

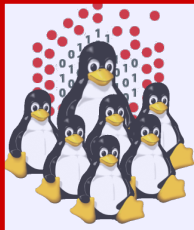
**Moreno Baricevic**

CNR-IOM DEMOCRITOS  
Trieste, ITALY

# Installation Procedures for Clusters

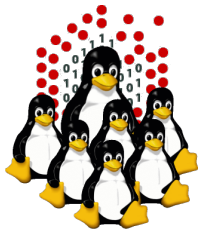
PART 1 – Cluster Services and  
Installation Procedures



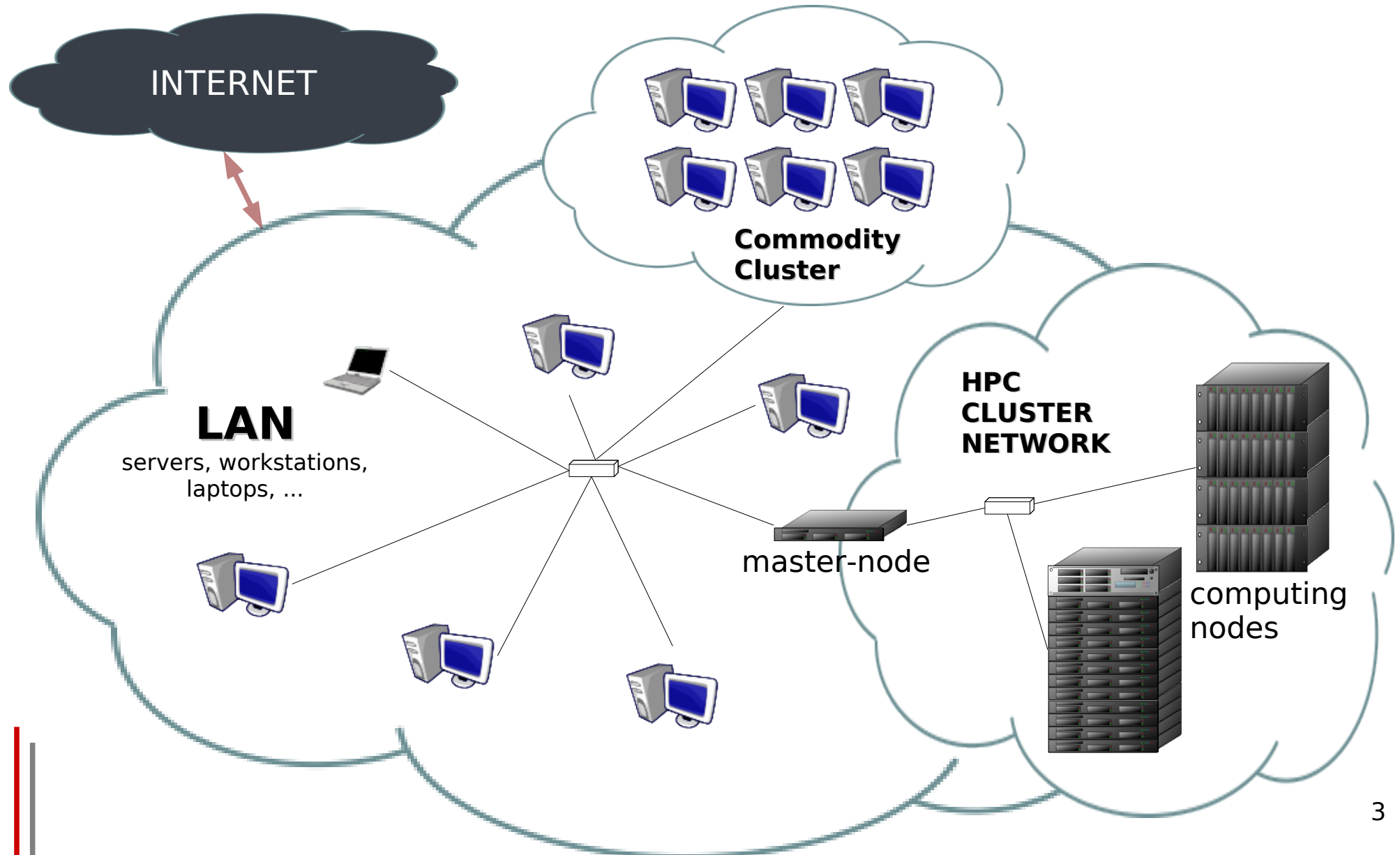


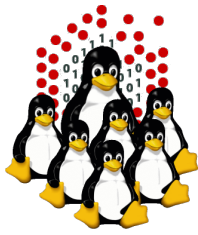
# Agenda

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



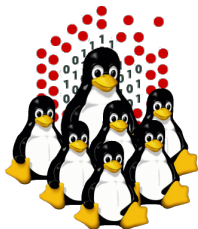
# What's a cluster?



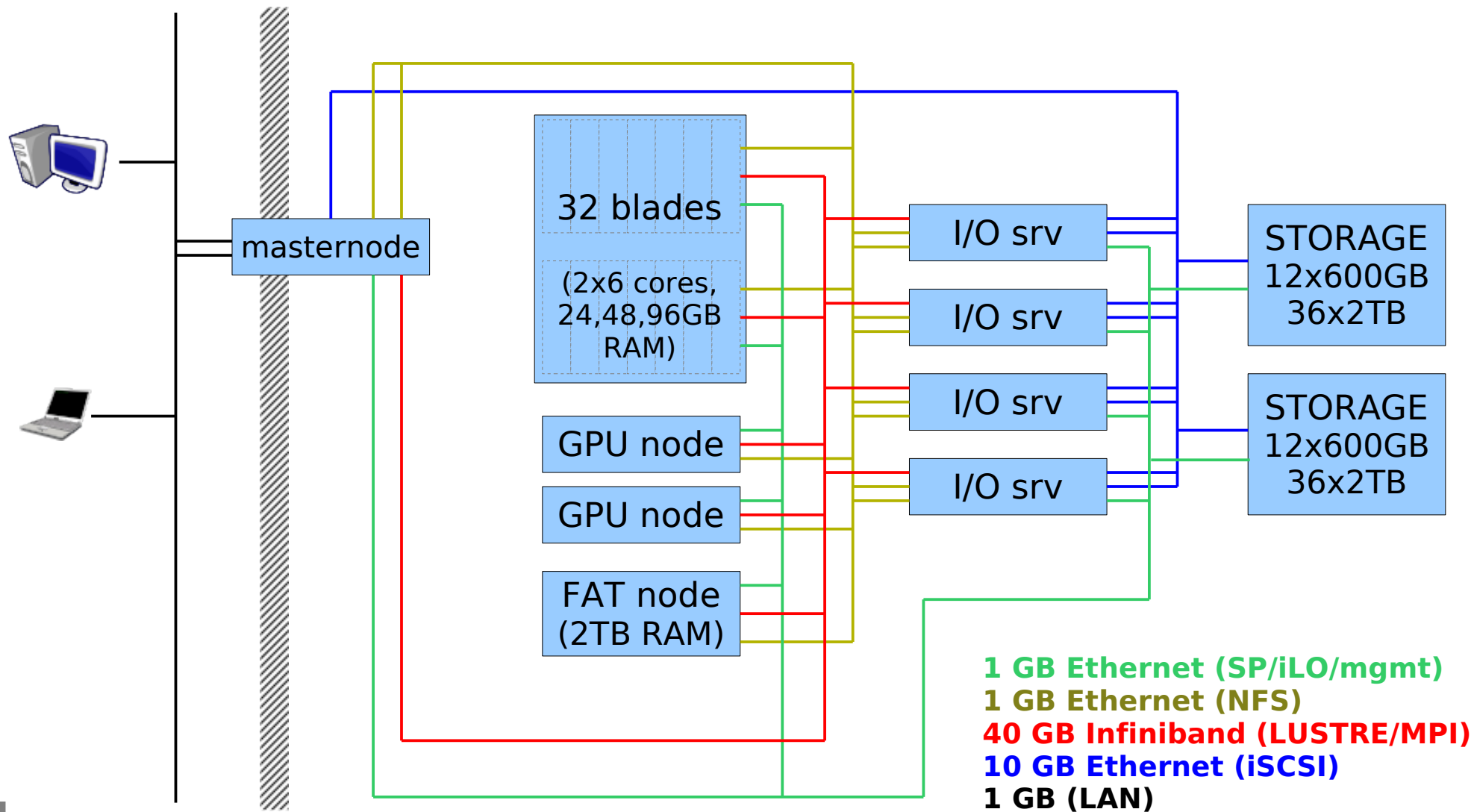


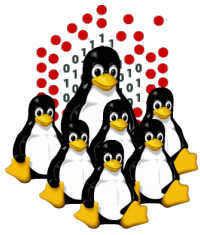
# What's a cluster?

- 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)





# What's a cluster from the HW side?

PC / WORKSTATION

LAPTOP



1U Server  
(rack mountable)

RACKs + rack mountable SERVERS



BLADE Servers



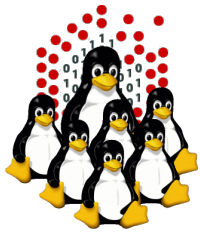
IBM Blade Center  
14 bays in 7U **2x**



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

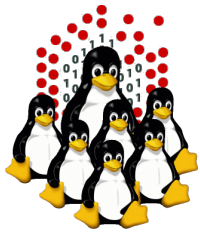


HP c7000  
8-16 bays in 10U 6

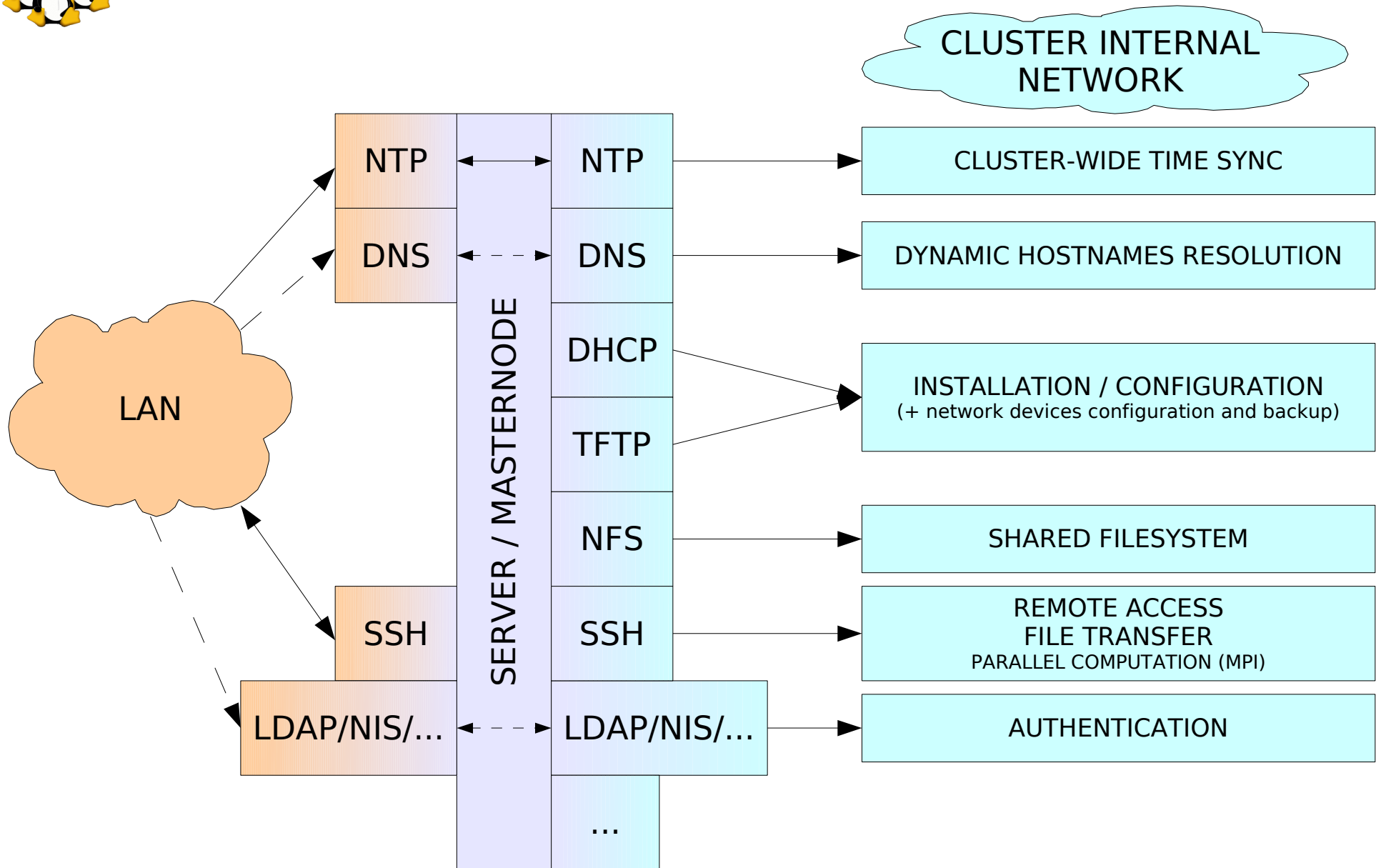


# What's a cluster from the HW side?

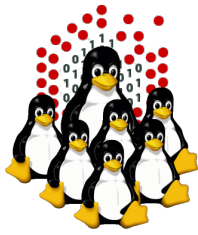




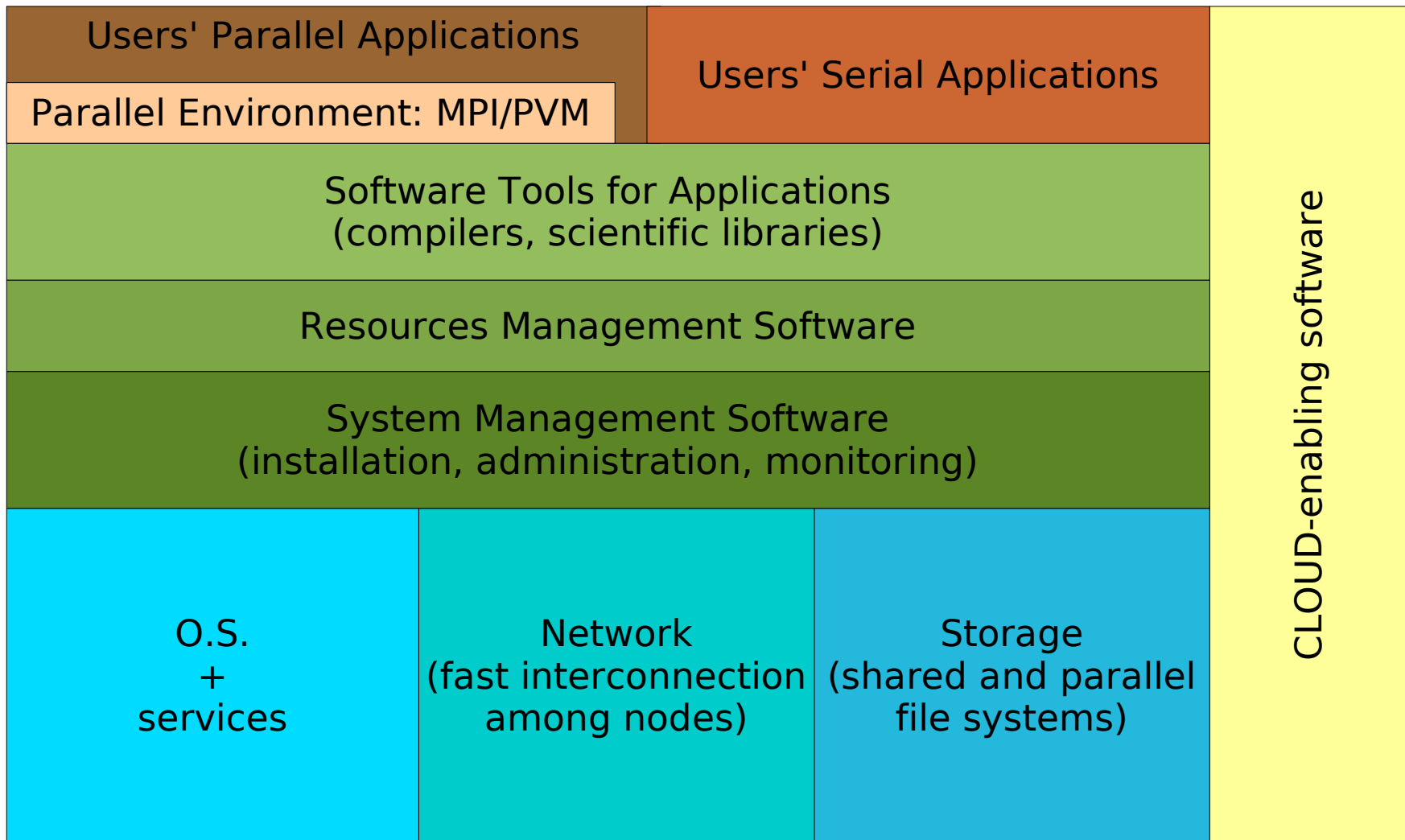
# CLUSTER SERVICES

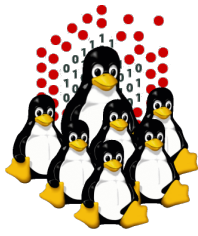




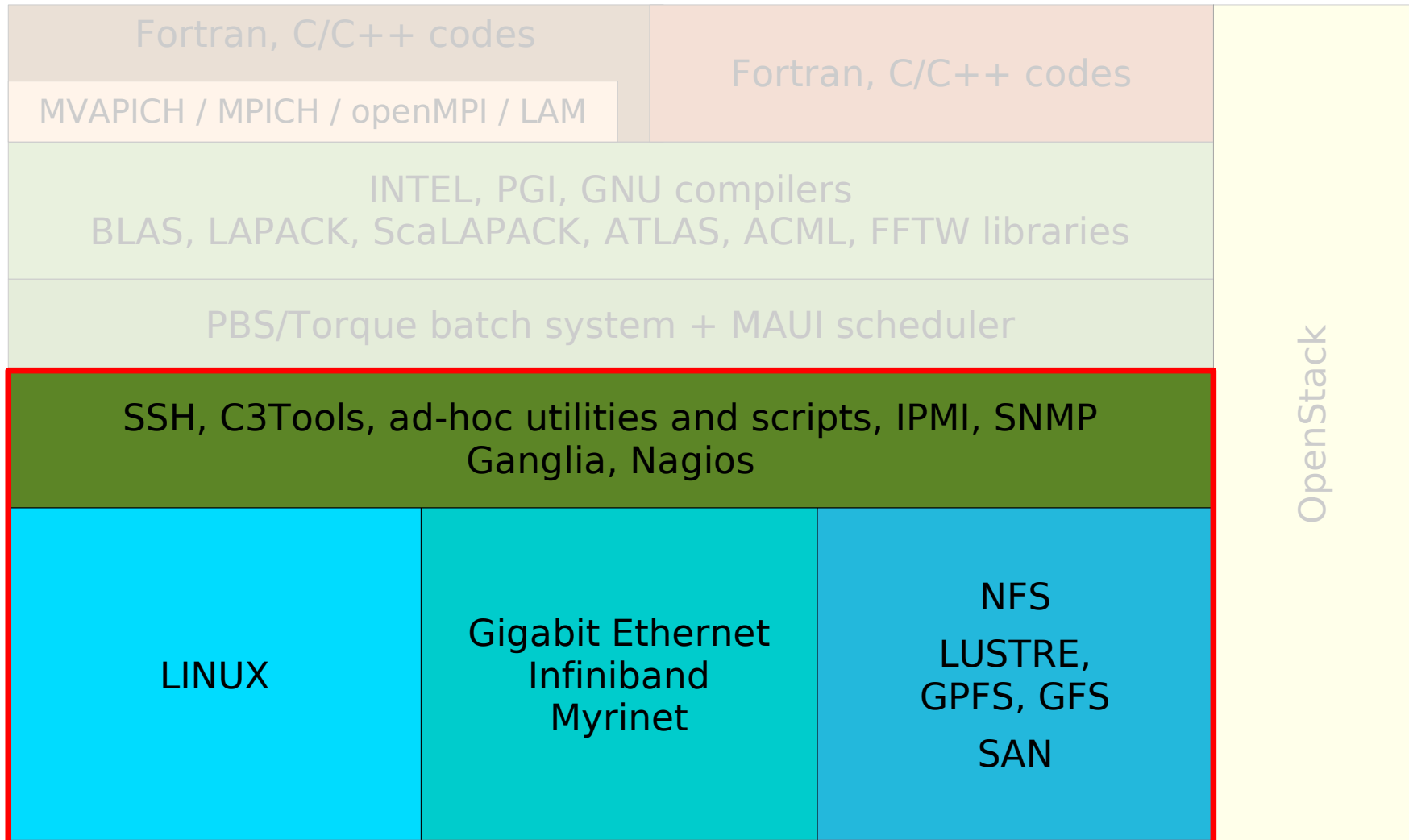


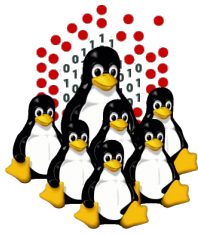
# HPC SOFTWARE INFRASTRUCTURE Overview





# HPC SOFTWARE INFRASTRUCTURE Overview (our experience)

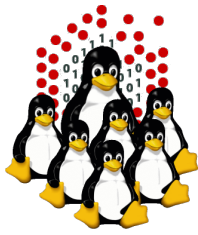




# CLUSTER MANAGEMENT Installation

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:

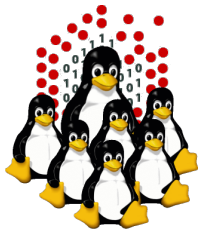
- local devices (CD-ROM, DVD-ROM, Floppy, ...)
- network based (PXE+DHCP+TFTP+NFS/HTTP/FTP)

## 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

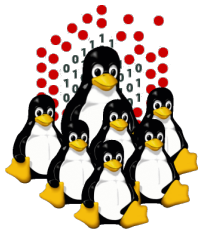
## Cluster Nodes Installation

### 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)**



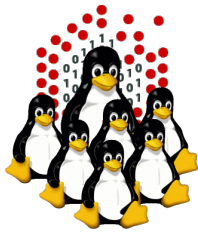
# CLUSTER MANAGEMENT

## Existent toolkits

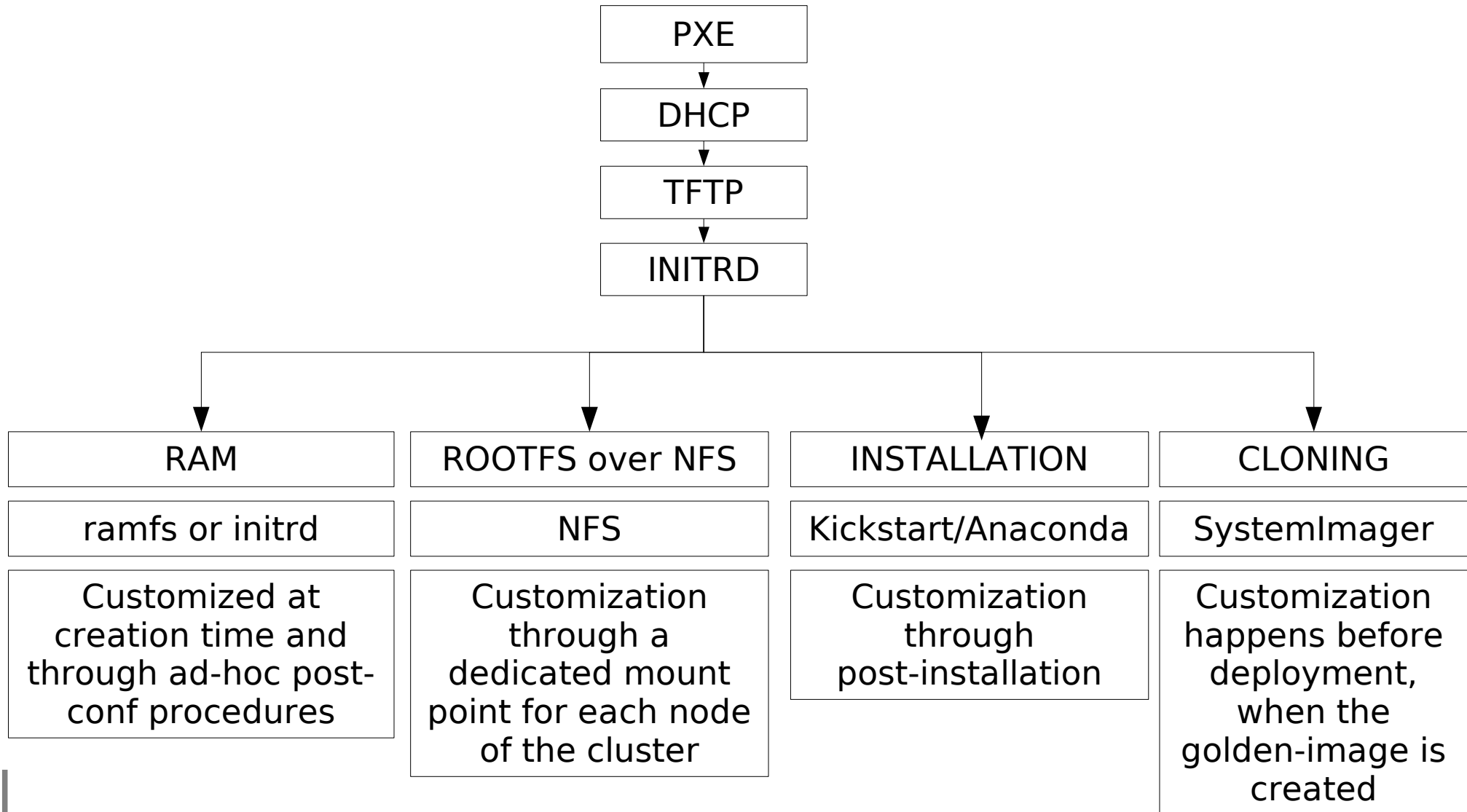
Are generally made of an ensemble of already available software packages thought for specific tasks, but configured to operate together, plus some add-ons.

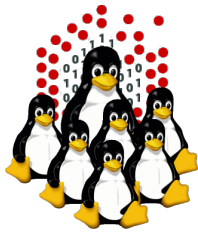
Sometimes limited by rigid and not customizable configurations, often bound to some specific LINUX distribution and version. May depend on vendors' hardware.

- 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





# Network-based Distributed Installation

## Basic services

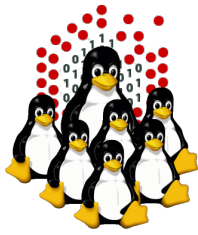
### Deployment

- **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

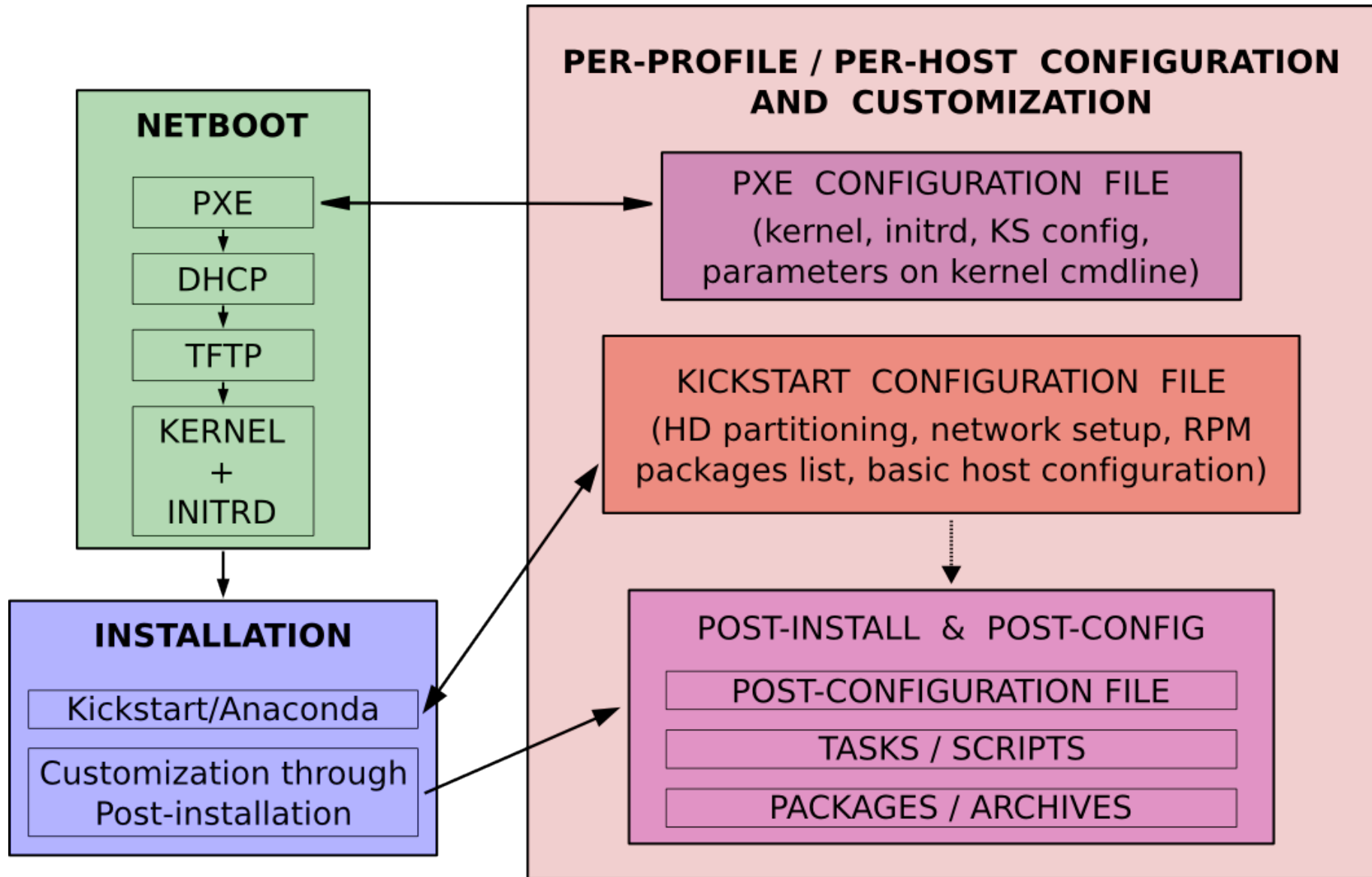
- passive updates: post-boot updates using port-knocking, ssh, distributed shells, wget, ...
- active configuration/package updates: ssh, distributed shells
- advanced IT automation tools: Ansible, CFEngine, ...

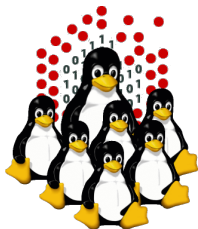




# Customization layers

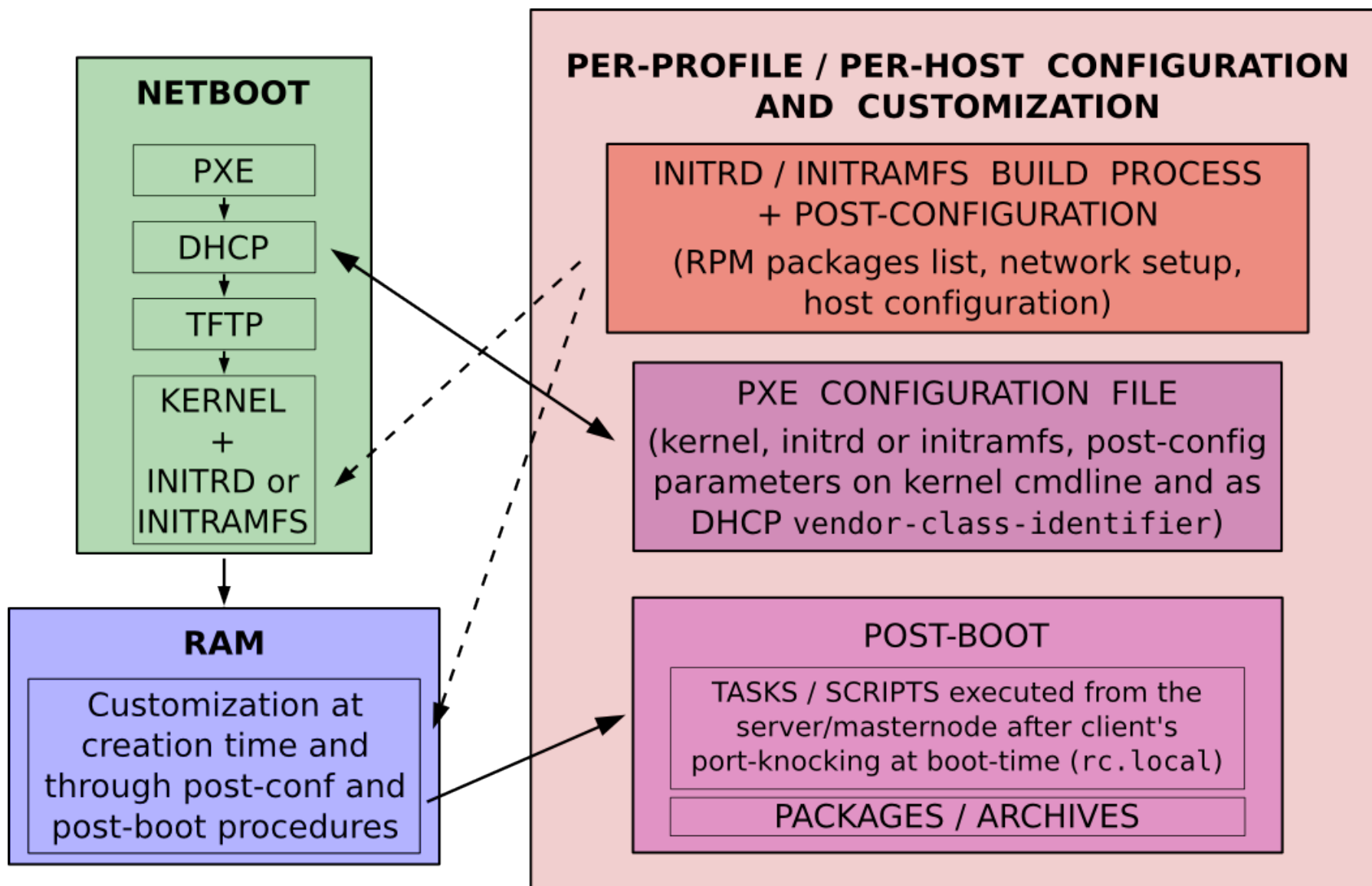
## Installation process

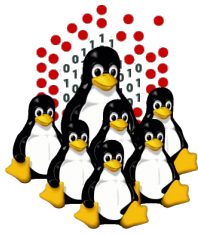




# Customization layers

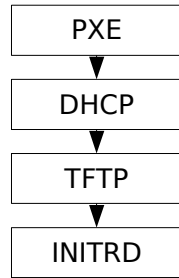
Ramdisk/Ramfs for disk-less nodes, rescue and HW test



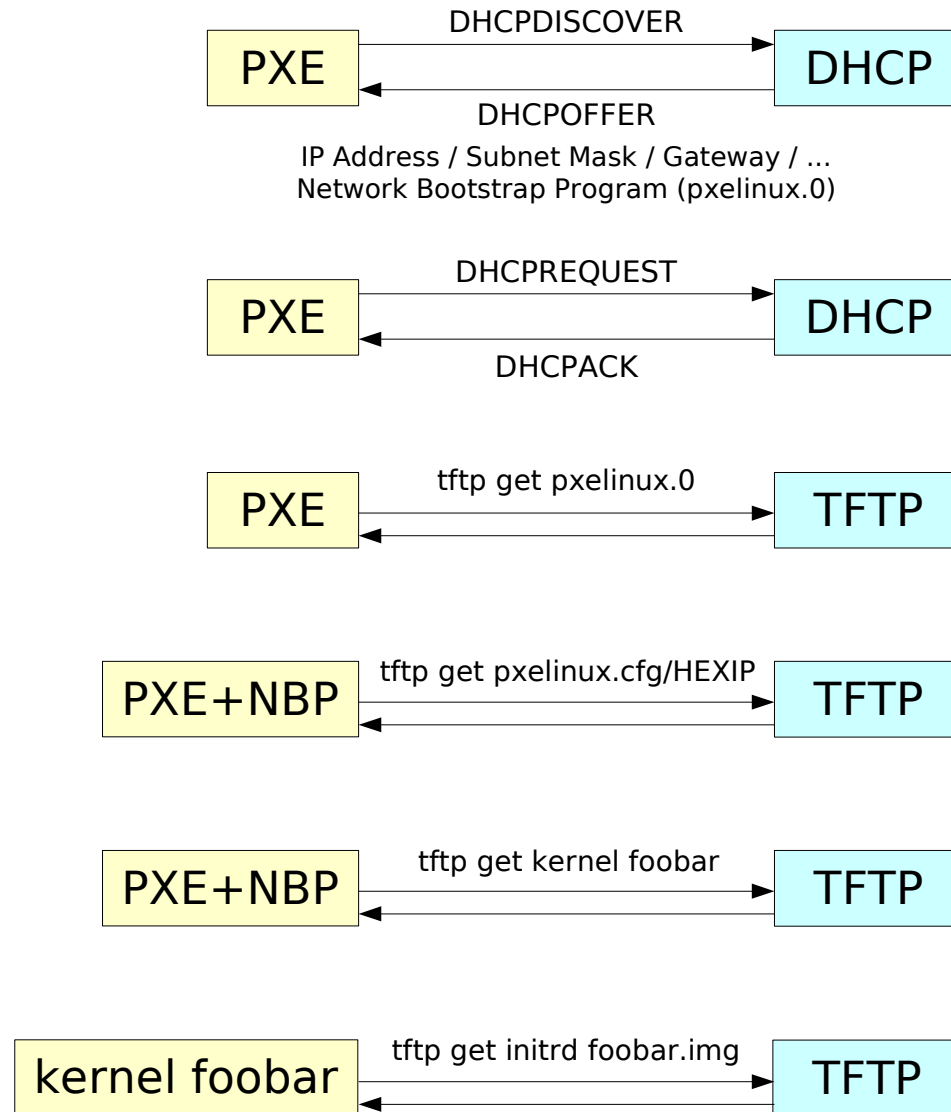


# Network booting (NETBOOT)

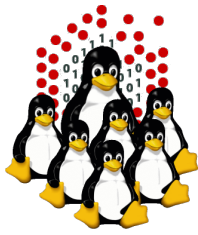
## PXE + DHCP + TFTP + KERNEL + INITRD



CLIENT / COMPUTING NODE



SERVER / MASTER NODE



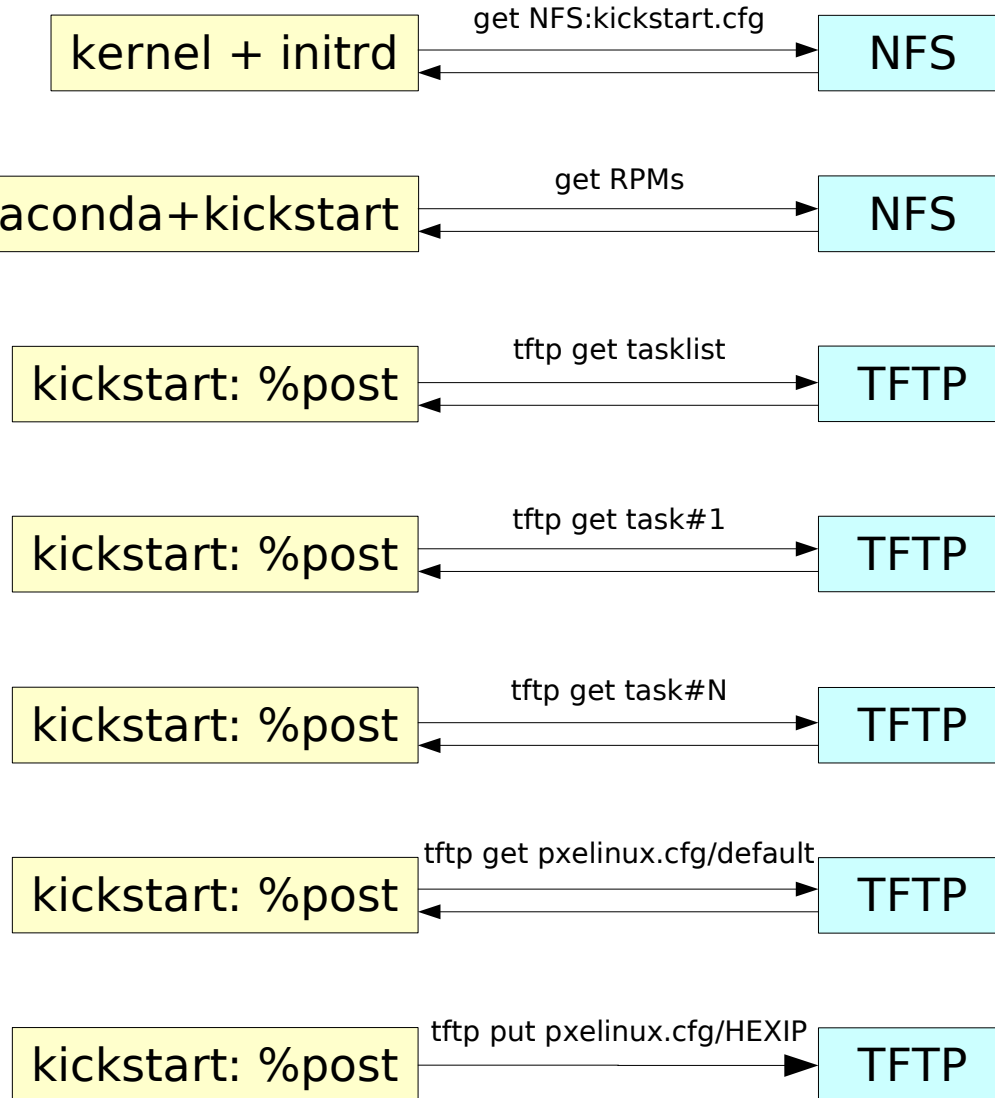
# Network-based Distributed Installation

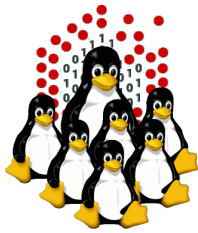
## NETBOOT + KICKSTART INSTALLATION

Installation

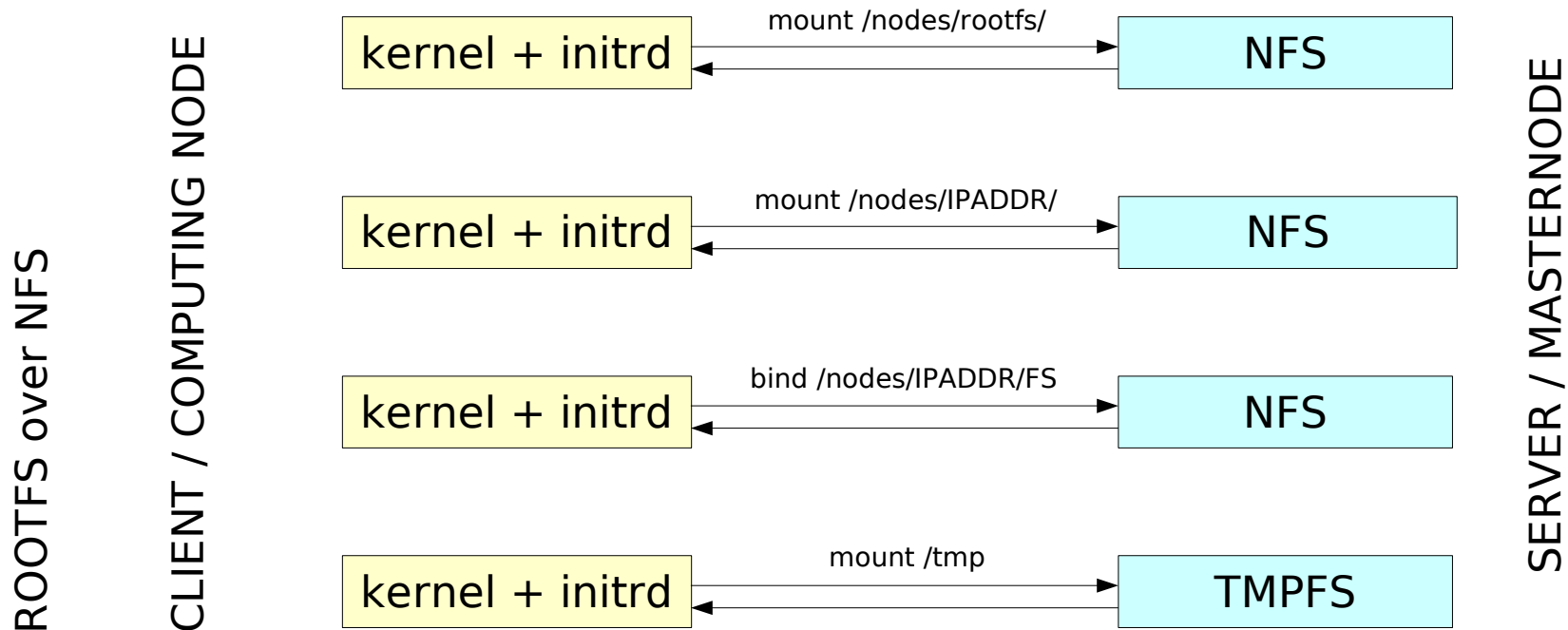
CLIENT / COMPUTING NODE

SERVER / MASTER NODE





# Diskless Nodes NFS Based NETBOOT + NFS



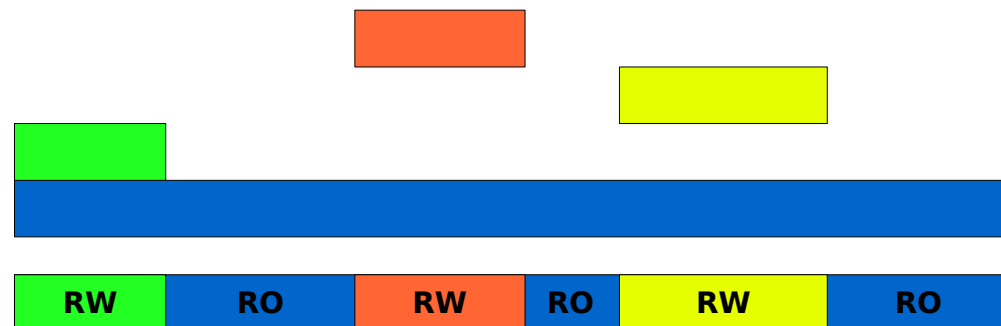
**/tmp/ as tmpfs (RAM)**

**/nodes/10.10.1.1/var/**

**/nodes/10.10.1.1/etc/**

**/nodes/rootfs/**

**Resultant file system**

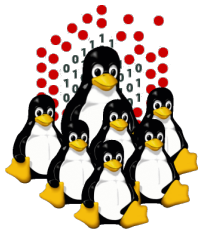


**RW (volatile)**

**RW (persistent)**

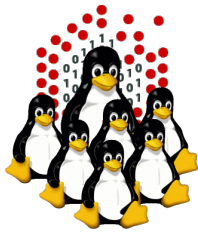
**RW (persistent)**

**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)

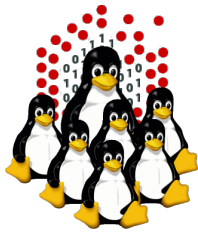


# That's All Folks!



```
( 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>

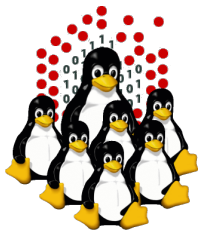
## UnionFS:

- 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>
- SYS LINUX <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