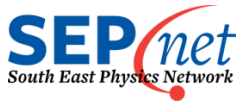
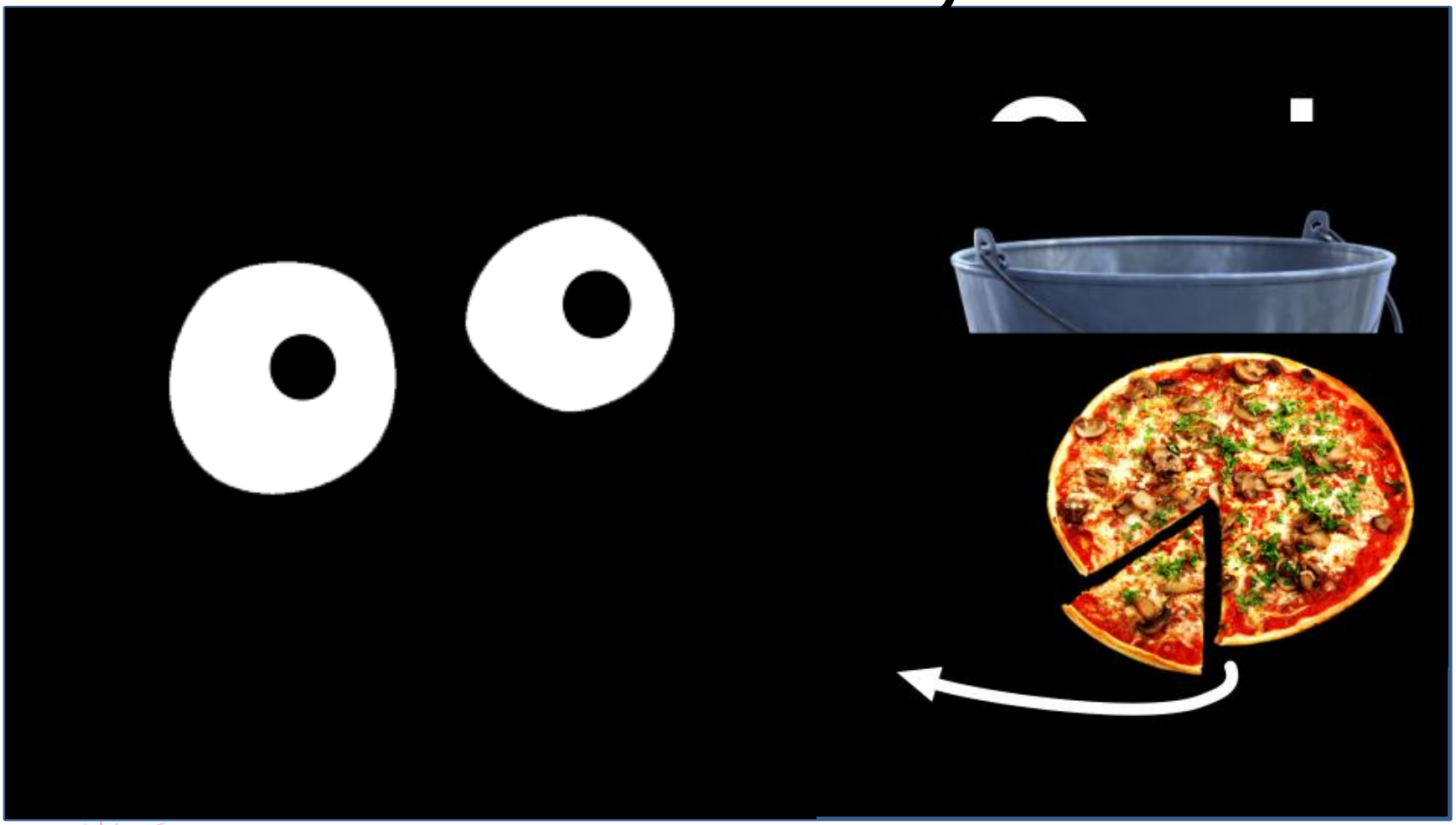
How to build a supercomputer

ON THE CHEAP

G Burton – ICG – Jan-14 – v1.2



Server Guy



SCAMA Portsmouth ICG Portsmouth ALCES SOFTWARE University of Portsmouth SEPnet South East Physics Network

HPC Facts

- Chinesse Tianhe-2 (MilkyWay -2) - China's National University of Defense Technology
  Number 1 in top 500 with 3.12 million Cores(processors).
- 34 petaFLOPS (Floating Point Operations / Second). Sciama 10 teraFLOPS.
- In Top 500 - USA=213, China=37, UK=29
- 75% of Top 500 use Intel processors.
- Race for first ExaFlop (exescale) .

Computer performance

Name	FLOPS
yottaFLOPS	10^{24}
zettaFLOPS	10^{21}
exaFLOPS	10^{18}
petaFLOPS	10^{15}
teraFLOPS	10^{12}
gigaFLOPS	10^{9}
megaFLOPS	10^{6}
kiloFLOPS	10^{3}

SCIAMA Portsmouth ICG Portsmouth ALCES SOFTWARE University of Portsmouth SEPnet South East Physics Network

To illustrate how we use an HPC we will set ourselves a problem …..

What is the answer to the ultimate question of Life, the Universe , and Everything ?

C: Hitch Hikers Guide to the Galaxy

SCAMA Portsmouth

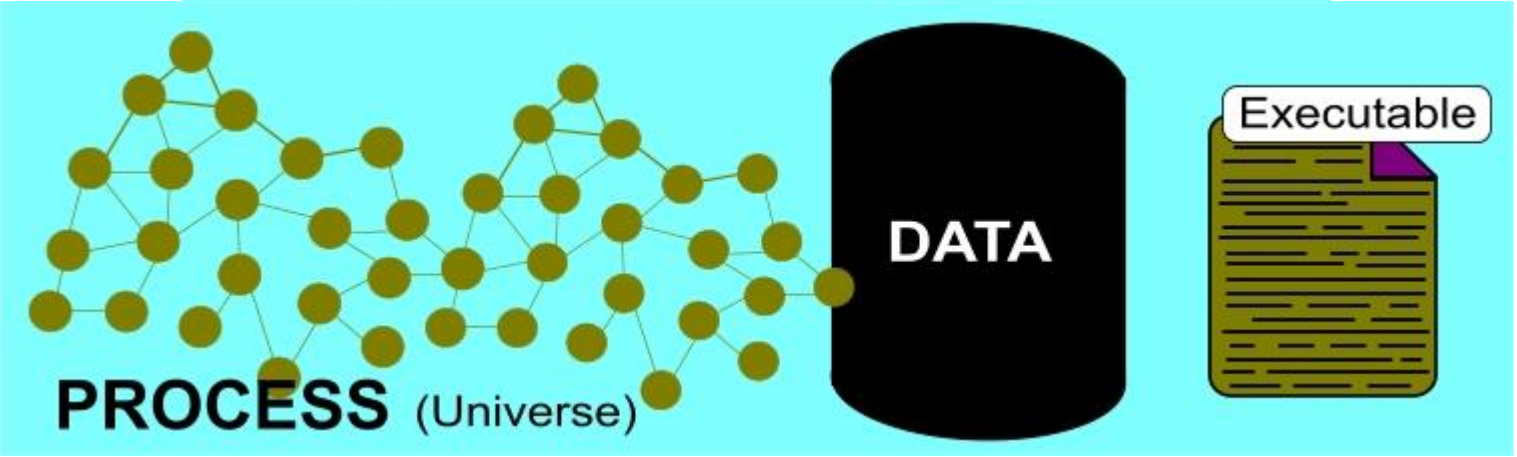
University of Portsmouth

ALCES SOFTWARE

ICG Portsmouth

SEPnet South East Physics Network

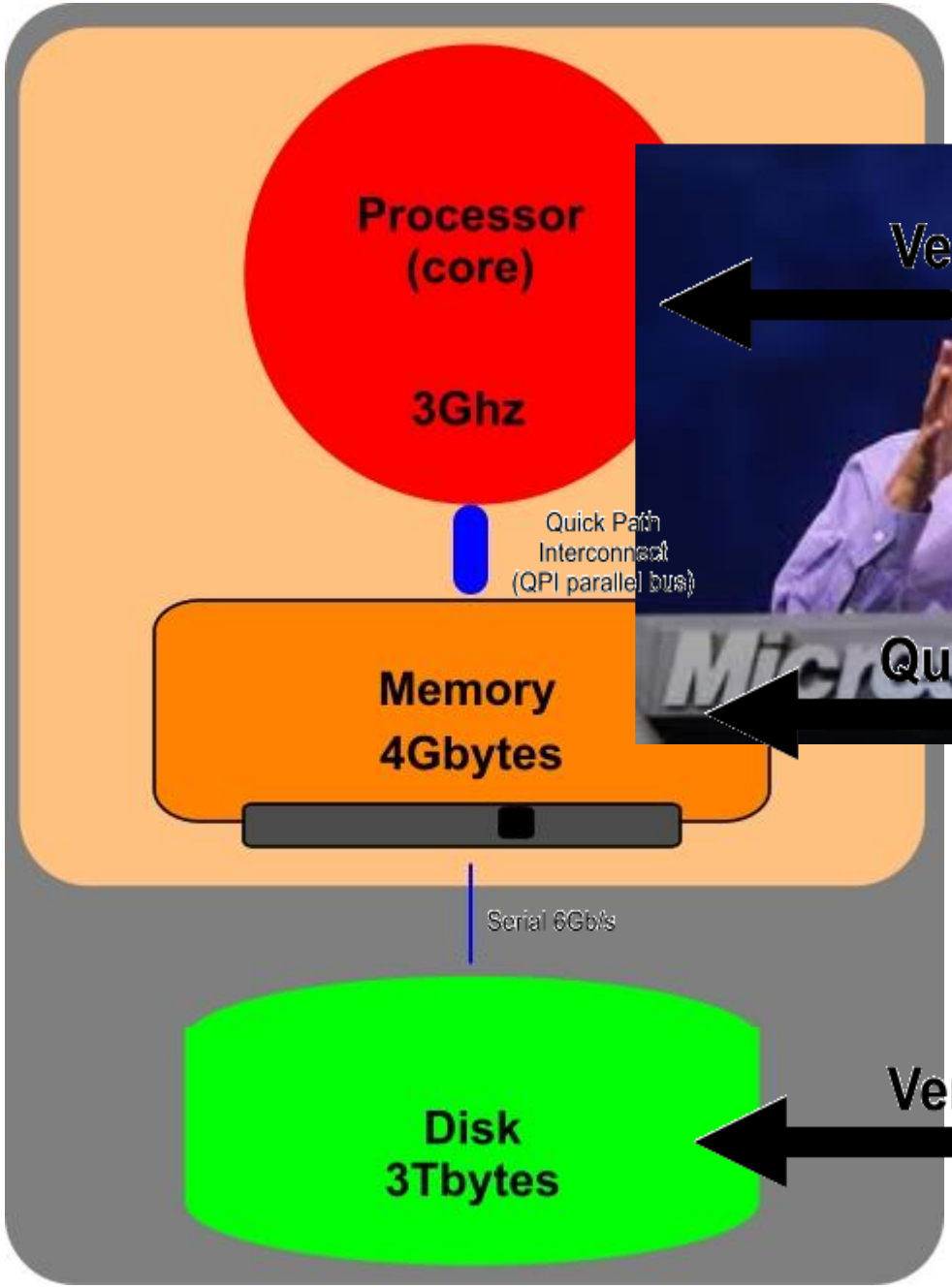
Start with small data, test with larger and then run with full…..



What is the answer to the ultimate question of Life, the Universe, and Everything ?

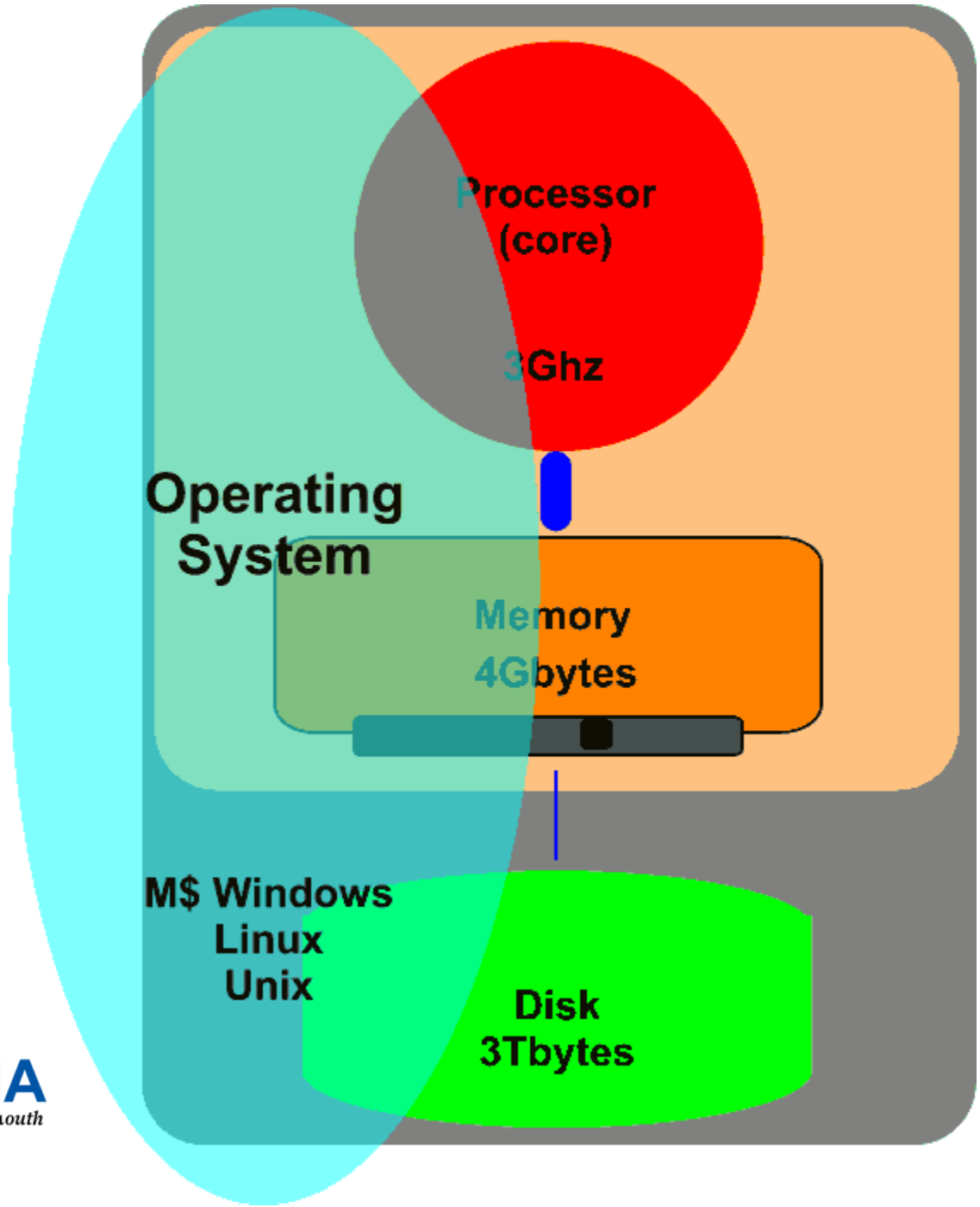
Solar System

Back to Basics



SCAMA Portsmouth

SEPnet South East Physics Network

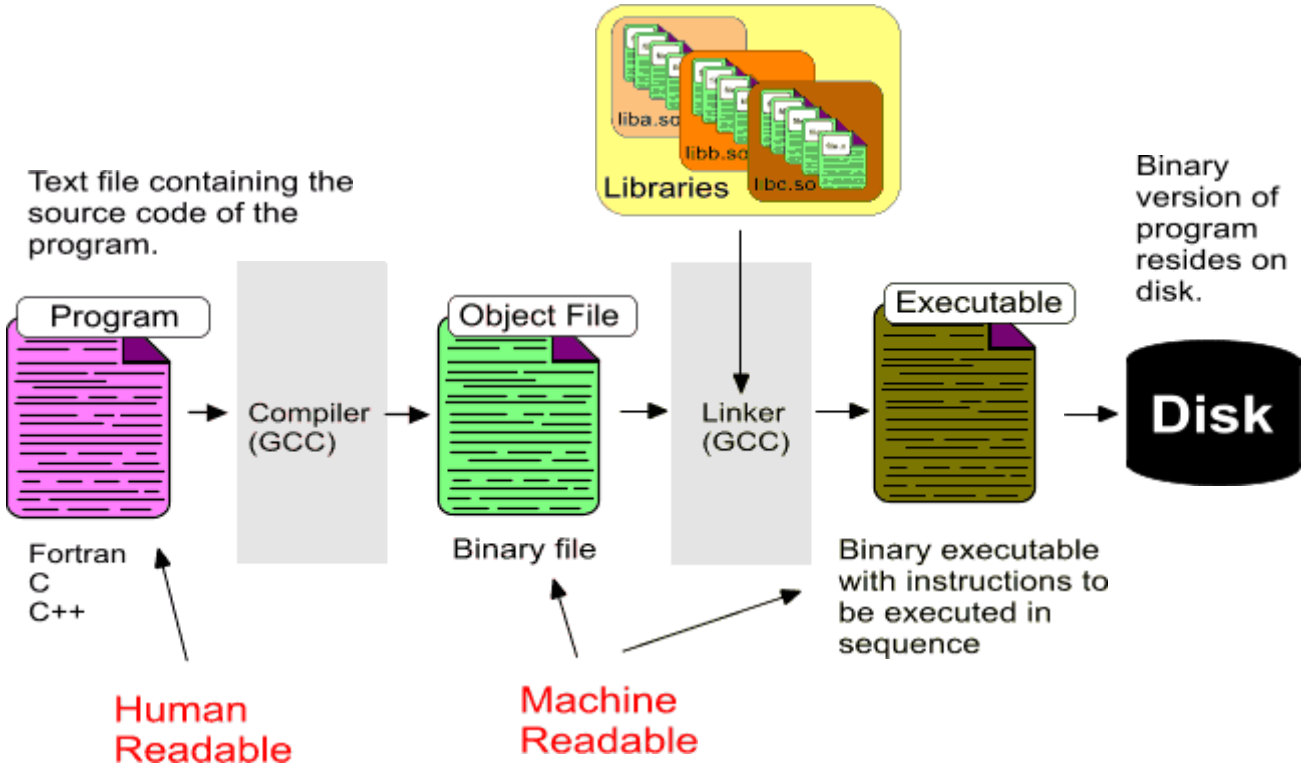


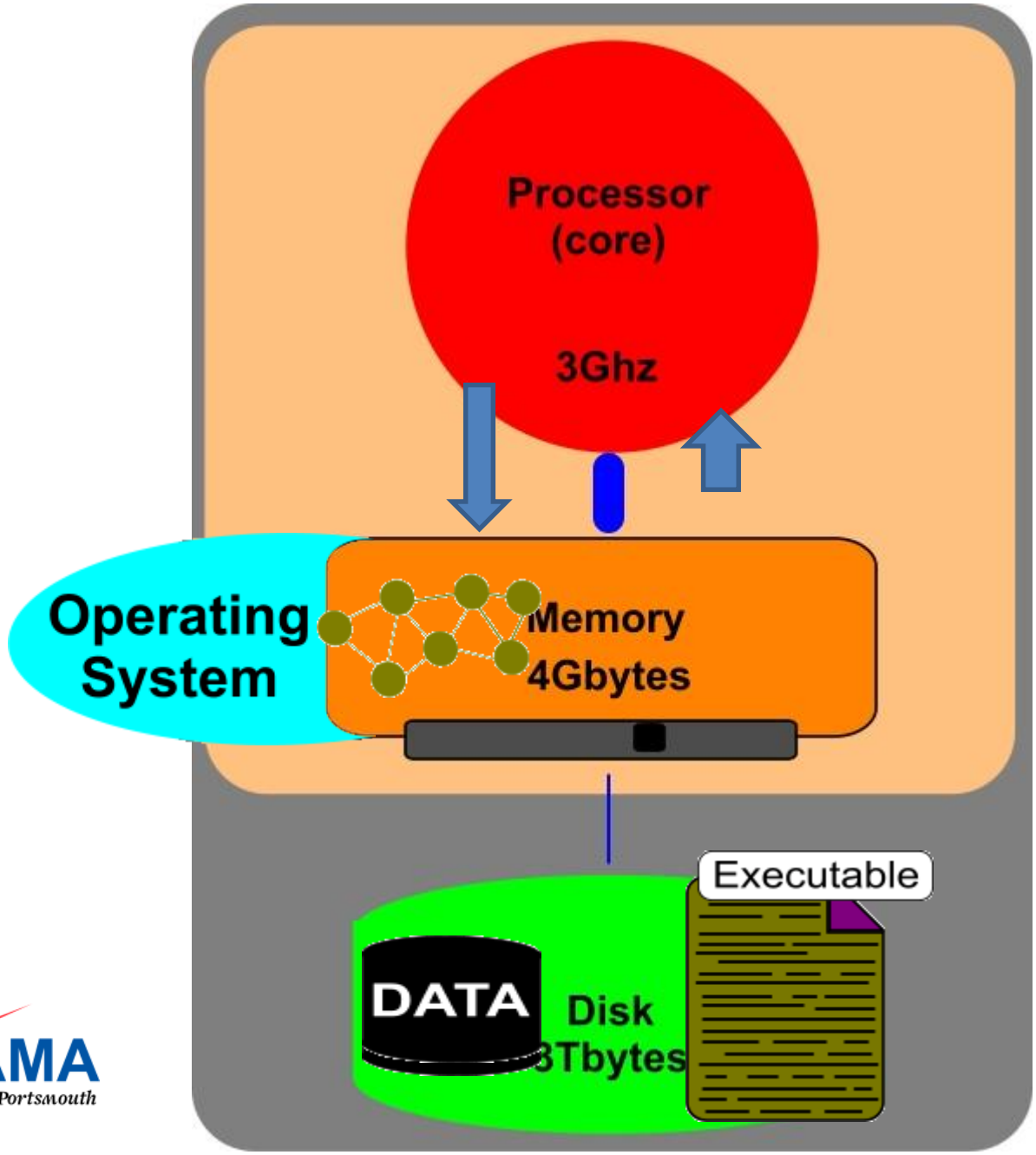
Back to Basics

SCAMA Portsmouth

SEPnet South East Physics Network

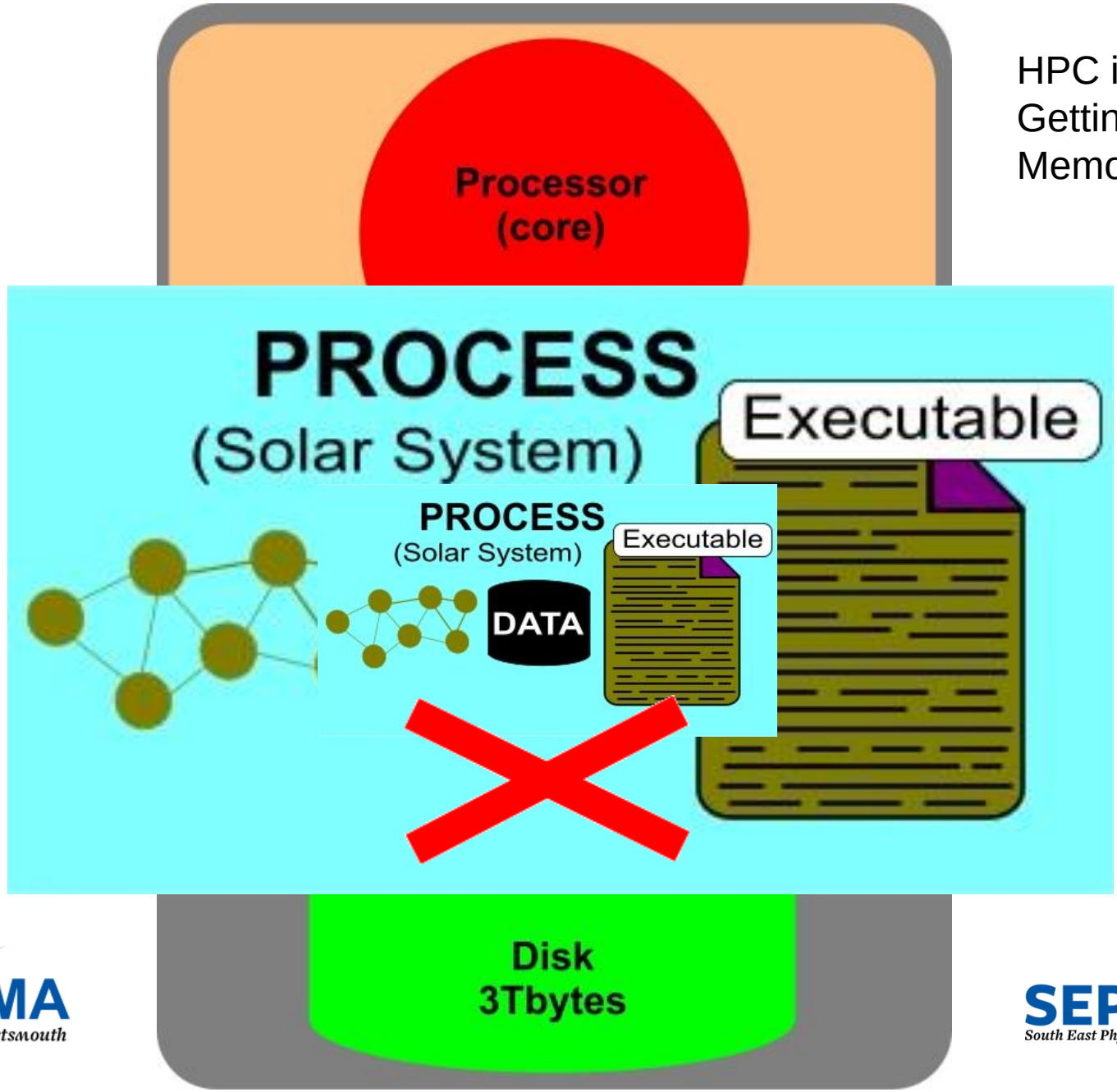
In order to solve our problem we need a “Program” to run.





SCAMA Portsmouth

SEPnet South East Physics Network

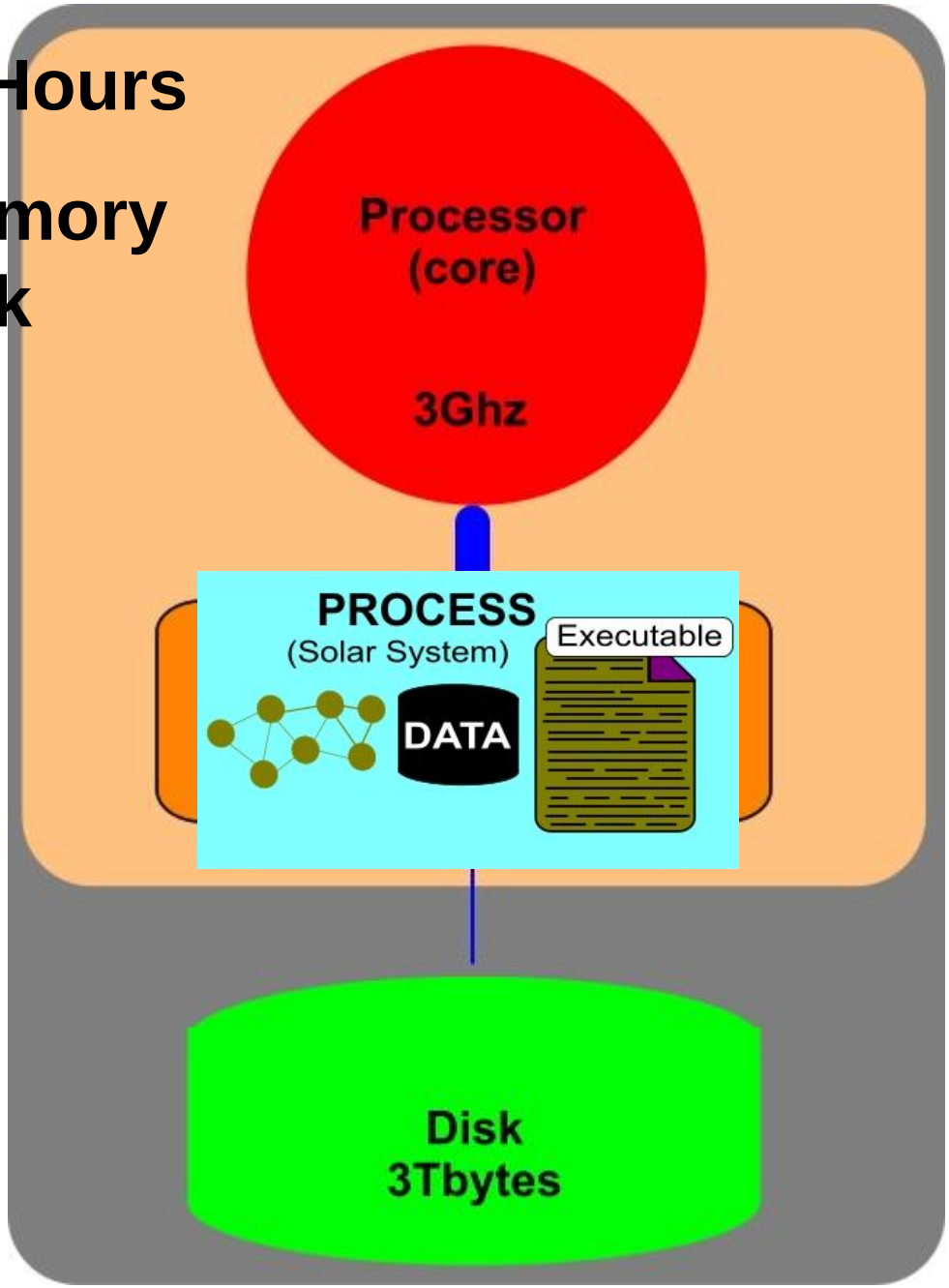


HPC is all about
Getting access to
Memory.

SCAMA Portsmouth

SEPnet South East Physics Network

10,000 CPU Hours
(Check Point Restarts)
3 Gbytes Memory
2 Tbytes Disk



Answer

63

SCAMA Portsmouth

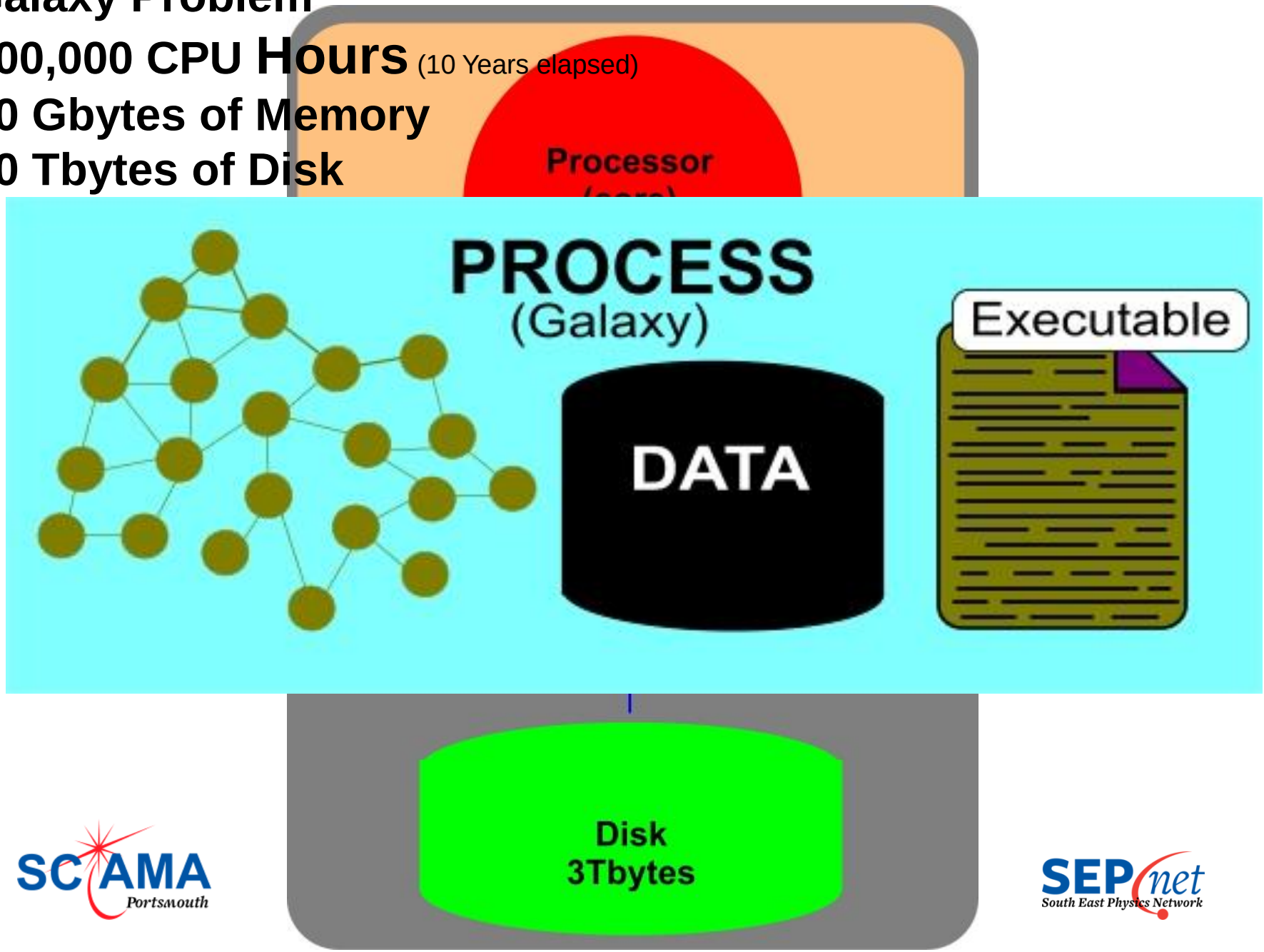
SEPnet South East Physics Network

Galaxy Problem

100,000 CPU Hours (10 Years elapsed)

30 Gbytes of Memory

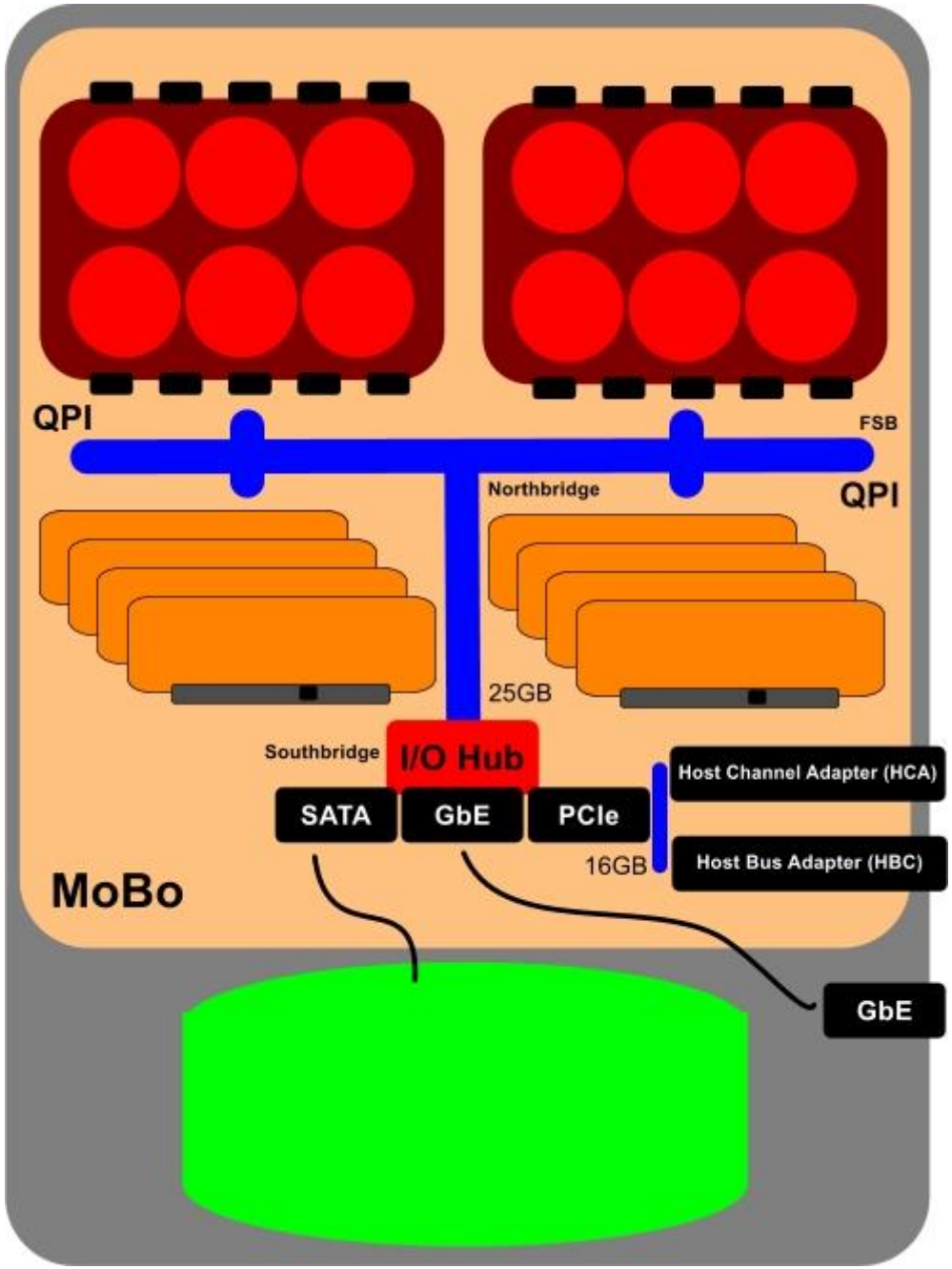
20 Tbytes of Disk



SCAMA Portsmouth

SEPnet South East Physics Network

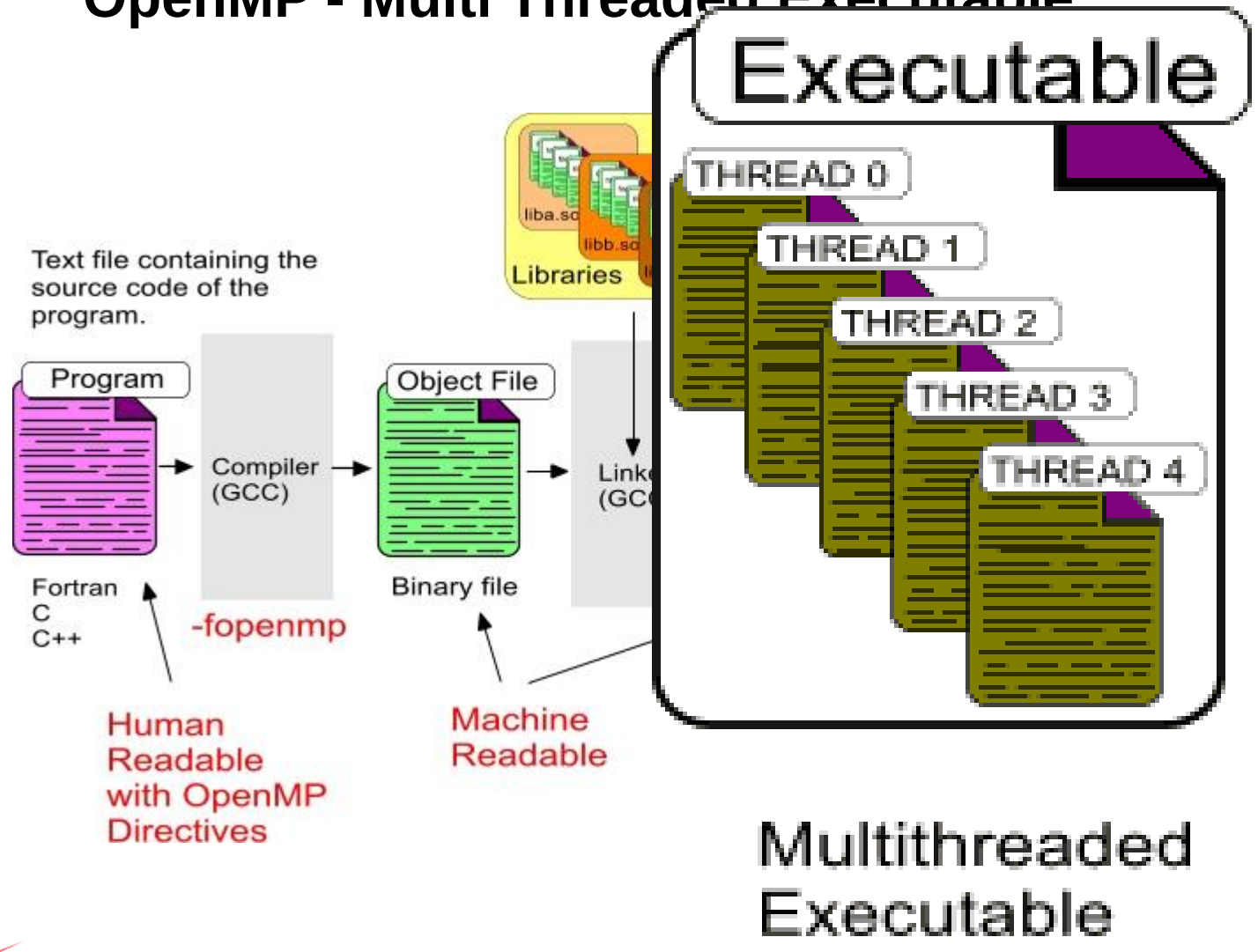
Galaxy Problem
100,000 CPU Hours
30 Gbytes of Memory
20 Tbytes of Disk



SCAMA Portsmouth

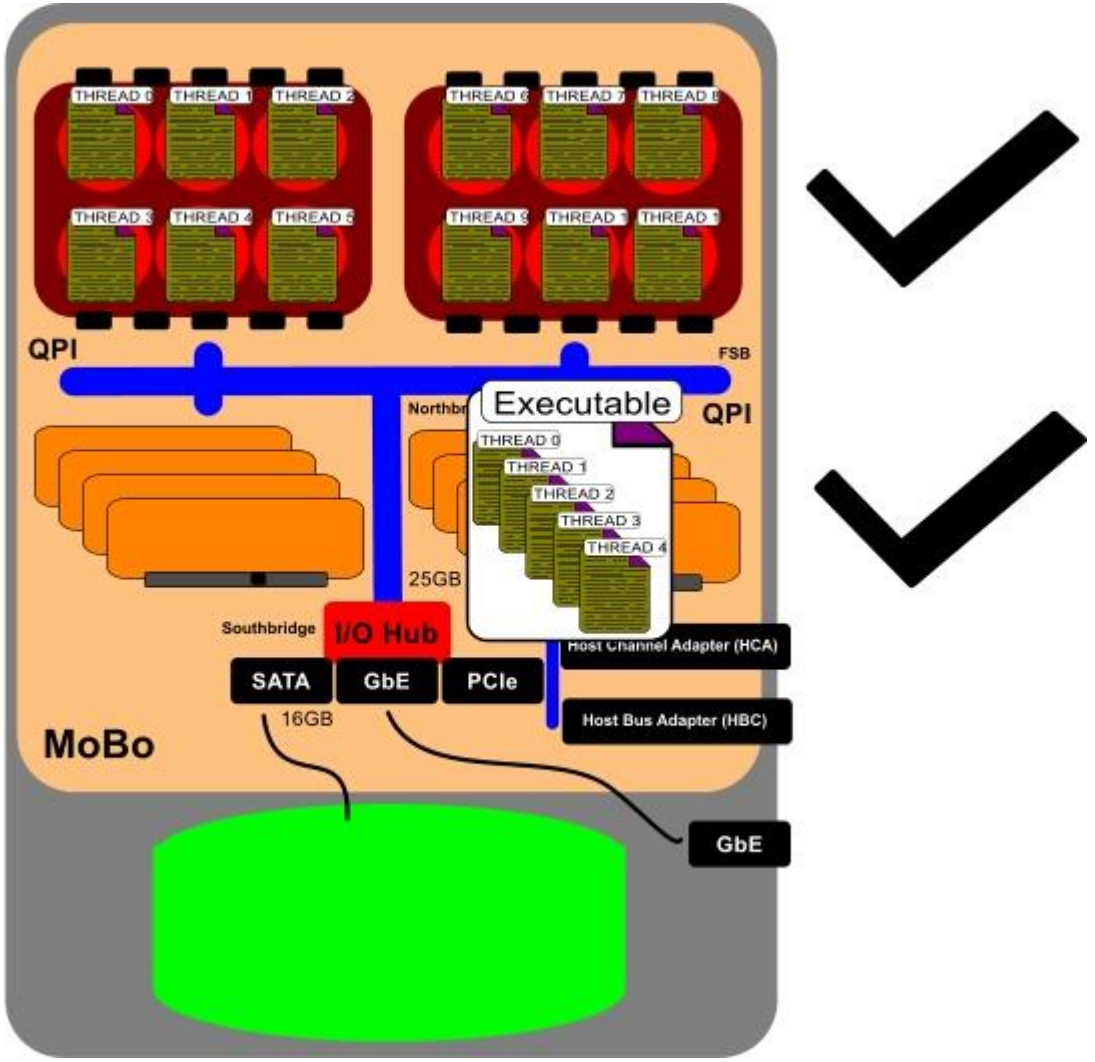
SEPnet South East Physics Network

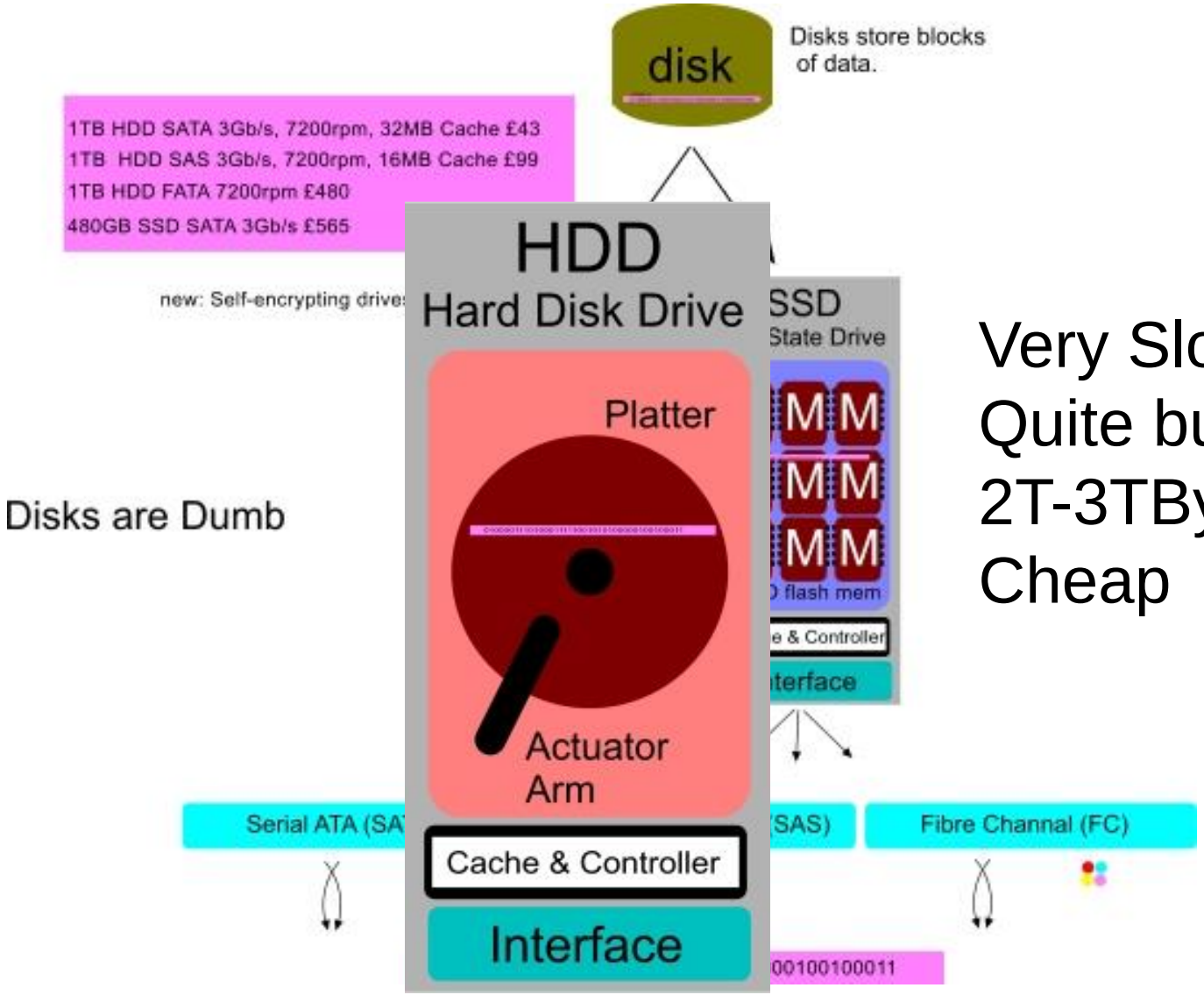
OpenMP - Multi Threaded Executable



SCAMA Portsmouth ICG Portsmouth ALCES SOFTWARE University of Portsmouth SEPnet South East Physics Network

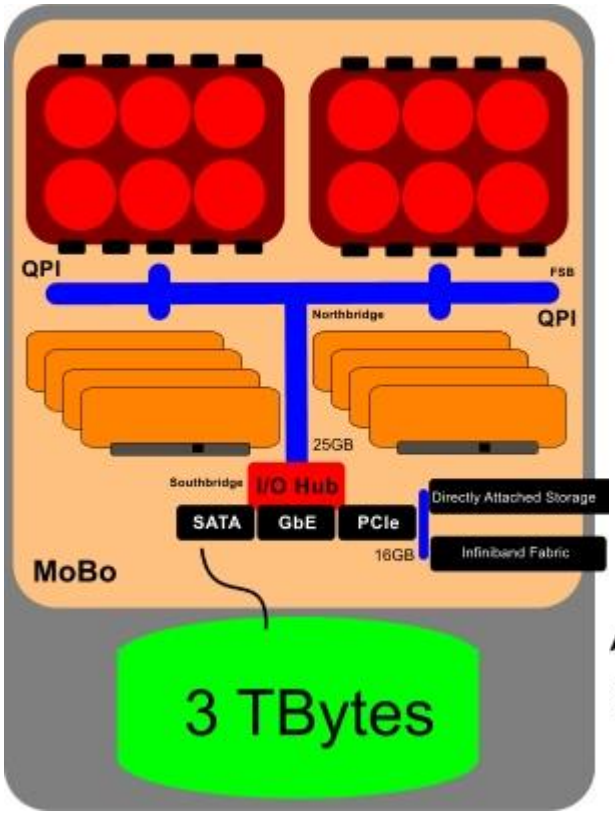
10 x Faster so
100,000 Cpu Hrs
goes to
10,000 Elapsed Hrs

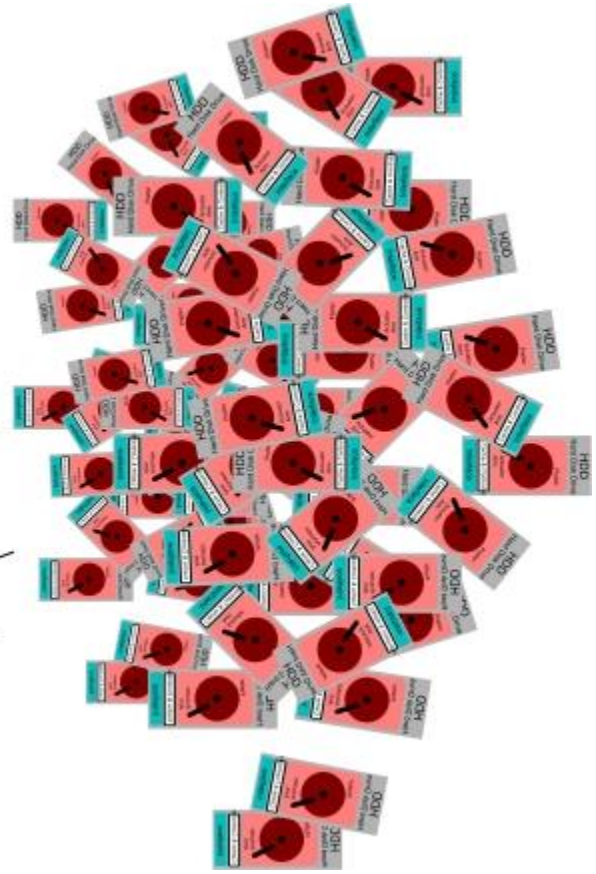




Very Slow
Quite bulky
2T-3TBytes
Cheap

SCAMA Portsmouth
ICG Portsmouth
ALCES SOFTWARE
University of Portsmouth
SEPnet South East Physics Network





SCAMA Portsmouth

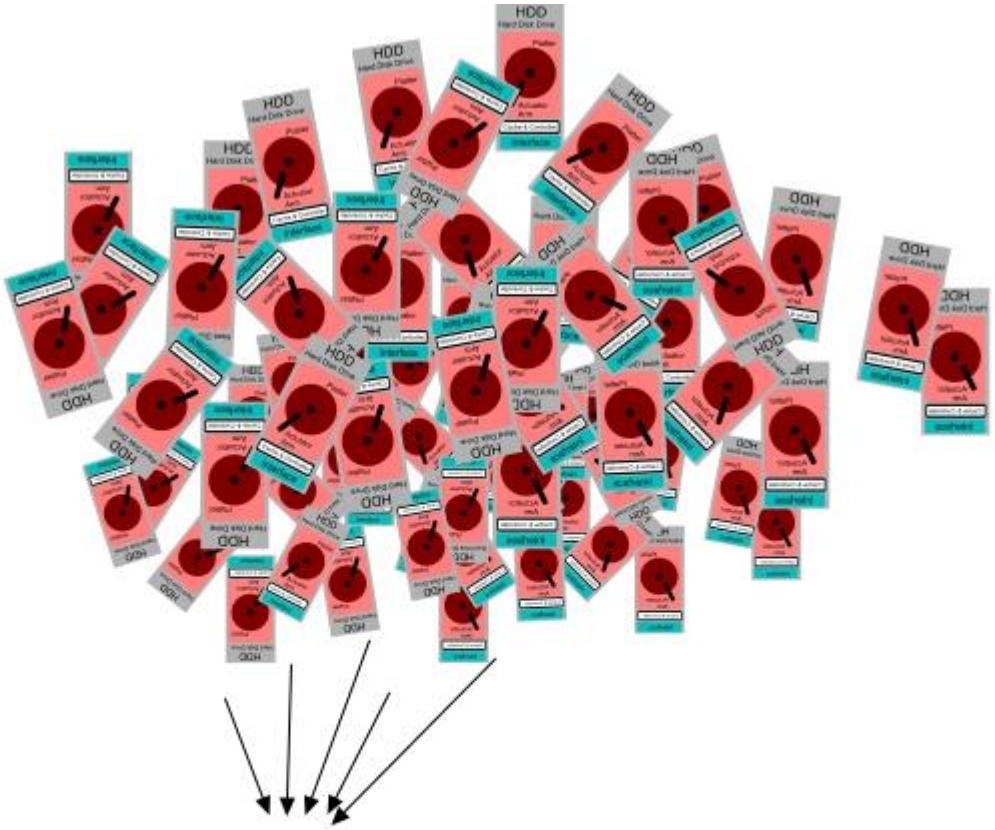
ICG Portsmouth

ALCES SOFTWARE

University of Portsmouth

SEPnet South East Physics Network

Commodity Disks



SCAMA Portsmouth

ICG Portsmouth

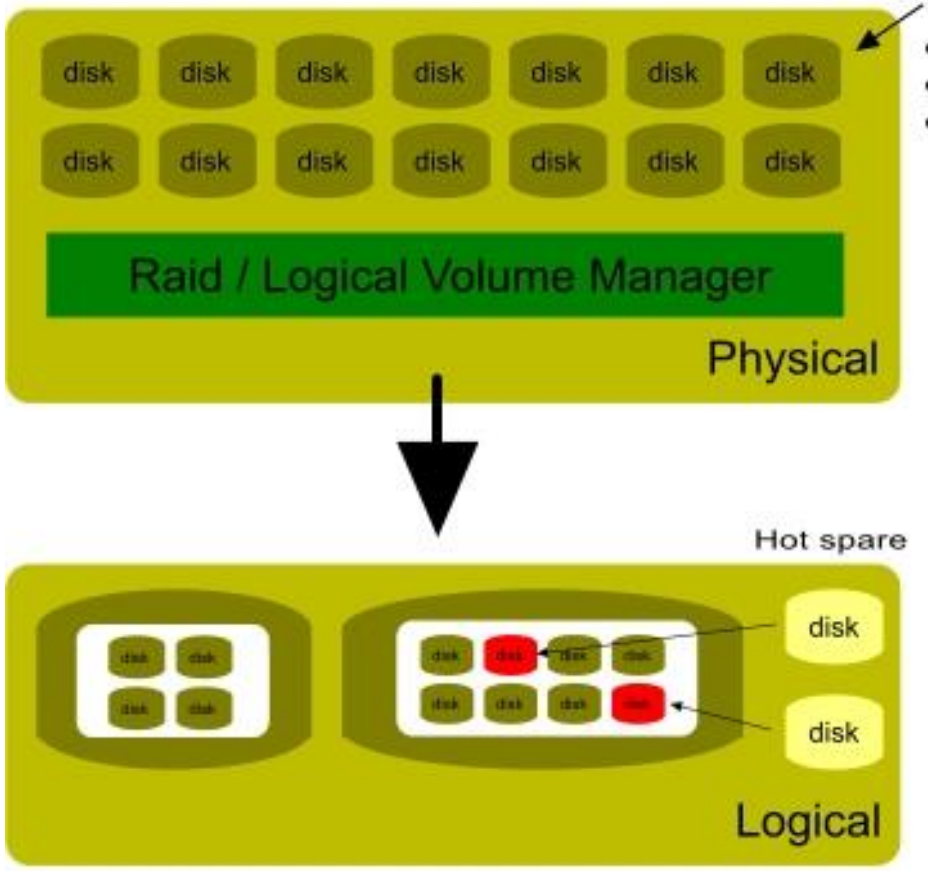
ALCES SOFTWARE

University of Portsmouth

SEPnet South East Physics Network

In HPC internal disks usually only contain the Operating System.
Some times two disks are "mirrored" for security.

Single Disk of little use for data :-

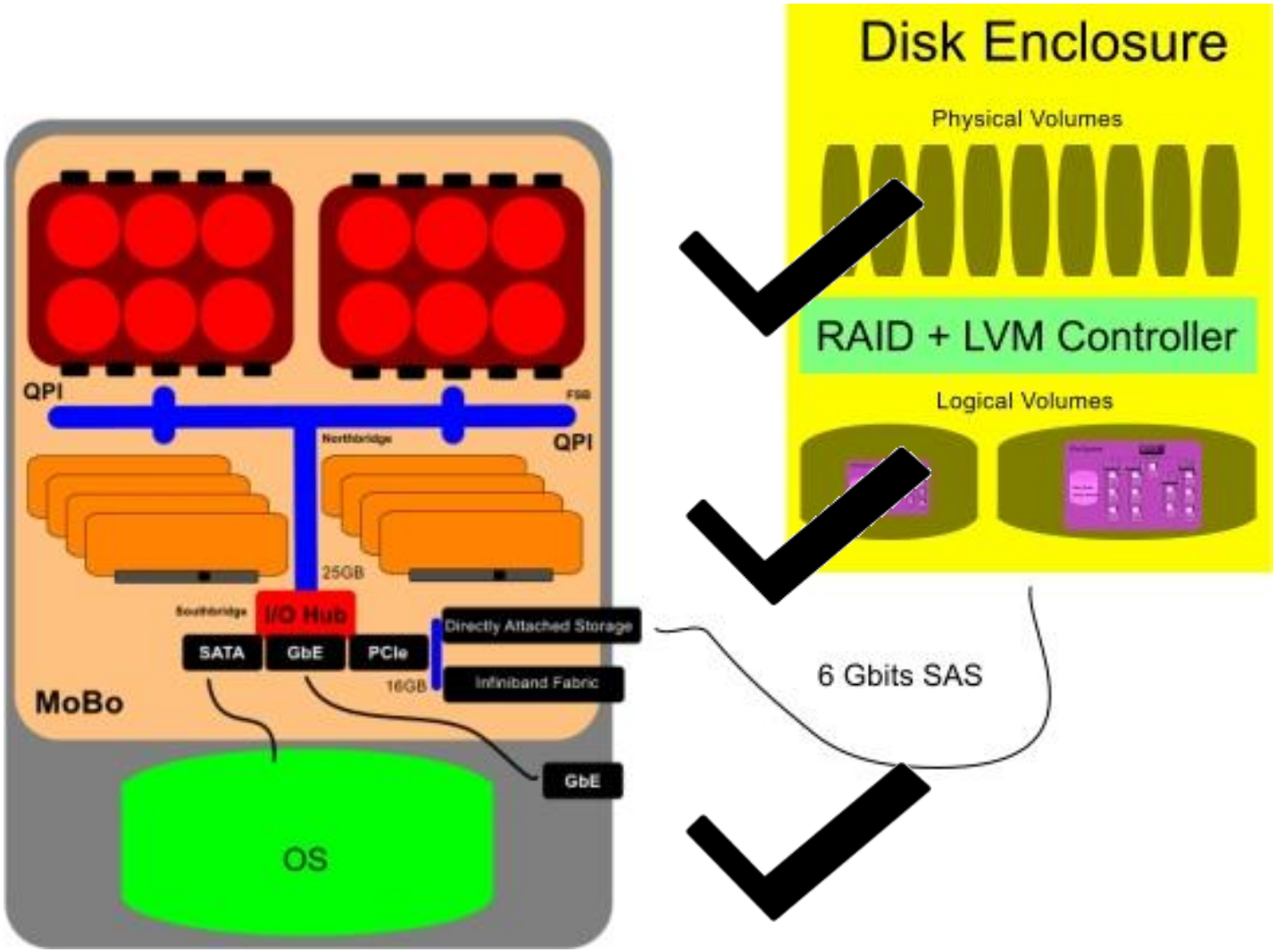


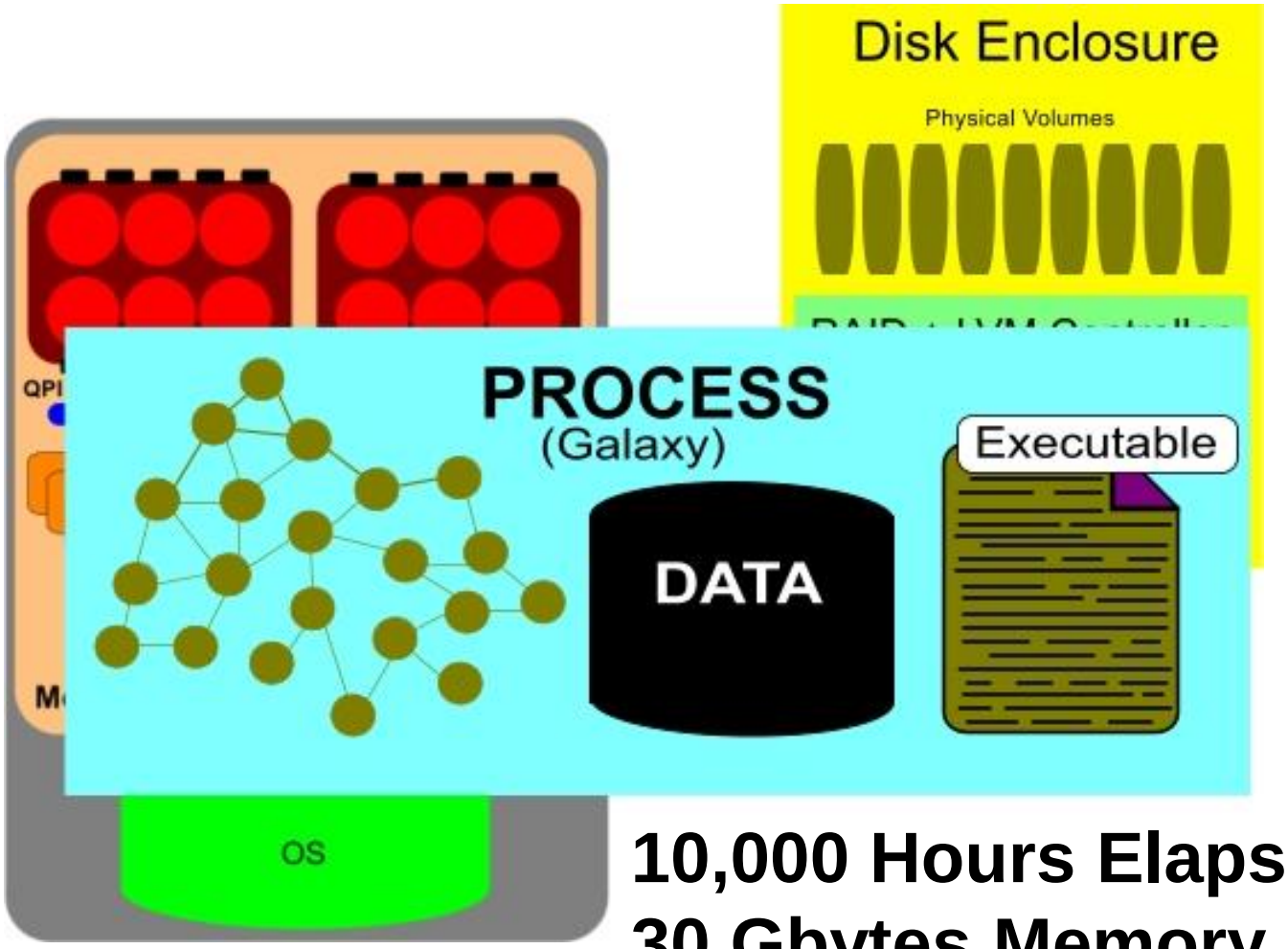
- Limited size.
- Limited performance
- Limited fault tolerance

Logical Volumes

- Increased size
- increased performance through stripping.
- Increased resilience through parity and mirroring.

SCAMA Portsmouth ICG Portsmouth ALCES SOFTWARE University of Portsmouth SEPnet South East Physics Network


Disk Enclosure
Physical Volumes
RAID + LVM Controller
Logical Volumes
QPI
FSB
QPI
Northbridge
25GB
Southbridge
I/O Hub
SATA
GbE
PCIe
16GB
Directly Attached Storage
Infiniband Fabric
MoBo
GbE
OS
6 Gbits SAS



Answer

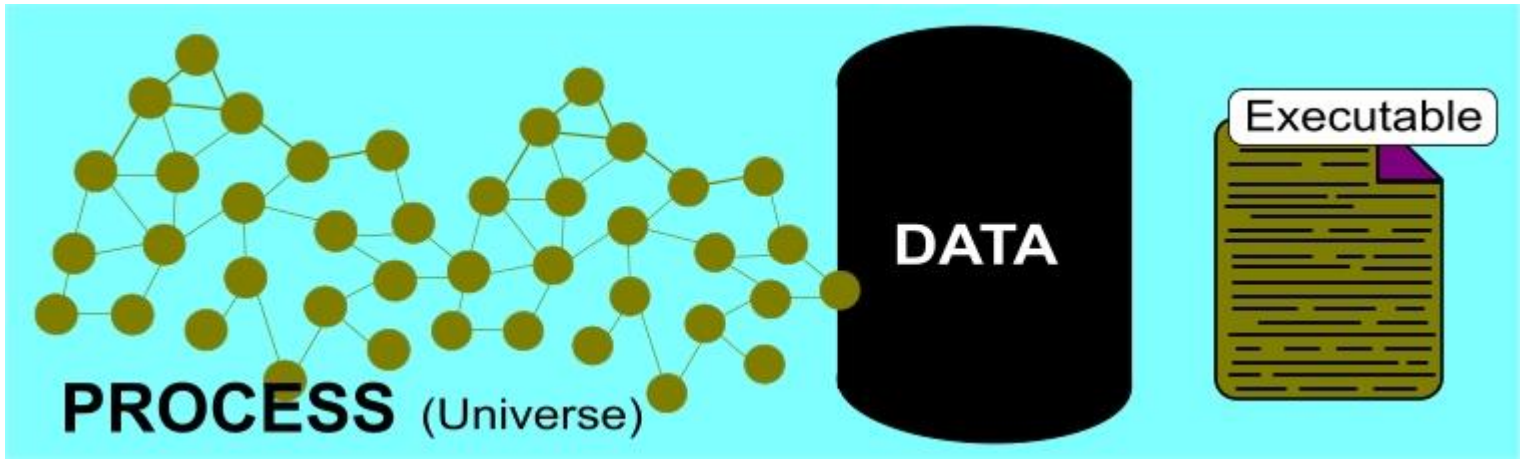
54

10,000 Hours Elapsed (Parallel Processing)

30 Gbytes Memory (Large memory boards)

20 Tbytes Disk (Direcly Attached Storage)

SCAMA Portsmouth ICG Portsmouth ALCES SOFTWARE University of Portsmouth SEPnet South East Physics Network

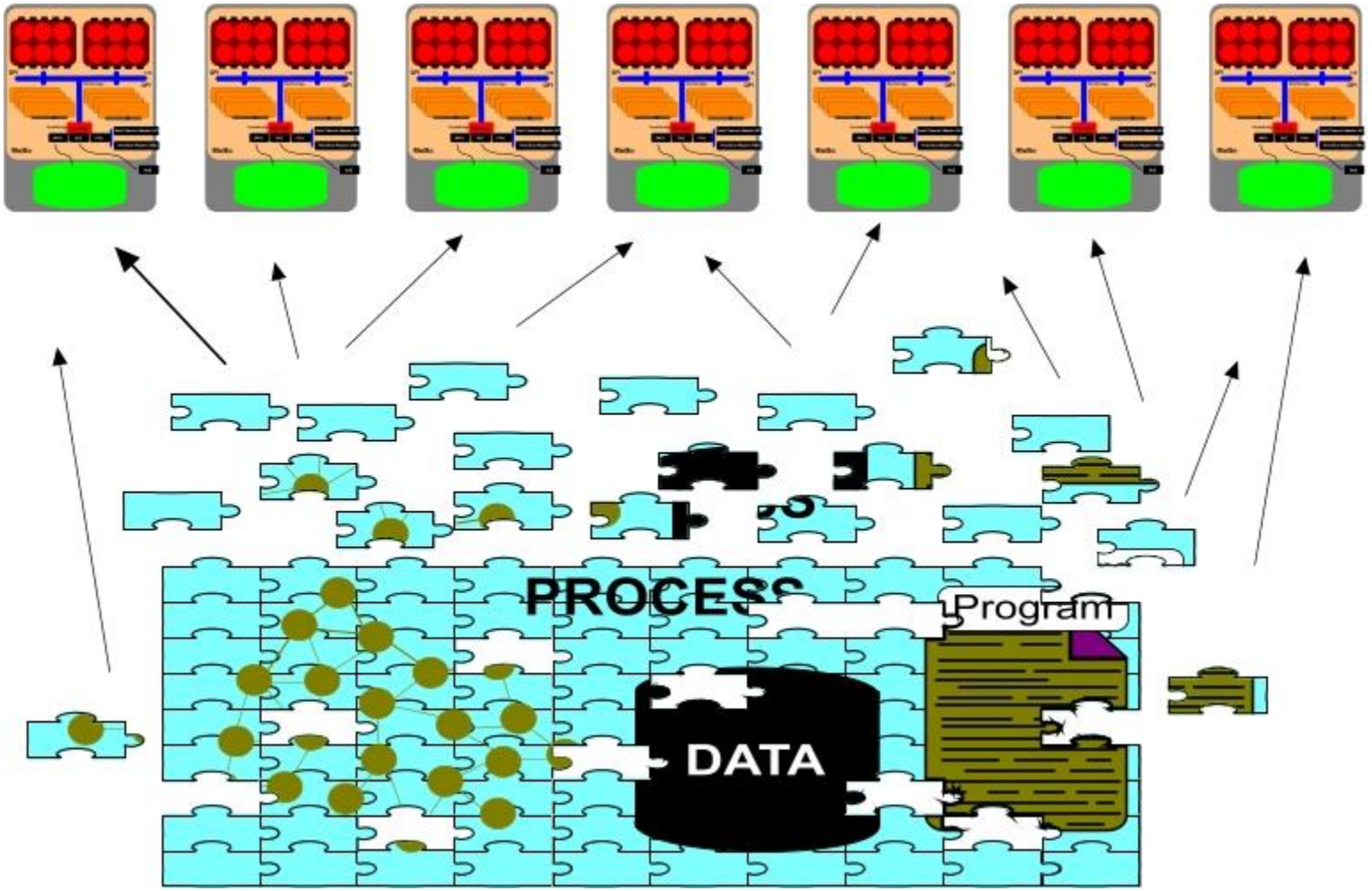


Universe Problem

10,000,000 CPU Hours (1000 years)
3T Bytes Memory
200 Tbytes Disk

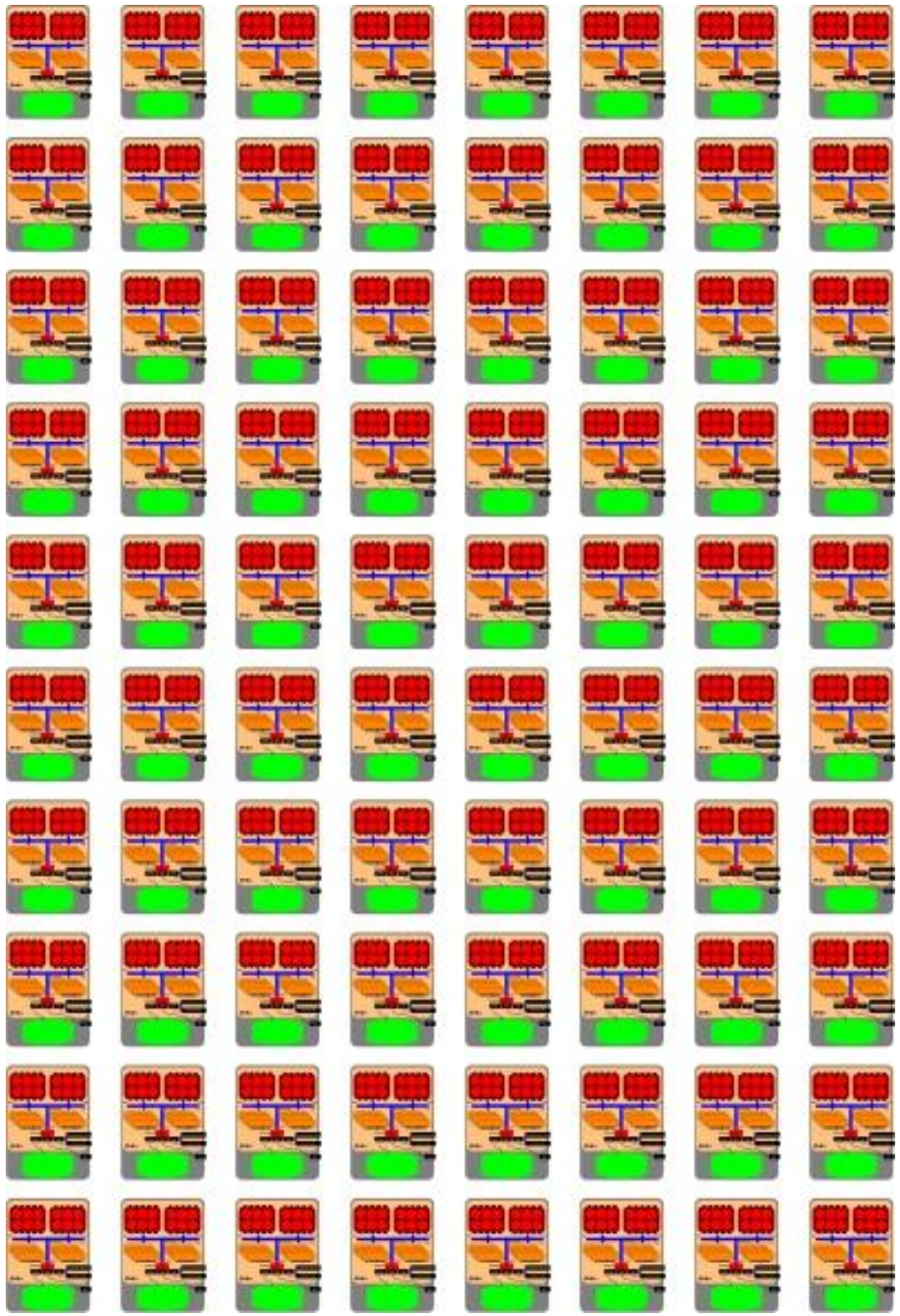
SCAMA Portsmouth
ICG Portsmouth
ALCES SOFTWARE
University of Portsmouth
SEPnet South East Physics Network

Commodity is Cheap – Specialised is Expensive



SCAMA Portsmouth ICG Portsmouth ALCES SOFTWARE University of Portsmouth SEPnet South East Physics Network

Nodes (10 Cores)



Required:-

10,000,000 CPU Hours
3T Bytes Memory

Distributed Over 100 Nodes (100x10 cores):-

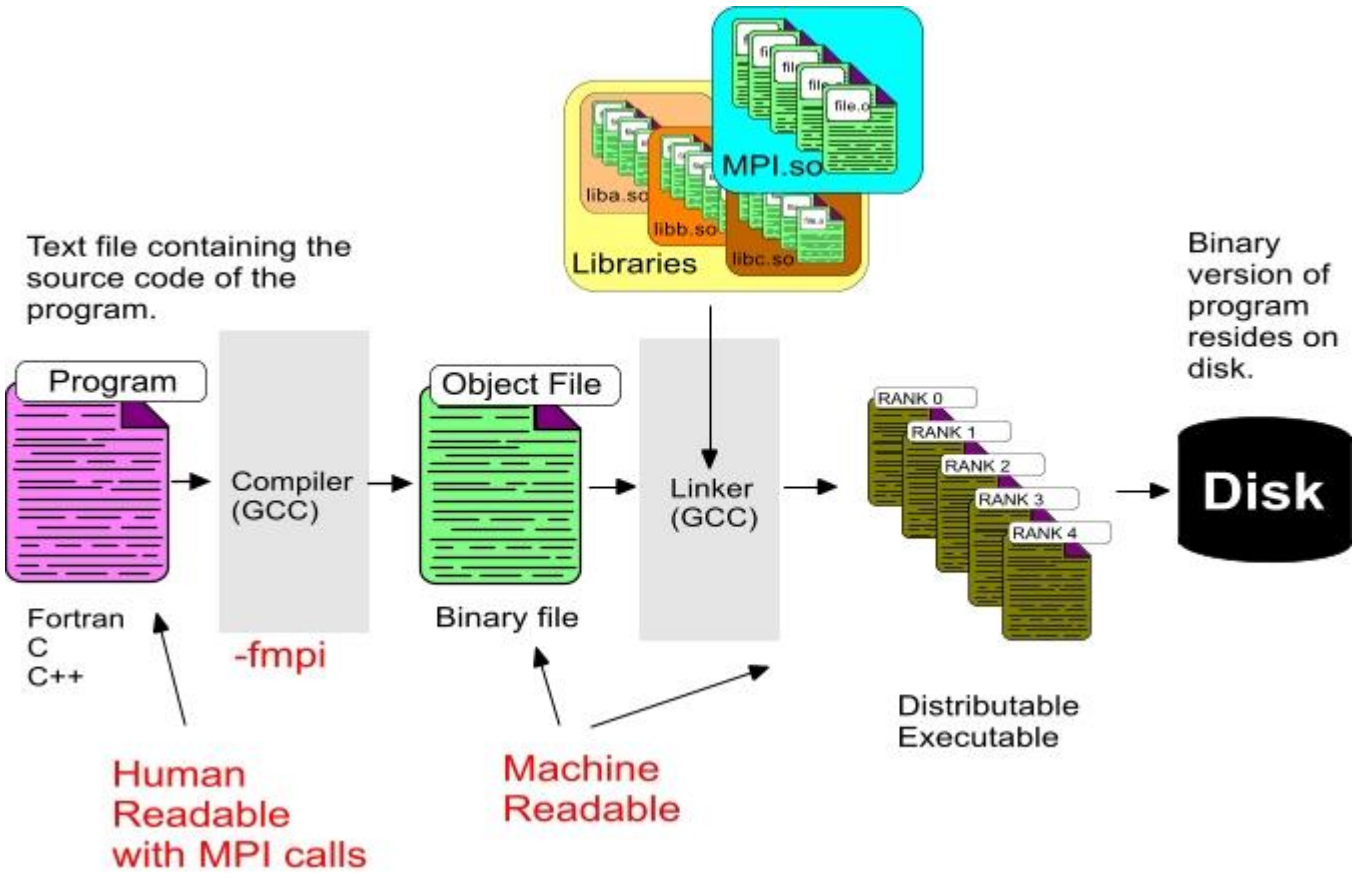
10,000 Hours Elapsed
30Gbytes / Node

Commodity Cluster

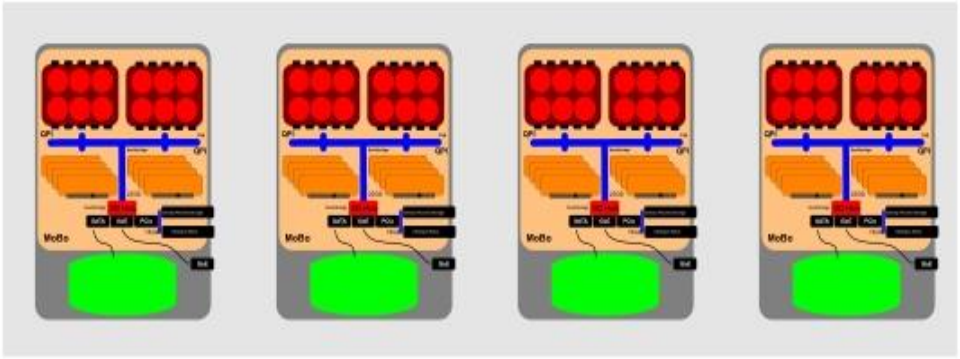
SCAMA Portsmouth

SEPnet South East Physics Network

Message Passing Interface - MPI



Dell 6100 Front View



Dell 6100
Rear View

High density - Google / Amazon

SCAMA Portsmouth

SEPnet South East Physics Network

SCAMA Portsmouth

ICG Portsmouth

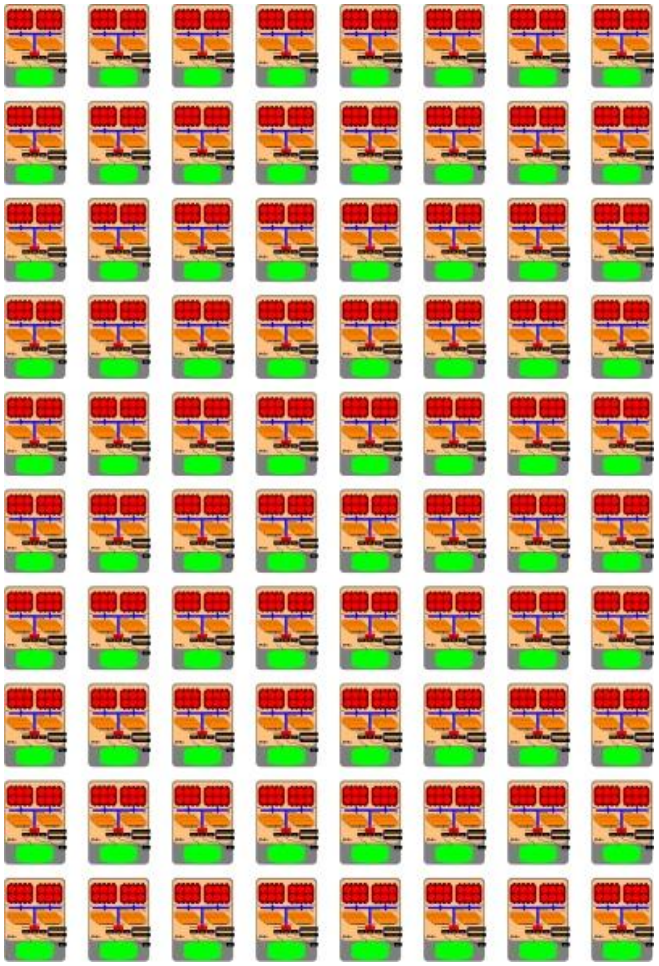
ALCES SOFTWARE

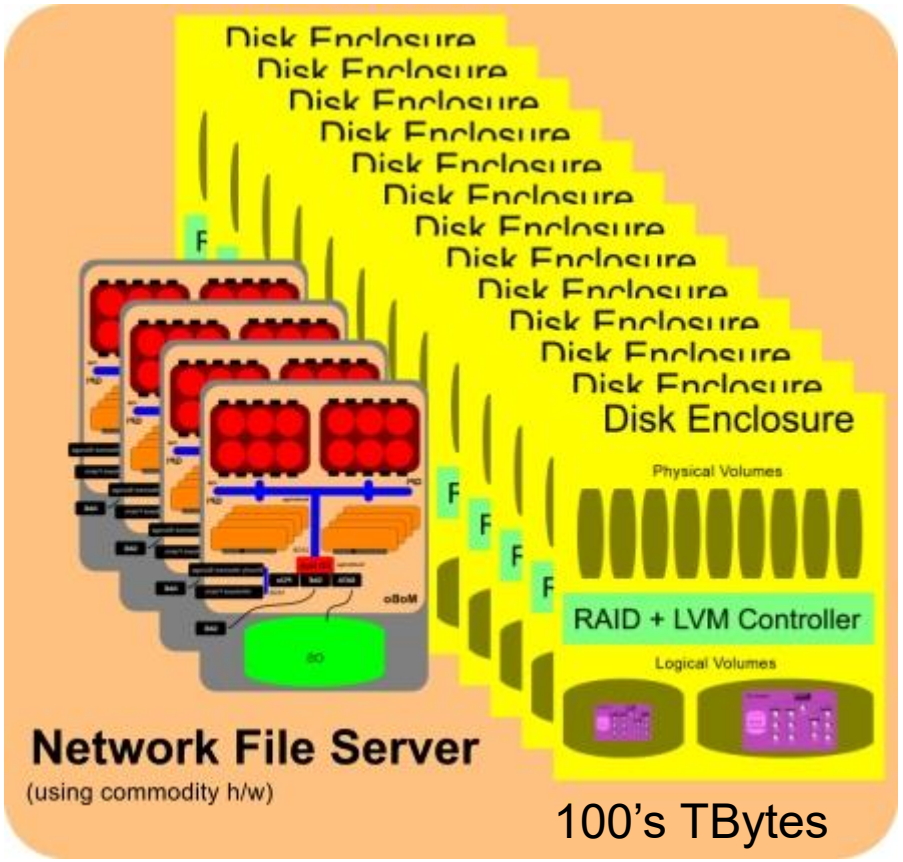
University of Portsmouth

SEPnet South East Physics Network

Shared Storage

(Require 200TBytes)





Network File System Clients

A change made by one Node will be seen by all others.

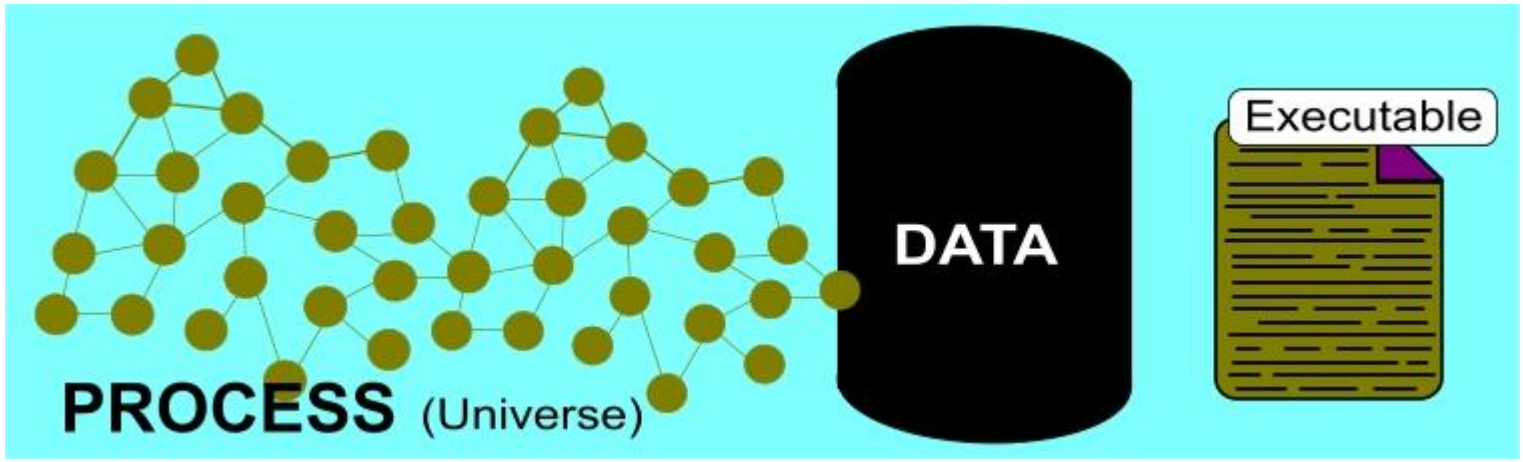
SCAMA Portsmouth

ICG Portsmouth

ALCES SOFTWARE

University of Portsmouth

SEPnet South East Physics Network



Universe Problem

Answer

42

10,000 Hours Elapsed (Distributed Cores)

3 Tbytes Memory (Distributed Memory)

200 Tbytes Disk (Network File Server)

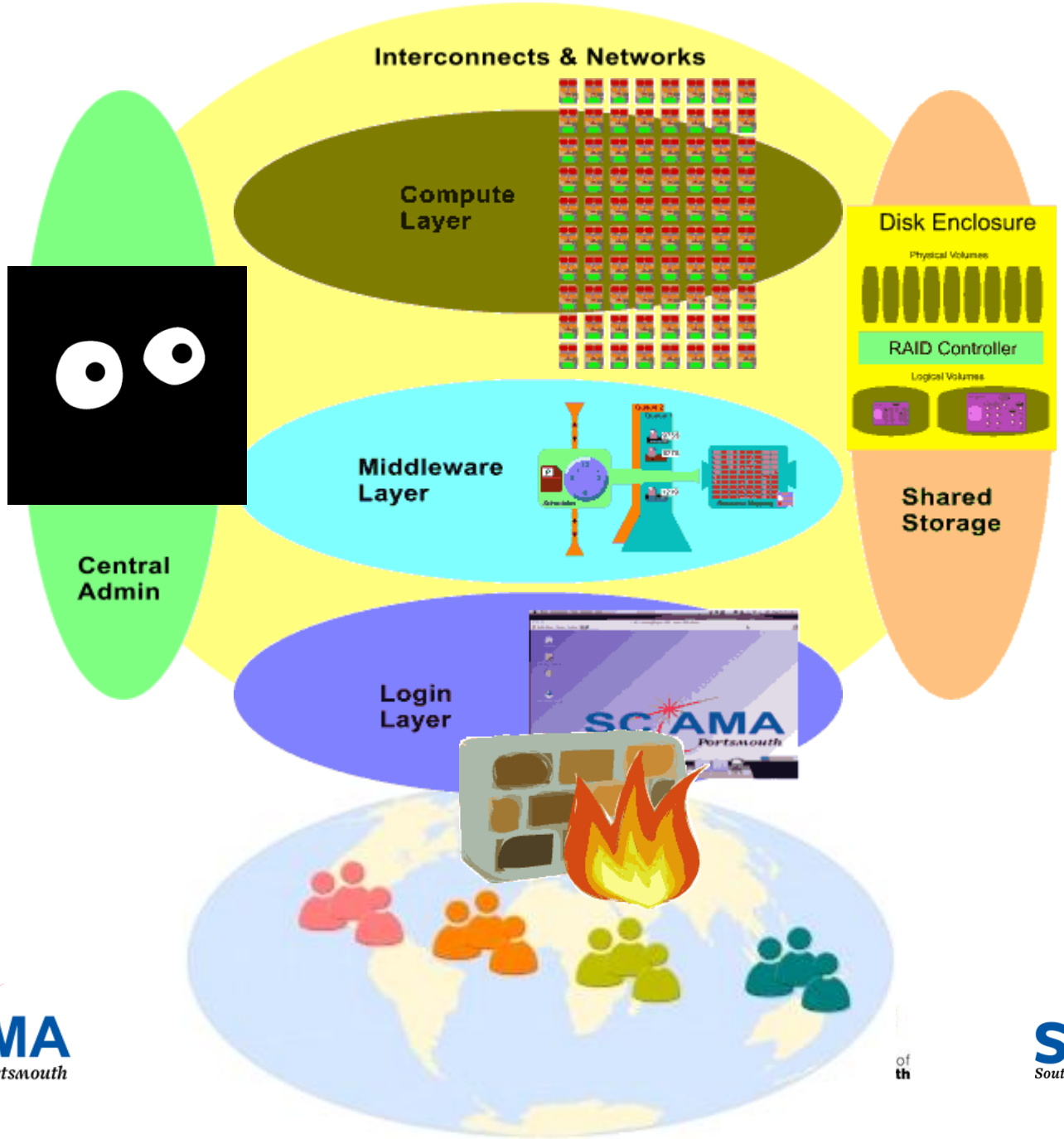
C: Hitch Hikers Guide to the Galaxy

SCAMA Portsmouth ICG Portsmouth ALCES SOFTWARE University of Portsmouth SEPnet South East Physics Network

So long …. and thanks for all the fish !!

C: Hitch Hikers Guide to the Galaxy

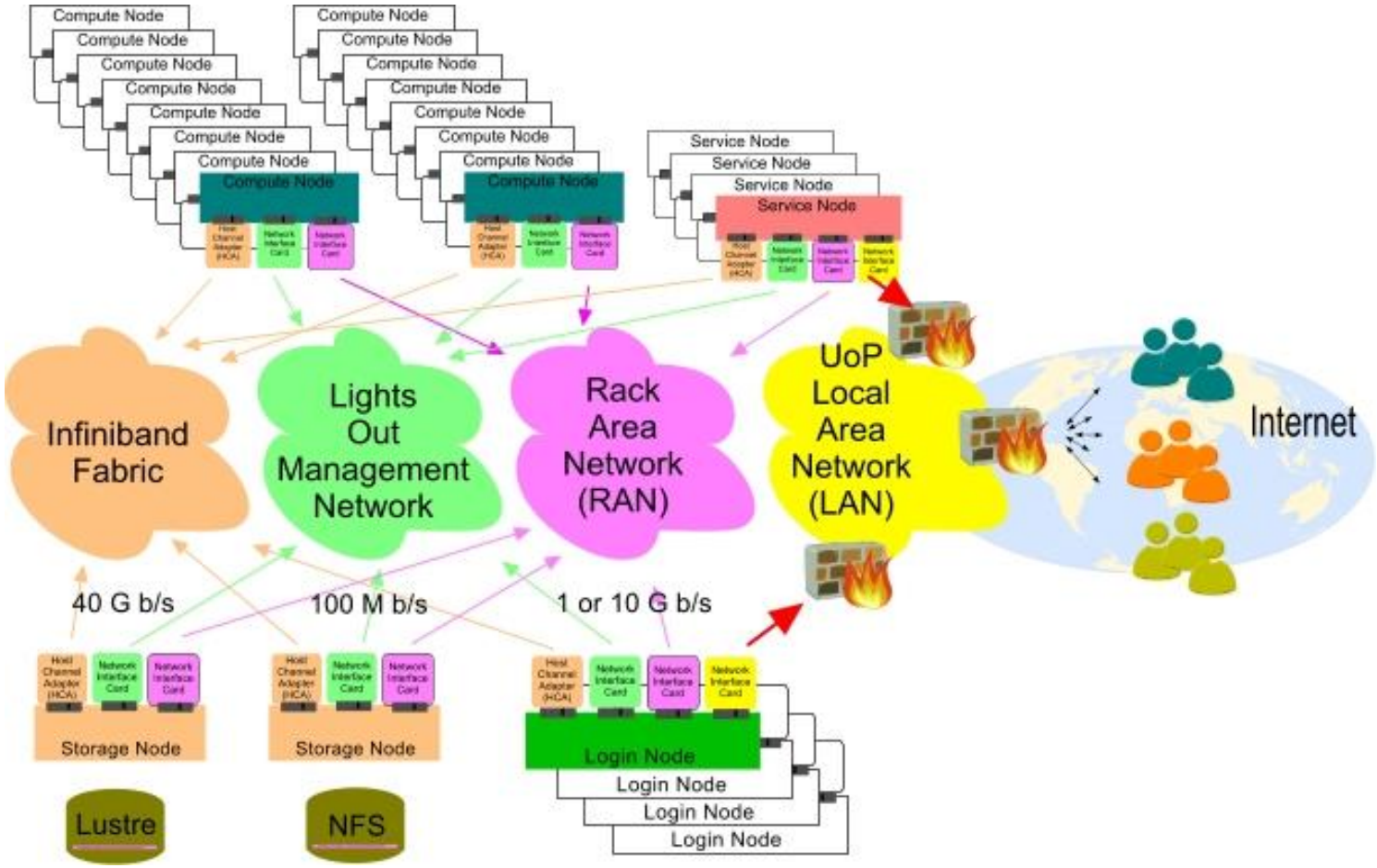
SCAMA Portsmouth ICG Portsmouth ALCES SOFTWARE University of Portsmouth SEPnet South East Physics Network



SCAMA Portsmouth

SEPnet South East Physics Network

Sciama Network



SCIAMA Portsmouth

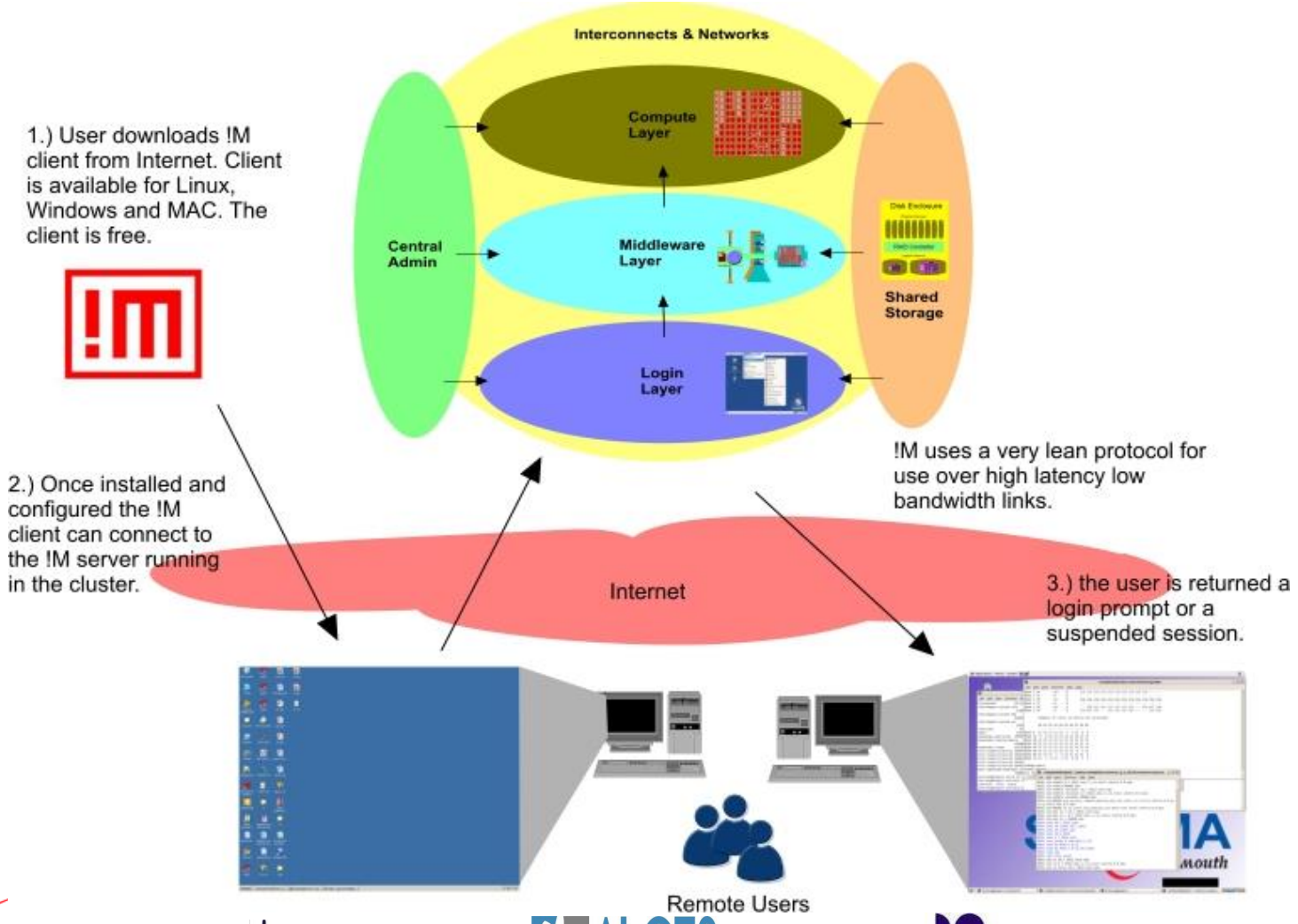
ICG Portsmouth

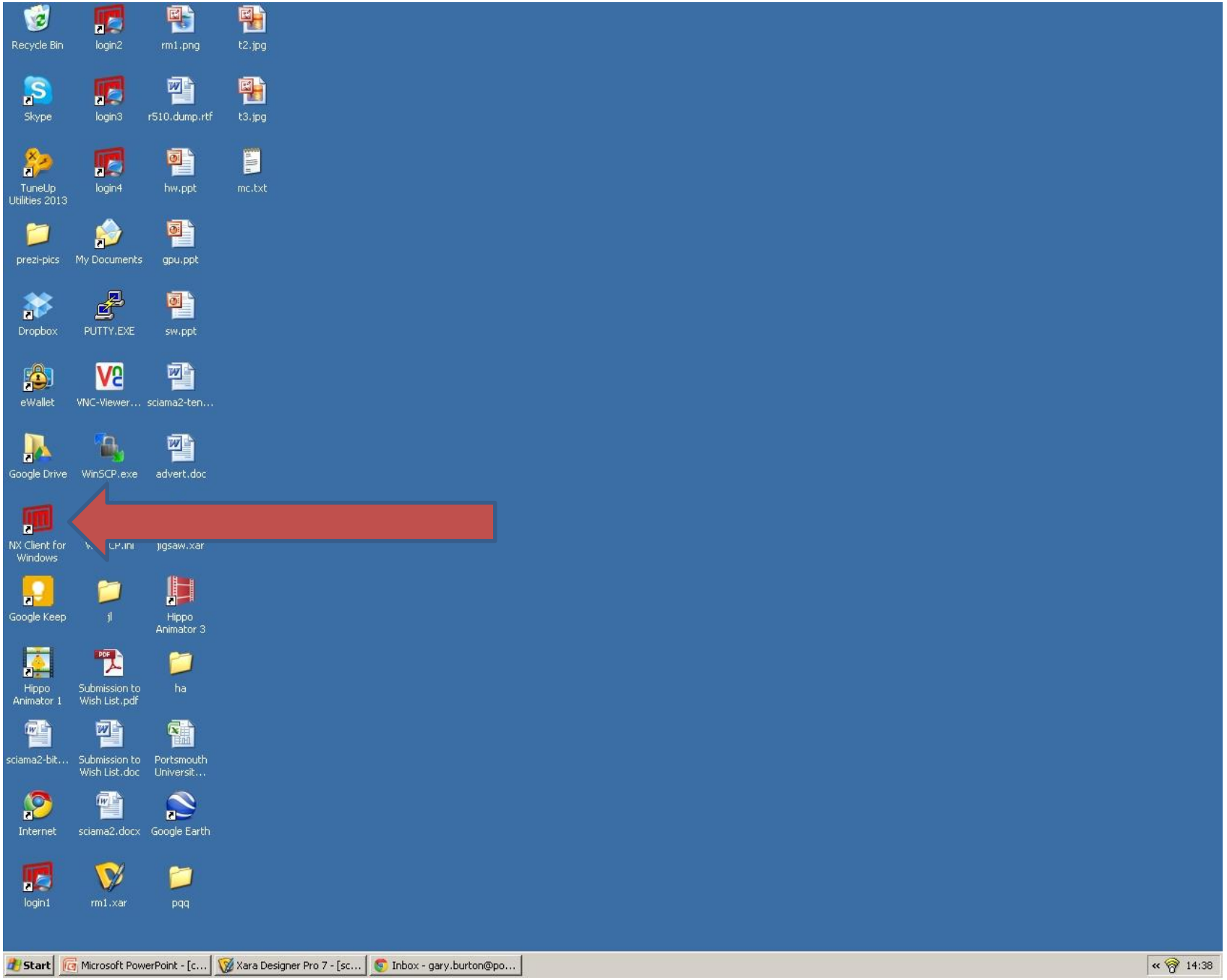
ALCES SOFTWARE

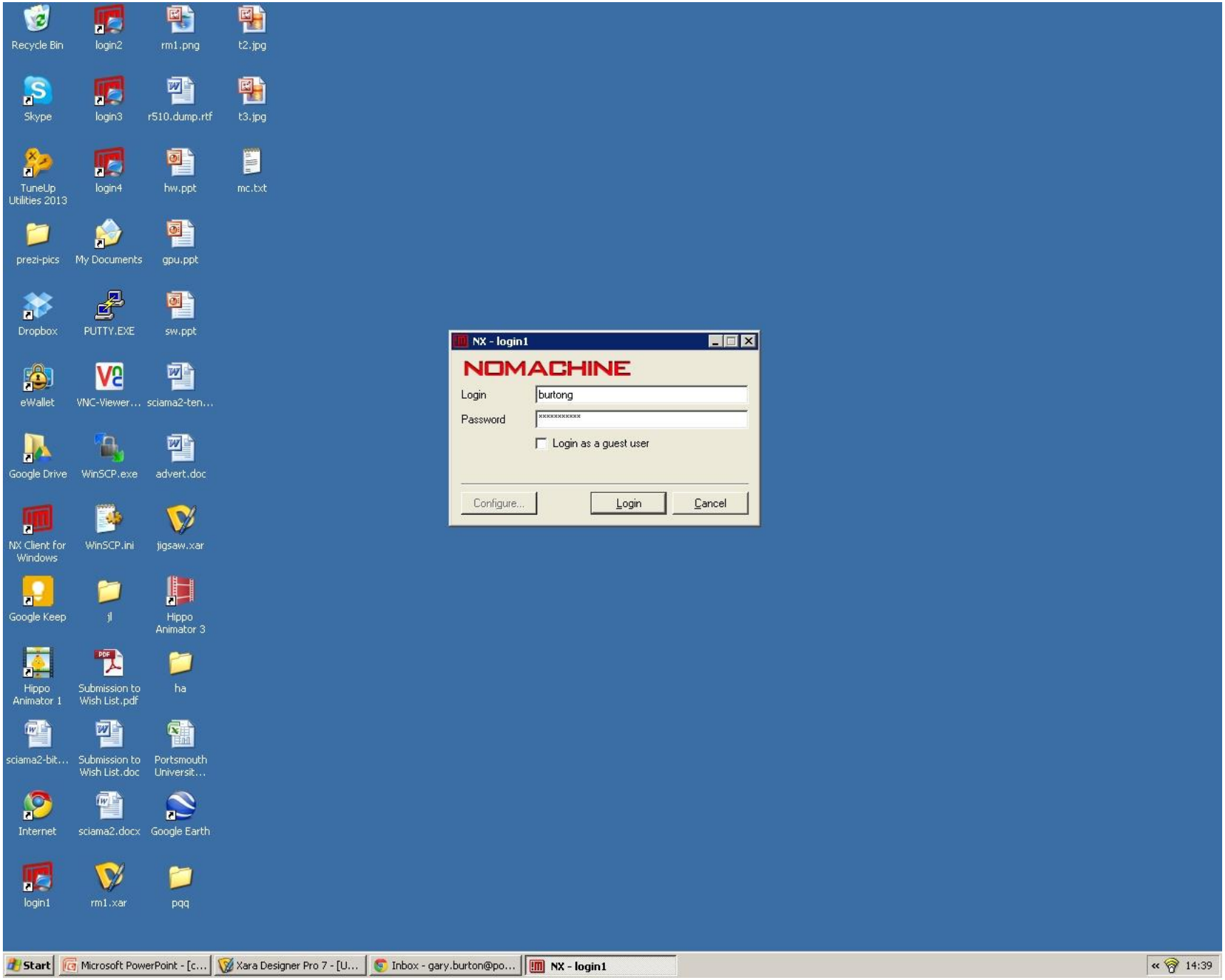
University of Portsmouth

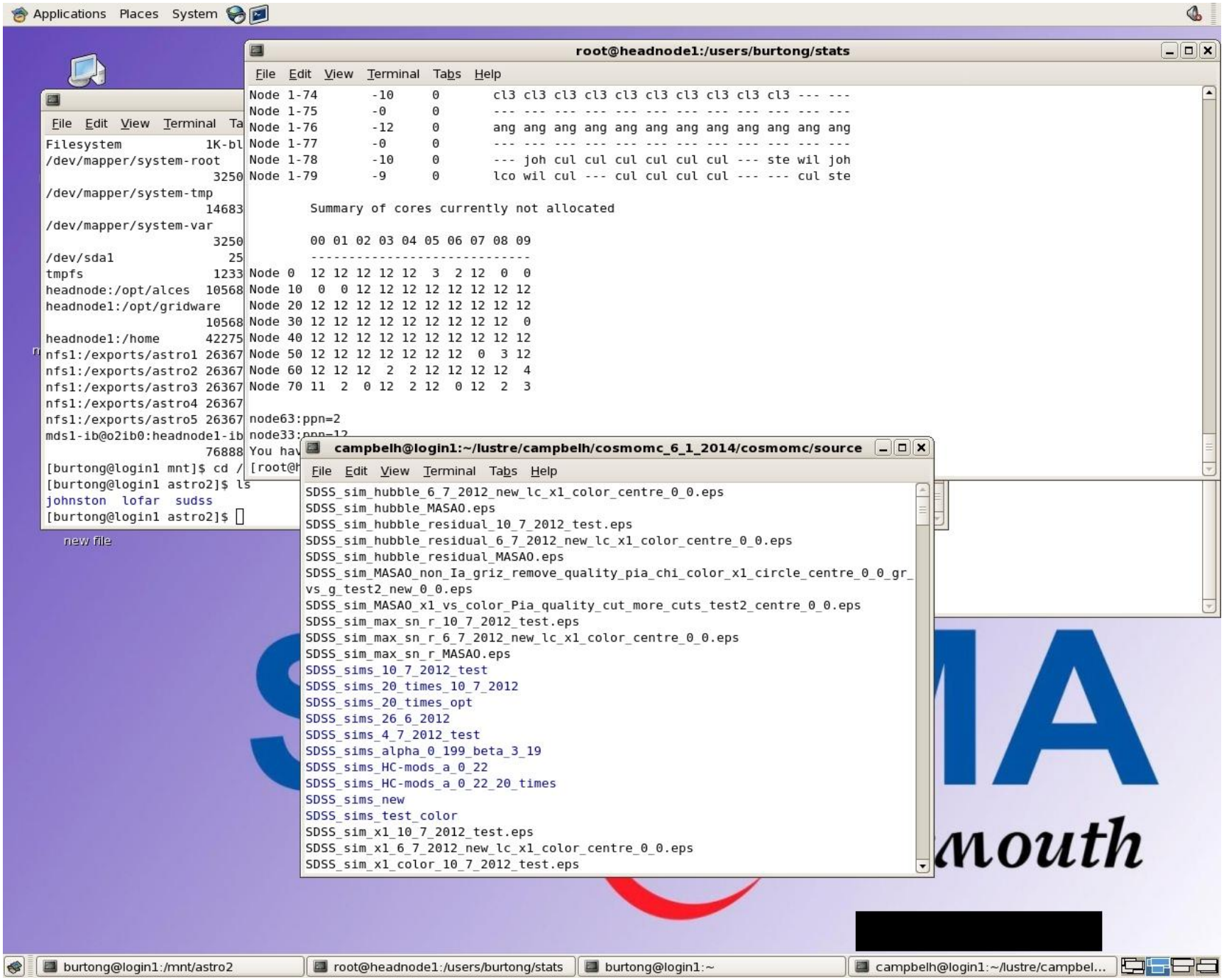
SEPnet South East Physics Network

Use of Remote Login Client

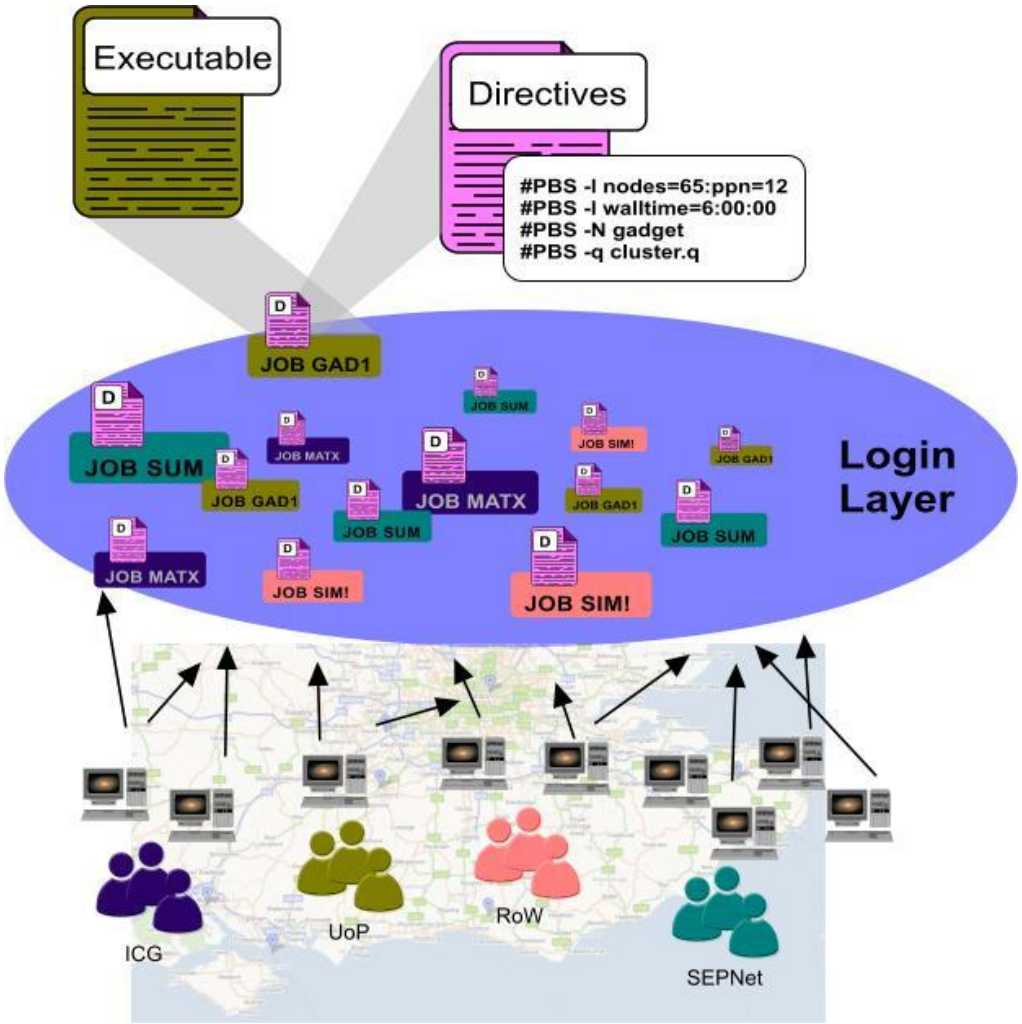



Recycle Bin
login2
rm1.png
t2.jpg
Skype
login3
r510.dump.rtf
t3.jpg
TuneUp Utilities 2013
login4
hw.ppt
mc.txt
prezi-pics
My Documents
gpu.ppt
Dropbox
PUTTY.EXE
sw.ppt
eWallet
VNC-Viewer...
sciama2-ten...
Google Drive
WinSCP.exe
advert.doc
NX Client for Windows
jigsaw.xar
Google Keep
jl
Hippo Animator 3
Hippo Animator 1
Submission to Wish List.pdf
ha
sciama2-bit...
Submission to Wish List.doc
Portsmouth Universit...
Internet
sciama2.docx
Google Earth
login1
rm1.xar
pqq
Start
Microsoft PowerPoint - [c...
Xara Designer Pro 7 - [sc...
Inbox - gary.burton@po...
14:38


Recycle Bin
login2
rm1.png
t2.jpg
Skype
login3
r510.dump.rtf
t3.jpg
TuneUp Utilities 2013
login4
hw.ppt
mc.txt
prezi-pics
My Documents
gpu.ppt
Dropbox
PUTTY.EXE
sw.ppt
eWallet
VNC-Viewer...
sciama2-ten...
Google Drive
WinSCP.exe
advert.doc
NX Client for Windows
WinSCP.ini
jigsaw.xar
Google Keep
jl
Hippo Animator 3
Hippo Animator 1
Submission to Wish List.pdf
ha
sciama2-bit...
Submission to Wish List.doc
Portsmouth Universit...
Internet
sciama2.docx
Google Earth
login1
rm1.xar
pqq
NX - login1
NOMACHINE
Login
burtong
Password
Login as a guest user
Configure...
Login
Cancel
Start
Microsoft PowerPoint - [c...
Xara Designer Pro 7 - [U...
Inbox - gary.burton@po...
NX - login1
14:39


Applications Places System
root@headnode1:/users/burtong/stats
File Edit View Terminal Tabs Help
Node 1-74 -10 0 cl3 cl3 cl3 cl3 cl3 cl3 cl3 cl3 cl3 cl3 --- ---
Node 1-75 -0 0 --- --- --- --- --- --- --- --- --- --- --- ---
Node 1-76 -12 0 ang ang ang ang ang ang ang ang ang ang ang ang
Node 1-77 -0 0 --- --- --- --- --- --- --- --- --- --- --- ---
Node 1-78 -10 0 --- joh cul cul cul cul cul cul --- ste wil joh
Node 1-79 -9 0 lco wil cul --- cul cul cul cul --- --- cul ste
Summary of cores currently not allocated
00 01 02 03 04 05 06 07 08 09
Node 0 12 12 12 12 12 3 2 12 0 0
Node 10 0 0 12 12 12 12 12 12 12 12
Node 20 12 12 12 12 12 12 12 12 12 12
Node 30 12 12 12 12 12 12 12 12 12 0
Node 40 12 12 12 12 12 12 12 12 12 12
Node 50 12 12 12 12 12 12 12 0 3 12
Node 60 12 12 12 2 2 12 12 12 12 4
Node 70 11 2 0 12 2 12 0 12 2 3
node63:ppn=2
node33:ppn=12
campbelh@login1:~/lustre/campbelh/cosmomc_6_1_2014/cosmomc/source
File Edit View Terminal Tabs Help
SDSS_sim_hubble_6_7_2012_new_lc_x1_color_centre_0_0.eps
SDSS_sim_hubble_MASA0.eps
SDSS_sim_hubble_residual_10_7_2012_test.eps
SDSS_sim_hubble_residual_6_7_2012_new_lc_x1_color_centre_0_0.eps
SDSS_sim_hubble_residual_MASA0.eps
SDSS_sim_MASA0_non_Ia_griz_remove_quality_pia_chi_color_x1_circle_centre_0_0_gr_vs_g_test2_new_0_0.eps
SDSS_sim_MASA0_x1_vs_color_Pia_quality_cut_more_cuts_test2_centre_0_0.eps
SDSS_sim_max_sn_r_10_7_2012_test.eps
SDSS_sim_max_sn_r_6_7_2012_new_lc_x1_color_centre_0_0.eps
SDSS_sim_max_sn_r_MASA0.eps
SDSS_sims_10_7_2012_test
SDSS_sims_20_times_10_7_2012
SDSS_sims_20_times_opt
SDSS_sims_26_6_2012
SDSS_sims_4_7_2012_test
SDSS_sims_alpha_0_199_beta_3_19
SDSS_sims_HC-mods_a_0_22
SDSS_sims_HC-mods_a_0_22_20_times
SDSS_sims_new
SDSS_sims_test_color
SDSS_sim_x1_10_7_2012_test.eps
SDSS_sim_x1_6_7_2012_new_lc_x1_color_centre_0_0.eps
SDSS_sim_x1_color_10_7_2012_test.eps
File Edit View Terminal Ta
Filesystem 1K-bl
/dev/mapper/system-root
3250
/dev/mapper/system-tmp
14683
/dev/mapper/system-var
3250
/dev/sda1 25
tmpfs 1233
headnode:/opt/alces 10568
headnode1:/opt/gridware
10568
headnode1:/home 42275
nfs1:/exports/astro1 26367
nfs1:/exports/astro2 26367
nfs1:/exports/astro3 26367
nfs1:/exports/astro4 26367
nfs1:/exports/astro5 26367
mds1-ib@o2ib0:headnode1-ib
76888
[burtong@login1 mnt]$ cd /
[burtong@login1 astro2]$ ls
johnston lofar sudss
[burtong@login1 astro2]$
new file
mouth
burtong@login1:/mnt/astro2
root@headnode1:/users/burtong/stats
burtong@login1:~
campbelh@login1:~/lustre/campbel...

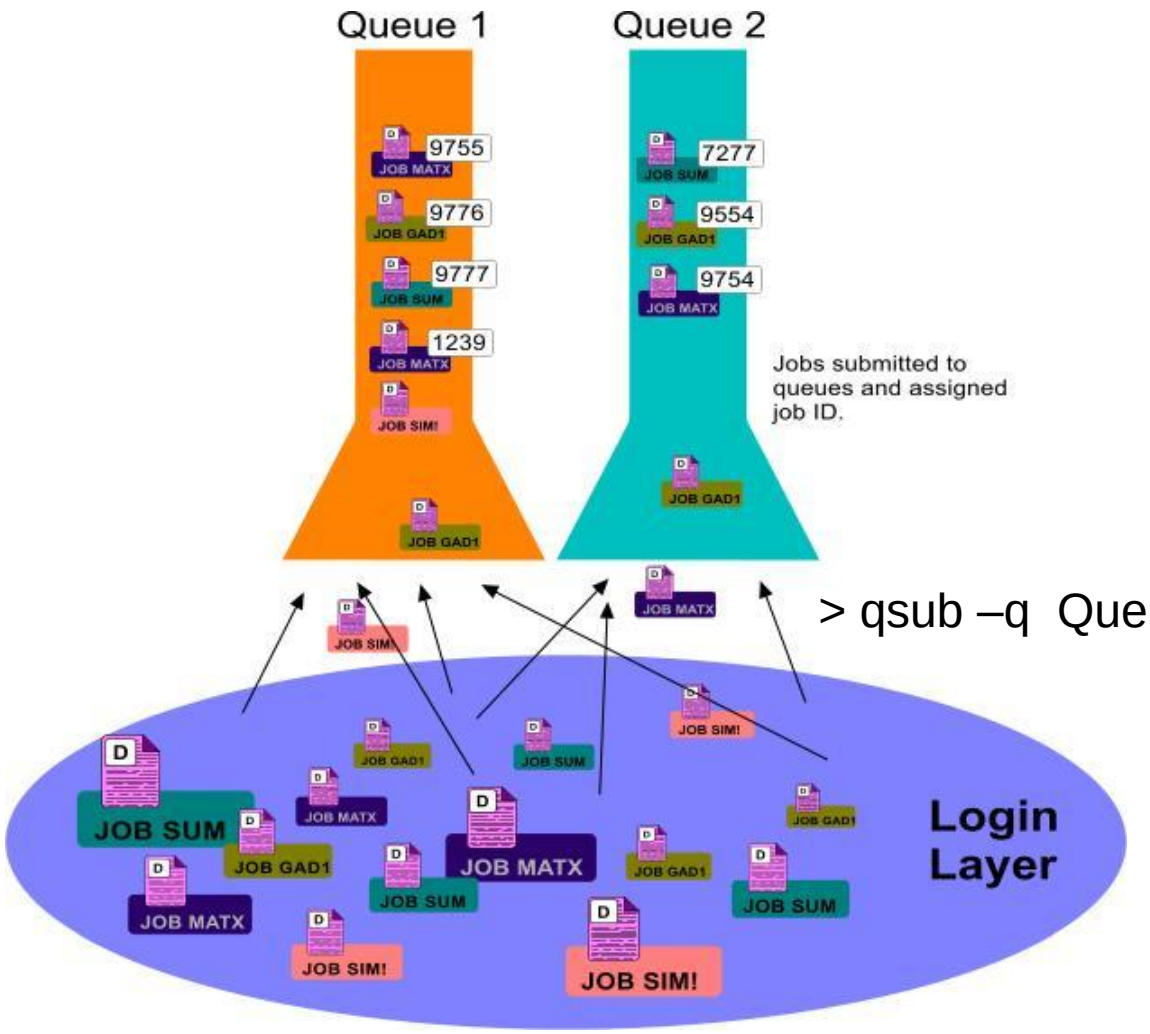
Executable and Jobscript setup in Login Layer



SCAMA Portsmouth

SEPnet South East Physics Network

Jobs submitted to the queues

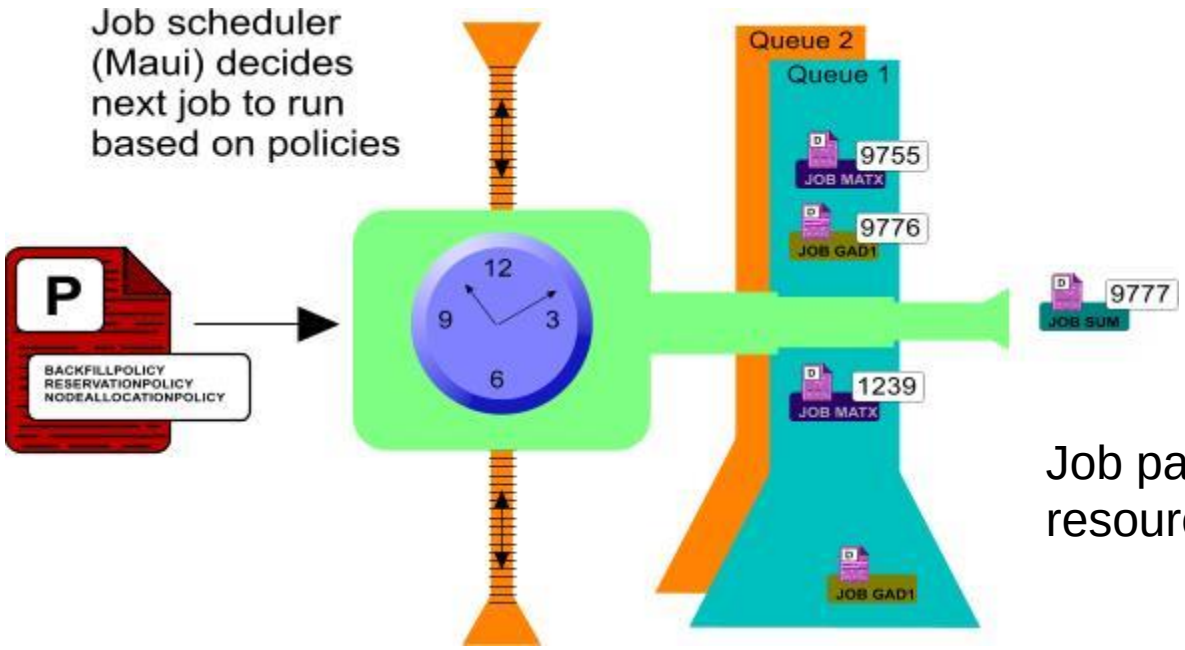


> qsub –q Queue1 run_script.sh

SCAMA Portsmouth

SEPnet South East Physics Network

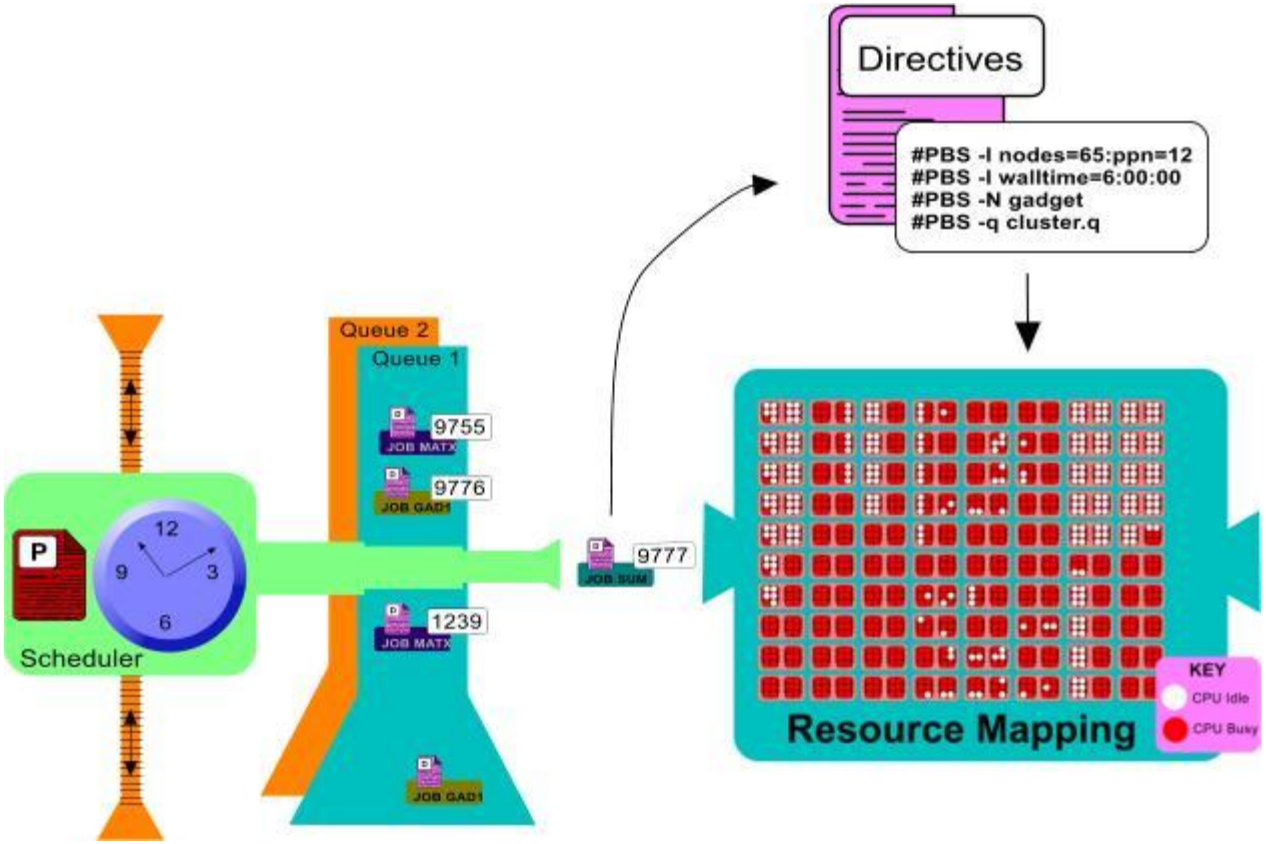
Scheduler prioritises and deems a job ready to run.



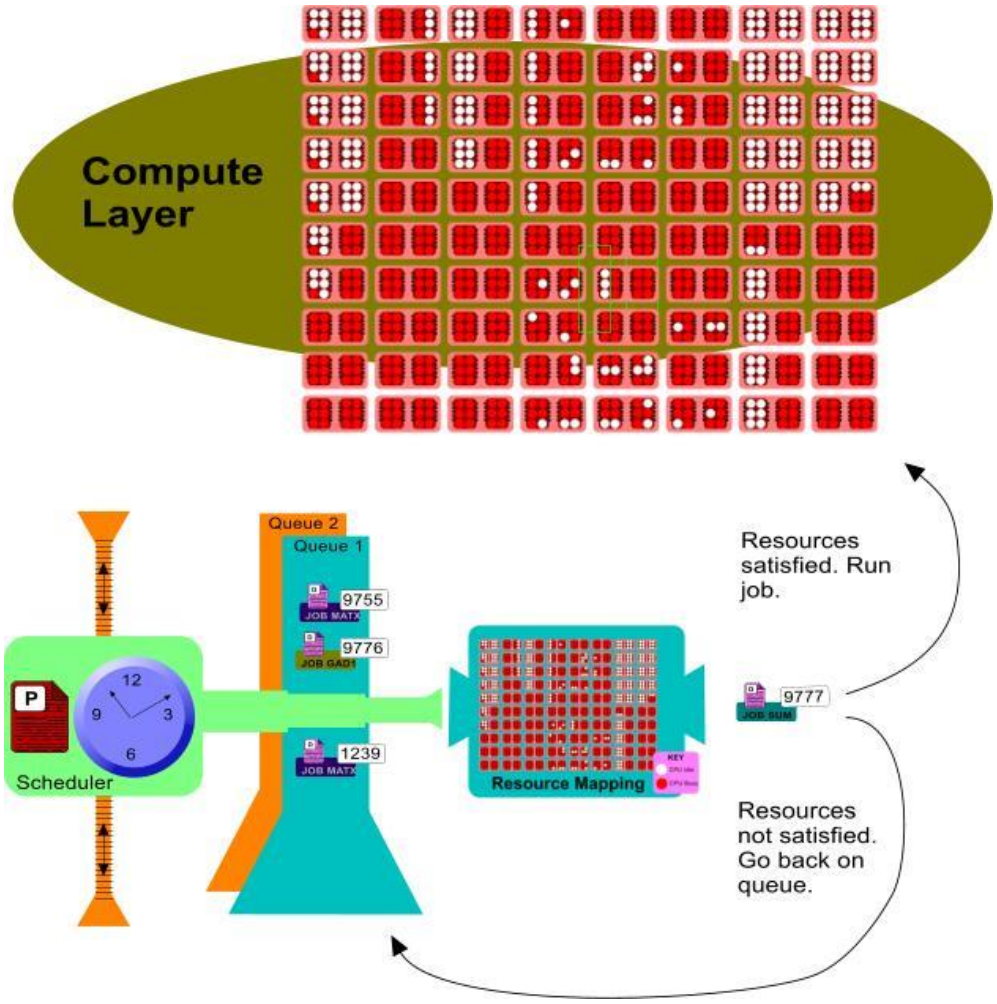
Job passed to resource manager

SCAMA Portsmouth ICG Portsmouth ALCES SOFTWARE University of Portsmouth SEPnet South East Physics Network

Resource manager (Torque) checks for available resources.



Job either runs in the compute pool or returns to queue



SCAMA Portsmouth

SEPnet South East Physics Network

So long …. and thanks for all the ~~fish~~ Coffee !!

C: Hitch Hikers Guide to the Galaxy

SCAMA Portsmouth ICG Portsmouth ALCES SOFTWARE University of Portsmouth SEPnet South East Physics Network