MHPC
Master in High Performance Computing

Moreno Baricevic

CNR-IOM DEMOCRITOS
Trieste, ITALY

Installation Procedures for Clusters

PART 1 – Cluster Services and Installation Procedures

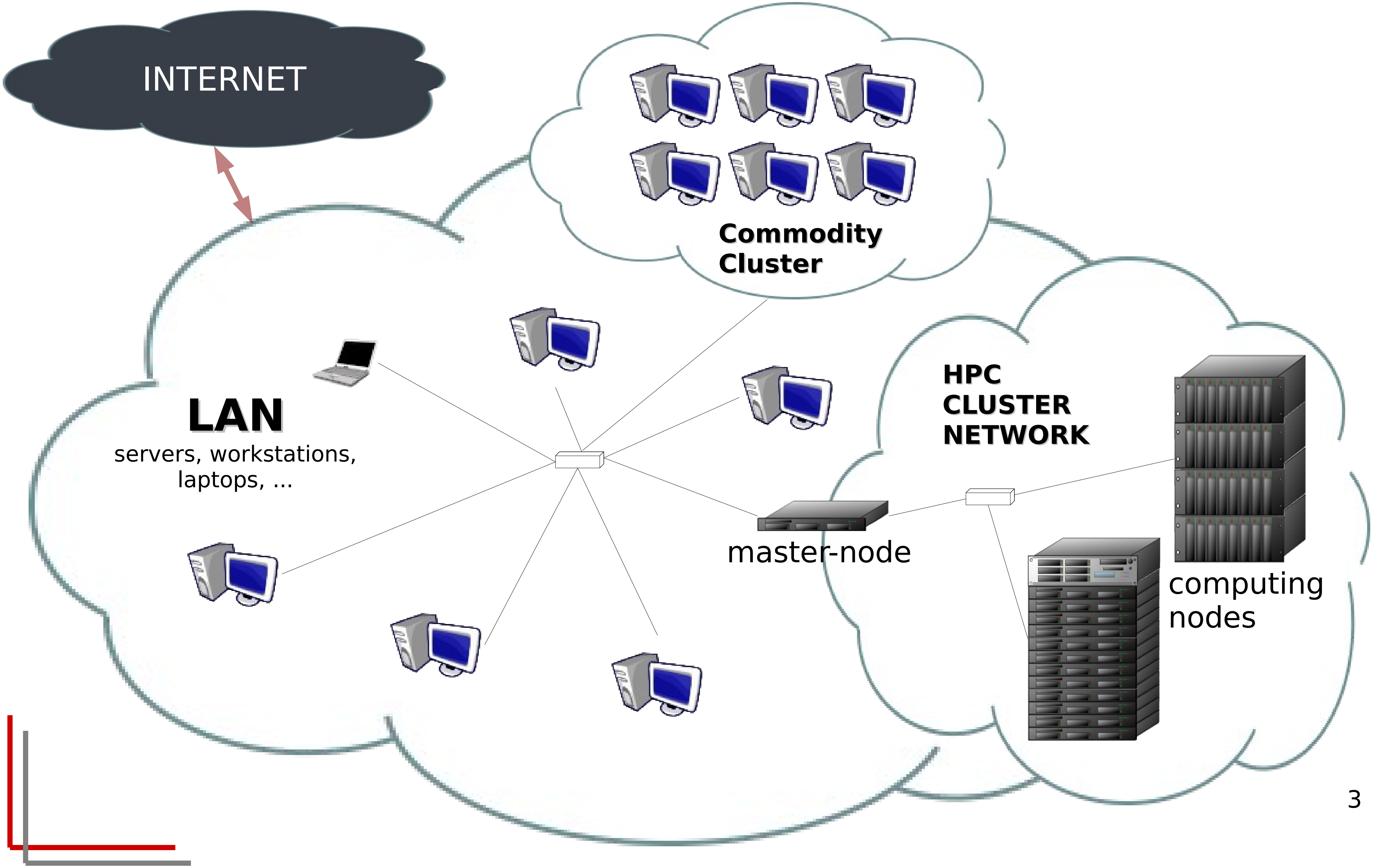
ma per seguir virtute e canoscenza
SISSA
Scuola Internazionale Superiore
di Studi Avanzati

ICTP
The Abdus Salam
International Centre
for Theoretical Physics

Agenda

- Introduction and 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

Some definitions (1/2)

Parallel computing

The simultaneous execution of a task split up on multiple processors in order to obtain results faster.

Distributed computing

Same thing but with many computers (concept of network).

Cluster

Group of linked computers working together (can be seen as a single computer).

Some definitions (2/2)

Node (aka Computing Node)

Computer used for its computational power.

Frontend

It's through this node that the users will submit/launch/manage jobs.

Access Node

A cluster is usually isolated from outside for security purpose, this node is the access gateway.

Master Node

Management server, that might as well act as frontend and access node.

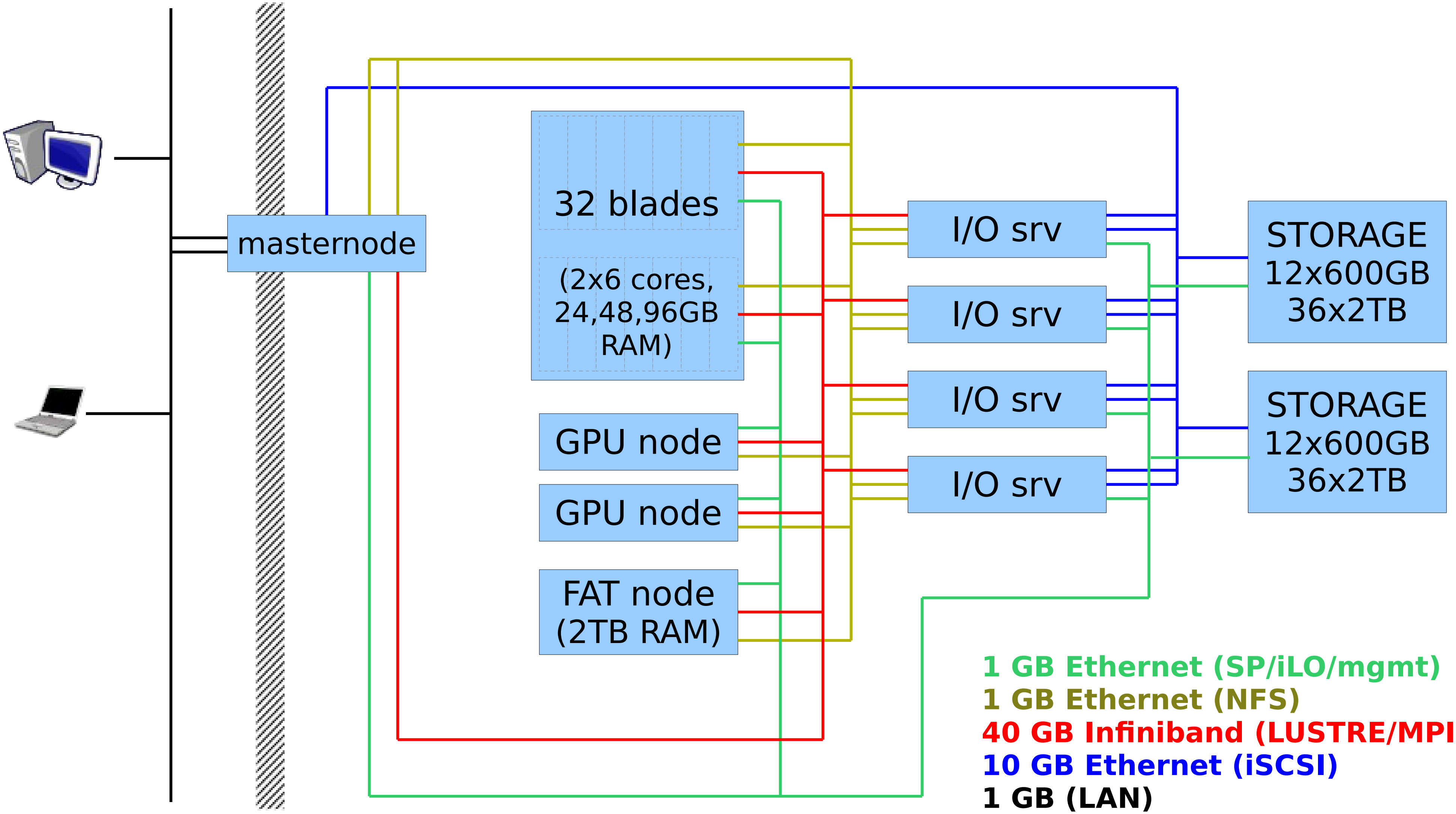
Batch Scheduler

Software responsible for scheduling the users' jobs on the cluster.

Resources Manager

Software that enable the jobs to connect the nodes and run.

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

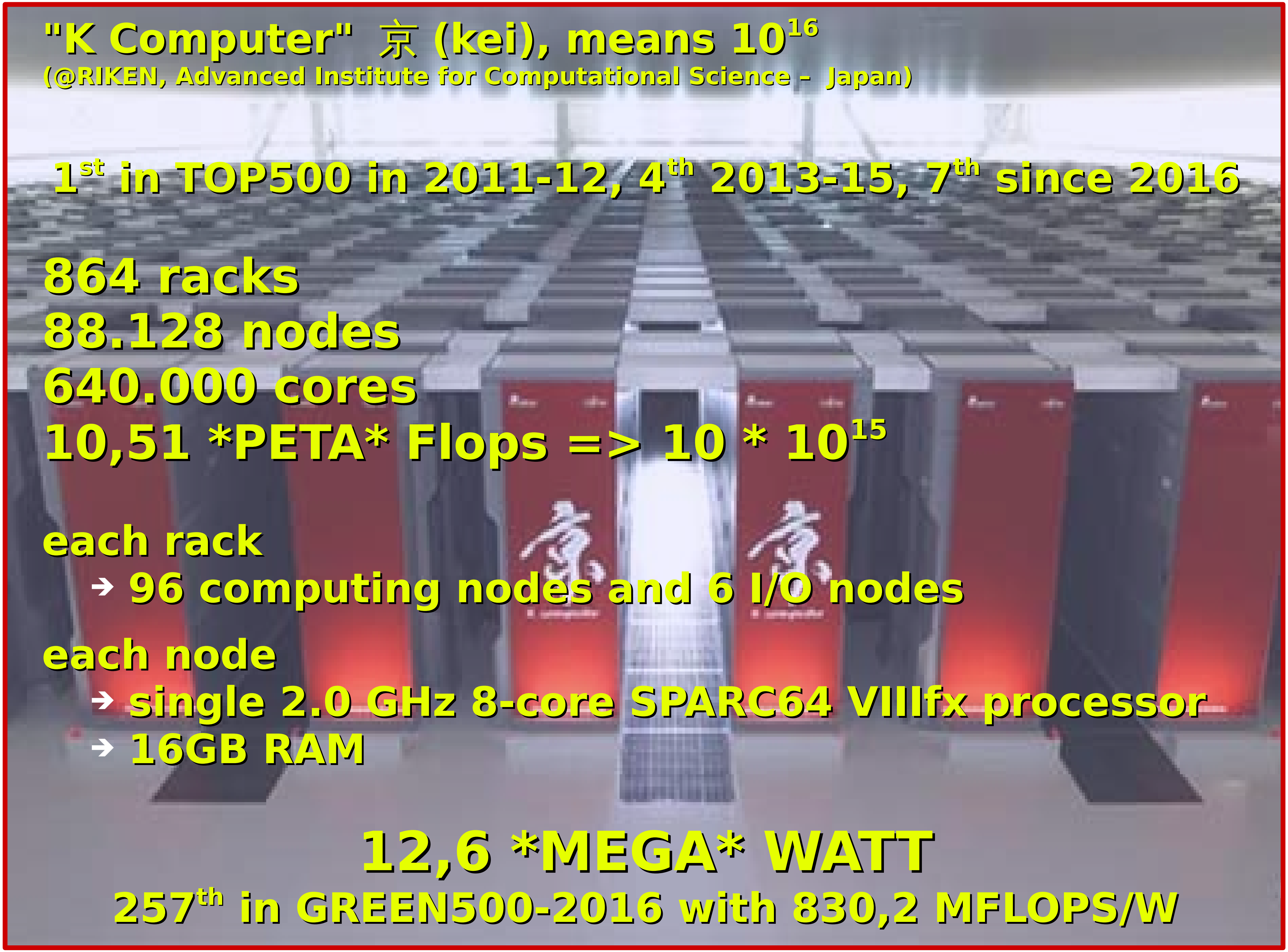
IBM Blade Center
14 bays in 7U ***2x***

:-(

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" 京 (kei), means 10^{16}

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

1st in TOP500 in 2011-12, 4th 2013-15, 7th since 2016

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

each rack

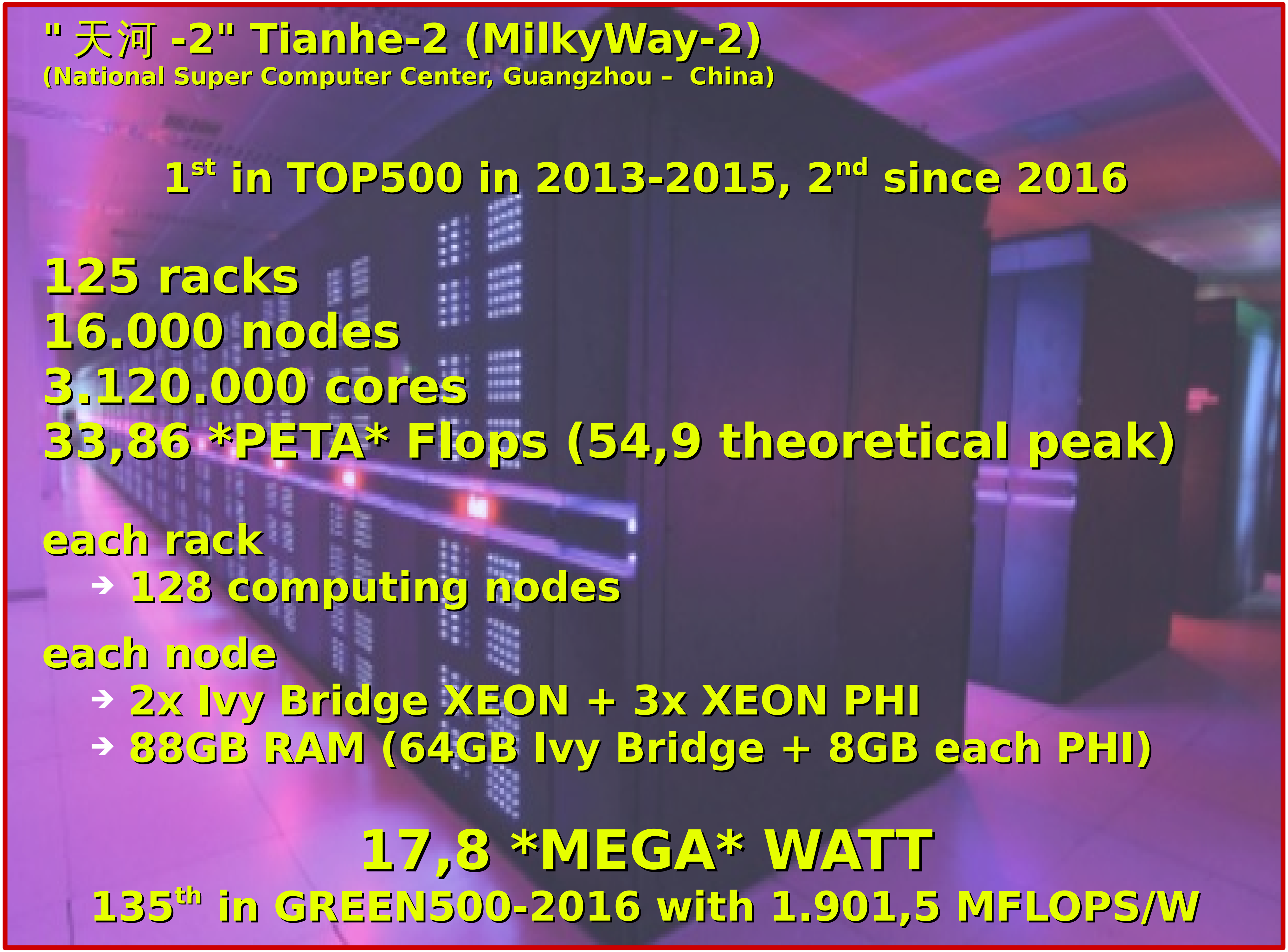
→ 96 computing nodes and 6 I/O nodes

each node

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

12,6 *MEGA* WATT

257th in GREEN500-2016 with 830,2 MFLOPS/W



" 天河 -2" Tianhe-2 (MilkyWay-2)

(National Super Computer Center, Guangzhou – China)

1st in TOP500 in 2013-2015, 2nd since 2016

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

each rack

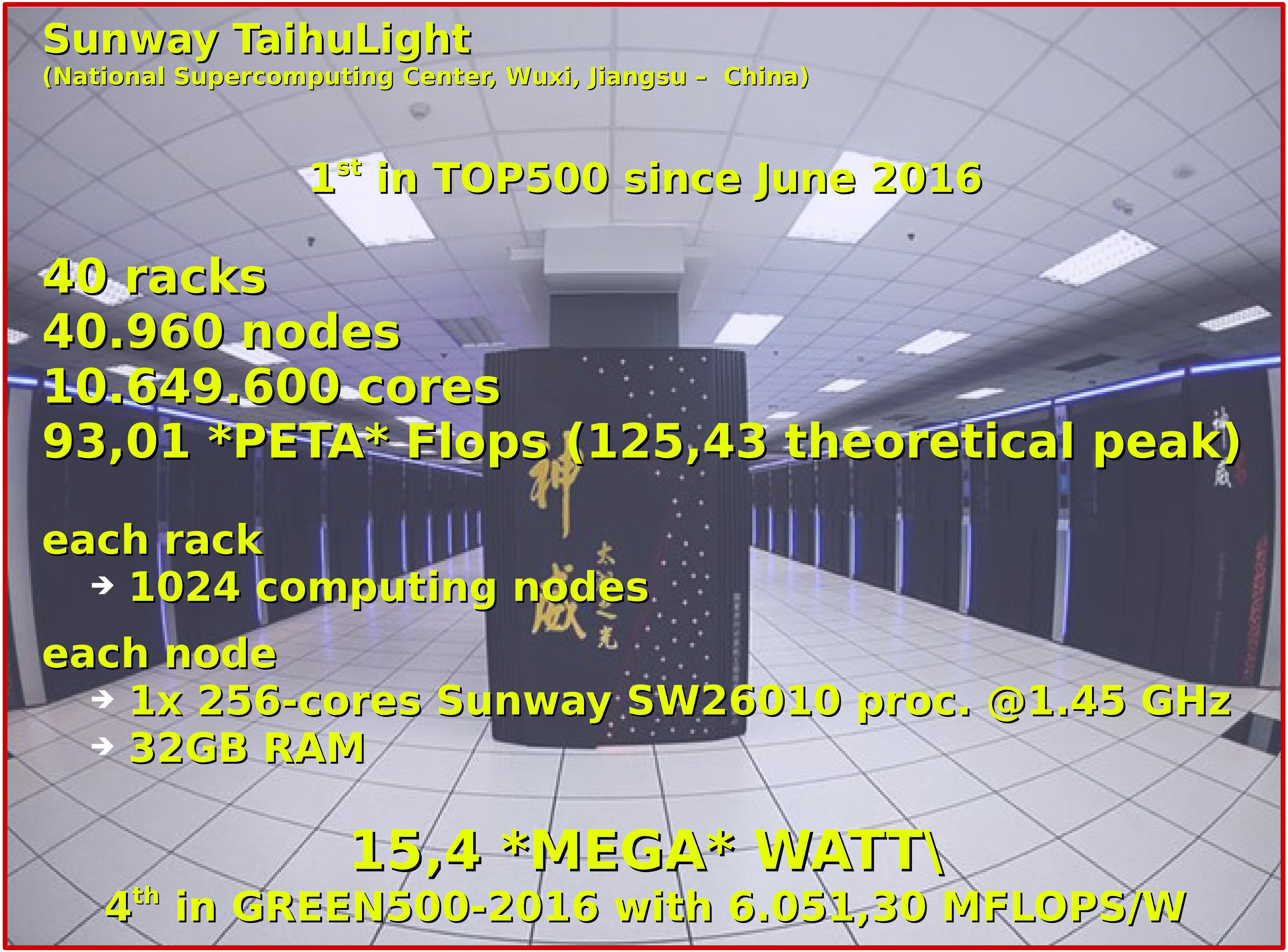
- ➔ 128 computing nodes

each node

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

17,8 *MEGA* WATT

135th in GREEN500-2016 with 1.901,5 MFLOPS/W



Sunway TaihuLight

(National Supercomputing Center, Wuxi, Jiangsu – China)

1st in TOP500 since June 2016

40 racks
40.960 nodes
10.649.600 cores
93,01 *PETA* Flops (125,43 theoretical peak)

each rack

- ➔ 1024 computing nodes

each node

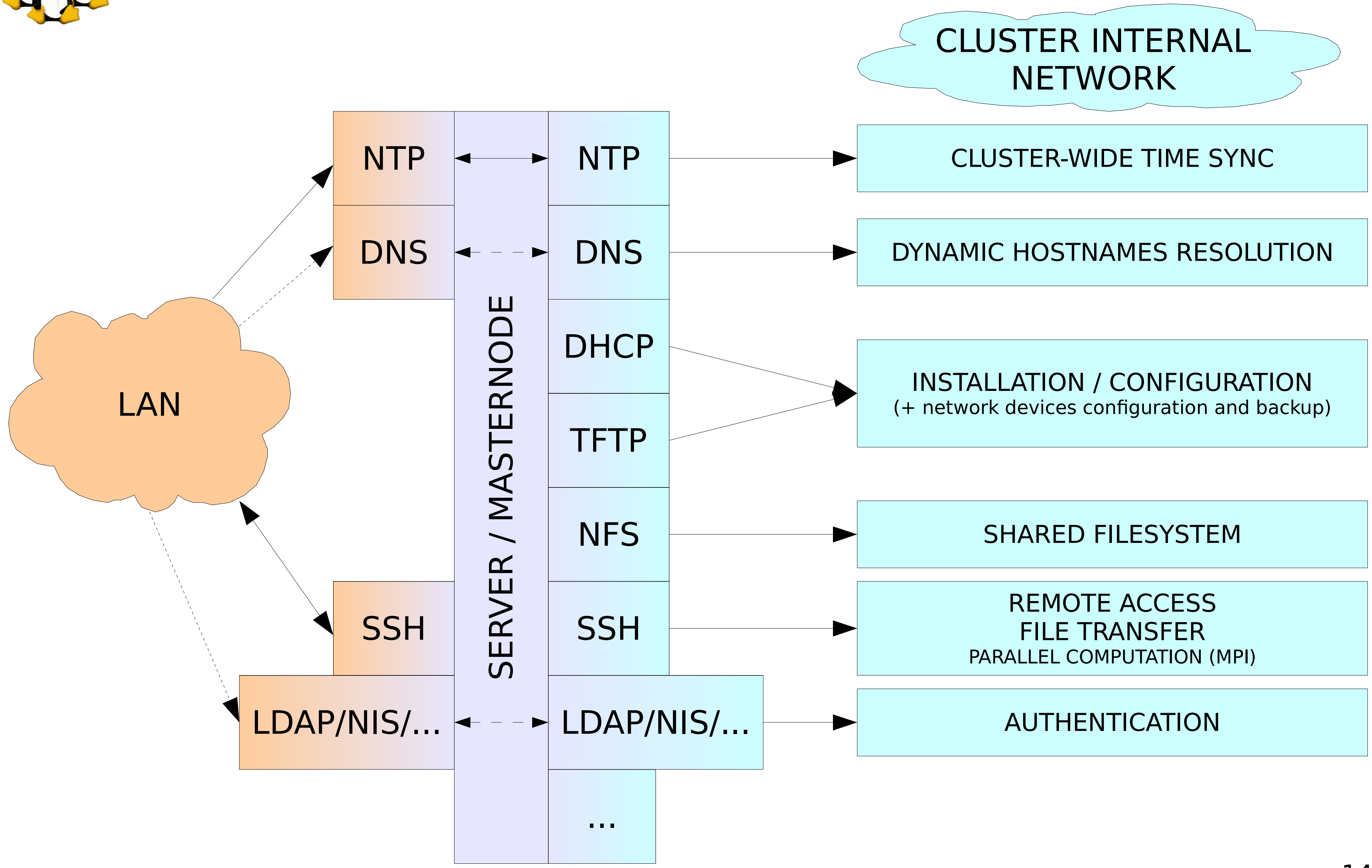
- ➔ 1x 256-cores Sunway SW26010 proc. @1.45 GHz
- ➔ 32GB RAM

15,4 *MEGA* WATT

4th in GREEN500-2016 with 6.051,30 MFLOPS/W

	K	**Tianhe-2**	**Sunway TaihuLight**
1st in top500 (year)	2011	2013	2016
top500-2016	7th	2nd	1st
green500-2016	257th	135th	4th
#racks	**864**	**125**	**40**
#nodes	**88.128**	**16.000**	**40.960**
#cores	**640.000**	**3.120.000**	**10.649.600**
#nodes/rack	**96+6**	**128**	**1.024**
Processor	1x8c@2GHz	2x12c@2.2GHz + 3xPHI	1x256c@1.5Ghz
Architecture	SPARC	Intel (Xeon IvyBridge + PHI)	Sunway
RAM (GB)	16	88	32
Rmax (PFLOPS)	10,51	33,86	93,01
Rpeak (PFLOPS)	11,28	54,9	125,43
Power (MW)	12,6	17,8	15,4
En.Eff.(MFLOPS/W)	**830,2**	**1.901,5**	**6.051,30**

CLUSTER SERVICES



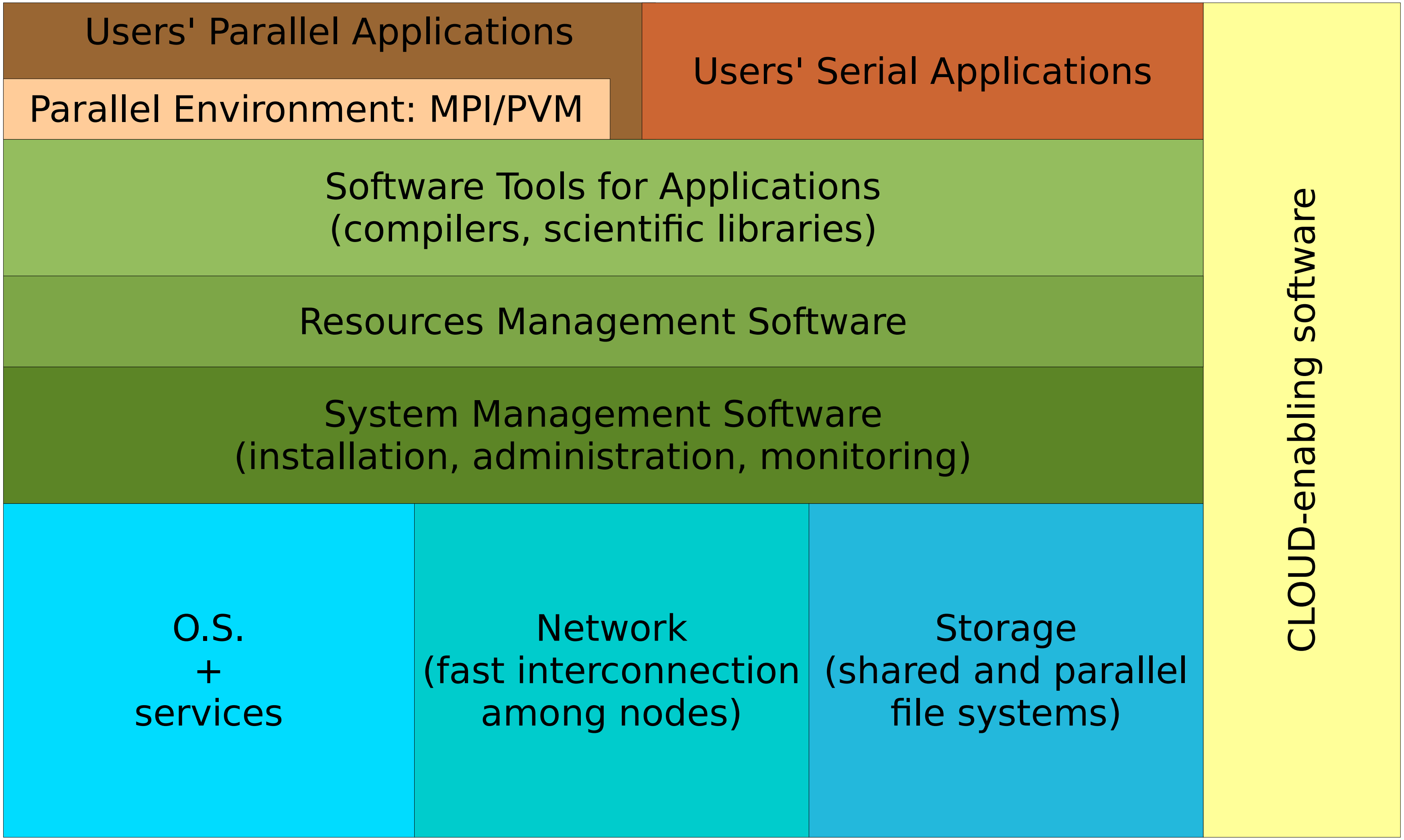
CLUSTER INTERNAL NETWORK

LAN

SERVER / MASTERNODE

LAN side		Cluster side	
NTP	↔	NTP	CLUSTER-WIDE TIME SYNC
DNS	↔	DNS	DYNAMIC HOSTNAMES RESOLUTION
		DHCP	INSTALLATION / CONFIGURATION (+ network devices configuration and backup)
		TFTP	INSTALLATION / CONFIGURATION (+ network devices configuration and backup)
		NFS	SHARED FILESYSTEM
SSH		SSH	REMOTE ACCESS FILE TRANSFER PARALLEL COMPUTATION (MPI)
LDAP/NIS/...	↔	LDAP/NIS/...	AUTHENTICATION
		...	

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)

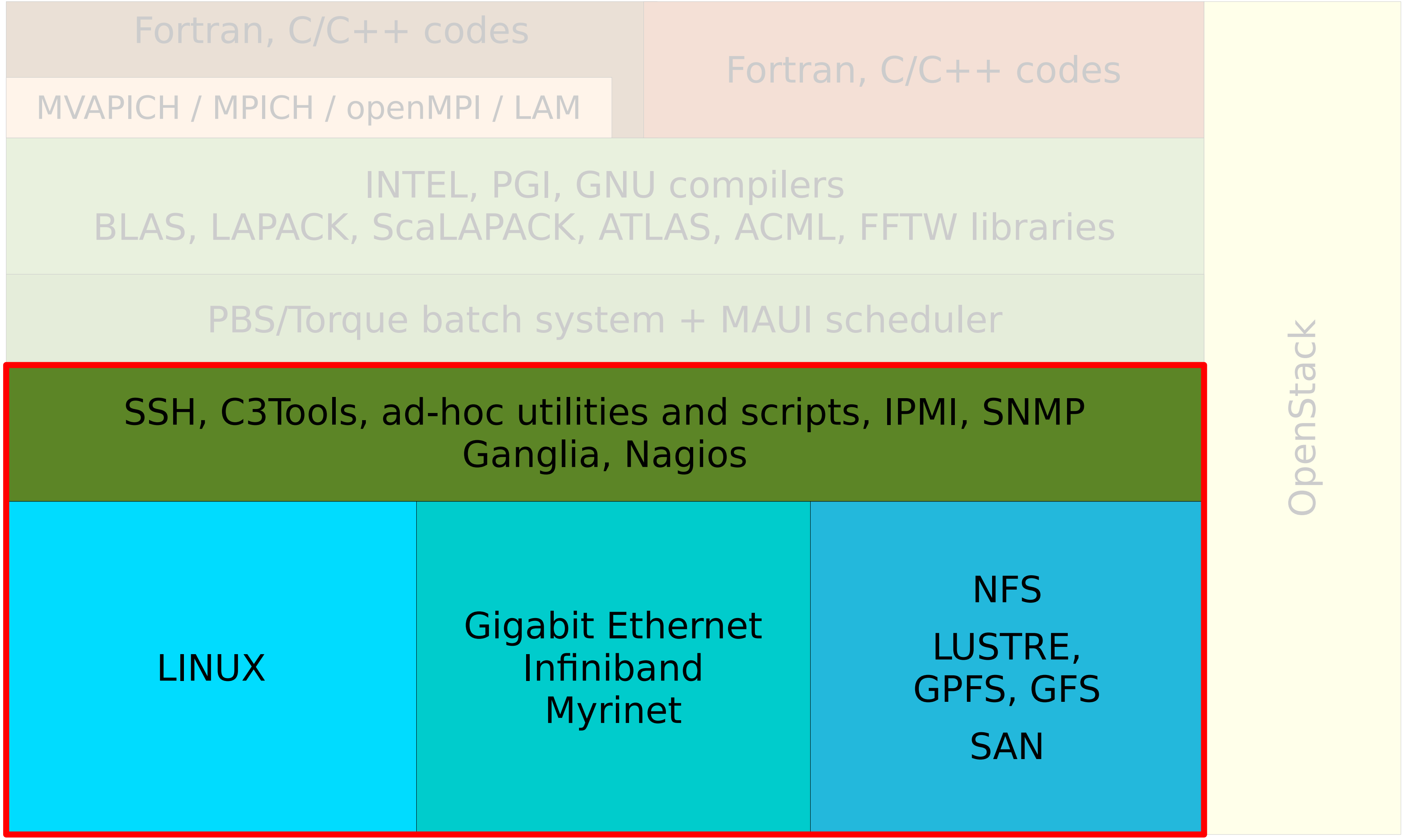
O.S.
+
services

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

<u>Installation can be performed:</u>

- 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

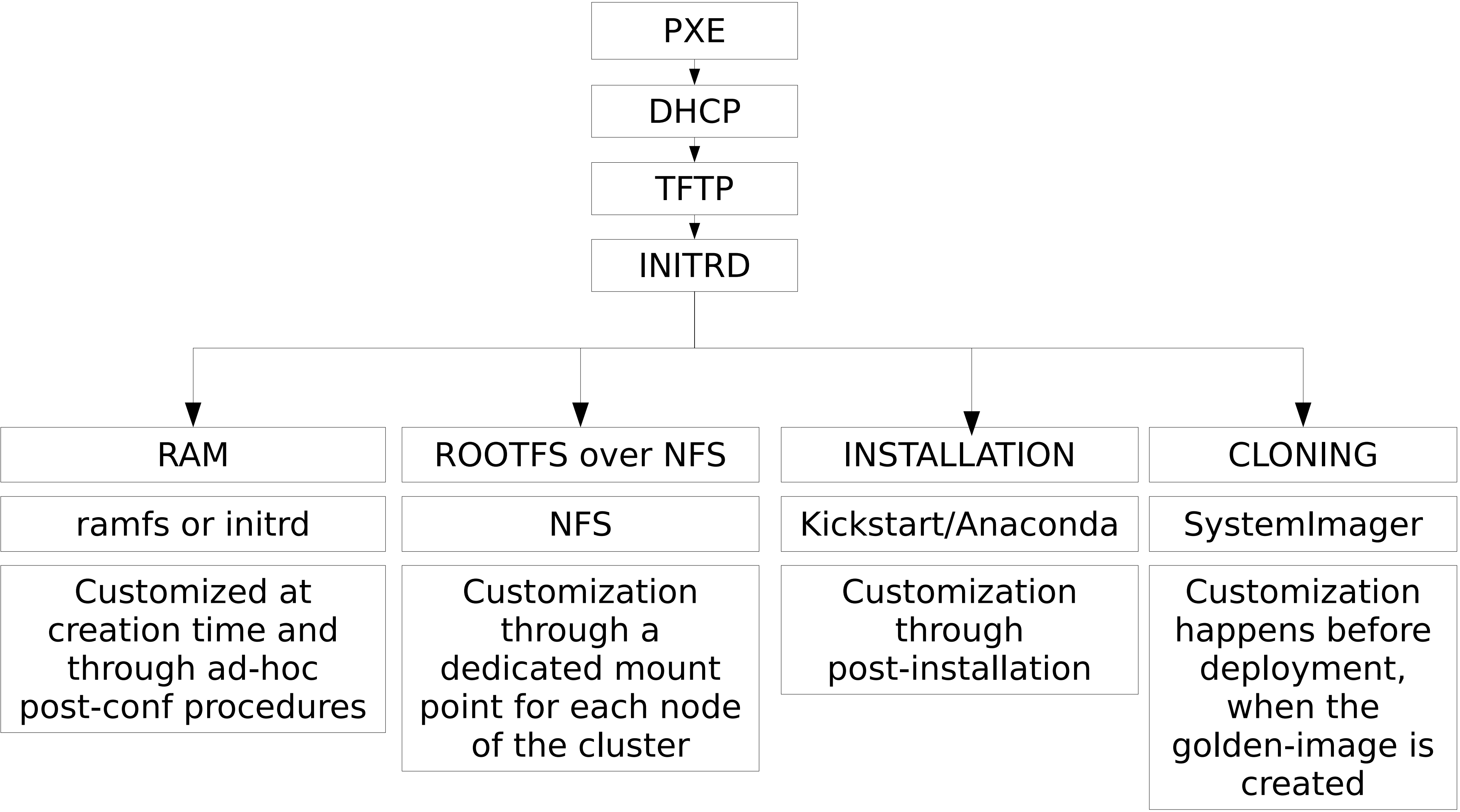
Are generally made of an ensemble of already available software packages thought for specific tasks, but configured to operate together, plus some add-ons.

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

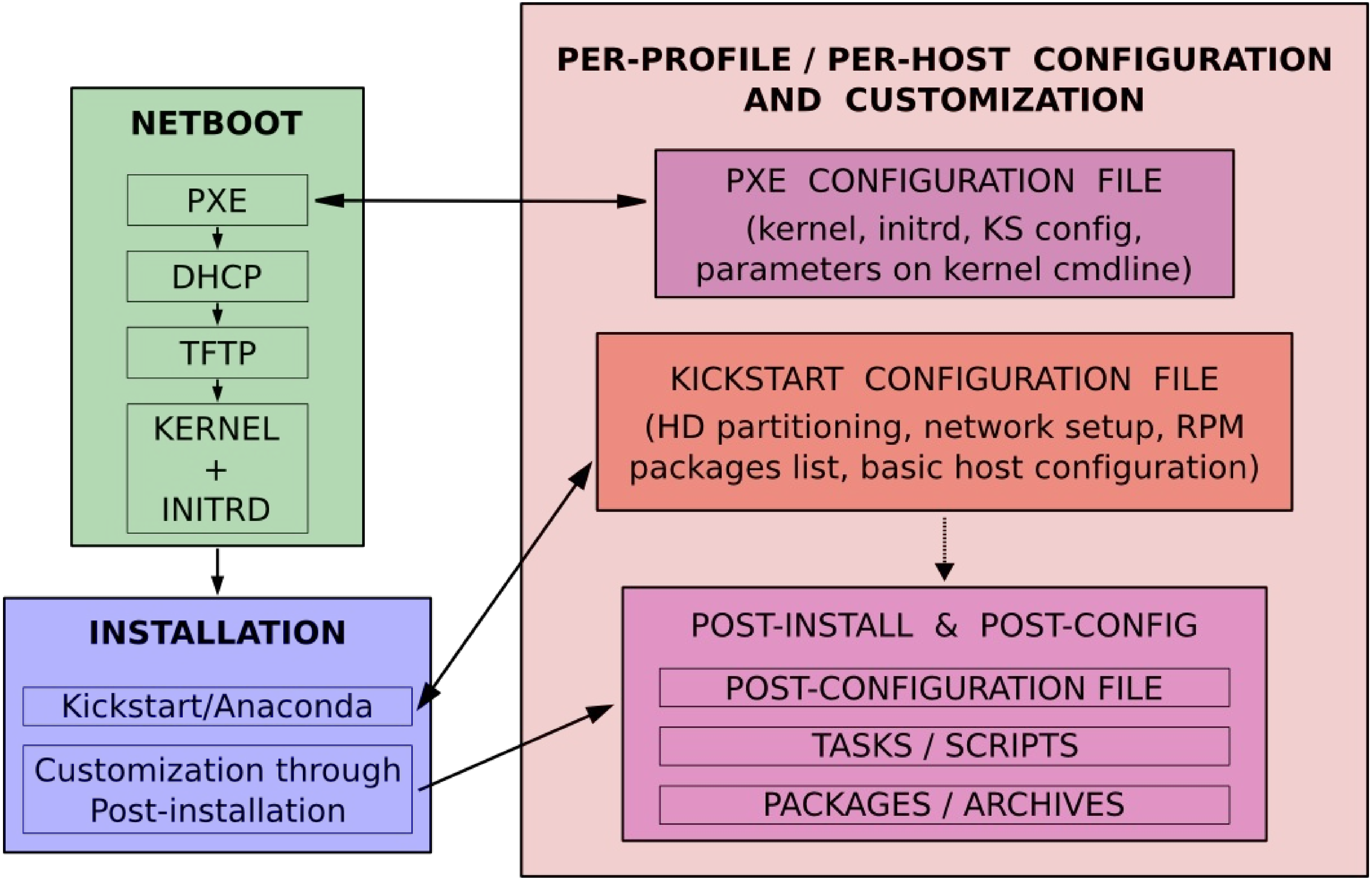
- **PXE**: network booting
- **DHCP**: IP binding + NBP (pxelinux.0)
- **TFTP**: pxe configuration file (pxelinux.cfg/<HEXIP>), alternative boot-up images (memtest, UBCD, ...)
- **NFS**: kickstart + RPM repository (with little modification **HTTP(S)** or **FTP** can be used too)

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, ...

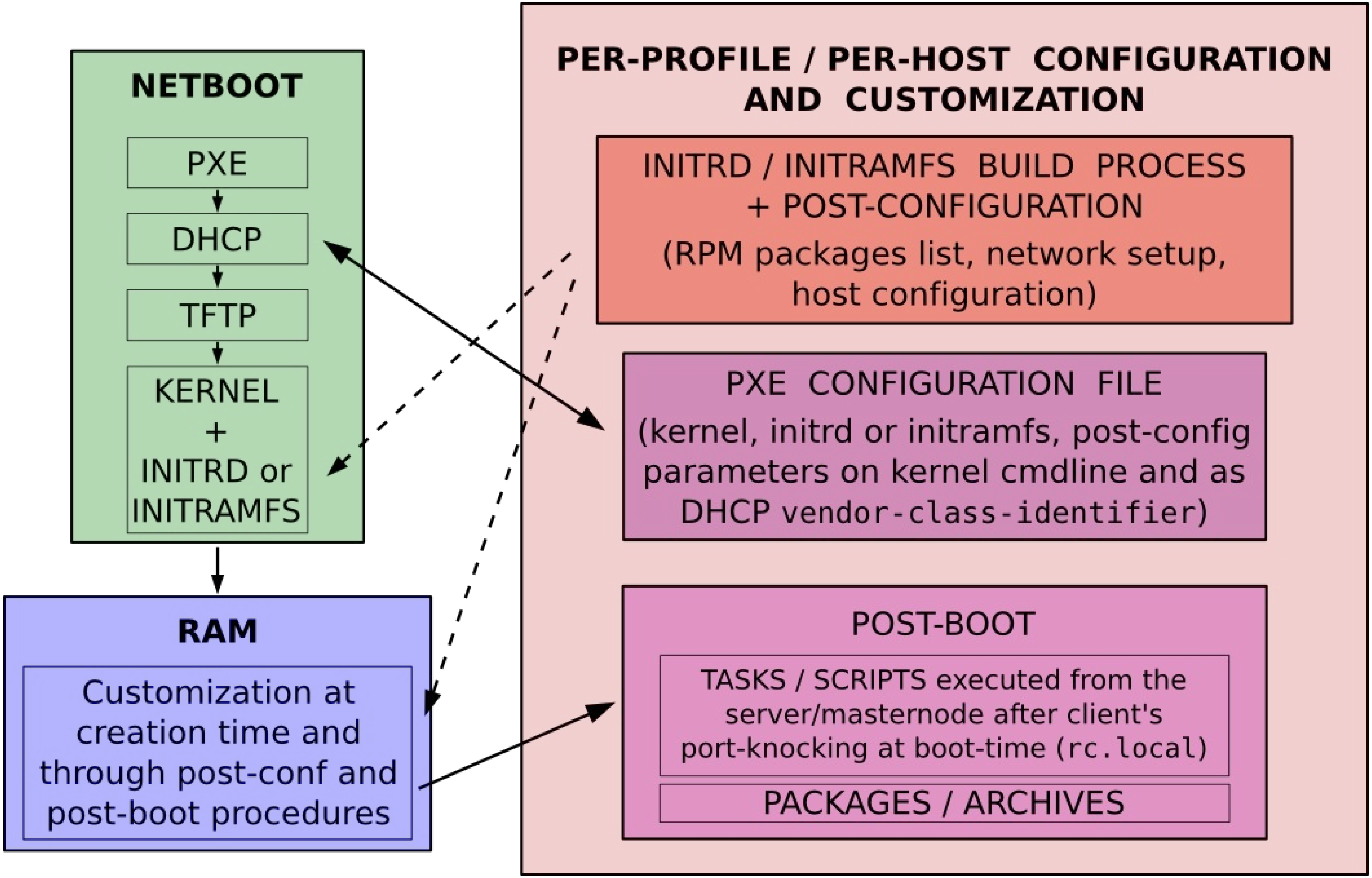
Customization layers

Installation process


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

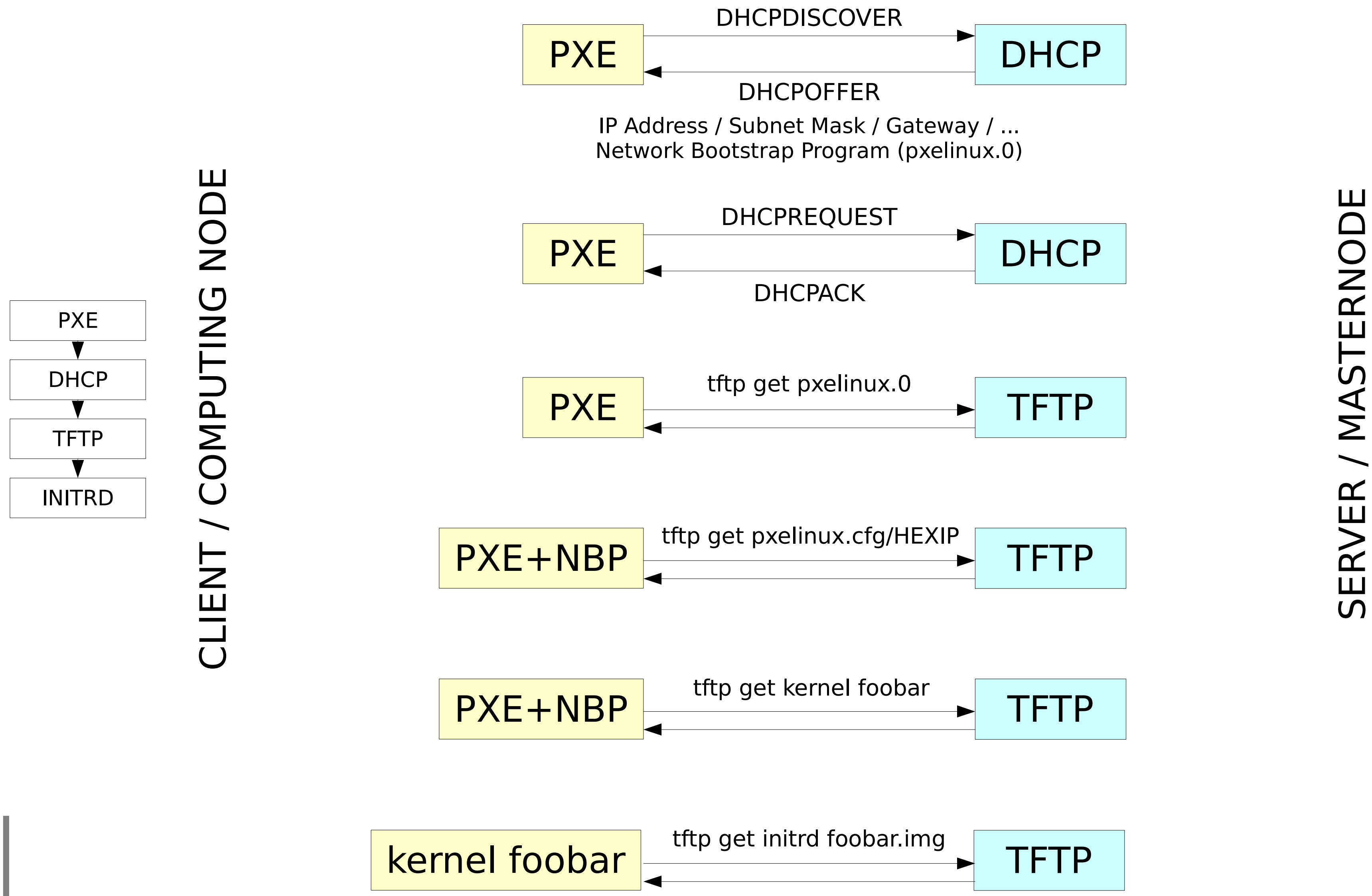
POST-BOOT

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

PXE —DHCPDISCOVER→ DHCP

DHCPOFFER

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

PXE —DHCPREQUEST→ DHCP

DHCPACK

PXE —tftp get pxelinux.0→ TFTP

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

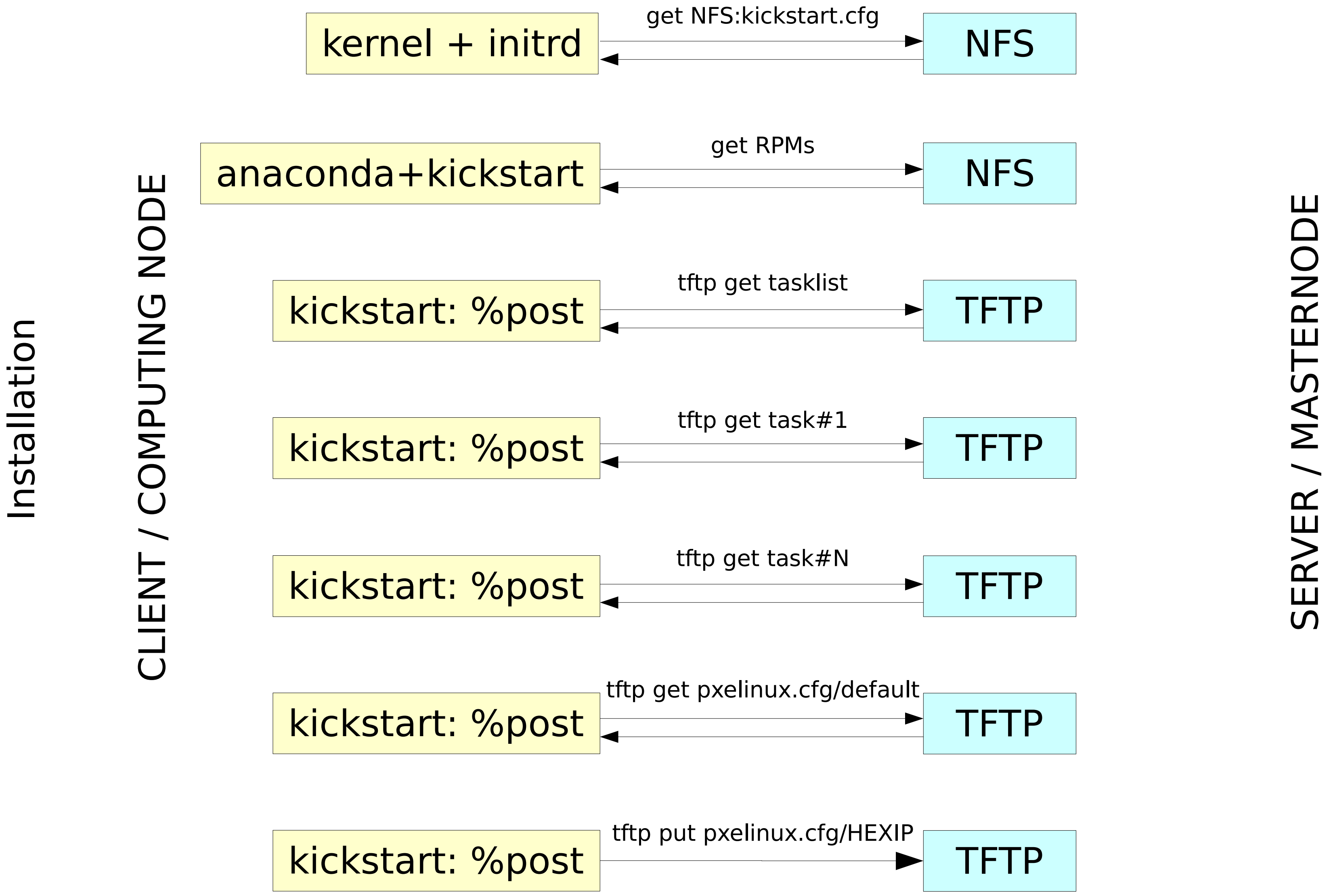
PXE+NBP —tftp get kernel foobar→ TFTP

kernel foobar —tftp get initrd foobar.img→ TFTP

SERVER / MASTERNODE

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

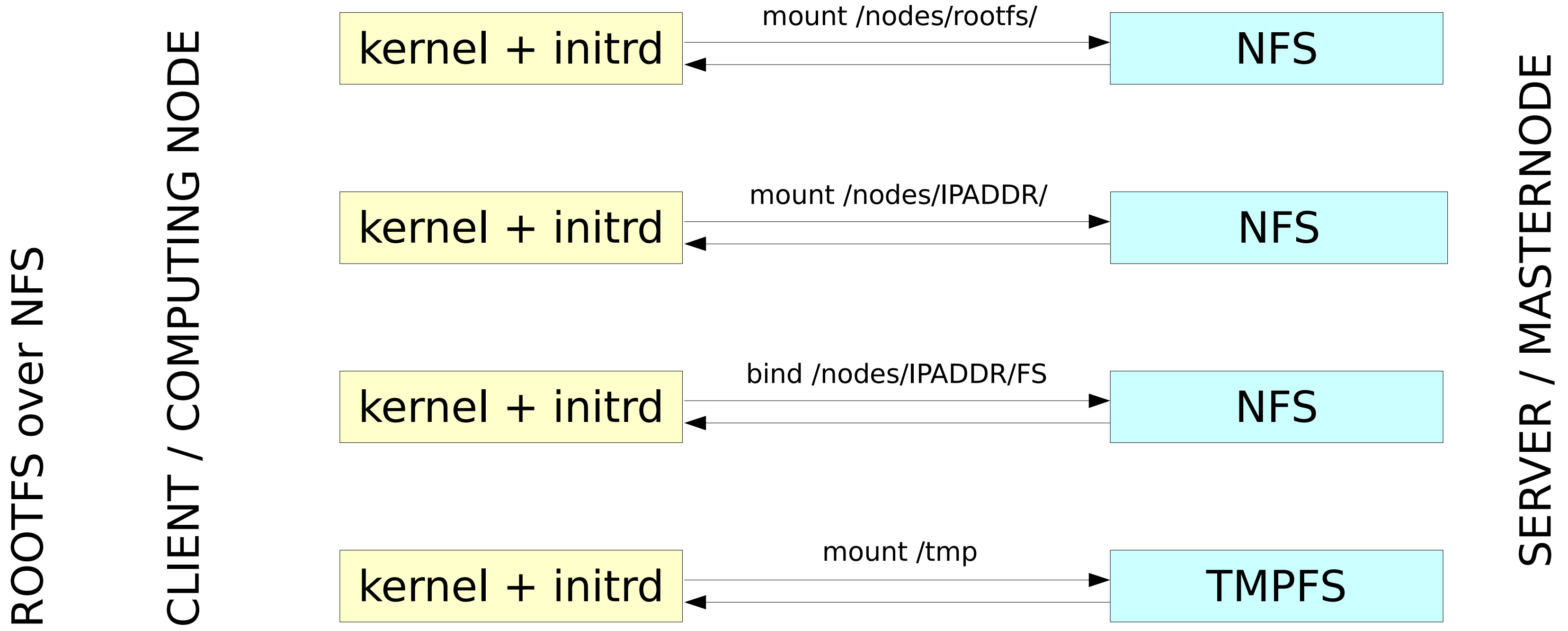
kickstart: %post → tftp get task#N → TFTP

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

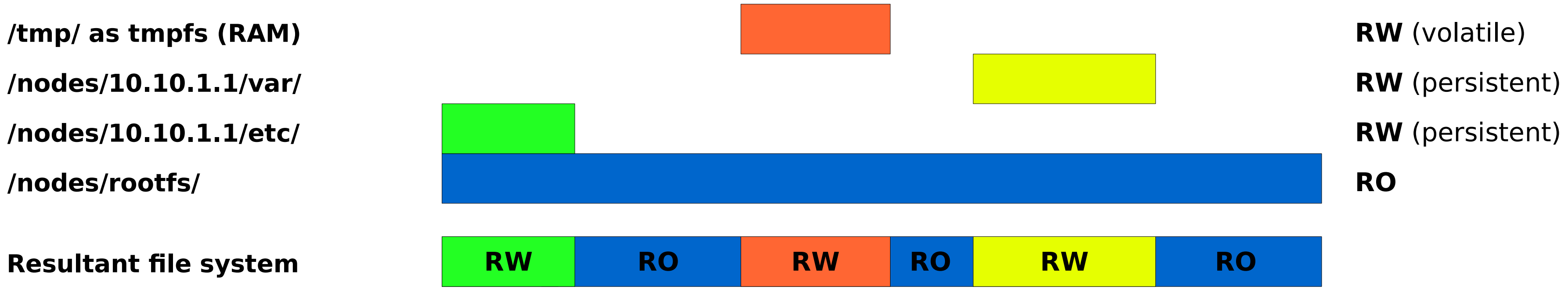
SERVER / MASTERNODE

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

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

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

kernel + initrd — mount /tmp — TMPFS



/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