MHPC
Master in High Performance Computing

Moreno Baricevic

CNR-IOM DEMOCRITOS
Trieste, ITALY

Installation Procedures for Clusters

PART 1 – Cluster Services and Installation Procedures

ma per seguir virtute e canoscenza
SISSA
Scuola Internazionale Superiore
di Studi Avanzati

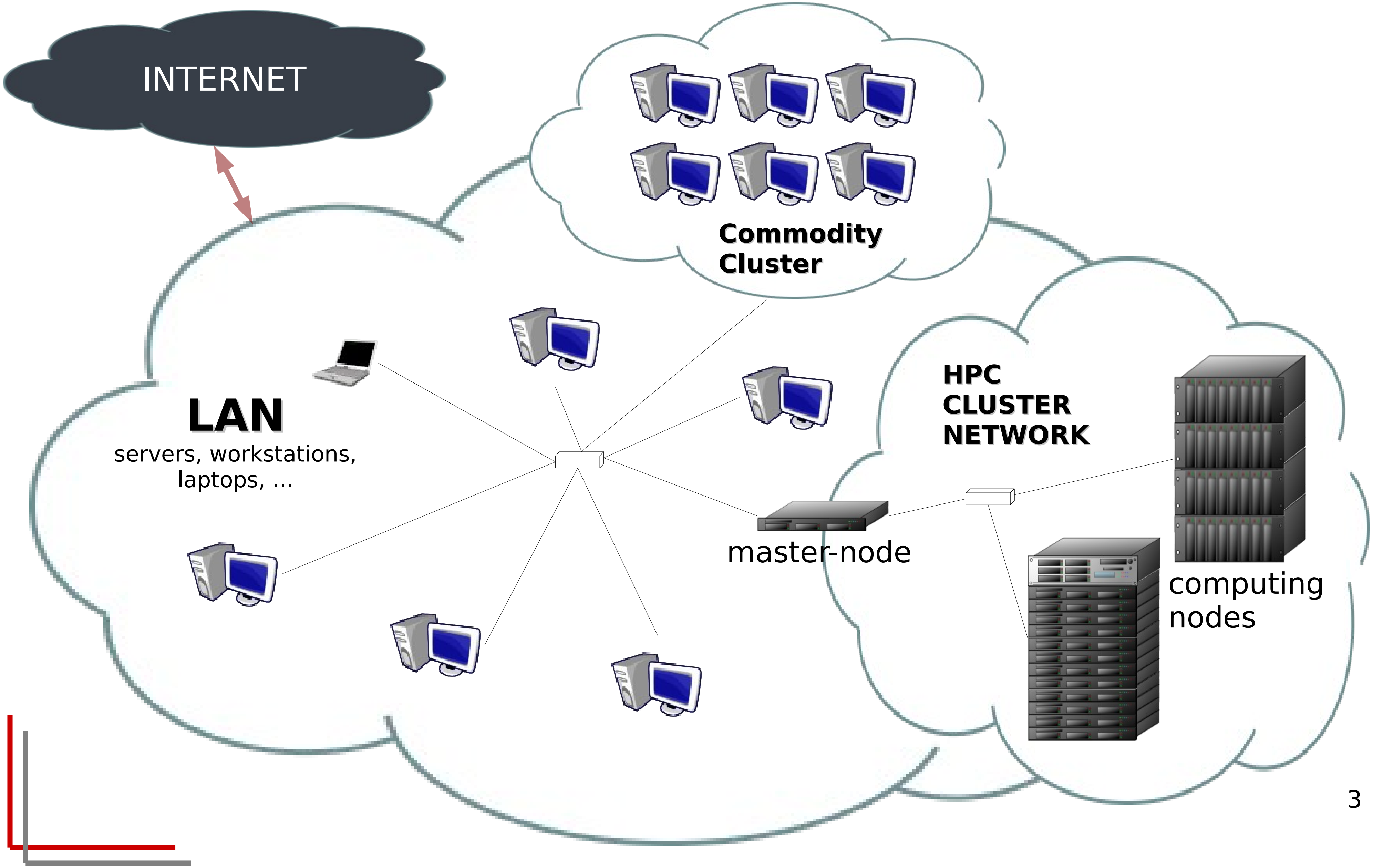
ICTP
The Abdus Salam
International Centre
for Theoretical Physics

Agenda

- Cluster Services
- Overview on Installation Procedures
- Configuration and Setup of a NETBOOT Environment
- Troubleshooting
- Cluster Management Tools
- Notes on Security

- Hands-on Laboratory Session

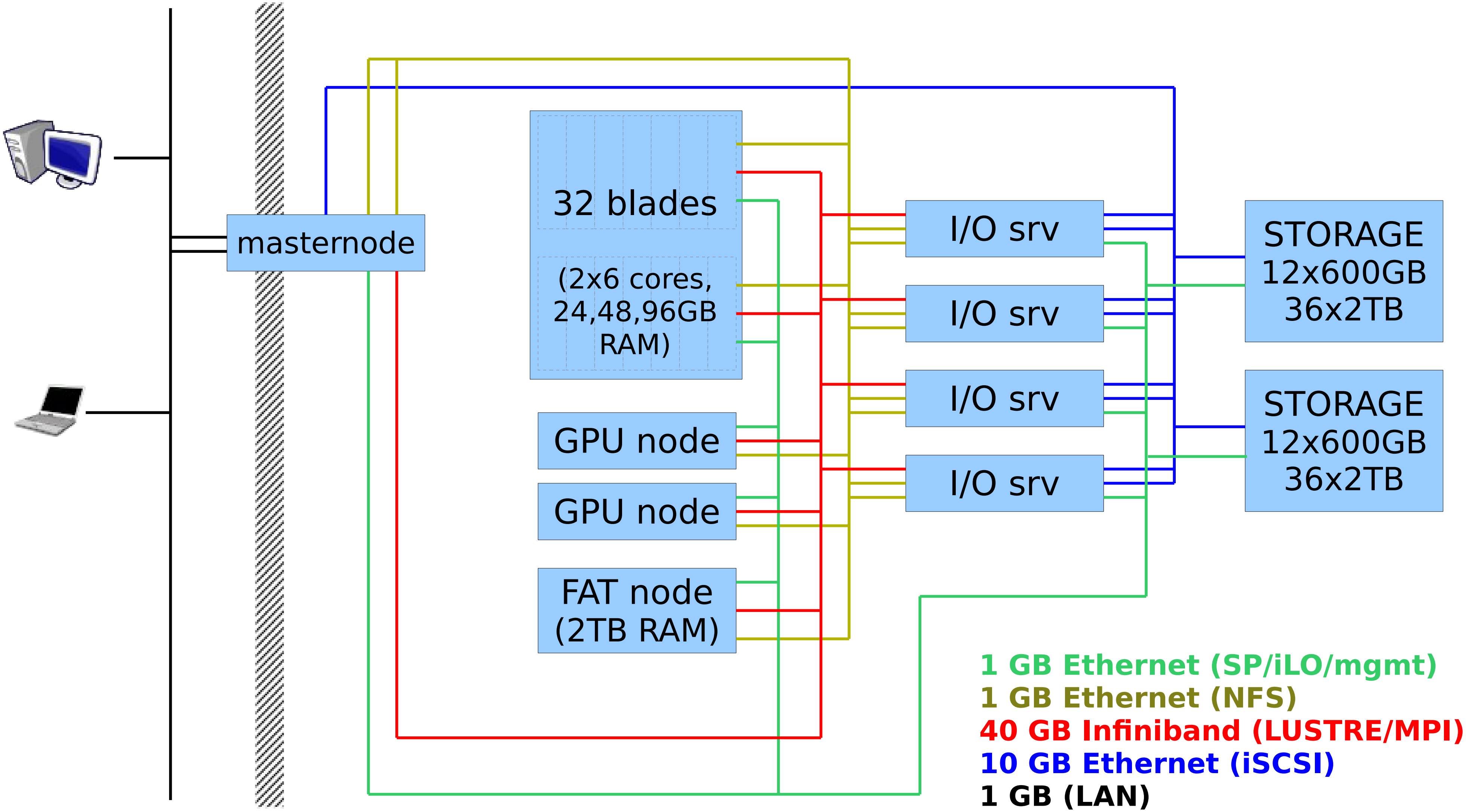
What's a cluster?



What's a cluster?

- A cluster **needs**:
  - Several computers, nodes, often in special cases for easy mounting in a rack
  - One or more networks (interconnects) to hook the nodes together
  - Software that allows the nodes to communicate with each other (e.g. MPI)
  - Software that reserves resources to individual users
- A cluster **is**: all of those components working together to form one big computer

Cluster example (internal network)



1 GB Ethernet (SP/iLO/mgmt)
1 GB Ethernet (NFS)
40 GB Infiniband (LUSTRE/MPI)
10 GB Ethernet (iSCSI)
1 GB (LAN)

What's a cluster from the HW side?

LAPTOP

PC / WORKSTATION

RACKs + rack mountable SERVERS

1U Server
(rack mountable)

BLADE Servers

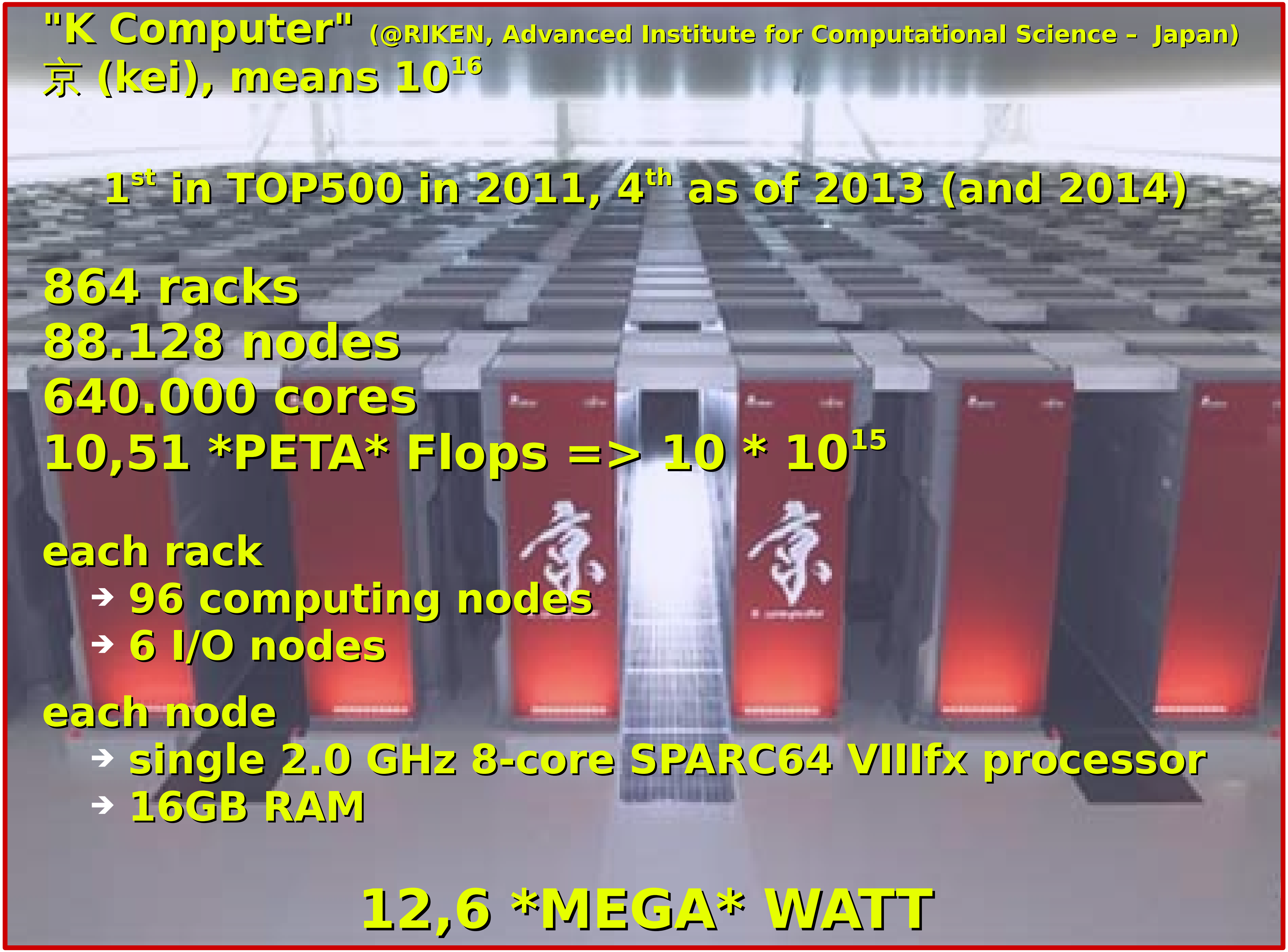
IBM Blade Center
14 bays in 7U ***2x***

:-(

SUN Fire B1600
16 bays in 3U ***5x***

HP c7000
8-16 bays in 10U

What's a cluster from the HW side?



"K Computer" (@RIKEN, Advanced Institute for Computational Science – Japan)

京 (kei), means 10^{16}

1st in TOP500 in 2011, 4th as of 2013 (and 2014)

864 racks
88.128 nodes
640.000 cores
10,51 *PETA* Flops => 10 * 10^{15}

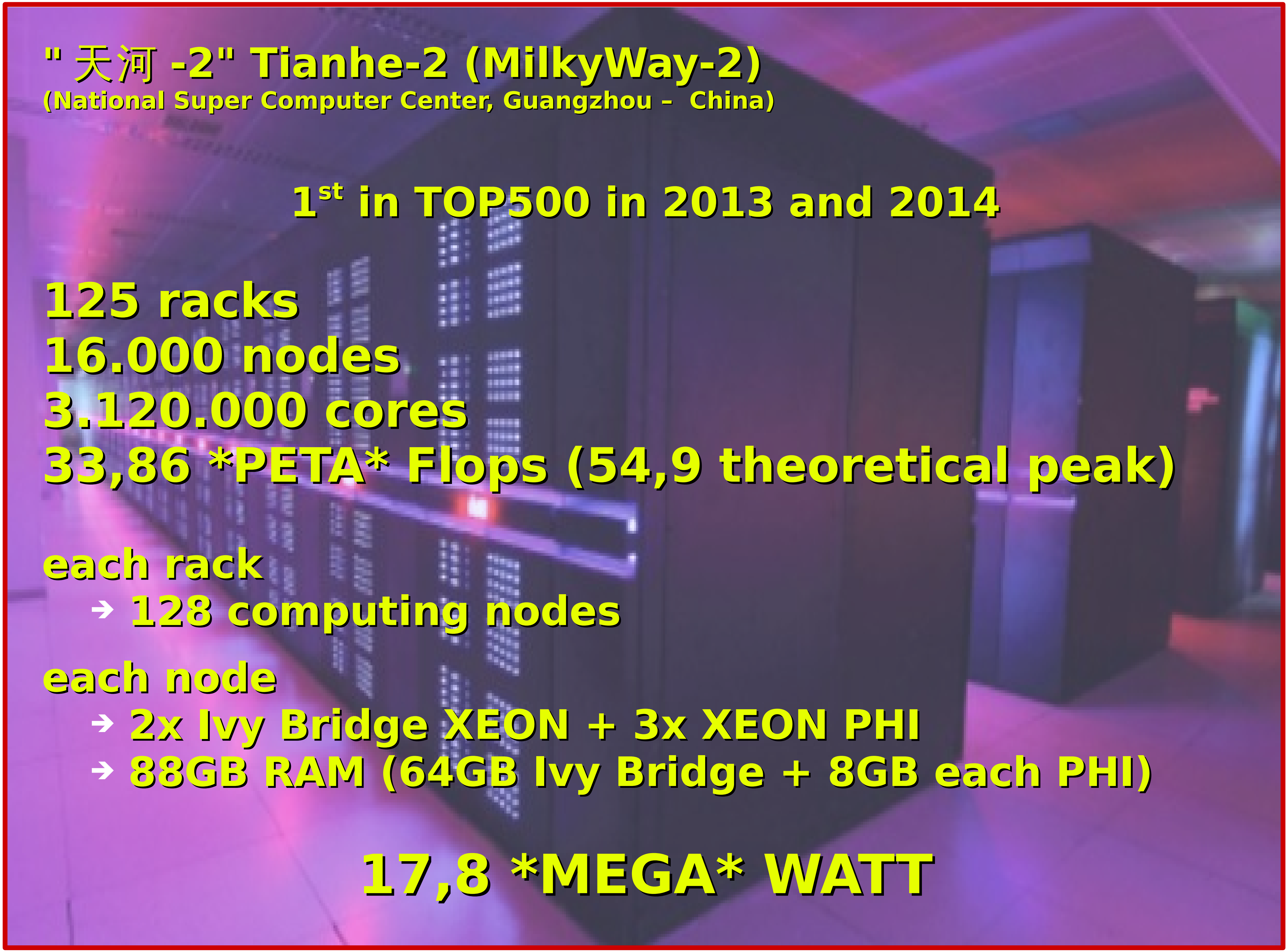
each rack

- 96 computing nodes
- 6 I/O nodes

each node

- single 2.0 GHz 8-core SPARC64 VIIIfx processor
- 16GB RAM

12,6 *MEGA* WATT



" 天河 -2" Tianhe-2 (MilkyWay-2)

(National Super Computer Center, Guangzhou – China)

1st in TOP500 in 2013 and 2014

125 racks
16.000 nodes
3.120.000 cores
33,86 *PETA* Flops (54,9 theoretical peak)

each rack

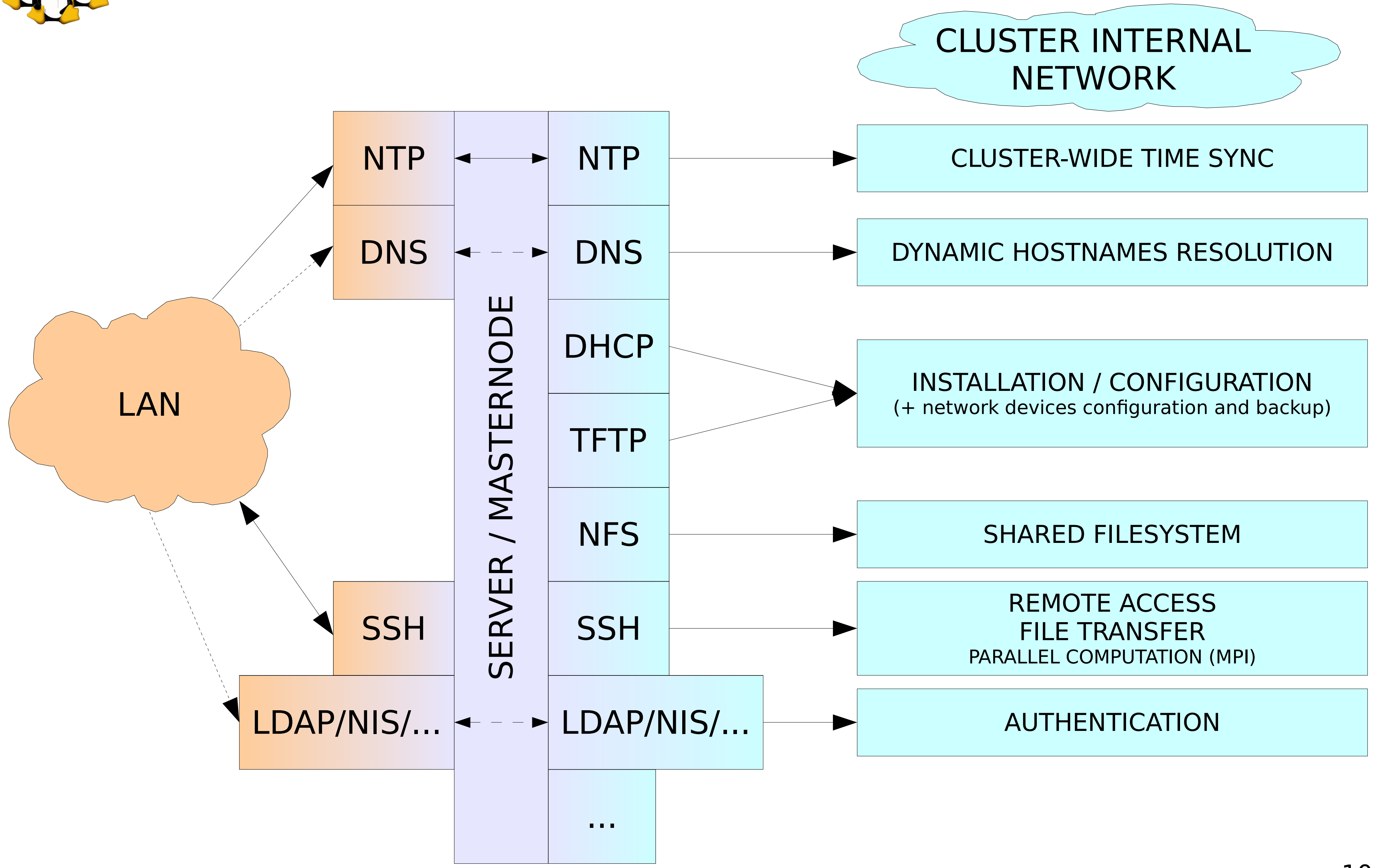
- 128 computing nodes

each node

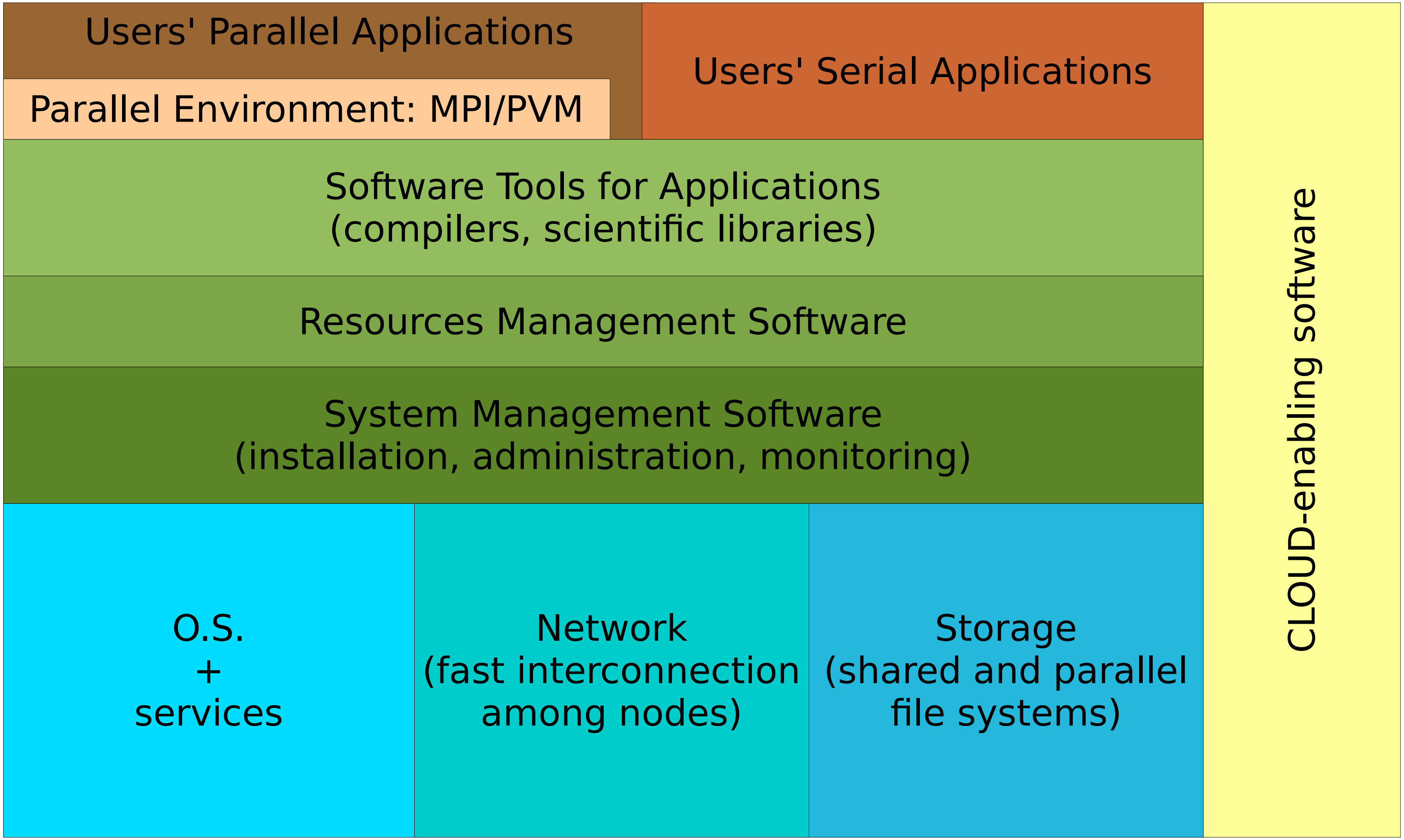
- 2x Ivy Bridge XEON + 3x XEON PHI
- 88GB RAM (64GB Ivy Bridge + 8GB each PHI)

17,8 *MEGA* WATT

CLUSTER SERVICES



HPC SOFTWARE INFRASTRUCTURE
Overview



Users' Parallel Applications

Parallel Environment: MPI/PVM

Users' Serial Applications

Software Tools for Applications
(compilers, scientific libraries)

Resources Management Software

System Management Software
(installation, administration, monitoring)

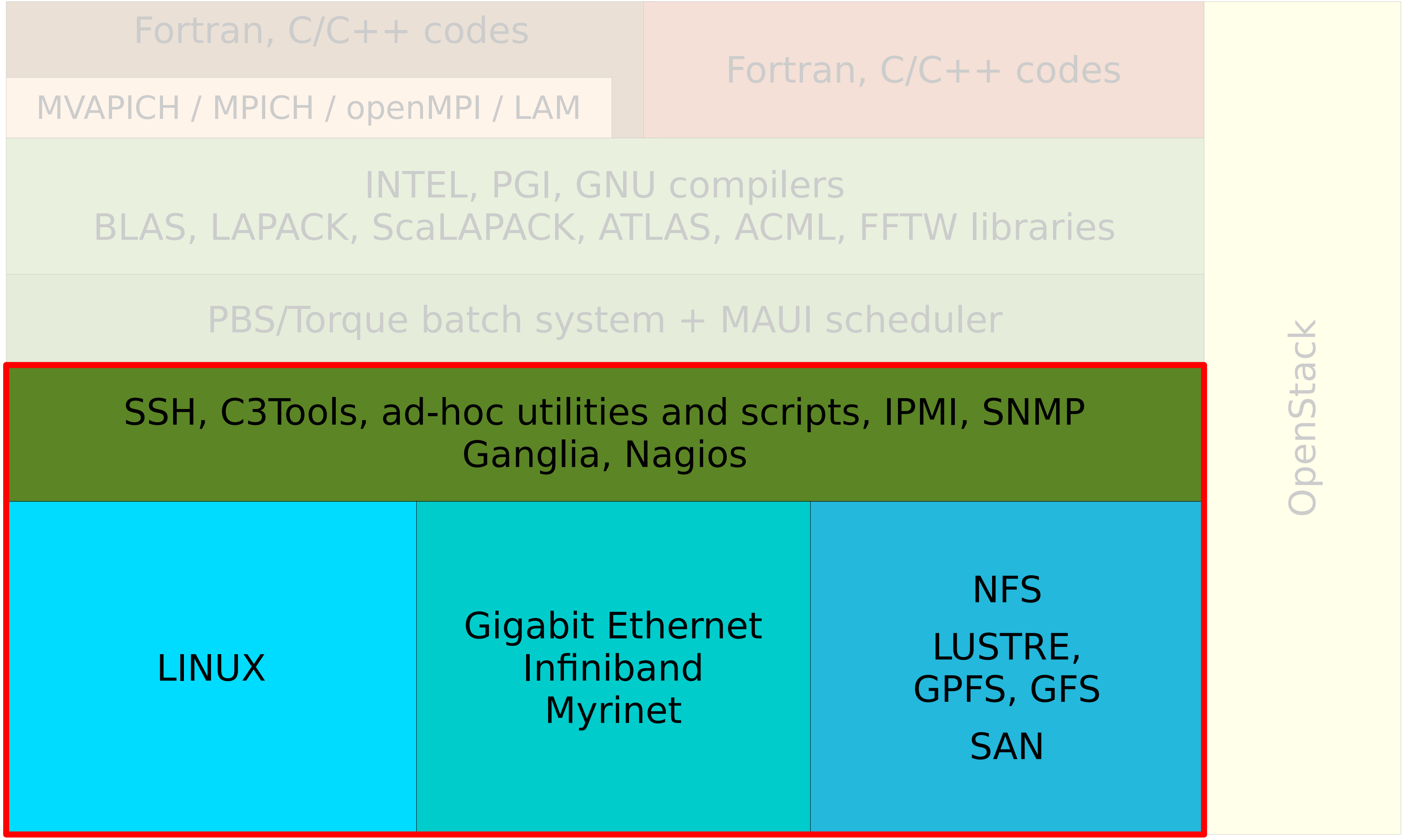
O.S.
+
services

Network
(fast interconnection
among nodes)

Storage
(shared and parallel
file systems)

CLOUD-enabling software

HPC SOFTWARE INFRASTRUCTURE
Overview (our experience)



Fortran, C/C++ codes

MVAPICH / MPICH / openMPI / LAM

Fortran, C/C++ codes

INTEL, PGI, GNU compilers
BLAS, LAPACK, ScaLAPACK, ATLAS, ACML, FFTW libraries

PBS/Torque batch system + MAUI scheduler

SSH, C3Tools, ad-hoc utilities and scripts, IPMI, SNMP
Ganglia, Nagios

LINUX

Gigabit Ethernet
Infiniband
Myrinet

NFS

LUSTRE,
GPFS, GFS

SAN

OpenStack

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/HTTP/FTP)

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 and the like (tar, rsync, ...)

  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/HTTP/FTP

  More efforts to make the first installation work properly (especially for heterogeneous clusters), (mostly) straightforward for the next ones

2) Disk-less nodes

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

CLUSTER MANAGEMENT
Existent toolkits

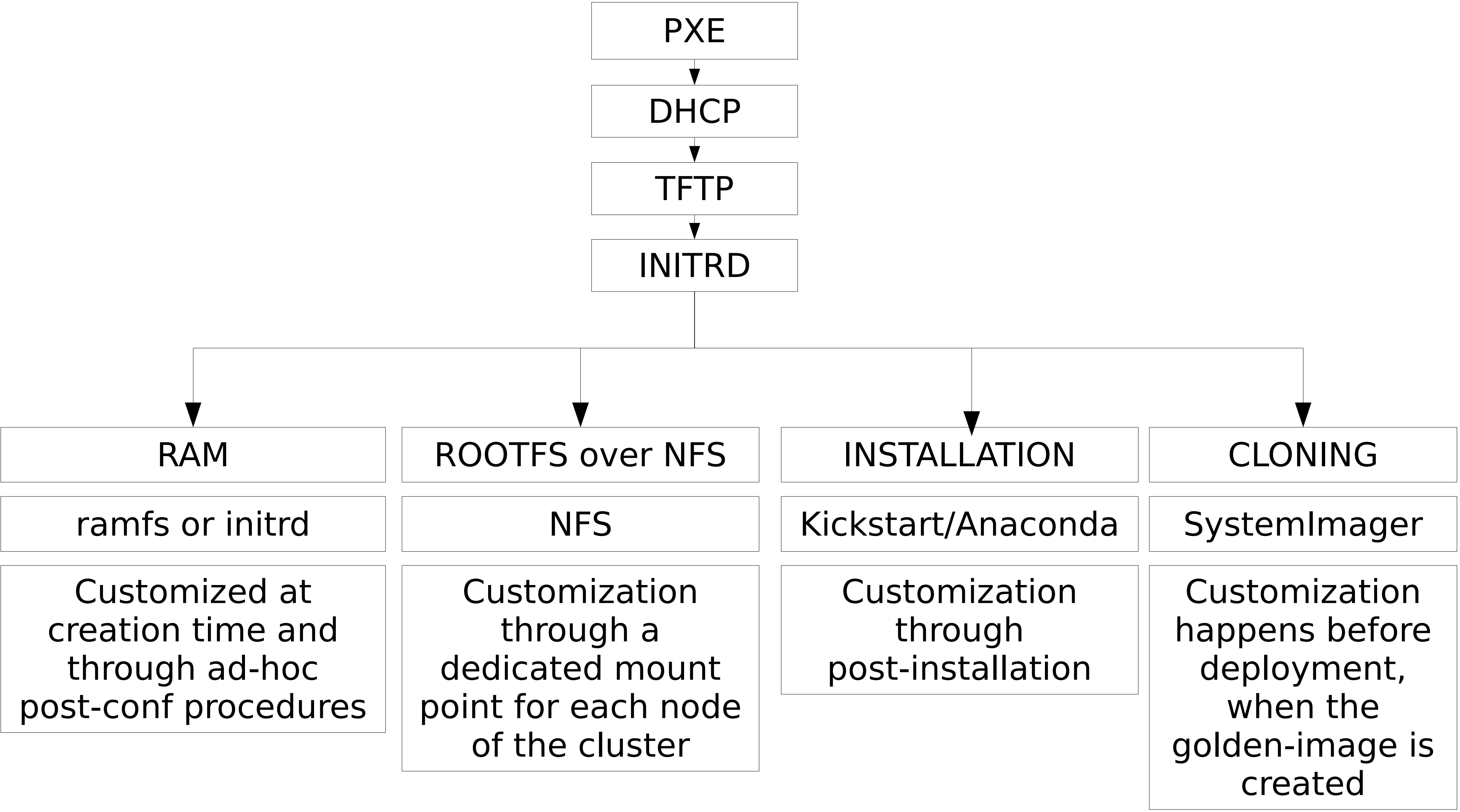
Are generally made of an ensemble of already available software packages thought for specific tasks, but configured to operate together, plus some add-ons.

Sometimes limited by rigid and not customizable configurations, often bound 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)
  - Warewulf/PERCEUS
  - SystemImager
  - Kickstart (RH/Fedora), FAI (Debian), AutoYaST (SUSE)
- Commercial
  - Scyld Beowulf
  - IBM CSM (Cluster Systems Management)
  - HP, SUN and other vendors' Management Software...

Network-based Distributed Installation

Overview



PXE

DHCP

TFTP

INITRD

RAM	ROOTFS over NFS	INSTALLATION	CLONING
ramfs or initrd	NFS	Kickstart/Anaconda	SystemImager
Customized at creation time and through ad-hoc post-conf procedures	Customization through a dedicated mount point for each node of the cluster	Customization through post-installation	Customization happens before deployment, when the golden-image is created

Network-based Distributed Installation
Basic services

Deployment

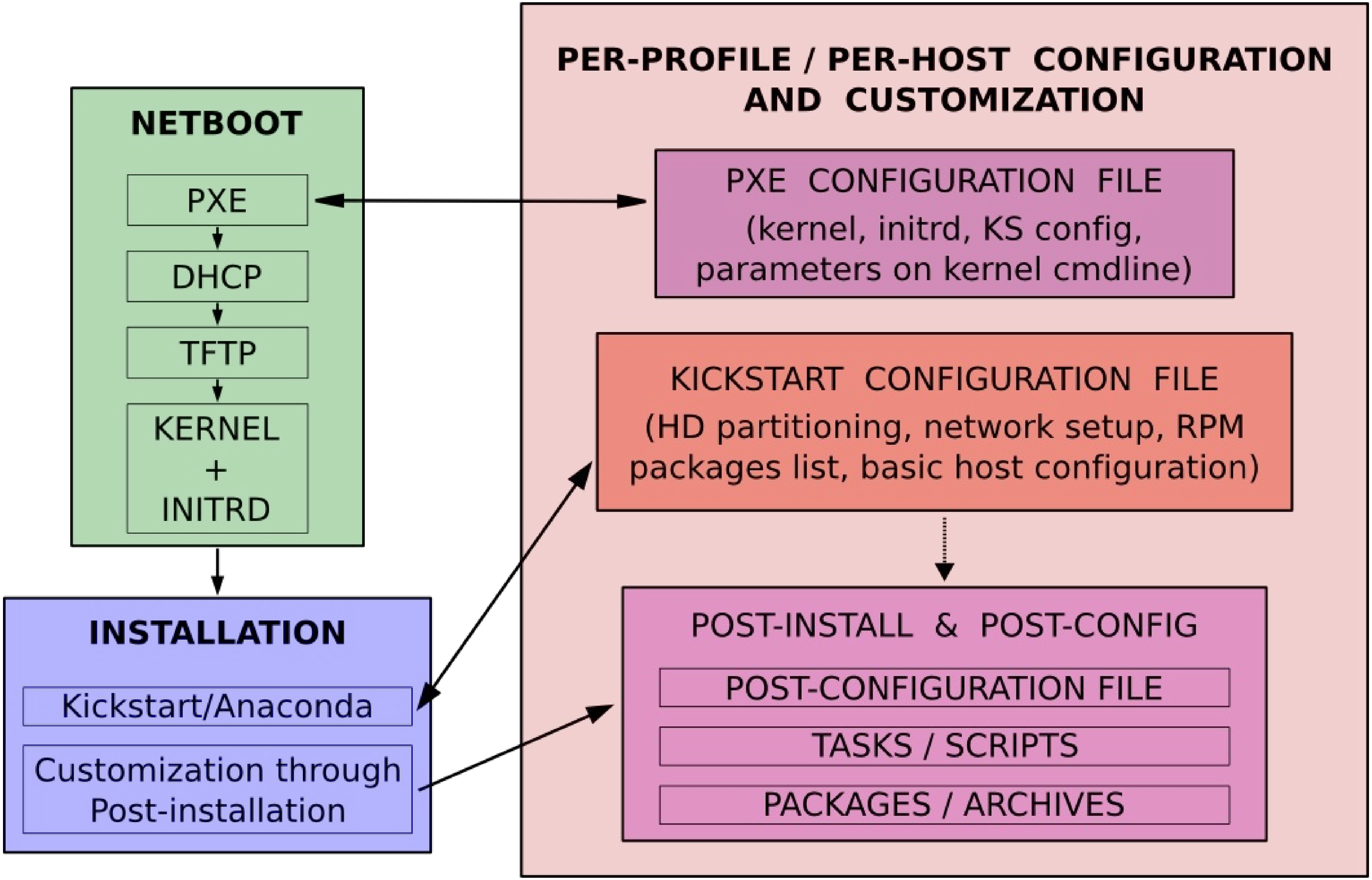
- **PXE**: network booting
- **DHCP**: IP binding + NBP (pxelinux.0)
- **TFTP**: pxe configuration file (pxelinux.cfg/<HEXIP>), alternative boot-up images (memtest, UBCD, ...)
- **NFS**: kickstart + RPM repository (with little modification **HTTP(S)** or **FTP** can be used too)

Maintenance

- passive updates: post-boot updates using port-knocking, ssh, distributed shells, wget, ...
- active configuration/package updates: ssh, distributed shells
- advanced IT automation tools: Ansible, CFEngine, ...

Customization layers

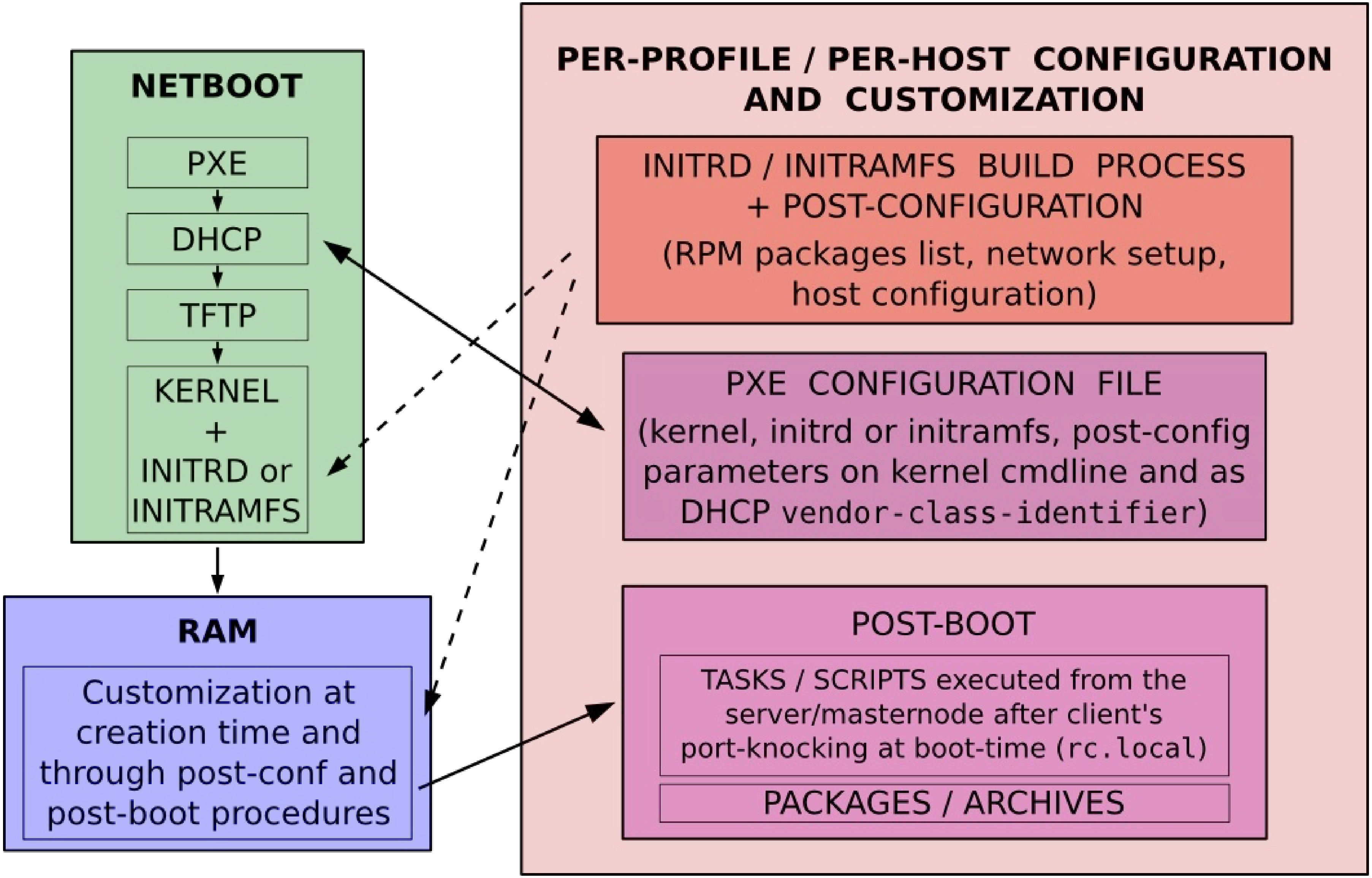
Installation process



NETBOOT
PXE
DHCP
TFTP
KERNEL
+
INITRD
INSTALLATION
Kickstart/Anaconda
Customization through
Post-installation
PER-PROFILE / PER-HOST CONFIGURATION
AND CUSTOMIZATION
PXE CONFIGURATION FILE
(kernel, initrd, KS config,
parameters on kernel cmdline)
KICKSTART CONFIGURATION FILE
(HD partitioning, network setup, RPM
packages list, basic host configuration)
POST-INSTALL & POST-CONFIG
POST-CONFIGURATION FILE
TASKS / SCRIPTS
PACKAGES / ARCHIVES

Customization layers

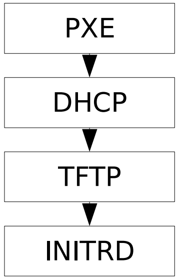
Ramdisk/Ramfs for disk-less nodes, rescue and HW test



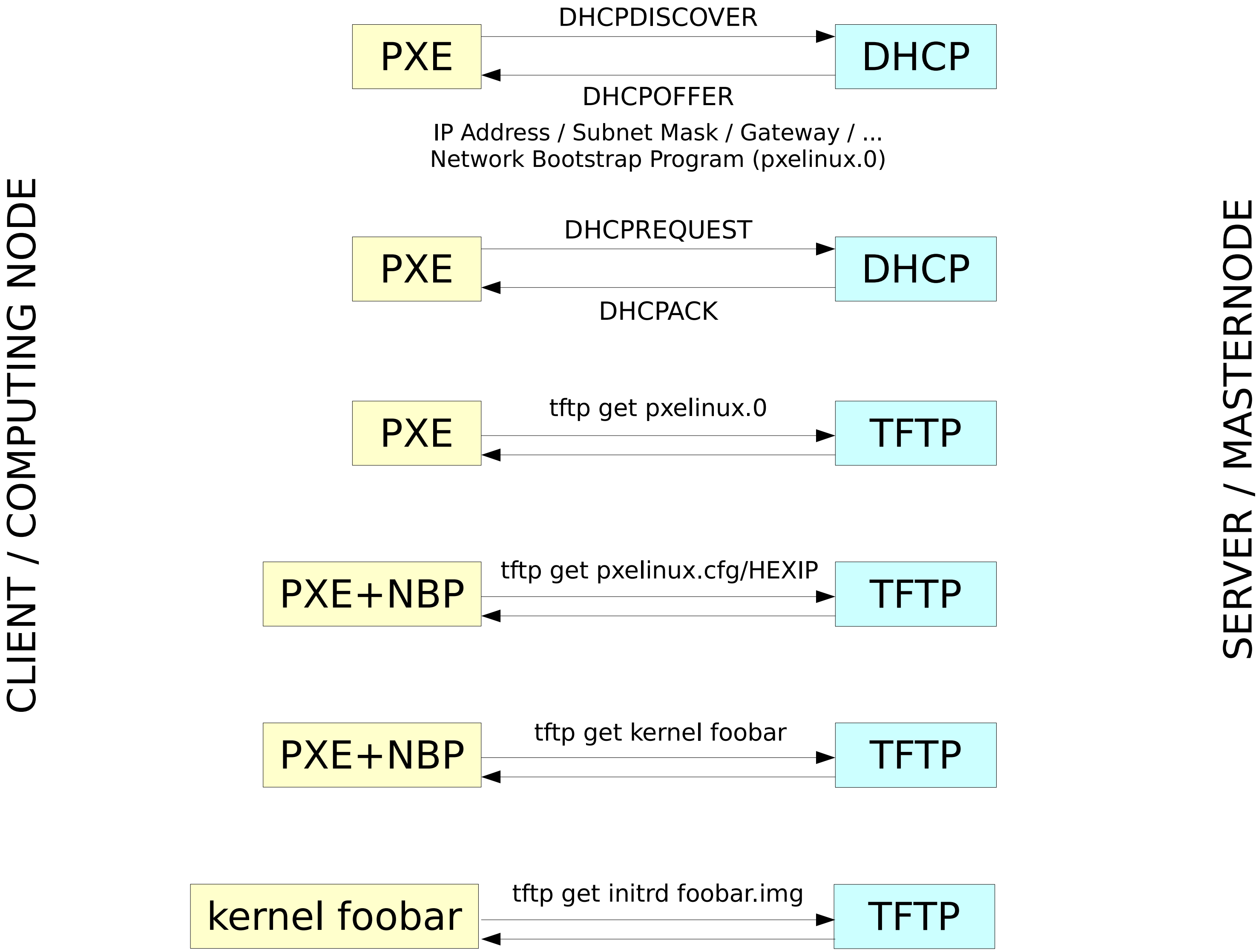
NETBOOT
PXE
DHCP
TFTP
KERNEL + INITRD or INITRAMFS
RAM
Customization at creation time and through post-conf and post-boot procedures
PER-PROFILE / PER-HOST CONFIGURATION AND CUSTOMIZATION
INITRD / INITRAMFS BUILD PROCESS + POST-CONFIGURATION
(RPM packages list, network setup, host configuration)
PXE CONFIGURATION FILE
(kernel, initrd or initramfs, post-config parameters on kernel cmdline and as DHCP vendor-class-identifier)
POST-BOOT
TASKS / SCRIPTS executed from the server/masternode after client's port-knocking at boot-time (rc.local)
PACKAGES / ARCHIVES

Network booting (NETBOOT)

PXE + DHCP + TFTP + KERNEL + INITRD



PXE
DHCP
TFTP
INITRD



CLIENT / COMPUTING NODE

SERVER / MASTERNODE

PXE → DHCP: DHCPDISCOVER
DHCP → PXE: DHCPOFFER

IP Address / Subnet Mask / Gateway / ...
Network Bootstrap Program (pxelinux.0)

PXE → DHCP: DHCPREQUEST
DHCP → PXE: DHCPACK

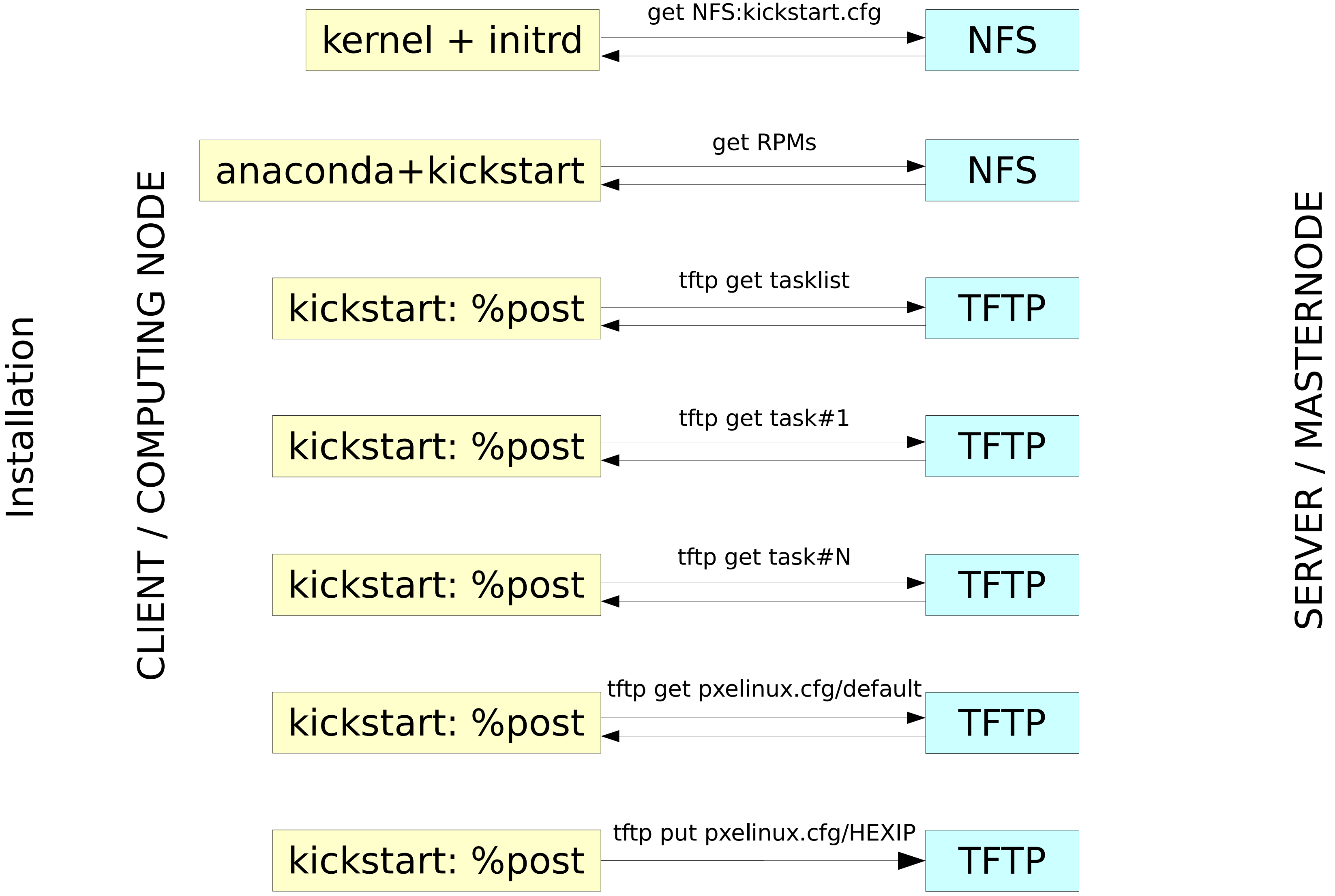
PXE → TFTP: tftp get pxelinux.0

PXE+NBP → TFTP: tftp get pxelinux.cfg/HEXIP

PXE+NBP → TFTP: tftp get kernel foobar

kernel foobar → TFTP: tftp get initrd foobar.img

Network-based Distributed Installation
NETBOOT + KICKSTART INSTALLATION



Installation

CLIENT / COMPUTING NODE

SERVER / MASTERNODE

kernel + initrd — get NFS:kickstart.cfg → NFS

anaconda+kickstart — get RPMs → NFS

kickstart: %post — tftp get tasklist → TFTP

kickstart: %post — tftp get task#1 → TFTP

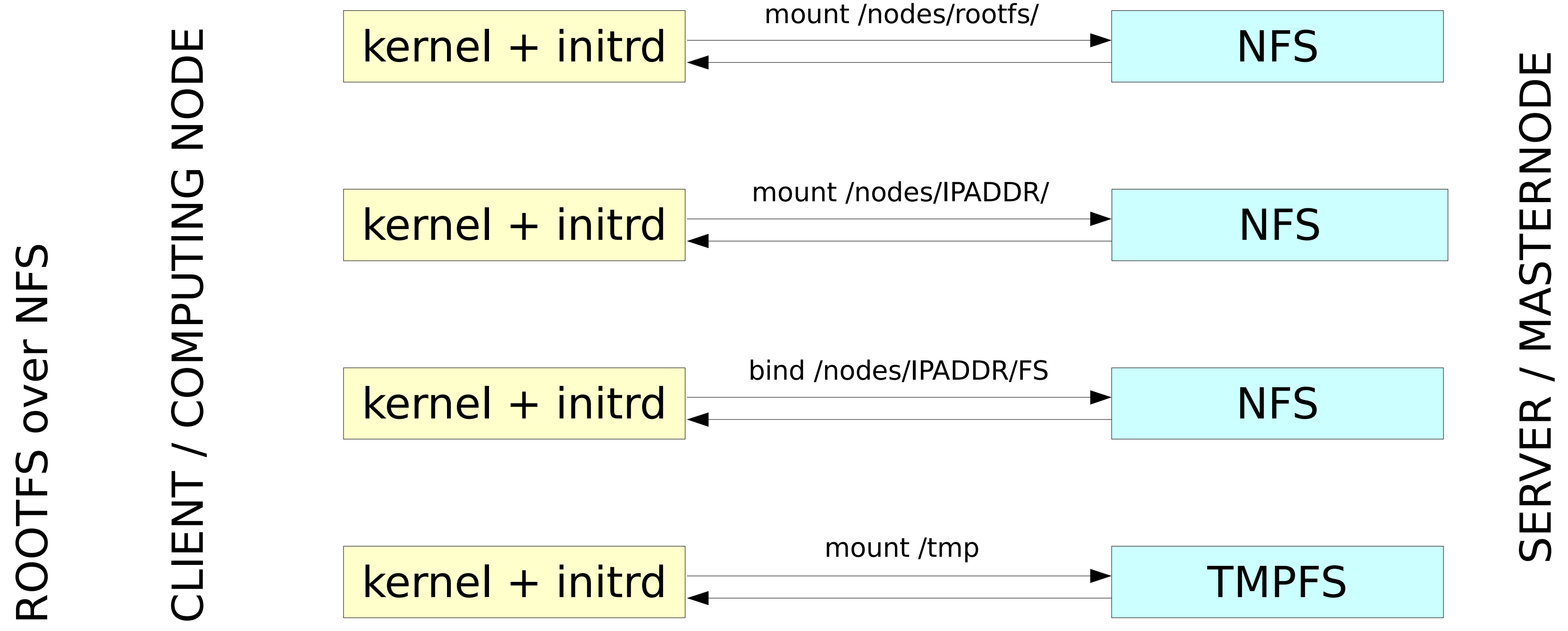
kickstart: %post — tftp get task#N → TFTP

kickstart: %post — tftp get pxelinux.cfg/default → TFTP

kickstart: %post — tftp put pxelinux.cfg/HEXIP → TFTP

Diskless Nodes NFS Based

NETBOOT + NFS



ROOTFS over NFS

CLIENT / COMPUTING NODE

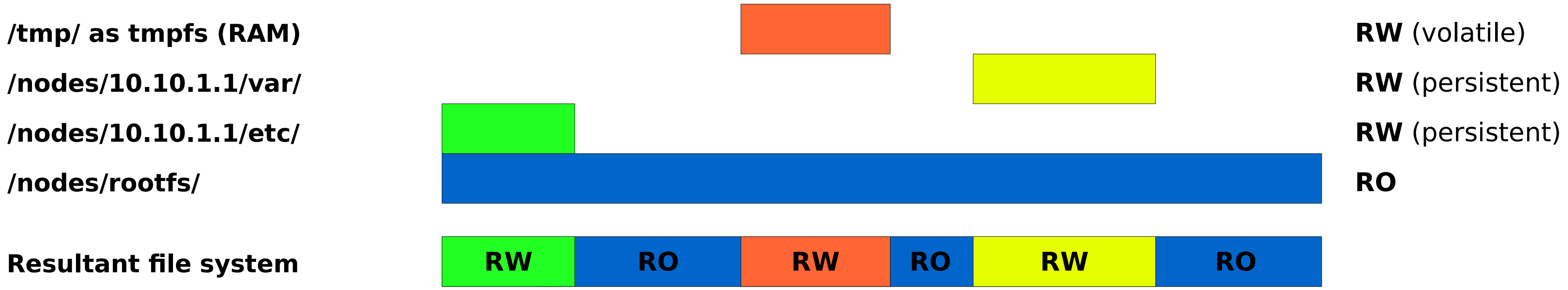
kernel + initrd — mount /nodes/rootfs/ — NFS

kernel + initrd — mount /nodes/IPADDR/ — NFS

kernel + initrd — bind /nodes/IPADDR/FS — NFS

kernel + initrd — mount /tmp — TMPFS

SERVER / MASTERNODE



/tmp/ as tmpfs (RAM) — **RW** (volatile)

/nodes/10.10.1.1/var/ — **RW** (persistent)

/nodes/10.10.1.1/etc/ — **RW** (persistent)

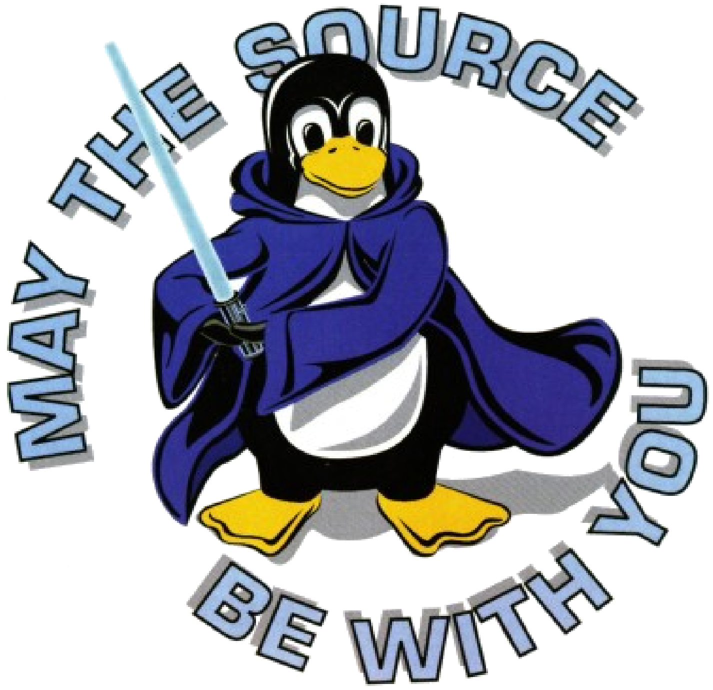
/nodes/rootfs/ — **RO**

Resultant file system — **RW** | **RO** | **RW** | **RO** | **RW** | **RO**

Drawbacks

- Removable media (CD/DVD/floppy):
  - not flexible enough
  - needs both disk and drive for each node (drive not always available)
- ROOTFS over NFS:
  - NFS server becomes a single point of failure
  - doesn't scale well, slow down in case of frequently concurrent accesses
  - requires enough disk space on the NFS server
- RAM disk:
  - need enough memory
  - less memory available for processes
- Local installation:
  - upgrade/administration not centralized
  - need to have an hard disk (not available on disk-less nodes)

That's All Folks!



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

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

REFERENCES AND USEFUL LINKS

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/
- Warewulf/PERCEUS
  http://www.warewulf-cluster.org/ http://www.perceus.org/

Installation Software:

• SystemImager	http://www.systemimager.org/
• FAI	http://www.informatik.uni-koeln.de/fai/
• Anaconda/Kickstart	http://fedoraproject.org/wiki/Anaconda/Kickstart

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/
- PDSH – Parallel Distributed SHell
  https://computing.llnl.gov/linux/pdsh.html
- DSH – Distributed SHell
  http://www.netfort.gr.jp/~dancer/software/dsh.html.en
- ClusterSSH
  http://clusterssh.sourceforge.net/
- C4 tools – Cluster Command & Control Console
  http://gforge.escience-lab.org/projects/c-4/

Monitoring Tools:

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

Network traffic analyzer:

• tcpdump	http://www.tcpdump.org
• wireshark	http://www.wireshark.org

UnionFS:

- Hopeless, a system for building disk-less clusters
  http://www.evolware.org/chri/hopeless.html
- UnionFS – A Stackable Unification File System
  http://www.unionfs.org
  http://www.fsl.cs.sunysb.edu/project-unionfs.html

RFC: (http://www.rfc.net)

- RFC 1350 – The TFTP Protocol (Revision 2)
  http://www.rfc.net/rfc1350.html
- RFC 2131 – Dynamic Host Configuration Protocol
  http://www.rfc.net/rfc2131.html
- RFC 2132 – DHCP Options and BOOTP Vendor Extensions
  http://www.rfc.net/rfc2132.html
- RFC 4578 – DHCP PXE Options
  http://www.rfc.net/rfc4578.html
- RFC 4390 – DHCP over Infiniband
  http://www.rfc.net/rfc4390.html

- PXE specification
  http://www.pix.net/software/pxeboot/archive/pxespec.pdf
- SYSLINUX http://syslinux.zytor.com/

Some acronyms...

HPC – High Performance Computing

OS – Operating System
LINUX – LINUX is not UNIX
GNU – GNU is not UNIX
RPM – RPM Package Manager

CLI – Command Line Interface
BASH – Bourne Again SHell
PERL – Practical Extraction and Report Language

PXE – Preboot Execution Environment
INITRD – INITial RamDisk

NFS – Network File System
SSH – Secure SHell
LDAP – Lightweight Directory Access Protocol
NIS – Network Information Service
DNS – Domain Name System

PAM – Pluggable Authentication Modules

LAN – Local Area Network
WAN – Wide Area Network

IP – Internet Protocol
TCP – Transmission Control Protocol
UDP – User Datagram Protocol
DHCP – Dynamic Host Configuration Protocol
TFTP – Trivial File Transfer Protocol
FTP – File Transfer Protocol
HTTP – Hyper Text Transfer Protocol
NTP – Network Time Protocol

NIC – Network Interface Card/Controller
MAC – Media Access Control
OUI – Organizationally Unique Identifier

API – Application Program Interface
UNDI – Universal Network Driver Interface
PROM – Programmable Read-Only Memory
BIOS – Basic Input/Output System

SNMP – Simple Network Management Protocol
MIB – Management Information Base
OID – Object IDentifier

IPMI – Intelligent Platform Management Interface
LOM – Lights-Out Management
RSA – IBM Remote Supervisor Adapter
BMC – Baseboard Management Controller