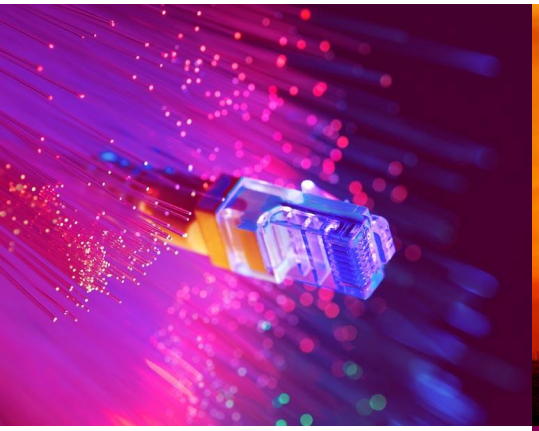


Running Python code in CSC Taito supercluster

Kylli Ek, CSC

Espoo, 13.11.2018

Non-profit state organization with special tasks

Turnover in 2017

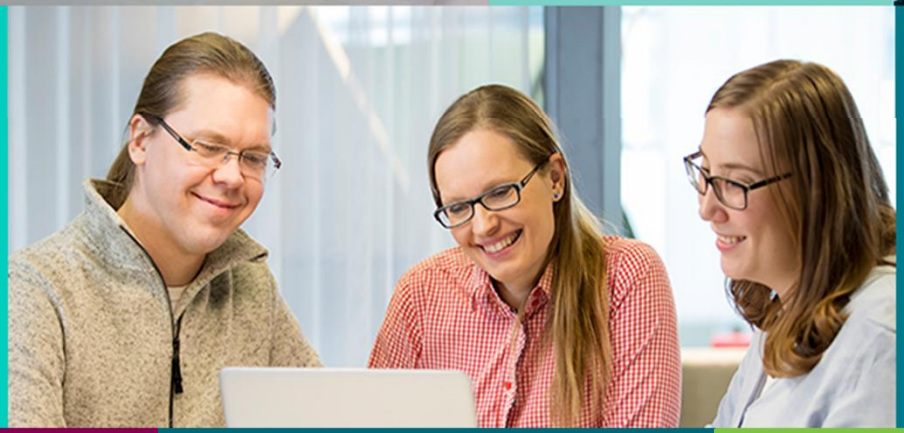
40,5 M€

CSC

Headquarters in Espoo, datacenter in Kajaani

Owned by state (70%)

and all Finnish higher education institutions (30%)

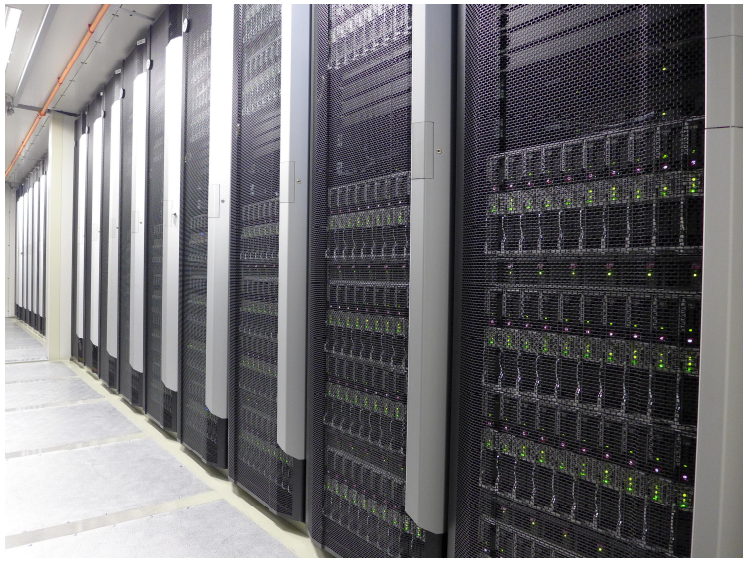


Circa

320

employees in 2017

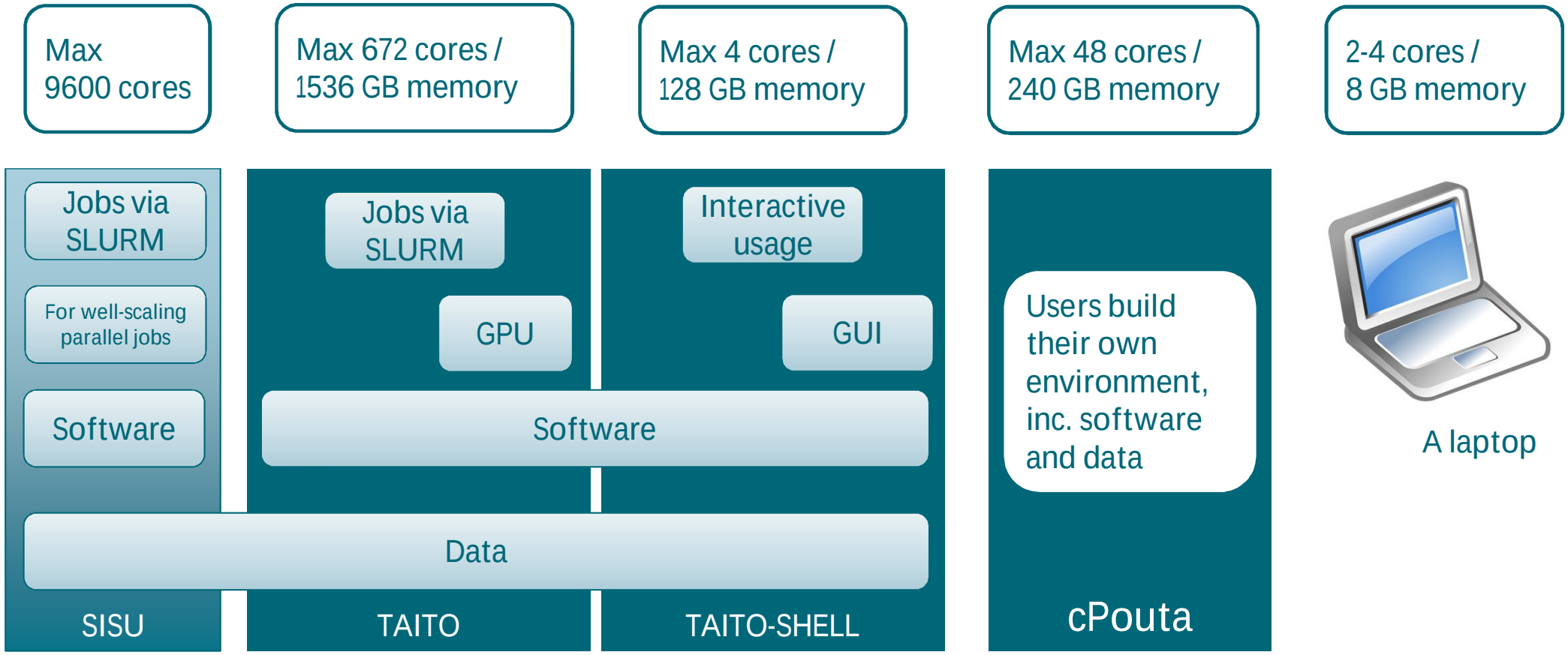
Reasons for using CSC computing resources



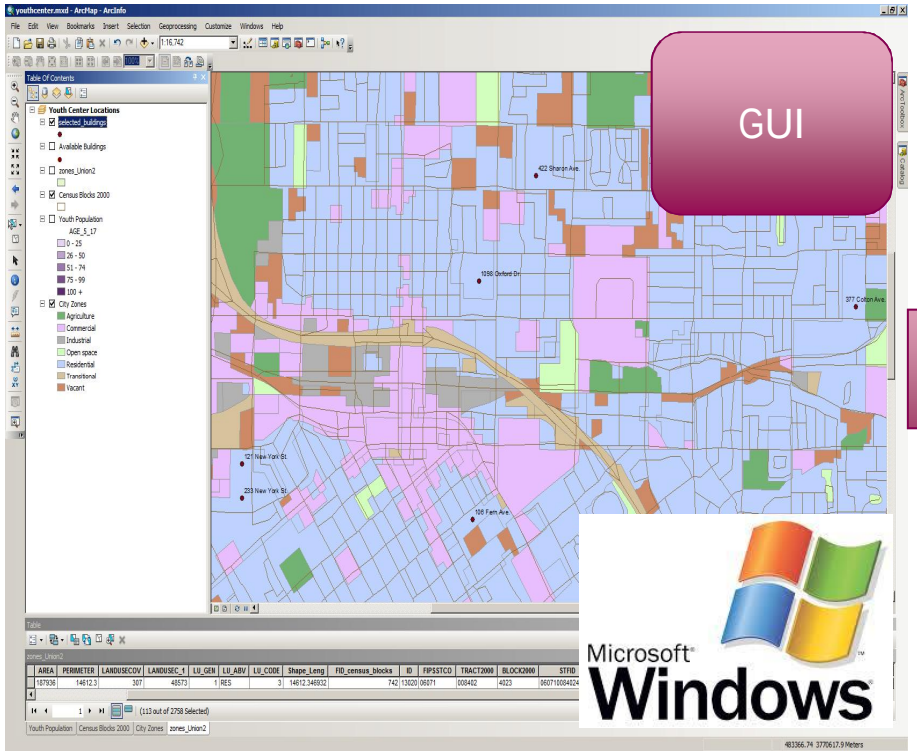
- Computing something takes more than 2-4 hours
- Need for more memory
- Very big datasets
- Keep your desktop computer for normal usage, do computation elsewhere
- Need for a server computer
- Need for a lot of computers with the same set-up (courses)

- Free for Finnish university users and for state research insitutes

CSC HPC resources



The keys to geocomputing: Change in working style & Linux




```python
from rasterstats import zonal_stats
import time
import fiona
import multiprocessing as mp
from shapely.geometry import mapping, shape

#input zones
zone_f = "zones.shp"
#output zonal stats
zonal_f = "zonal_stats.shp"
#Raster you want to use to compute zonal stastics from.
vrt = "dem2m.vrt"

statistics = ['count', 'min' ,'mean', 'max','median']

#yields n sized chunks from list l (used for splitting
task to multiple processes)
def chunks(l, n):
    for i in range(0, len(l), n):
        yield l[i:i + n]


#calculates zonal stats and adds res
def worker(z,vrt,d):
```

Ln : 1 Col : 1 Sel : 0 | 0 Unix (LF)

Scripts

Linux

ArcGIS, QGIS, …

R, Python, shell scripts, Matlab, …

Taito / Taito-shell pre-installed software for GIS

- R
- Python
- MatLab / Octave
- GDAL/OGR
- GRASS GIS
- LasTools (some)
- PDAL
- Proj4
- QGIS
- SagaGIS
- SNAP, sen2cor
- Taudem
- Zonation

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

Geoconda

- NumPy, Scipy, Pandas etc.
- GIS Specific packages
  - Geopandas
  - Fiona
  - Shapely
  - Rasterio
  - Rasterstats
  - cartopy
  - GDAL/OGR
  - Networkx
  - Skimage
  - Pyproj
  - Pysal
  - Rtree
  - Descartes

```
module load geoconda
```

Installing software

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

- You can add also packages to Python with pip
- You can also make your own conda environments

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.

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
SYKE
    All open GIS data

https://research.csc.fi/gis_data_in_taito

Access to Taito from Windows

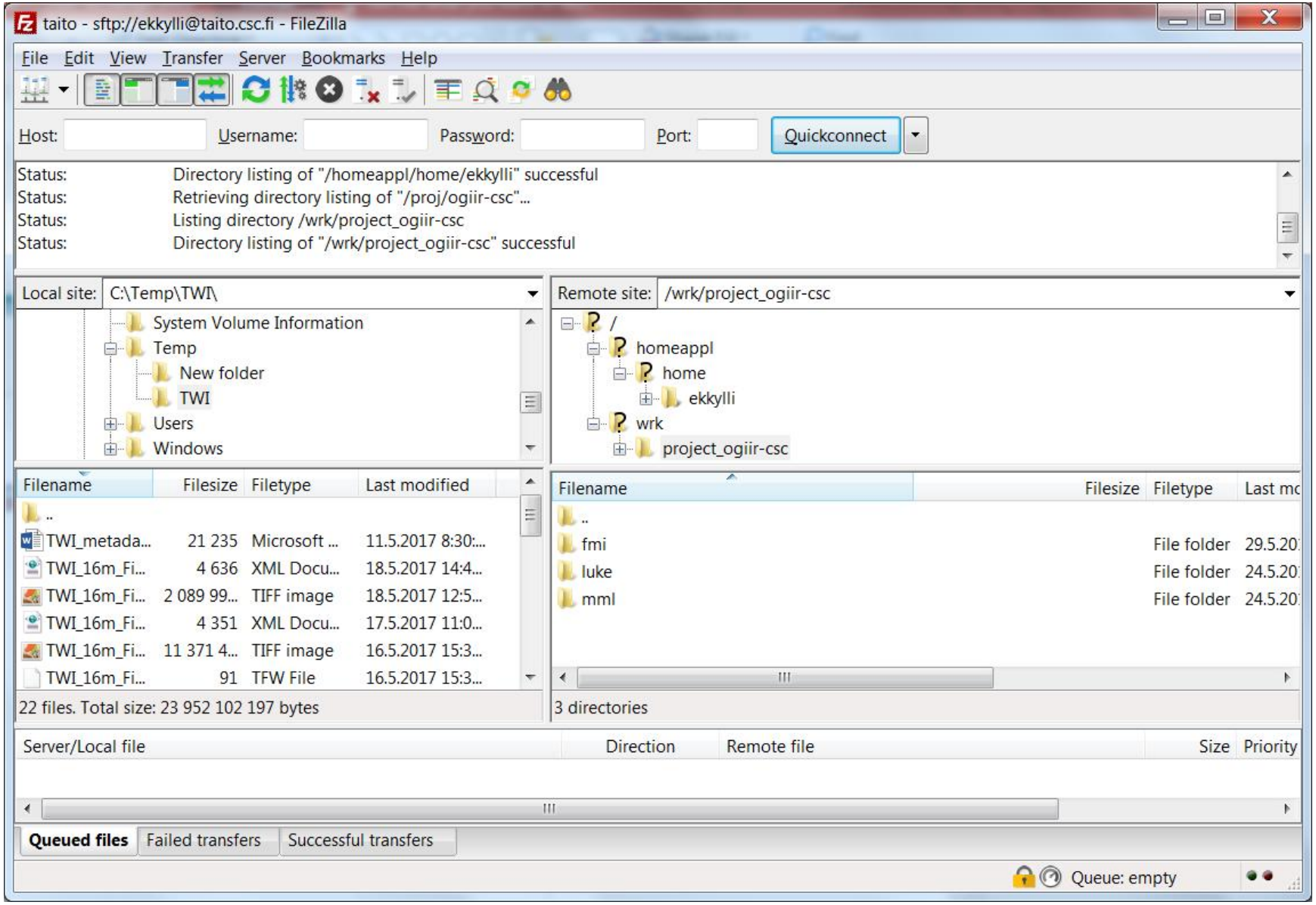
- Putty for ssh connection
- FileZilla/WinSCP for moving data
- NoMachine for GUI
- Find about other access options and more information at: https://research.csc.fi/taito-connecting

Putty

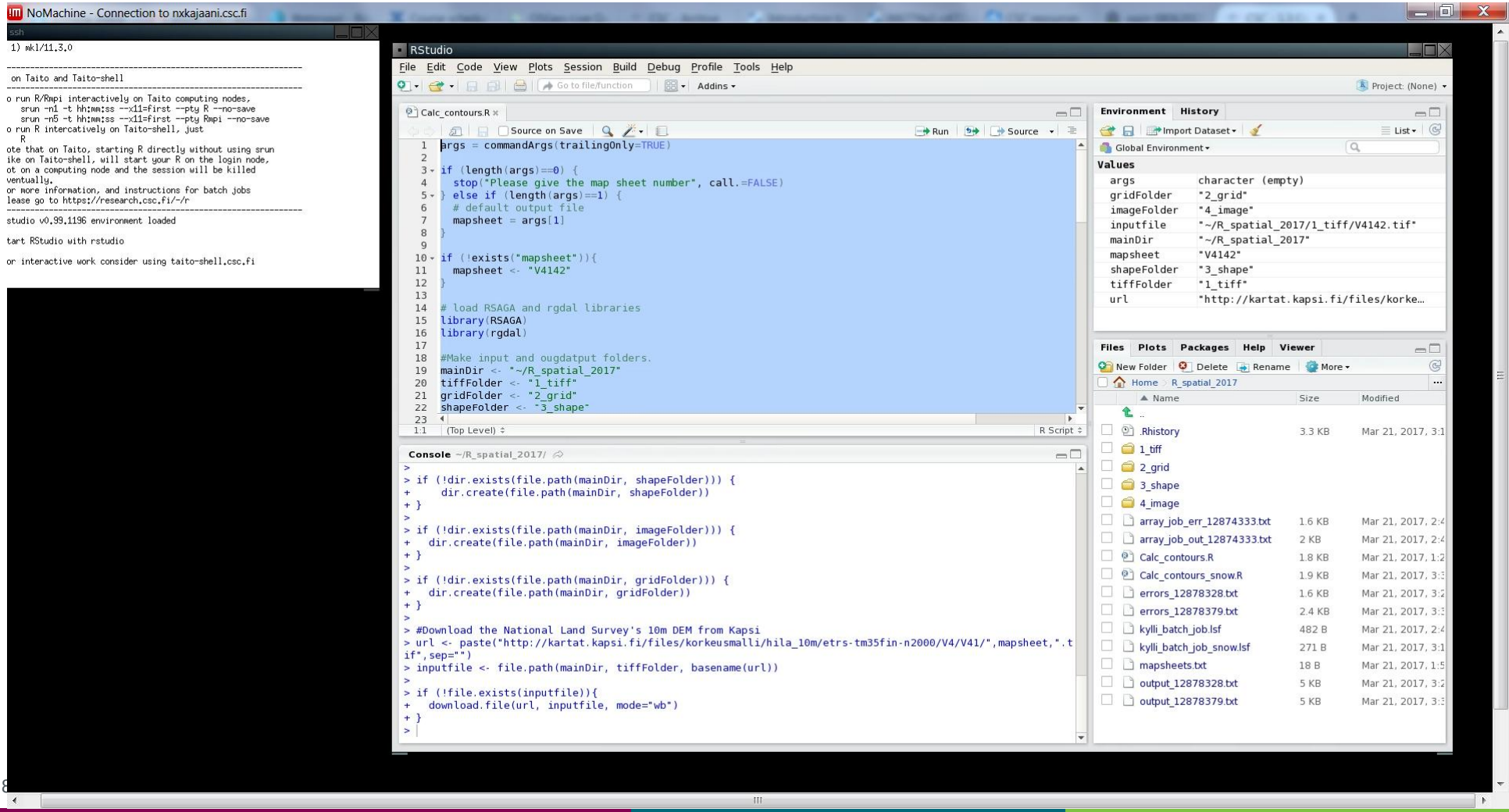
ekkylli@taito-login4:~

```
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 ~]$
```

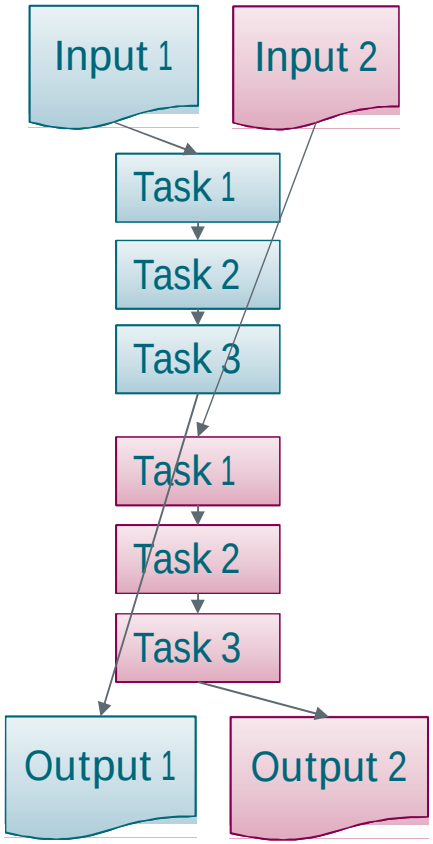
FileZilla


taito - sftp://ekkylli@taito.csc.fi - FileZilla
File Edit View Transfer Server Bookmarks Help
Host: Username: Password: Port: Quickconnect
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
Remote site: /wrk/project_ogiir-csc
/
homeappl
home
ekkylli
wrk
project_ogiir-csc
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
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
Queued files Failed transfers Successful transfers
Queue: empty

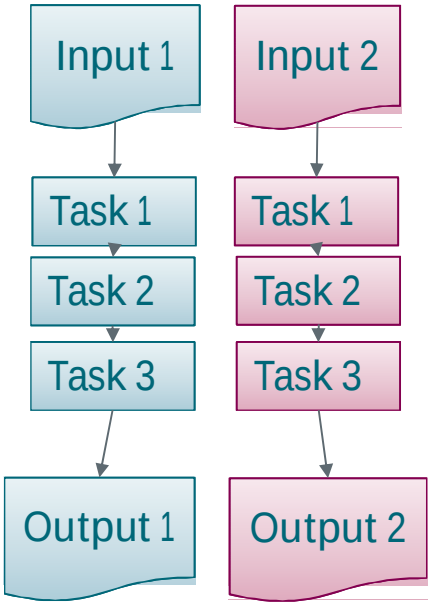
NoMachine


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
File Edit Code View Plots Session Build Debug Profile Tools Help
Calc_contours.R
Source on Save
Run
Source
1 args = commandArgs(trailingOnly=TRUE)
3 if (length(args)==0) {
4 stop("Please give the map sheet number", call.=FALSE)
5 } else if (length(args)==1) {
6 # default output file
7 mapsheet = args[1]
8 }
10 if (!exists("mapsheet")){
11 mapsheet <- "V4142"
12 }
14 # load RSAGA and rgdal libraries
15 library(RSAGA)
16 library(rgdal)
18 #Make input and ougdatput folders.
19 mainDir <- "~/R_spatial_2017"
20 tiffFolder <- "1_tiff"
21 gridFolder <- "2_grid"
22 shapeFolder <- "3_shape"
1:1 (Top Level) R Script
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,".tif",sep="")
> inputfile <- file.path(mainDir, tiffFolder, basename(url))
> if (!file.exists(inputfile)){
+ download.file(url, inputfile, mode="wb")
+ }
Environment History
Import Dataset
Global Environment
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...
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home R_spatial_2017
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

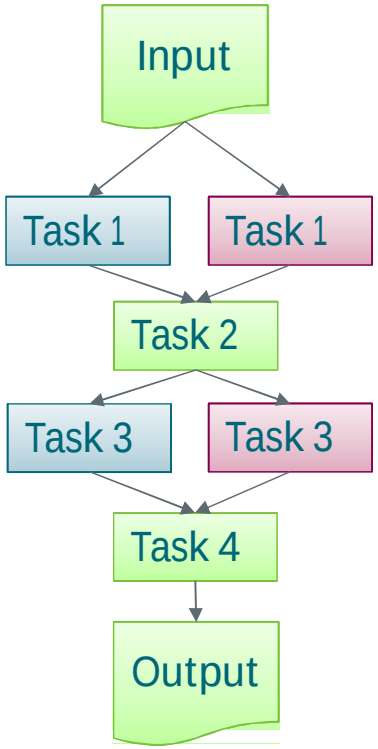
Simple job



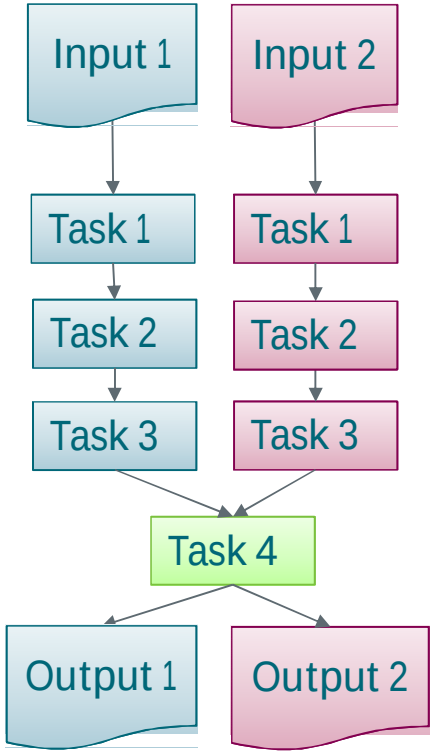
Array job



Parallel job 1



Parallel job 2



Example: steps for running your Python 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 Spyder in Taito-shell with NoMachine.
4. Check which Python 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 multiprocessing package in your Python code or with array jobs.)

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. 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

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

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.
- Example: initialize Python with GIS packages

```
$ module load geoconda
```

Accounts

- University users can create an account themselves in SUI: https://research.csc.fi/accounts-and-projects
- University users can start using Taito without project with the default quota.
- Research institute users have to ask for account from supportdesk.
- For serious work create a project and apply for resources.
- 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

Example code in CSC training Github

- Examples for doing spatial analysis in CSC computing environment with:
  - Python
  - R
- Examples include also batch job scripts suitable for Taito.
- Some of the examples include samples for serial, array and parallel jobs.

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

GIS training 2018/19

https://www.csc.fi/training

- Geospatial data analysis with R
- Google Earth Engine
- Lidar data analysis
- Webinar: Paituli, cPouta for GIS

Further information

- https://research.csc.fi/taito-user-guide
- https://research.csc.fi/geocomputing
- CSC Python documentation: research.csc.fi/-/python
- Geoconda module: https://research.csc.fi/-/geoconda
- Geo-env module: https://research.csc.fi/-/geo-env
- Multiprocessing: https://docs.python.org/3/library/multiprocessing.html

Training archive:
https://www.csc.fi/web/training/-/geocomputing-in-taito

Support: servicedesk@csc.fi

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

giscoord@csc.fi