rijksuniversiteit
groningen

1 December 2015

Using the Peregrine cluster

Fokke Dijkstra, Bob Dröge
Research and Innovation Support
Center for Information Technology

General introduction

- Course aimed at beginners
  - This part assumes knowledge about the Linux command line, file transfers and editing files
- Topics
  - What is a cluster
  - Cluster storage
  - Module environment
  - Submitting jobs

rijksuniversiteit groningen

Research and Innovation Support (RIS)

- HPC Facilities
  - Peregrine cluster
  - Grid cluster
- Visualisation
  - Cave
  - Theatre
  - Scientific data
  - Virtual reality
- Geo Services
- Research Support in IT

rijksuniversiteit groningen

Computer cluster

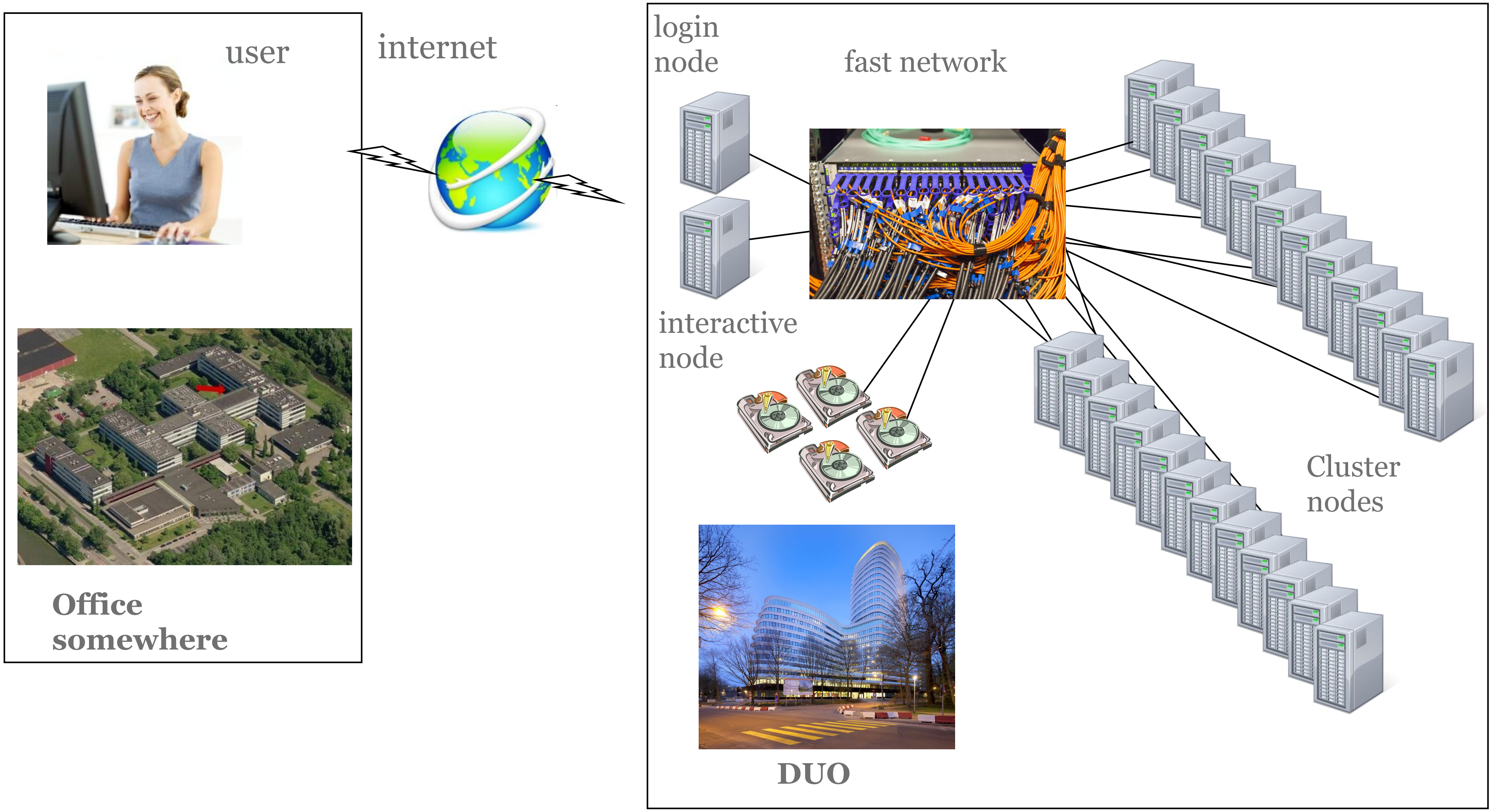
- A cluster is a collection of computers connected by a network
- A single front-end
- Lots of computers in the background for running tasks
- 1994 first cluster of commodity PCs at NASA

- Peregrine cluster looks quite different
- Most clusters run the Linux operating system

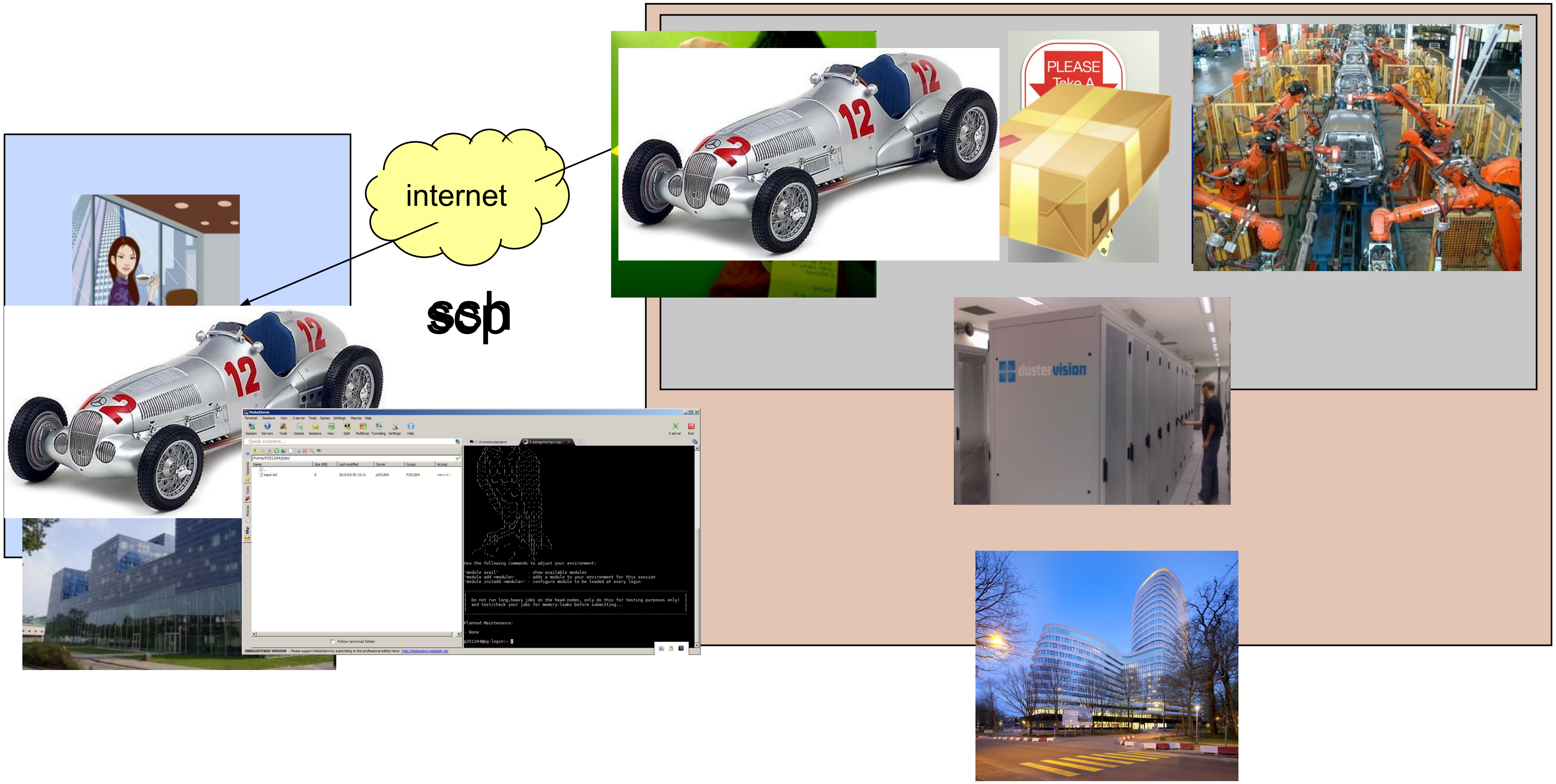
Peregrine falcon

- Fastest animal on earth
- Stoops down on prey from the air

Peregrine cluster



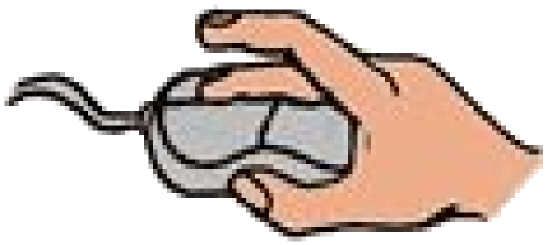
How does it work?



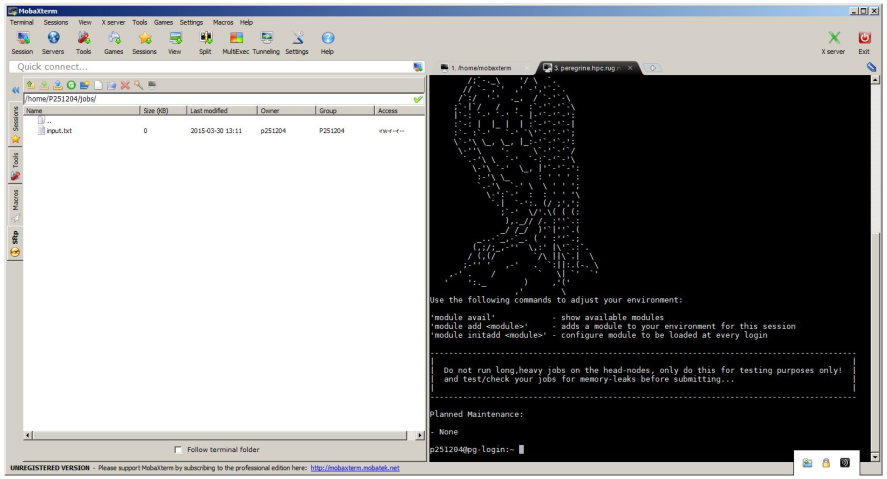
```
Use the following commands to adjust your environment:

'module avail'              - show available modules
'module add <module>'       - adds a module to your environment for this session
'module initadd <module>'   - configure module to be loaded at every login

-----------------------------------------------------------------------------
|                                                                           |
|  Do not run long,heavy jobs on the head-nodes, only do this for testing purposes only! |
|  and test/check your jobs for memory-leaks before submitting...           |
|                                                                           |
-----------------------------------------------------------------------------

Planned Maintenance:

- None

p251204@pg-login:~
```

Why a command line?

- Easier over the network
- Scripts are easier to run many times
- Instead of:



- One can use a small program



MobaXterm
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Settings Help
X server Exit
Quick connect...
1. /home/mobaxterm
3. peregrine.hpc.rug.n
Sessions Tools Macros Sftp
Follow terminal folder
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net

/home/P251204/jobs/

Name	Size (KB)	Last modified	Owner	Group	Access
..					
input.txt	0	2015-03-30 13:11	p251204	P251204	-rw-r--r--

```
Use the following commands to adjust your environment:

'module avail'              - show available modules
'module add <module>'       - adds a module to your environment for this session
'module initadd <module>'   - configure module to be loaded at every login

---------------------------------------------------------------------------
|                                                                         |
|  Do not run long,heavy jobs on the head-nodes, only do this for testing purposes only! |
|  and test/check your jobs for memory-leaks before submitting...         |
|                                                                         |
---------------------------------------------------------------------------

Planned Maintenance:

- None

p251204@pg-login:~
```

SSH (Secure SHell)

- SSH protocol used to connect to the cluster
  - Standard interface for Unix systems
  - Encrypted network traffic
- Software for Windows called MobaXterm:
  - http://mobaxterm.mobatek.net/
  - Freely available for personal use, already installed on UWP 2
- Use terminal window under Linux or Mac OS X

- University P/S/F/G account and password
- Hostname login node: **peregrine.hpc.rug.nl**
- Interactive node: **pg-interactive.hpc.rug.nl**

File transfer clients

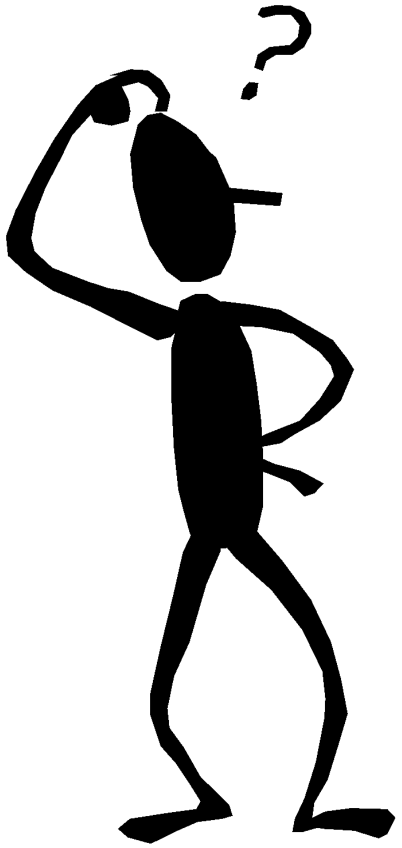
- **Windows**
  - MobaXterm has a file transfer interface
    - Included in UWP 2
- **Linux**
  - Command line scp or sftp
  - File manager, FileZilla
- **Mac OS X** (not tested)
  - Command line scp or sftp
  - FileZilla (Open source)

rijksuniversiteit groningen

Storage

File system	Space (TB)	Quota (GB)	Backup	Shared	Cleanup	Use case
/home	26	20	yes	yes	No	Programs Code Small data sets
/data	206	250	no	yes	No	Large reused data sets
/scratch	231	-	no	yes	Yes, 30 days retention	Temporary data shared between nodes manual clean up
/local	1	-	no	per node	Yes, automatically after job	Temporary data for single node automatic clean up

- What can it do for me?



rijksuniversiteit groningen

For whom?

- Account available for University staff & students
  - P, S, or F account is required
  - Undergraduate students through supervisor/teacher
  - Provide contact details and short description of planned use
- Access through SSH protocol
  - peregrine.hpc.rug.nl (login node)
  - pg-interactive.hpc.rug.nl (interactive node)

What can it be used for?

- Applications must be able to run under Linux
  - Compile the application yourself
  - Preinstalled applications
    - matlab, R, gromacs, …
  - Run anywhere languages
    - Java, Python, Perl, ….
- No user interaction
  - Input/output through files
- No graphical interface

General scenarios

- Long-running calculations
- Parallel calculations
- Many calculations

rijksuniversiteit groningen

Licenses

- If you want to use commercial software a license may be necessary
- Provided by the university
  - May be installed system wide
  - Problem with Linux versions or number of licenses
- Provided by yourself
  - Install software in home directory
- Other cases may be more difficult

- Where is <insert my favourite software package>?
- Many applications already available
- Organized through a “module” system

rijksuniversiteit groningen

Module environment

› Environment set for software when module is loaded
› Useful commands:
  - module avail [name]
  - module list
  - module add / load <module>
  - module del / remove / rm / unload <module>
  - module save/restore [name]
  - module purge

rijksuniversiteit groningen

Module environment (2)

- Software built using toolchains:
  - goolfc:
    - GNU compilers, OpenMPI, OpenBLAS, Lapack, FFTW, CUDA
  - ictce:
    - Intel compilers, MKL, Intel MPI
  - Module name contains name of toolchain used to build the software
- Dependencies automatically loaded

Module examples (1)

```
$ module avail
...
--------------------------------- /software/modules/bio ------------------------
   ABySS/1.5.2-goolfc-2.7.11-Python-2.7.9
   BCFtools/1.2-goolfc-2.7.11
   BEDTools/2.22.1-goolfc-2.7.11
   BEDTools/2.23.0-goolfc-2.7.11                          (D)

...

--------------------------------- /software/modules/math -----------------------
   CPLEX/12.6.2
   Eigen/3.2.4-goolfc-2.7.11

...

$ bedtools
-bash: bedtools: command not found
$ module add BEDTools/2.22.1-goolfc-2.7.11
$ bedtools --version
bedtools v2.22.1
```

rijksuniversiteit groningen

Module examples (2)

```
$ module list
Currently Loaded Modules:
  1) GCC/4.8.4
  2) CUDA/6.5.14-GCC-4.8.4
  3) gcccuda/2.7.11
  4) OpenMPI/1.8.4-gcccuda-2.7.11
  5) gompic/2.7.11
  6) OpenBLAS/0.2.13-gompic-2.7.11-LAPACK-3.5.0
  7) FFTW/3.3.4-gompic-2.7.11
  8) ScaLAPACK/2.0.2-gompic-2.7.11-OpenBLAS-0.2.13-LAPACK-3.5.0
  9) goolfc/2.7.11
 10) BEDTools/2.22.1-goolfc-2.7.11
$ module del BEDTools
$ module list
Currently Loaded Modules:
  1) GCC/4.8.4
  2) CUDA/6.5.14-GCC-4.8.4
  3) gcccuda/2.7.11
  4) OpenMPI/1.8.4-gcccuda-2.7.11
  5) gompic/2.7.11
  6) OpenBLAS/0.2.13-gompic-2.7.11-LAPACK-3.5.0
  7) FFTW/3.3.4-gompic-2.7.11
  8) ScaLAPACK/2.0.2-gompic-2.7.11-OpenBLAS-0.2.13-LAPACK-3.5.0
  9) goolfc/2.7.11
```

Module examples (3)

```
$ module purge
$ module list
No modules loaded
```

rijksuniversiteit groningen

Installation of new software

- Into your own home directory:
  - Keep control over the software yourself
  - No special privileges required
  - Cannot be used by other users (unless you grant permission)

- Into a new module:
  - Can be used by other users
  - Installation requires special privileges
  - Contact us, see “Support” slide

- But I only see a single computer!

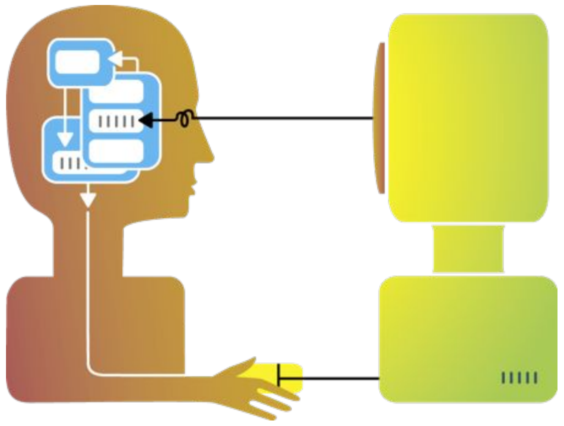
rijksuniversiteit groningen

Login node

- Front-end node *peregrine.hpc.rug.nl*
- Used for access to the cluster
  - Login
  - Data transfers
  - Job submission
  - Editing & Compiling programs
  - (Very) small tests

Interactive node

- Interactive node: *pg-interactive.hpc.rug.nl*
- Used for access to the cluster
  - Testing and porting software
  - Data transfers
  - Job submission
  - Editing & Compiling programs
- Shared machine, be careful about what you do!



rijksuniversiteit groningen

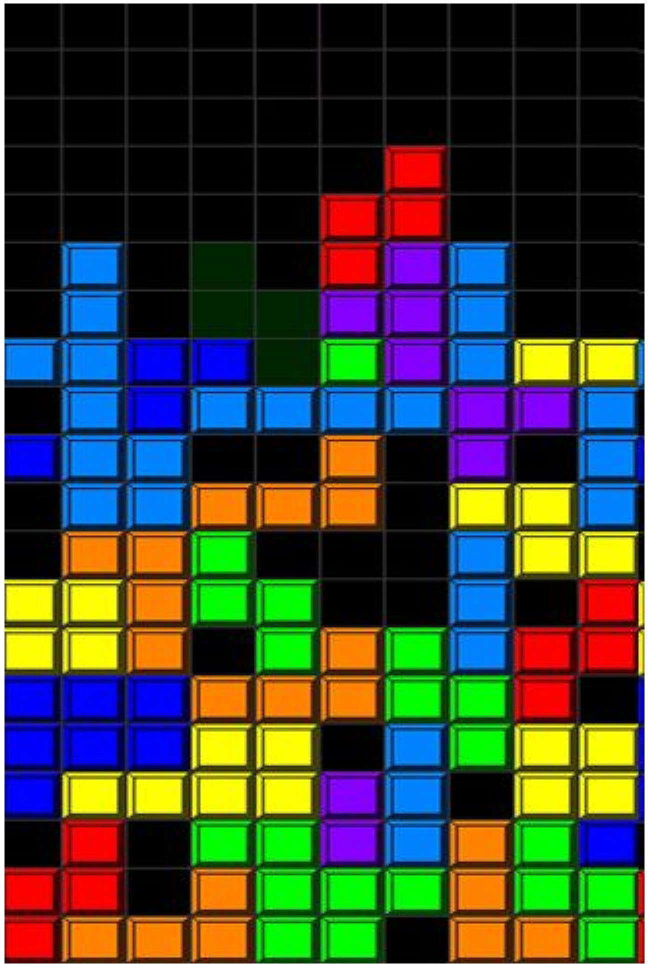
Compute nodes

	CPU	Memory	Internal disk	Network	Accelerator
160 Standard nodes	2x Intel Xeon E5 2680v3: 24 cores @ 2.5 GHz	128 GB	1 TB	56 Gbps Infiniband + 10 Gbps ethernet	-
4 GPU nodes	2x Intel Xeon E5 2680v3: 24 cores @ 2.5 GHz	128 GB			
2 Xeon Phi nodes	2x Intel Xeon E5 2680v3: 24 cores @ 2.5 GHz	128 GB			
7 Big memory nodes	4x Intel Xeon E7 4860v2: 48 cores @ 2.6 GHz	1024, 1536 or 2048 GB	1 TB	56 Gbps Infiniband, 10 Gbps ethernet	-
Standard desktop PC	~4 cores	~4-8GB	~1 TB	1 Gbps ethernet	Desktop GPU

4320 CPU cores,
23040 CUDA cores,
244 Xeon Phi cores

Scheduling system

- Users write job descriptions
- Scheduler finds matching resource
- Scheduler tries to make optimal use of the resources
- No resources: wait in a queue
- Priority determined by usage of system in the recent past
- SLURM: http://slurm.schedmd.com
  - Scheduler
  - Resource manager



rijksuniversiteit groningen

Scheduler: partitions

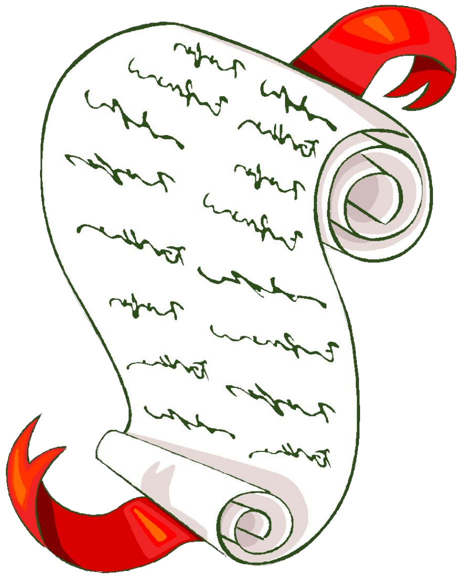
	Name	Max walltime
Standard nodes	nodes (default)	10 days
Big memory	himem	10 days
GPU	gpu	3 days
Xeon Phi	phi	3 days
Short	short	30 minutes

- Only half of the cores can be used for running long jobs (> 3 days)

Multiple jobs per node

- Scheduling full nodes to jobs a waste of resources
- Users can still request a complete node if necessary
- Jobs run in an isolated and limited set of resources
  - Not possible to use more resources than requested
- Jobs running on the node can still influence each other:
  - Disk space (space and bandwidth)
  - Memory (bandwidth)
  - Can sometimes be problematic

› What is a job script?



Job script

- Job script tells the system what you want to do
  - List of steps / commands to run
- First line should always point to the right interpreter that will run your script
  - Typically:
    `#!/bin/bash`
- Includes requirements needed to be able to run it
  - Amount of memory
  - Number of cores and/or nodes
  - Amount of time needed to complete the job

Job requirements/options

- Can be put in job script using lines that start with #SBATCH
- These lines should be at the top of the script, right after the #!/bin/bash line!

- Can also be passed as command-line arguments to sbatch command
  - Not explained in detail here
  - One simple example:

```
sbatch --time=10:00 --ntasks=4 jobscript.sh
```

rijksuniversiteit groningen

Job requirements

› Requirements specified using various options:
  - Wall clock time
    - #SBATCH --time=*days-hh:mm:ss*
  - Number of tasks / cores:
    - #SBATCH --ntasks=*n*
  - Specific number of nodes and tasks per node:
    - #SBATCH --nodes=*m*
    - #SBATCH --ntasks-per-node=*n*

› Specific node types using --partition option:
    - #SBATCH --partition=himem

rijksuniversiteit groningen

Memory requirements

- Memory requirements can be specified using
  - #SBATCH --mem=*n*
    n is the total amount of memory per node (!) in megabytes
  - #SBATCH --mem-per-cpu=*n*
    n is the amount of memory per CPU core in megabytes
  - Suffix K or KB, M or MB, G or GB, T or TB for other units
- Average memory limits:
  - 21.3, 32 or 42.7 GB per core on big memory nodes
  - 5.3 GB per core on all other nodes
- Default 2000MB memory limit!
- **Exceeding the limit will kill your application/job**

rijksuniversiteit groningen

Job properties

- Also using #SBATCH lines or on the command line
- Name of the job
  - #SBATCH --job-name=*name*
- Name of the output file of the job
  - #SBATCH --output=*filename*
  - Default is: slurm-<*jobid*>.out
- Email notifications
  - #SBATCH --mail-user=*user@mail.com*
  - #SBATCH --mail-type=*event1,event2,…,eventN*
    - *event* can be any of: ALL, BEGIN, END, FAIL, REQUEUE, TIME_LIMIT, TIME_LIMIT_50, TIME_LIMIT_80, TIME_LIMIT_90
    - Mail can be problematic in case of cluster problems

rijksuniversiteit groningen

Job script itself

› Contains Linux commands
  • cd, mkdir, etc.

› Run some application

› Sample script:

```
#!/bin/bash
#SBATCH --ntasks=2
#SBATCH --time=00:30:00
#SBATCH --job-name=testjob

cd $HOME/bin
./myprog
```

Useful environment variables

$HOME: your home directory

$USER: your username

$SCRATCHDIR: Temporary directory created for your job on /scratch. Removed after your job has finished!

$TMPDIR: Temporary directory created for your job on /local. Removed after your job has finished!

$SLURM_JOB_ID: Id of job, useful for creating unique files or directories for a job

$SLURM_JOB_NAME: Name of the job, as specified in the jobscript

…

Submitting jobs

› sbatch *jobscript*

```
$ sbatch testjob.sh
Submitted batch job 2865
```

Job id

› Job will start in the directory from which it was submitted
› Your complete environment will be transferred to the job; this includes all loaded modules.
  › Can be adjusted using:
    #SBATCH --export=VAR1,VAR2, … (or: NONE)

Checking job status (1)

› squeue

```
$ squeue
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
              4983     nodes  testjob  p456789 PD       0:00     20 (Resources)
              4984     nodes  testjob  p456789 PD       0:00     20 (Priority)
              4985     nodes  testjob  p456789 PD       0:00     20 (Priority)
              4986     nodes  testjob  p456789 PD       0:00     20 (Priority)
              4987     nodes  testjob  p456789 PD       0:00     20 (Priority)
              4978     nodes  testjob  p456789  R       0:01     20 pg-node[041-060]
              4979     nodes  testjob  p456789  R       0:01     20 pg-node[061-080]
              4980     nodes  testjob  p456789  R       0:01     20 pg-node[081-100]
              4981     nodes  testjob  p456789  R       0:01     20 pg-node[101-120]
              4982     nodes  testjob  p456789  R       0:01     20 pg-node[121-140]
              4976     nodes  testjob  p456789  R       0:04     20 pg-node[001-020]
              4977     nodes  testjob  p456789  R       0:04     20 pg-node[021-040]
```

Checking job status (2)

```
$ squeue -u p456789

     JOBID PARTITION     NAME       USER ST       TIME  NODES NODELIST(REASON)
      3018     nodes hpl.128.    p456789  R       3:26    128 pg-node[001-120,122-129]
```

Status:

PD: pending

R: running

CA: cancelled

CG: completing

CD: completed

F: failed

NF: node failure

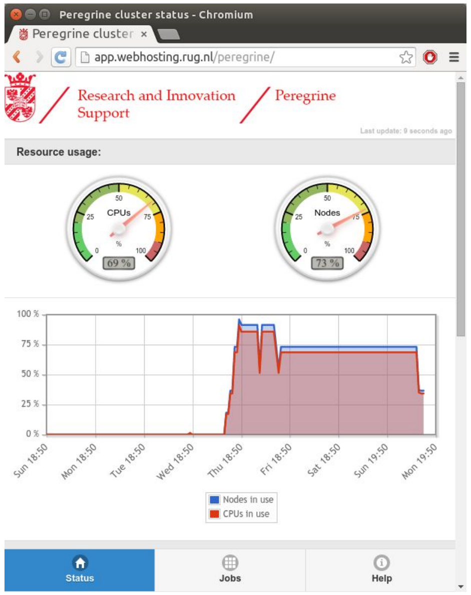
TO: timeout (reached time limit)

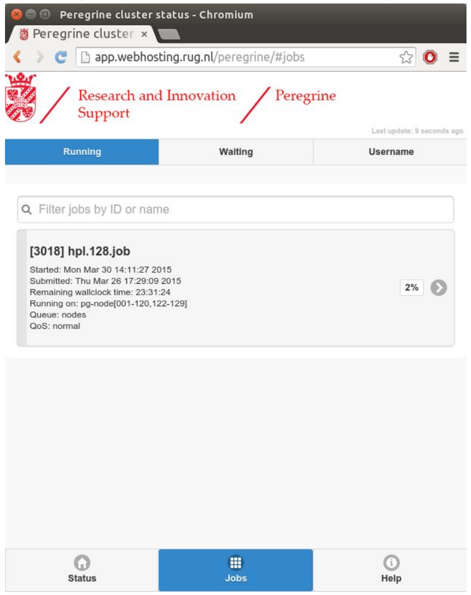
Checking job status (3)

- More information about a particular job:
  - Still running or just completed:
    - `scontrol show job <job id>`
  - Accounting information for completed jobs:
    - `sacct --jobs=<one or more job ids>`
    - Add `--long` option for even more information
    - Use `--format` option if you want specific information for your job, e.g. for memory usage: `sacct --jobs=2781 --format=JobID,jobname,MaxRSS`
  - Same information for running jobs:
    - `sstat --jobs=<job id(s)>`

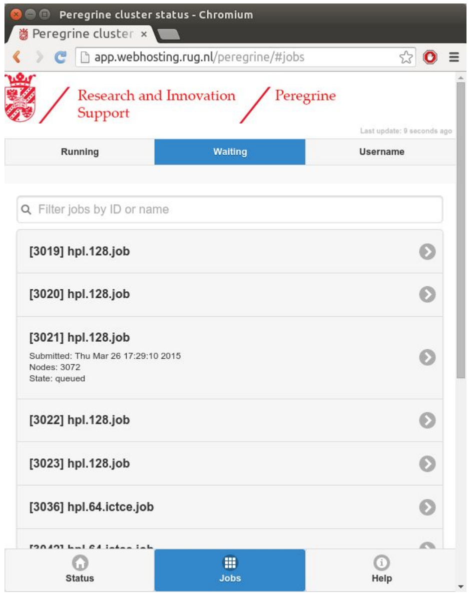
Checking job status: web app

- http://app.webhosting.rug.nl
- Monitor cluster status and usage
- Monitor job status, progress and information
- Intended for smartphones, but also works on desktop
- Also available as MyUniversity widget




Peregrine cluster status - Chromium
Peregrine cluster
app.webhosting.rug.nl/peregrine/#jobs
Research and Innovation Support
Peregrine
Last update: 9 seconds ago
Running
Waiting
Username
Filter jobs by ID or name
2%
Status
Jobs
Help

[3018] hpl.128.job
Started: Mon Mar 30 14:11:27 2015
Submitted: Thu Mar 26 17:29:09 2015
Remaining wallclock time: 23:31:24
Running on: pg-node[001-120,122-129]
Queue: nodes
QoS: normal


Peregrine cluster status - Chromium
Peregrine cluster
app.webhosting.rug.nl/peregrine/#jobs
Research and Innovation Support
Peregrine
Last update: 9 seconds ago
Running
Waiting
Username
Filter jobs by ID or name
Status
Jobs
Help

[3019] hpl.128.job

[3020] hpl.128.job

[3021] hpl.128.job
Submitted: Thu Mar 26 17:29:10 2015
Nodes: 3072
State: queued

[3022] hpl.128.job

[3023] hpl.128.job

[3036] hpl.64.ictce.job

Checking the results

- Unless specified otherwise, output file is written to same directory as from which the job was submitted
- *slurm-<jobid>.out*
- Created when job starts running
- While job is running, new output gets appended

- If the job has state CD or has disappeared from squeue, it has finished

Oops, I did not want to run that!

- scancel

```
$ sbatch testjob.sh
Submitted batch job 2870
```

```
$ squeue -u peter
     JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
      2870     nodes  testjob    peter  R       0:03      1 pg-node021
```

```
$ scancel 2870
```

- Cancel multiple jobs at once:

```
$ scancel --state=PENDING --user=bob --partition=short
```

A real example

› A job script that runs R code:

```
#!/bin/bash
#SBATCH --job-name=R_job
#SBATCH --time=00:01:00
#SBATCH --ntasks=1
#SBATCH --mem=1000
#SBATCH --partition=short

pwd
module load R/3.1.2-goolfc-2.7.11-default
module list
Rscript myscript.r
```

rijksuniversiteit groningen

Another real example

› A job script that runs Matlab code:

```
#!/bin/bash
#SBATCH --job-name=matlab_job
#SBATCH --time=00:02:00
#SBATCH --ntasks=1
#SBATCH --mem=1000
#SBATCH --partition=short

module load MATLAB/2014b-GCC-4.8.4
module list
matlab -nodisplay -r mycode
```

Code in file mycode.m

rijksuniversiteit groningen

Support

- Support through CIT central service desk
  - Phone: 3232
  - E-mail: citservicedesk@rug.nl
- Online documentation and account request form:
  - https://redmine.hpc.rug.nl
- Comments and questions are always welcome

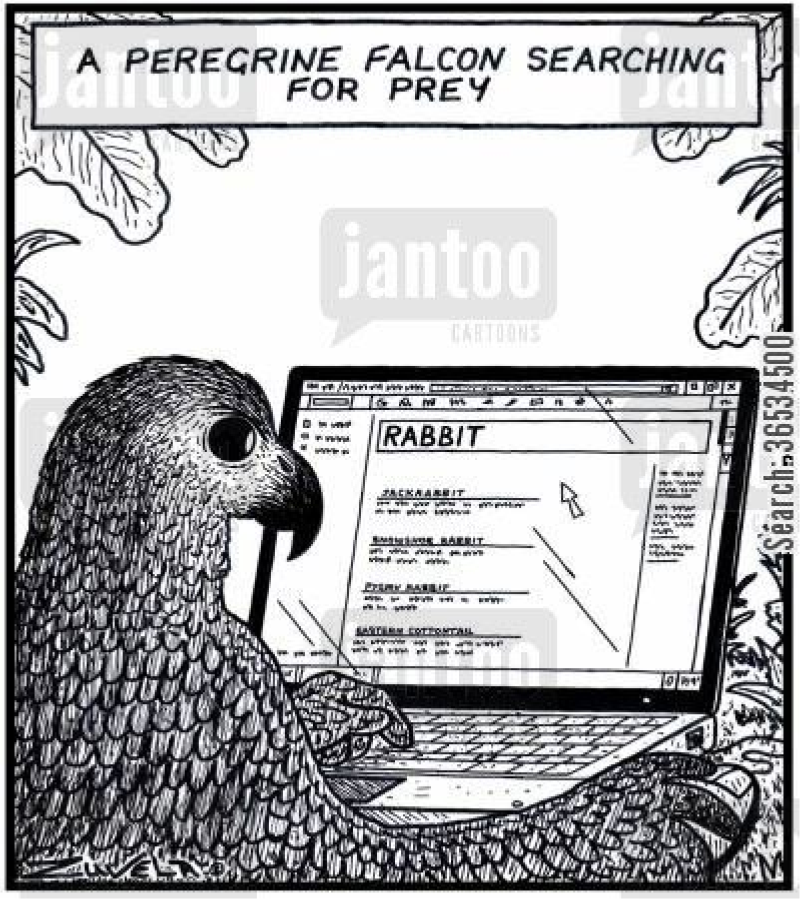
› I want to know more!

Useful documents

- Introduction to Linux by Machteld Garrels: http://tldp.org/LDP/intro-linux/html/index.html
- Bash shell guide by Machteld Garrels: http://tldp.org/LDP/Bash-Beginners-Guide/html/index.html
- Linux guide: http://www.tuxfiles.org
- Linux command line: http://linuxcommand.org/

- Documentation and more details about SLURM: http://slurm.schedmd.com
- Manual pages for all SLURM commands: http://slurm.schedmd.com/man_index.html

Questions?



http://www.jantoo.com/cartoon/36534500

rijksuniversiteit groningen

Peregrine team

Wietze Albers

Fokke Dijkstra, Niels Idsinga, Ger Strikwerda, Robin Teeninga, Bob Dröge, Laurent Jensma, Henk-Jan Zilverberg, Wim Nap

Exercises

- Hostname: peregrine.hpc.rug.nl
- Username & password have been handed out

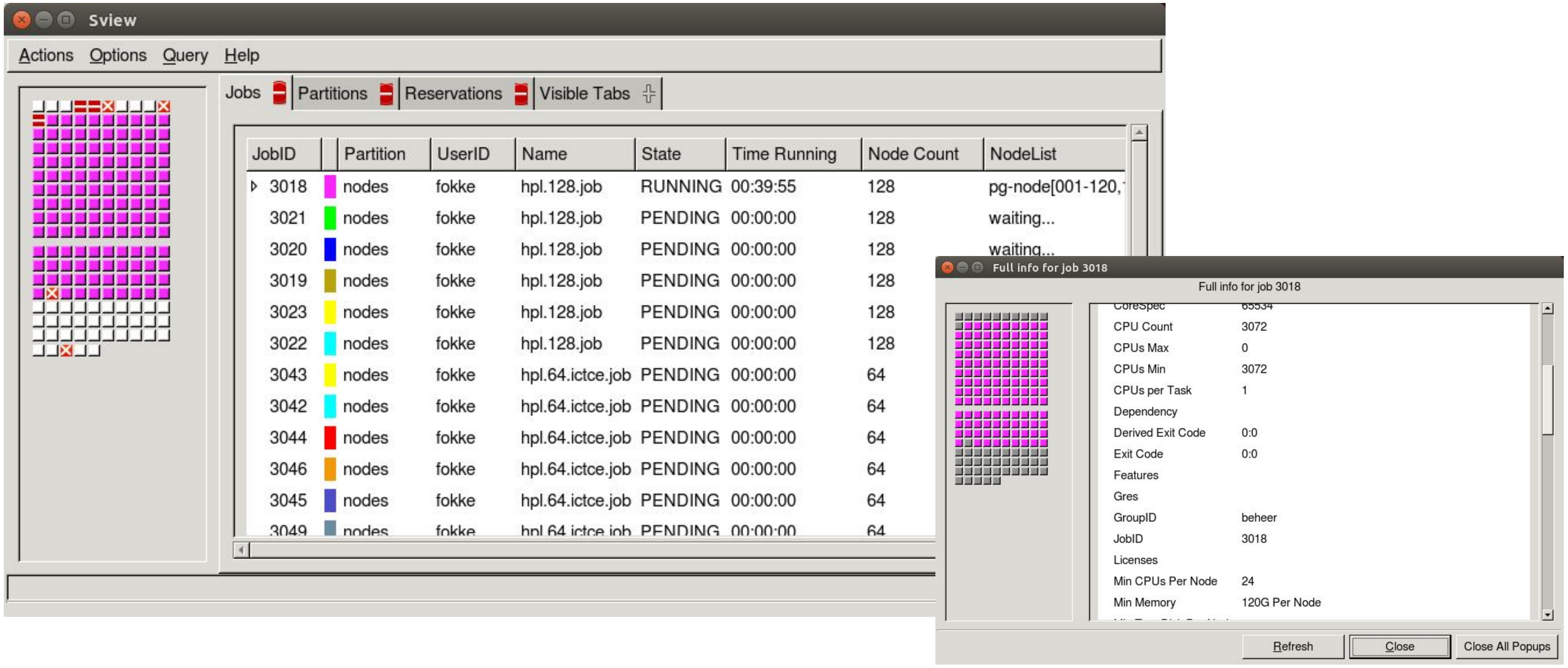
Checking partition details

› sinfo

```
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
nodes*       up 10-00:00:0      2  down* pg-node[006,141]
nodes*       up 10-00:00:0      2  drain pg-node[068,111]
nodes*       up 10-00:00:0     27   resv pg-node[005,007-015,018-021,023-035]
nodes*       up 10-00:00:0     28  alloc pg-node[104-110,142-162]
nodes*       up 10-00:00:0     95   idle pg-node[016-017,022,036-067,069-099,112-140]
short        up      30:00      4    mix pg-node[100-103]
short        up      30:00      2   idle pg-node[003-004]
gpu          up 3-00:00:00      4   comp pg-gpu[01-04]
phi          up 3-00:00:00      2   idle pg-phi[01-02]
himem        up 10-00:00:0      7   idle pg-memory[01-07]
```

GUI for showing all cluster details

- sview on the login node
- See slides about running applications with a GUI



JobID		Partition	UserID	Name	State	Time Running	Node Count	NodeList
3018	🟪	nodes	fokke	hpl.128.job	RUNNING	00:39:55	128	pg-node[001-120,
3021	🟩	nodes	fokke	hpl.128.job	PENDING	00:00:00	128	waiting...
3020	🟦	nodes	fokke	hpl.128.job	PENDING	00:00:00	128	waiting...
3019	🟫	nodes	fokke	hpl.128.job	PENDING	00:00:00	128	
3023	🟨	nodes	fokke	hpl.128.job	PENDING	00:00:00	128	
3022	🟦	nodes	fokke	hpl.128.job	PENDING	00:00:00	128	
3043	🟨	nodes	fokke	hpl.64.ictce.job	PENDING	00:00:00	64	
3042	🟦	nodes	fokke	hpl.64.ictce.job	PENDING	00:00:00	64	
3044	🟥	nodes	fokke	hpl.64.ictce.job	PENDING	00:00:00	64	
3046	🟧	nodes	fokke	hpl.64.ictce.job	PENDING	00:00:00	64	
3045	🟦	nodes	fokke	hpl.64.ictce.job	PENDING	00:00:00	64	
3049	🟦	nodes	fokke	hpl.64.ictce.job	PENDING	00:00:00	64	

Full info for job 3018

CoreSpec	65534
CPU Count	3072
CPUs Max	0
CPUs Min	3072
CPUs per Task	1
Dependency	
Derived Exit Code	0:0
Exit Code	0:0
Features	
Gres	
GroupID	beheer
JobID	3018
Licenses	
Min CPUs Per Node	24
Min Memory	120G Per Node