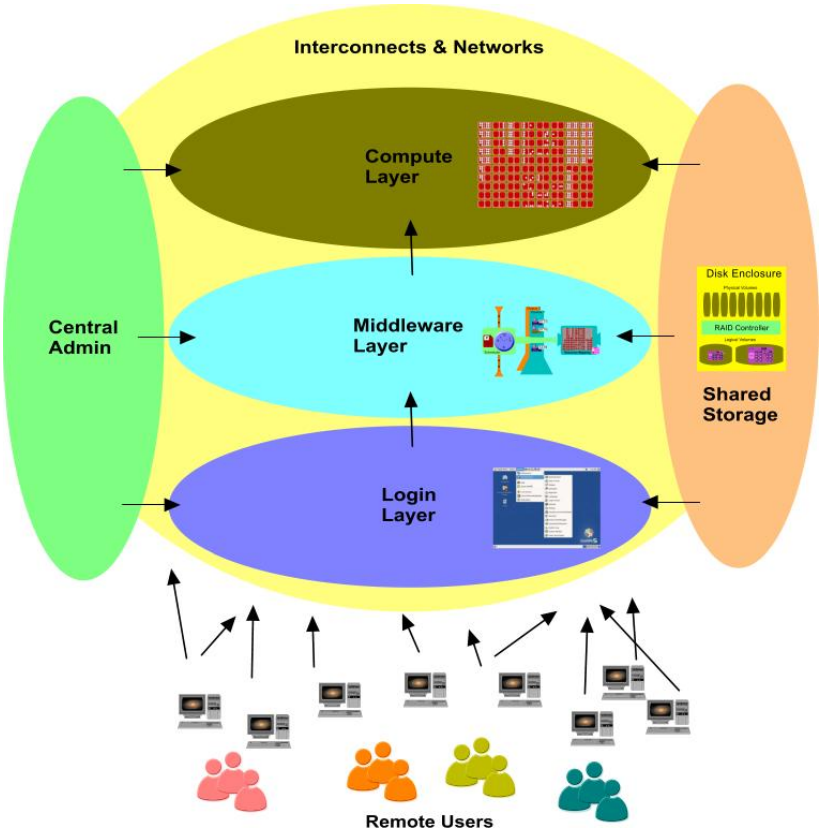
HPC Software and Management



Overview

- Common phrases and building blocks
- Memory models (reviewed)
- Managing software / Modules
- Schedulers and resource managers
- Being a good HPC citizen

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Common Phrases

- Login Server – The machine you connect to from your desktop (laptop).
- Computer server/node – A machine in the compute pool (3700 cores). Node = PC
- Job – A task to be completed by the compute pool. Will comprise of a program and some meta data (size,time etc)
- Queue – The place where jobs wait before they start.
- Lustre storage – Large project data area
- NFS storage - $HOME, the area you log into.
- Batch / Interactive jobs – job type sent to queues.

SCAMA Portsmouth

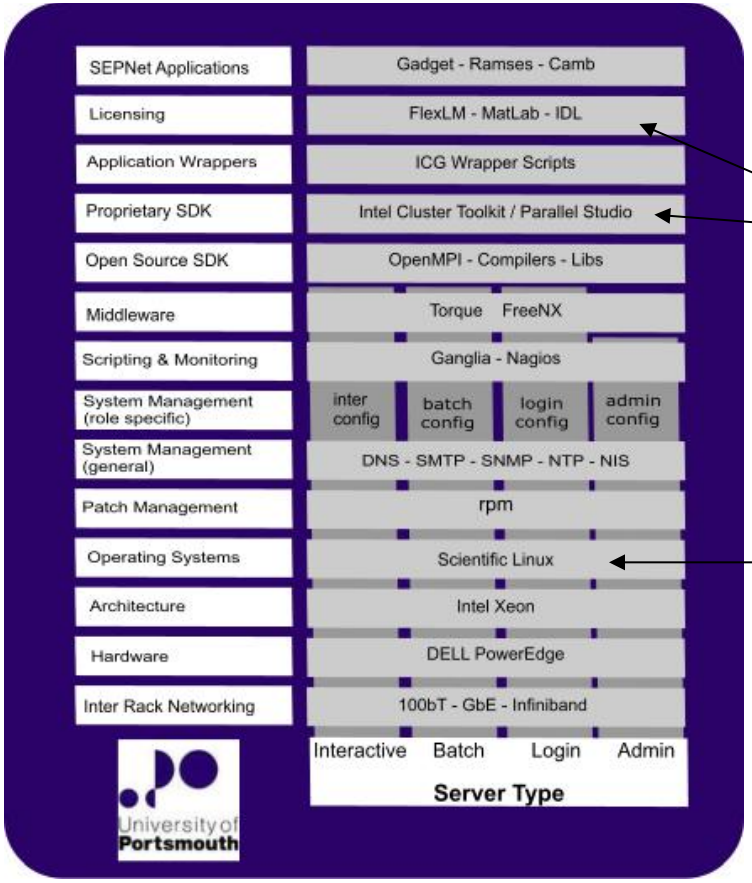
UNIVERSITY OF PORTSMOUTH

The STACK - Open Source HPC Stack

Two main Enterprise Linux players:-

RedHat & SUSE

Stability is all important. No need for latest and greatest (Fedora 16)



Licensed

RedHat Clone

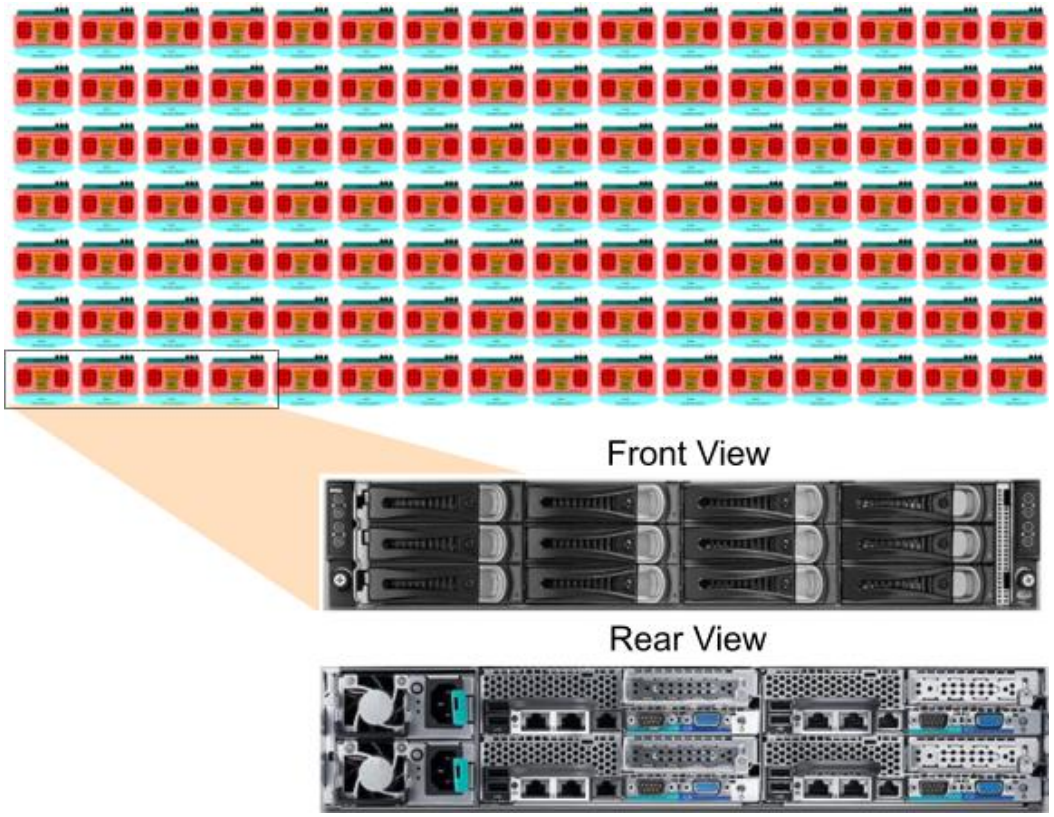
(CERN)

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Memory Models

Distributed Memory Model



SCAMA Portsmouth

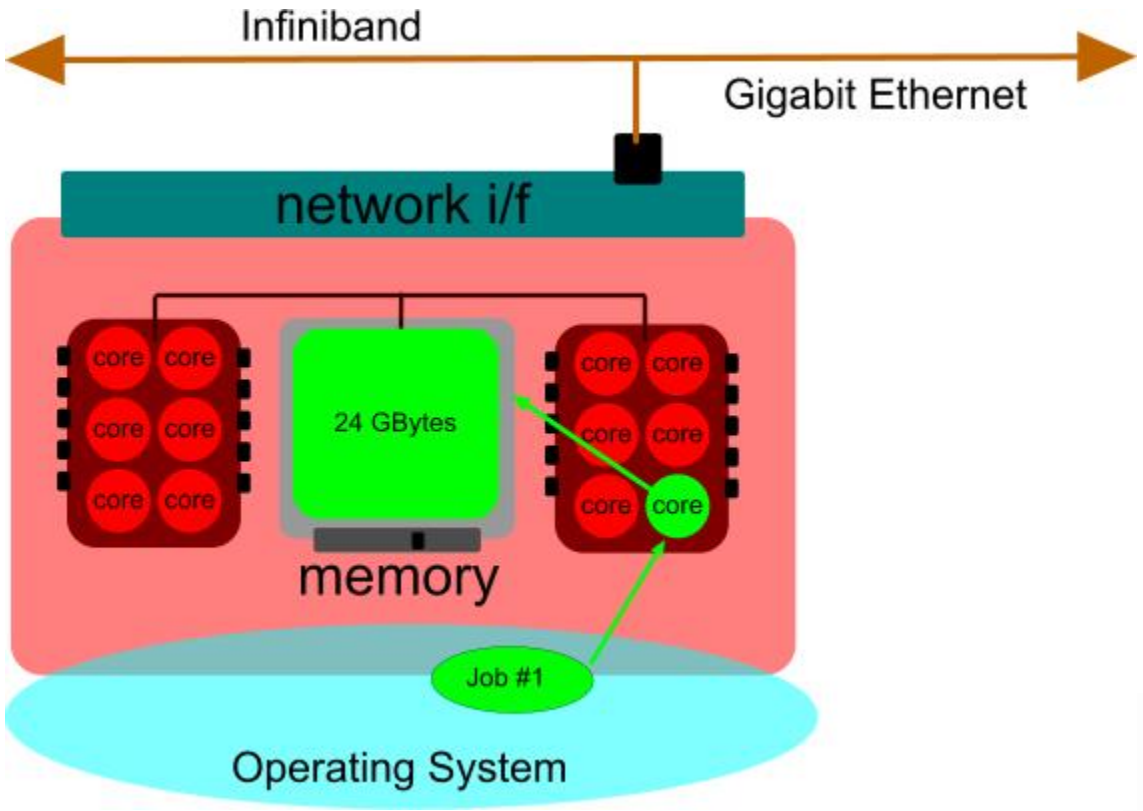
UNIVERSITY OF PORTSMOUTH

Largest (sensible) job is 24/64Gbytes in this distributed memory model

Paging

Swapping

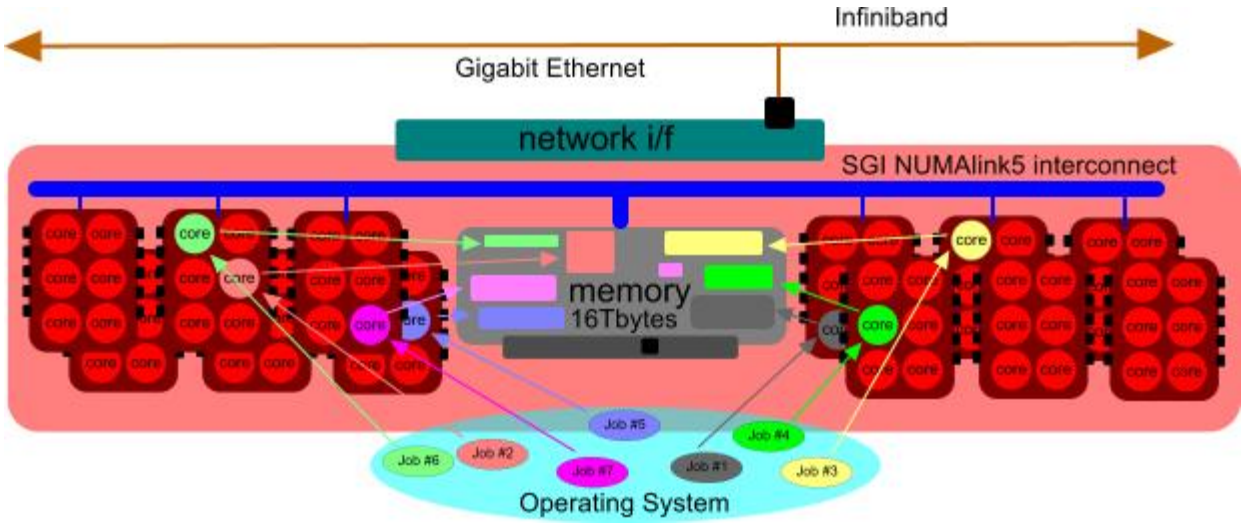
OOM killer



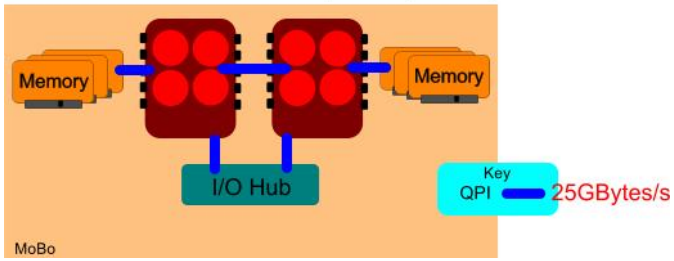
SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Shared Memory Model



Intel Quick Path Interconnect (Formally - FSB)



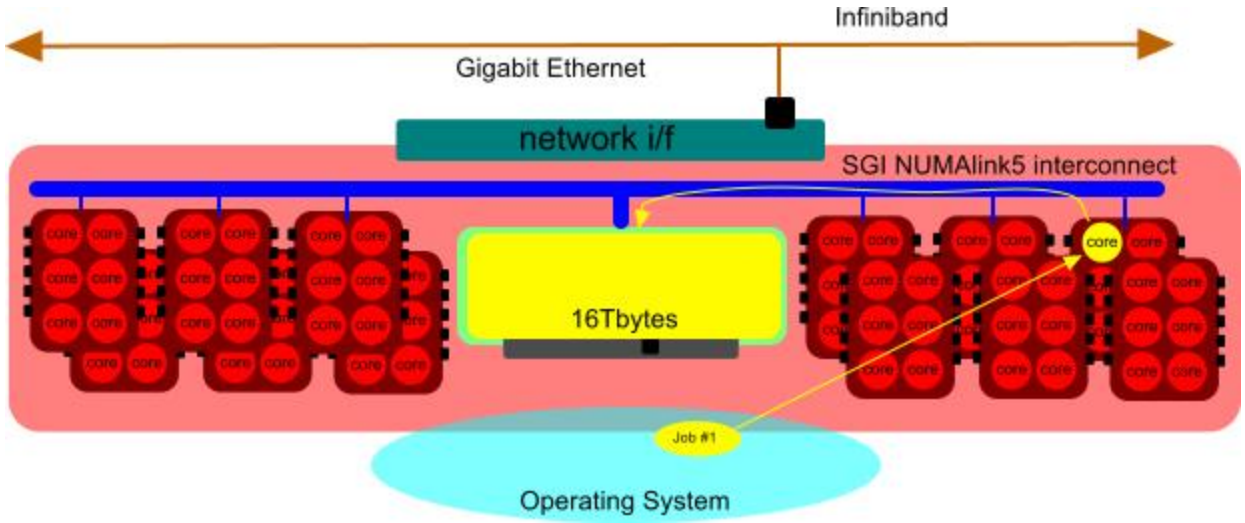
15GB/s peak

(3.2Gbytes)

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Shared Memory Models Can Support very large processes

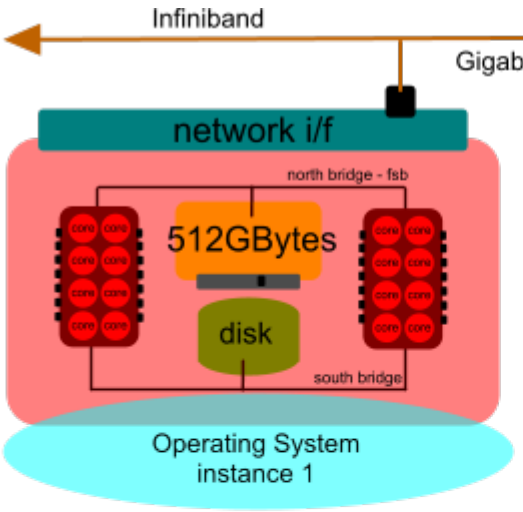


Cosmos
Universe

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Fat Node



Fat Node
16 Cores
512 GBytes
32 GBytes / Core

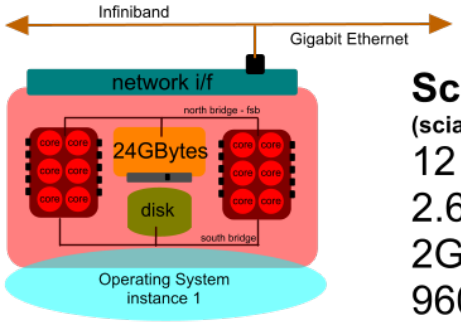
Can install 2TBytes
In stand node.

SCAMA Portsmouth

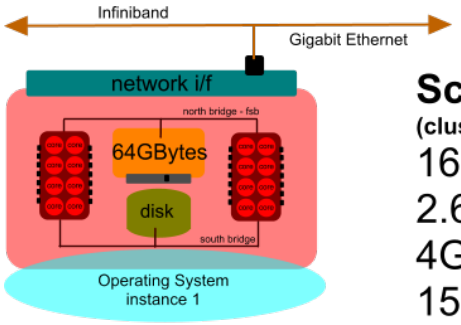
UNIVERSITY OF PORTSMOUTH

You need to know your HPC Building Block

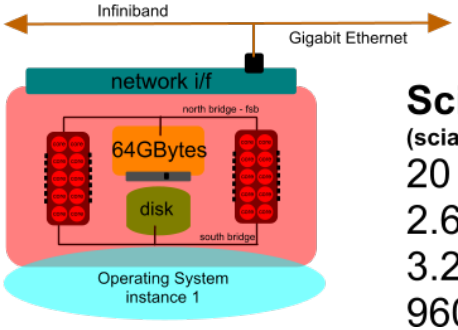
You need to know :-
Architecture – Intel/AMD
No. Cores
How much memory
Clock rate



Sciama 1 Node
(sciama1.q)
12 cores
2.6G Westmere
2GByte / Core
960 Cores



Sciama 2 Node
(cluster.q)
16 cores
2.6G Ivy Bridge
4GByte / Core
1536 Cores



Sciama 3 Node
(sciama3.q)
20 cores
2.6G Haswell
3.2GByte / Core
960 Cores

SCIAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Managing software / Modules

Running Programs on an HPC - 3 Scenarios

- Run an application already loaded in the environment.
  - Use selection tool (“module avail”) to check what is installed and select.
- Get a new application installed so you can run it.
  - Unless very specific its best to get administrators to install centrally.
  - Others may want – saves space, often quicker.
- Compile your own application from source code.
  - You will need to select the compiler(s) and libraries using the “module” command described. Request dependencies to be installed but not always possible.

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Many “modules” (probably) available

```
(Training)[burtong@login6(sciama) ~]$module avail

------------------------------------------------------------------------------------------------------ /opt/apps/modulefiles/core -------------------------------------------------------------------------------------------------------
bundles/2016.0/parallel     bundles/2016.0/serial       bundles/tools              services/torque/2.5.13      system/intel64(default)     system/sciama-2
bundles/2016.0/parallel_oss bundles/2016.0/serial_oss   services/maui/3.3.1        system/ia32                 system/sciama-1             system/sciama-3

---------------------------------------------------------------------------------------------------- /opt/apps/modulefiles/compilers ----------------------------------------------------------------------------------------------------
gnu_comp/4.3.5    gnu_comp/4.4.7    gnu_comp/4.9.4    intel_comp/2011.0 intel_comp/2013.2 intel_comp/2016.2 intel_comp/2018.2
gnu_comp/4.4.5    gnu_comp/4.8.4    gnu_comp/5.4.0    intel_comp/2013.1 intel_comp/2016.1 intel_comp/2017.2

---------------------------------------------------------------------------------------------------- /opt/apps/modulefiles/libraries ----------------------------------------------------------------------------------------------------
arpack-ng/3.5.0              cfitsio/3.37                glibc/2.17                  healpix/3.20                lapack/3.7.0                openmpi/1.8.2               Qhull/2015.2
atlas/3.10.2                 cfitsio/3.39                glibc/2.21                  healpix/3.31                libconfig/1.4.9             openmpi/1.8.4               qrupdate/1.1.2
atlas/3.10.3                 cfitsio/3.41                glibc/2.23                  intel_mpi/2017.2            libconfig/1.5               openmpi/2.0.1               root/6.10.8
bdwgc/7.2                    copter/0.8.7                glpk/4.61                   intel_mpi/4.1.0(default)    libpng/1.5.28               openmpi/2.0.2               ssh2/1.4.3
blas/3.7.0                   cuba/3.0                    graphicsmagic/1.3.19        intel_mpi/4.1.3             libz/1.2.11                 oracle-jdk/1.8.0_131        SuiteSparse/4.2.1
blitz/0.10                   cuba/4.2                    graphicsmagic/1.3.25        intel_mpi/5.1.2             llvm/4.0.1                  oracle-jre/1.8.0_131        SuiteSparse/4.3.1
blitz/0.9                    curl/7.54.0                 gsl/1.16                    intel_mpi/5.1.3             llvm/5.0.0                  pcre/8.35                   SuiteSparse/4.4.1
boost/1.53.0                 ffi/3.1                     gsl/2.3                     lalsuite/lal/6.18.0         log4cplus/1.1.1             pcre/8.40                   SuiteSparse/4.4.4
boost/1.59.0                 fftw/2.1.5                  hdf4/4.2.12                 lalsuite/lalburst/1.4.4     metaio/8.4.0                pcre2/10.23                 SuiteSparse/4.5.5
boost/1.62.0                 fftw/3.3.6                  hdf5/1.10.0-patch1          lalsuite/lalframe/1.4.3     mpich/3.2                   petsc/3.0.0                 swig/3.0.12
boost/1.63.0                 fftw_mpi/2.1.5              hdf5/1.10.1                 lalsuite/lalinspiral/1.7.7  netcdf/4.3.2                petsc/3.4.5                 unistring/0.9.3
boost_mpi/1.57.0             fftw_mpi/3.3.6              hdf5/1.6.10                 lalsuite/lalmetaio/1.3.1    openmpi/1.10.2              petsc/3.5.3(default)        wcs/4.16
boost_mpi/1.63.0             fgsl/0.9.4                  hdf5/1.8.17                 lalsuite/lalpulsar/1.16.0   openmpi/1.10.6              petsc/3.6.3                 wcs/4.23
bzip2/1.0.6                  fltk/1.3.4                  hdf5_mpi/1.10.0-patch1      lalsuite/lalsimulation/1.7.3 openmpi/1.4.3              plplot/5.10                 wcs/4.24
cblas/3.7.0                  frame/8.30                  hdf5_mpi/1.6.10             lalsuite/lalstochastic/1.1.20 openmpi/1.6.4             pqxx/4.0.1                  wxwidgets/3.0.1
cfitsio/2.5.10               glew/1.13                   hdf-java/3.2.1              lalsuite/lalxml/1.2.4       openmpi/1.8.1               qdbm/1.8.77                 xml2/2.9.7

-------------------------------------------------------------------------------------------------- /opt/apps/modulefiles/applications --------------------------------------------------------------------------------------------------
abaqus/6.10-2                bison/3.0/gcc-4.4.7         emacs/24.5                  gromacs/5.1.2               matlab/R2017a               Qt/5.3.1                    tex/2015
abaqus/6.12-1                blast/2.3.0                 enzo/2.1.1                  gtkplus/2.12.121            meqtrees/aug14              R/3.2.2                     tips/1.0
abaqus/6.14                  bowtie/1.1.2                enzo/2.4                    guile/1.8.8                 mercurial/aug14             R/3.4.1                     tkdiff/4.2
abaqus/6.9-2                 bowtie2/2.2.7               enzo/2.5                    gv/3.5.8                    minuit2/5.34.14             radmc-3d/0.27               tmux/1.6
anaconda/1.0                 bwa/0.7.12                  enzo/2.5-mc                 heasoft/6.16                montepython/2.2.2           rclone/1.38                 tmv/0.70
anaconda/2.2.0               casa/4.2.1                  enzo/2.5-mc-intel           help2man/1.47.5             mpss/3.0                    rna-seq/1.0                 topcat/4.2
anaconda/2.4.0               casa-core-rest/1.7.0/gcc-4.4.7 ez/0.3.1                 htslib/1.3                  multinest/3.6               rockstar/0.99.9-RC3+        tophat/2.1
anaconda/2.4.1               cgal/4.8                    ffmpeg/3.1                  idl/8.0                     music/oct15                 rrdtool/1.6.0               vcftools/0.1.13
anaconda3/2.1.0              class/2.6.1                 fluidstructures/17.1        idl/8.3                     nlopt/2.4.2                 samtools/1.3                visit/2.10.2
anaconda3/2.5.0              cloudy/13.04                freetypes/2.6.2             idl/8.5                     octave/4.2.1                scamp/2.0.4                 vmd/1.9.2
anaconda3/2.5.0-2            cmake/2.8.3/gcc-4.4.7       fv/5.3                      igv/2.3.68                  OpenCV/2.4.10               sextractor/2.19.5           vtk/6.1.0
anaconda-atlas/2.2.0         cmake/3.0.0/gcc-4.4.7       gadgetviewer/1.0.4          ImageMagick/6.9.2-8         parallel/1.0                SNANA/v10_36e               vtk/6.1.0-a
anaconda-intel/2.2.0         cmake/3.10.0                gadgetviewer/1.0.7          iraf/2.16.1                 paraview/5.0.1              SNANA/v10_40b               vtk/6.1.0-mod
ant/1.10.1                   cmake/3.10.1                galsim/1.3                  java/1.8.0_131              parosol/jun16               splotch/4.4                 wcstools/3.9.0
appspack/5.0.1-C3            cmake/3.6.2                 gdl/0.9.4                   java-ide/apr15              perl/5.26                   splotch/6.0                 weka/3.6.11
autoconf/2.69                cpython/2.7.13              ghostscript/9.14            knime/3.2.1                 pgplot/5.2                  splotch/6.0a                xv/3.10a
autogen/5.13                 cpython/3.6.1               git/2.9.4                   libtool/2.4.6               pgplot/5.2-gcc              starccm/10.06.010           xvfb/2.4
autogen/5.18.1               crime/1.3                   gnuplot/4.4.3               lofar/oct13                 pgplot/5.2-sept15           starccm/10.06.010-r8        xz/5.2.3
automake/1.13.1              cufflinks/2.2.1             gnuplot/4.6.6               lofar/oct13-pybdsm          pigz/2.3.3                  starccm/9.06.011            yasm/1.3.0
automake/1.15                doxygen/1.8.13              go/1.4                      m4/1.4.17                   plc/2.0                     swarp/2.38.0
axesim/apr15                 dtfe/1.1.1                  go/1.9.1                    make/4.2                    pybdsm/1.8.2                syn++/0.98
bcftools/1.3                 ecl/jul17                   grace/5.1.25                mathematica/10.4.0          Qt/4.5.3                    tcllib/1.17
binutils/2.28                elements/3.8                gromacs/3.3                 mathematica/11.0.0          Qt/4.8.4                    tcllib/1.18
```

SCIAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Application, Libraries and Compilers

- Use the “module” command (>man module)
- “> module avail” to see available modules.
- “>module add <mod path>”
- “>module delete <mod path>”
- “>module list”
- “>module initadd <mod path>”
- $HOME/.modules

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Ex. Selecting a different version of Python

1.) Current Python version

```
(Training)[burtong@login6(sciama) ~]$
(Training)[burtong@login6(sciama) ~]$python -V
Python 2.6.6
(Training)[burtong@login6(sciama) ~]$module avail anaconda

------------------------------------------------------------------------ /opt/apps/modulefiles/applications -
anaconda/1.0         anaconda/2.4.0       anaconda3/2.1.0      anaconda3/2.5.0-2    anaconda-intel/2.2.0
anaconda/2.2.0       anaconda/2.4.1       anaconda3/2.5.0      anaconda-atlas/2.2.0
(Training)[burtong@login6(sciama) ~]$
(Training)[burtong@login6(sciama) ~]$module add anaconda3/2.5.0
(Training)[burtong@login6(sciama) ~]$
(Training)[burtong@login6(sciama) ~]$python -V
Python 3.5.2 :: Anaconda custom (64-bit)
(Training)[burtong@login6(sciama) ~]$
(Training)[burtong@login6(sciama) ~]$module list
Currently Loaded Modulefiles:
  1) anaconda3/2.5.0
(Training)[burtong@login6(sciama) ~]$
```

2.) Check available versions

3.) Select required version

4.) New version

5.) Modules loaded

SCIAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Environmental Variables

- Env's customise your environment.
- Run the command "env" .
- $PATH is important – search path for commands.
- "echo $PATH"
- "export PATH="/usr/bin:/usr/local/bin/:/bin"

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

“module show”

```
[burtong@login8(sciama) ~]$ module show libs/wcs/4.16/gcc-4.4.7
-------------------------------------------------------------------
/opt/apps/modules//libs/wcs/4.16/gcc-4.4.7:

module-whatis   loads the necessary `wcs-4.16' library paths
setenv          WCSINCLUDE /opt/apps/libs/wcs/4.16/gcc-4.4.7/include
setenv          WCSLIB /opt/apps/libs/wcs/4.16/gcc-4.4.7/lib
setenv          WCSBIN /opt/apps/libs/wcs/4.16/gcc-4.4.7/bin
prepend-path    PATH /opt/apps/libs/wcs/4.16/gcc-4.4.7/bin
prepend-path    LD_LIBRARY_PATH /opt/apps/libs/wcs/4.16/gcc-4.4.7/lib/
prepend-path    MANPATH /opt/apps/libs/wcs/4.16/gcc-4.4.7/share/man/
prepend-path    C_INCLUDE_PATH /opt/apps/libs/wcs/4.16/gcc-4.4.7/include/
prepend-path    CPLUS_INCLUDE_PATH /opt/apps/libs/wcs/4.16/gcc-4.4.7/include/
prepend-path    --delim  CPPFLAGS  -I/opt/apps/libs/wcs/4.16/gcc-4.4.7/include
prepend-path    --delim  CFLAGS  -I/opt/apps/libs/wcs/4.16/gcc-4.4.7/include
prepend-path    --delim  LDFLAGS  -L/opt/apps/libs/wcs/4.16/gcc-4.4.7/lib/
-------------------------------------------------------------------
```

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Prepending $PATH

[burtong@login8(sciama) ~]$ echo $PATH
/usr/lib64/qt-3.3/bin:/usr/local/sbin:/usr/local/bin:/usr/sbin:
/usr/bin:/sbin:/bin:/bin:/usr/bin:/usr/X11R6/bin:/users/
burtong/bin:/users/burtong/bin

[burtong@login8(sciama) ~]$ module add libs/wcs/4.16/gcc-4.4.7

[burtong@login8(sciama) ~]$ echo $PATH
/opt/apps/libs/wcs/4.16/gcc-4.4.7/bin:/usr/lib64/qt-3.3/bin:
/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/bin:
/usr/bin:/usr/X11R6/bin:/users/burtong/bin:/users/burtong/bin

SCIAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Compilers, libraries and bundles

1.) Select the system (hardware specific)

```
services/torque/2.5.13      system/intel64(default)     system/sciama-2
system/ia32                 system/sciama-1            system/sciama-3
```

```
------------------------------------------------------------------------------ /opt/apps/modulefiles/compilers ------------------
gnu_comp/4.3.5    gnu_comp/4.4.7    gnu_comp/4.9.4    intel_comp/2011.0 intel_comp/2013.2 intel_comp/2016.2 intel_comp/2018.2
gnu_comp/4.4.5    gnu_comp/4.8.4    gnu_comp/5.4.0    intel_comp/2013.1 intel_comp/2016.1 intel_comp/2017.2
```

2.) Select Compiler

```
intel_comp/2016.2 \
intel_mpi/5.1.3 \
fftw_mpi/3.3.6 \
hdf5_mpi/1.10.0-patch1 \
gsl/2.3 \
boost_mpi/1.63.0 \
```

3.) Select the libraries

4.) Alternatively select Bundle

```
------------------------------------------------------------------------------
bundles/2016.0/parallel     bundles/2016.0/serial       bundles/tools
bundles/2016.0/parallel_oss bundles/2016.0/serial_oss   services/maui/3.3.1
```

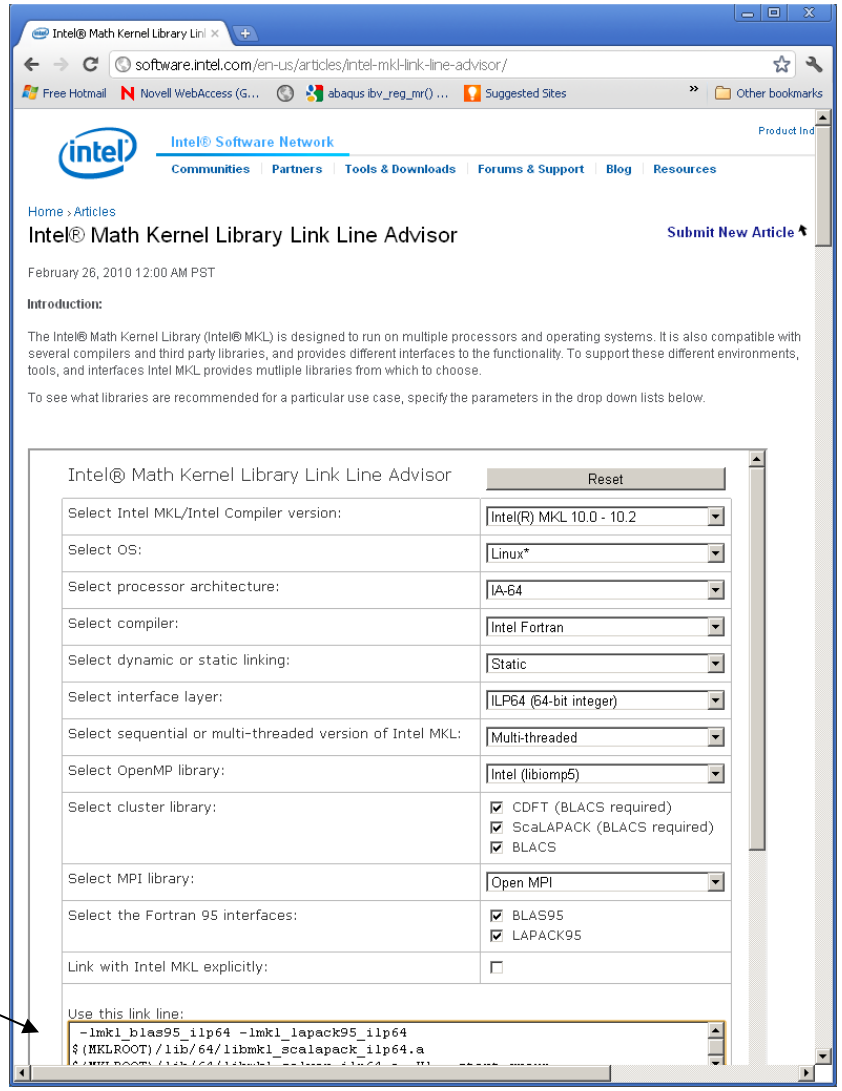
5.) Compatible libaries selected

```
[jschewts@login7(sciama) ~]$ module add system/intel64
[jschewts@login7(sciama) ~]$ module add bundles/2016.0/parallel_oss
[jschewts@login7(sciama) ~]$ module list
Currently Loaded Modulefiles:
  1) services/torque/2.5.13       3) system/intel64             5) openmpi/2.0.2              7) hdf5_mpi/1.10.0-patch1      9) boost_mpi/1.63.0
  2) services/maui/3.3.1          4) gnu_comp/5.4.0             6) fftw_mpi/3.3.6             8) gsl/2.3                    10) bundles/2016.0/parallel_oss
[jschewts@login7(sciama) ~]$
```

SCIAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Compile Options



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Schedulers and Resource managers

Creating a job

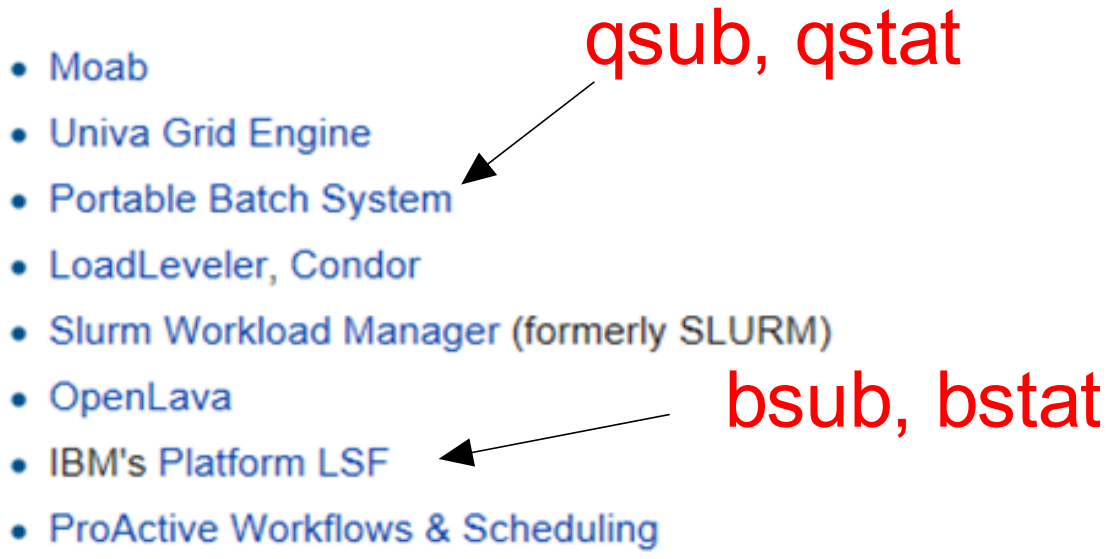
- Once an application has been selected or compiled it can be run.
- This is done by setting up a “job”.
- A jobs has two components:-
  - The program
  - The “meta data”
- A job script is needed (bash script)
- We need to submit the job

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Schedulers and Resource Managers (The Queueing System)

- We use Torque and Maui based on PBS (Portable Batch System)
- Mainly because its free (open source).
- We only use basic functionality.



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Job Script Example

```
#!/bin/bash
#PBS -l nodes=10:ppn=12
#PBS -l walltime=6:00:00
#PBS -N gadget
#PBS -o out_gadget_blind8001.o$PBS_JOBID
#PBS -e err_gadget_blind8001.e$PBS_JOBID
#PBS -m abe
#PBS -M gary.burton@port.ac.uk
#PBS -V
#PBS -q cluster.q
```

Queuing System Directives

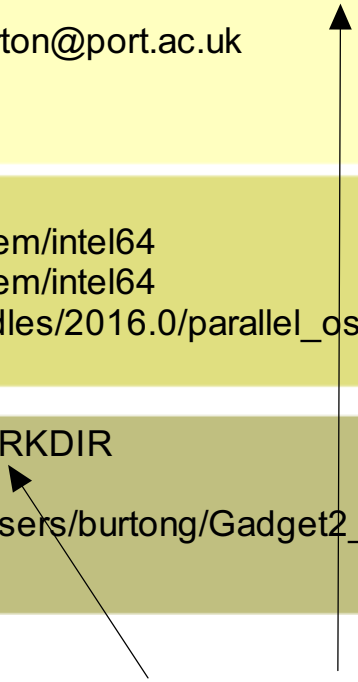
```
module purge
module load system/intel64
module load system/intel64
module add  bundles/2016.0/parallel_oss
```

Module selection

```
cd $PBS_O_WORKDIR

mpirun -np 780 /users/burtong/Gadget2_grid2400  /users/burtong/Gadget_Oriana_Bli
nd_ir8001.param
```

Program to Run



Internal Variables

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Running Jobs

- You should never run any large processes on a login server. Login servers are shared by others so any intense processing will slow down the responsiveness others experience.
- All major applications / compilations should be done on a compute node
- There are two ways to do this both using the “qsub” command to submit a job to a queue :-
  - Create an interactive shell on a compute node:-

    qsub –IX # I=Interactive X=allow X server windows

    This will give you a command prompt in the current shell.
  - Run a batch job on a compute node(s) :-

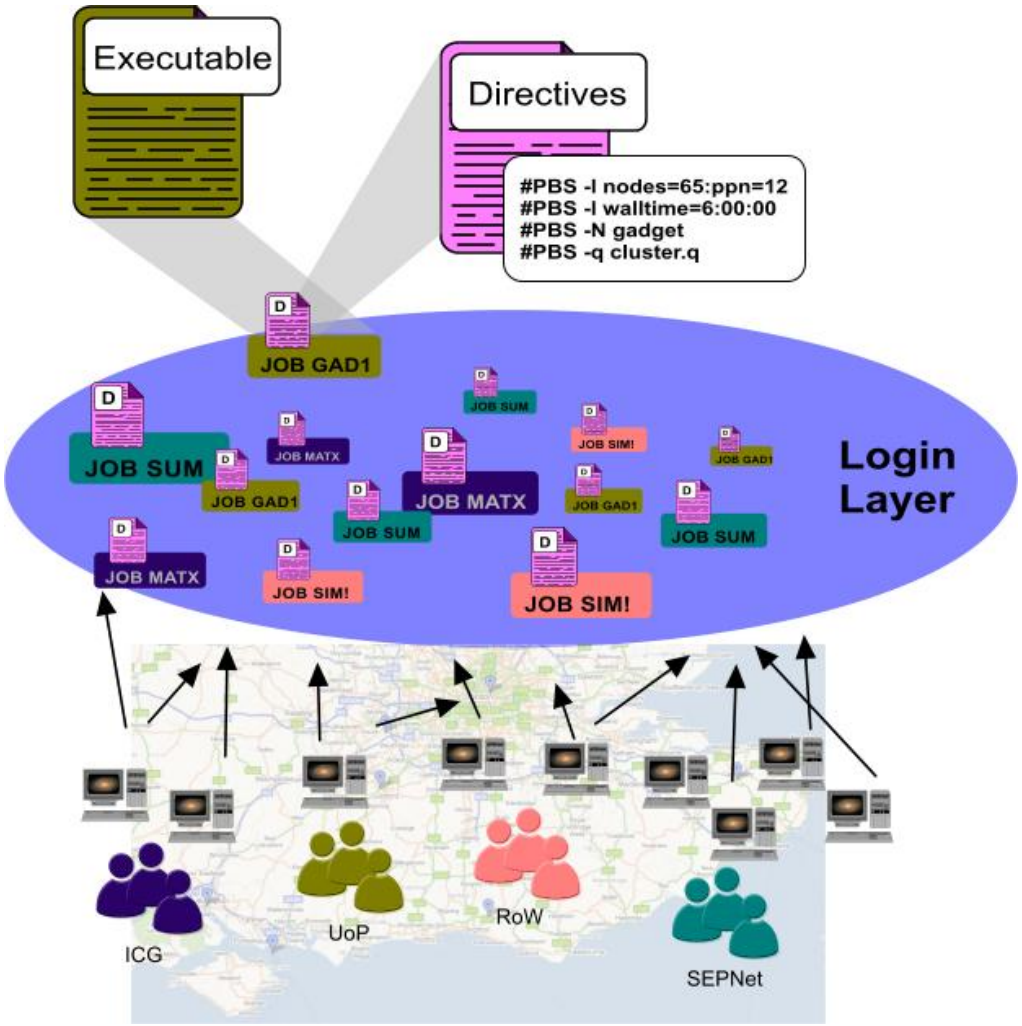
    qsub <job-script>

    This will start / queue a job according to the directives specified in the job script.

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

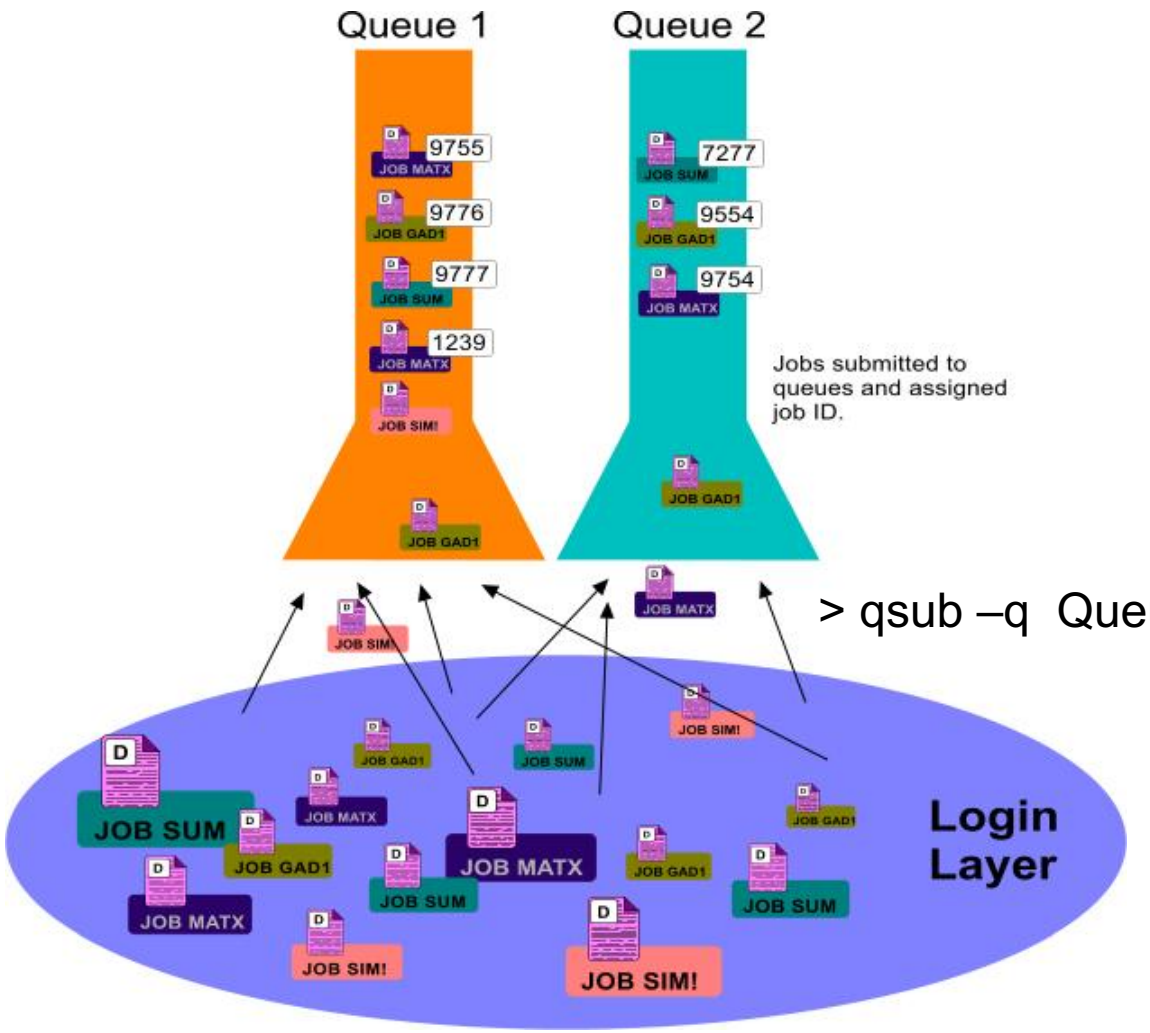
Executable and Jobscript setup in Login Layer



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

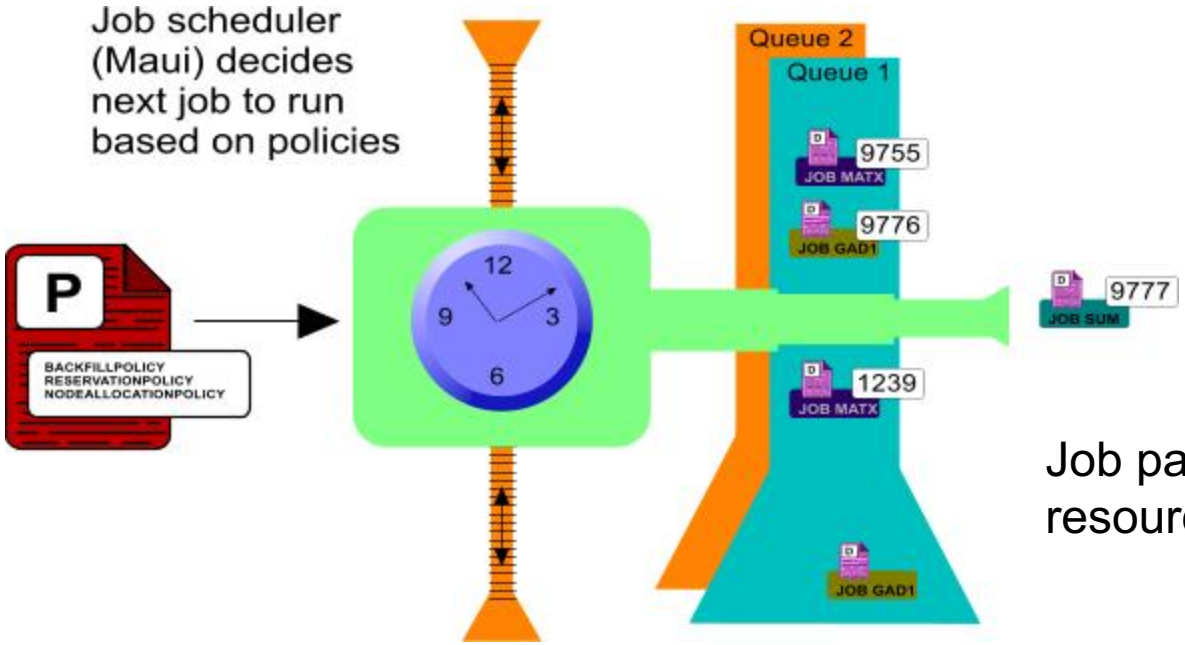
Jobs submitted to the queues



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

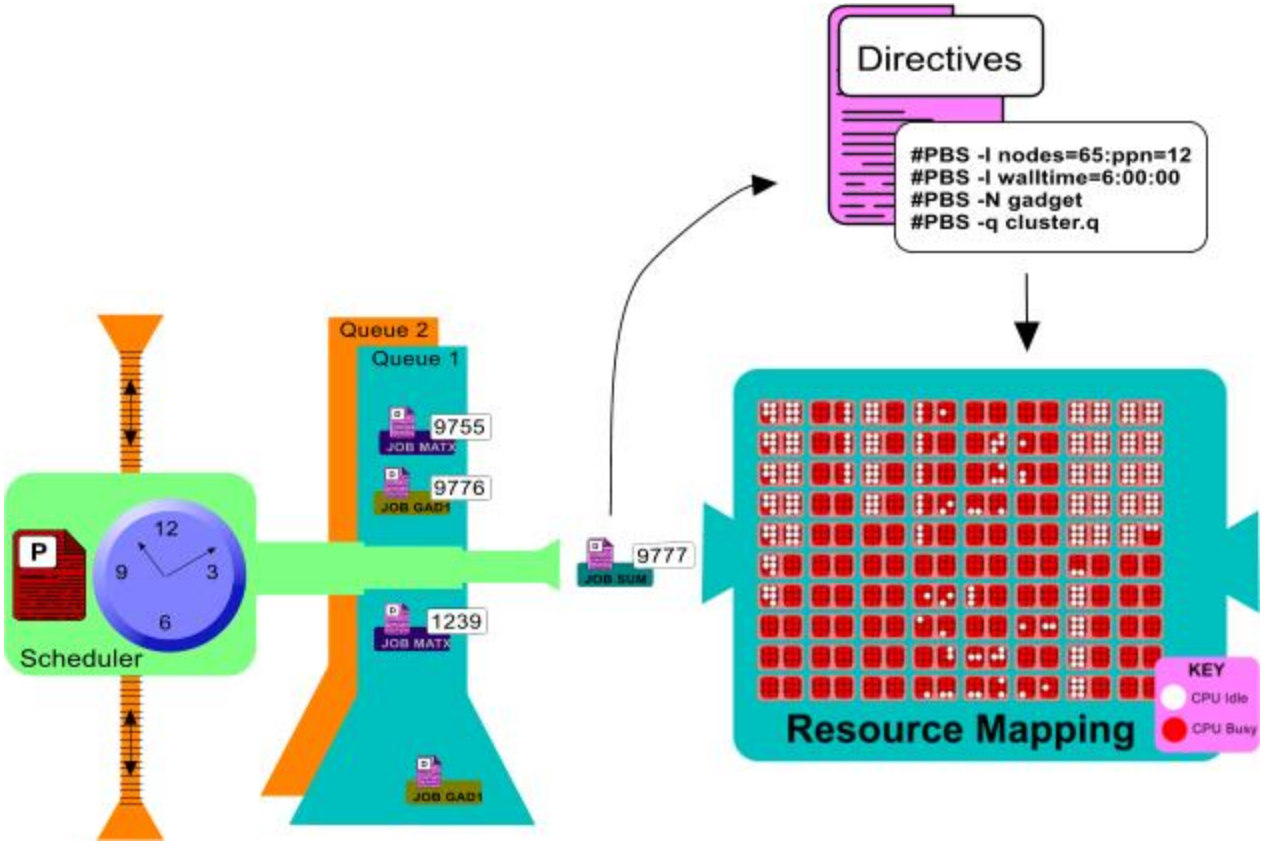
Scheduler prioritises and deems a job ready to run.



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

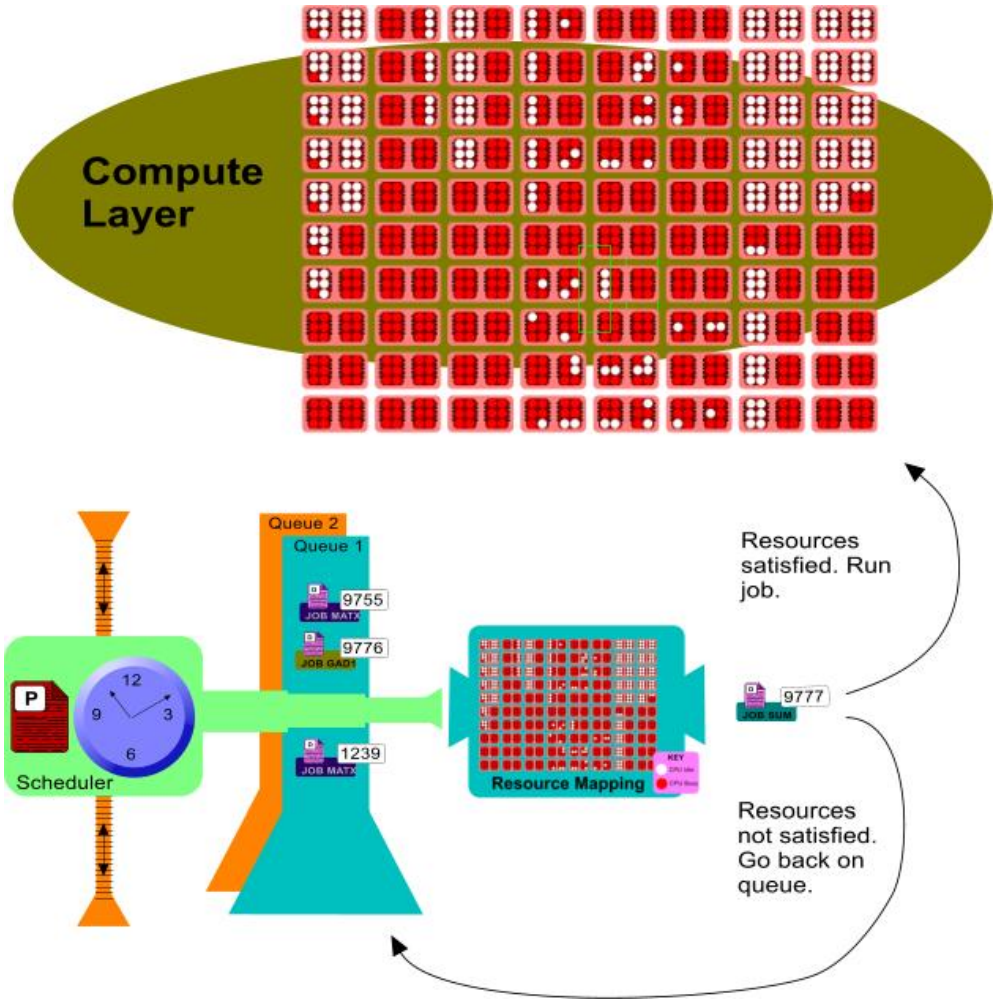
Resource manager (Torque) checks for available resources.



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Job either runs in the compute pool or returns to queue



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Common Queue Commands

- Qsub, Qstat, Qdel

-tn1ru

```
[burtong@login8(sciama) ~]$ qstat                                cksim

headnode1.prv.sciama.cluster:
                                                                  Req'd    Req'd       Elap
Job ID                  Username    Queue    Jobname         SessID  NDS   TSK   Memory  Time    S  Time
----------------------- ----------- -------- --------------- ------ ----- ------ ------ --------- - ---------
2839001.headnode1.prv.  cksim       cluster. job2_1           67977     1    16    --  400:00:00 R 366:25:56
node105/15+node105/14+
2862199.headnode1.prv.  cksim       cluster. job3_1            2734     1    16    --  400:00:00 R 279:32:25
node158/15+node158/14+
2868380.headnode1.prv.  cksim       cluster. job3             75690     1    16    --  400:00:00 R 208:20:13
node193/15+node193/14+
2868381.headnode1.prv.  cksim       cluster. job4             40446     1    16    --  400:00:00 R 208:20:13
node159/15+node159/14+
```

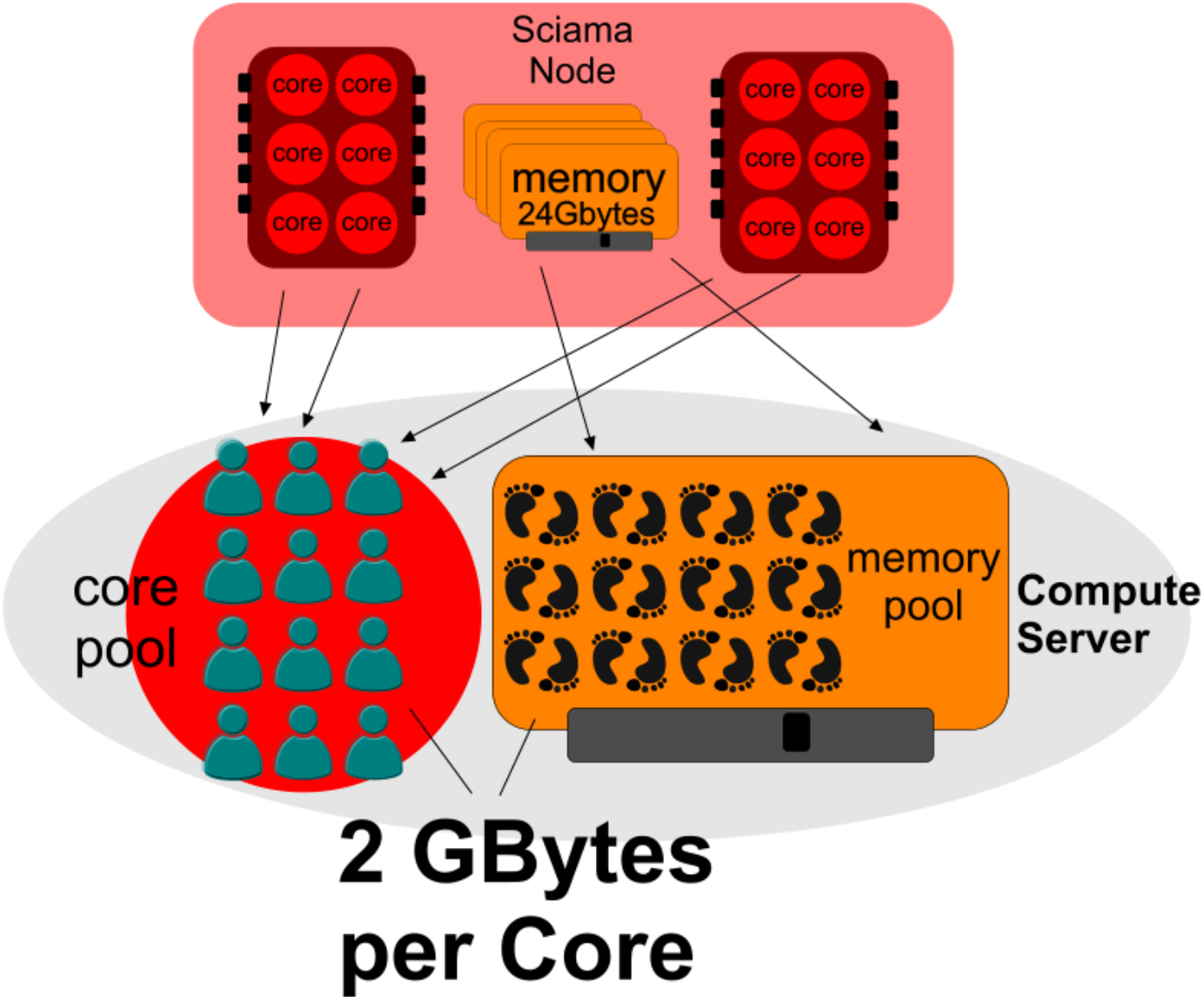
SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Being a Good HPC Citizen

SCAMA Portsmouth

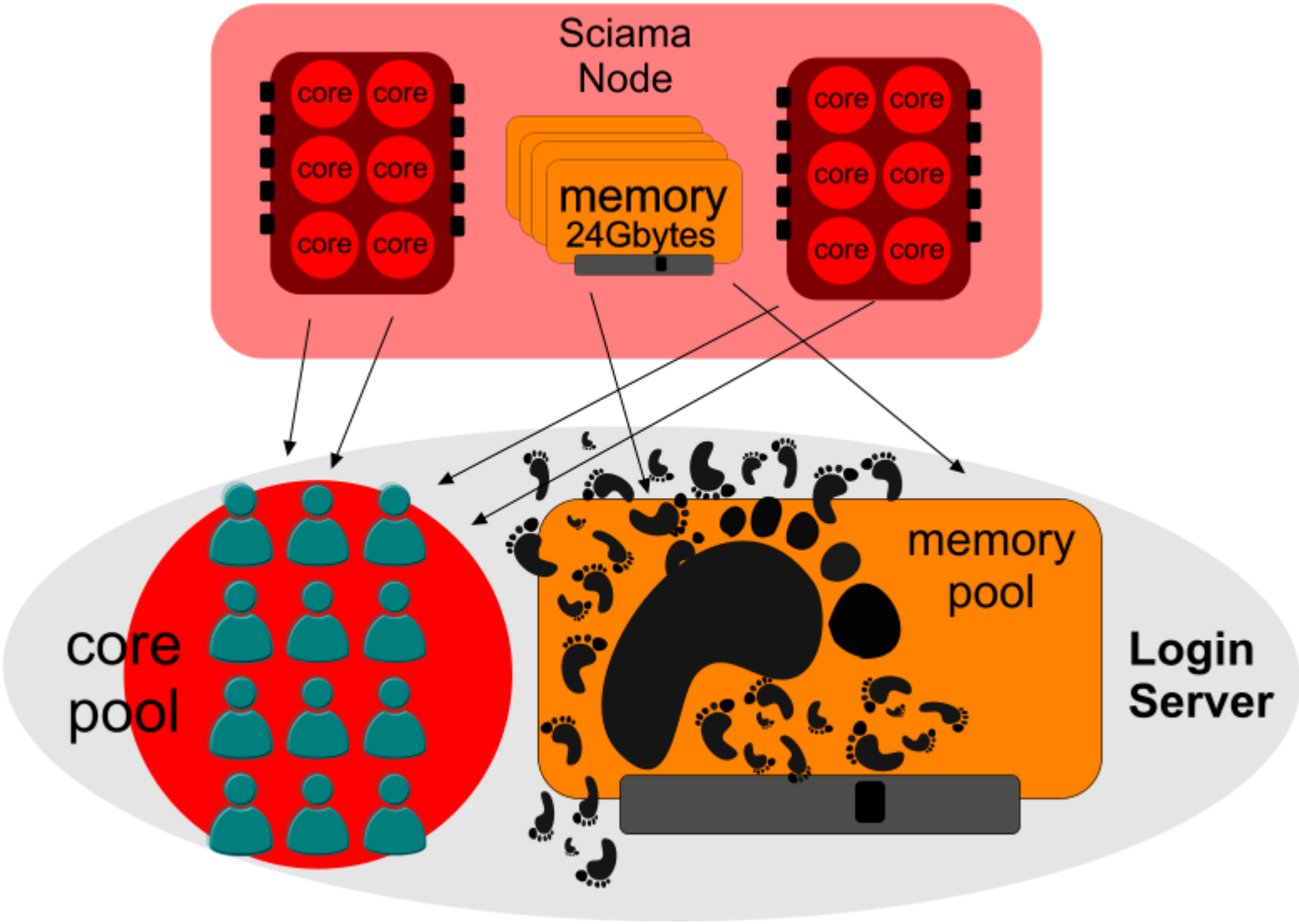
UNIVERSITY OF PORTSMOUTH



2 GBytes per Core

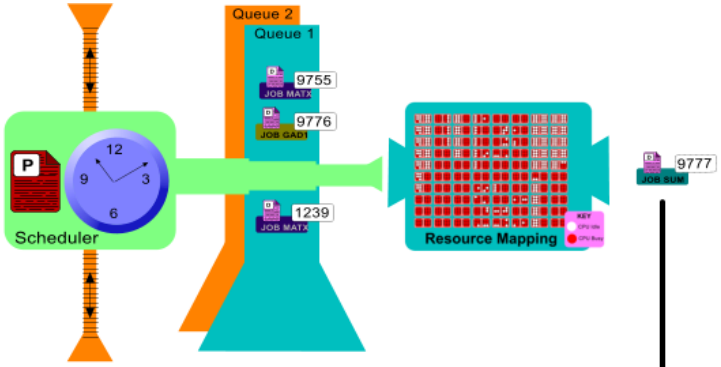
SCIAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

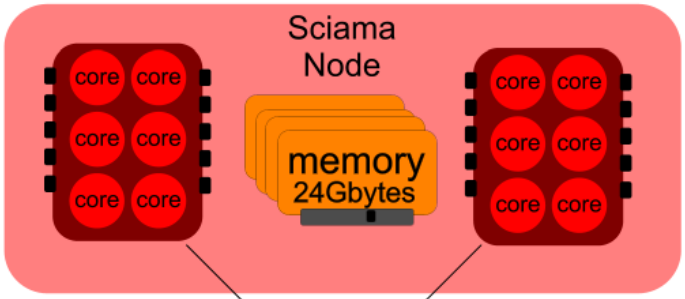


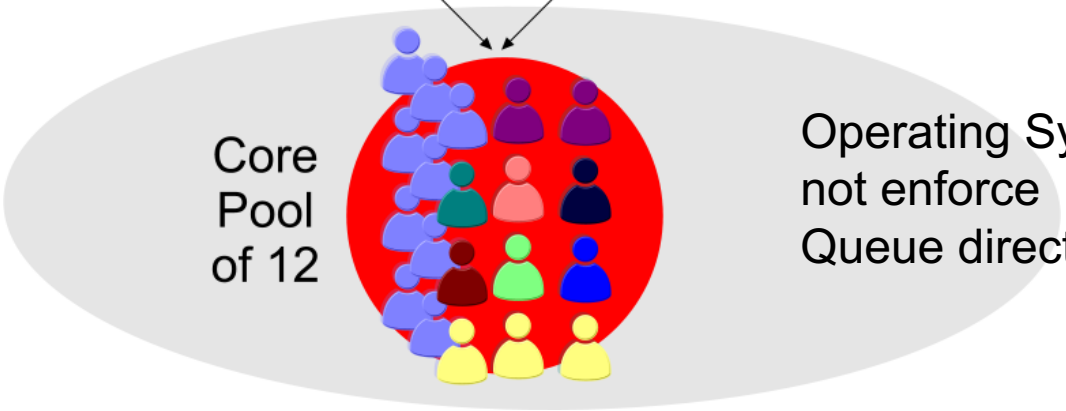
SCIAMA Portsmouth

UNIVERSITY OF PORTSMOUTH



Beware Multi-threaded Programs





Operating System does not enforce Queue directives.

SCIAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Sciama has few enforced limits

- Don’t run big jobs on the login nodes.
- Stick to the memory / core limits (over allocate).
- Restrict Openmp cores (omp_num_cores)
- If unsure allocated the whole node to check the resources used (-l nodes=1:ppn=16)
- If you require more than 100GBytes storage then request a “Lustre” (project) area.
- Don’t install packages without checking.
- If you sense something is wrong then tell us.

SCIAMA Portsmouth

UNIVERSITY OF PORTSMOUTH