UNIVERSITY OF Southampton

Introduction to cluster computing on Iridis 4

Administrative

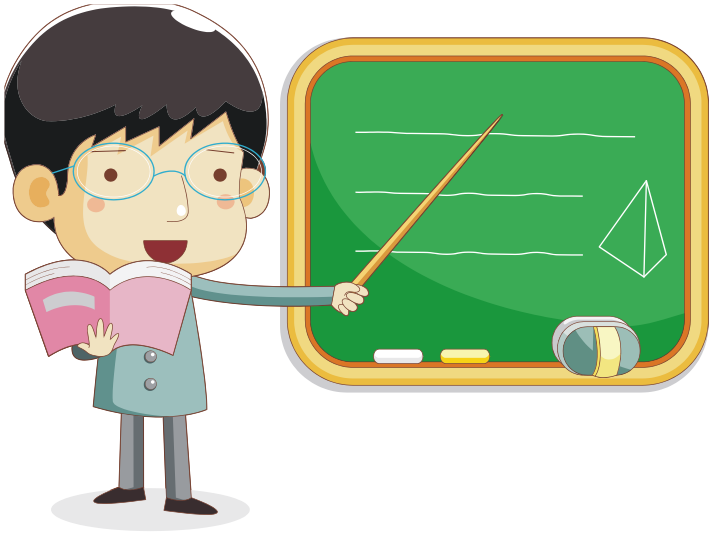
- Fire Alarms/exits/toilets
- Coffee break
- Phones
- Handouts

Tutorial Objectives

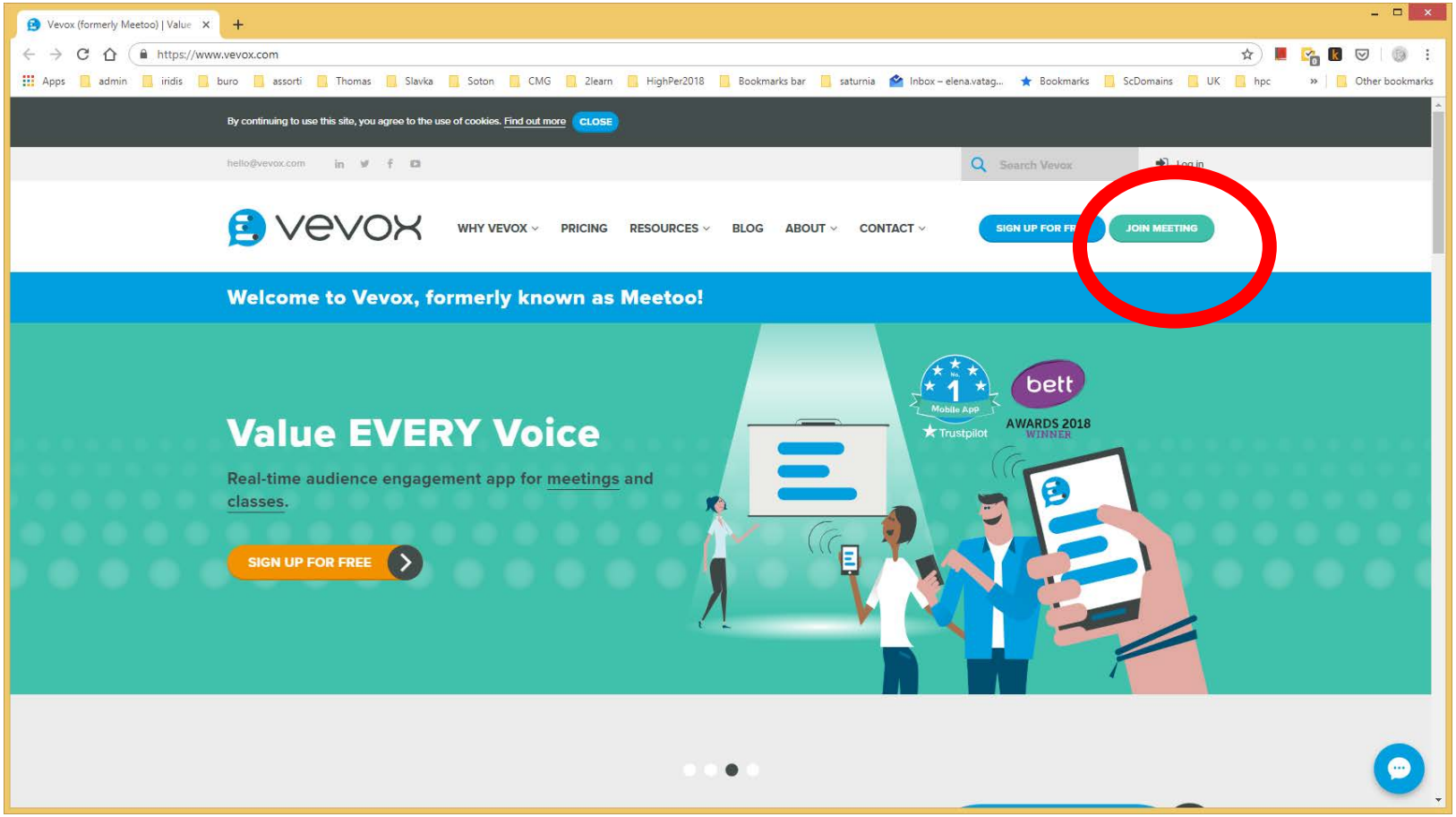
By the end of this training course, participants will be familiar with:

- Iridis 4 cluster structure
- How to find your software
- Different job types
- Moab/PBS environment: how to
  - submit
  - monitor
  - delete

  a batch job on Iridis



Meetoo: www.meetoo.com
meeting ID: **164-979-246**



UNIVERSITY OF Southampton

History of HPC in UoS

– Operator at console. User approaches operator direct. 'My job finished yet?'

HPC @ Southampton

UNIVERSITY OF Southampton

IBM3090 (1980's)

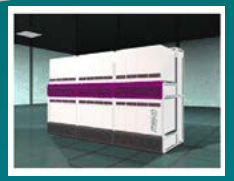
IBM SP2 (1990's)

Meiko CS2 Origin 2000 (1990's)

Iridis 1-4 (2001-present)

ICL (1960's & 70's)

Pegasus (1950's)



Anatomy of HPC cluster

UNIVERSITY OF Southampton

Compute nodes – as many as possible

Interconnect – as fast as possible

Storage – as much as possible

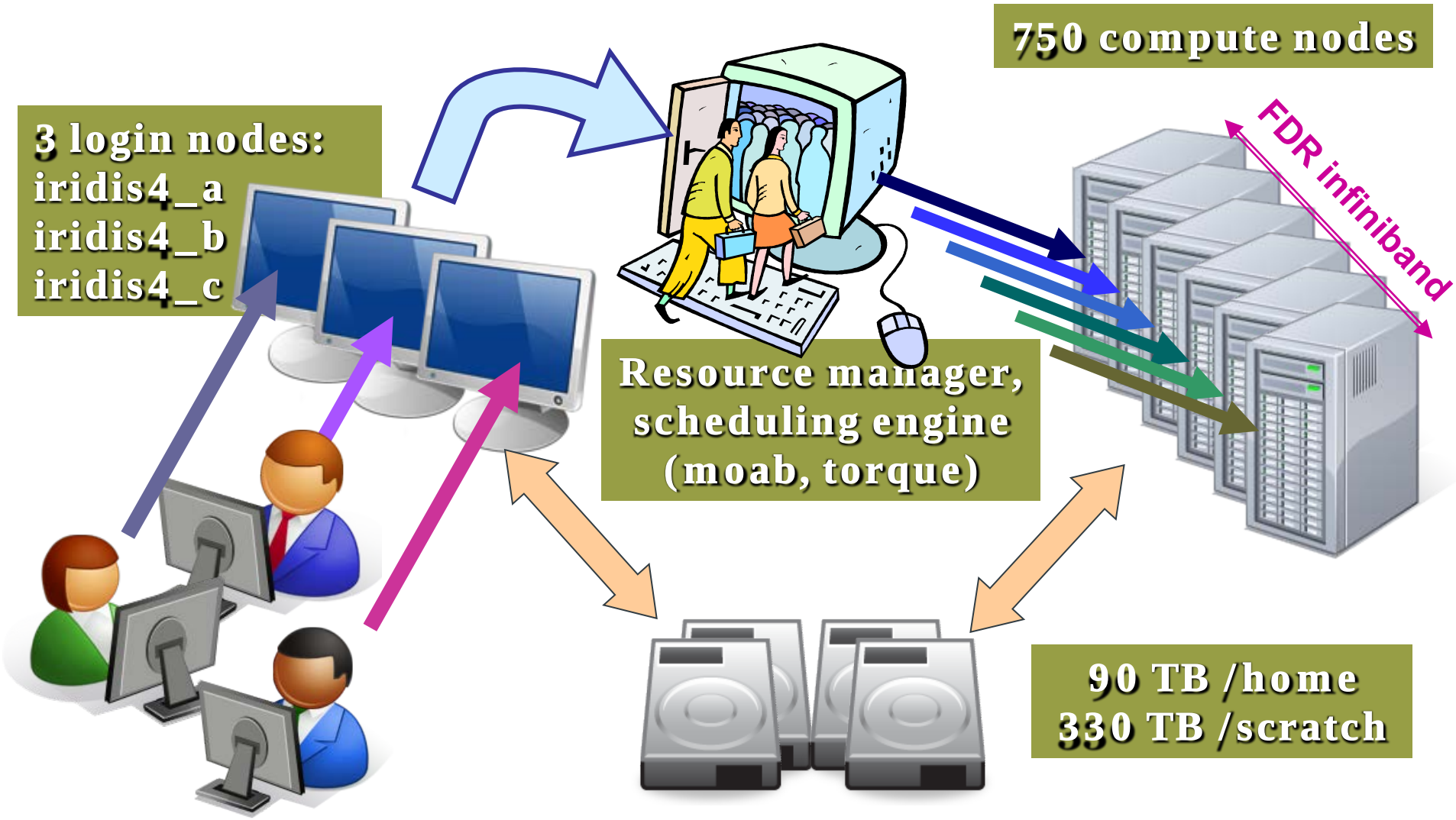
Software

Few management and login nodes

Happy users

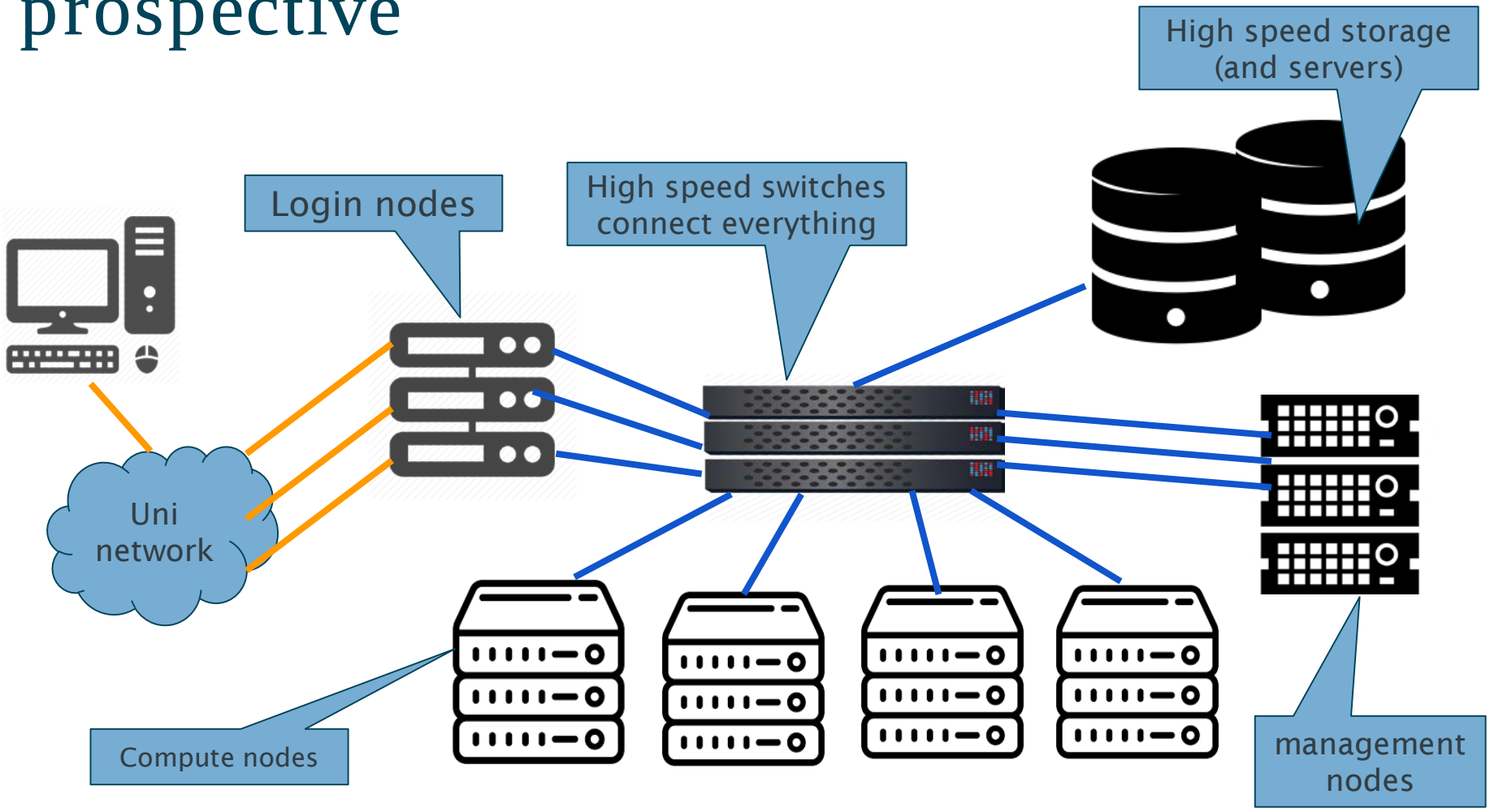
Iridis 4: components

UNIVERSITY OF Southampton



Cluster component - admin prospective

UNIVERSITY OF Southampton



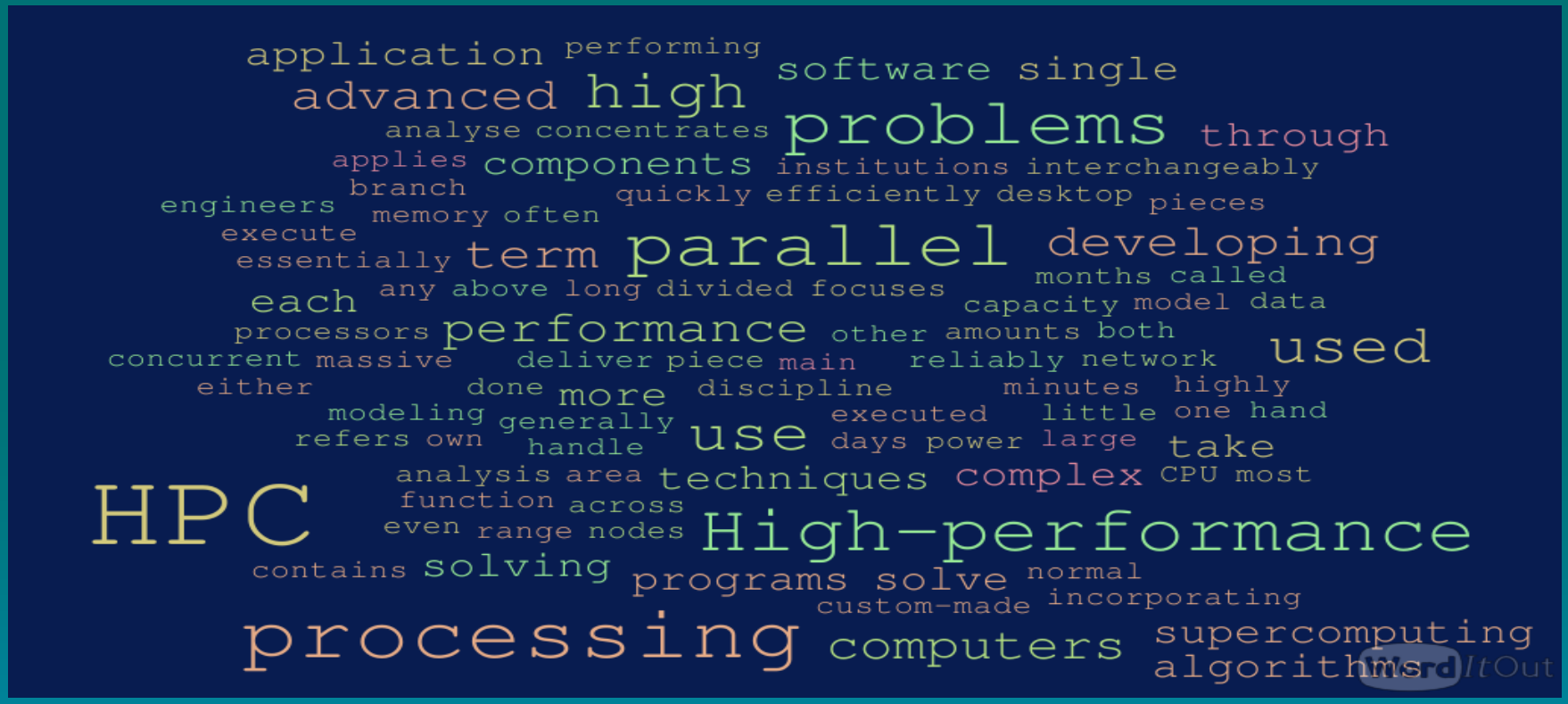
Iridis 4 specifications

- **#216 in the world (based on July 2014 TOP500 list) with R_{peak} ~ 249 TFlops/s**
- **750 2.6 GHz nodes with 16 cores per node – 4 GB memory per core, 64 GB per node**
- **4 FAT nodes with 32 cores and 256 GB of memory**
- **800 TB disk with parallel file system** (>12 GB/s)
- **£3M Project delivered by OCF/IBM**
- **Iridis 5 added in 2018, more computational power, less nodes**

http://www.top500.org/

UNIVERSITY OF Southampton

High Performance Computing =

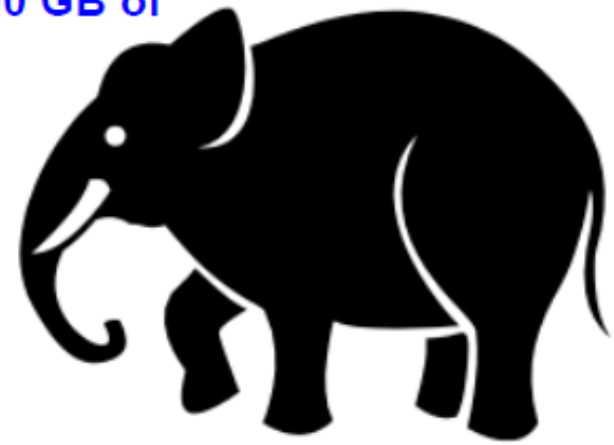


The term High Performance Computing (HPC) refers to any computational activity requiring more than a single computer to execute a task.

Typical HPC problem...

> Can't fit on a PC – usually because they need more than a few GB of RAM, or more than a few 100 GB of disk.

Size



> Take a very very long time to run on a PC: months or even years. But a problem that would take a month on a PC might take only a few hours on a supercomputer.

Speed

Laptop

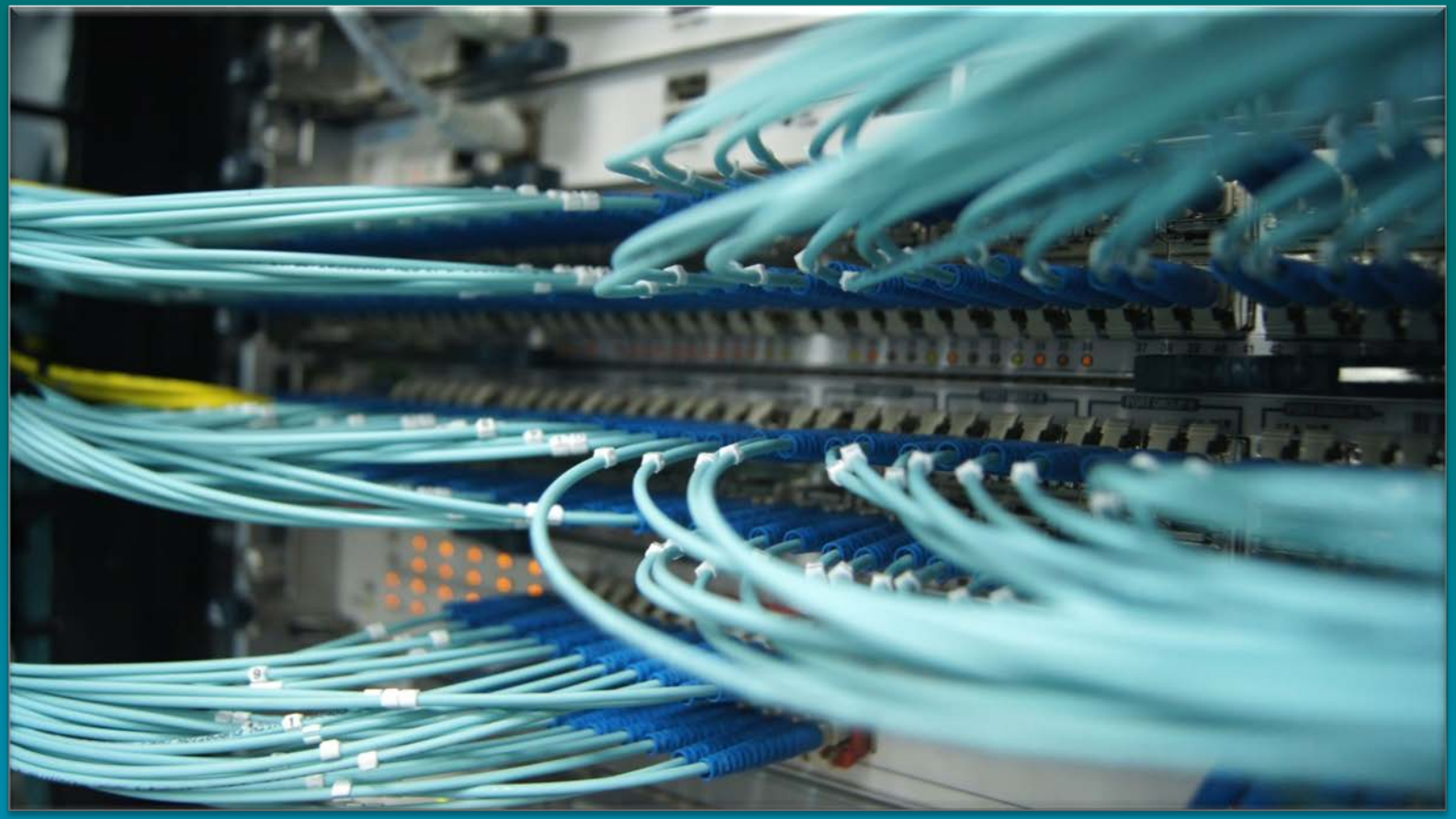
From Jason Shih

UNIVERSITY OF Southampton

High-performance computing is really different. No matter what you have learned about coding, using HPC clusters takes a whole new skill set.

UNIVERSITY OF Southampton

Getting ready to roll



Using Iridis: Security

- Boring but very, very important
- out IT infrastructure is under constant attack by would-be intruders;
- Your data and research career is threatened by intruders;
- Big systems are big, juicy targets;
- Don't let intruders in:
  - Keep your password (or private key passphrase) safe.
  - Choose a strong password.
  - Keep the software on your laptops/tablets/PCs up to date.
  - Don't share accounts (this is against the rules anyway).

Connecting to Iridis: SSH

- Usually available out of the box on Linux or MAC

  $ ssh –Y abc123@iridis4_a.soton.ac.uk

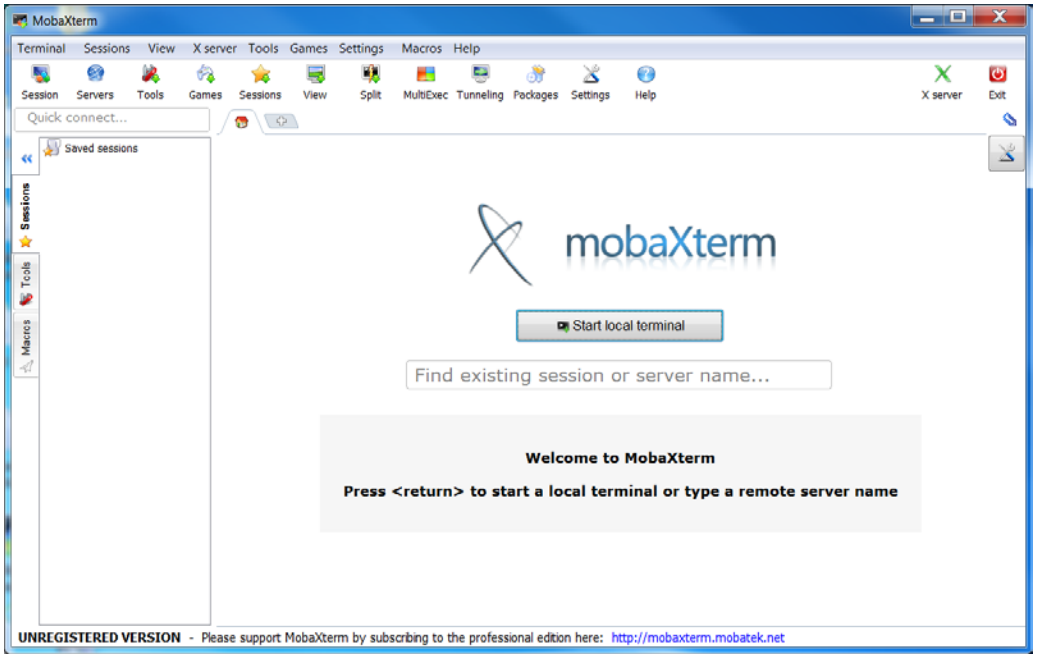
- Windows clients:
  - SSH client + Exceed
  - Cygwin

    http://cygwin.com/install.html

    Includes X server for displaying graphical applications running remotely.
  - MobaXterm
- Need to be on internal network or use VPN

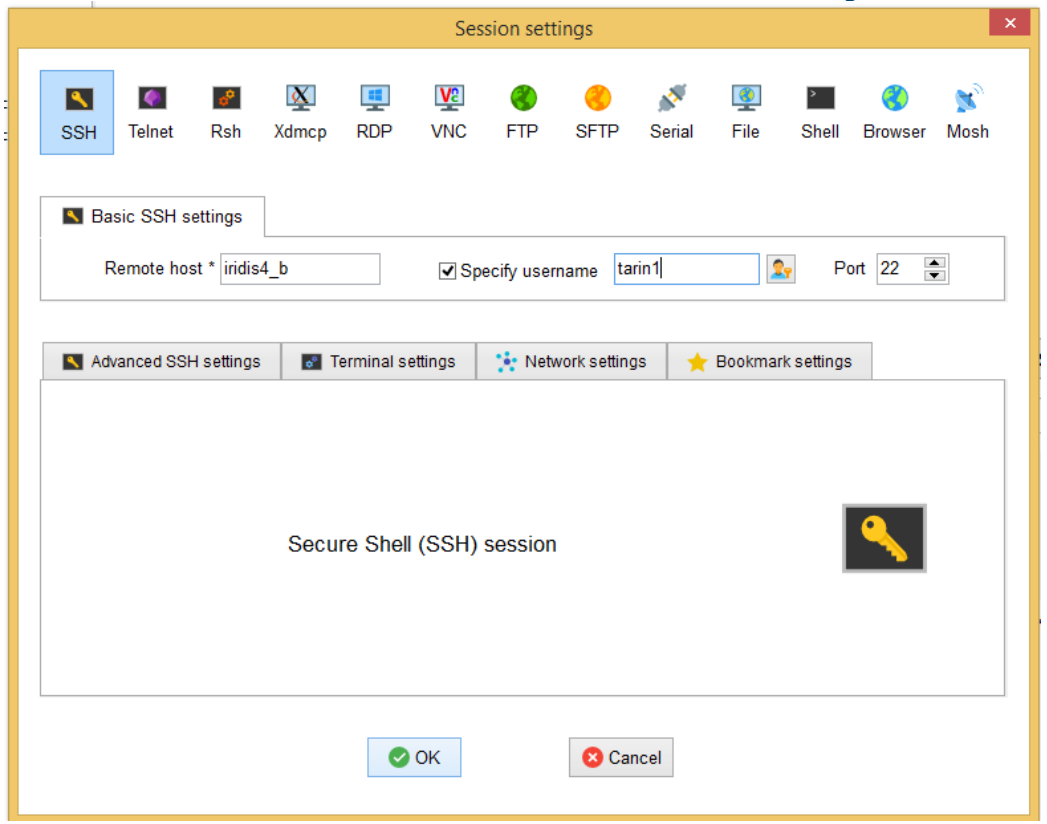
SSH client - MobaXterm

- http://mobaxterm.mobatek.net/
- Go to Download – Portable edition



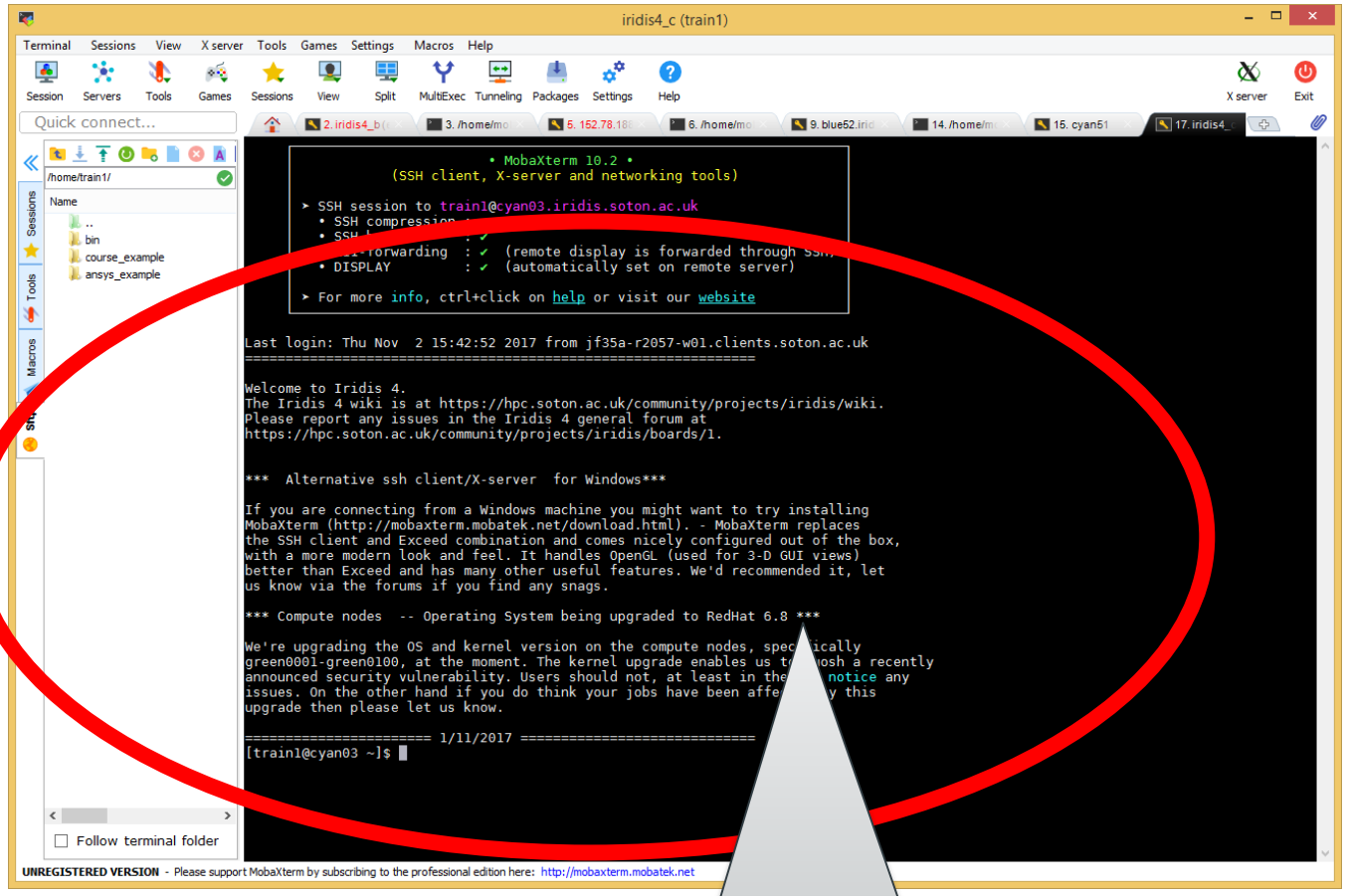
Login to Iridis

- Start ssh client MobaXterm
- Go to Sessions -> SSH
- Remote host =
  - iridis4_a
  - Iridis4_b
  - Iridis4_c
- Use your iSolutions username/password
- Press OK
- You may want to save password
- Try to log out (“exit”) and login again



The authenticity of host 'iridis4_b (152.78.188.134)' can't be established.
RSA key fingerprint is 01:60:2d:58:1b:24:05:e9:a2:d6:17:3b:25:c5:be:ef.
Are you sure you want to continue connecting (yes/no)?

MOTD= Message Of The Day



Read latest system announcement/status/news

UNIVERSITY OF Southampton

Introducing Iridis 4

New user kit

UNIVERSITY OF Southampton

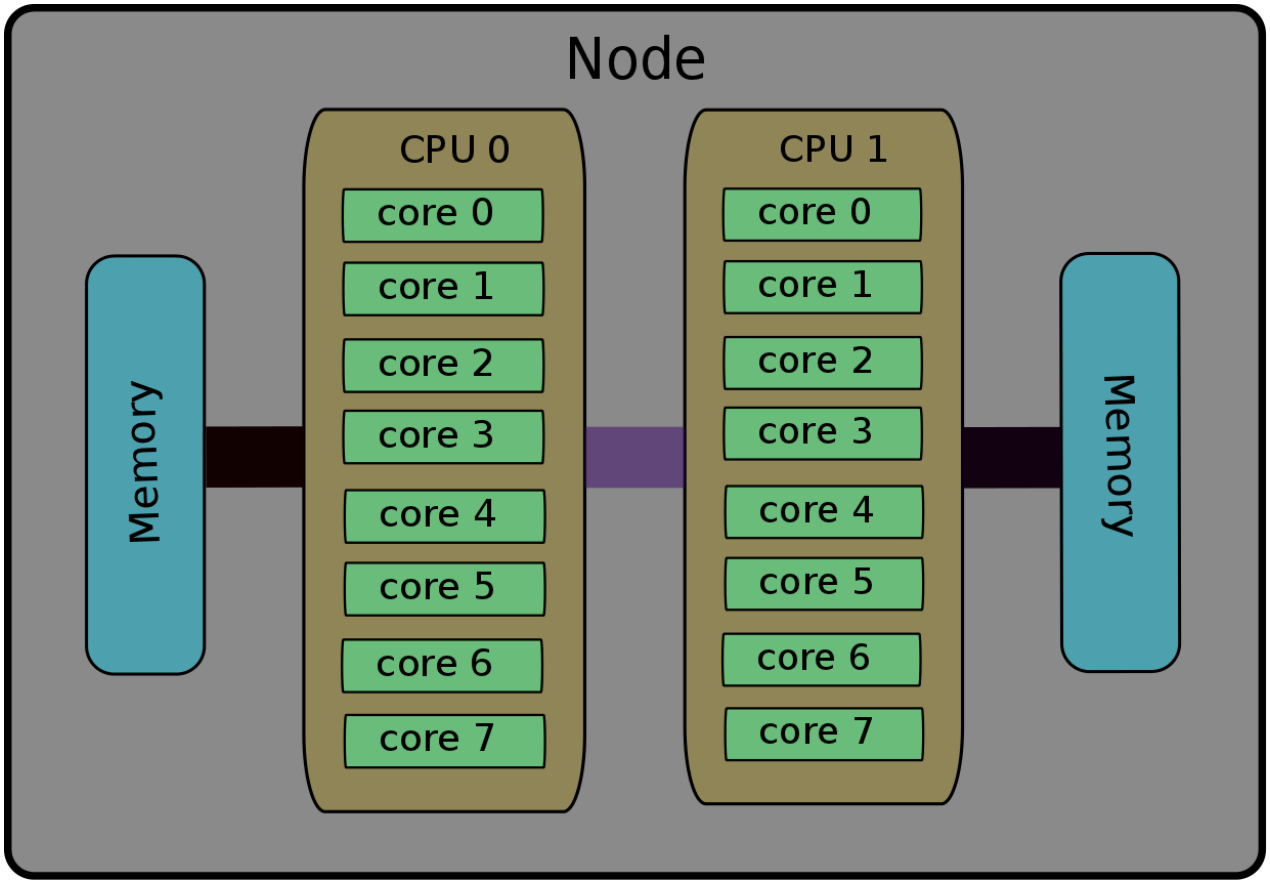
/home 100 GB (backed up)

/scratch 1 TB

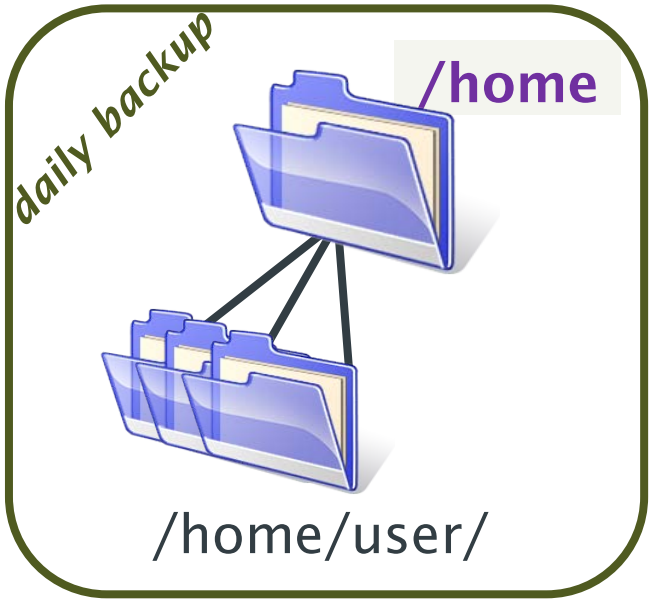
Up to 512 cores (32 nodes)

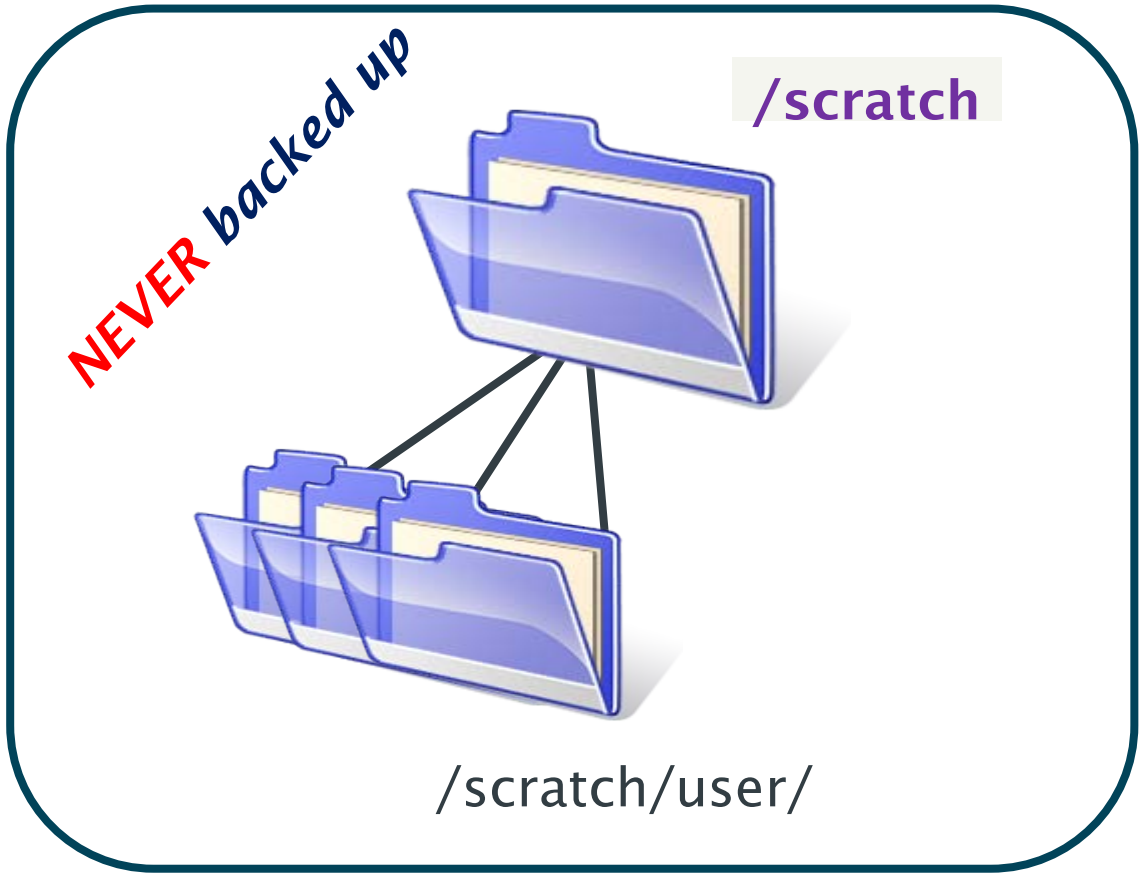
Software

Compute node (greenNNNN): 16 CPUs 64 GB



Iridis File System





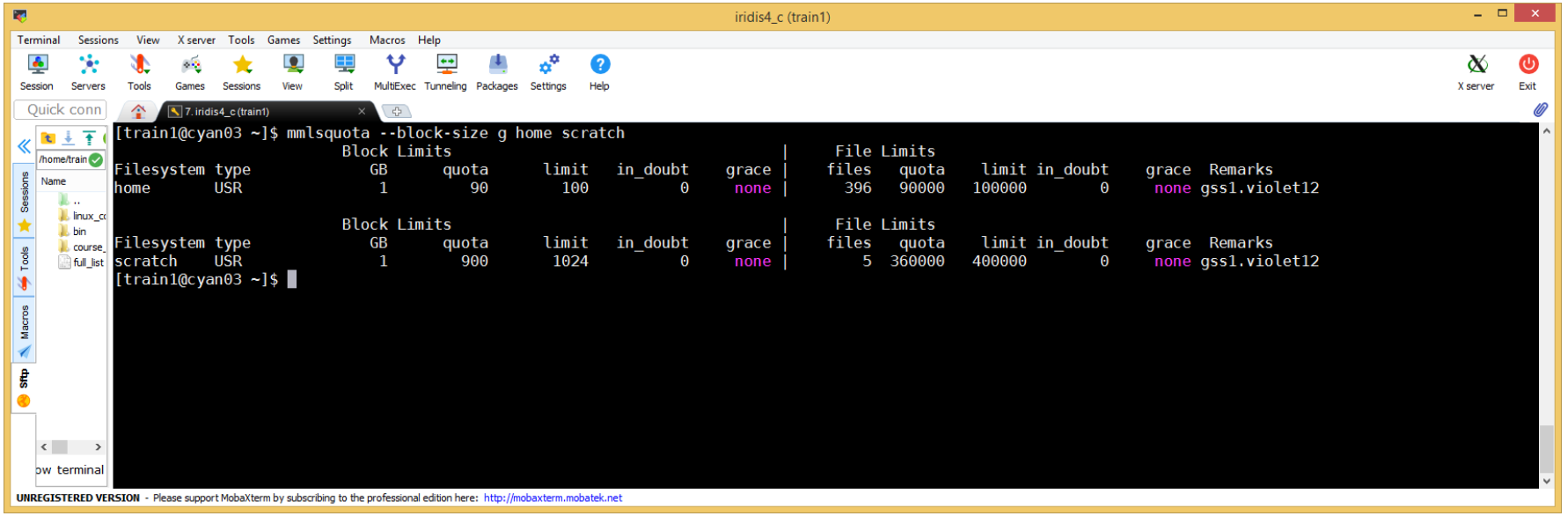
File system	Size	inodes
/home	100 GB	100 000
/scratch	1000 GB	200 000

```
$mmlsquota home scratch --block-size g
```

Filesystems: Quota

University of Southampton

```
$ mmlsquota --block-size g home scratch
```



- Aim to stay below the soft limit (quota)
- Once over the soft limit, you have grace period of 30 days to return below quota

POLL OPEN

Which of the following statement is true:

Vote for up to 6 choices

1. It's better to keep your research data in /scratch then in /home because /scratch quota is 10 times bigger;
2. If I exceed my quota in both scratch and home I will not be able to login to Iridis;
3. I can keep as many files as I want in my home directory as long as their total size is under 100GB;
4. I will get email notification when I am close to my quota;
5. Once grace period is over, I need to shrink my files under quota to be able to use my allocation;
6. All Iridis storage is backed up: /home is backed up every day and /scratch - once per month;

UNIVERSITY OF Southampton

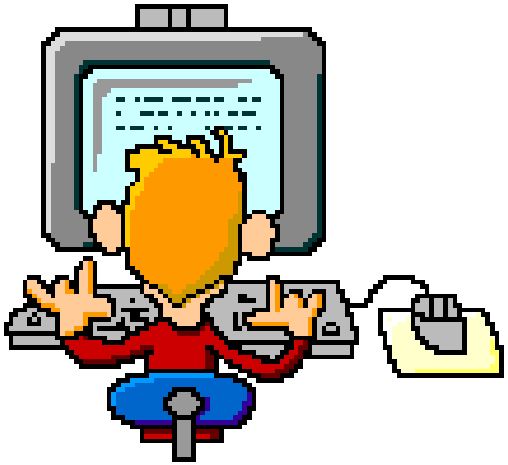
Software Environment Management - Modules

A module system is used to manage the software on Iridis 4

Try:

```
$ module av[ail]
$ module li[st]
$ module help matlab
$ module load matlab
$ module show R/2.15.3
$ module load R
$ module switch R/3.0.2
$ module --help
```

How to use cluster?



Few easy steps:

- Connect to Iridis
- Determine required software
- Transfer input files & source code
- Compile programs (if needed)
- Write submission script
- Test software/programs/scripts
- **Submit batch jobs**
- Get results, write papers & get PhD!

UNIVERSITY OF Southampton

Managing Batch Jobs

Jobs

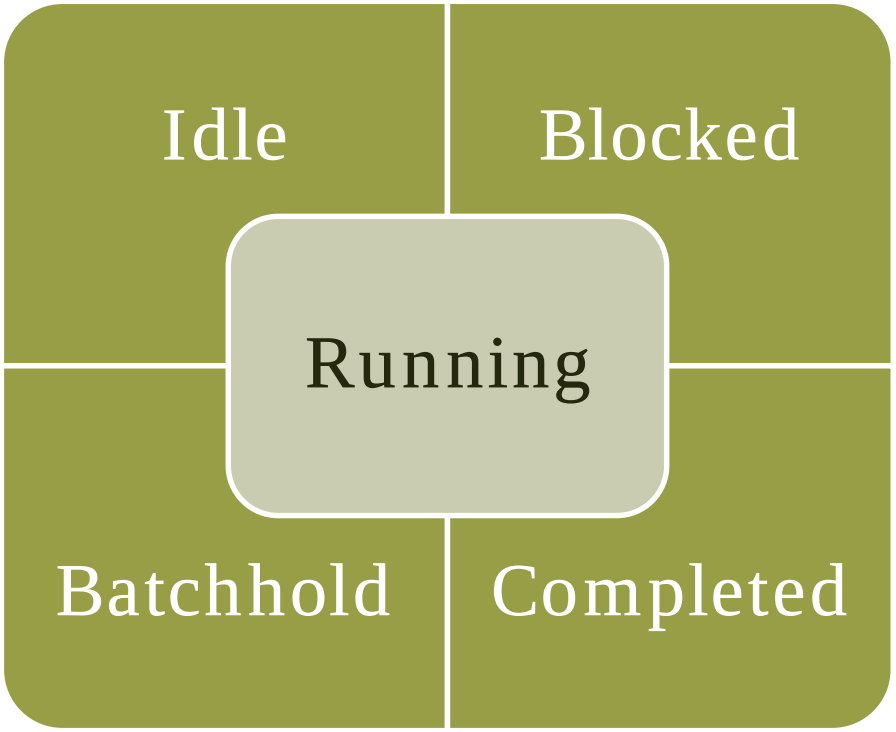
- Job = user application
- Several types of jobs:
  - Batch vs. interactive
    - batch = does not require user's interaction
    - interactive = useful for testing codes, scripts & input, using graphical interfaces (GUI)
  - Serial vs. parallel
    - Serial = job requires only one processor for execution
    - Parallel = job requires more then one processor for execution, probably over multiple nodes

UNIVERSITY OF Southampton

Job Management System

- Resource manager: TORQUE-PBS – manages jobs
- Job scheduler: MOAB – decides what jobs run and when
  - Fairness between users (but it's not easy)
  - Optimizes cluster node utilization
- User interface for submitting and monitoring jobs
- A submission script is used to tell the resource manager the resources required, how to run the job

Job state – how to check it



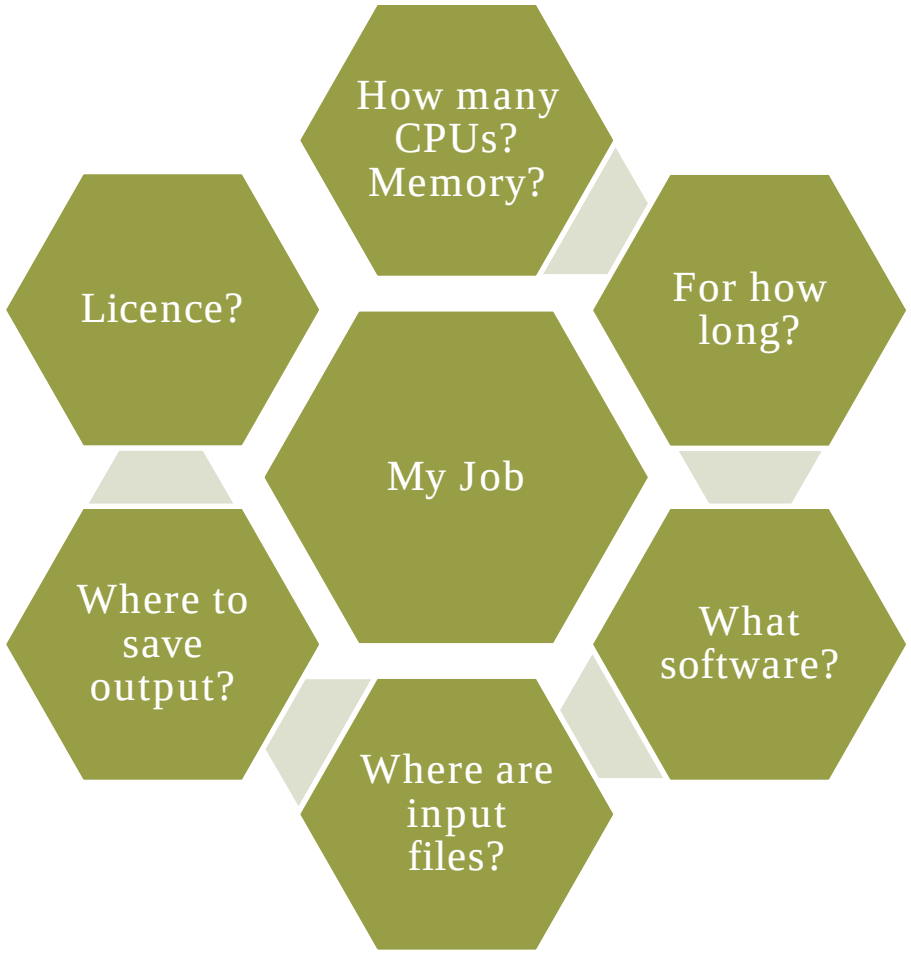
All jobs:

```
$ showq |less
$ showq [-r/b/i/c]
```

Your jobs:

```
$ qstat
$ qstat –a
$ qstat -f
$ checkjob JID
$ checkjob -v JID
```

Specifying job resources



Job Submission: qsub

- Submits batch job defined in script
- Returns **job ID** used for job management
- Syntax

```
$> qsub [options] <job_script>
```

- Options – job description options, equivalent to options in script (command line take precedence)
- Each job produces 2 files with output & error messages, `jobname.o<jobid>` & `jobname.e<jobid>`

Let's try Interactive Job:

- To request an interactive session:

```
$ qsub –I -X
```

- Start an interactive job and
  - Note node name: `green0NNN` as opposite to `cyan0N`
  - Monitor it from another terminal with `qstat, checkjob <JID>`
  - Try to figure out requested resources
  - Run on login and compute nodes commands `who` and `top` and compare output
  - Kill this job with `qdel <JID>`
- **You share login nodes with other users – do not run CPU intensive jobs on them! Use batch system instead**

And now practice!

```
$ copy_example
$ copy_example course2
$ cd course2_example
$ ls -l
```

Each example has README file

Let's do examples:

- 01_simple_example
- 02_simple_exercise
- job_array_example
- Any other example

Default values and Environment Variables

- `PBS -l walltime=2:00:00`
- `PBS -l nodes=1:ppn=1`
- `PBS –l pmem=4gb`
- `$PBS_O_WORKDIR  # dir from which job was submitted`
- `Default queue: batch`

Anatomy of job script

```
#!/bin/bash

#PBS -l walltime=00:30:00

cd $PBS_O_WORKDIR

# Say hello to the user

echo "Hello $USER“

sleep 300

echo "Goodbye $USER“
```

Important things to remember:

All lines starting from

#PBS

.....are called PBS directives

PBS will scan all the initial comment lines for PBS directives until the first executable line. Any directives embedded in the script after some executable lines will be ignored.

POLL OPEN

My job is single-threaded and requires 20GB of memory for about 10 hours. Which pbs command/directive will garantee my job to complete?

Vote for up to 4 choices

1. #PBS –l walltime=12:00:00,pmem=20g
2. qsub -l walltime=10:00:00,mem=20g
3. #PBS -l nodes=1:ppn=6, walltime=12:00:00
4. qsub -l nodes=1:ppn=16,walltime=20:00:00

POLL OPEN

My job is multi-threaded (up to 16) and requires 20GB of memory for under 10 hours. Which pbs command/directive shell I use?

Vote for up to 4 choices

1. #PBS –l walltime=10:00:00,mem=20g
2. qsub -l walltime=10:00:00,mem=64g
3. #PBS -l nodes=1:ppn=5, walltime=10:00:00
4. qsub -l nodes=1:ppn=16,walltime=10:00:00

Job Arrays

- Best solution when you need to run many very similar jobs (for example process many input files)
- Single input script functions as a template
- $PBS_ARRAYID – array index of the job

```
$ qsub –t 0-10 my_job.pbs
$ qsub –t 100 my_job.pbs
$ qsub –t 4-9,12,800 my_job.pbs
```

Check your job output files:

<scriptname>.e<jobid>: Standard Error
<scriptname>.o<jobid>: Standard Output

For example script pbs_example.sh would generate the files:

pbs_example.sh.e10245
pbs_example.sh.o10245

Resource Analysis

Requested resource limits are
neednodes=1:ppn=16,nodes=1:ppn=16,pmem=4000mb,walltime=10:00:00
Used resource limits are
cput=71:49:50,mem=5086416kb,vmem=6293876kb,walltime=05:10:35

```
Job efficiency = CPU_time / (#CPUs x wall_time)
```

Why won't my job run?

Possible reasons could be:

- error in job script => check error file
- resource issue => not enough memory?
- cluster is full & you'll have to wait
- Iridis is down or in maintenance => check Iridis status page
- you have exceeded your disk quota:
  `$ mmlsquota home scratch`

Few reminders...

- Check MOTD for coming maintenance period/known problems
- Check Iridis status pages:

https://hpc.soton.ac.uk/community/projects/iridis/wiki/System_Status

- If reporting a problem – provide us with
  - ✓ your job ID
  - ✓ path to submission script
  - ✓ instructions on how to reproduce the problem
- Do not run CPU intensive jobs on login nodes.

Iridis 4 or Iridis 5 ?

Iridis 5:

- 464 compute nodes with dual Intel Skylake processors;
- Each compute node = 40 CPUs and 192 GB of memory;
- Newer OS => RHEL 7;
- Another resource manager: SLURM instead of PBS;
- Latest GPUs 20 Nvidia Tesla V100;
- 4 High memory nodes: 64 cores & 768 GB of RAM;
- Hadoop/Spark data service;
- Private Cloud (coming soon);
- Data Visualisation;

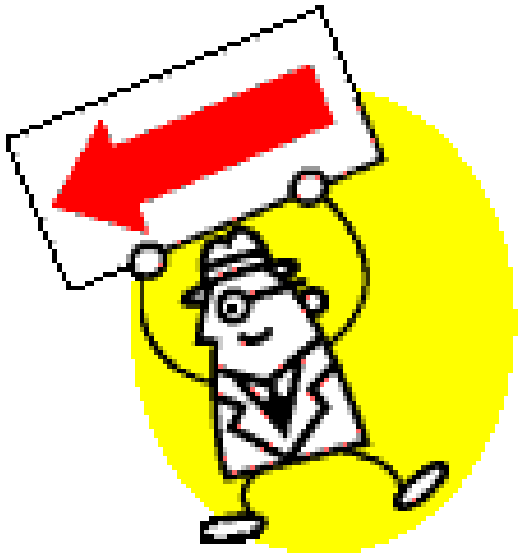
Request transfer to Iridis 5 on Iridis forum

Tutorial Outcome

By now participants are familiar with:

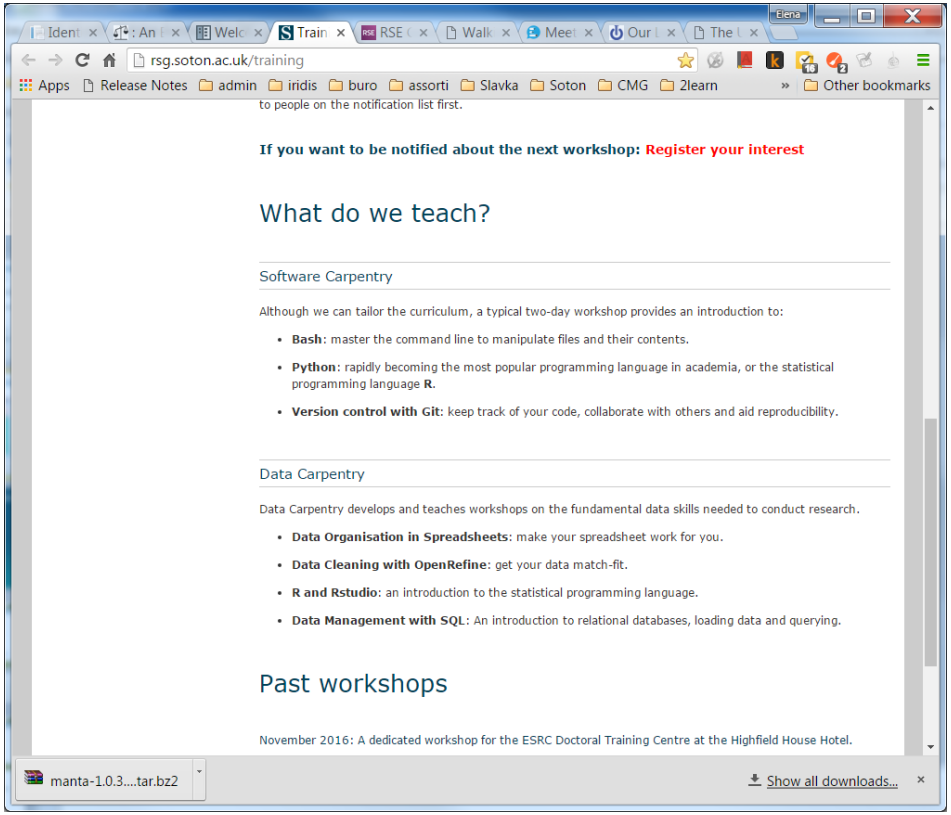
- Iridis 4 cluster components
- Basic Linux commands
- Software modules
- Different job types
- Moab/PBS environment: how to
  - submit
  - monitor
  - control

  a batch job on Iridis



Further training

- Ask questions on Iridis Forum
- Request training on Iridis forum
- Follow one of Software Carpentry courses:
  - http://rsg.soton.ac.uk/training



UNIVERSITY OF Southampton

Backup