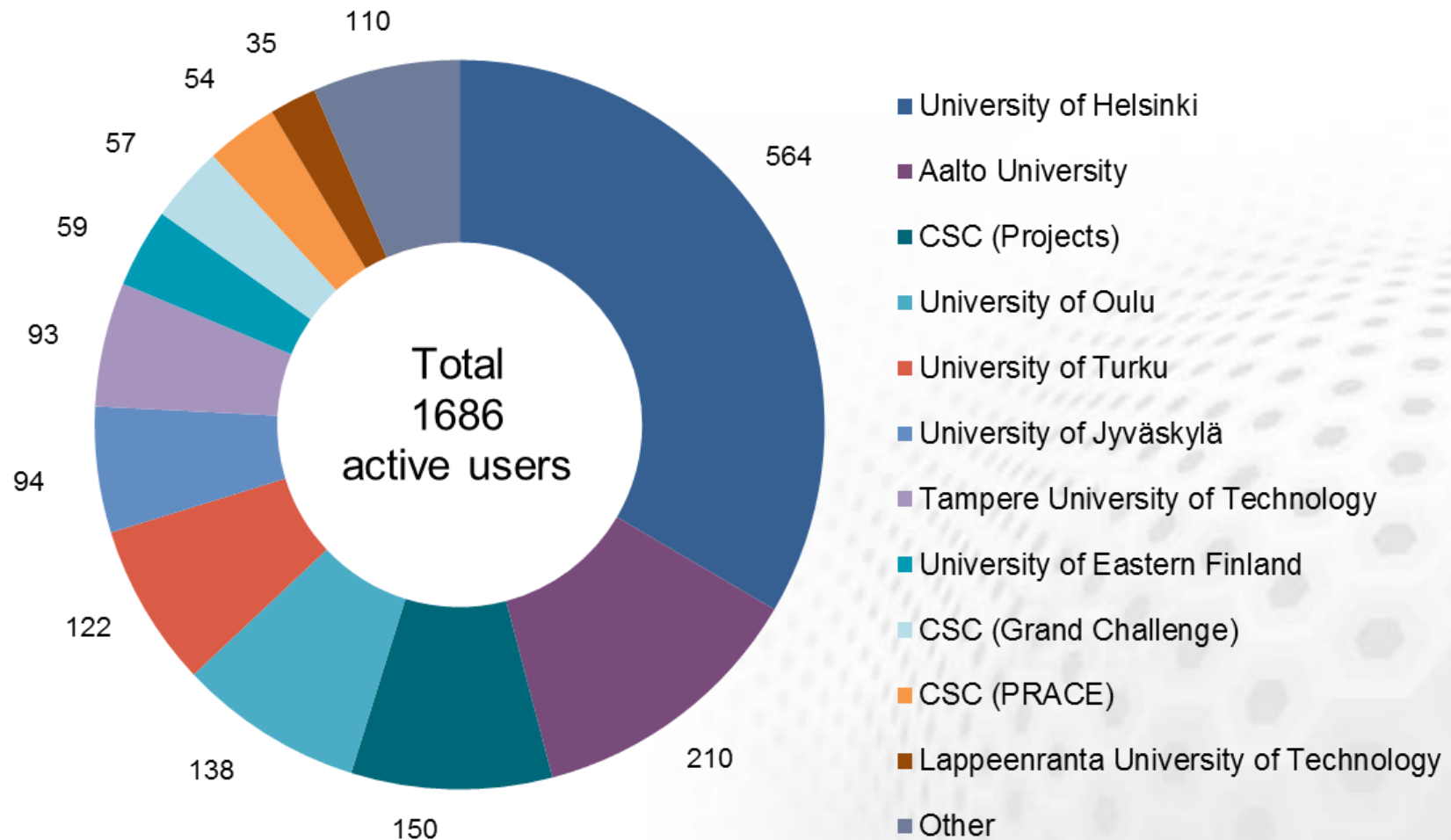
# SUPPORT FOR RESEARCH



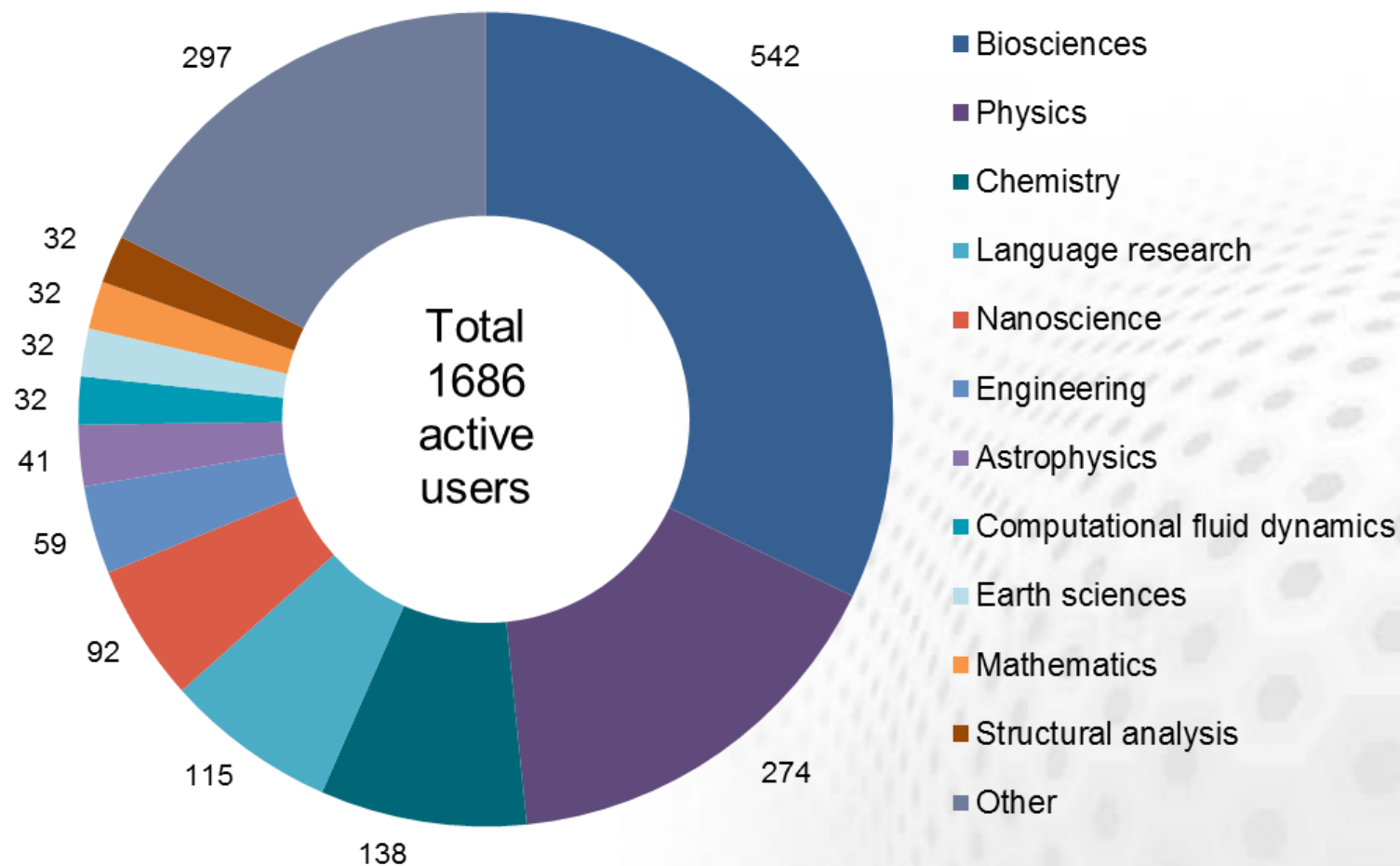
## CSC's services in numbers in 2014

- 1686 active users on CSC's computing servers
  - 5200 users in Scientist's User Interface
- 3054 participants in 92 courses/events
- 7466 user tickets resolved
- 331M billing units of computing capacity (~2x core hours)
- 164 officially supported applications (+ more)

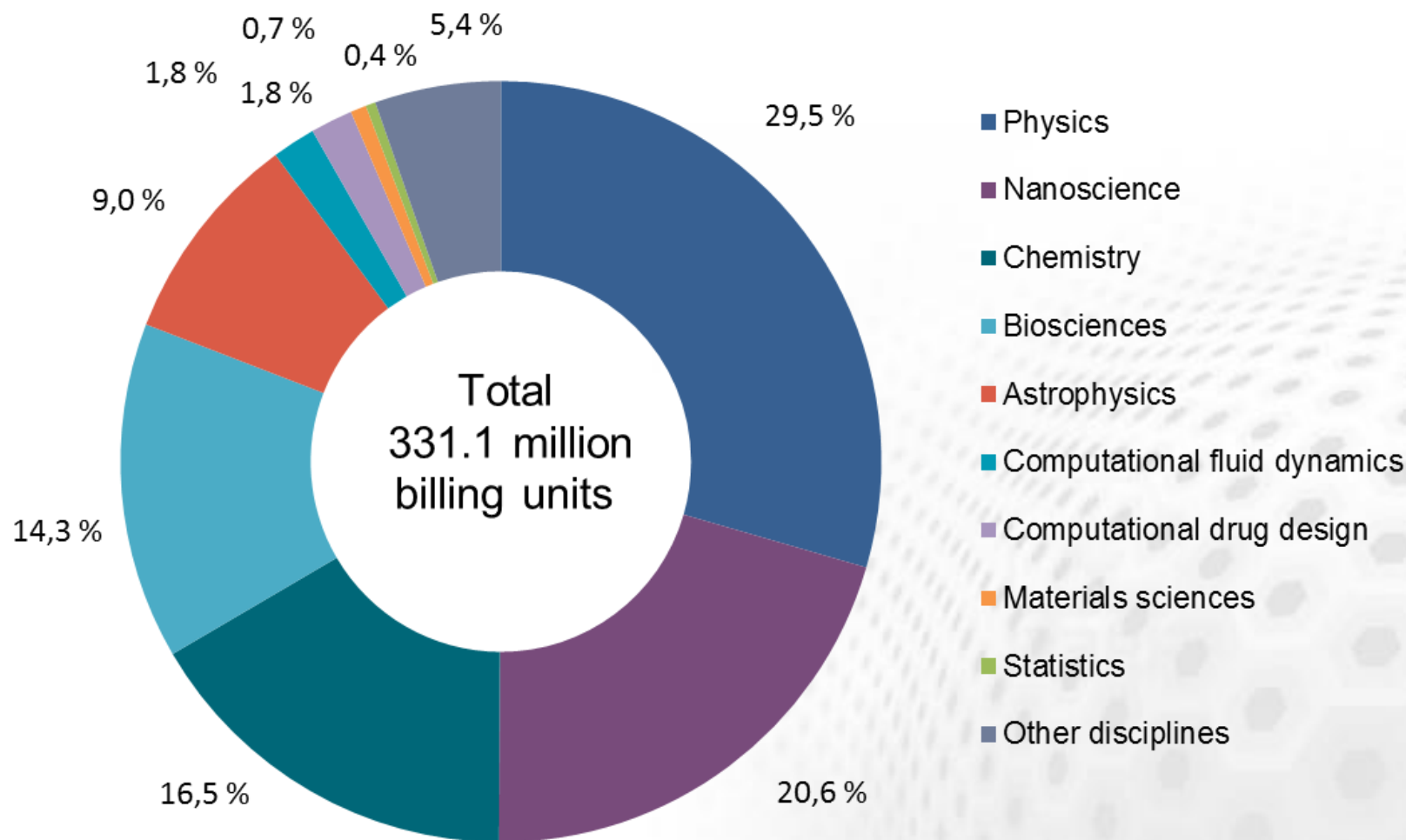
# Users of computing resources by organization 2014



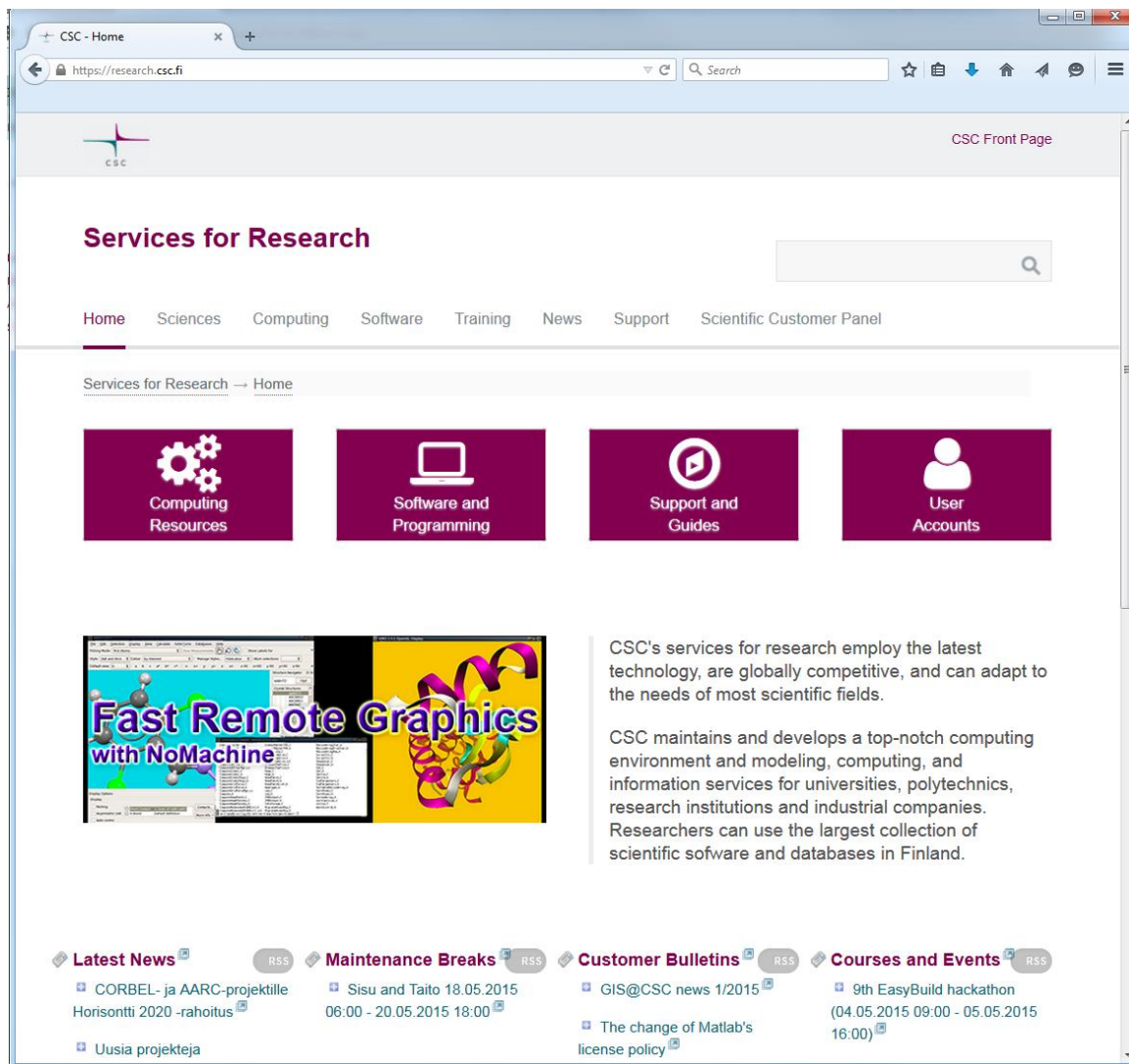
## Users of computing resources by discipline 2014



## Computing usage by discipline 2014



# https://research.csc.fi



The screenshot shows the CSC Front Page in a web browser. The browser's address bar displays "https://research.csc.fi". The page features the CSC logo in the top left and a search bar in the top right. Below the logo, the text "Services for Research" is prominently displayed. A navigation menu includes links for Home, Sciences, Computing, Software, Training, News, Support, and Scientific Customer Panel. The main content area is divided into four purple boxes with white icons and text: "Computing Resources" (gears icon), "Software and Programming" (laptop icon), "Support and Guides" (play button icon), and "User Accounts" (person icon). Below these boxes, there is a section titled "Fast Remote Graphics with NoMachine" featuring a screenshot of a remote desktop session and a 3D molecular model. To the right of this section, a paragraph states: "CSC's services for research employ the latest technology, are globally competitive, and can adapt to the needs of most scientific fields. CSC maintains and develops a top-notch computing environment and modeling, computing, and information services for universities, polytechnics, research institutions and industrial companies. Researchers can use the largest collection of scientific software and databases in Finland." At the bottom of the page, there are four RSS feed links: "Latest News", "Maintenance Breaks", "Customer Bulletins", and "Courses and Events". Each link is accompanied by a small RSS icon and a list of recent updates.

CSC Front Page

## Services for Research

Home Sciences Computing Software Training News Support Scientific Customer Panel

Services for Research → Home

- Computing Resources
- Software and Programming
- Support and Guides
- User Accounts

### Fast Remote Graphics with NoMachine

CSC's services for research employ the latest technology, are globally competitive, and can adapt to the needs of most scientific fields.

CSC maintains and develops a top-notch computing environment and modeling, computing, and information services for universities, polytechnics, research institutions and industrial companies. Researchers can use the largest collection of scientific software and databases in Finland.

[Latest News](#) [RSS](#)
[Maintenance Breaks](#) [RSS](#)
[Customer Bulletins](#) [RSS](#)
[Courses and Events](#) [RSS](#)

- CORBEL- ja AARC-projektille Horisontti 2020 -rahoitus
- Sisu and Taito 18.05.2015 06:00 - 20.05.2015 18:00
- GIS@CSC news 1/2015
- The change of Matlab's license policy
- 9th EasyBuild hackathon (04.05.2015 09:00 - 05.05.2015 16:00)
- Uusia projekteja



# Software and database offered by CSC



- Large selection (over 200) of software and database packages for research <https://research.csc.fi/software>
- Mainly for academic research in Finland
- Centralized national offering: software consortia, better licence prices, continuity, maintenance, training and support

## Services for Research

Home Sciences Computing **Software** News Support Sci

Services for Research — Software — Software Packages

### Software

#### Software Packages

#### Programming

#### Parallel Computing

#### Code Optimization

#### Open Source Software

#### Development at CSC

### Software Package

All software packages in alphabet

#### Title

Abaqus

ABYSS

Acquis Communautaire Multiling

ADF

afterburner

Ajatella, Miettä, Pohtia, Harkita

ALLPATHS-LG

Amber

ANSYS Academic Research

ANSYS Academic Teaching Intr

ANSYS CFX

ANSYS Fluent

ANSYS ICEM CFD

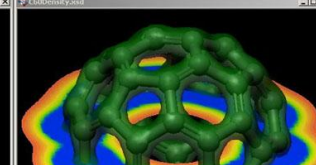
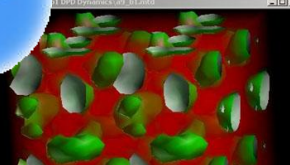
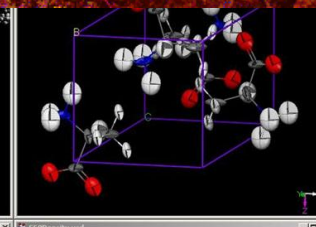
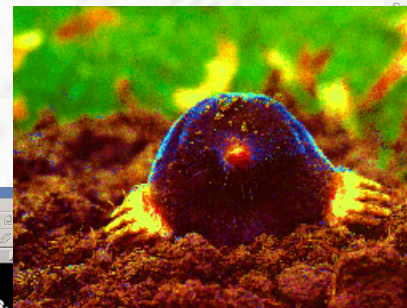
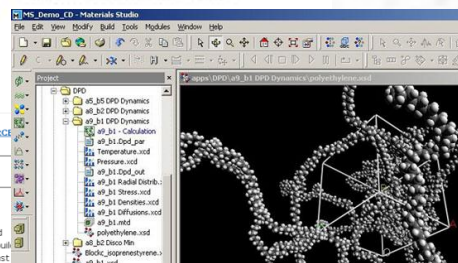
ARB

ArcGIS

AutoDock

Babel

BEAST



# Services for Research

Home Sciences Computing **Software** Training News Support Scientific Customer Panel

Services for Research → Software → Software Packages → Software details → R

## Software

Software Packages +

Programming

Parallel Computing

Code Optimization +

Visualization +

Open Source  
Software  
Development at CSC

## R

### Description

R is a free software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R also provides a wide range of tools and methods for the analysis and comprehension of next generation sequencing (NGS), microarray and proteomics data in [Bioconductor](#) packages. For more information, see [CRAN's R Manuals](#).

In Taito you can find the Rmpi-package, which provides an interface to MPI APIs. This package allow you to run R programs in parallel across multiple processors and to accomplish a goal more quickly than running a single program on one machine.

Taito also includes RStudio which is a powerful and productive graphical user interface for R. To use RStudio you must have a graphical connection to the server. The recommended way is to use [NoMachine remote desktop](#).

### Versions

The following versions are on Taito:

- R 2.15.3
- R 3.0.0
- R 3.0.1
- R 3.0.2
- R 3.1.0
- R 3.1.1
- R 3.1.2

### Usage

In Taito and Taito-shell, to load the latest installed version of R use the following command. Please note that execution of the command unloads all current modules. Thus, if your R session depends on some atypical modules, these need to be setup following the R-environment startup.

## Support

Guides +

FAQ Knowledge Base +

## Courses

User Accounts and Projects +

Open Calls for Resources +

Grid Certificates +

Service Breaks

## Courses

### ■ 9th EasyBuild hackathon (04.05.2015 09:00 - 05.05.2015 16:00)

Introduction to the EasyBuild framework for installing software on HPC systems followed by a hackathon to develop EasyBuild support for new applications or systems.

### ■ ELIXIR Finland launch (04.05.2015 15:00 - 18:00)

Let's celebrate the launch of the ELIXIR Finland node! In the event, you will hear about ELIXIR - European infrastructure for biological information, and how our areas of specialization in Finland integrate to the European whole.

### ■ Introduction to data science for researchers (05.05.2015 09:00 - 06.05.2015 16:15)

This course aims to give researchers a basic understanding of what data science is and how it could be used as part of research. Examples and exercises are given in R and Python, starting knowledge of one or the other (but not both) is assumed.

### ■ Digitaalisten ihmistieteiden aamu (12.05.2015 10:00 - 12:00)

Digitaalisten ihmistieteiden aamu tiistaina 12.5. klo 10-12 Helsingin yliopiston Tiedekulmassa (Porthania)

### ■ Webinar: Introduction to PANNZER functional annotation server and how use it at CSC (12.05.2015 10:30 - 11:00)


This webinar introduces the new functional annotation server developed at the University of Helsinki and shows how to run PANNZER jobs at CSC's Taito server.

### ■ Introduction to C programming (18.05.2015 09:00 - 20.05.2015 16:00)


This course will cover the basics of C programming language.

## Courses and Events on Twitter!


Tweets
Follow

 @CSCfi  
You call this the Spring Festival? Our #datacenter in #Kajaani #Finland keeps it cool.

Happy May Day!  
#vappu2015  
pic.twitter.com/uWU2jdPu2o  
Retweeted by Courses at CSC



Expand

 **Courses at CSC** 27 Apr  
@CoursesAtCSC

Tweet to @CoursesAtCSC

[iCal export](#) [More](#)

## Spring School in Computational Chemistry 2015 @ CSC

10-13 March 2015  
CSC - IT Center for Science  
Europe/Helsinki timezone

[Overview](#)[Timetable](#)[Support](#)[patc@csc.fi](mailto:patc@csc.fi)

### Slides

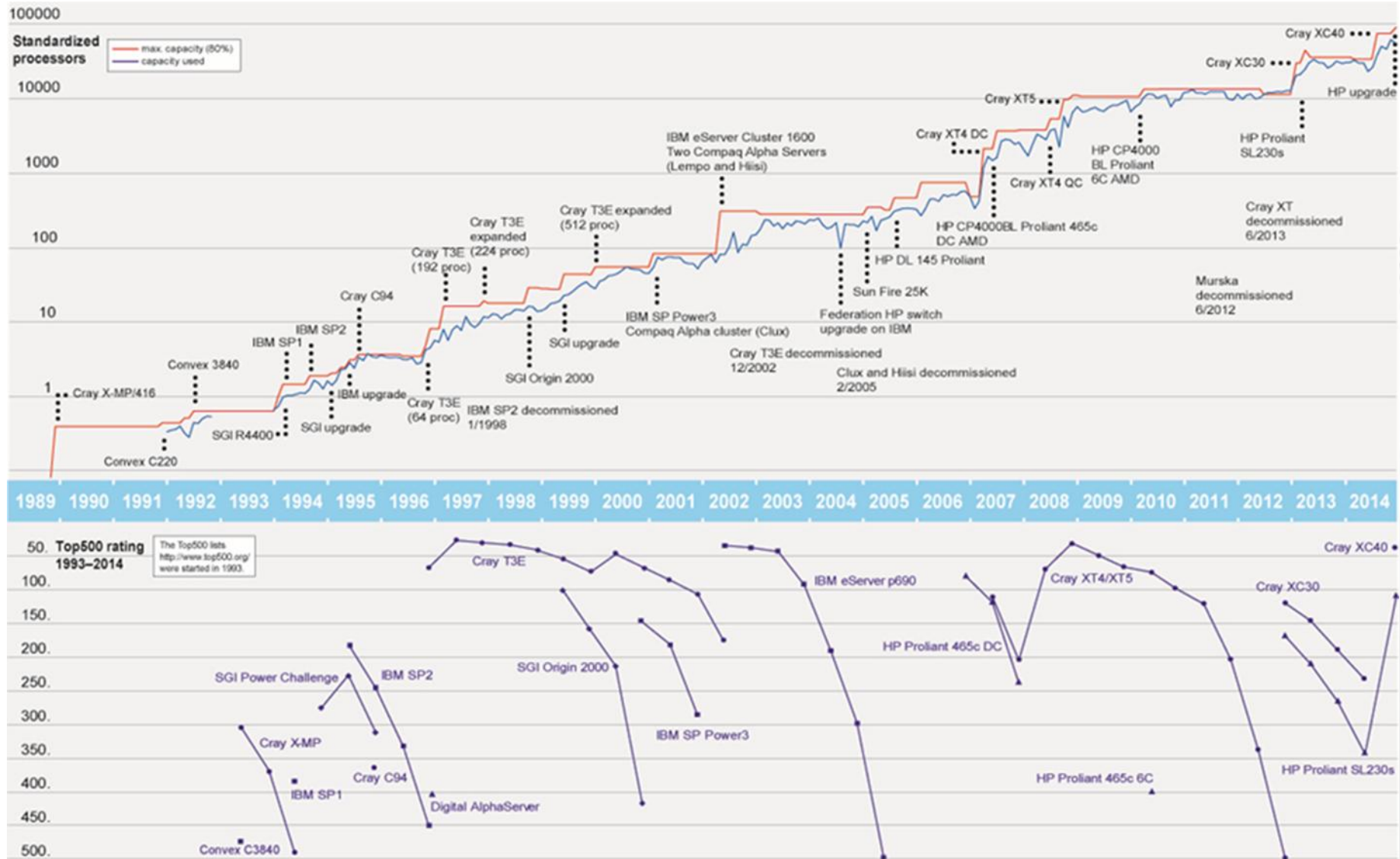
-  [CSC2015-EnhancedSampling.pdf](#)  
Accelerated/Enhanced Sampling Techniques  
**File name:** CSC2015-EnhancedSampling.pdf  
**File size:** 4 MB  
**File creation date:** 12 Mar 2015 10:18
-  [CSC2015-IntroMD.pdf](#)  
Introduction to Classical MD  
**File name:** CSC2015-IntroMD.pdf  
**File size:** 5 MB  
**File creation date:** 10 Mar 2015 09:44
-  [CSC\\_SSCC\\_2015\\_Karttunen.pdf](#)  
Predicting physical properties of crystalline solids  
**File name:** CSC\_SSCC\_2015\_Karttunen.pdf  
**File size:** 3 MB  
**File creation date:** 13 Mar 2015 14:36
-  [IES1.pdf](#)  
Introduction to electronic structure theory 1  
**File name:** IES1.pdf  
**File size:** 5 MB  
**File creation date:** 13 Mar 2015 14:37
-  [IES2.pdf](#)  
Introduction to electronic structure theory 2  
**File name:** IES2.pdf  
**File size:** 9 MB  
**File creation date:** 13 Mar 2015 14:37
-  [Patzschke\\_csc\\_vmd\\_2015.pdf](#)  
Introduction to VMD  
**File name:** Patzschke\_csc\_vmd\_2015.pdf



# HARDWARE, ACCESS, DEMO



# CSC Computing Capacity 1989–2014



# 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 XE40)
- Parallel from 72 up to thousands of cores
- Scaling tests 1008+

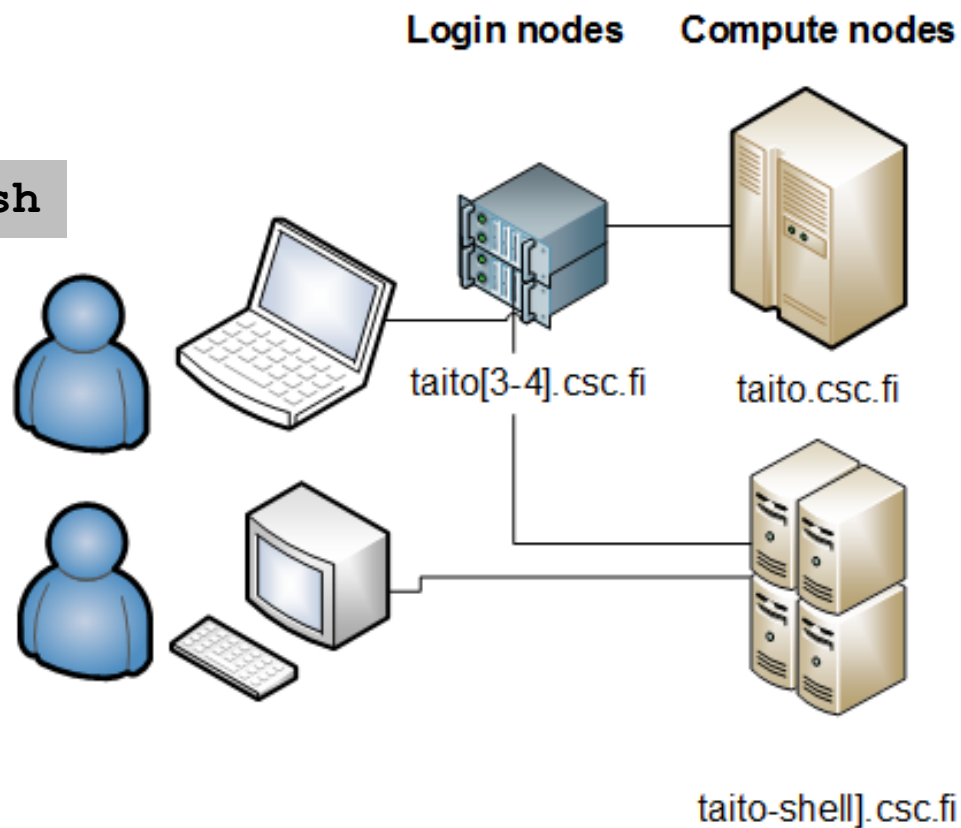
- cPouta (HP) Cloud
- Serial and parallel upto 16 cores

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

# Compute nodes are used via queuing system

```
$ sbatch job_script.sh
```

```
$ ./my_prog &
```





# Batch jobs: what and why



- User has to specify necessary resources
  - Can be added to the batch job script or given as command line options for sbatch (or a combination of script and command line options)
- Resources need to be adequate for the job
  - Too small memory reservation will cause the job to fail
  - When the time reservation ends, the job will be terminated whether finished or not
- But: Requested resources can affect the time the job spends in the queue
  - Especially number of cores and memory reservation
  - Don't request extra "just in case" (time is less critical than memory wrt this)
- So: Realistic resource requests give best results
  - Not always easy to know beforehand
  - Usually best to try with smaller tasks first and check the used resources
  - You can check what was actually used with the `sacct` command

## Example serial batch job script on Taito

```
#!/bin/bash -l
#SBATCH -J myjob
#SBATCH -e myjob_err_%j
#SBATCH -o myjob_output_%j
#SBATCH --mail-type=END
#SBATCH --mail-user=a.user@foo.net
#SBATCH --mem-per-cpu=4000
#SBATCH -t 02:00:00
#SBATCH -n 1
#SBATCH -p serial
#SBATCH --constraint=snb
```

Submit with:  
sbatch script.bash

```
module load R.latest
srun R --no-save < myscript.R
```

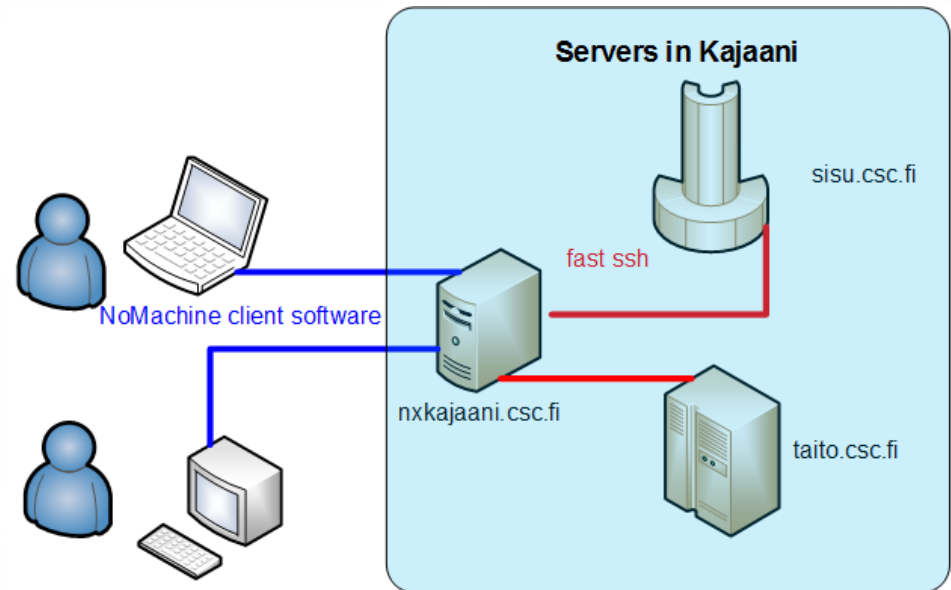
## Example parallel batch job script on Taito

```
#!/bin/bash -l
#SBATCH -J myjob
#SBATCH -e myjob_err_%j
#SBATCH -o myjob_output_%j
#SBATCH --mail-type=END
#SBATCH --mail-user=a.user@foo.net
#SBATCH --mem-per-cpu=4000
#SBATCH -t 02:00:00
#SBATCH -n 32
#SBATCH -p parallel
#SBATCH --constraint=snb
```

```
module load R.latest
srun Rmpi --no-save < myscript.R
```

# NoMachine Remote Desktop

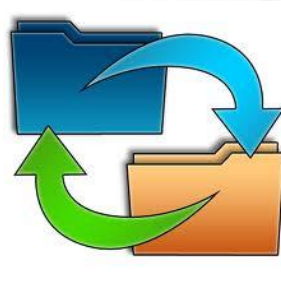
- Client connection between user and gateway
- Good performance even with slow network
- Ssh from gateway to server (fast if local)
- Persistent connection
- Suspendable
  - Continue later at another location
- Read the [instructions](#)...
  - ssh-key, keyboard layout, mac specific workarounds, ...
- Choose an application or server to use (right click)



## Demo: interactive R at CSC

- Log in to taito-shell (using NoMachine)
- Load R and RStudio modules
- Start interactive RStudio session
- Useful if you need more resources than you have on a local computer
- If you need dedicated resources or want to run a batch job, check:
- <https://research.csc.fi/-/r>

# STORING AND SHARING DATA



# Moving files, best practices



Space!

A purple arrow originates from the 'Space!' text and points towards the command line in the second list item.

- tar & bzip first (bzip more error tolerant)
- rsync, not scp (when lots of/big files)
  - \$ `rsync -P username@taito-login3.csc.fi:/tmp/huge.tar.gz .`
- Funet FileSender (max 50 GB [50GB as an attachment? No!])
  - <https://filesender.funet.fi>
  - Files can be downloaded also with **wget**
- iRODS, batch-like process, staging
- IDA: <http://www.tdata.fi/ida>
- CSC can help to tune e.g. TCP/IP parameters
  - <http://www.csc.fi/english/institutions/funet/networkservices/pert>
- FUNET backbone 10 Gbit/s
- More info in [CSC computing environment Guide](#)

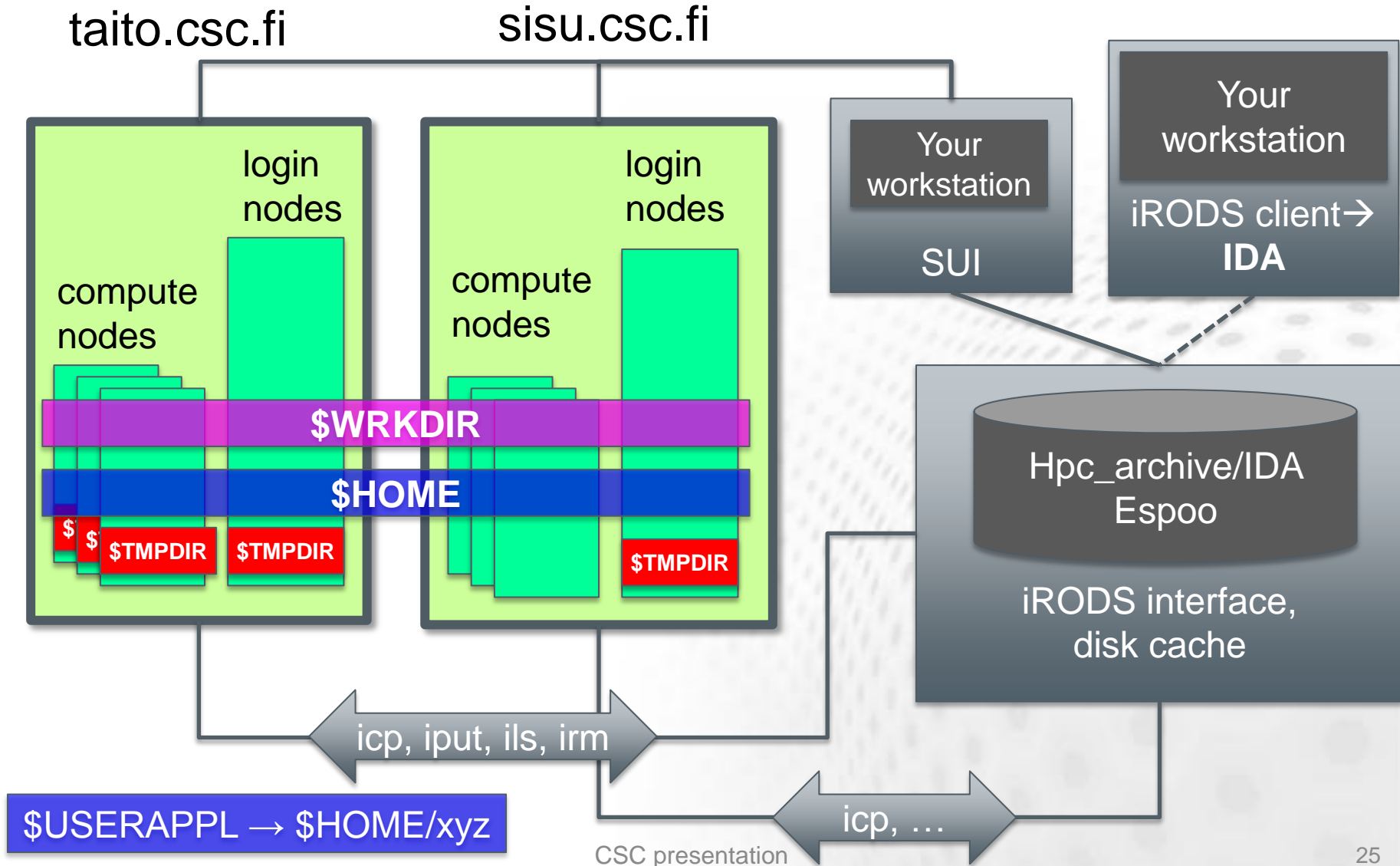
# Directories at CSC Environment (1)

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

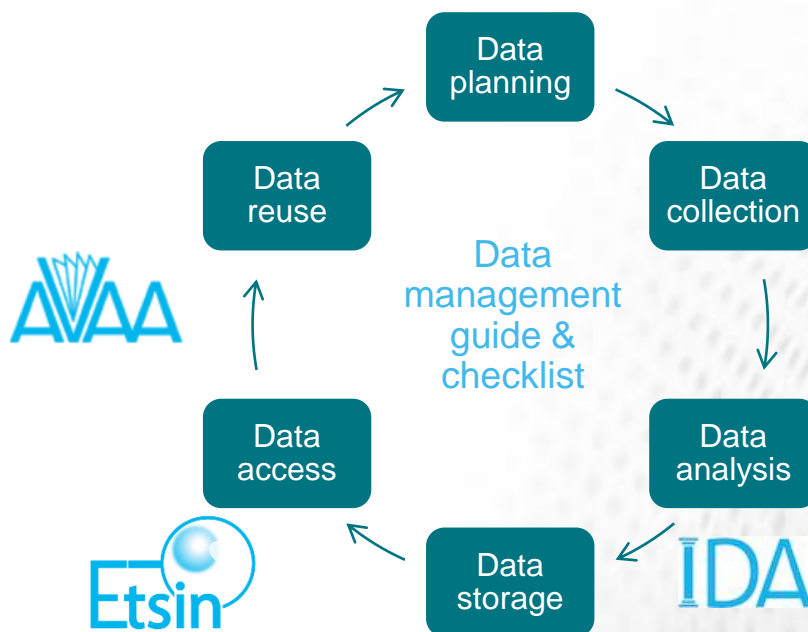
<sup>1</sup>: Lustre parallel (<sup>3</sup>:local) file system in Kajaani    <sup>2</sup>: iRODS storage system in Espoo



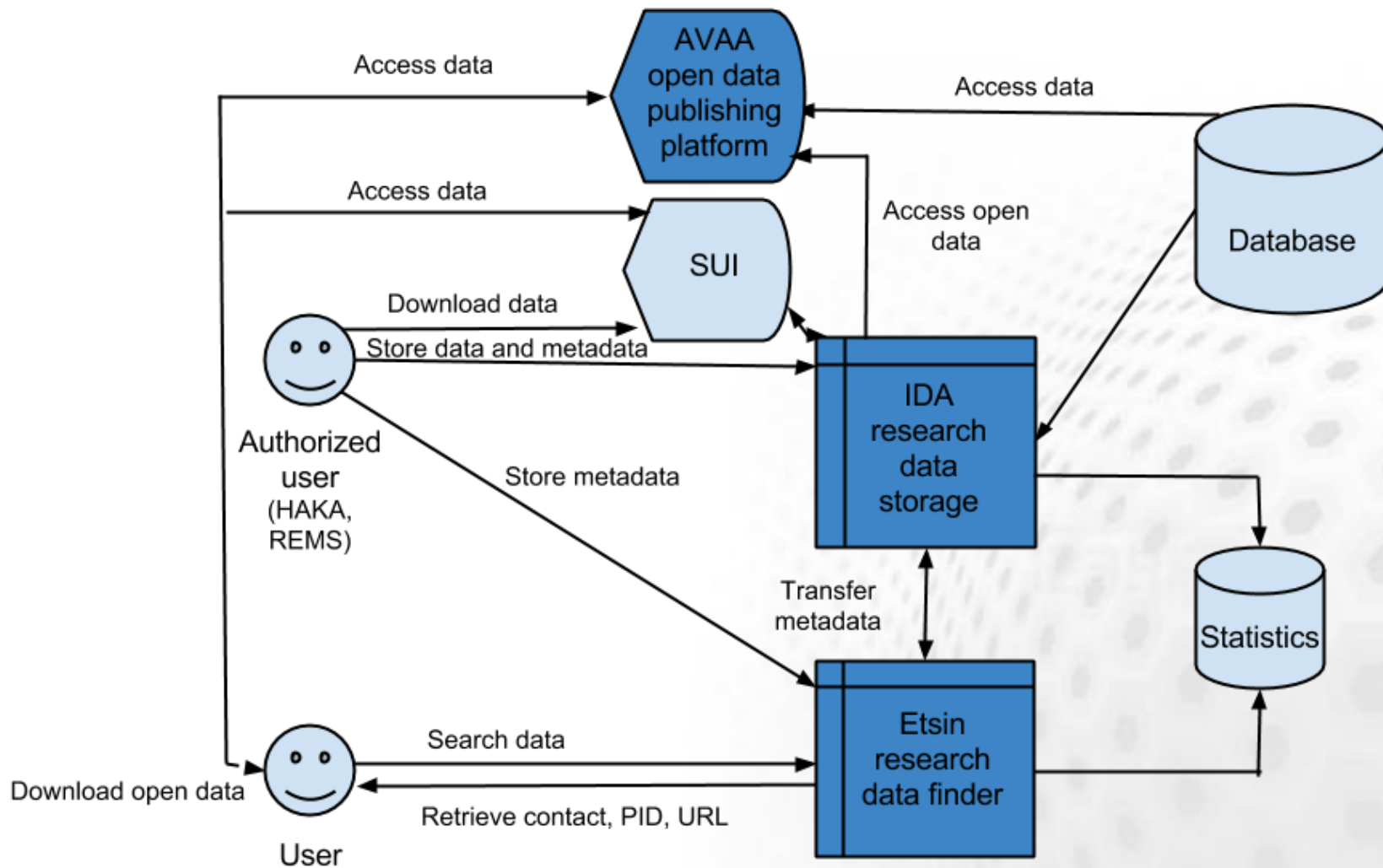
# Directories at CSC Environment (2)



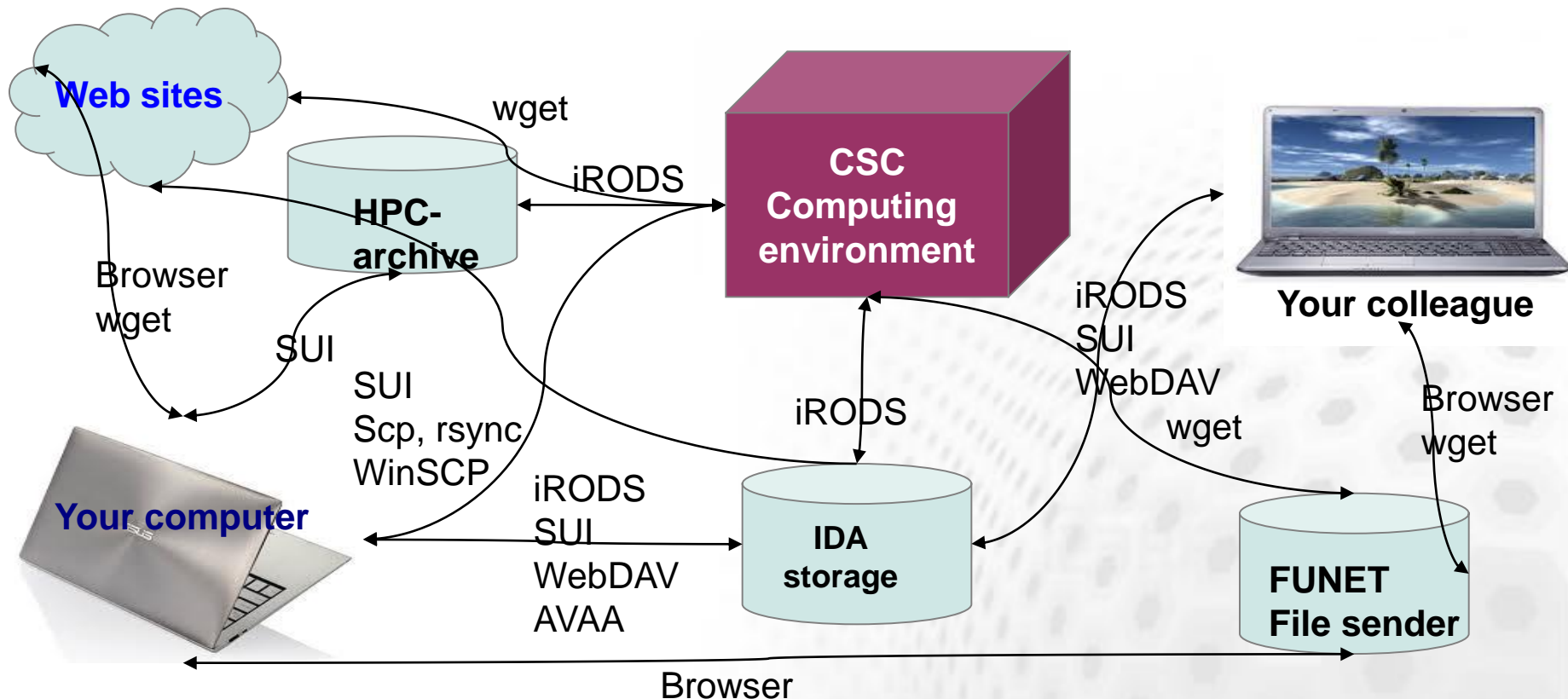
# Research data lifecycle and open science services produced by CSC



# Use of open science services



# Moving data to and from CSC



# Thanks, questions?

