High Performance Computer Hardware

G Burton – ICG – May18 – v1.1

Overview

- What is an HPC
- Fundamentals of executing a program
- Commodity clusters
- Shared Storage
- Interconnects and Networks
- Login layer
- GPU's and Intel Phi's

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

What is a High Performance Computer ?

- This depends on who you are and what you want to do …
  - Are you a graphics designer that needs high end graphics on a work station.
  - Are you a financial trader that requires super fast networking.
  - Are you my Mum who thinks her iPhone is the most powerful computer in the world.
  - Are you weather forecaster that needs to run a weather simulation.
  - Are you a Scientist wanting to experiment with a quantum computer.
  - Are you a banker requiring mission critical transaction guarantee.
  - Are you mining for bit coins.
- We need to narrow things down

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Specifically

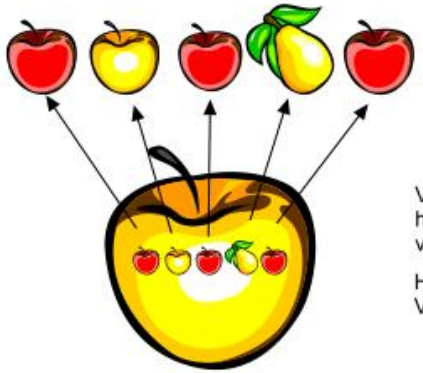
High Performance Computing (HPC) Hardware for Scientific Numerical Analyses

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Demystifying the Techno Babble

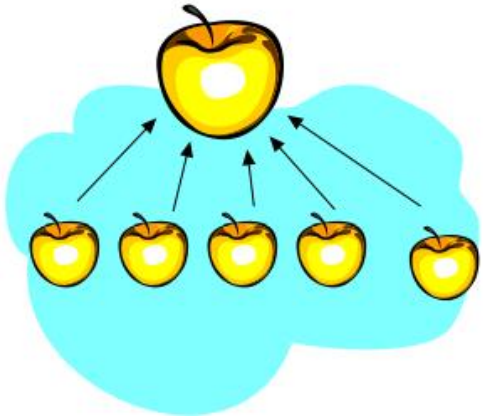
Virtualisation - One machine appears as many



Virtualisation maximises hardware resources through virtual machines (VM's)

Host machine may support VM's with different OS's

Clouds - Many machines appear as one



Clouds are all about scalability

Being able to switch (automatically) resources in or out.

Sometimes called SOA
Service Orientated Architecture.
IaaS - Infrastructure as a Service
PaaS - Platform as a Service
SaaS - Software as a Service

Clouds use virtualisation.

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Demystifying the Techno Babble (2)

Clusters - Many homogeneous machines at one location

Grid - Many heterogeneous machines at disparate locations

SCAMA Portsmouth

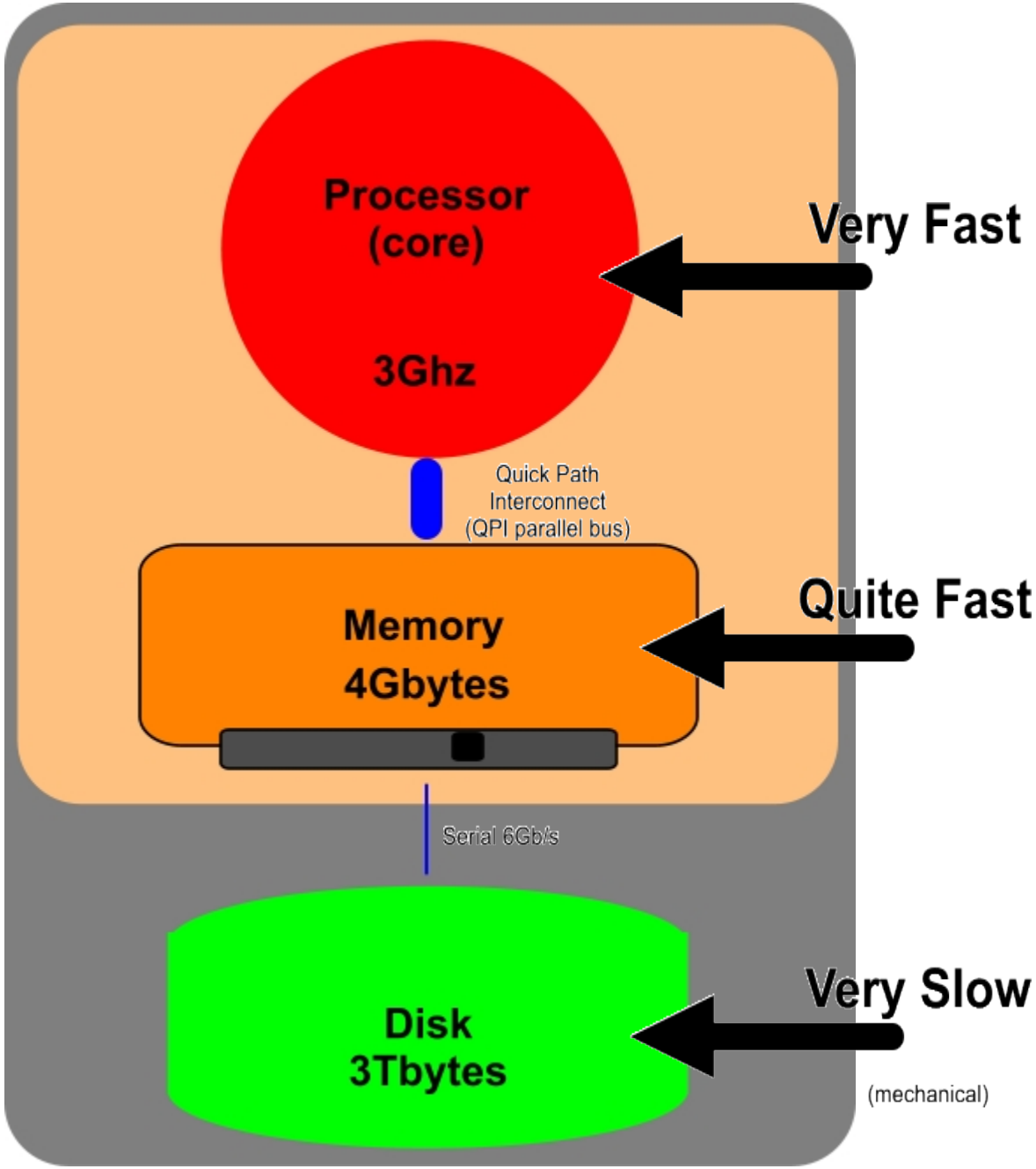
UNIVERSITY OF PORTSMOUTH

Fundamentals of executing a program

SCAMA Portsmouth

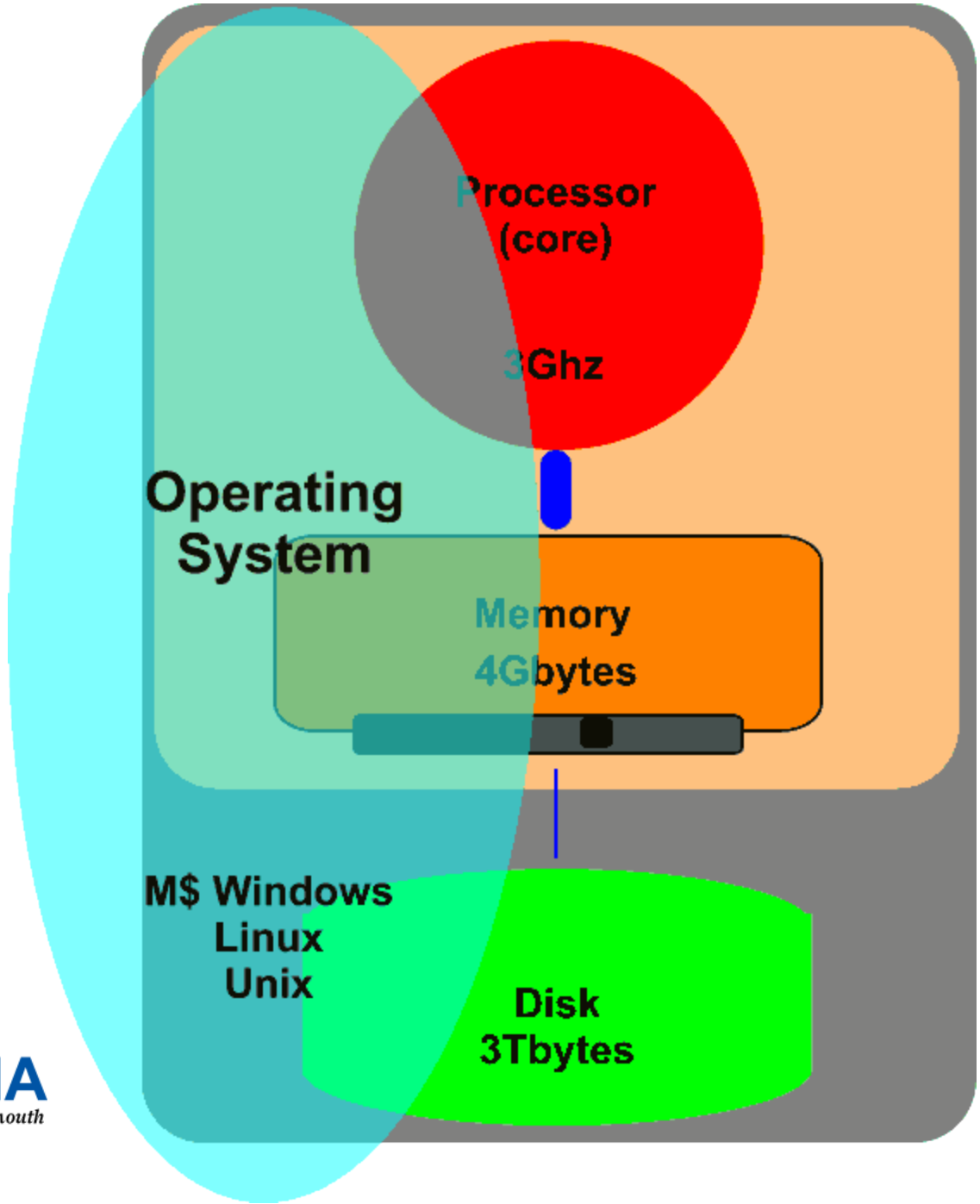
UNIVERSITY OF PORTSMOUTH

Back to Basics



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

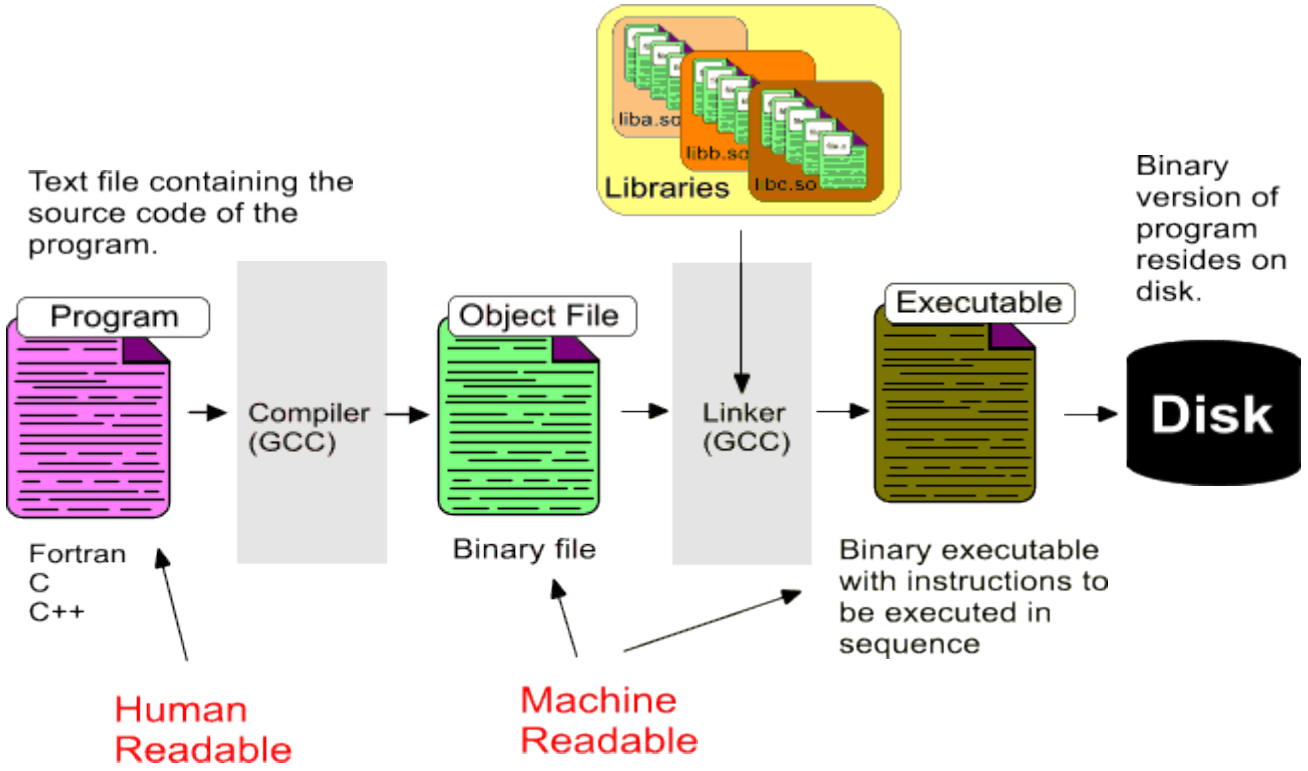


Back to Basics

SCAMA Portsmouth

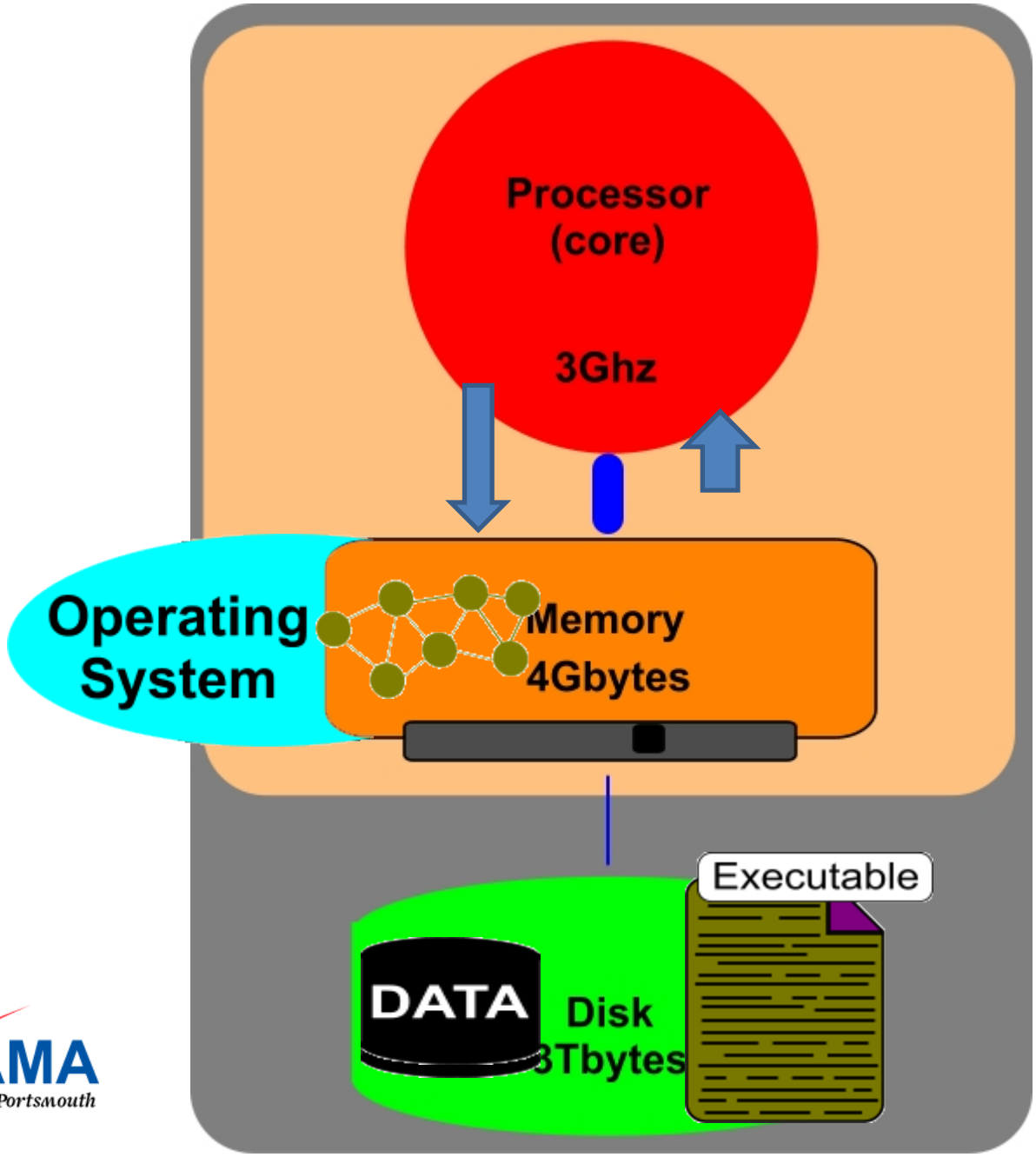
UNIVERSITY OF PORTSMOUTH

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



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH



Paging

Swapping

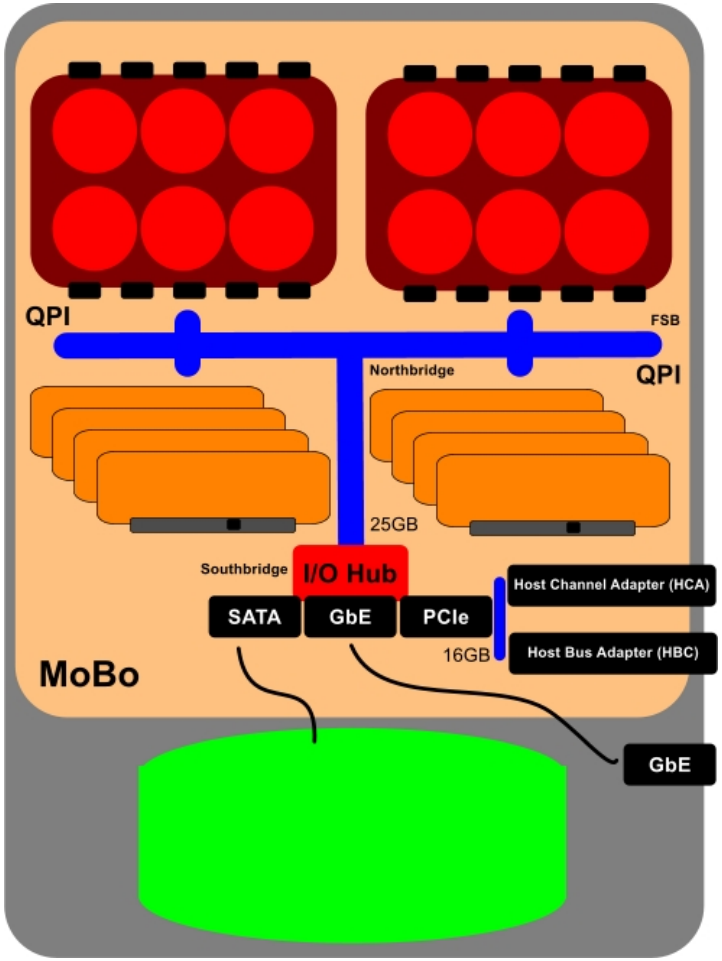
OOM killer

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

… these days much more packed into the same space … but basically the same!

Its doesn't matter how many cores the standard executable will execute sequentially one instruction at a time.



Programming Frameworks

- Shared memory (cooperating threads – or multi-threading)
  - OpenMP
  - POSIX Threads
  - Cilk
  - Threaded Building Blocks
  - etc, etc
- Distributed memory (cooperating processes)
  - MPI (PVM, etc)
  - Co-array Fortran, UPC, etc
  - Global Array Toolkit (etc)
  - Adlib and HPspmd?!
  - etc, etc

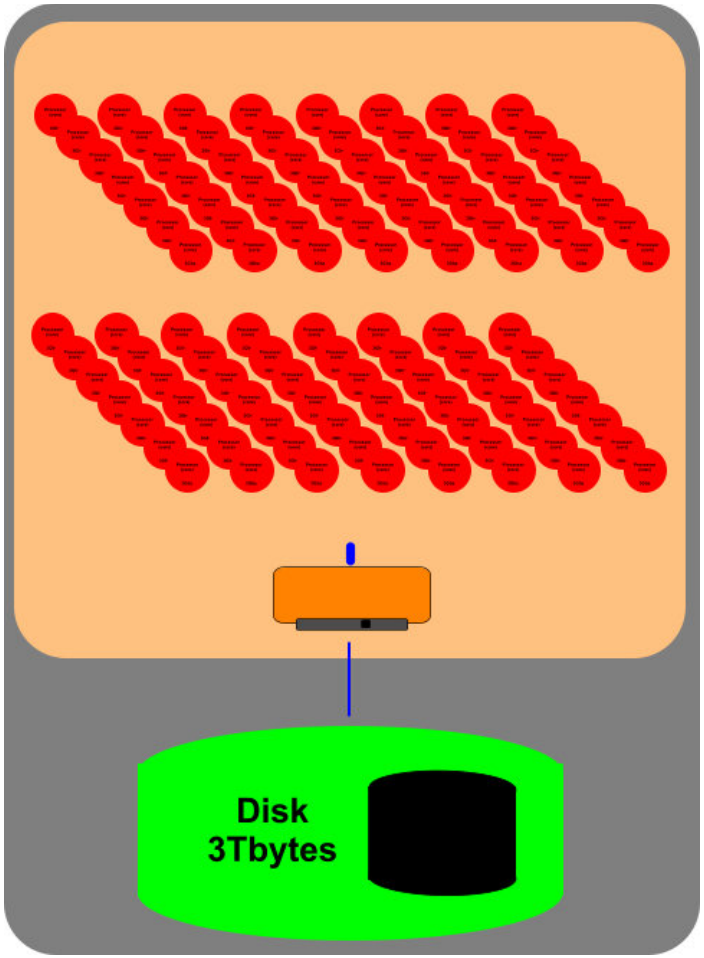
How effective…..
Depends on the nature of what you are trying to solve.

Not always Possible !!

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

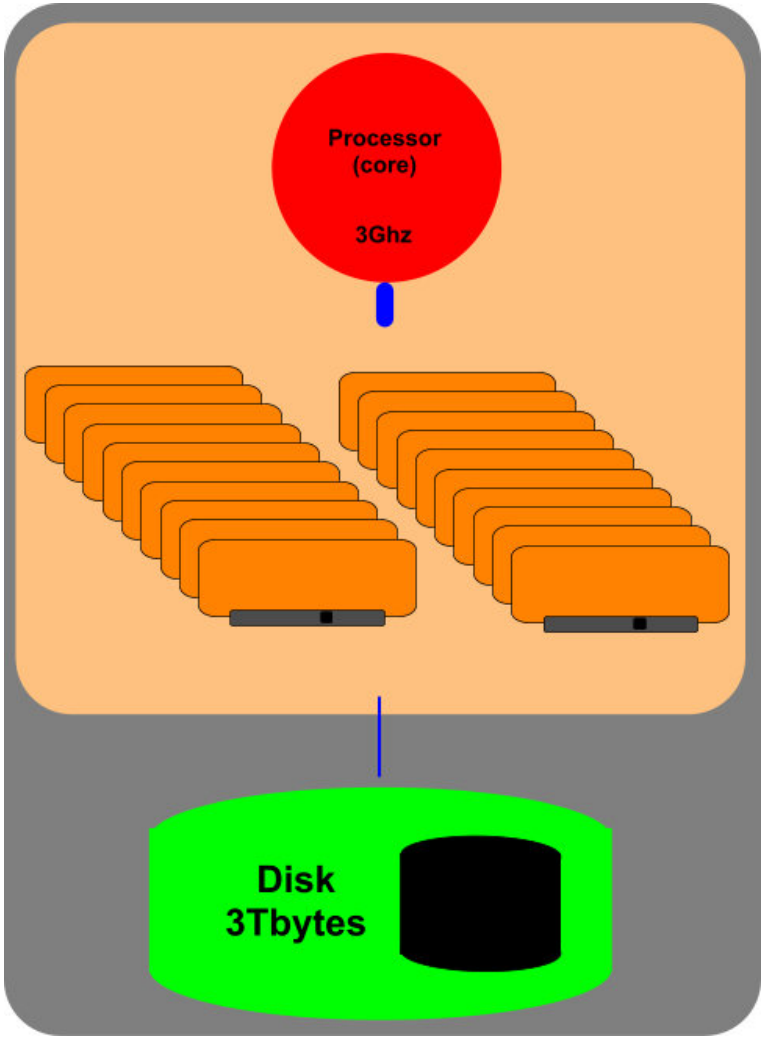
In Reality – A program may require many 100's of cores



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

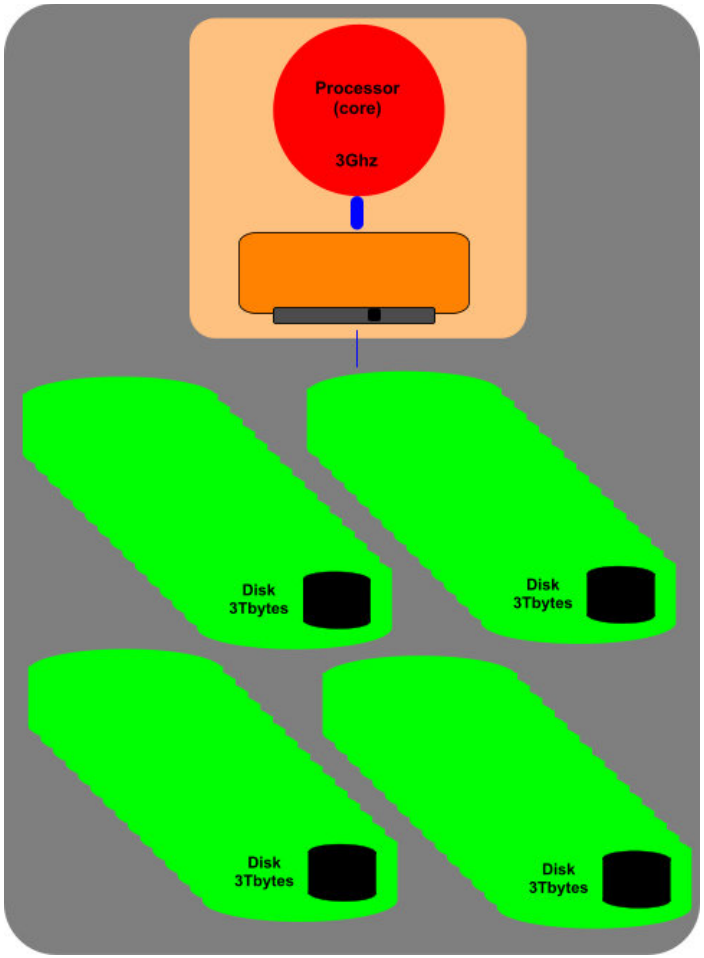
… or many Gbytes of memory



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

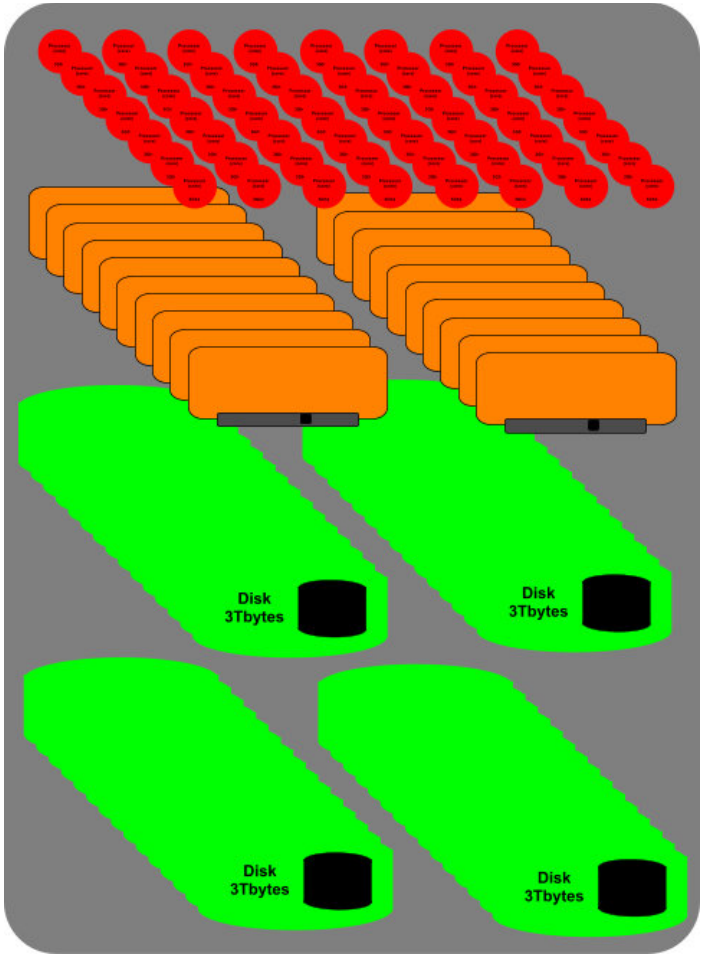
….or a lot of disk space



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

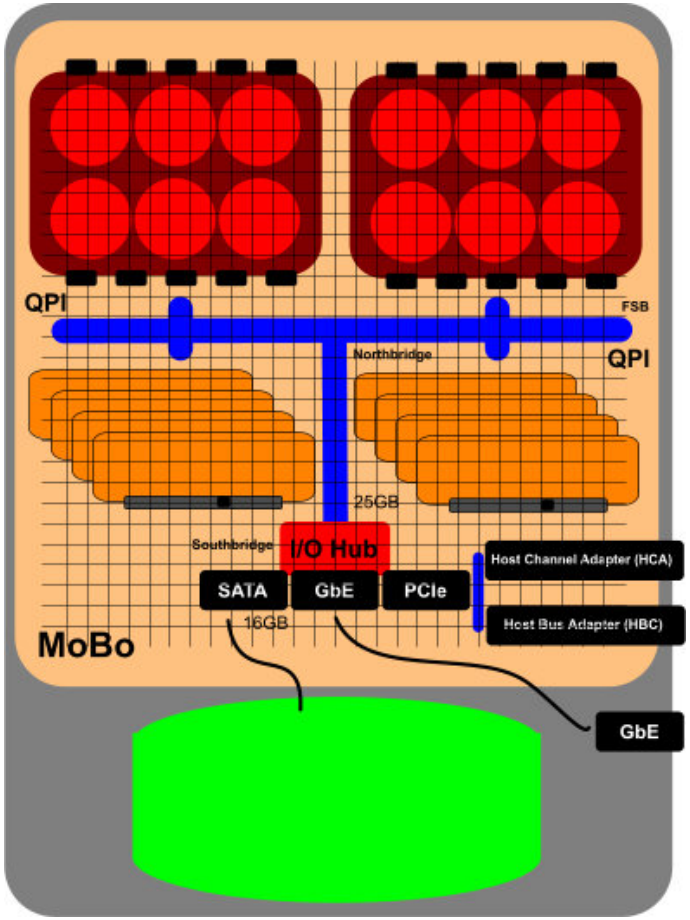
….or a combination of all.



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

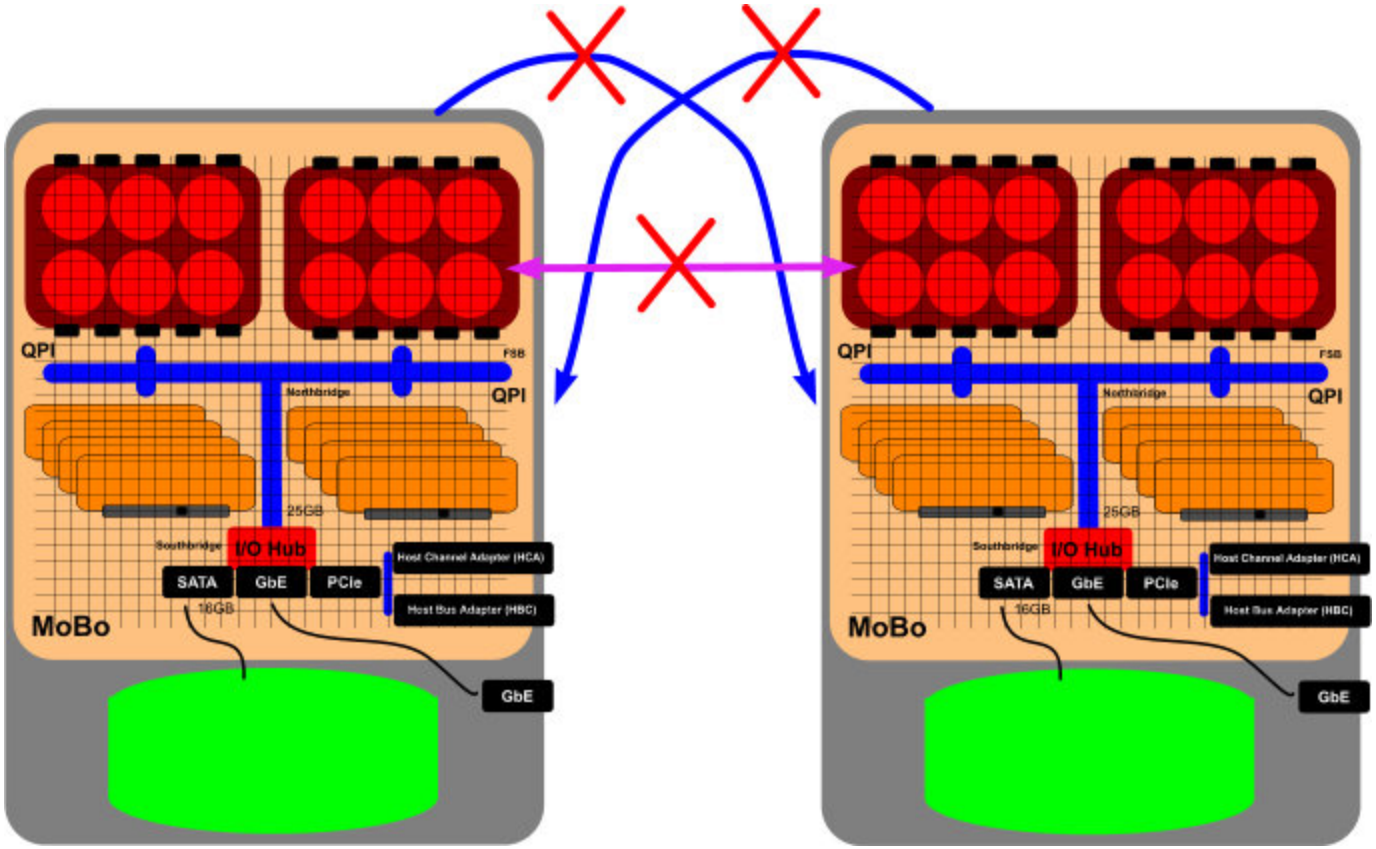
Internal Communications Possible



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

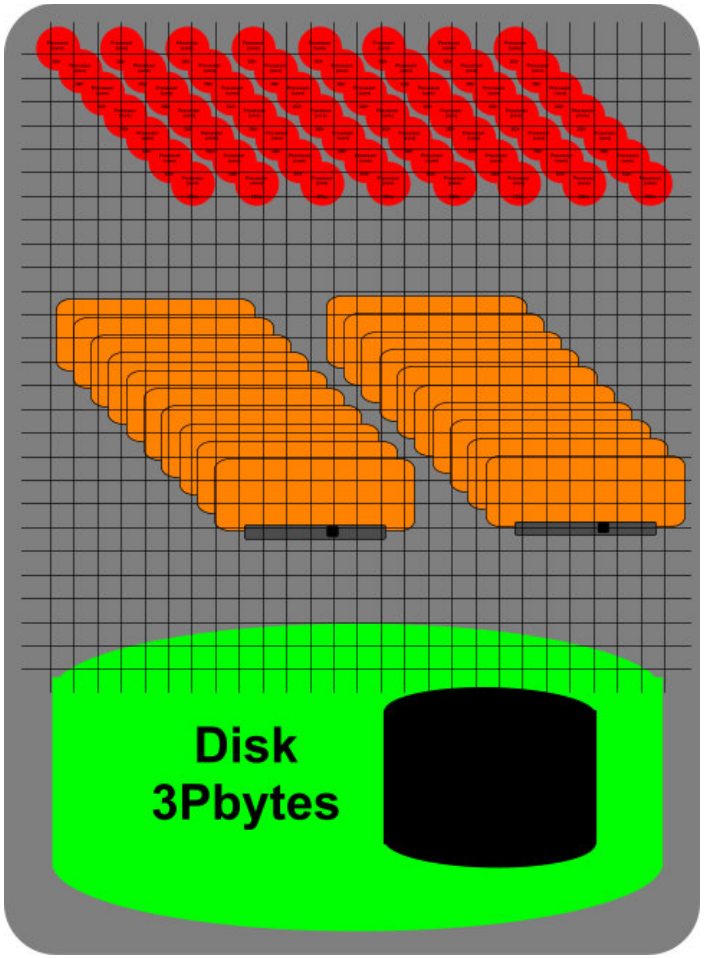
No direct comms between Pc's



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

We could build a very expensive machine ……..

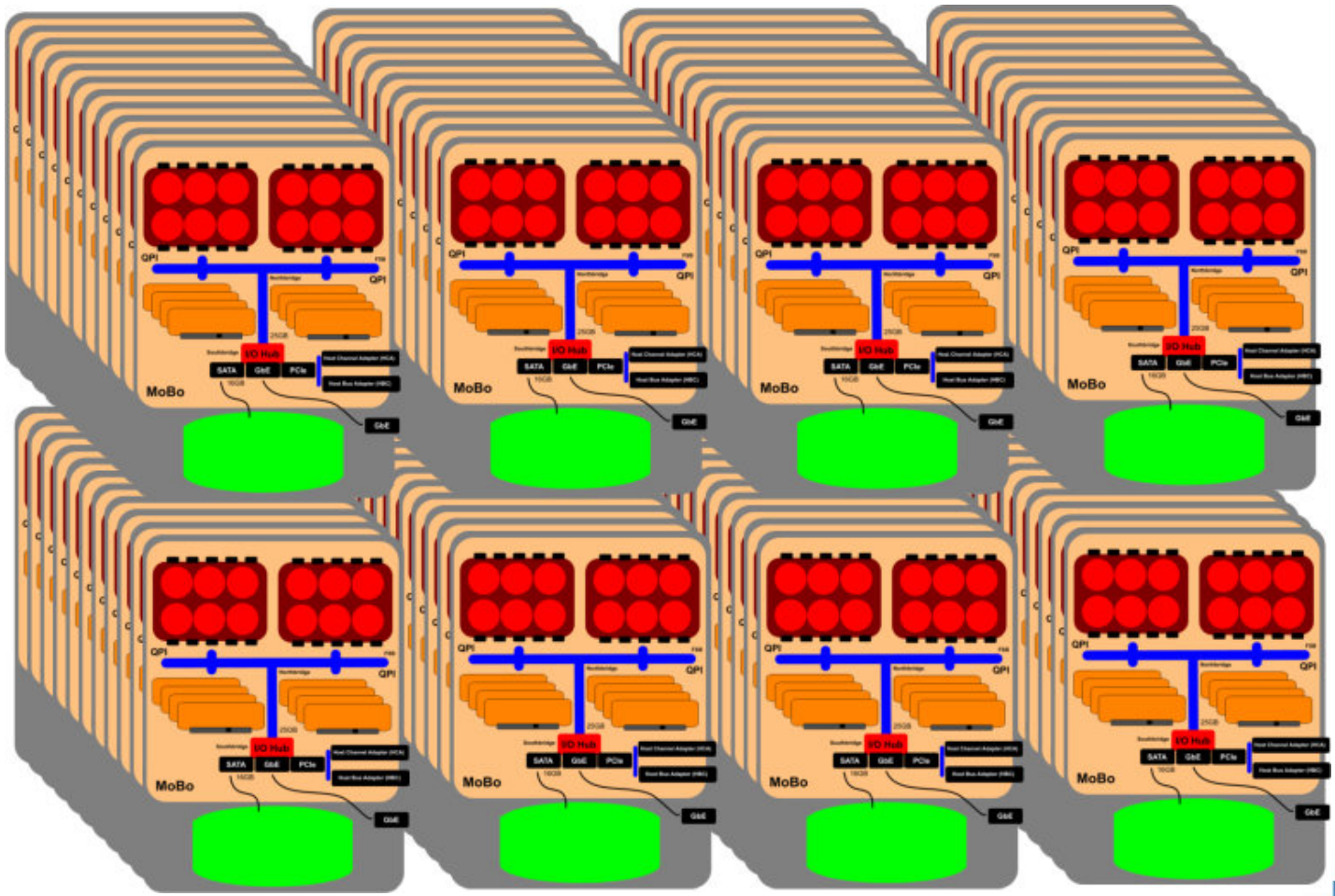


Referred to as a Shared Memory Machine.

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

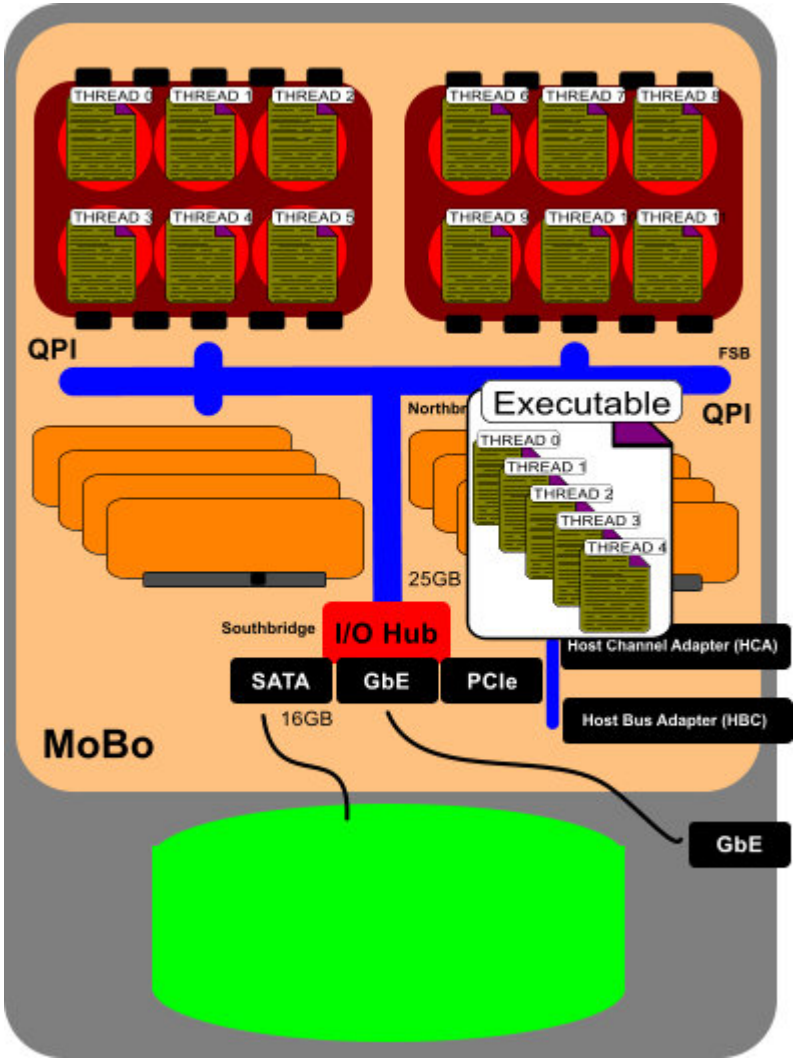
Commodity cluster using high end PCs



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Multithreading takes advantage of onboard cores…(but not across PC's)

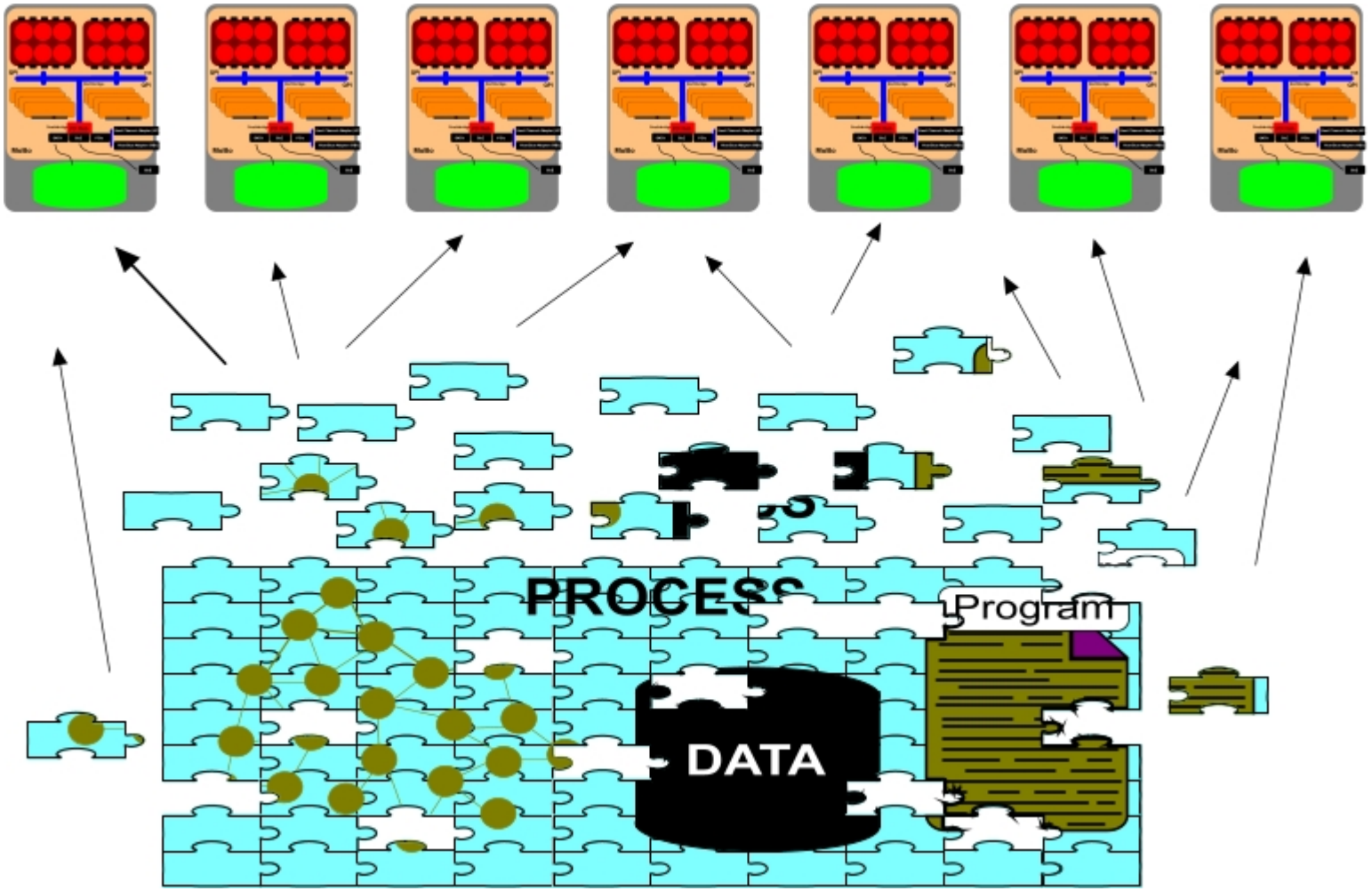


Limited to number Of cores on Mobo.

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Distributed Memory works between PC's



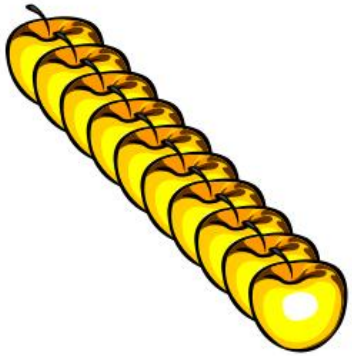
SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Summary - HPC alternatives for Scientific Computing

HPC through
Specialised Hardware

HPC using
Commodity Hardware



Examples:-

Highend Graphics Workstation
for 3D modelling.

Shared Memory or Symmetric
Multiprocessor (SMP) machines
like COSMOS & Universe

Vector Processors using, for
example, expensive Cray
or Convex Mainframes.

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

High Performance Computer
=
Commodity Cluster

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

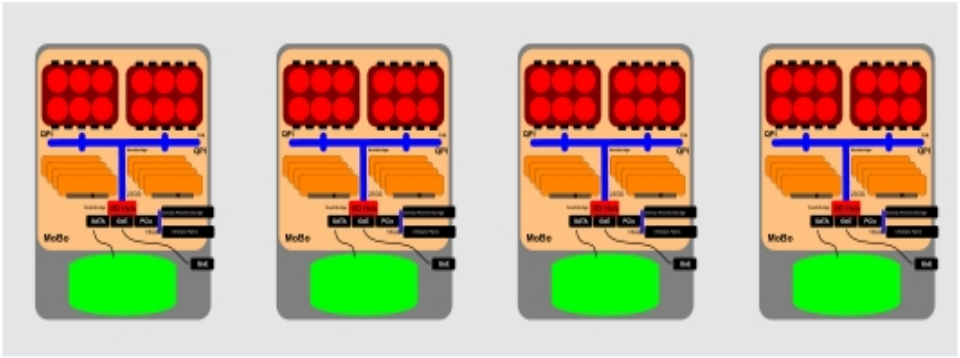
Commodity Clusters

- Made from commodity (off-the-shelf) components.
- Consequently (relatively) cheap.
- Usually Linux based
- High availability storage (no single point of failure)
- Generic compute pool (cloned servers that can easily be replaced).

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Dell 6100 Front View



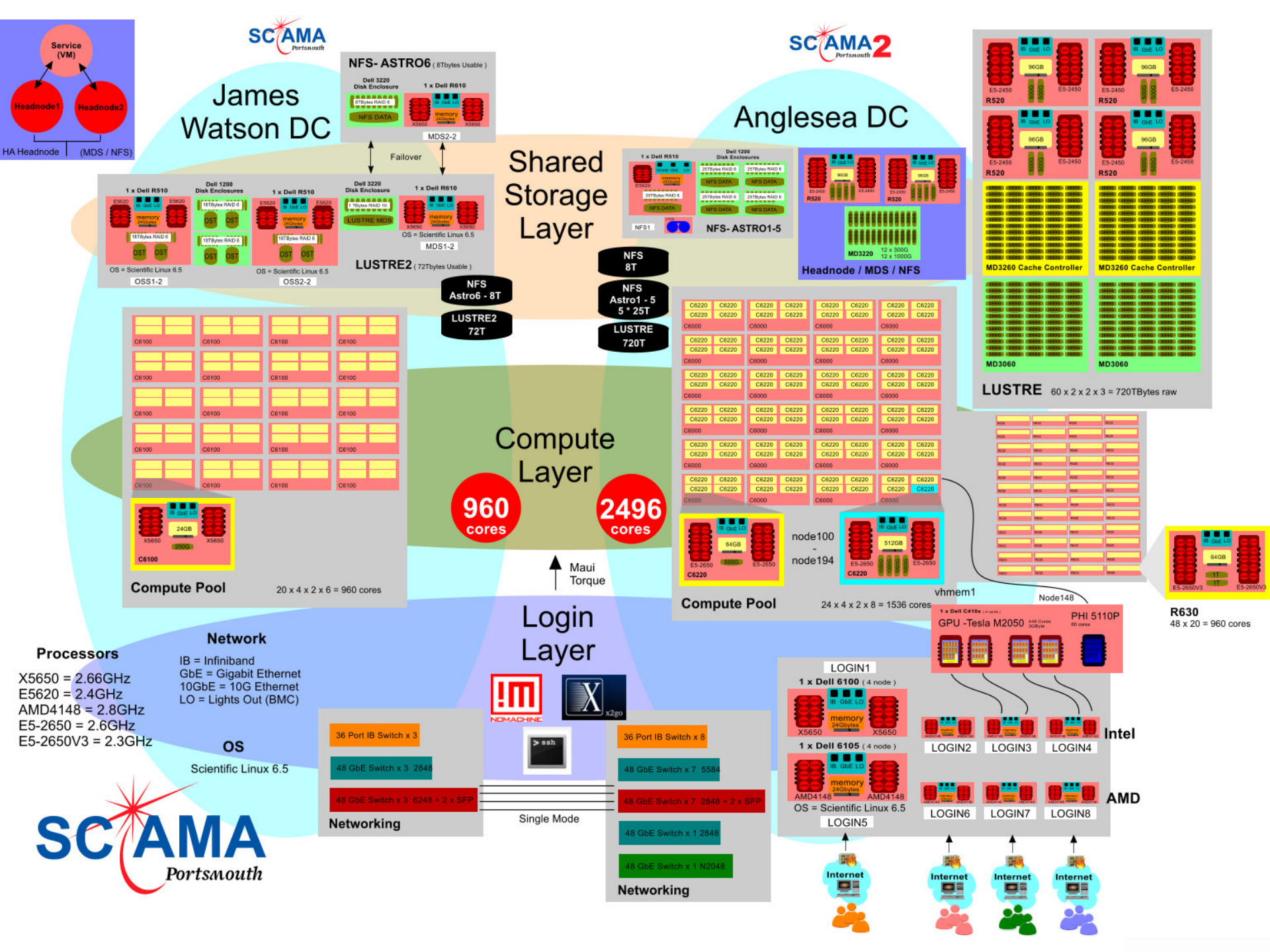
Google
Amazon
Ebay

Dell 6100 Rear View

High density - Google / Amazon

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH


Service (VM)
Headnode1
Headnode2
HA Headnode
(MDS / NFS)
SCAMA Portsmouth
James Watson DC
NFS- ASTRO6 (8Tbytes Usable)
Dell 3220 Disk Enclosure
1 x Dell R610
NFS DATA
MDS2-2
Failover
1 x Dell R510
Dell 1200 Disk Enclosures
OST
OS = Scientific Linux 6.5
OSS1-2
OSS2-2
LUSTRE MDS
LUSTRE2 (72Tbytes Usable)
Shared Storage Layer
NFS Astro6 - 8T
LUSTRE2 72T
NFS 8T
NFS Astro1 - 5 5 * 25T
LUSTRE 720T
NFS-ASTRO1-5
SCAMA2 Portsmouth
Anglesea DC
Headnode / MDS / NFS
MD3220
12 x 300G
12 x 1000G
R520
MD3260 Cache Controller
MD3060
LUSTRE 60 x 2 x 2 x 3 = 720TBytes raw
Compute Layer
960 cores
2496 cores
C6100
C6220
C6000
Compute Pool
20 x 4 x 2 x 6 = 960 cores
24 x 4 x 2 x 8 = 1536 cores
node100 - node194
vhmem1
Node148
GPU -Tesla M2050
PHI 5110P
R630
48 x 20 = 960 cores
Maui Torque
Login Layer
NOMACHINE
x2go
ssh
LOGIN1
LOGIN2
LOGIN3
LOGIN4
LOGIN5
LOGIN6
LOGIN7
LOGIN8
1 x Dell 6100 (4 node)
1 x Dell 6105 (4 node)
Intel
AMD
OS = Scientific Linux 6.5
Internet
Processors
X5650 = 2.66GHz
E5620 = 2.4GHz
AMD4148 = 2.8GHz
E5-2650 = 2.6GHz
E5-2650V3 = 2.3GHz
Network
IB = Infiniband
GbE = Gigabit Ethernet
10GbE = 10G Ethernet
LO = Lights Out (BMC)
OS
Scientific Linux 6.5
36 Port IB Switch x 3
48 GbE Switch x 3 2848
48 GbE Switch x 3 6248 + 2 x SFP
Networking
Single Mode
36 Port IB Switch x 8
48 GbE Switch x 7 5584
48 GbE Switch x 7 2848 + 2 x SFP
48 GbE Switch x 1 2848
48 GbE Switch x 1 N2048
Networking
SCAMA Portsmouth

HPC Facts

- China's Sunway TaihuLight now number 1 in top 500 with 10.6 million Cores.
- 125 petaFLOPS (Floating Point Operations / Second). Sciama 10 teraFLOPS.
- China now has 202 in top 500 with US second with 143.
- Japan=35,Germany=20,France=18,UK=15
- Exascale expected 2020. Race between Japan, France, China, USA
- Uk – top 5 to do with weather, followed by Cambridge and Edinburgh m/cs

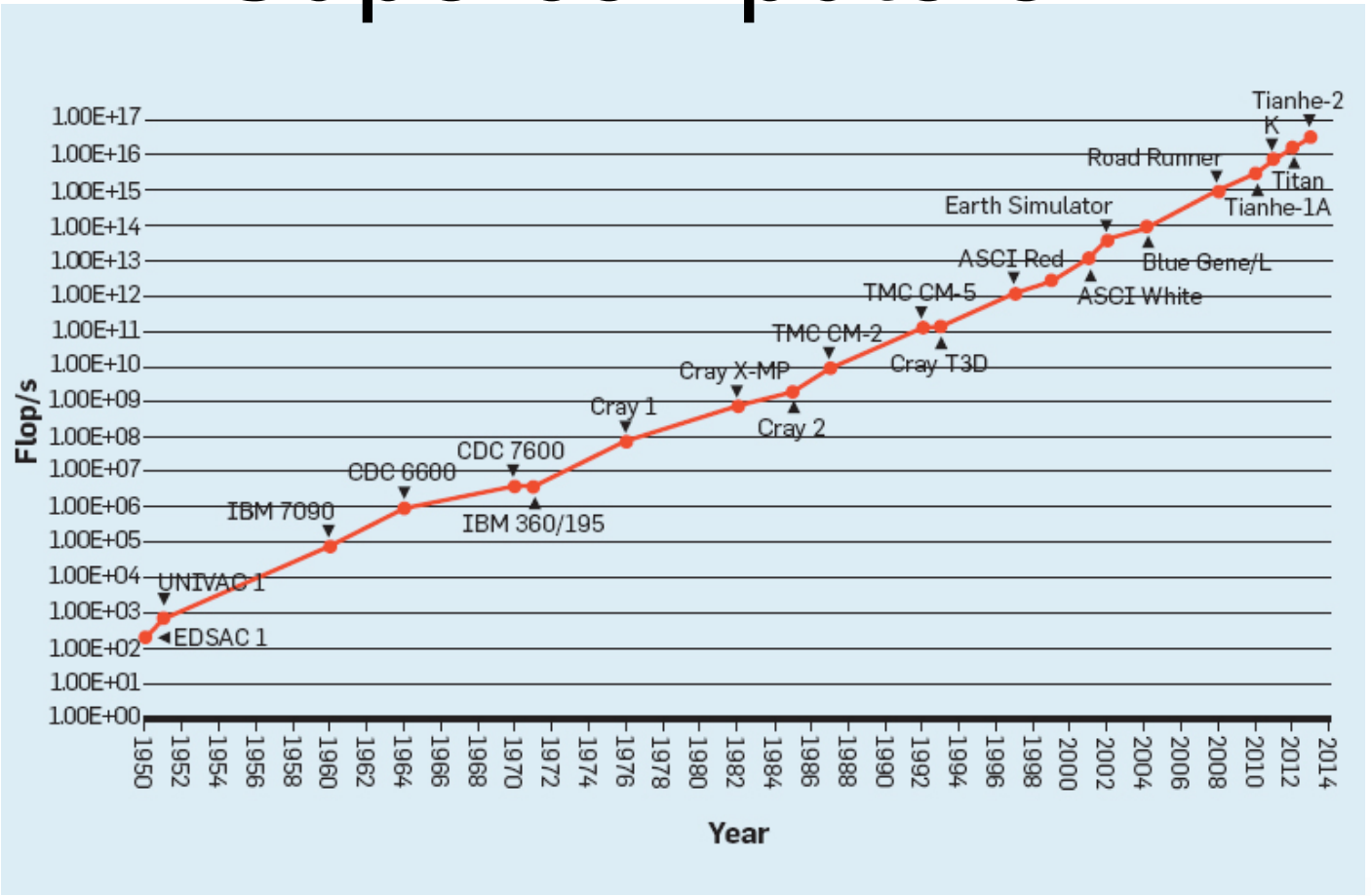
Computer performance

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

UNIVERSITY OF PORTSMOUTH

Moore's Law for Supercomputers



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

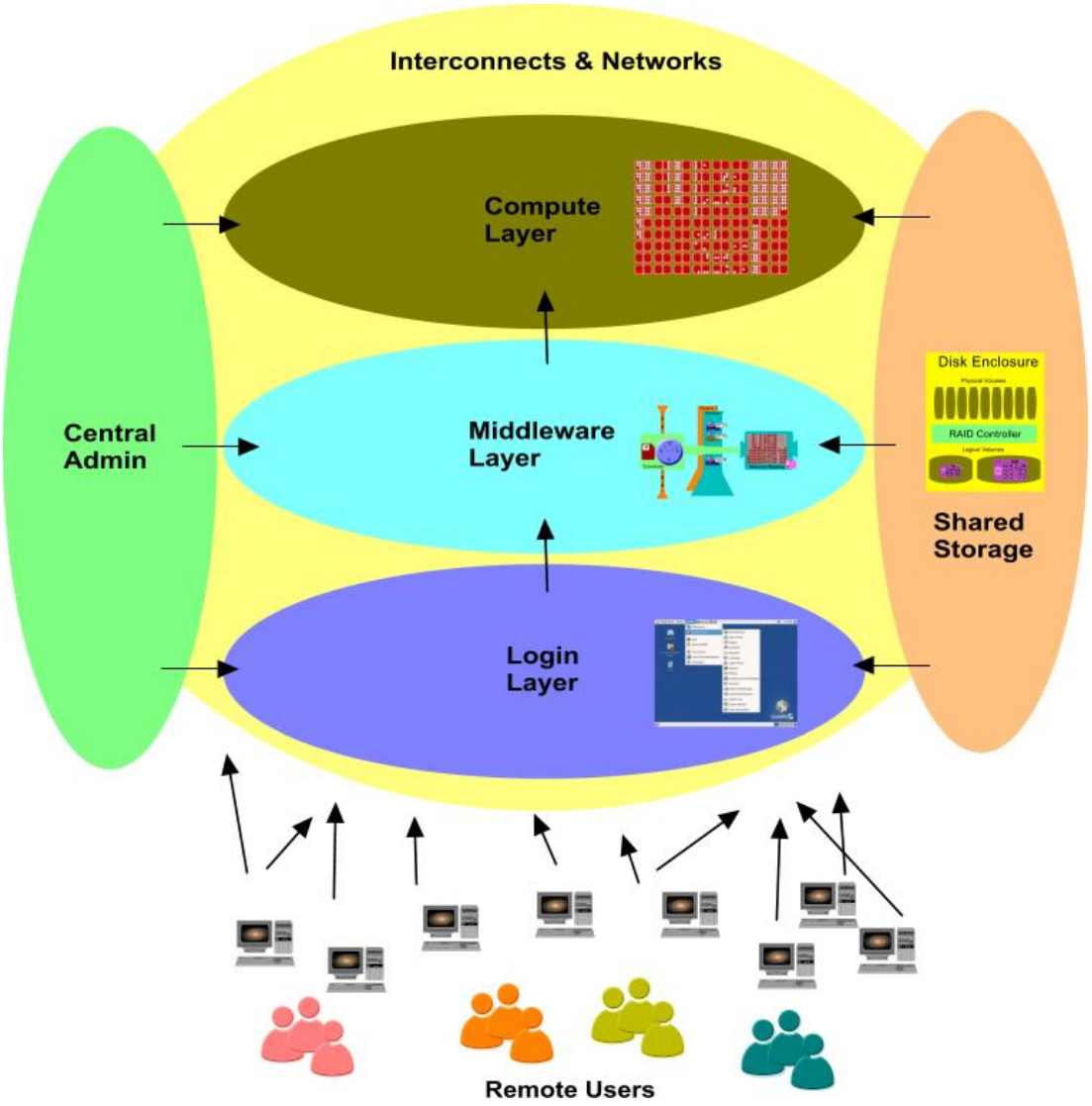
Common Misconception

- A super computer does not necessarily mean a program will run faster.
- Commodity processors typically slower than average desktop / laptop (2.6GHz vs 3.0GHz)
- Unless a program can be parallelised it may run slower.
- Disk access may also be slower (directly attached).
- However super computer much more stable for long runs.

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

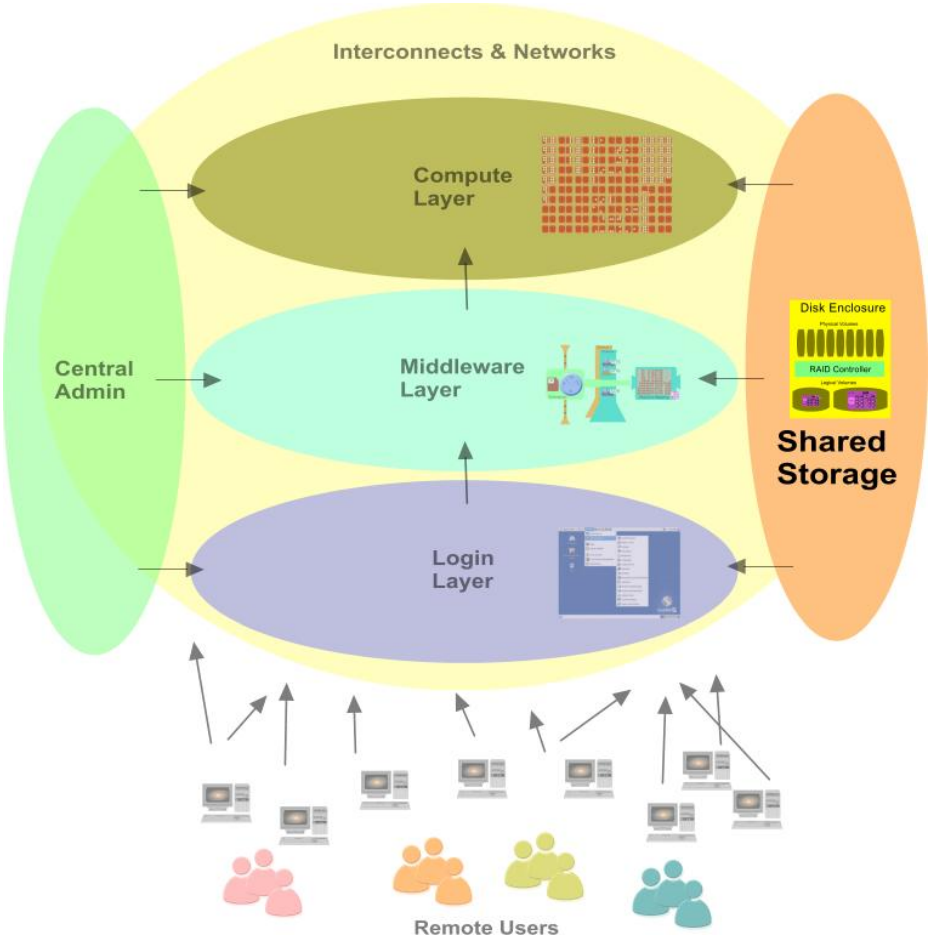
Cluster Concept



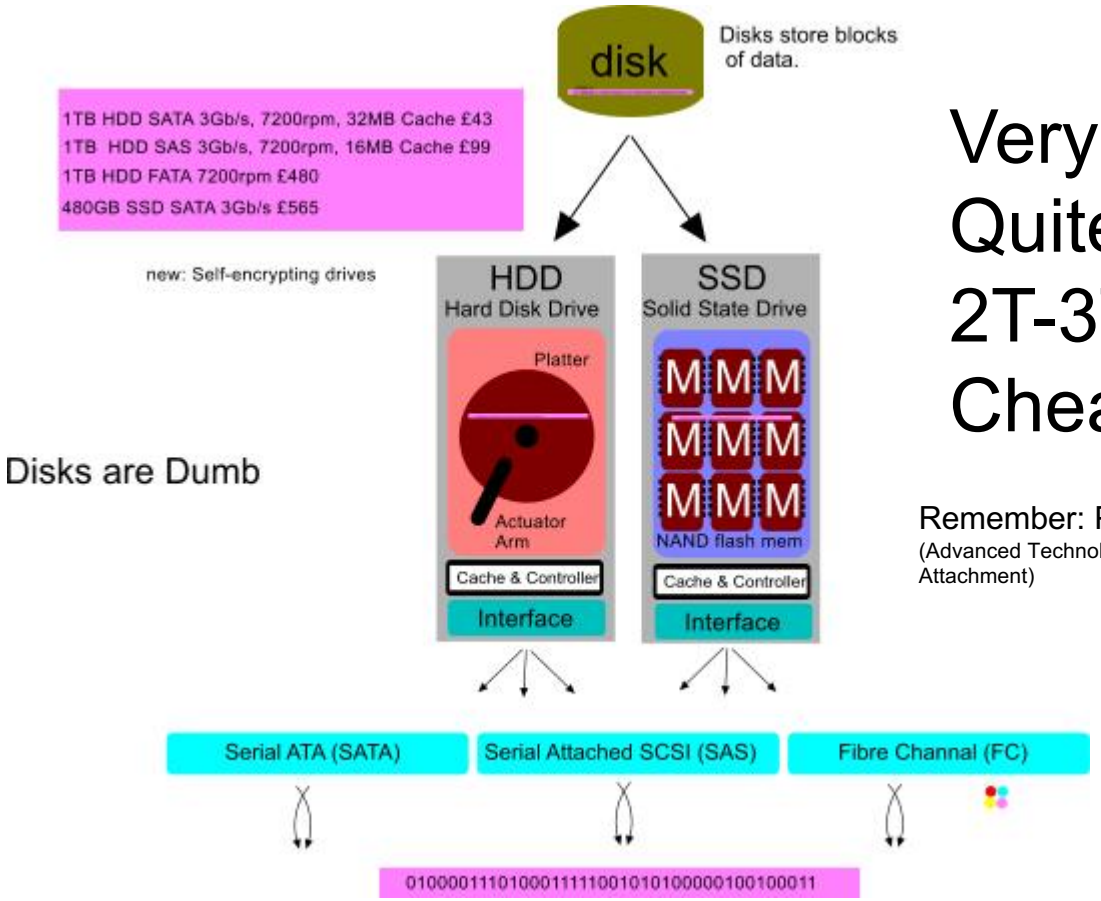
SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Shared Storage



Raw Disks are Dumb

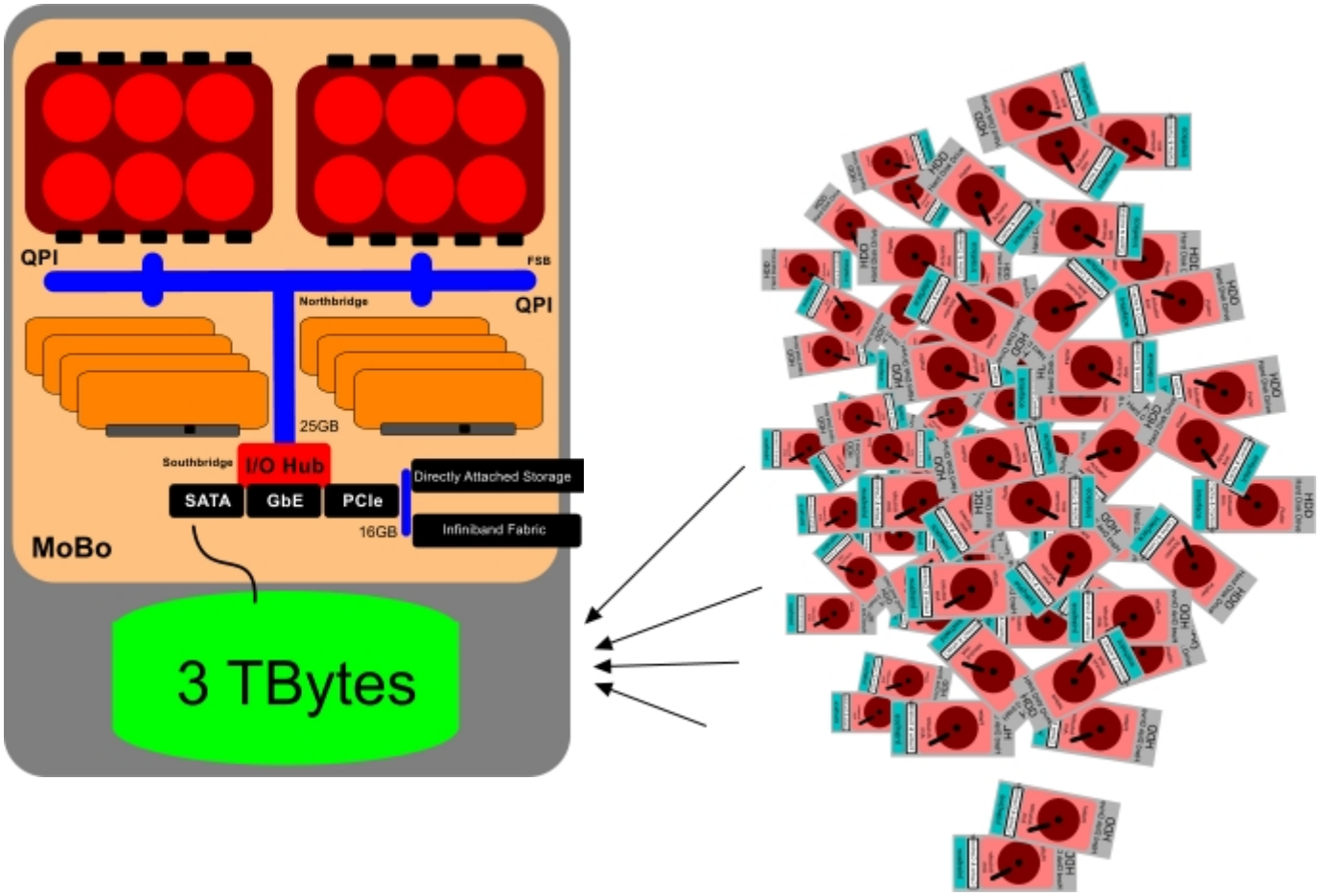


Very Slow
Quite bulky
2T-3TBytes
Cheap

Remember: PATA, IDE
(Advanced Technology Attachment)

SCAMA Portsmouth

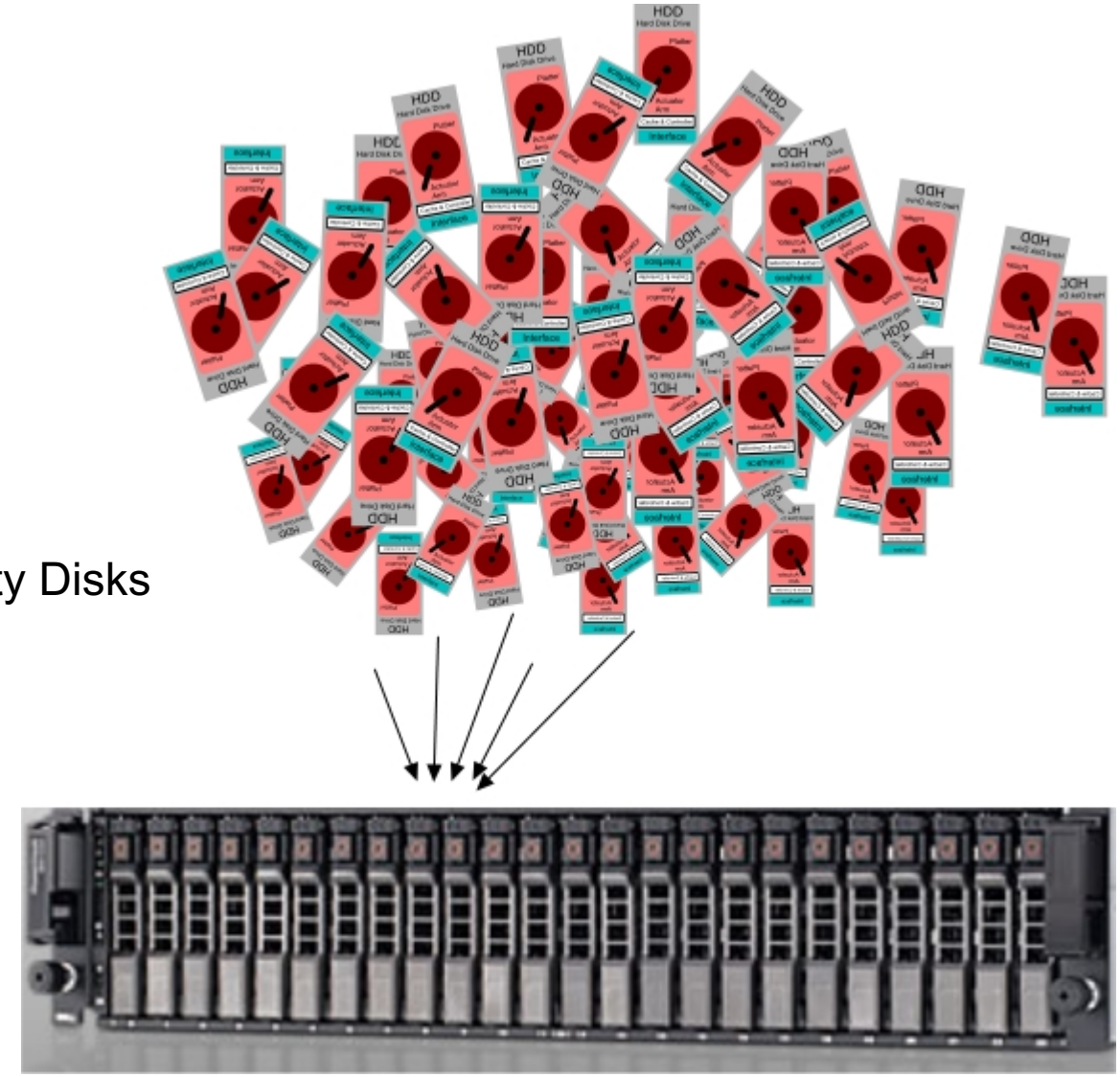
UNIVERSITY OF PORTSMOUTH



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Commodity Disks



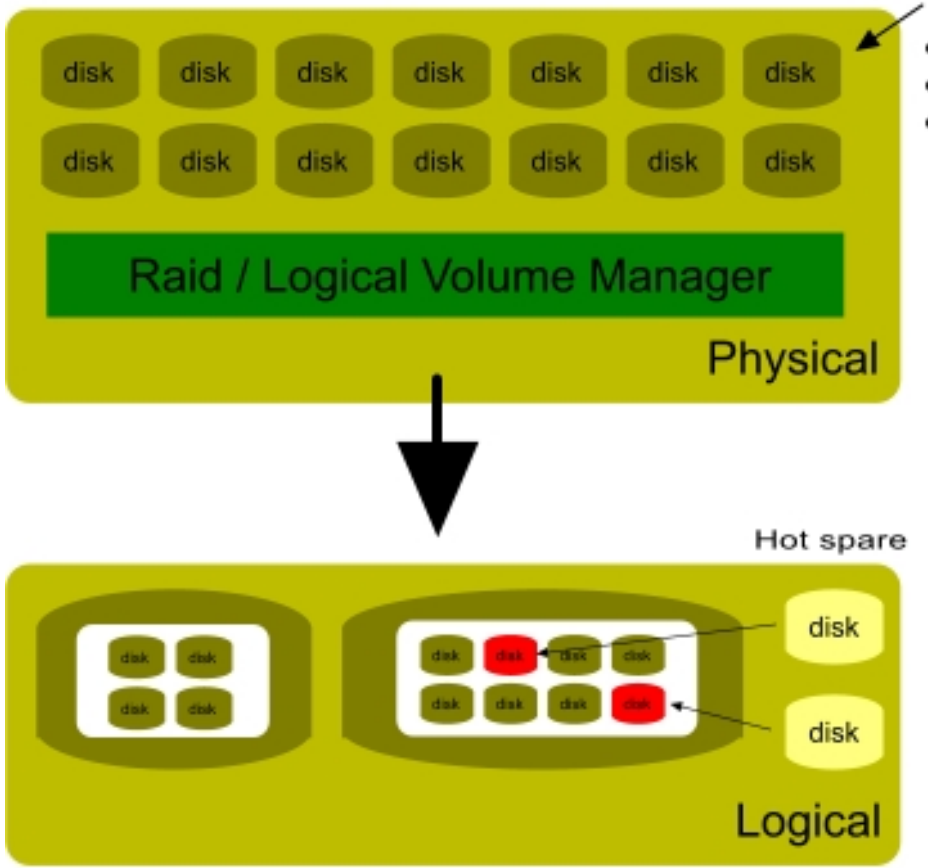
SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

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



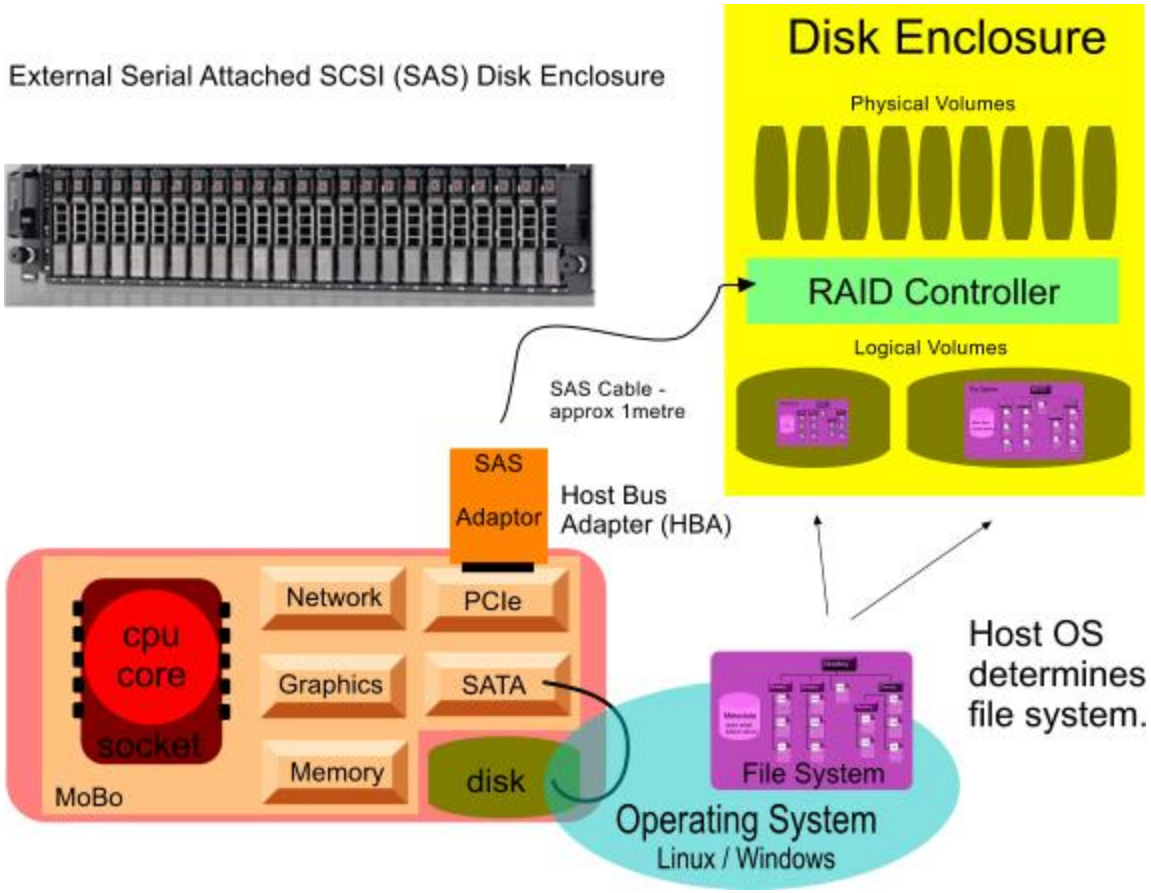
Logical Volumes

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

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Directly Attached Storage (DAS)

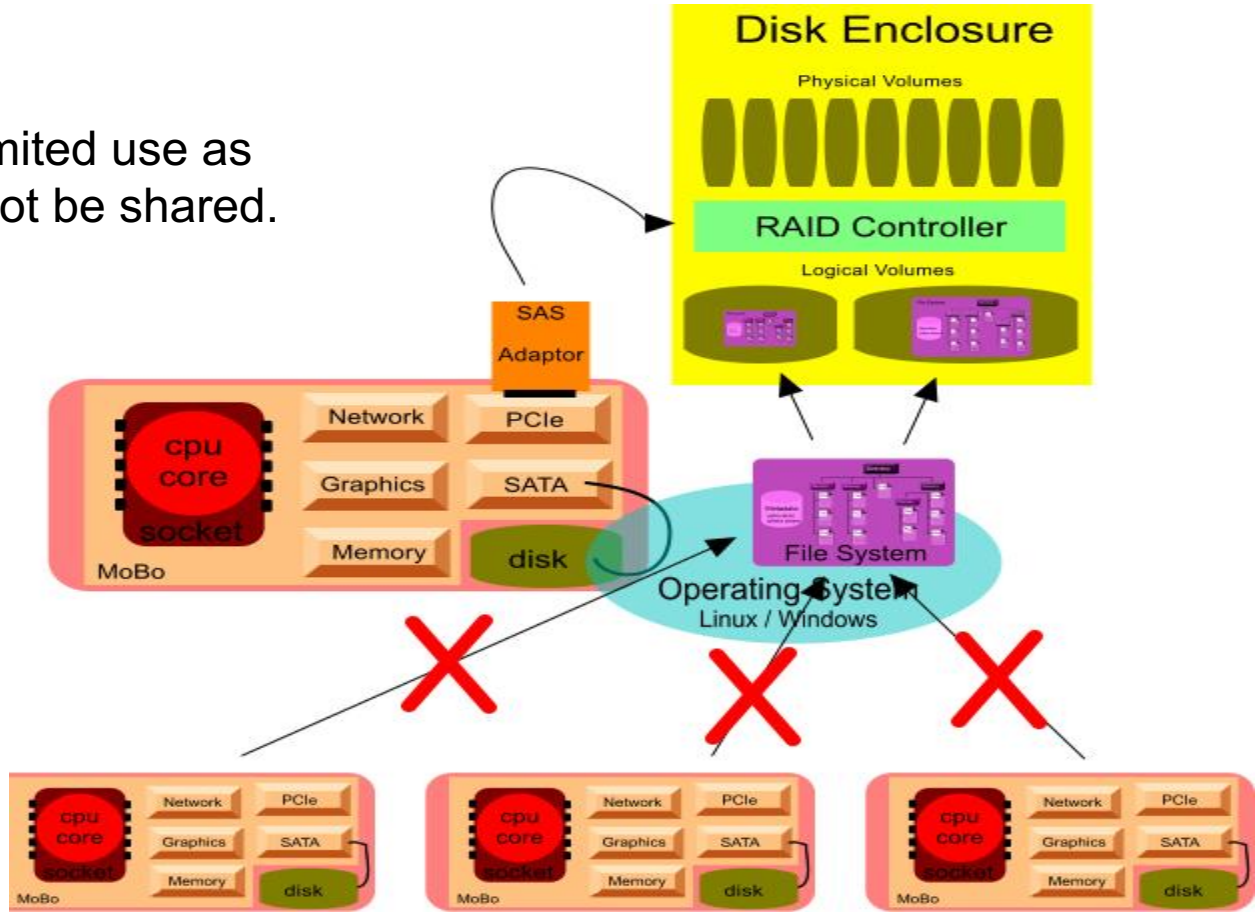


SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Directly Attached Storage (DAS)

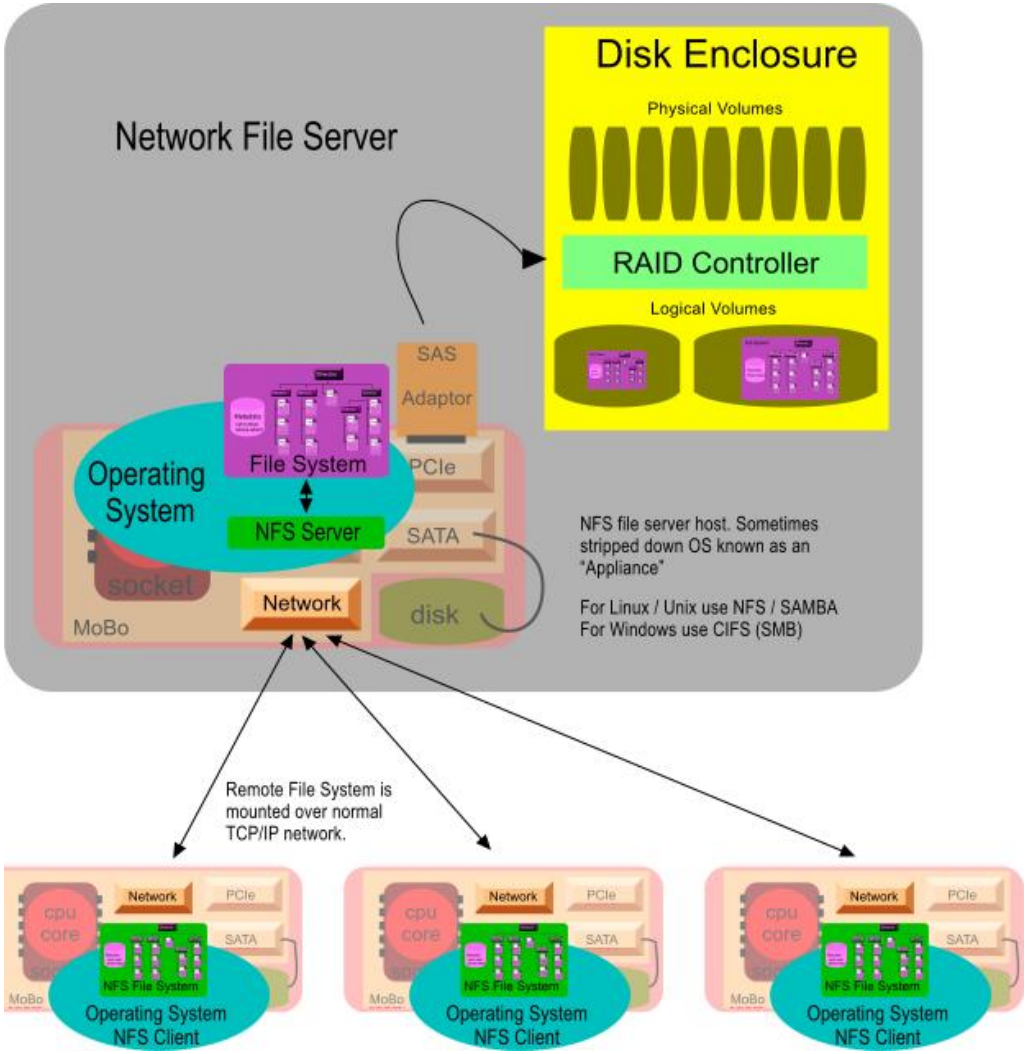
Of limited use as cannot be shared.



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

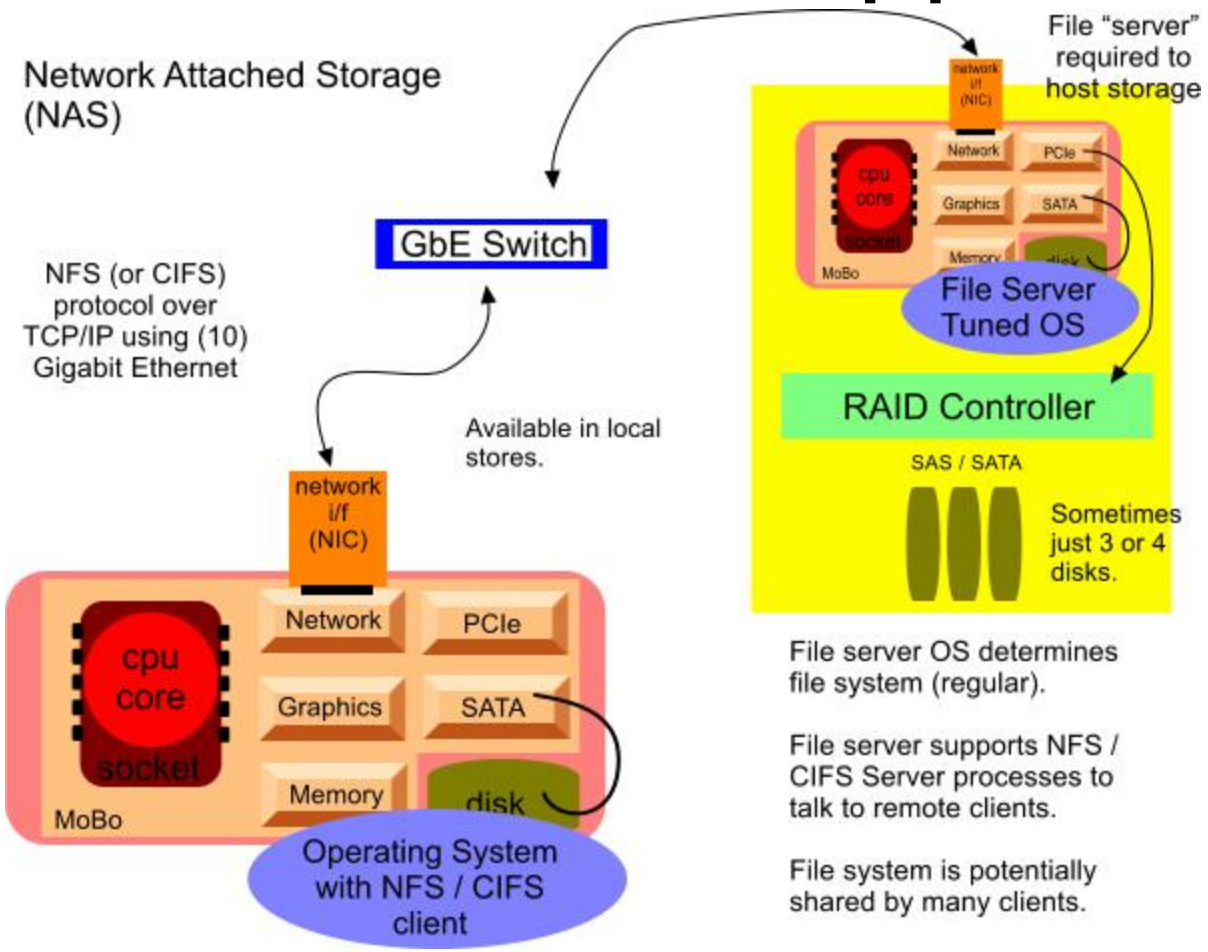
Network Attached Storage (NAS)



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

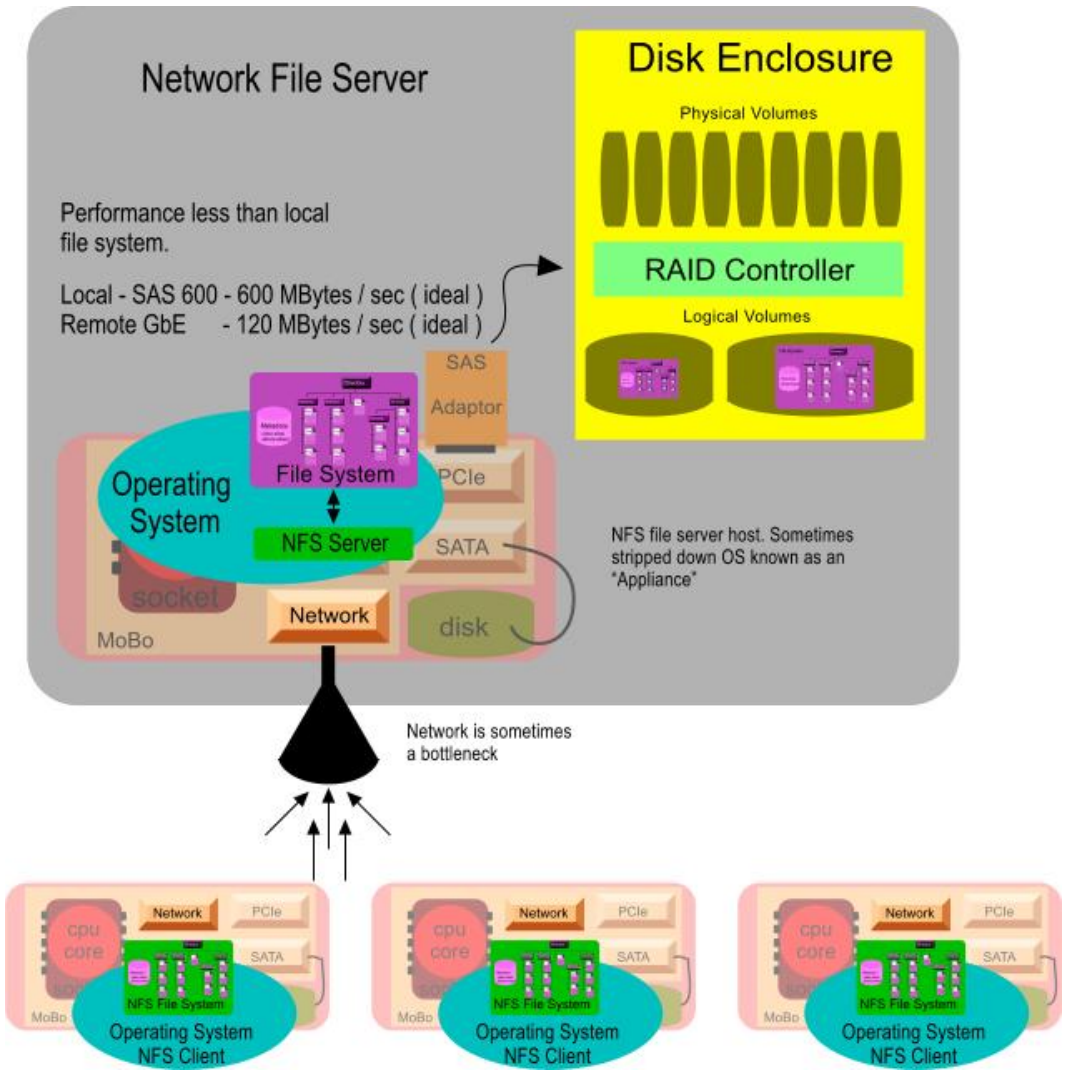
NAS or Network Appliance



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Network BW is often the bottleneck



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Distributed File Systems – quick glance from Wikipedia

Examples [edit]

Main article: List of distributed file systems

- Alluxio
- BeeGFS (Fraunhofer)
- Ceph (Inktank, Red Hat, SUSE)
- Windows Distributed File System (DFS) (Microsoft)
- Infinit
- GfarmFS
- GlusterFS (Red Hat)
- GFS (Google Inc.)
- HDFS (Apache Software Foundation)
- IPFS
- iRODS
- LizardFS (Skytechnology)
- MapR FS
- MooseFS (Core Technology / Gemius)
- ObjectiveFS
- OneFS (EMC Isilon)
- OpenIO
- OrangeFS (Clemson University, Omnibond Systems), formerly Parallel Virtual File System
- Panfs (Panasas)
- Parallel Virtual File System (Clemson University, Argonne National Laboratory, Ohio Supercomputer Center)
- RozoFS (Rozo Systems)
- Torus (CoreOS)
- XtreemFS

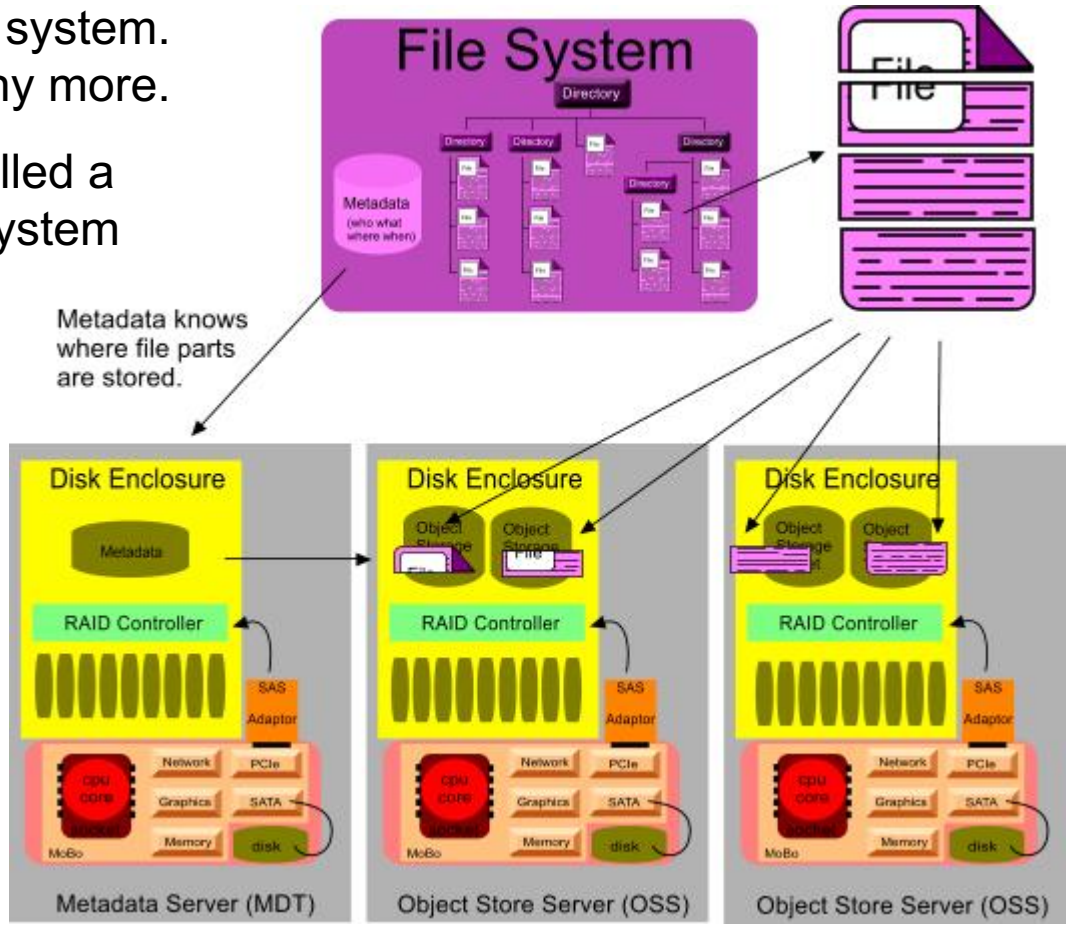
SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

NAS - Lustre File System

Lustre is an example of a distributed file system. There are many more.

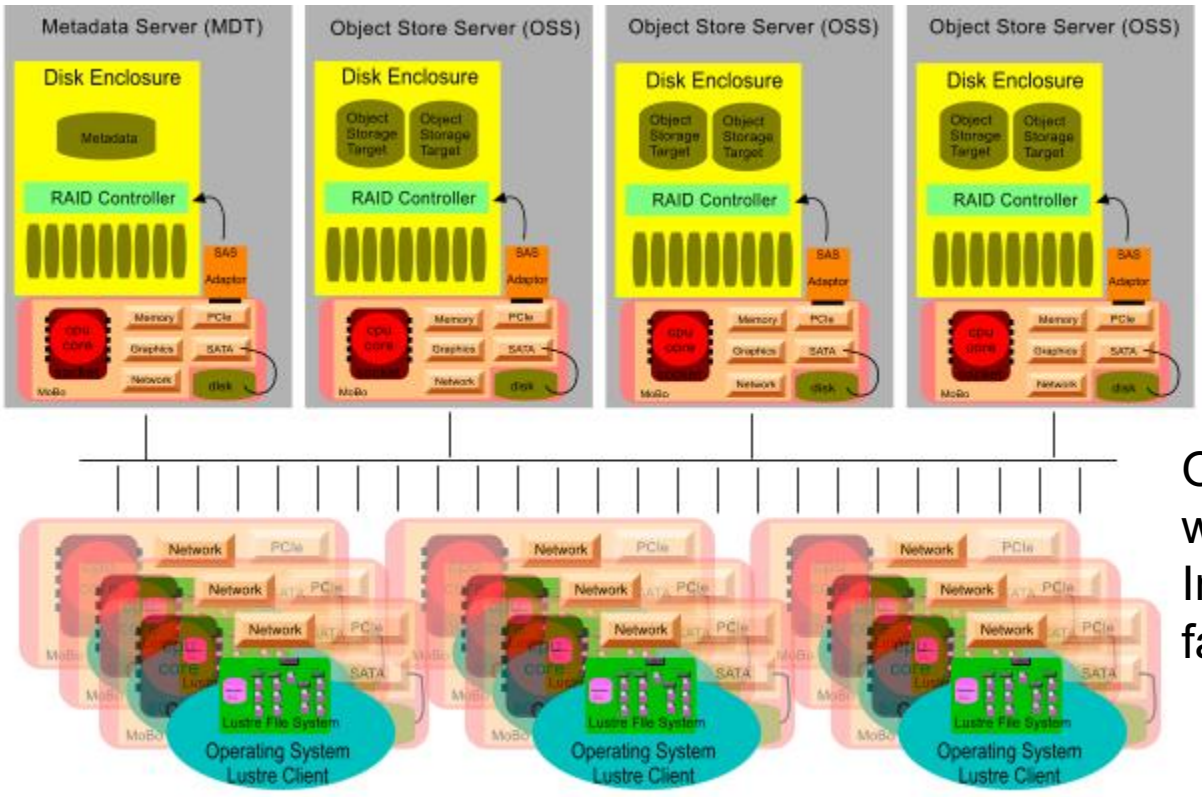
Sometimes called a "Cluster" file system



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

NAS – Lustre



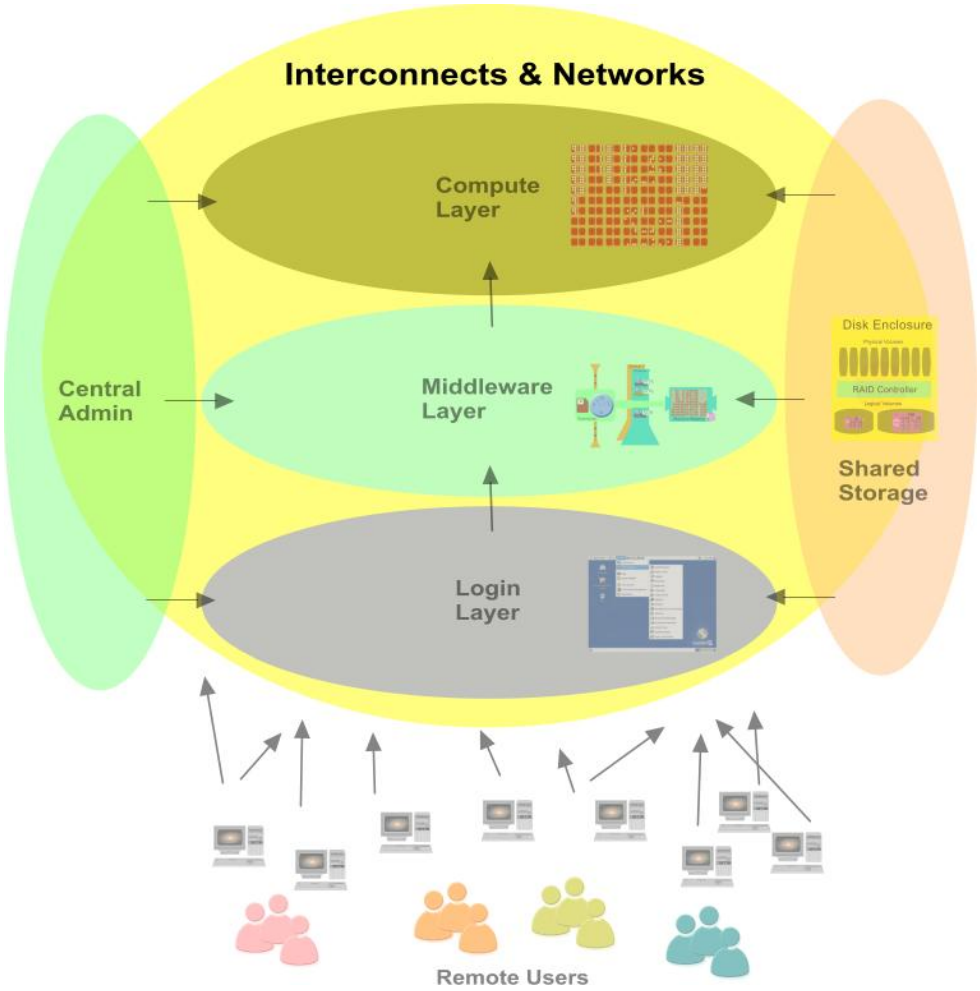
Often used with an Infiniband fabric.

LUSTRE - Massively Parallel Distributed Shared File System

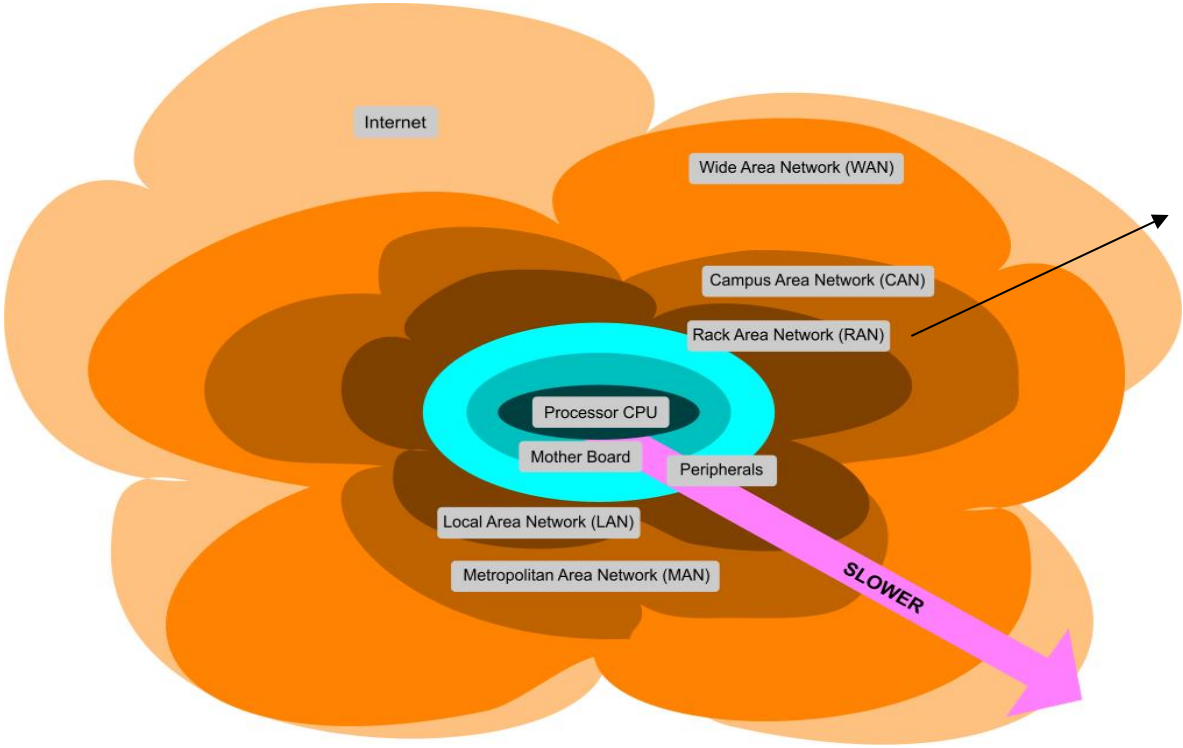
SCAMA Portsmouth

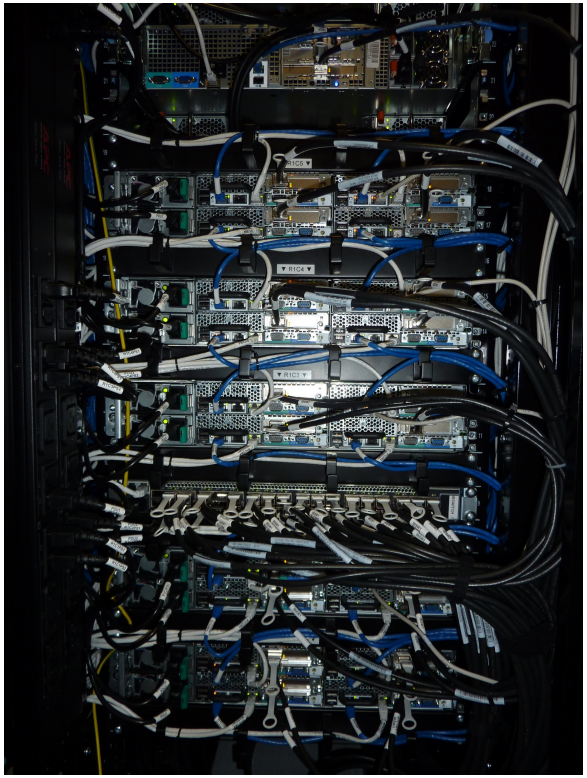
UNIVERSITY OF PORTSMOUTH

Interconnects and Networks



Interconnects and Networks



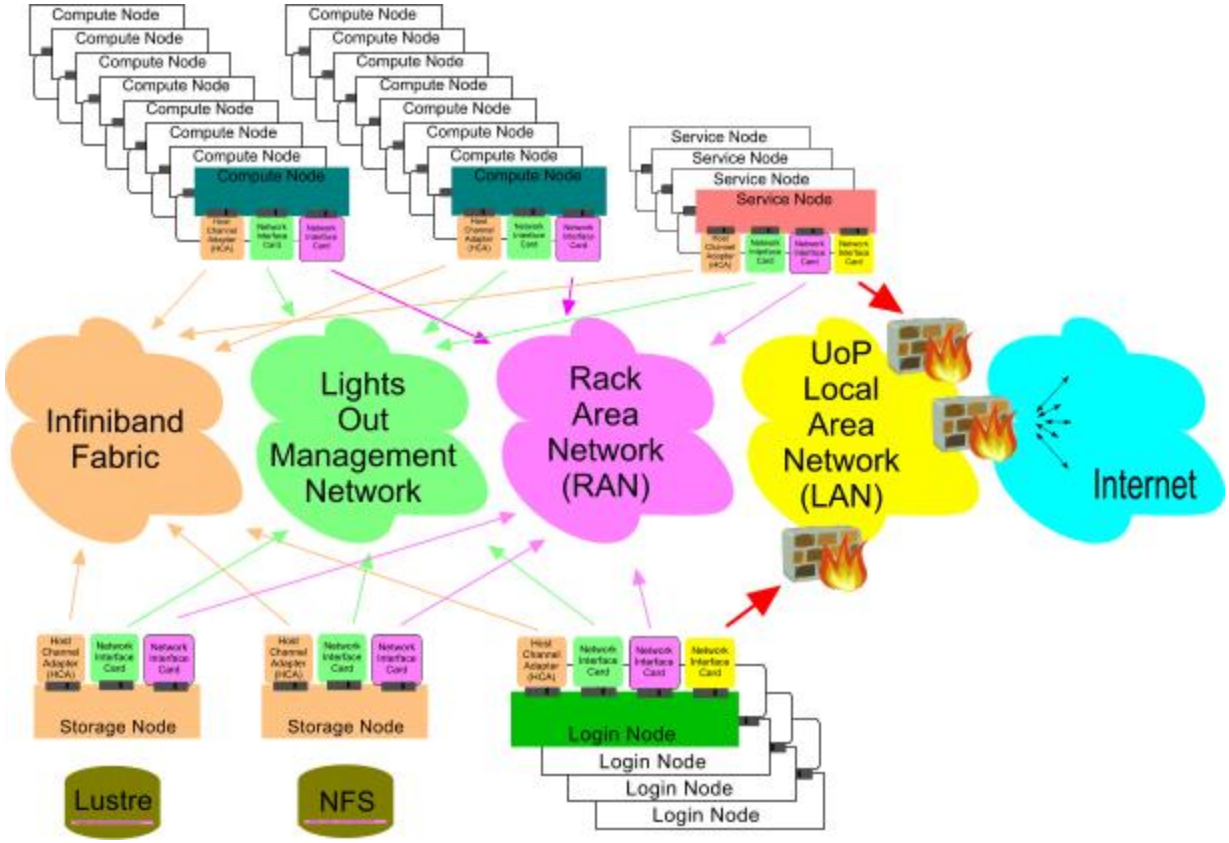


Moving away from the Processor towards the Internet you get slower and slower due to Increased Latency and Reduced Bandwidth

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

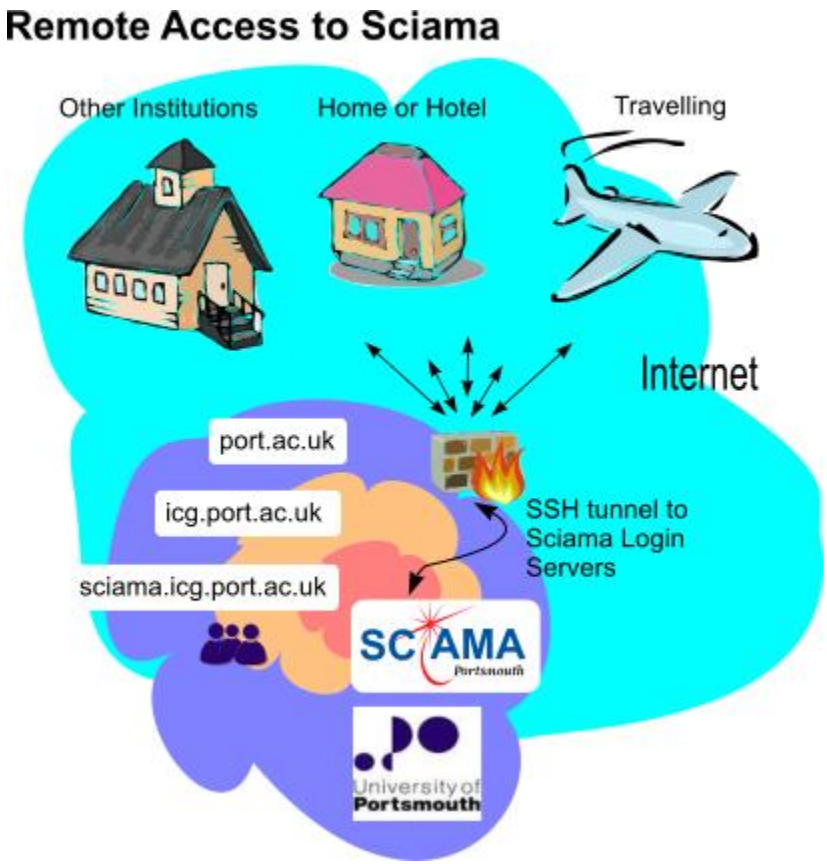
Commodity Cluster Networks



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Connecting to an HPC

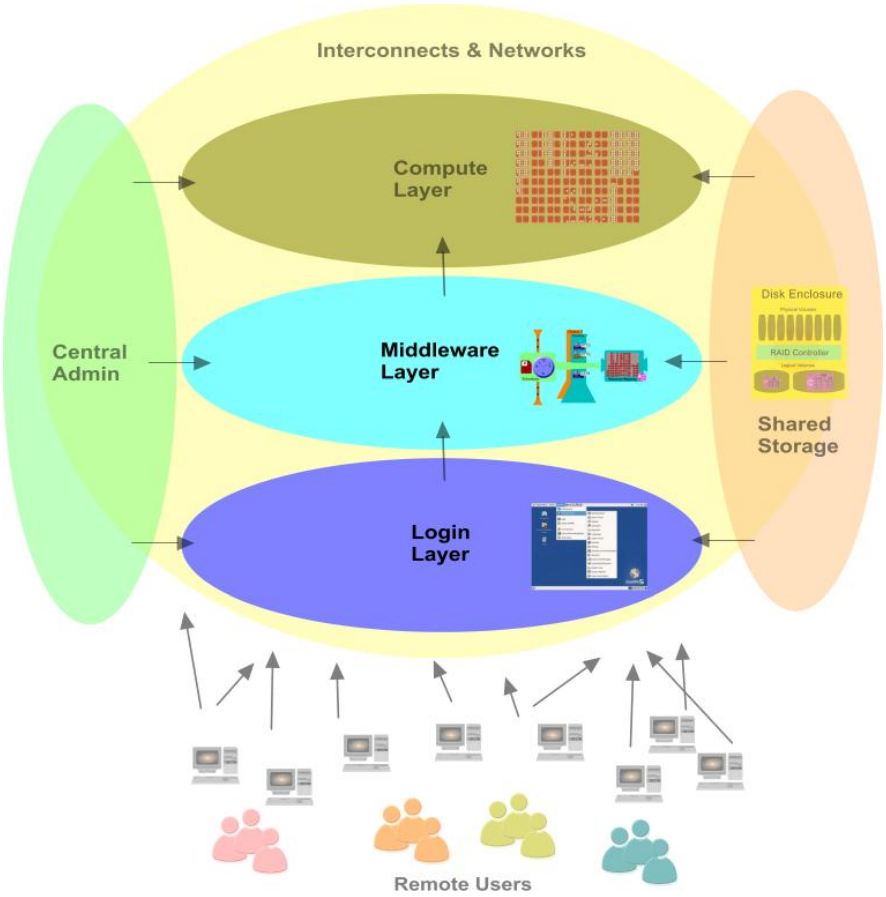


SCIAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

SCAMA Portsmouth

Login Layer



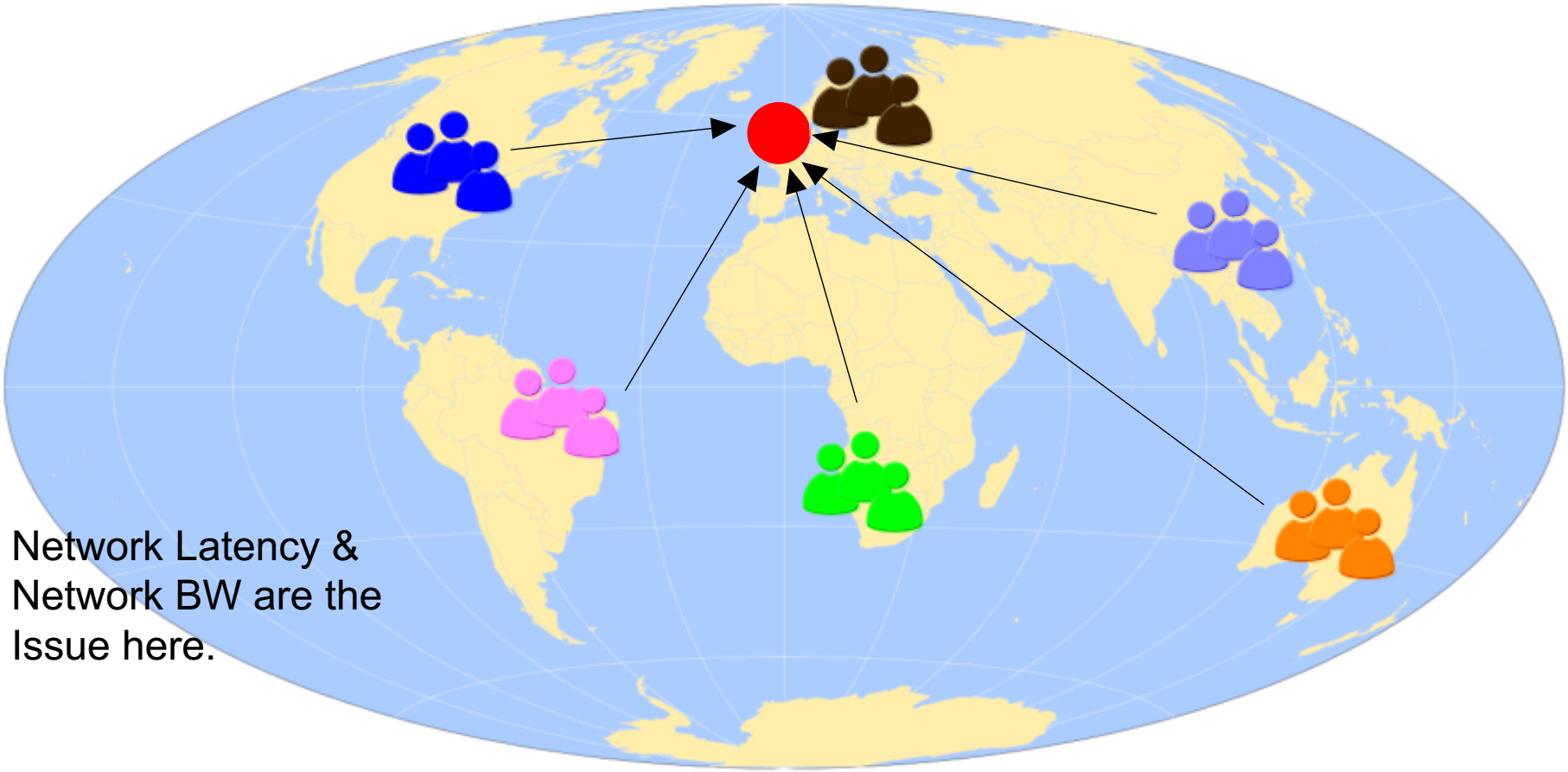
Why Login Servers

- Login servers will provide the gateway to the cluster.
- Users can remotely login into the servers using “ssh” or a Remote Desktop Client.
- A desktop client gives a full working desktop in the environment (can full screen)

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Global User Base

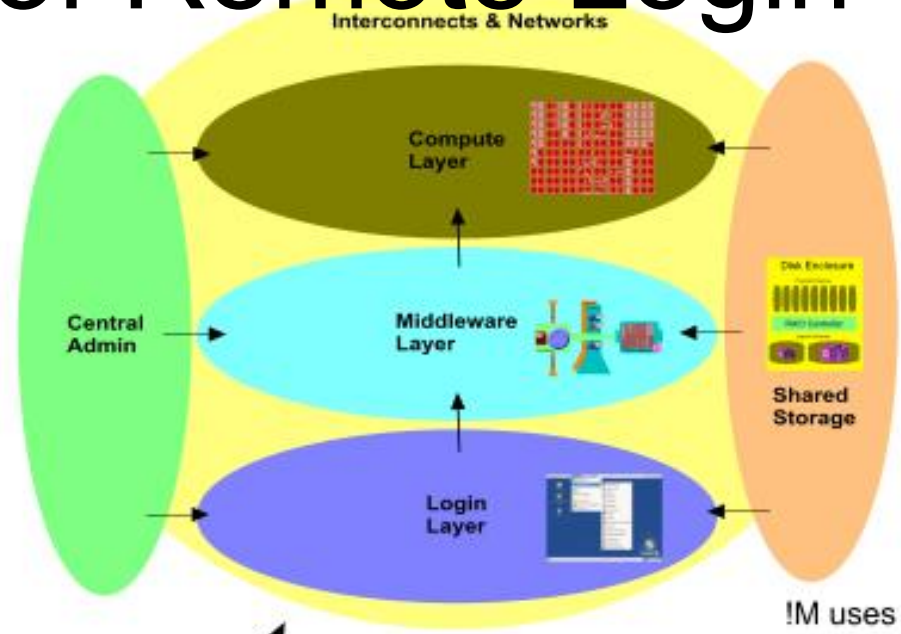


Network Latency & Network BW are the Issue here.

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Use of Remote Login Client



1.) User downloads !M client from Internet. Client is available for Linux, Windows and MAC. The client is free.

2.) Once installed and configured the !M client can connect to the !M server running in the cluster.

!M uses a very lean protocol for use over high latency low bandwidth links.

Internet

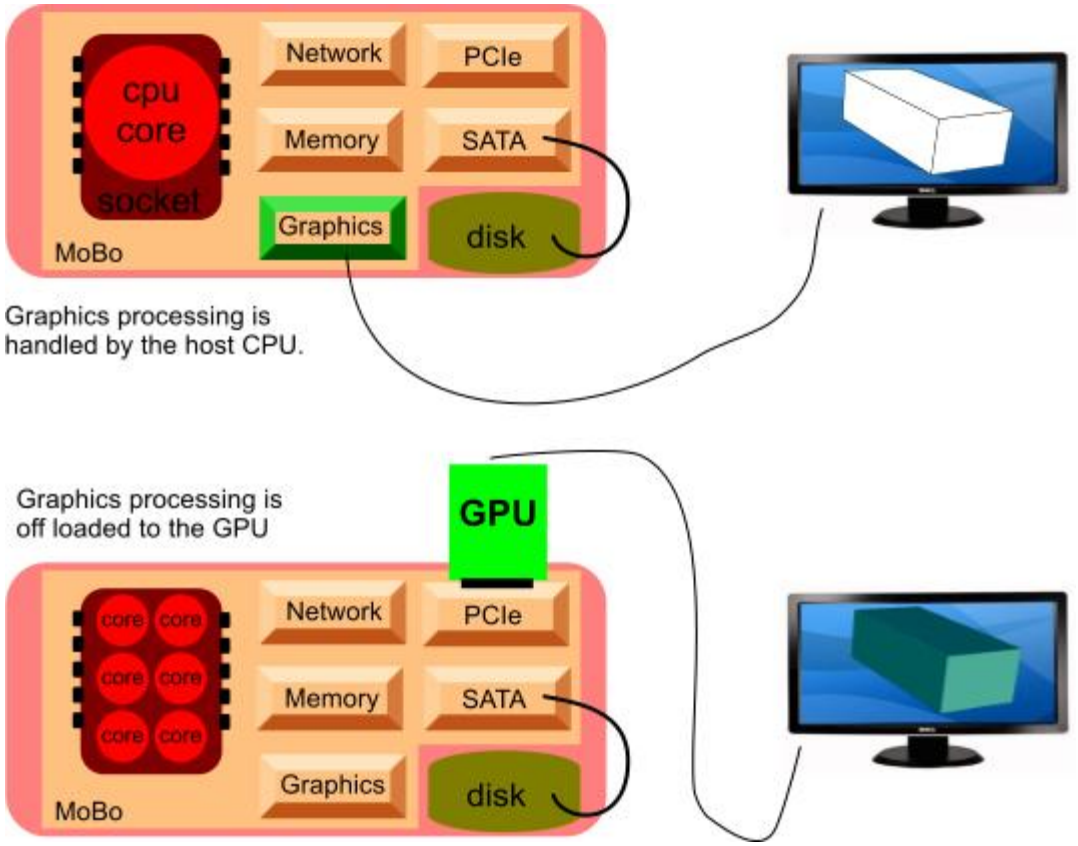
3.) the user is returned a login prompt or a suspended session.

GPGPU's

General Purpose Graphics Processing Units

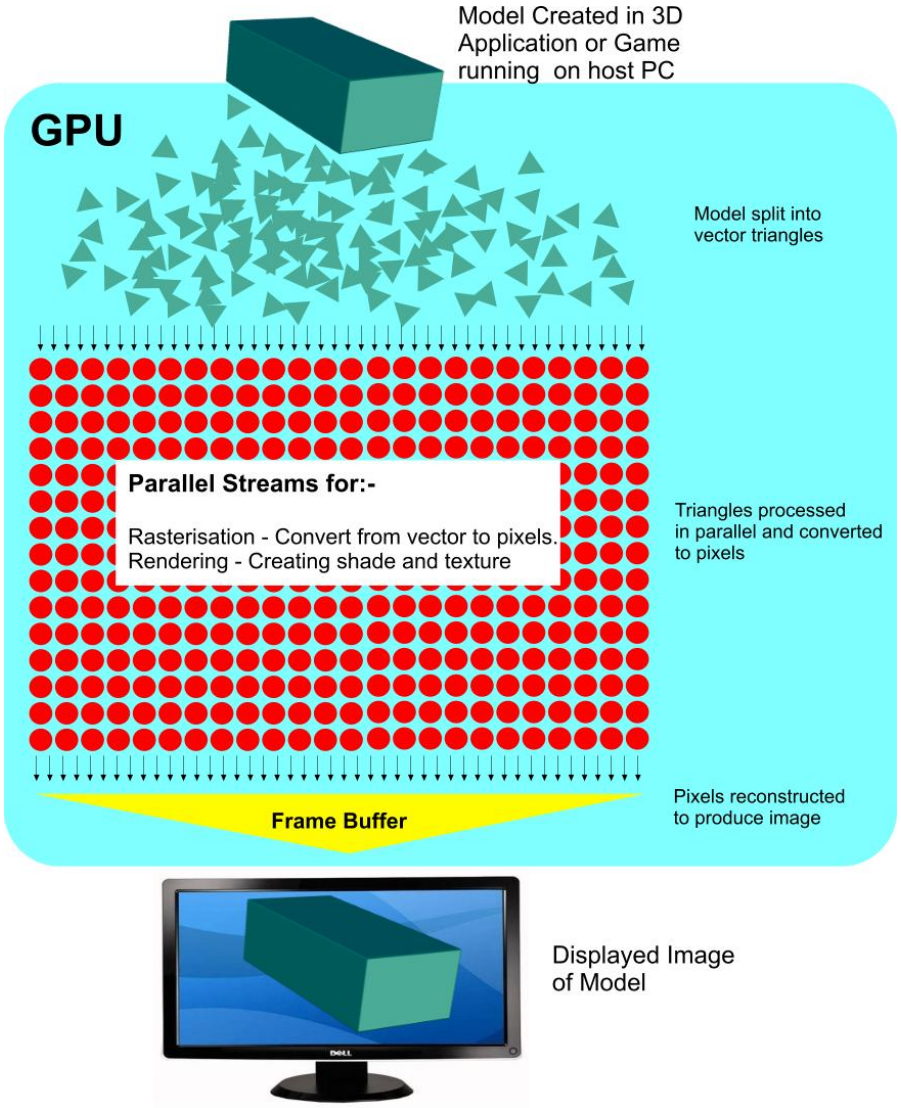
G Burton – ICG – Oct12 – v1.0

Graphics Processing



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH



Cheap commodity hardware mainly from Nvida, AMD and Intel for home PC's.

Driven by gaming.

Bespoke architecture replaced by generic programmable architecture.

SCAMA Portsmouth

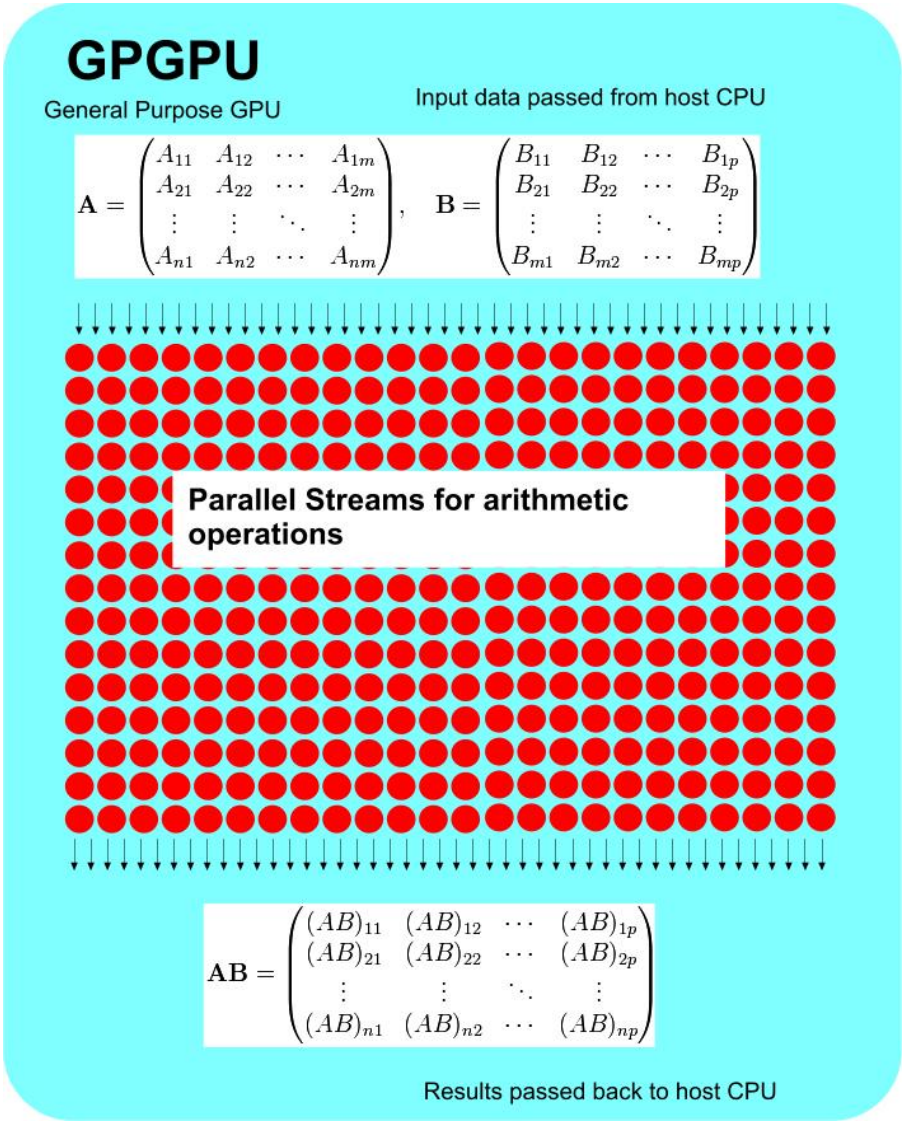
UNIVERSITY OF PORTSMOUTH

The Birth of Cuda

(compute Unified Device Architecture) and OpenCL.

Cuda (Nvida) is cutting edge, OpenCL follows.

If you know Cuda then OpenCL Is easy.

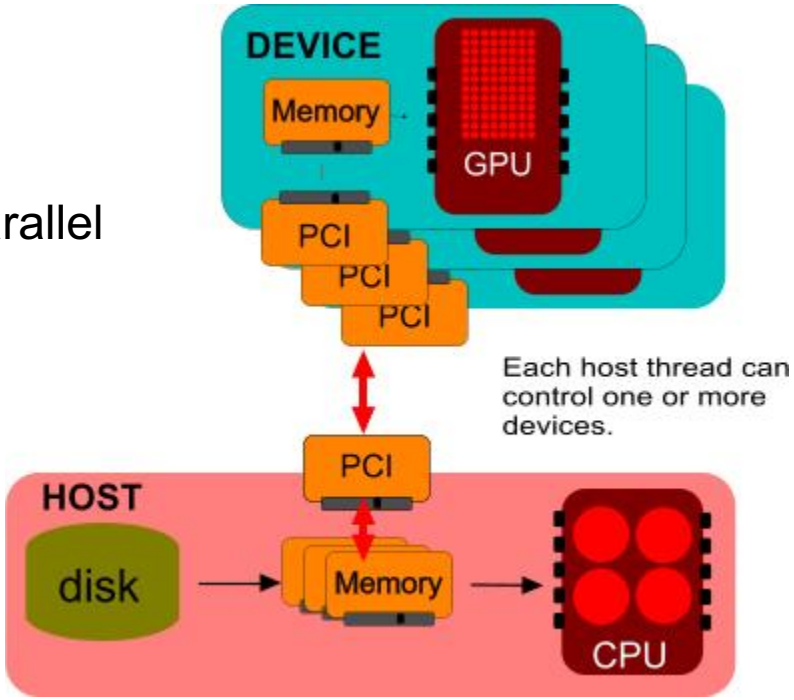


SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Several GPU cards can be connected in Parallel

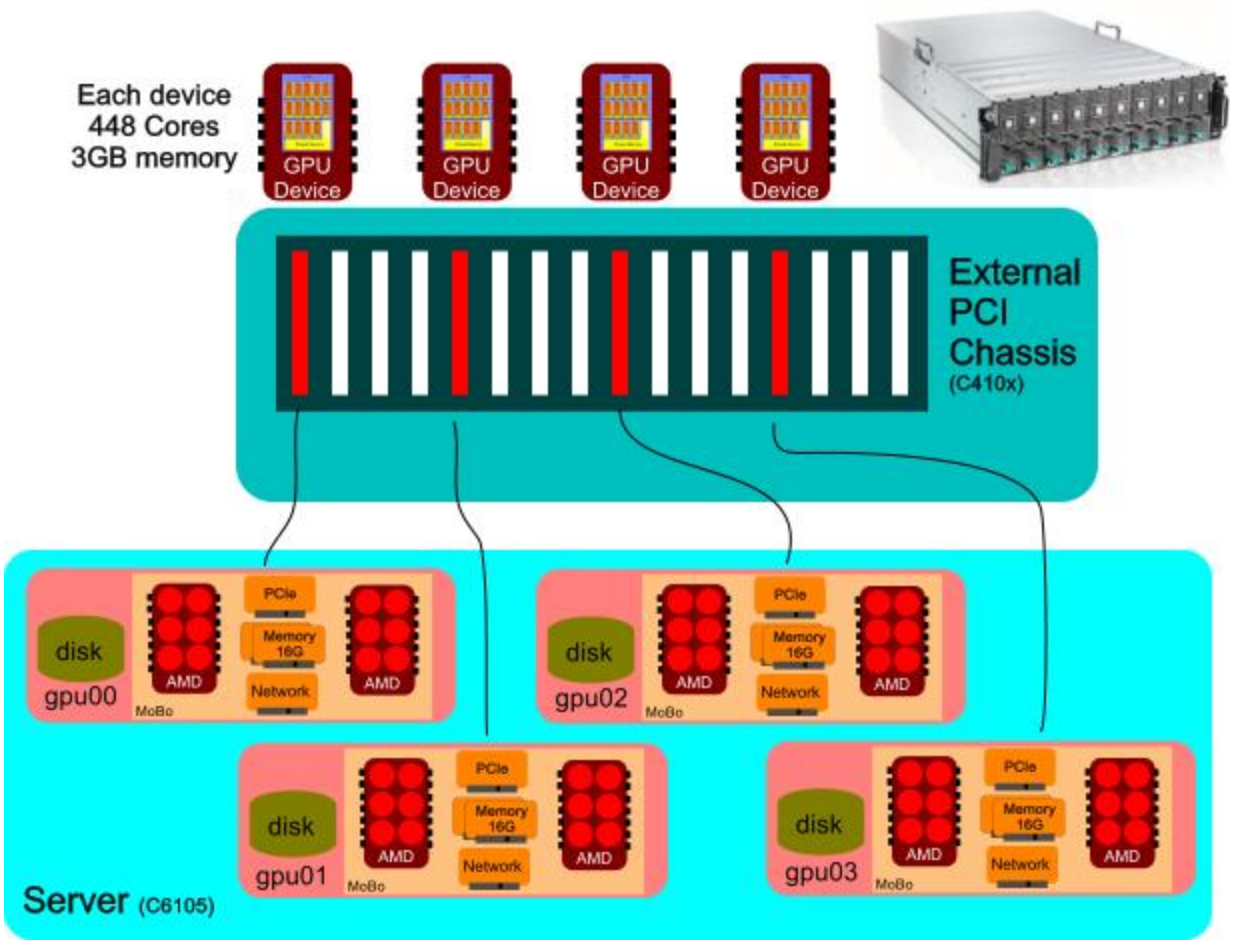
Allows 1000's of GPU cores for massively parallel computation.



SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Sciama Specifics



SCIAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Sciama GPU Specification

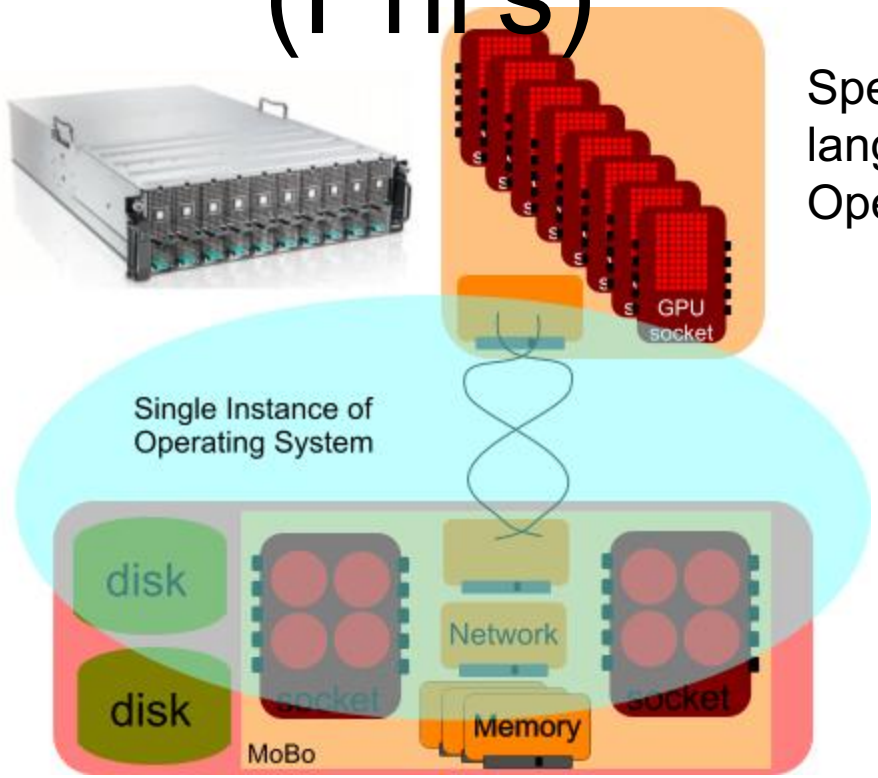
```
CUDA Device Query (Runtime API) version (CUDART static linking)

Found 1 CUDA Capable device(s)

Device 0: "Tesla M2050"
  CUDA Driver Version / Runtime Version          4.20 / 4.0
  CUDA Capability Major/Minor version number:    2.0
  Total amount of global memory:                 2687 MBytes (2817982464 bytes)
  (14) Multiprocessors x (32) CUDA Cores/MP:     448 CUDA Cores
  GPU Clock Speed:                               1.15 GHz
  Memory Clock rate:                             1546.00 Mhz
  Memory Bus Width:                              384-bit
  L2 Cache Size:                                 786432 bytes
  Max Texture Dimension Size (x,y,z)             1D=(65536), 2D=(65536,65535), 3D=(2048,2048,2048)
  Max Layered Texture Size (dim) x layers        1D=(16384) x 2048, 2D=(16384,16384) x 2048
  Total amount of constant memory:               65536 bytes
  Total amount of shared memory per block:       49152 bytes
  Total number of registers available per block: 32768
  Warp size:                                     32
  Maximum number of threads per block:           1024
  Maximum sizes of each dimension of a block:    1024 x 1024 x 64
  Maximum sizes of each dimension of a grid:     65535 x 65535 x 65535
  Maximum memory pitch:                          2147483647 bytes
  Texture alignment:                             512 bytes
  Concurrent copy and execution:                 Yes with 2 copy engine(s)
  Run time limit on kernels:                     No
  Integrated GPU sharing Host Memory:            No
  Support host page-locked memory mapping:       Yes
  Concurrent kernel execution:                   Yes
  Alignment requirement for Surfaces:            Yes
  Device has ECC support enabled:                Yes
  Device is using TCC driver mode:               No
  Device supports Unified Addressing (UVA):      Yes
  Device PCI Bus ID / PCI location ID:           15 / 0
  Compute Mode:
    < Default (multiple host threads can use ::cudaSetDevice() with device simultaneously) >

deviceQuery, CUDA Driver = CUDART, CUDA Driver Version = 4.20, CUDA Runtime Version = 4.0, NumDevs = 1,
Device = Tesla M2050
[gputest] test results...
PASSED
```

SCIAMA Portsmouth

UNIVERSITY OF PORTSMOUTH

Graphical Processing Units (GPU's) and Intel CoProcessors (Phi's)

Three players:-
Intel
AMD
Nvidia



Special programming language. CUDA and OpenCL

Cpu – multiple cores
Gpu – 100's of cores

CPU's still in charge

SCAMA Portsmouth

UNIVERSITY OF PORTSMOUTH