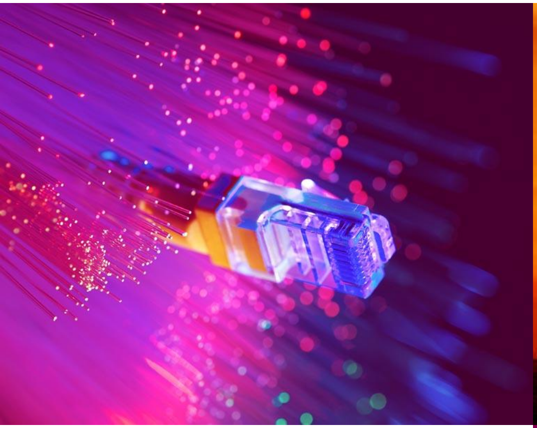
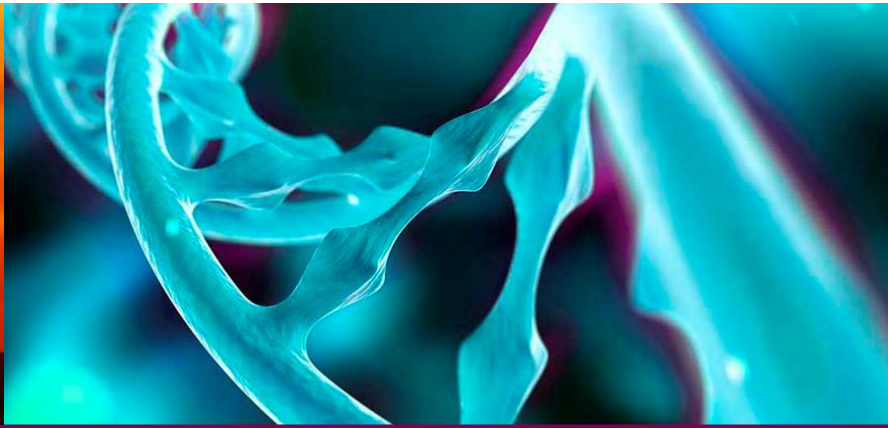




# CSC computing resources for GIS
# Kylli Ek, Eduardo Gonzalez, CSC

CSC, 8.10.20018

*CSC – Suomalainen tutkimuksen, koulutuksen, kulttuurin ja julkishallinnon ICT-osaamiskeskus*

Non-profit state organization with special tasks

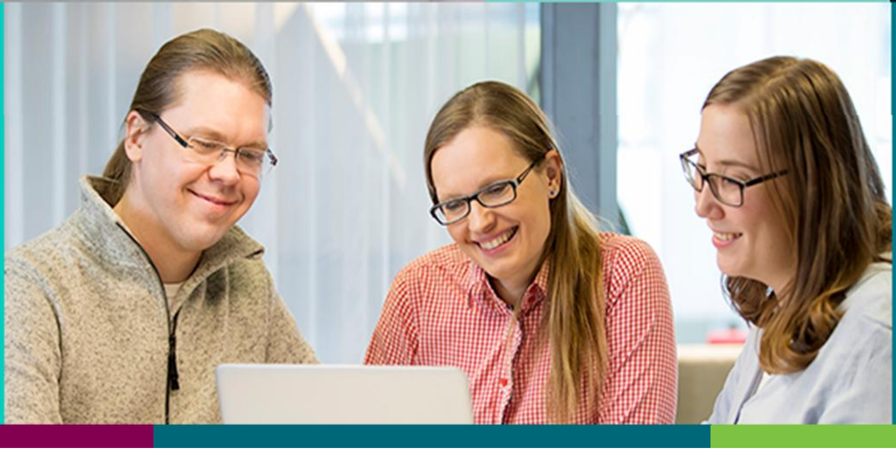
Turnover in 2017

40,5 M€

Headquarters in Espoo, datacenter in Kajaani

Owned by state (70%)

and all Finnish higher education institutions (30%)

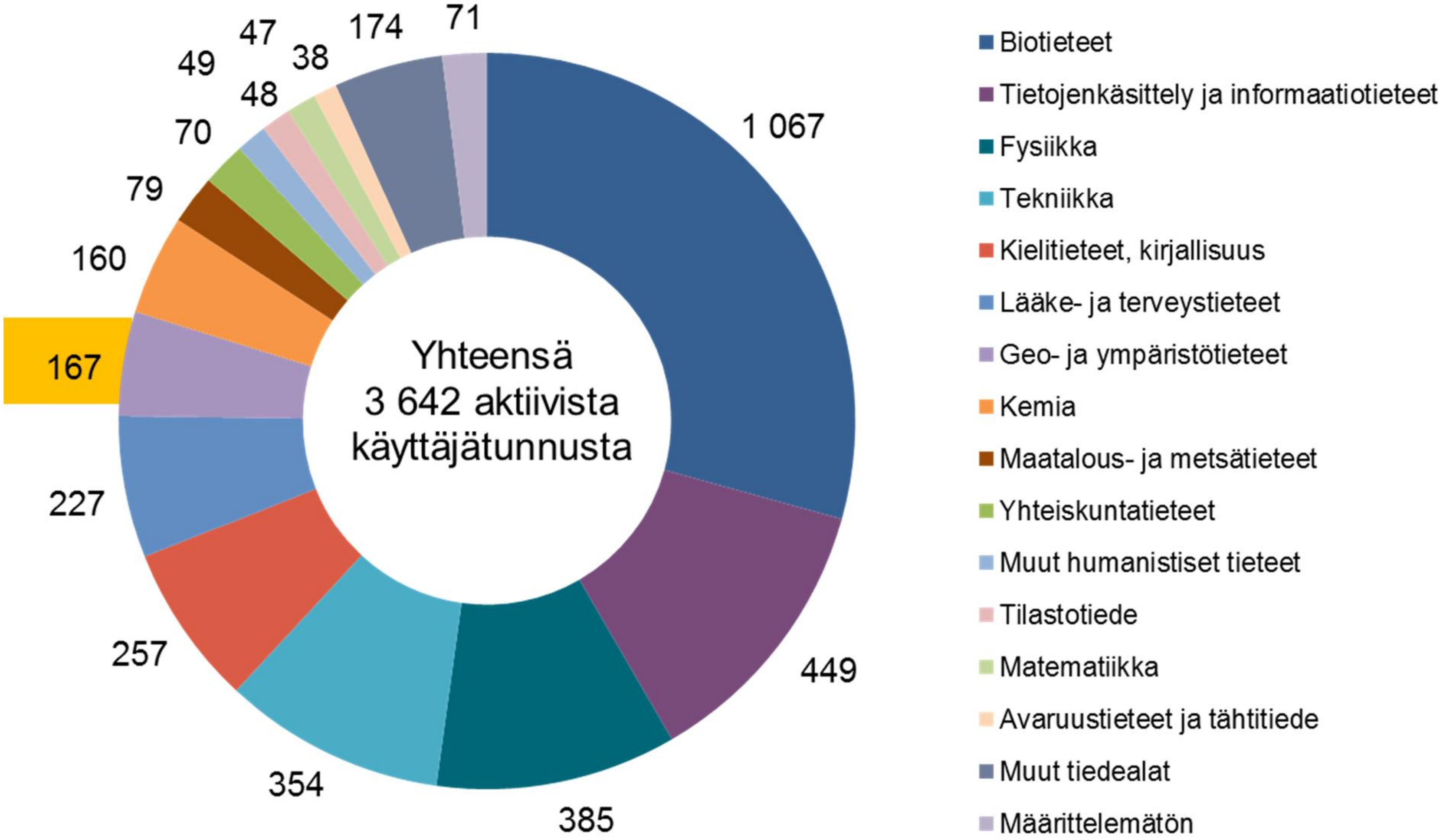


Circa

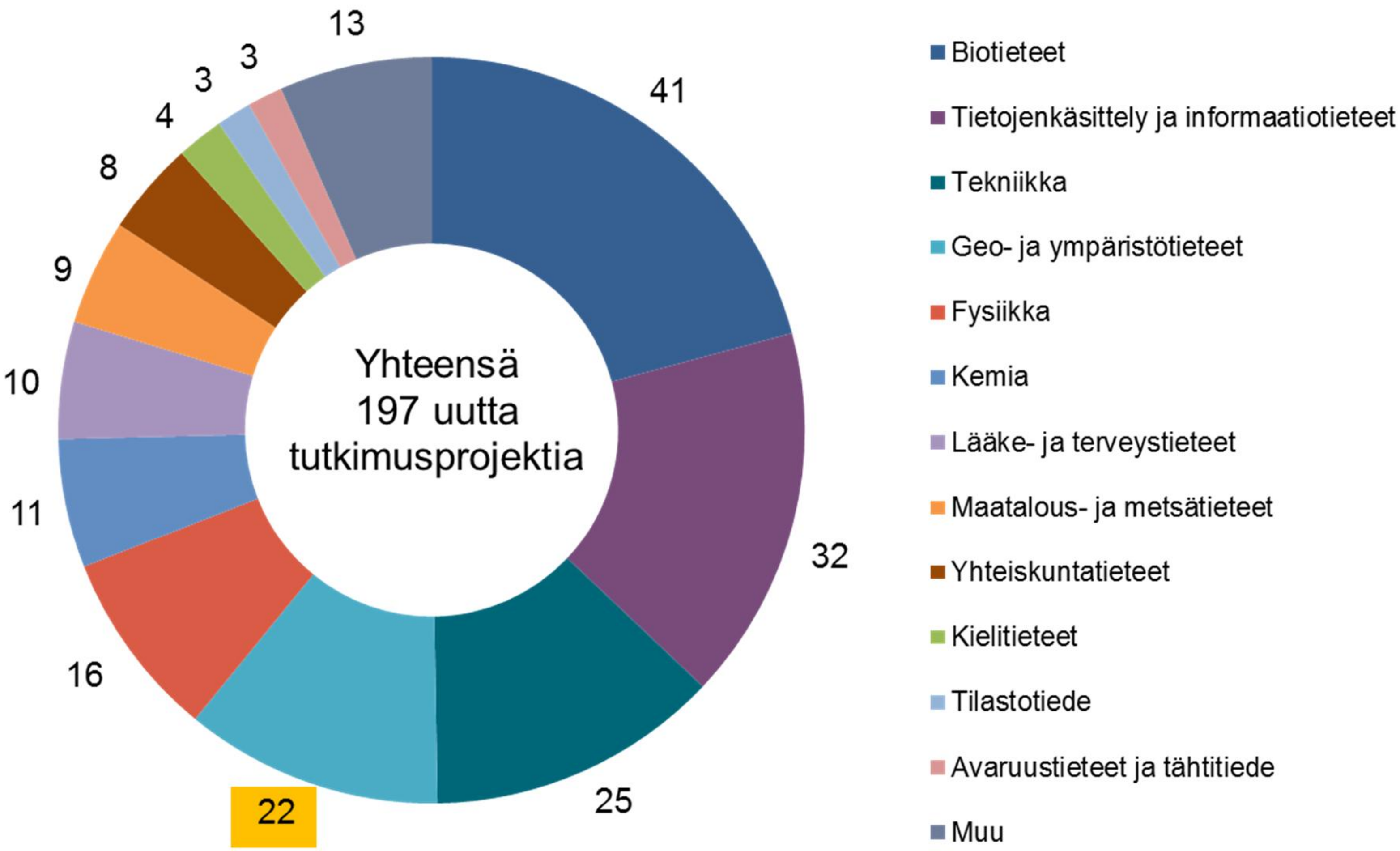
320

employees in 2017

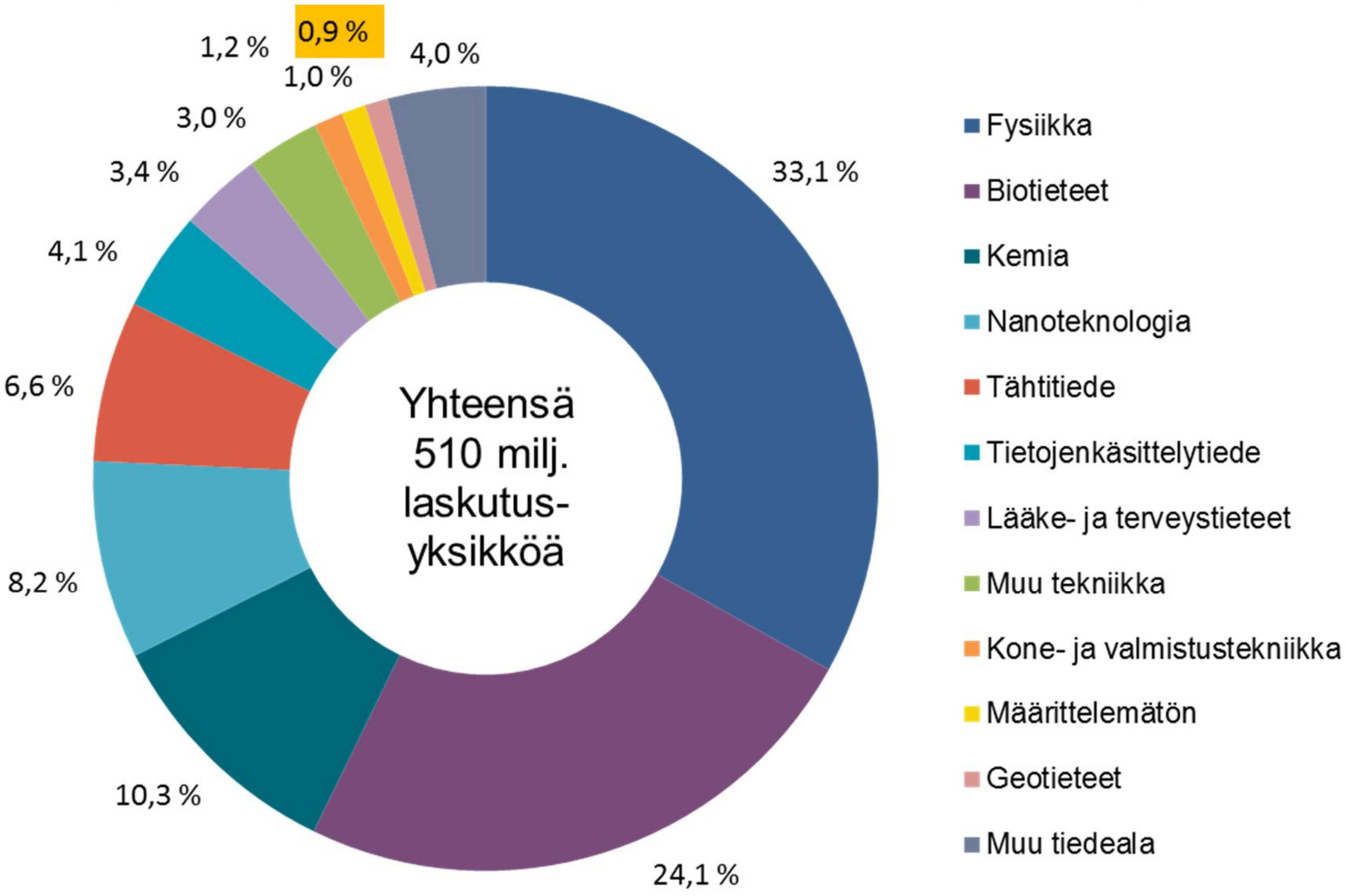
# Aktiiviset käyttäjätunnusasiakkaat tiedealoittain kesäkuun 2018 lopussa



# Uudet jaksolla H1/2018 avatut tutkimusprojektit tiedealoittain



# Tietokoneresurssien käyttö tiedealoittain kaudella H1/2018
## (sisältää Sisu-, Taito-, Taito-shell, cPouta ja ePouta-käytön)

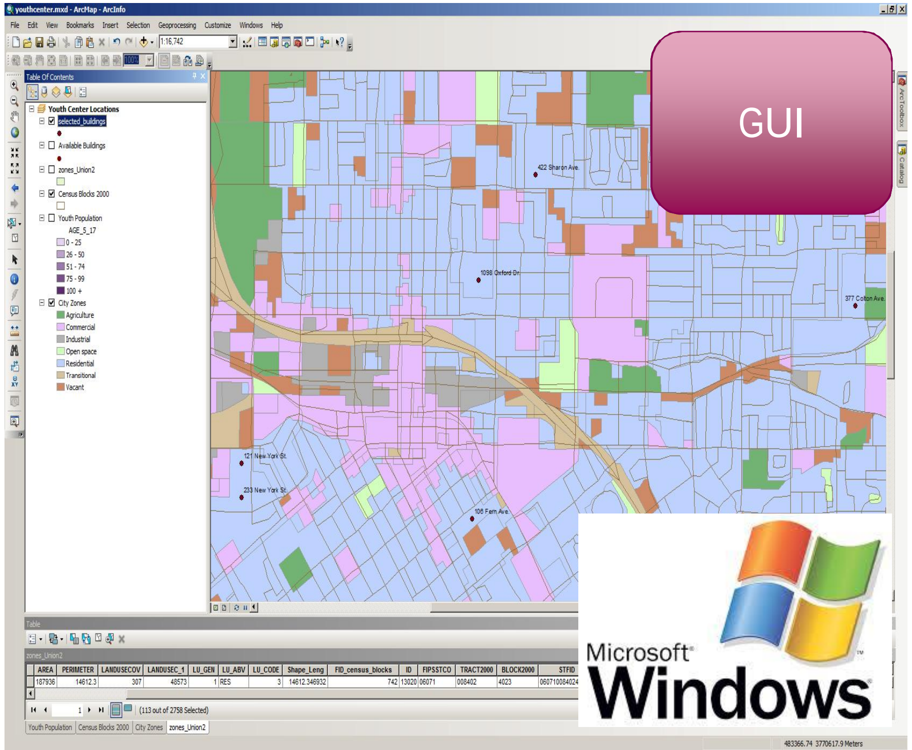


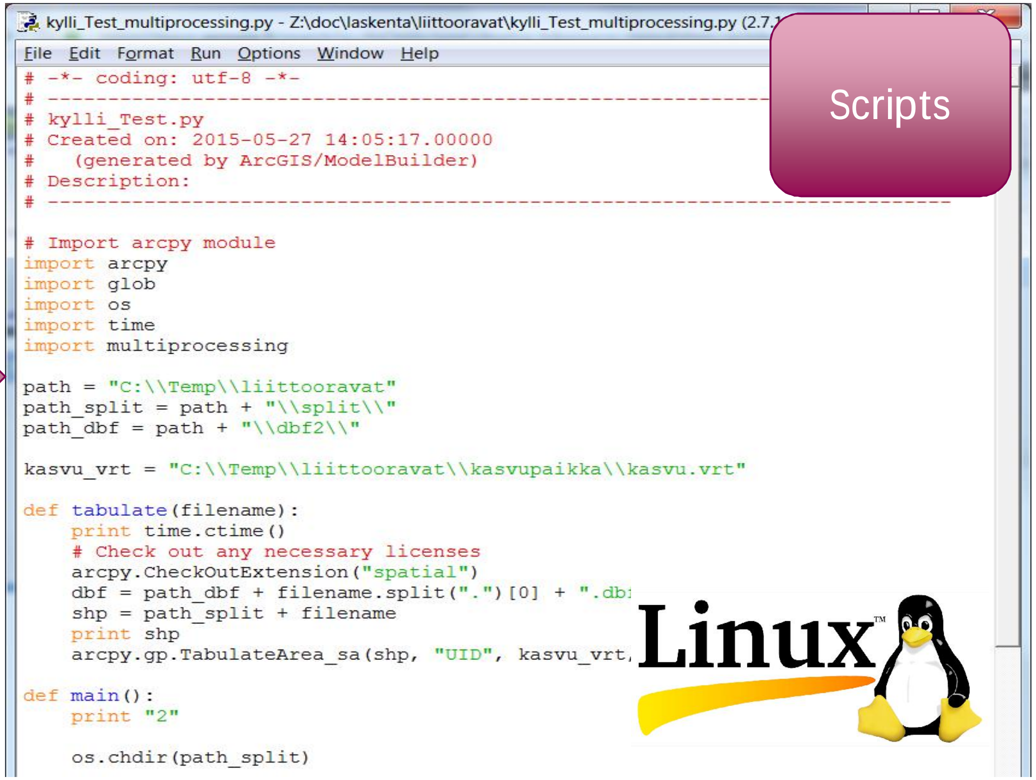
# Reasons for using CSC computing resources

- Computing something takes more than 2-4 hours
- Need for more memory
- Very big datasets
- Keep your desktop computer for normal usage, do computation elsewhere
- Need for a server computer
- Need for a lot of computers with the same set-up (courses)

- Free for Finnish university users / will be free for state research insitutes

# The keys to geocomputing: Change in working style & Linux




kylli_Test_multiprocessing.py - Z:\doc\laskenta\liittooravat\kylli_Test_multiprocessing.py (2.7.1
File Edit Format Run Options Window Help
Scripts
Linux

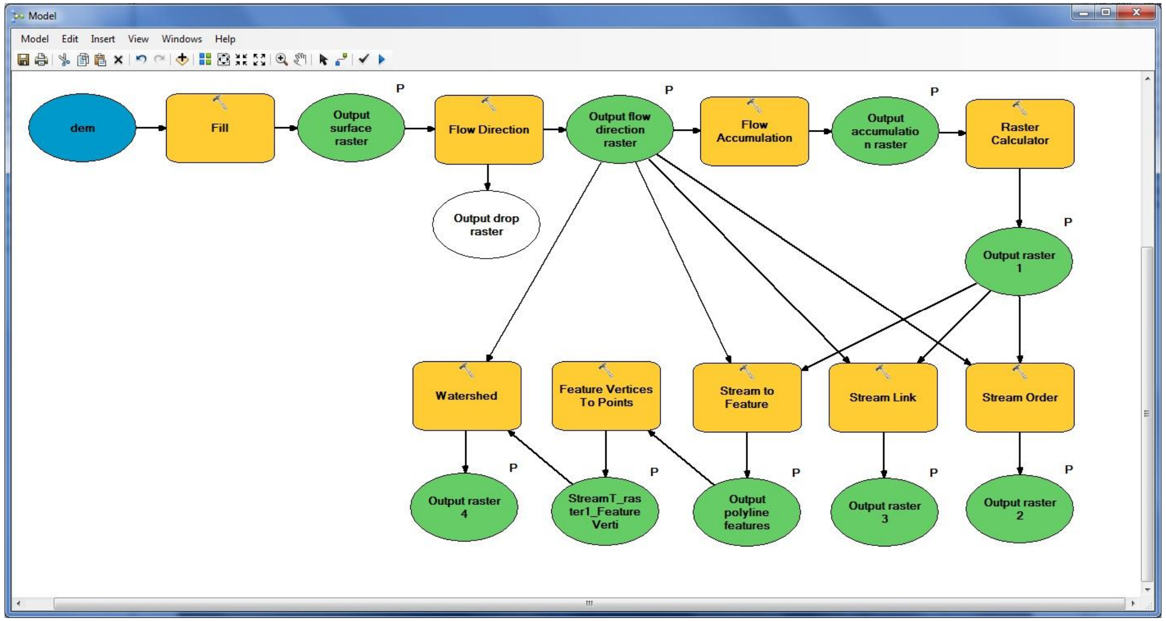
```
# -*- coding: utf-8 -*-
# ---------------------------------------------------------------------------
# kylli_Test.py
# Created on: 2015-05-27 14:05:17.00000
#   (generated by ArcGIS/ModelBuilder)
# Description:
# ---------------------------------------------------------------------------

# Import arcpy module
import arcpy
import glob
import os
import time
import multiprocessing

path = "C:\\Temp\\liittooravat"
path_split = path + "\\split\\"
path_dbf = path + "\\dbf2\\"

kasvu_vrt = "C:\\Temp\\liittooravat\\kasvupaikka\\kasvu.vrt"

def tabulate(filename):
    print time.ctime()
    # Check out any necessary licenses
    arcpy.CheckOutExtension("spatial")
    dbf = path_dbf + filename.split(".")[0] + ".db
    shp = path_split + filename
    print shp
    arcpy.gp.TabulateArea_sa(shp, "UID", kasvu_vrt,

def main():
    print "2"

    os.chdir(path_split)
```

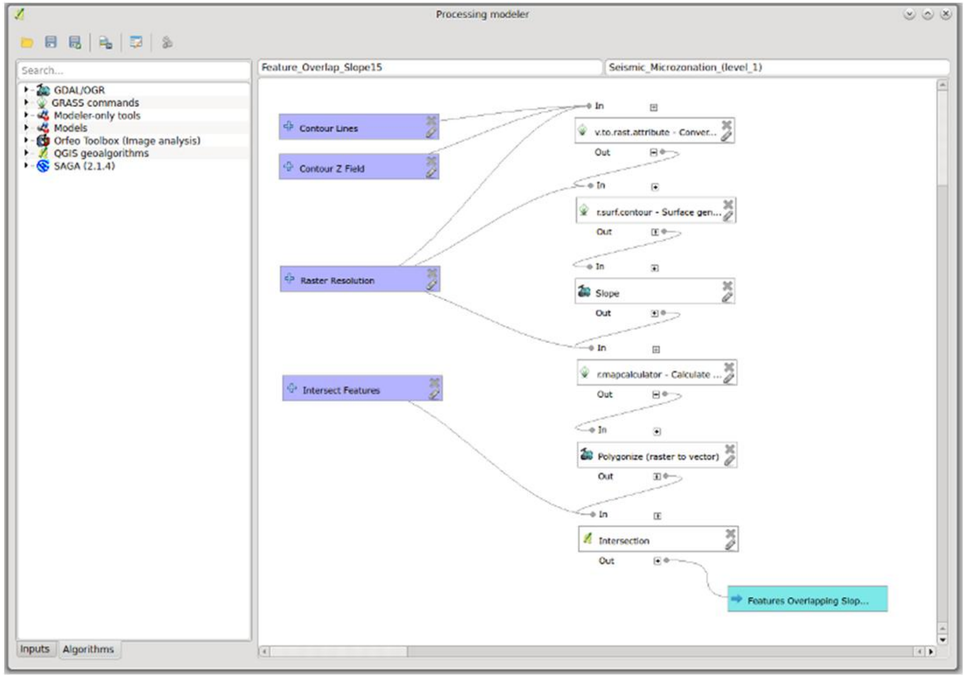
ArcGIS, QGIS, …

R, Python, shell scripts, Matlab, …

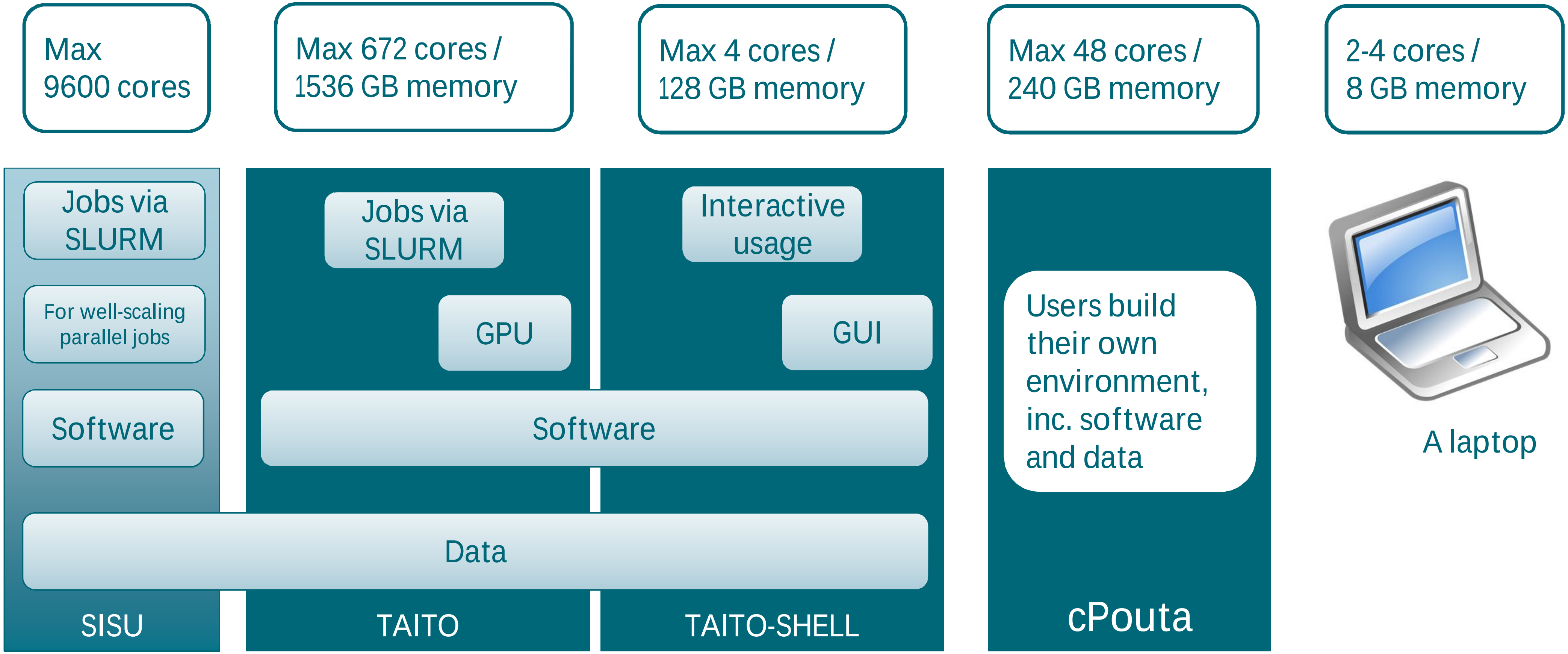
# Support for creating Python scripts

- ArcGIS Model Builderer -> ArcPy Python script
- QGIS Graphical (Processing) Modeler -> PyQGIS Python script





# CSC HPC resources



Max 9600 cores

Max 672 cores / 1536 GB memory

Max 4 cores / 128 GB memory

Max 48 cores / 240 GB memory

2-4 cores / 8 GB memory

SISU
- Jobs via SLURM
- For well-scaling parallel jobs
- Software

TAITO
- Jobs via SLURM
- GPU

TAITO-SHELL
- Interactive usage
- GUI

Software (TAITO and TAITO-SHELL)

Data (SISU, TAITO and TAITO-SHELL)

cPouta
- Users build their own environment, inc. software and data

A laptop

# Realistic expectations

- A single core of a CSC machine is about as fast as one of a basic laptop.
- It has just a lot of them.
- .. and more memory and faster input-output.

➢Just running your single core script at CSC does not make it much faster.

➢For clear speed-ups you have to use several cores.

➢… or optimize your script.

# Taito / Taito-shell pre-installed software for GIS

- R
- Python
- MatLab / Octave
- GDAL/OGR
- GRASS GIS
- LasTools (some)
- PDAL
- Proj4
- QGIS
- SagaGIS
- Taudem
- Zonation

https://research.csc.fi/software -> Geosciences

# GIS Software not available in Taito

Windows software:

- ArcGIS
- MapInfo
- LasTools Windows tools

Server software

- GeoServer, MapServer
- PostGIS

Web map libraries

- OpenLayers, Leaflet

# Using different GIS-software in Taito

| | Bash | R | Python | QGIS |
|---|---|---|---|---|
| GDAL | x | x | x | x |
| GRASS | x | x | (x) | x |
| LasTools | x | (x) | (x) | x |
| SagaGIS | x | x | (x) | x |
| Taudem | x | (x) | (x) | ? |
| R spatial packages | - | x | - | - |
| Python geo packages | - | - | x | - |

# Installing software for own use

- Possibility to install software for own use
  - o The software must be available for Linux
  - o .. and installation must be possible without root access

- You can add also packages to R and Python

# Shared data area in Taito

- Hosts large commonly used datasets
- Reduces the need to transfer data to Taito
- Located at /proj/ogiir-csc/
- All Taito users have read access.
- Only CSC personnel have write access.
- For data with open license

- If you think some other dataset should be included here, ask from servicedesk@csc.fi

All Paituli open data
+
LUKE
  Multi-source national forest inventory
NLS
  Virtual rasters for DEMs

https://research.csc.fi/gis_data_in_taito

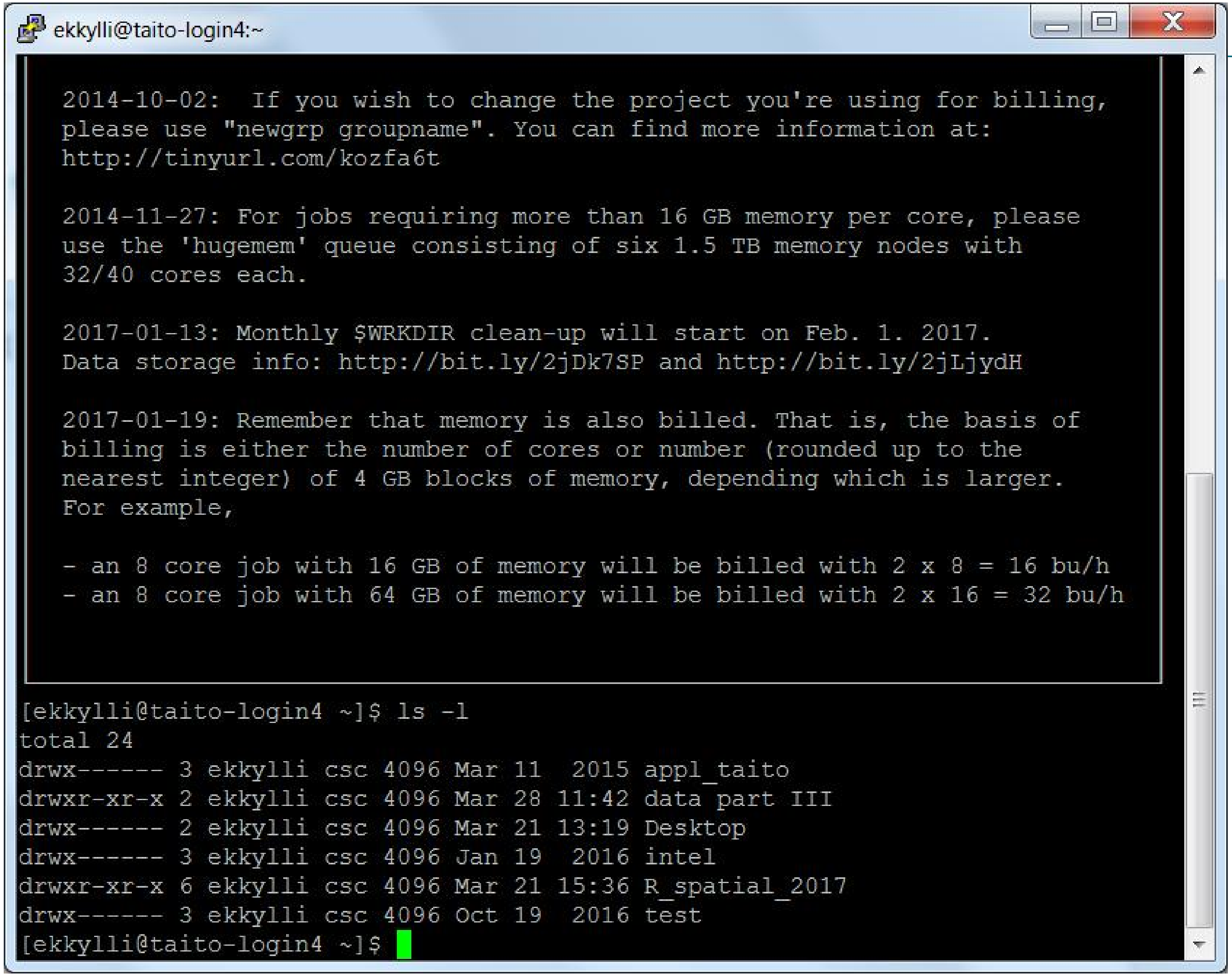
# Virtual rasters

- Allows working with dataset of multiple files as if they were a single file.
- XML pointing to actual raster files
  - The virtual file doesn't need to be rectangular, it can have holes and the source files can even have different resolutions
- Taito has ready made virtual rasters for elevation models and a python tool to create your own for a specific area.

# Access to Taito from Windows

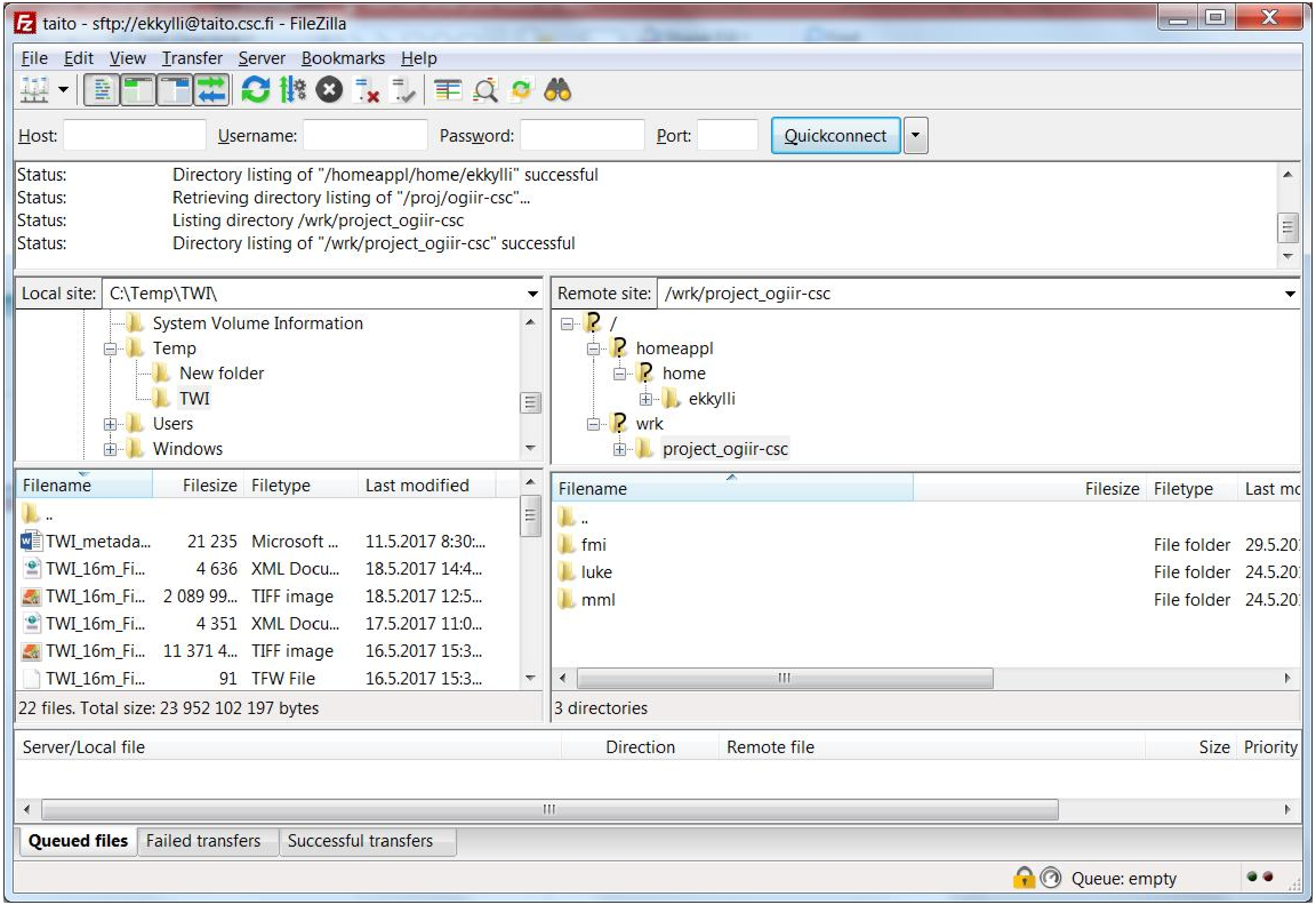
- Putty for ssh connection
- FileZilla/WinSCP for moving data
- NoMachine for GUI
- Find about other access options and more information at: https://research.csc.fi/taito-connecting

# Putty



```
2014-10-02:  If you wish to change the project you're using for billing,
please use "newgrp groupname". You can find more information at:
http://tinyurl.com/kozfa6t

2014-11-27: For jobs requiring more than 16 GB memory per core, please
use the 'hugemem' queue consisting of six 1.5 TB memory nodes with
32/40 cores each.

2017-01-13: Monthly $WRKDIR clean-up will start on Feb. 1. 2017.
Data storage info: http://bit.ly/2jDk7SP and http://bit.ly/2jLjydH

2017-01-19: Remember that memory is also billed. That is, the basis of
billing is either the number of cores or number (rounded up to the
nearest integer) of 4 GB blocks of memory, depending which is larger.
For example,

- an 8 core job with 16 GB of memory will be billed with 2 x 8 = 16 bu/h
- an 8 core job with 64 GB of memory will be billed with 2 x 16 = 32 bu/h

[ekkylli@taito-login4 ~]$ ls -l
total 24
drwx------ 3 ekkylli csc 4096 Mar 11  2015 appl_taito
drwxr-xr-x 2 ekkylli csc 4096 Mar 28 11:42 data part III
drwx------ 2 ekkylli csc 4096 Mar 21 13:19 Desktop
drwx------ 3 ekkylli csc 4096 Jan 19  2016 intel
drwxr-xr-x 6 ekkylli csc 4096 Mar 21 15:36 R_spatial_2017
drwx------ 3 ekkylli csc 4096 Oct 19  2016 test
[ekkylli@taito-login4 ~]$
```

# FileZilla



taito - sftp://ekkylli@taito.csc.fi - FileZilla

Host:
Username:
Password:
Port:

```
Status:        Directory listing of "/homeappl/home/ekkylli" successful
Status:        Retrieving directory listing of "/proj/ogiir-csc"...
Status:        Listing directory /wrk/project_ogiir-csc
Status:        Directory listing of "/wrk/project_ogiir-csc" successful
```

Local site: C:\Temp\TWI\

```
System Volume Information
Temp
    New folder
    TWI
Users
Windows
```

| Filename | Filesize | Filetype | Last modified |
|---|---|---|---|
| .. | | | |
| TWI_metada... | 21 235 | Microsoft ... | 11.5.2017 8:30:... |
| TWI_16m_Fi... | 4 636 | XML Docu... | 18.5.2017 14:4... |
| TWI_16m_Fi... | 2 089 99... | TIFF image | 18.5.2017 12:5... |
| TWI_16m_Fi... | 4 351 | XML Docu... | 17.5.2017 11:0... |
| TWI_16m_Fi... | 11 371 4... | TIFF image | 16.5.2017 15:3... |
| TWI_16m_Fi... | 91 | TFW File | 16.5.2017 15:3... |

22 files. Total size: 23 952 102 197 bytes

Remote site: /wrk/project_ogiir-csc

```
/
    homeappl
        home
            ekkylli
    wrk
        project_ogiir-csc
```

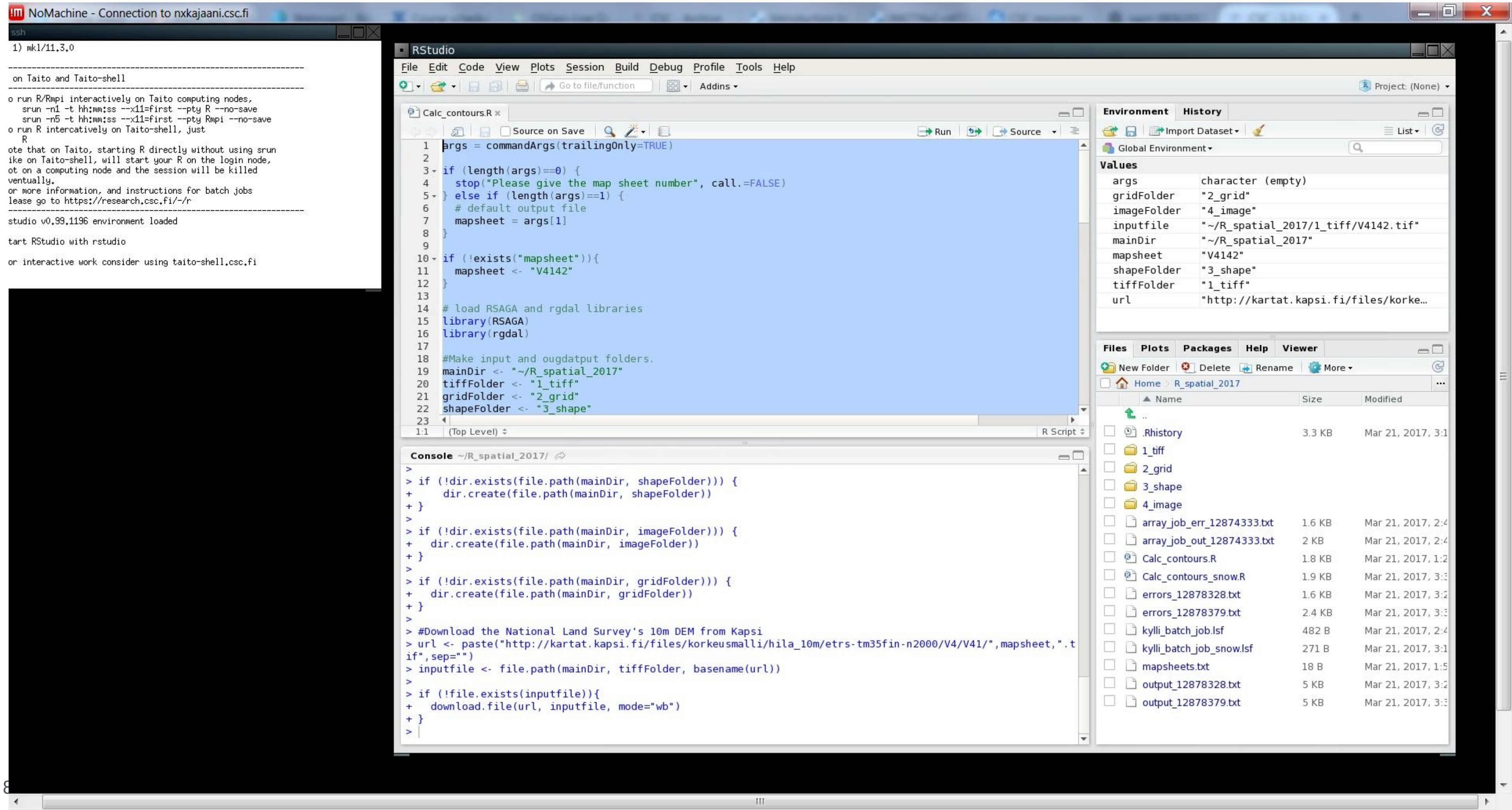
| Filename | Filesize | Filetype | Last mo |
|---|---|---|---|
| .. | | | |
| fmi | | File folder | 29.5.20 |
| luke | | File folder | 24.5.20 |
| mml | | File folder | 24.5.20 |

3 directories

| Server/Local file | Direction | Remote file | Size | Priority |
|---|---|---|---|---|

Queue: empty

# NoMachine



NoMachine - Connection to nxkajaani.csc.fi

ssh

```
 1) mkl/11.3.0

---------------------------------------------------------------------
 on Taito and Taito-shell
---------------------------------------------------------------------
o run R/Rmpi interactively on Taito computing nodes,
   srun -n1 -t hh:mm:ss --x11=first --pty R --no-save
   srun -n5 -t hh:mm:ss --x11=first --pty Rmpi --no-save
o run R intercatively on Taito-shell, just
   R
ote that on Taito, starting R directly without using srun
ike on Taito-shell, will start your R on the login node,
ot on a computing node and the session will be killed
ventually.
or more information, and instructions for batch jobs
lease go to https://research.csc.fi/-/r
---------------------------------------------------------------------
studio v0.99.1196 environment loaded

tart RStudio with rstudio

or interactive work consider using taito-shell.csc.fi
```

RStudio

```r
args = commandArgs(trailingOnly=TRUE)

if (length(args)==0) {
  stop("Please give the map sheet number", call.=FALSE)
} else if (length(args)==1) {
  # default output file
  mapsheet = args[1]
}

if (!exists("mapsheet")){
  mapsheet <- "V4142"
}

# load RSAGA and rgdal libraries
library(RSAGA)
library(rgdal)

#Make input and ougdatput folders.
mainDir <- "~/R_spatial_2017"
tiffFolder <- "1_tiff"
gridFolder <- "2_grid"
shapeFolder <- "3_shape"
```

Console ~/R_spatial_2017/

```
>
> if (!dir.exists(file.path(mainDir, shapeFolder))) {
+     dir.create(file.path(mainDir, shapeFolder))
+ }
>
> if (!dir.exists(file.path(mainDir, imageFolder))) {
+   dir.create(file.path(mainDir, imageFolder))
+ }
>
> if (!dir.exists(file.path(mainDir, gridFolder))) {
+   dir.create(file.path(mainDir, gridFolder))
+ }
>
> #Download the National Land Survey's 10m DEM from Kapsi
> url <- paste("http://kartat.kapsi.fi/files/korkeusmalli/hila_10m/etrs-tm35fin-n2000/V4/V41/",mapsheet,".t
if",sep="")
> inputfile <- file.path(mainDir, tiffFolder, basename(url))
>
> if (!file.exists(inputfile)){
+   download.file(url, inputfile, mode="wb")
+ }
>
```

```
Values
  args          character (empty)
  gridFolder    "2_grid"
  imageFolder   "4_image"
  inputfile     "~/R_spatial_2017/1_tiff/V4142.tif"
  mainDir       "~/R_spatial_2017"
  mapsheet      "V4142"
  shapeFolder   "3_shape"
  tiffFolder    "1_tiff"
  url           "http://kartat.kapsi.fi/files/korke…
```

| Name | Size | Modified |
|---|---|---|
| .. | | |
| .Rhistory | 3.3 KB | Mar 21, 2017, 3:1 |
| 1_tiff | | |
| 2_grid | | |
| 3_shape | | |
| 4_image | | |
| array_job_err_12874333.txt | 1.6 KB | Mar 21, 2017, 2:4 |
| array_job_out_12874333.txt | 2 KB | Mar 21, 2017, 2:4 |
| Calc_contours.R | 1.8 KB | Mar 21, 2017, 1:2 |
| Calc_contours_snow.R | 1.9 KB | Mar 21, 2017, 3:3 |
| errors_12878328.txt | 1.6 KB | Mar 21, 2017, 3:2 |
| errors_12878379.txt | 2.4 KB | Mar 21, 2017, 3:3 |
| kylli_batch_job.lsf | 482 B | Mar 21, 2017, 2:4 |
| kylli_batch_job_snow.lsf | 271 B | Mar 21, 2017, 3:1 |
| mapsheets.txt | 18 B | Mar 21, 2017, 1:5 |
| output_12878328.txt | 5 KB | Mar 21, 2017, 3:2 |
| output_12878379.txt | 5 KB | Mar 21, 2017, 3:3 |

# Directories at CSC Environment

https://research.csc.fi/data-environment

| Directory or storage area | Intended use | Default quota/user | Storage time | Backup |
|---|---|---|---|---|
| $HOME [1] | Initialization scripts, source codes, small data files.<br>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 | 90 days | No |
| $WRKDIR/DONOTREMOVE | Temporary data storage. | Incl. in above | Permanent | 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] | Storage and sharing of stable data. | On request | Permanent | No, multiple storage copies |

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

# Taito module system

- Tool to set up your environment
  - Load libraries, adjust path, set environment variables
  - Needed on a server with hundreds of applications and several compilers etc.
- Check the module names from [https://research.csc.fi/software](https://research.csc.fi/software)
- In NoMachine some tools with GUI are added to the context menu
- Example: initialize R and RStudio statistics packages

```
$ module load rspatial-env
$ module load rstudio
```

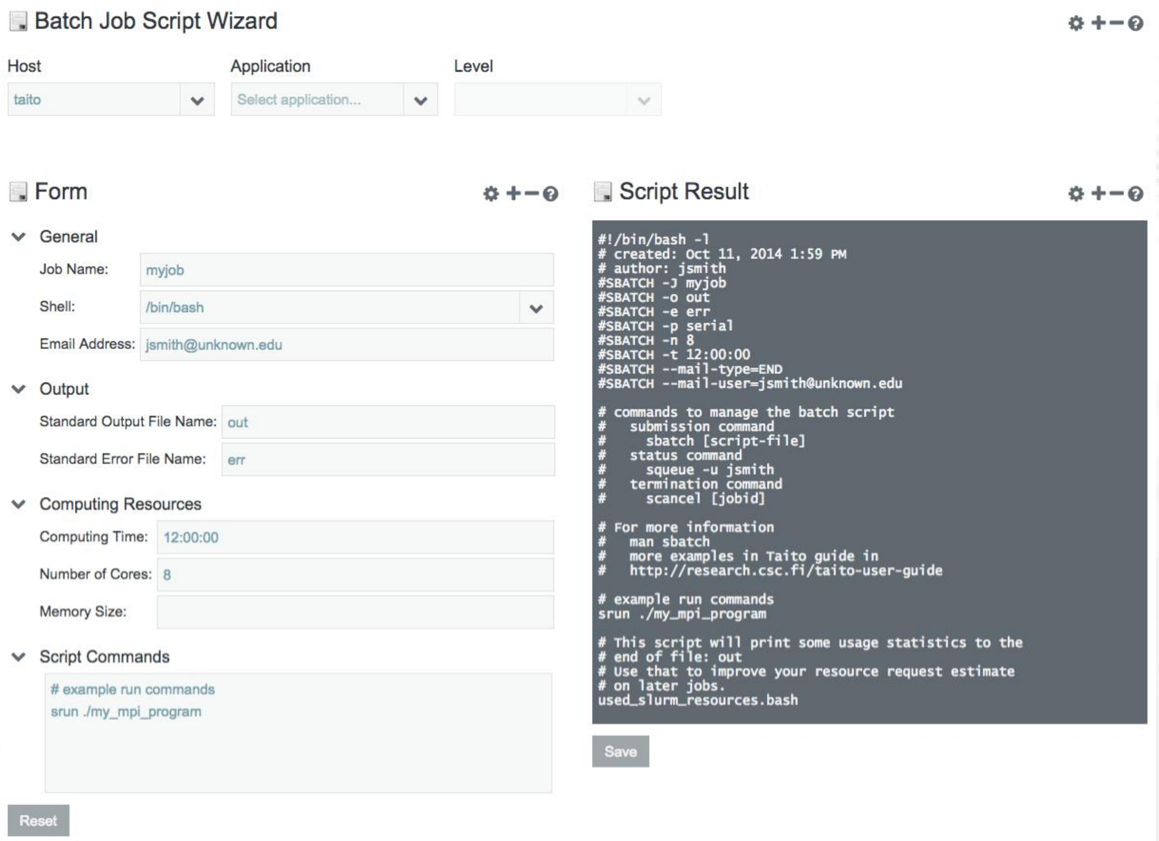
# Batch system

- Has to be used on Taito (not in Taito-shell)
- Optimizes resource usage by filling the server with jobs
- You have to reserve time, cores and memory for your job
- Several queues: parallel, serial, longrun, test and hugemem
- You have to write a batch job script
- https://research.csc.fi/taito-batch-jobs

# Scientist’s User Interface (SUI)

## Batch Job Script Wizard

- Create job scripts with easy to use forms
- Save scripts locally or in CSC $HOME
- Instructions of how to submit and monitor



### Batch Job Script Wizard

Host: taito

Application: Select application...

Level:

### Form

General

Job Name: myjob

Shell: /bin/bash

Email Address: jsmith@unknown.edu

Output

Standard Output File Name: out

Standard Error File Name: err

Computing Resources

Computing Time: 12:00:00

Number of Cores: 8

Memory Size:

Script Commands

```
# example run commands
srun ./my_mpi_program
```

### Script Result

```
#!/bin/bash -l
# created: Oct 11, 2014 1:59 PM
# author: jsmith
#SBATCH -J myjob
#SBATCH -o out
#SBATCH -e err
#SBATCH -p serial
#SBATCH -n 8
#SBATCH -t 12:00:00
#SBATCH --mail-type=END
#SBATCH --mail-user=jsmith@unknown.edu

# commands to manage the batch script
#   submission command
#     sbatch [script-file]
#   status command
#     squeue -u jsmith
#   termination command
#     scancel [jobid]

# For more information
#   man sbatch
#   more examples in Taito guide in
#   http://research.csc.fi/taito-user-guide

# example run commands
srun ./my_mpi_program

# This script will print some usage statistics to the
# end of file: out
# Use that to improve your resource request estimate
# on later jobs.
used_slurm_resources.bash
```

# Example: steps for running your R script in Taito

(0. Get yourself CSC user account)

1. Move your data and scripts to Taito (with FileZilla).
2. Log in to Taito (with Putty).
3.  Open RStudio in Taito-shell with NoMachine.
4. Check which R packages do you need and if they are available in Taito.

* If needed, install it yourself or ask CSC - servicedesk@csc.fi.

5. Fix the paths of your input/output files.
6. Test your script in Taito-shell with some test data.
7. Run your scripts with all data interactively on Taito-shell or in Taito as batch job.

(8. Make use of several cores using snow, foreach or rmpi packages in your R code.)

# Example code in CSC training Github

- Spatial analysis, with batch job scripts suitable for Taito. Examples for serial, array and parallel jobs
  - Python
  - R
- cPouta installation guidelines:
  - PostGIS
  - GeoServer
  - ArcGIS Server for ArcPy

https://github.com/csc-training/geocomputing

# GIS projects on Taito

- UH/CBIG: Global protected area expansion and conservation prioritization analysis (R, Zonation)
- UH: Weather modelling (R)
- UH: Travelling times (custom code)
- UH: Climate impact on bird populations (R)
- UTU: Forest mapping in the Amazon (R)
- FGI: Catchment area calculations for whole Finland (custom GPU code)
- SYKE: Species modelling in the sea (R)
- LUKE: Several forestry related (Matlab, custom code)

# cPouta cloud

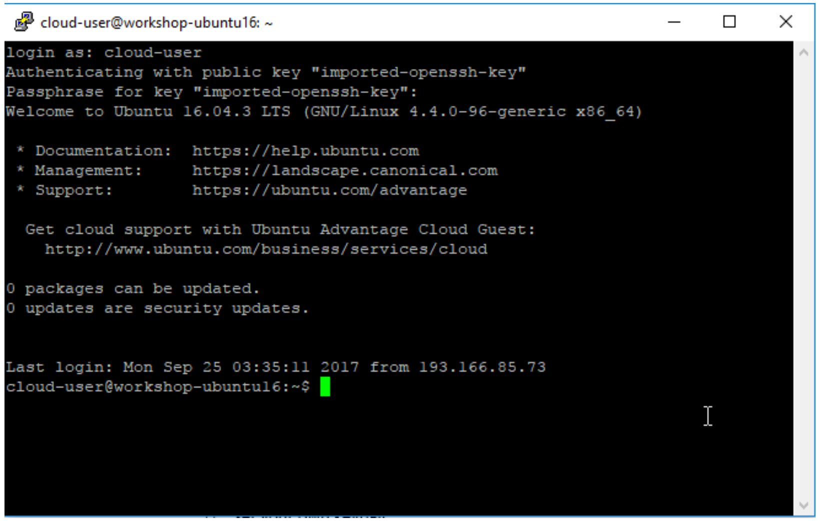
# Pouta Clouds in general

- Serviced offered by CSC (hardware in Finland)
- True self-service cloud IaaS powered by OpenStack
  - o Deploy your own virtual machines, storage and networks as your requirements evolve
  - o No proprietary software to limit scalability
- Simple to create and modify virtual resources
  - o Choose from Web UI, CLI or RESTful APIs
- Designed to serve scientific as well as other use cases
  - o General purpose
  - o High Performance Computing
  - o Data Intensive Computing
  - o Sensitive data

# cPouta

- The user is responsible of setting up the virtual machine and has to install everything
- Almost anything is possible
- A lot of freedom, but also more responsibility
- Linux is the default and easy way, but Windows is also available.

KE1



```
login as: cloud-user
Authenticating with public key "imported-openssh-key"
Passphrase for key "imported-openssh-key":
Welcome to Ubuntu 16.04.3 LTS (GNU/Linux 4.4.0-96-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

0 packages can be updated.
0 updates are security updates.


Last login: Mon Sep 25 03:35:11 2017 from 193.166.85.73
cloud-user@workshop-ubuntu16:~$
```

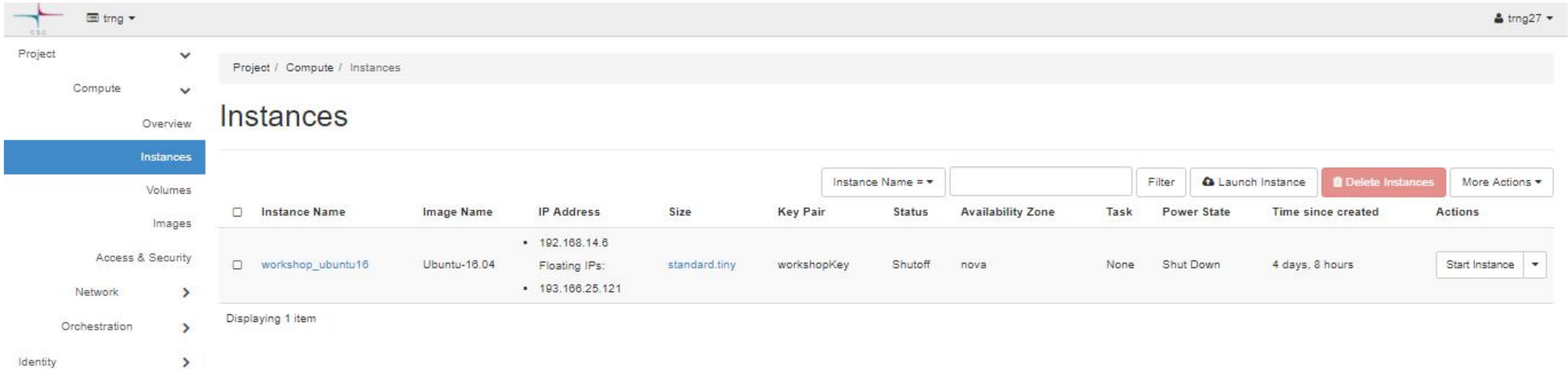
Slide 30

KE1 Some more intro here?
Kylli Ek; 11.12.2017

# Traditional HPC (Taito) vs. IaaS (cPouta)

| | Traditional HPC environment | Cloud environment<br>Virtual Machine |
|---|---|---|
| Operating system | Same for all: CSC's cluster OS | Chosen by the user |
| Software installation | Done by cluster administrators<br>Customers can only install software to their own directories, no administrative rights | Installed by the user<br>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 = SLURM in Taito) | 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 |

# cPouta web interface



Project / Compute / Instances

## Instances

| ☐ | Instance Name | Image Name | IP Address | Size | Key Pair | Status | Availability Zone | Task | Power State | Time since created | Actions |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ☐ | workshop_ubuntu16 | Ubuntu-16.04 | • 192.168.14.6<br>Floating IPs:<br>• 193.166.25.121 | standard.tiny | workshopKey | Shutoff | nova | None | Shut Down | 4 days, 8 hours | Start Instance |

Displaying 1 item

# Options for installing software

- From an virtual machine image or Docker container with ready installed software packages
  - o for ex. OSGeoLive, opendronemap

- Installing sofware manually
  - o for ex. using apt-get command line

- Scripting tools
  - o for ex. Ansible

# GIS projects on cPouta

- UH: Driving times (ArcPy -> PostGIS)
- UH: Automating GIS processes course (Pebbles + remote desktop -> Notebooks, JypiterLab)
- UH: GeoServer and PostGIS for course use
- Aalto: Water simulations (MIKE, Windows)
- FGI: Catchment area calculations on the fly (custom code, Leaflet)
- FGI: SNAP+Python for Sentinel image analysis
- FGI: GeoCubes
- Oulu, UEF: Sharing research results with GeoServer
- UTU: Data sharing GeoNode
- UEF: drone image analysis with opendronemap
- Aalto: Spatial analysis with Spark

# Object storage

- Used for storing and sharing data / files.
- Included in Pouta projects
- Ready to use (no need to set up a virtual machine)
- Manage via the Pouta Web interface or via API (s3, swift, python…)
- Data can be accessed from anywhere using URL or via API
- Data can be private, public and temporarily shared
- Limitations:
  - Object storage file can not be edited (you can delete and make a new copy)
  - Not suitable for databases
  - Can not (efficiently) be mounted as file system

# Rahti

- Platform-as-a-service (based on OpenShift, Red Hat's distribution of Kubernetes)
- Used for running and orchestrating containers that run applications
- Still you need to install your software and pack it as containers
- Same end goal as cPouta: enable end users to run their own software in the cloud
  - web applications
  - APIs/microservices for science
  - Apache Spark
  - Jupyter notebooks
- Compared to Pouta, you don’t need to manage virtual machines but you need to manage containers

# notebooks.csc.fi

- Easy-to-use environments for working with data and programming.
- Primarily for teaching and course use
- Jupyter notebooks: R, Python, Julia, Spark, Machine learning
- New or coming:
  - JupyterLab
  - Custom containers from Rahti
- Limitations:
  - Time limited (some hours)
  - Data can not be saved
- Login: HAKA + CSC usernames + e-mail invitations

# How to choose: Taito, cPouta or Rahti

- Taito:
  - Heavy computing with tools that can be installed to Taito
- cPouta
  - Server software: PostGIS, GeoServer, GeoNode etc
  - Heavy computing with tools that can not be installed to Taito
- Rahti:
  - Launch containeraized software as single or distributed applications
  - Can host web based applications
  - For server software and tools that can not be installed to Taito

# Accounts

- Using CSC resources is mostly free of charge for univeristy users and research institues for open research
- HAKA-users can create an account themselves in SUI: https://research.csc.fi/accounts-and-projects
- Research institutes have to ask for account from servicedesk.
- HAKA-users can start using Taito without project with the default quota.
- For cPouta you always need a project.

# Billing units

- Each project is given certain amount of so-called billing units (BU).
- On Taito, if you are using batch jobs, the billing is based on actual time used, but on the number of cores and memory reserved.
  - If you need help with estimating your job resource needs, see the seff command from the end of this page or see the webinar about estimating needed memory: https://www.youtube.com/watch?v=4ThGRZq1G8U
  - Changing billing project: https://research.csc.fi/billing-and-monitoring
  - Project saldo, to see how much BUs you have used: https://research.csc.fi/saldo
- In cPouta, billing is based on virtual machine size/type and its life time.
- You can ask for more quota, if you need.

# Support

CSC service desk:

servicedesk@csc.fi

Add giscoord@csc.fi as cc, for a little bit faster reply.

- Help
- Installation requests
- Code optimization

# Guidelines and news

- Guidelines

https://research.csc.fi/geocomputing

- GIS@CSC news
- GIS@CSC e-mail list: gis-hpc

http://research.csc.fi/gis-csc-news

# Contact

http://research.csc.fi/geosciences

Kylli Ek, +358 50 38 12 838

Eduardo Gonzalez, +358 40 848 8989

giscoord@csc.fi