DEMOCRITOS

DEmocritos MOdeling Center for Research In aTOmistic Simulation

EXADRON

NETTAB 2006

Distributed Applications, Web Services, Tools and GRID Infrastructures for Bioinformatics

HPC Infrastructures

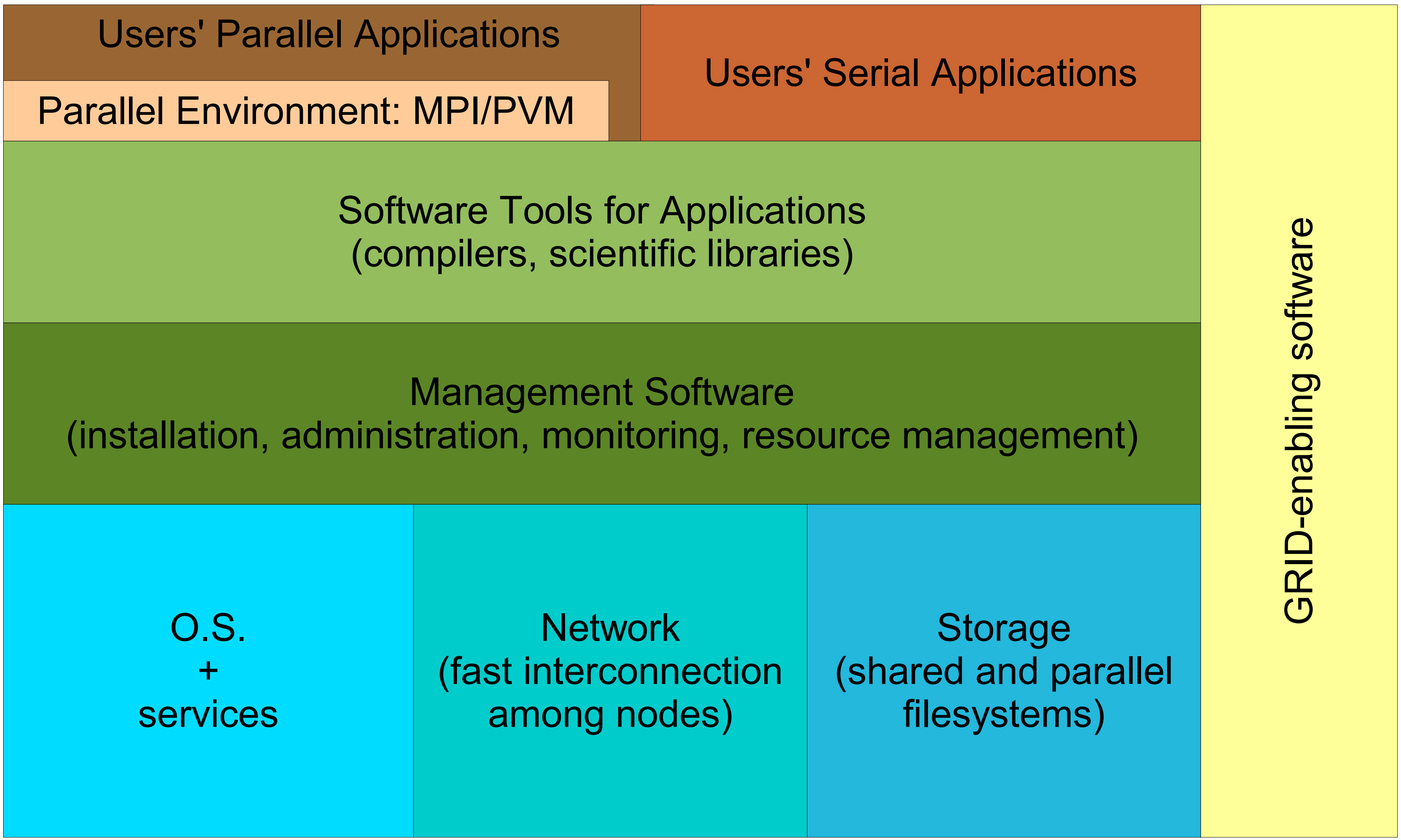
Moreno Baricevic

CNR-INFM DEMOCRITOS, Trieste

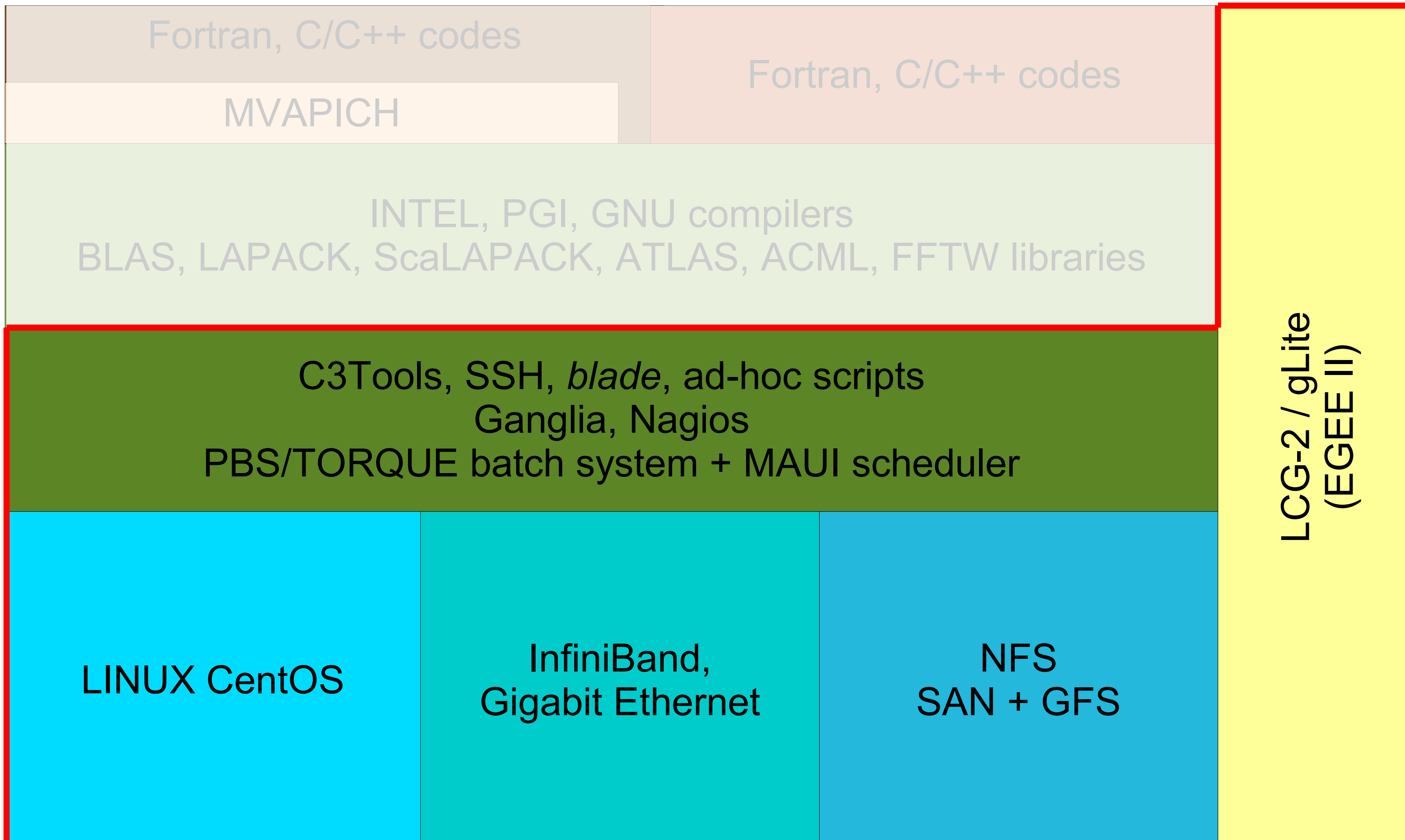
Matteo Vit

EXADRON, Amaro (UD)

SOFTWARE INFRASTRUCTURE
Overview



SOFTWARE INFRASTRUCTURE Overview (Michelangelo @ CILEA)



COMPATIBILITY ISSUES
Kernel vs new hardware

- Latest vanilla kernel: **2.6.16**.9
- CentOS kernel: **2.6.9**-22
- UnionFS v1.1.4 ↔ kernel **2.6.9** ÷ **2.6.14**
- InfiniBand IBGD-1.8.2 ↔ kernel ≤ **2.6.11**
- GFS cluster 1.01 ↔ kernel ≤ **2.6.14**
- GFS cluster 1.02 ↔ kernel **2.6.15** <u>patched by FC5</u>
- Qlogic qla2xxx (severe bug fixed) ↔ kernel ≥ **2.6.15**
- AMD CPU Dual Core 275 ↔ kernel ≥ **2.6.12**

[Up to May 2006]

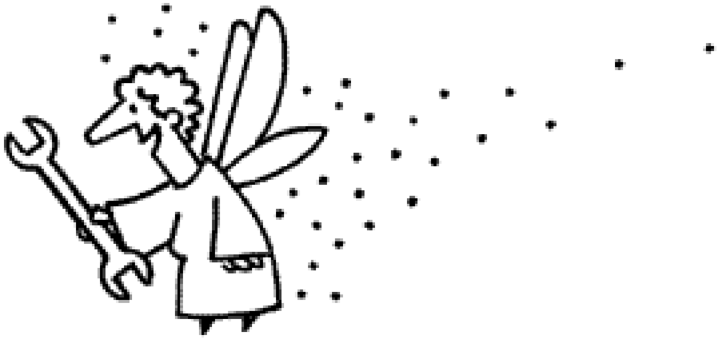
COMPATIBILITY ISSUES
Kernel vs new hardware

kernel version 2.6.	...	9	10	11	12	13	14	15	16	...

- Latest vanilla kernel 2.6.16
- CentOS kernel 2.6.9
- UnionFS v1.1.4 2.6.9 ÷ 2.6.14
- InfiniBand IBGD-1.8.2 ≤ 2.6.11
- GFS cluster 1.01 ≤ 2.6.14
- GFS cluster 1.02 2.6.15 (patched by FC5)
- Qlogic qla2xxx (severe bug fixes) ≥ 2.6.15
- AMD CPU Dual Core 275 ≥ 2.6.12

[Up to May 2006]

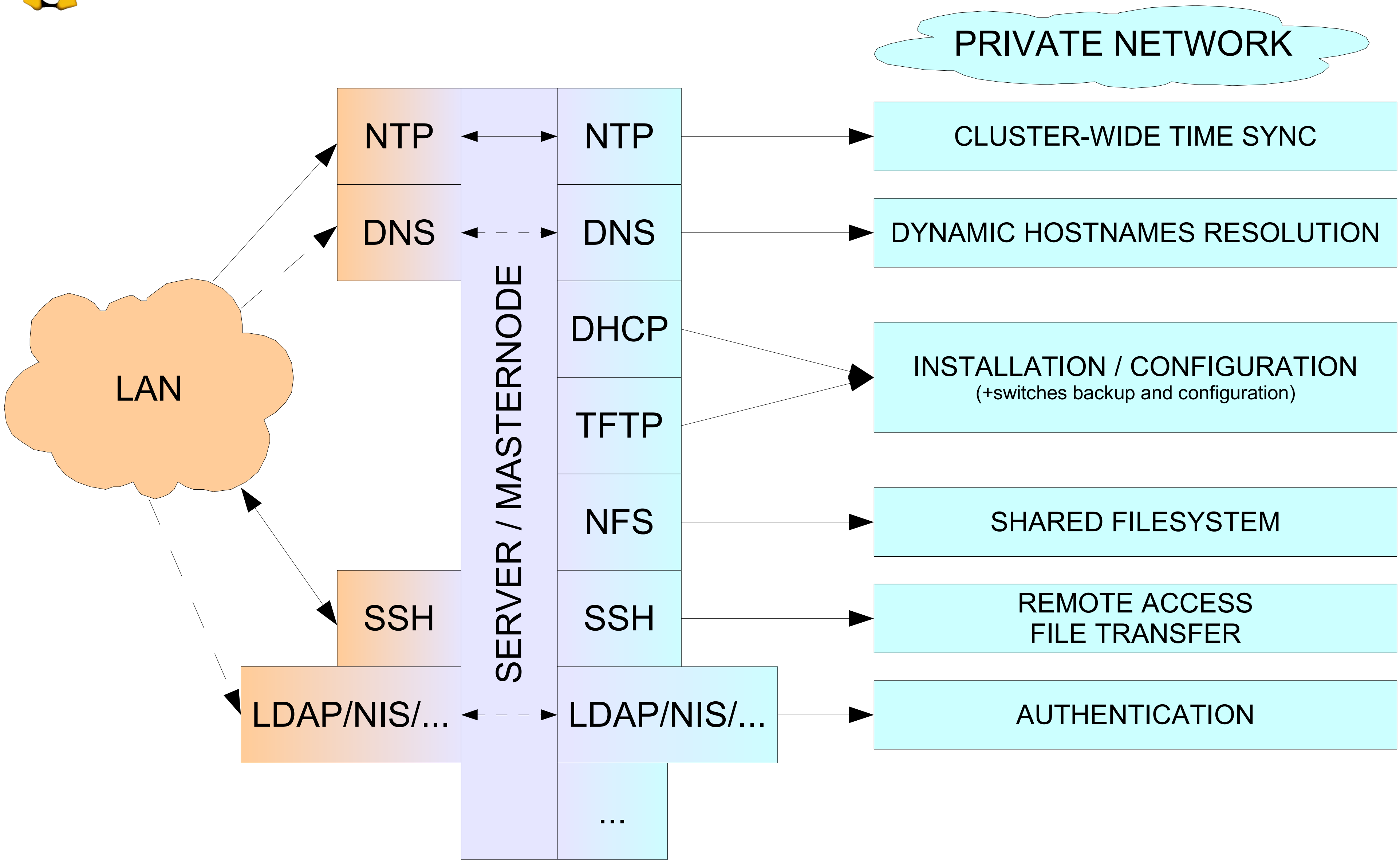
COMPATIBILITY ISSUES
Kernel vs new hardware



Roll up your own kernel
and patch as needed!

- vanilla kernel **2.6.16.16**
- UnionFS 1.1.4 (patched)
- IBGD 1.8.2 (patched)
- GFS Cluster-1.02 (patched)
- Qlogic qla2xxx (asis)

CLUSTER SERVICES



CLUSTER MANAGEMENT
Installation

Installation can be performed:

- interactively
- non-interactively

- **Interactive** installations:
  - finer control
- **Non-interactive** installations:
  - minimize human intervention and let you save a lot of time
  - are less error prone
  - are performed using programs (such as RedHat Kickstart) which:
    - “simulate” the interactive answering
    - can perform some post-installation procedures for customization

CLUSTER MANAGEMENT
Installation

MASTERNODE

Ad-hoc installation once forever (hopefully), usually interactive:

- local devices (CD-ROM, DVD-ROM, Floppy, ...)
- network based (PXE+DHCP+TFTP+NFS)

CLUSTER NODES

One installation reiterated for each node, usually non-interactive.

Nodes can be:

1) disk-based
2) disk-less (not to be really installed)

CLUSTER MANAGEMENT
Cluster Nodes Installation

1) Disk-based nodes

- CD-ROM, DVD-ROM, Floppy, ...
  Time expensive and tedious operation
- HD cloning: mirrored raid, dd, Ghost and the like
  A “template” hard-disk needs to be swapped or a disk image needs to be available for cloning, configuration needs to be changed either way
- Distributed installation: PXE+DHCP+TFTP+NFS
  More efforts to make the first installation work properly (especially for heterogeneous clusters), (mostly) straightforward for the next ones

2) Diskless nodes

- Live CD/DVD/Floppy
- NFS
- NFS + UnionFS
- initrd (RAM disk)

CLUSTER MANAGEMENT
Existent toolkits

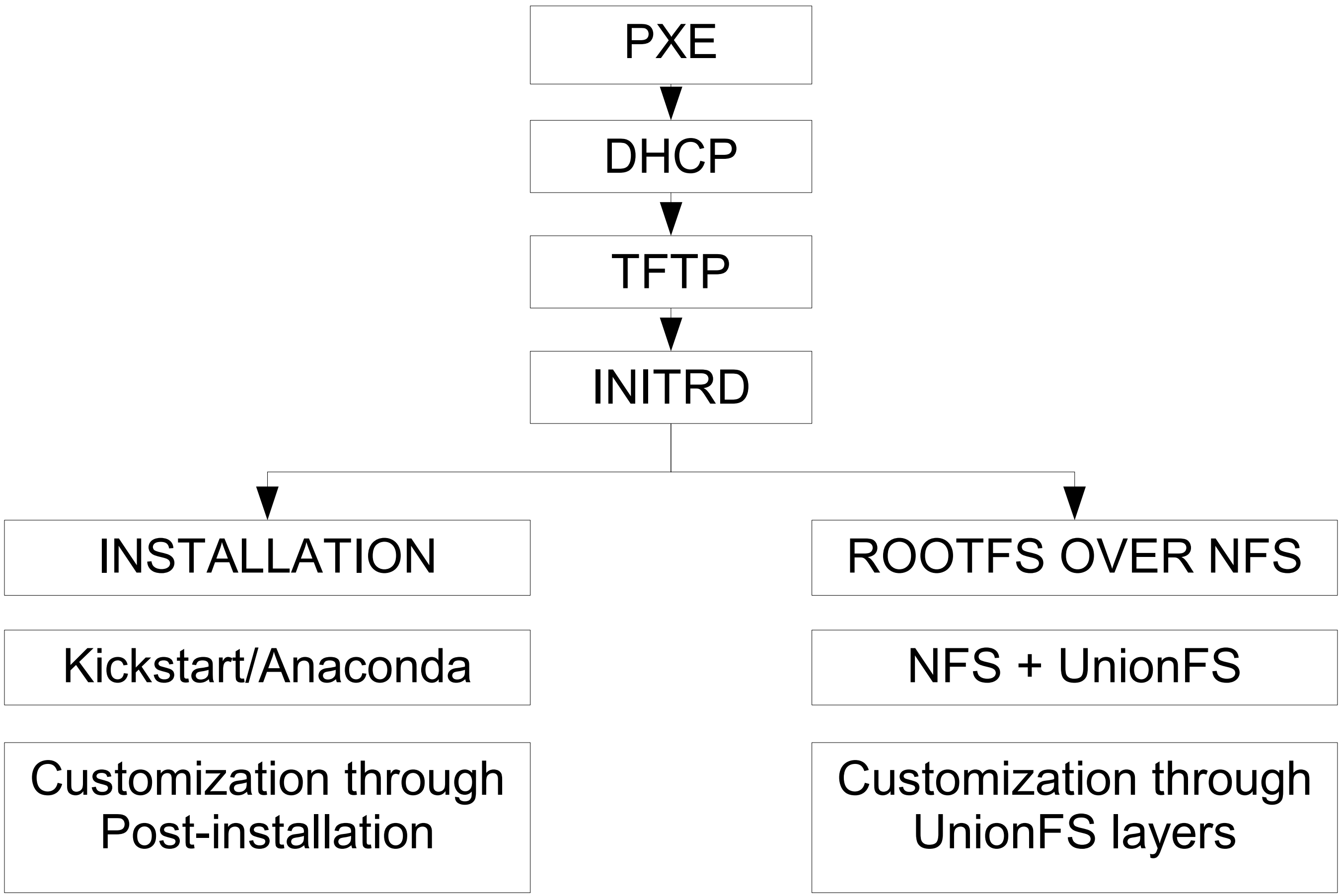
Are generally made of an ensemble of already available software packages thought for specific tasks, but configured to operate together, plus some add-ons.

Sometimes limited by rigid and not customizable configurations, often bounded to some specific linux distribution and version. May depend on vendors' hardware.

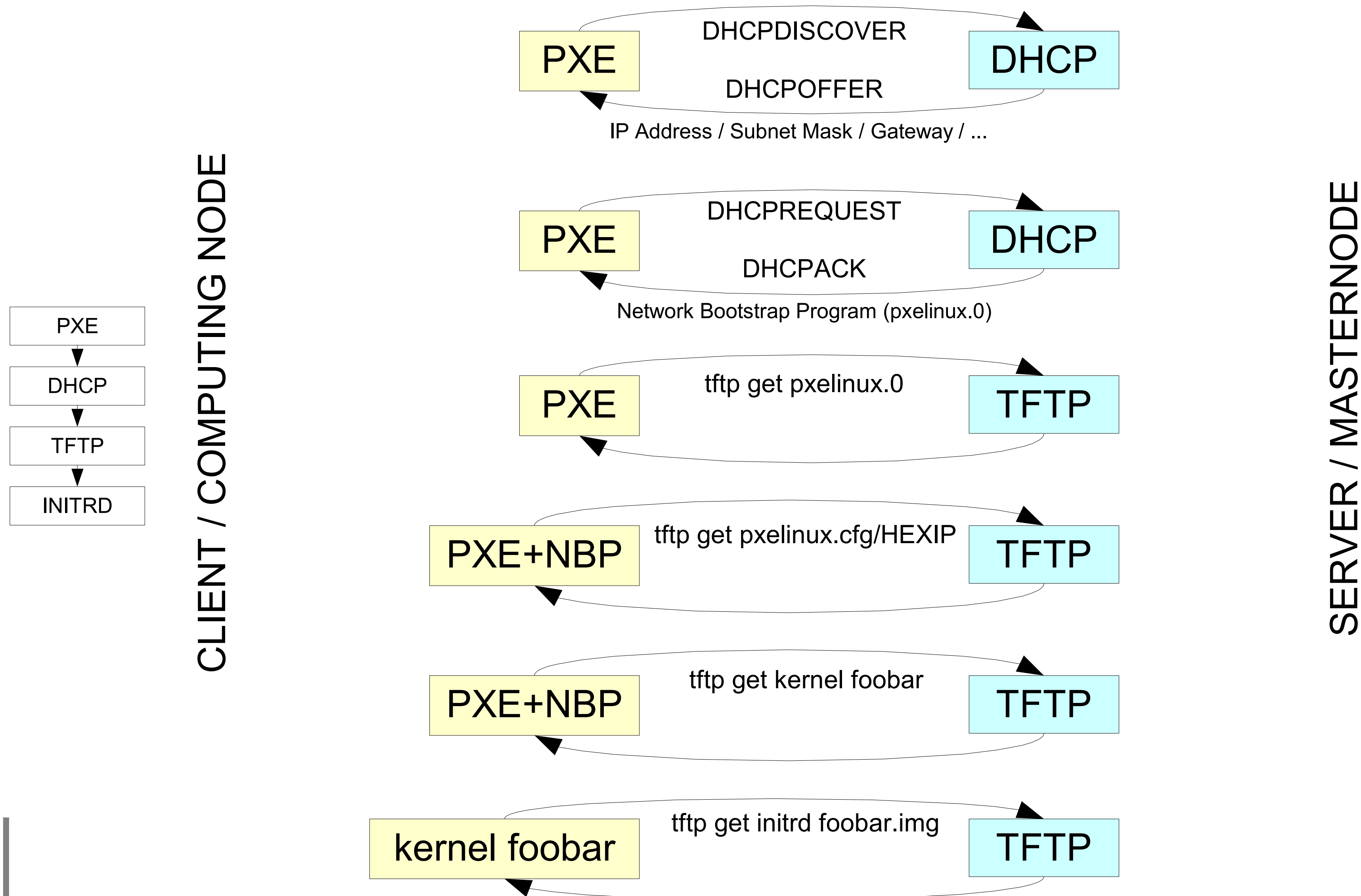
- Free and Open
  - OSCAR (Open Source Cluster Application Resources)
  - NPACI Rocks
  - xCAT (eXtreme Cluster Administration Toolkit)
  - OpenSCE (Open Scalable Cluster Environment)
  - Warewulf

- Commercial
  - IBM CSM (Cluster Systems Management)
  - Scyld Beowulf
  - HP, SUN and other vendors' Management Software...

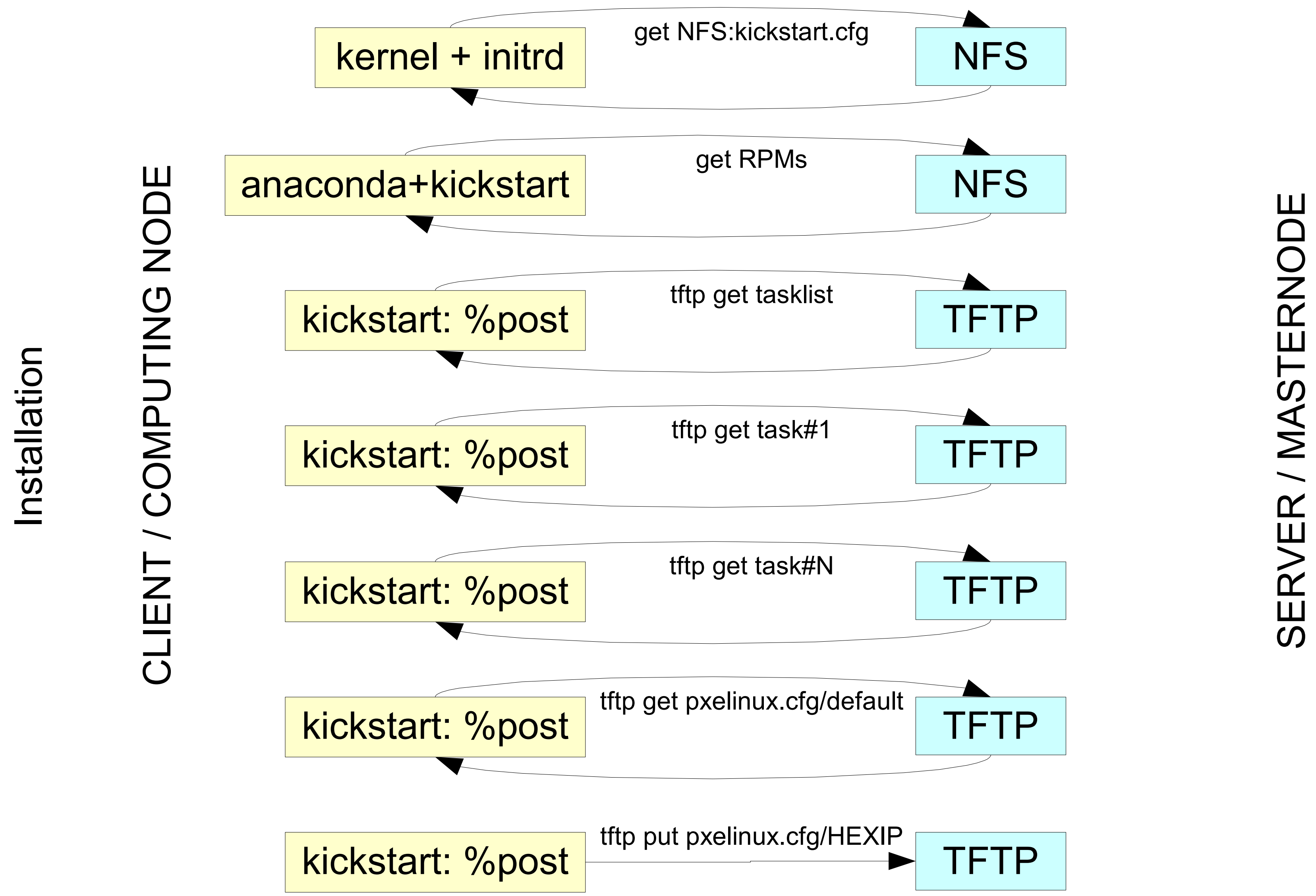
CLUSTER MANAGEMENT
Network-based Distributed Installation



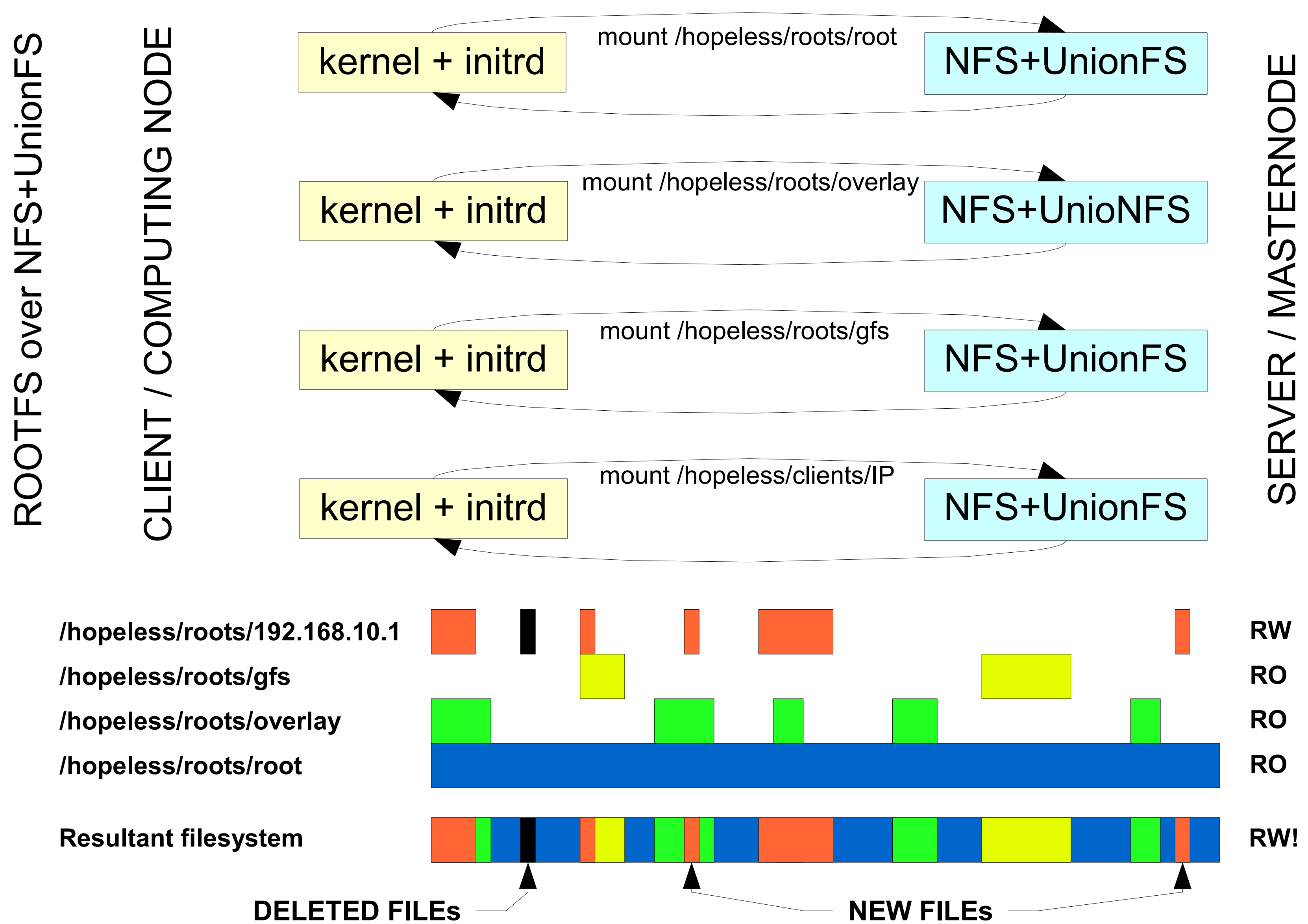
CLUSTER MANAGEMENT
Network-based Distributed Installation



CLUSTER MANAGEMENT
Network-based Distributed Installation



CLUSTER MANAGEMENT
Network-based Distributed Installation



CLUSTER MANAGEMENT
Administration Tools

Requirements:

- cluster-wide command execution
- cluster-wide file distribution and gathering
- <u>must be simple, efficient, easy to use for CLI addicted</u>

- C3 tools - The Cluster Command and Control tool suite
  - allows configurable clusters and subsets of machines
  - concurrently execution of commands
  - supplies many utilities
    - cexec (parallel execution of standard commands on all cluster nodes)
    - cexecs (as the above but serial execution, useful for troubleshooting and debugging)
    - cpush (distribute files or directories to all cluster nodes)
    - cget (retrieves files or directory from all cluster nodes)
    - crm (cluster-wide remove)
    - ... and many more
  - http://www.csm.ornl.gov/torc/C3/

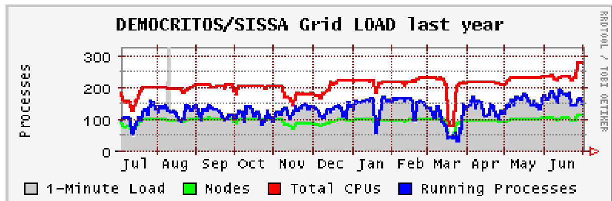
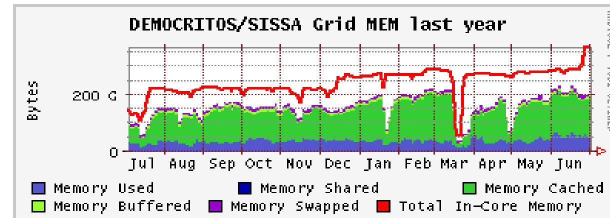
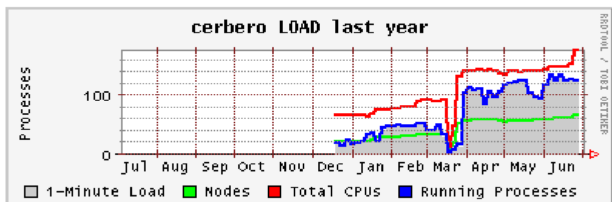
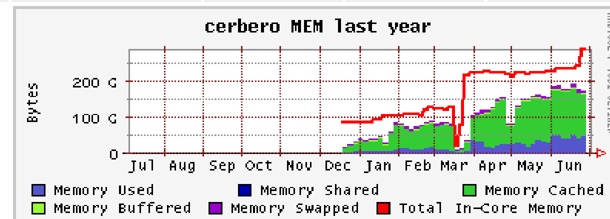
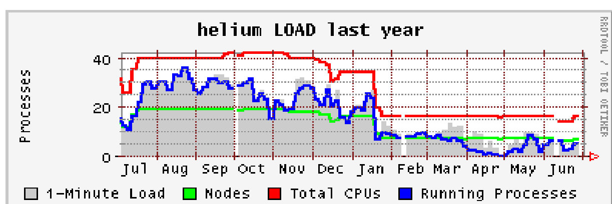
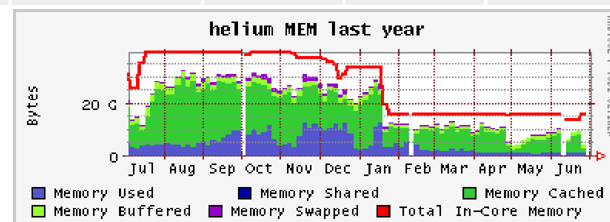
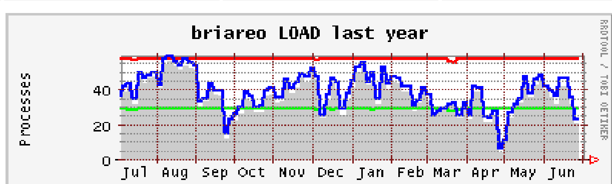
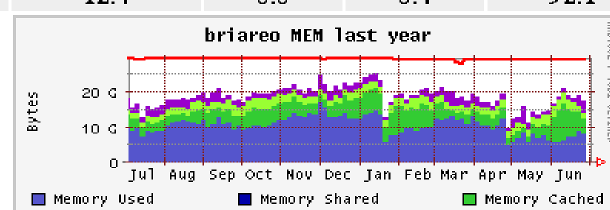
- DSH - Distributed Shell
    - http://www.netfort.gr.jp/~dancer/software/dsh.html.en

CLUSTER MANAGEMENT
Monitoring Tools

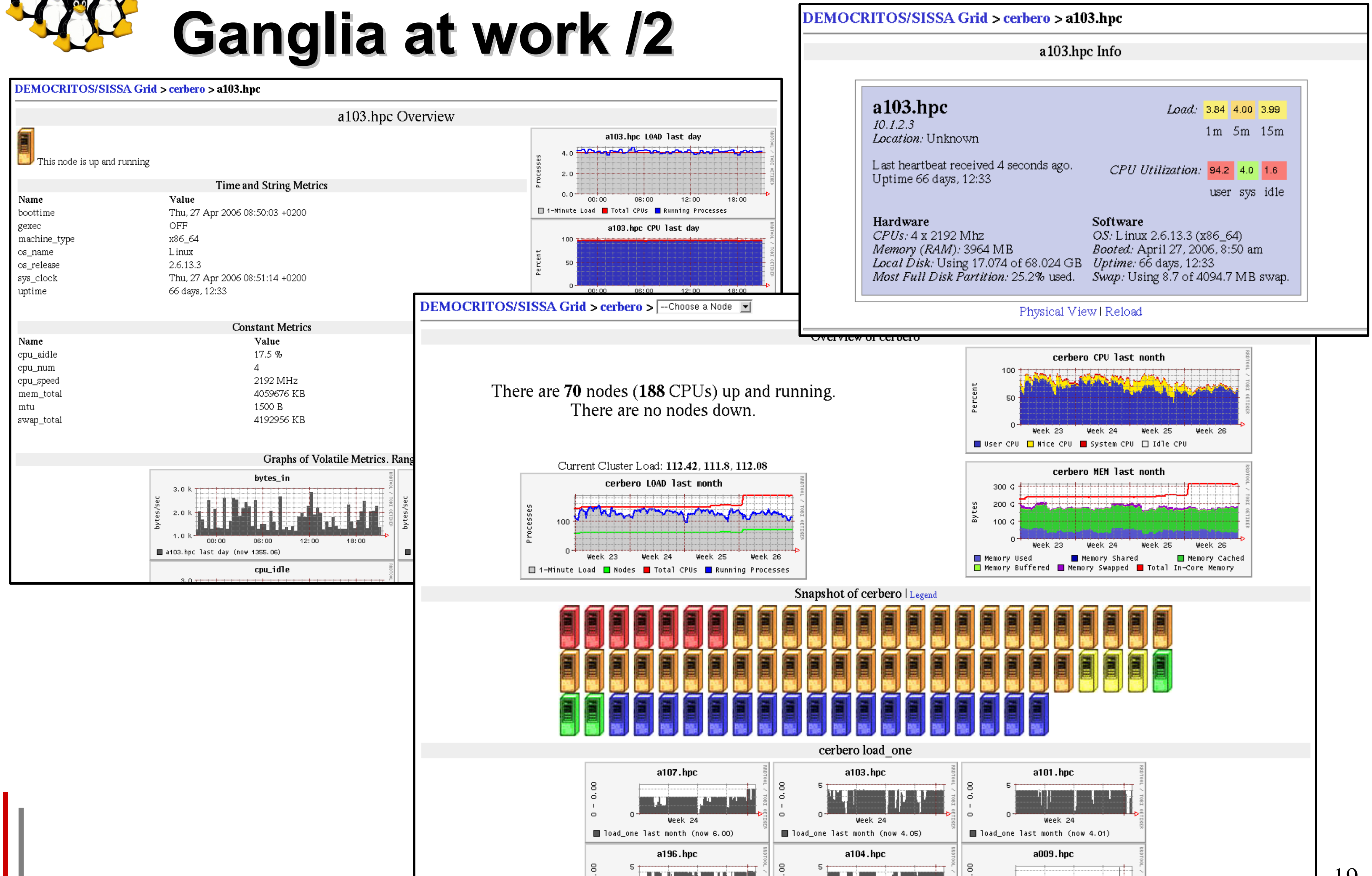
- Ad-hoc scripts (shell, perl, ...) + cron
- Ganglia
  - excellent graphic tool
  - XML data representation
  - web-based interface for visualization
  - http://ganglia.sourceforge.net/
- Nagios
  - complex but can interact with other software
  - configurable alarms, SNMP, e-mail, sms, ...
  - optional web interface
  - http://www.nagios.org/

CLUSTER MANAGEMENT
Ganglia at work /1

DEMOCRITOS/SISSA Grid > --Choose a Source

Name / Info	Load Averages			%CPU User, Nice, System, Idle			
DEMOCRITOS/SISSA Grid (4 sources) (tree view)	124.76	124.33	124.26	45.5	1.3	1.0	52.6
Hosts up: **113** (**276** CPUs Total) Hosts down: **1**							
cerbero (physical view)	111.72	111.80	112.15	65.4	2.1	1.5	29.7
Cluster Localtime: **July 2, 2006, 9:19 pm** Hosts up: **70** (**188** CPUs Total) Hosts down: **0**							
helium (physical view)	4.00	4.00	3.75	28.6	0.0	0.0	71.4
Cluster Localtime: **July 2, 2006, 9:19 pm** Hosts up: **7** (**16** CPUs Total) Hosts down: **0**							
briareo (physical view)	8.73	8.49	8.35	12.4	0.0	0.4	92.1
Cluster Localtime: **July 2, 2006, 9:19 pm** Hosts up: **29** (**58** CPUs Total) Hosts down: **0**							

CLUSTER MANAGEMENT
Ganglia at work /2



STORAGE
Shared and Parallel Filesystems

A shared filesystem to ease management and supply a centralized repository:

Performance is not an issue!

- NFS – Network File System

A filesystem to deal with intensive I/O operations both serial and parallel (parallel filesystem).

Available choices:

Performance IS an issue!

- GFS – Global File System
- GPFS – Global Parallel File System
- PVFS – Parallel Virtual File System
- Lustre

STORAGE
Shared Filesystems: NFS

Central repository for:

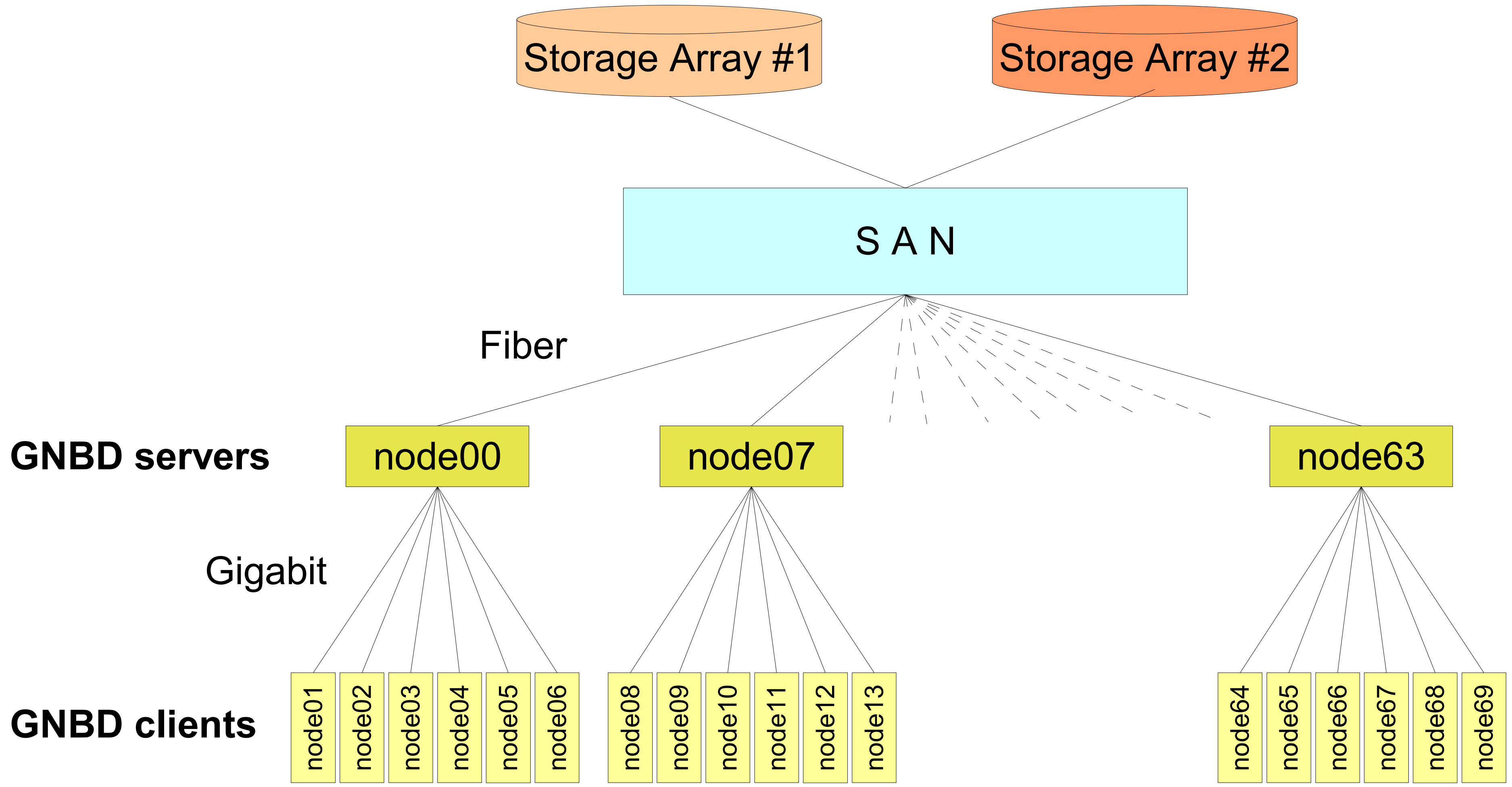
- packages (installation/updates)
- cluster-wide configurations
- libraries
- non-critical executables (not needed at boot-up)
- sporadic non I/O intensive operations
- ...

Can supply the root filesystem (and/or UnionFS layers) for disk-less nodes and can export the /home filesystem as well.

STORAGE
Parallel Filesystem: GFS

- works in a SAN/LAN environment
- single system image style view of the filesystem
- fully 64bit
- journaled
- works with LVM volume managers
- scalable

STORAGE
Parallel Filesystem: GFS



STORAGE
Parallel Filesystem: GFS

Quorum – CMAN (Cluster MANager)

- gives the ability to maintain the cluster through flaky network spikes
- it will keep you from prematurely removing nodes from the cluster
- each node has configurable number of votes
- doesn't scale well
- if the quorum is lost, the filesystem becomes unavailable and most cluster applications (GFS related) will not operate...

Fencing

- ensures data integrity of shared storage devices
- makes sure that a node is gone before recovering data (power fencing!)
- if heartbeats among machines are lost, the nodes will attempt to fence each other...

> Note: we wrote our own fence agents (bash and perl scripts) that interact with a small utility, *blade*, that allows remote hardware control of the blade chassis.

Locking – DLM (Distributed Lock Manager)

- ensures that nodes in the cluster who share the data on the SAN don't corrupt each other's data

Device mapper - LVM2 (Logical Volume Manager, GFS-aware)

- handle physical volumes providing software RAID (striping, mirroring)

RESOURCES MANAGEMENT

We have a pool of users and a pool of resources, then what?

- some software that control available resources
- some other software that decide which application to execute based on available resources
- some other software devoted to actually execute applications

The resource manager allows:

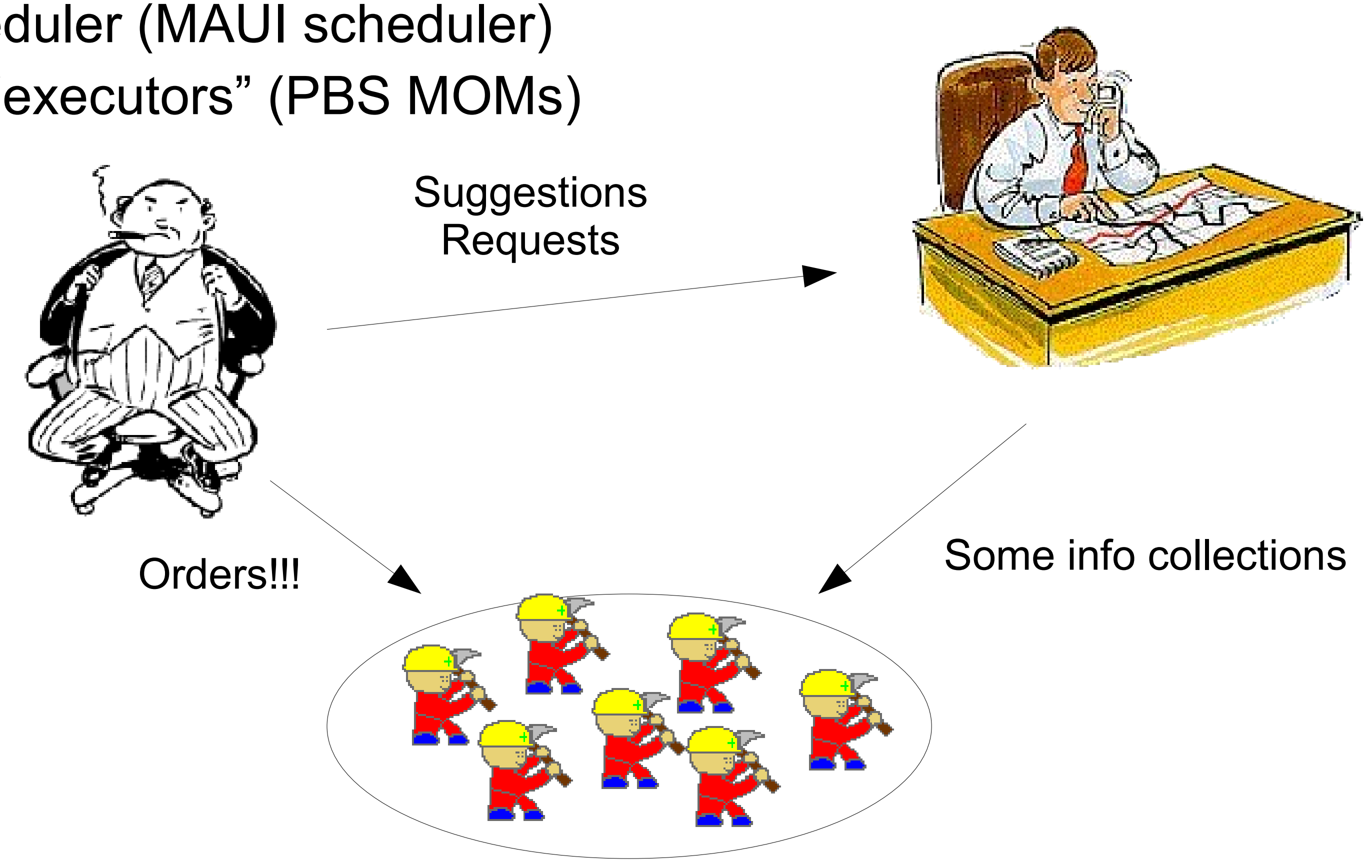
- better resource control
- better resource utilization
- better access control

The scheduler should have:

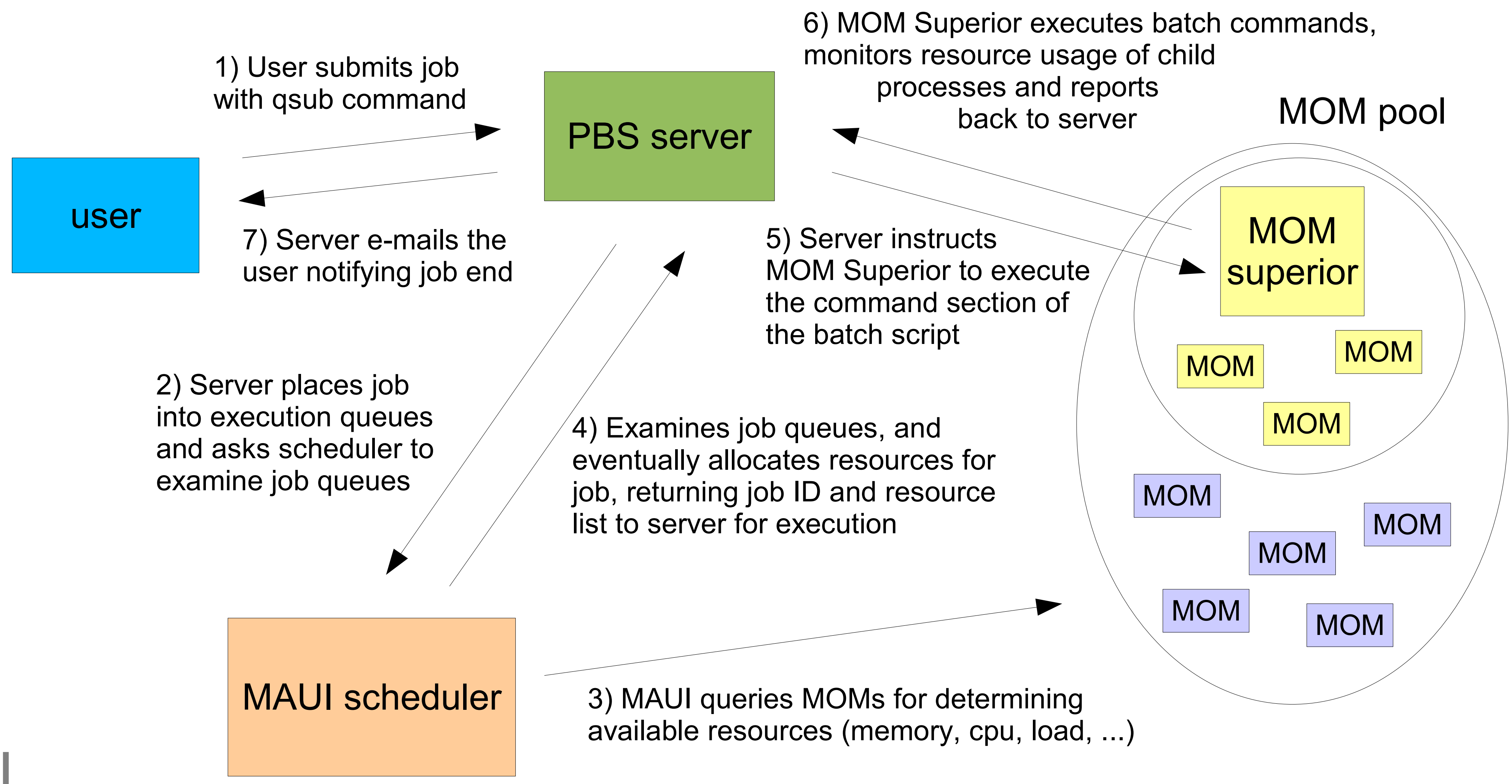
- **Fair Share mechanism**
- **Backfill scheduling algorithm**
- reservations for high priority jobs
- more control parameters on users
- commands for querying the scheduler

RESOURCES MANAGEMENT
The Queue System - PBS/TORQUE + MAUI

- General Components
  - A resource manager (PBS server)
  - A scheduler (MAUI scheduler)
  - Many “executors” (PBS MOMs)



RESOURCES MANAGEMENT
A typical job session



RESOURCES MANAGEMENT
Fair sharing

Fairshare is a mechanism which allows historical resource utilization information to be incorporated into job feasibility and priority decisions.

Fairshare information only affects the job's priority relative to other jobs.

Using the standard fairshare target

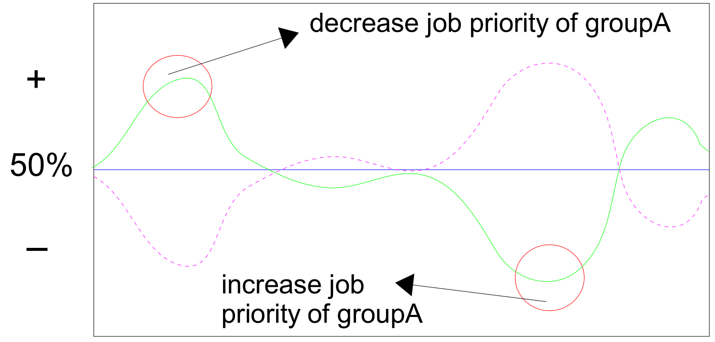
- the priority of jobs of a particular group which has used too many resources over the specified fairshare window is lowered.
- the priority of jobs which have not received enough resources will be increased

RESOURCES MANAGEMENT
Fair sharing – How it works

- At the beginning all the jobs are created equals (in term of priority)
- However some jobs are more/less equal than others
- Priority is increased/decreased when the fair sharing quota is below/above from its target
- Gain/lost in priority:
  - is configurable
  - 1% far from fair share means 4 hours on queues (DEMOCRITOS example)

GROUPCFG[groupA] FSTARGET=50% PRIORITY=5000
GROUPCFG[groupB] FSTARGET=50% PRIORITY=5000

Assume groupA has 50% of fairshare usage.
When it use more resources than those assigned, the priority of the jobs will be decreased; when it uses less resources, the priority of its jobs will be increased.
When a group is not computing, other group can benefit from the available resources



- better resource utilization
  - no idle CPUs

RESOURCES MANAGEMENT
Backfill /1

Backfill is a scheduling optimization which allows a scheduler to make better use of available resources by running jobs out of order.

Consider this example with a 10 CPUs machine:

Job1 (priority=20 walltime=10 nodes=6)
Job2 (priority=50 walltime=30 nodes=4)
Job3 (priority=40 walltime=20 nodes=4)
Job4 (priority=10 walltime=10 nodes=1)

1) When Maui schedules, it prioritizes the jobs in the queue according to a number of factors and then orders the jobs into a 'highest priority first' sorted list.

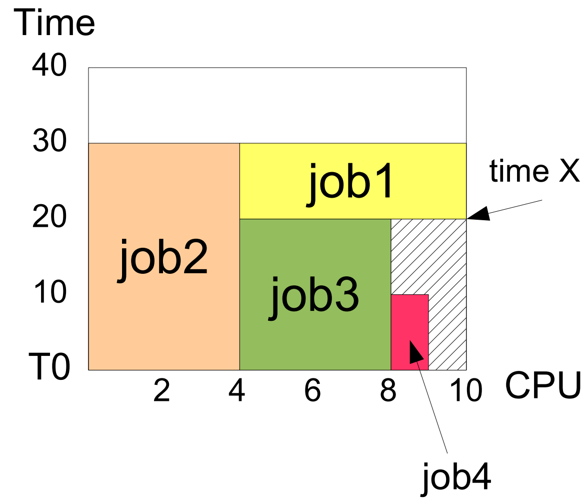
Sorted list:

Job2 (priority=50 walltime=30 nodes=4)
Job3 (priority=40 walltime=20 nodes=4)
Job1 (priority=20 walltime=10 nodes=6)
Job4 (priority=10 walltime=10 nodes=1)

RESOURCES MANAGEMENT
Backfill /2

2) It starts the jobs one by one stepping through the priority list until it reaches a job which it cannot start.

3) All jobs and reservations possess a start time and a wallclock limit, so MAUI can determine:
   - the completion time of all jobs in the queue
   - the earliest the needed resources will become available for the highest priority job to start (time X)
   - which jobs can be started without delaying this job (job4)



→ Enabling backfill allows the scheduler to start other, lower-priority jobs so long as they do not delay the highest priority job, essentially filling in holes in node space.

Job2 (priority=50 walltime=30 nodes=4)
Job3 (priority=40 walltime=20 nodes=4)
Job1 (priority=20 walltime=10 nodes=6)
Job4 (priority=10 walltime=10 nodes=1)

→ Backfill offers significant scheduler performance improvement:
- increased system utilization by around 20% and improved turnaround time by an even greater amount in a typical large system
- backfill tends to favor smaller and shorter running jobs more than larger and longer running ones: It is common to see over 90% of these small and short jobs backfilled.

COMPUTATIONAL SOFTWARE

Compilers

- INTEL → icc, ifc/ifort
- PGI → pgcc, pgf77
- GNU → gcc, g77, g95

Scientific Libraries

- BLAS / LAPACK / ScaLAPACK / ...
- ATLAS / ACML (optimized)
- FFTW

Parallel Environment

- MVAPICH (MPI over InfiniBand)

COMPUTATIONAL SOFTWARE
User's Environment

How can the complexity of an heterogeneous compilation environment be handled?

- shell variables set by system (of all the nodes) in:
  - /etc/profile
  - /etc/csh.login, /etc/csh.cshrc
  - /etc/bashrc
- and consider files in /etc/profile.d/
- shell variables set by users in users' profile files:
  - $HOME/.bash_profile, $HOME/.bashrc
  - $HOME/.tchsrc
- for new users, modify prototype profile files in /etc/skel/

What if one needs to change the environment during the same session?

```
$ export PATH=/some/bin/dir/:/some/other/bin/dir/:$PATH
$ export LD_LIBRARY_PATH=/some/lib/dir/:/some/other/lib/dir/:$LD_LIBRARY_PATH
$ export SOME_LICENCE_FILE=/some/license/file
$ export VOODOO_ENV_VAR=1
...
```

COMPUTATIONAL SOFTWARE
User's Environment

Modules Environment Project

http://modules.sourceforge.net/

“The Modules package is a set of scripts and information files that provides a simple command interface for modifying the environment.”

- The administrator can setup some configuration files (in TCL) that allows *module* (when invoked) to set the needed environment variables for the running shell.
- Users can configure their own *modulefiles* with personalized environments and can switch environment with just few user-friendly commands.

```
$ module avail
---------------- /opt/modules-3.1.6/versions -----------------
3.1.6
--------------- /opt/modules-3.1.6/modulefiles ---------------
gnu                  mpi                  mpich-intel-p4     pgi-6.05
icc-9.0              mpich-gnu-gm         mpich-intel-shmem  pgi-6.12
icc64-9.0            mpich-gnu-p4         mpich-pgi-gm
ifc-9.0              mpich-gnu-shmem      mpich-pgi-p4
ifc64-9.0            mpich-intel-gm       mpich-pgi-shmem
```

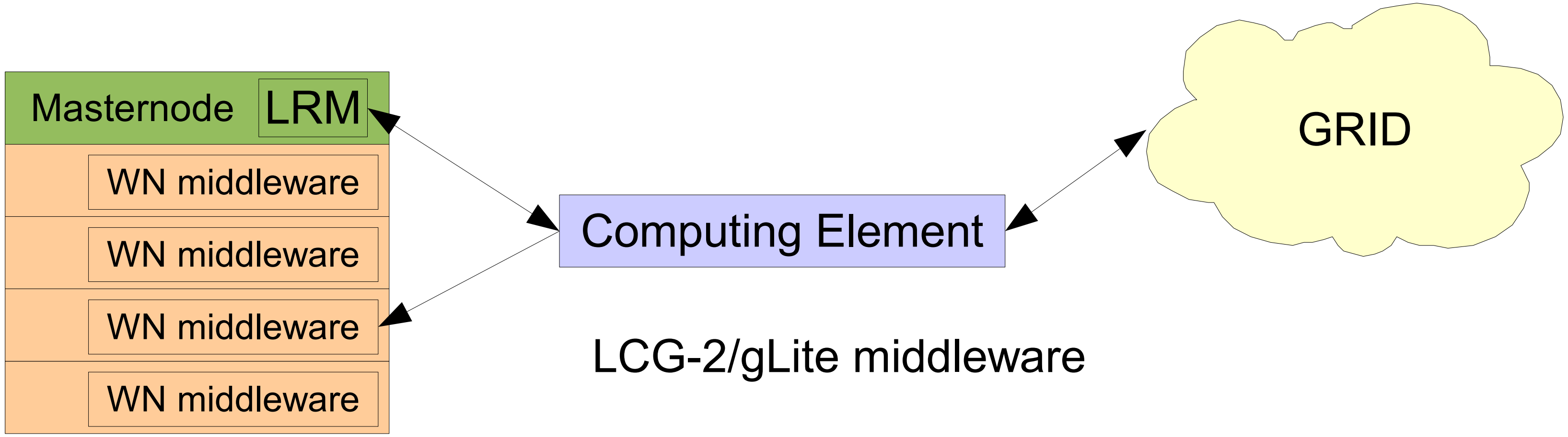
```
$ module load icc-9.0
$ module load ifc-9.0
$ module load mpich-intel-gm
$ module list
Currently Loaded Modulefiles:
  1) icc-9.0   2) ifc-9.0   3) mpich-intel-gm
$ module unload icc-9.0 ifc-9.0
$ module load icc64-9.0 ifc64-9.0
$ module list
Currently Loaded Modulefiles:
  1) mpich-intel-gm   2) icc64-9.0   3) ifc64-9.0
```

GRID ENVIRONMENT
World Wide Computing

EGEE II

The Michelangelo cluster is integrated into the LCG-2/gLite GRID using the following mechanism:

- an external server is acting as a Computing Element (CE)
  - standard Scientific LINUX 3.0.6 + LCG-2/gLite middleware
  - the Local Resource Manager system used by the CE is on the masternode (PBS/Torque)
- Each node of the cluster has installed the Worker Node (WN) middleware needed to run jobs coming from the CE.



ACKNOWLEDGMENS

LITBIO
Laboratorio Interdisciplinare di Tecnologie Bioinformatiche

The Laboratory is funded by Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR - Italy) through a FIRB 2003 grant for a period 2005 - 2010.

That's All Folks!


```
( questions ; comments ) | mail baro@democritos.it -s uheilaaa

( complaints ; insults ) &>/dev/null
```

REFERENCES AND USEFUL LINKS /1

•Hopeless, a system for building disk-less clusters (Christian Pellegrin, November 2005)
http://sole.infis.univ.ts.it/~chri/hopeless.html
•CentOS - RH-based Linux distribution
http://www.centos.org/
•UnionFS - A Stackable Unification File System
http://www.unionfs.org
http://www.fsl.cs.sunysb.edu/project-unionfs.html

Cluster File Systems:
•CLUSTER/GFS - RH Cluster Suite and the Global File System
http://sources.redhat.com/cluster/
http://sources.redhat.com/cluster/gfs/
•PVFS - The Parallel Virtual File System
http://www.parl.clemson.edu/pvfs/
•Lustre
http://www.lustre.org/
•GPFS - The IBM Global Parallel File System
http://www.ibm.com/servers/eserver/clusters/software/gpfs.html

Management Tools:
•openssh/openssl
http://www.openssh.com
http://www.openssl.org
•C3 tools - The Cluster Command and Control tool suite
http://www.csm.ornl.gov/torc/C3/
•DSH - Distributed SHell
http://www.netfort.gr.jp/~dancer/software/dsh.html.en

Cluster Toolkits:
•OSCAR - Open Source Cluster Application Resources
http://oscar.openclustergroup.org/
•NPACI Rocks
http://www.rocksclusters.org/
•Scyld Beowulf
http://www.beowulf.org/
•CSM - IBM Cluster Systems Management
http://www.ibm.com/servers/eserver/clusters/software/
•xCAT - eXtreme Cluster Administration Toolkit
http://www.xcat.org/
•OpenSCE - Open Scalable Cluster Environment
http://www.opensce.org/
•Warewulf
http://www.warewulf-cluster.org/

Resources Management:
•MAUI - Cluster Scheduler / TORQUE - Resource Manager
http://www.clusterresources.com/pages/products.php
•PBS/OpenPBS - Portable Batch System
http://www.openpbs.org/
•SGE – Sun Grid Engine
http://gridengine.sunsource.net/

Monitoring Tools:
•Ganglia
http://ganglia.sourceforge.net/
•Nagios
http://www.nagios.org/
•Zabbix
http://www.zabbix.org/

REFERENCES AND USEFUL LINKS /2

Compilers:
•GNU – gcc/g77
http://gcc.gnu.org/
•G95 – GNU f95 Compiler
http://www.g95.org/
•PGI – Portland Group
http://www.pgroup.com/
•Intel – icc/ifort
http://www.intel.com/
•NAG – Numerical Algorithms Group
http://www.nag.com/

Scientific Libraries:
•Netlib Repository
http://www.netlib.org/
•LAPACK - Linear Algebra PACKage
http://www.netlib.org/lapack/
•ScaLAPACK – Scalable LAPACK
http://www.netlib.org/scalapack/
•BLAS - Basic Linear Algebra Subprograms
http://www.netlib.org/blas/
•ATLAS - Automatically Tuned Linear Algebra Software
http://math-atlas.sourceforge.net/
•FFTW - Fastest Fourier Transform in the West
http://www.fftw.org/
•ACML - AMD Core Math Library
http://developer.amd.com/acml.aspx
•MKL – Intel Math Kernel Library
http://www.intel.com/

Modules - Environment Modules Project
http://modules.sourceforge.net/

Parallel Environment:
•MPI - The Message Passing Interface standard
http://www-unix.mcs.anl.gov/mpi/
•OpenMPI – A High Performance Message Passing Library
http://www.open-mpi.org/
•LAM / MPI – Parallel Computing
http://www.lam-mpi.org/
•PVM – Parallel Virtual Machine
http://www.csm.ornl.gov/pvm/

GRID Projects:
•EGEE II
http://www.eu-egee.org/
•CERN Datagrid
http://eu-datagrid.web.cern.ch/eu-datagrid/
•GRID.IT
http://www.grid.it/
•EGRID
http://www.egrid.it/

GRID Middleware
•LCG-2 / gLite
http://lcg.web.cern.ch/LCG/
http://glite.web.cern.ch/
•GLOBUS
http://www.globus.org/

Some acronyms...

HPC – High Performance Computing

OS – Operating System
LINUX – LINUX is not UNIX
GNU – GNU is not UNIX

PXE – Preboot Execution Environment
DHCP – Dynamic Host Configuration Protocol
TFTP – Trivial File Transfer Protocol
NFS – Network File System
INITRD – INITial RamDisk

LDAP – Lightweight Directory Access Protocol
NIS – Network Information Service
DNS – Domain Name System
NTP – Network Time Protocol

LAPACK - Linear Algebra PACKage
ScaLAPACK - Scalable LAPACK
BLAS - Basic Linear Algebra Subprograms
ATLAS - Automatically Tuned Linear Algebra Software
FFTW - Fastest Fourier Transform in the West
ACML - AMD Core Math Library

PVM – Parallel Virtual Machine
MPI – Message Passing Interface
MPICH – Message Passing Interface/CHameleon
MVAPICH – MPI over VAPI
VAPI – Verbs Level Interface

PBS – Portable Batch System
MOM – Machine Oriented Mini-server

EGEE – Enabling Grids for E-sciencE
LCG – LHC Computing P`roject
LHC – Large Hadron Collider

CE – Computing Element
WN – Worker Node
SE – Storage Element
LRM – Local Resource Manager
GRM – Global Resource Manager

SAN – Storage Area Network
NAS – Network Attached Storage

DEMOCRITOS – Democritos Modeling Center for Research In aTOmistic Simulations
INFM – Istituto Nazionale per la Fisica della Materia (Italian National Institute for the Physics of Matter)
CNR – Consiglio Nazionale delle Ricerche (Italian National Research Council)