IHEP cluster for Grid and distributed computing

V. Ezhova[1], A. Kotliar[1], V. Kotliar[1*], G. Latyshev[1] , E. Popova[1]

[1]State Research Center of Russian Federation Institute for High Energy Physics, RU-142281, Protvino, Moscow region, Russia

E-mail: {Victoria.Ezhova, Anna.Kotliar, Viktor.Kotliar, Grigory.Latyshev, Ekaterina.Popova}@ihep.ru

[*] Corresponding author

To build a computer cluster for the Grid and distributed computing is a highly complex task. Such cluster has to seamlessly combine grid middleware and different types of the other software in one system with shared cpu, storage, network and engineering infrastructure. To be able to run effectively and be flexible for the still unknown future usage patterns many software systems must be gathered together to build a complete system with high level of complexity. This work present a general possible architecture for such systems and a cluster software stack which could be used to build and operate it using IHEP computer cluster as an example.

What is a computing cluster

Cluster **WHAT**
- compute hardware
- storage hardware
- network hardware
- security and usage policies
- engineering infrastructure
- place

Cluster for **WHAT**
- for computation in more or less one field of science or more or less using similar computation technologies

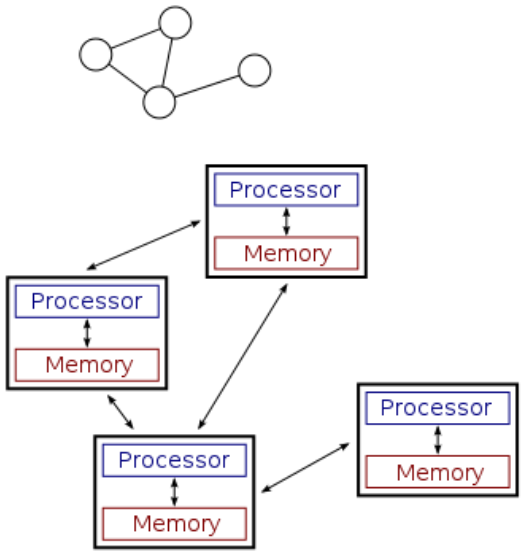
Cluster HOW
- computer technology
- storage technology
- network technology
- management technology
- engineering technology

For high energy physics
- Cuda
- Mathematica
- mpi
- OpenMP
- sse optimization
- caffe
- …

Cluster software **Debian, RHEL**
- DCIM
- puppet
- git
- AFS
- Lustre
- FAI
- Kerberos
- Openldap
- Maui
- torque
- xrootd
- dCache
- HA clusters drbd+pacemaker
- nagios
- Splunk
- elastic search + kibana
- munin
- pmacct
- collectl
- Grid middleware
- XEN, KVM

What is Grid and distributed computing

- Distributed computing is a field of computer science that studies distributed systems.
- A distributed system is a model in which components located on networked computers communicate and coordinate their actions by passing messages.
- The components interact with each other in order to achieve a common goal.
- In distributed computing, a problem is divided into many tasks, each of which is solved by one or more computers, which communicate with each other by message passing.



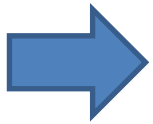
What is Grid and distributed computing

• Grids are a form of distributed computing whereby a “super virtual computer” is composed of many networked loosely coupled computers acting together to perform large tasks. Complete computers (with onboard CPUs, storage, power supplies, network interfaces, etc.) connected to a computer network (private or public) by a conventional network interface, such as Ethernet. This is in contrast to the traditional notion of a supercomputer, which has many processors connected by a local high-speed computer bus..

• Grid computers tend to be more heterogeneous and geographically dispersed (thus not physically coupled) than cluster computers.
• Grids are often constructed with general-purpose grid middleware software libraries.

Cluster step by step

Cluster **WHAT**
- compute hardware
- storage hardware
- network hardware
- security and usage policies
- engineering infrastructure
- place



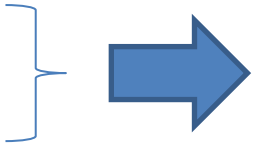
Infrastructure manager

Describe in human readable form:
- compute hardware
- storage hardware
- network hardware
- management hardware
- engineering infrastructure
- place

Has program interface.

Cluster step by step

Cluster **WHAT**
- engineering infrastructure
- place

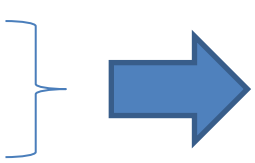


Place, power, cooling

Define resources installation:
- capacity
- reliability
- connectivity

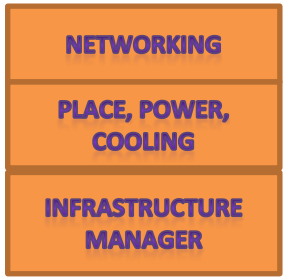
Cluster step by step

Cluster **WHAT**
• network hardware



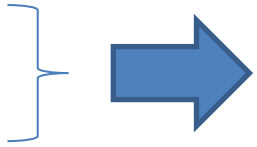
Networking

The core of the distributed computing
• must be scalable
• very reliable
• high throughput network for data transfers
• independent from general purpose network as much as possible (dns, gateways, dhcp, proxy)
• could be several networks (computing, storage, infrastructure, power)



Cluster step by step

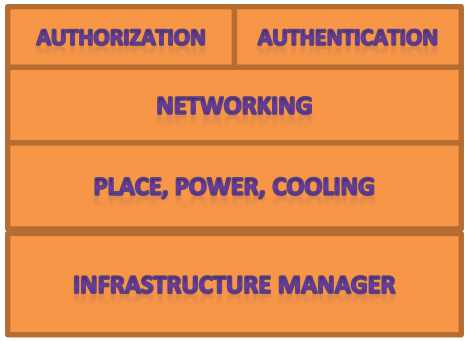
Cluster **WHAT**
- security and usage policies



Authentication is the process of ascertaining that somebody really is who he claims to be.

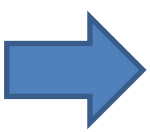
Authorization refers to rules that determine who is allowed to do what.

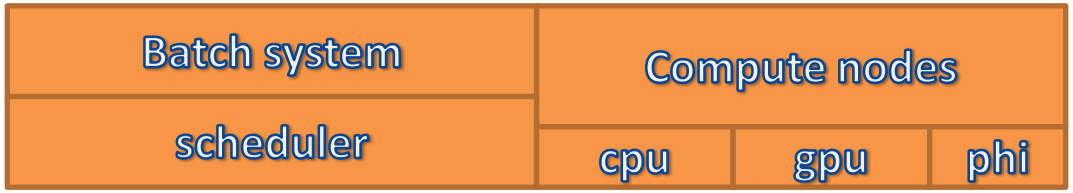
Authentication = login + password (who you are)
Authorization = permissions (what you are allowed to do)

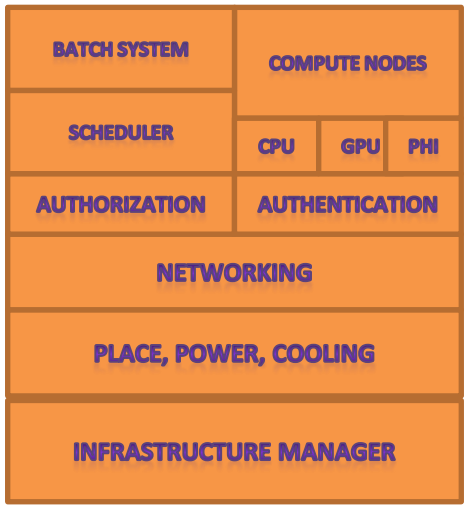


Cluster step by step

Cluster **WHAT**
- compute hardware



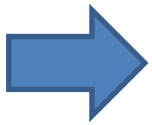


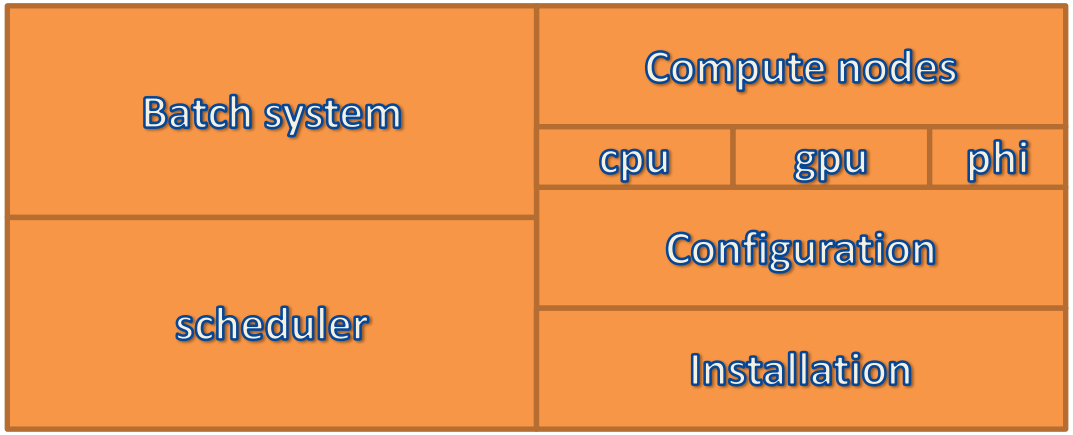


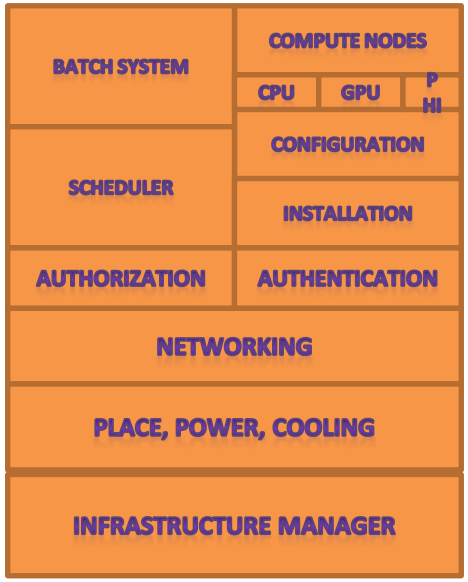
The main resource for computing – compute nodes
- batch system convenient way of running tasks
- orchestrator for compute nodes of different types
- scheduler is a brain of effective and fairly use the resources

Cluster step by step

Cluster **WHAT**
- compute hardware



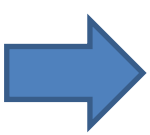


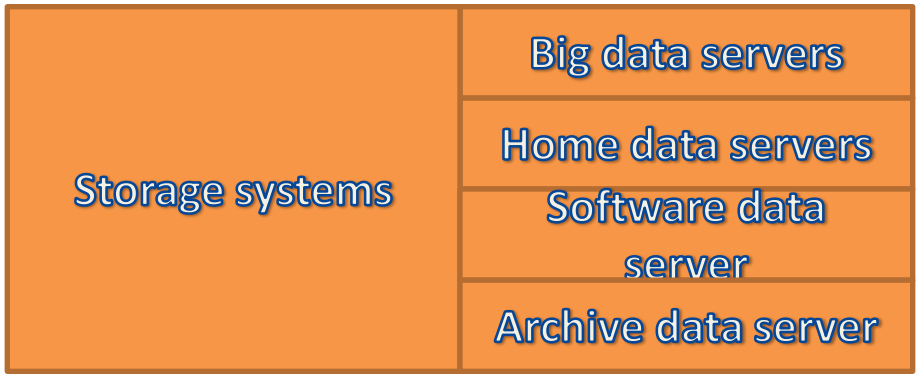


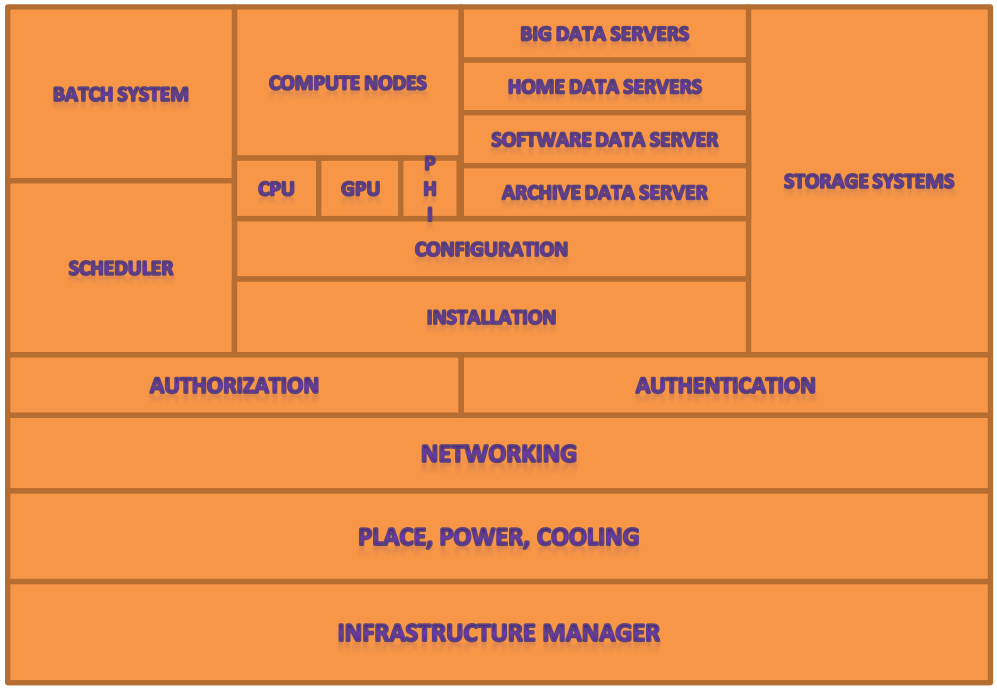
Many nodes need:
- Automatic installation
- Automatic configuration

Cluster step by step

Cluster **WHAT**
- storage hardware





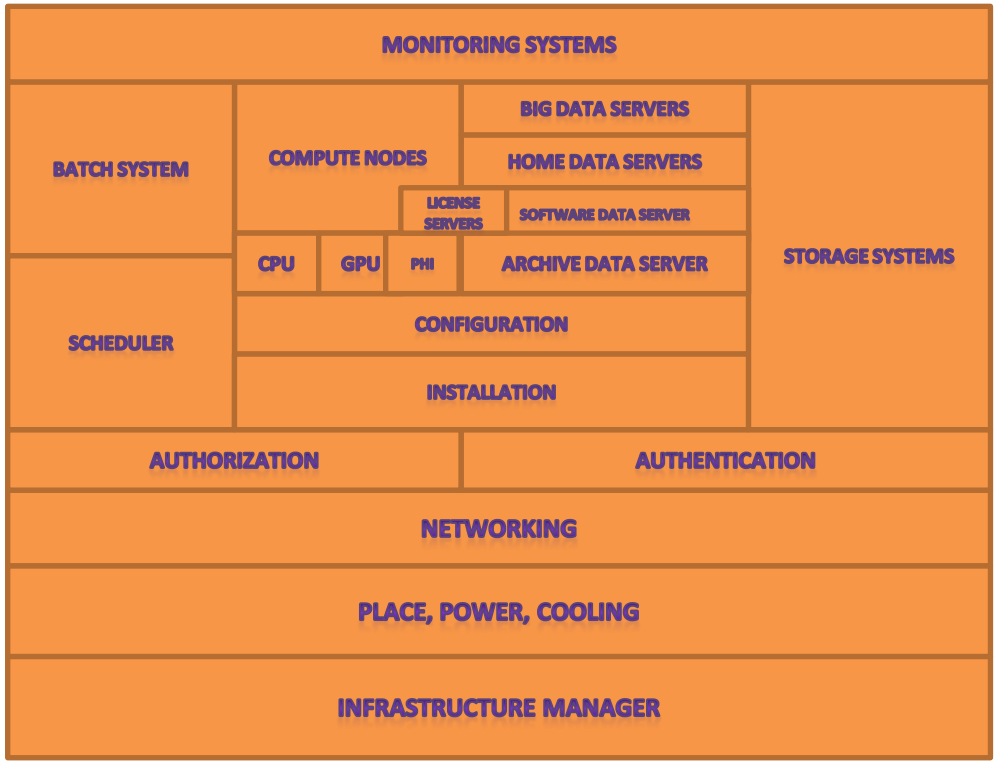


Store different types of data:
- home dirs with auto-backup
- big data for fast analysis
- software area for small files
- archive storage for long term storage and backup

Cluster step by step

Cluster

Monitoring systems

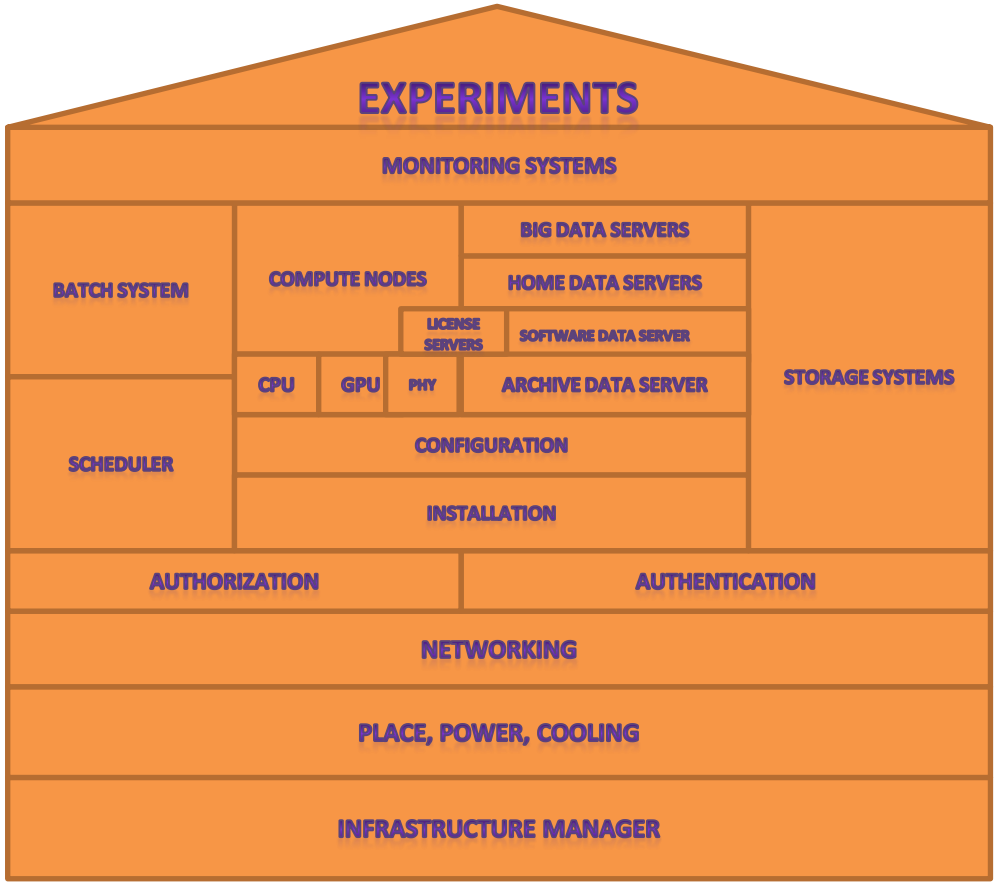


Many types of monitoring:

- you know exactly how the cluster works;
- accounting/billing for the cluster usage.

Cluster step by step

Cluster for WHAT

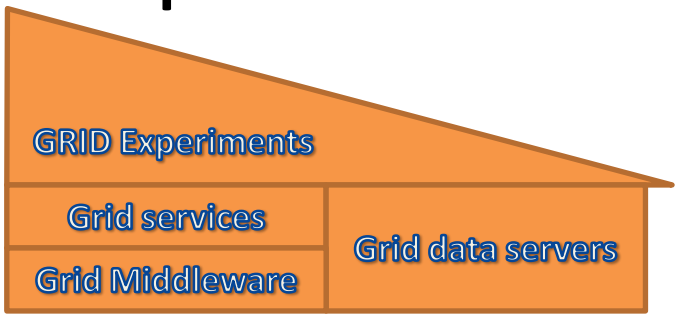


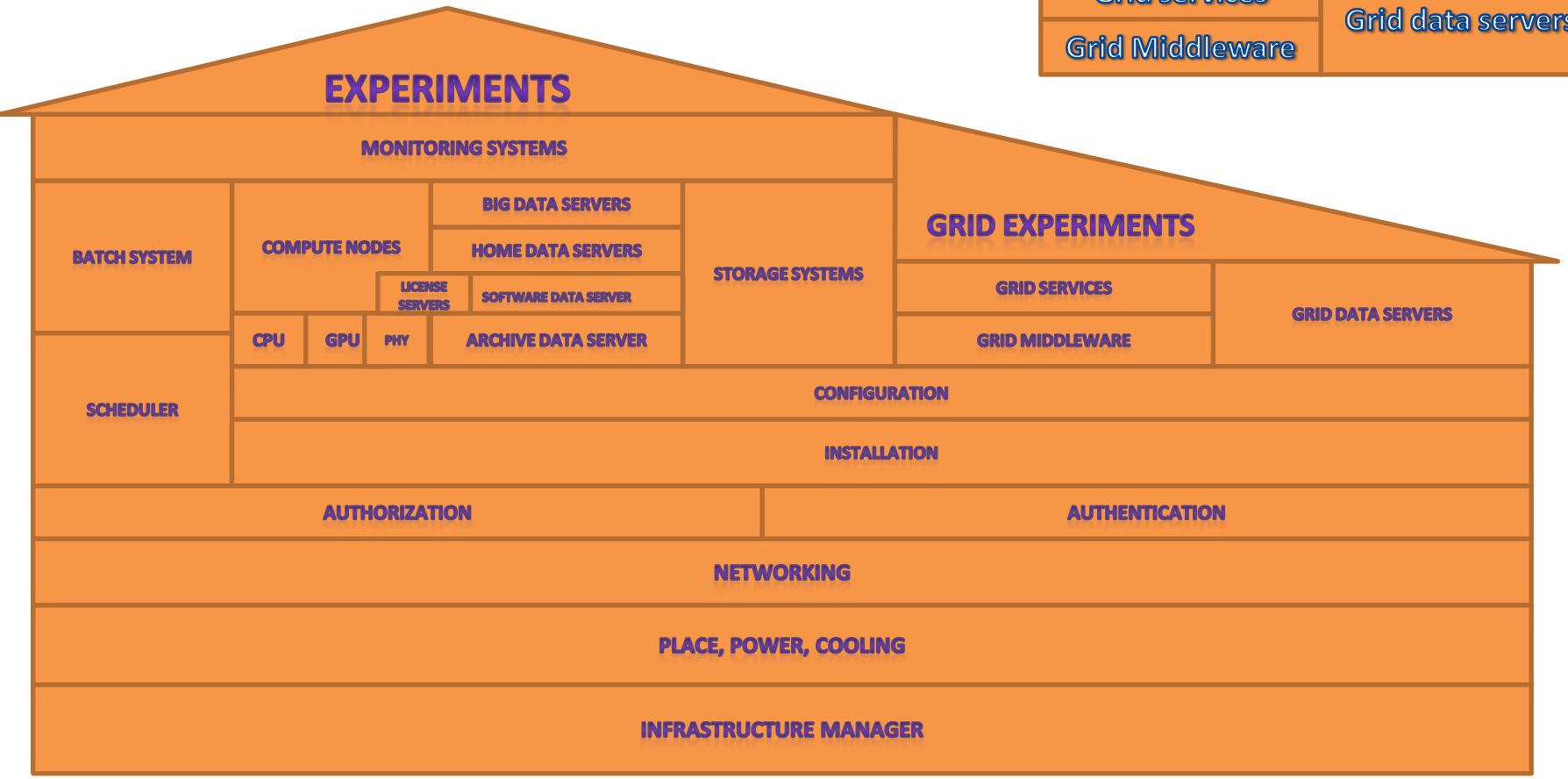
For whom cluster created:

- all experiments have their specific usage patterns
- in any time they can ask about something new
- all time communication about admins and users how to use the cluster in the best way to achieve users goals.

Cluster step by step

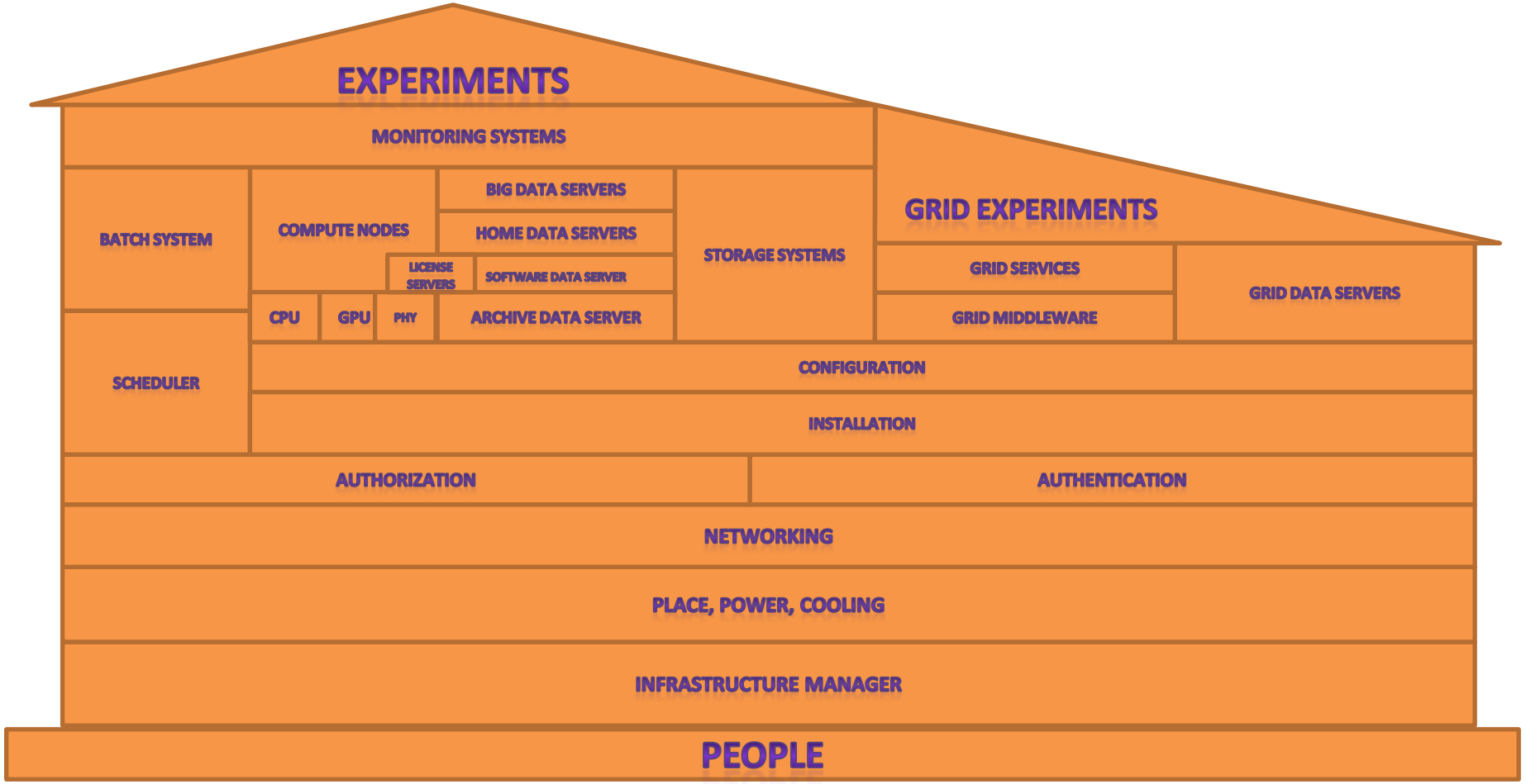
Cluster GRID





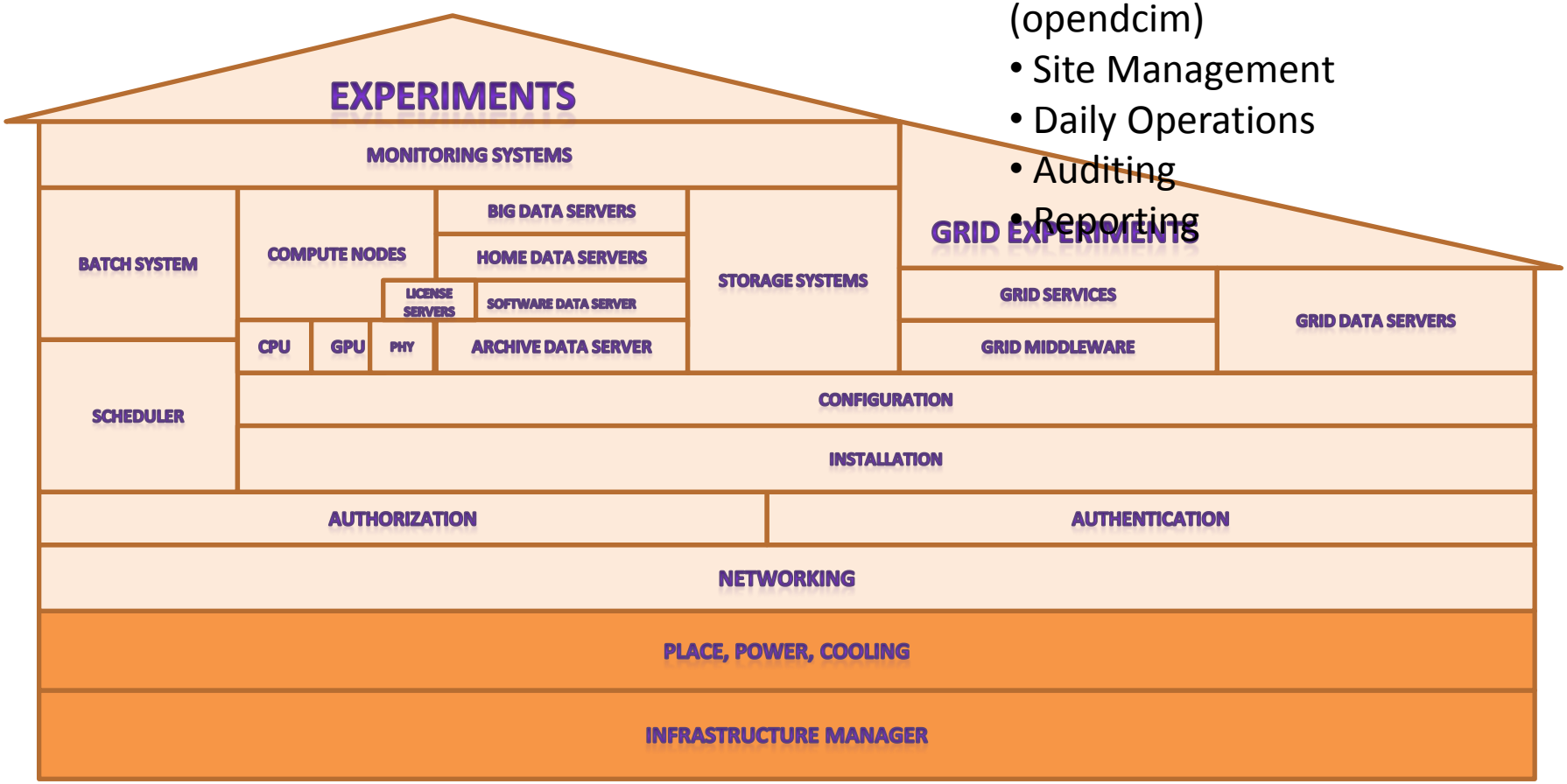
Cluster step by step

Cluster – people



Cluster software step by step

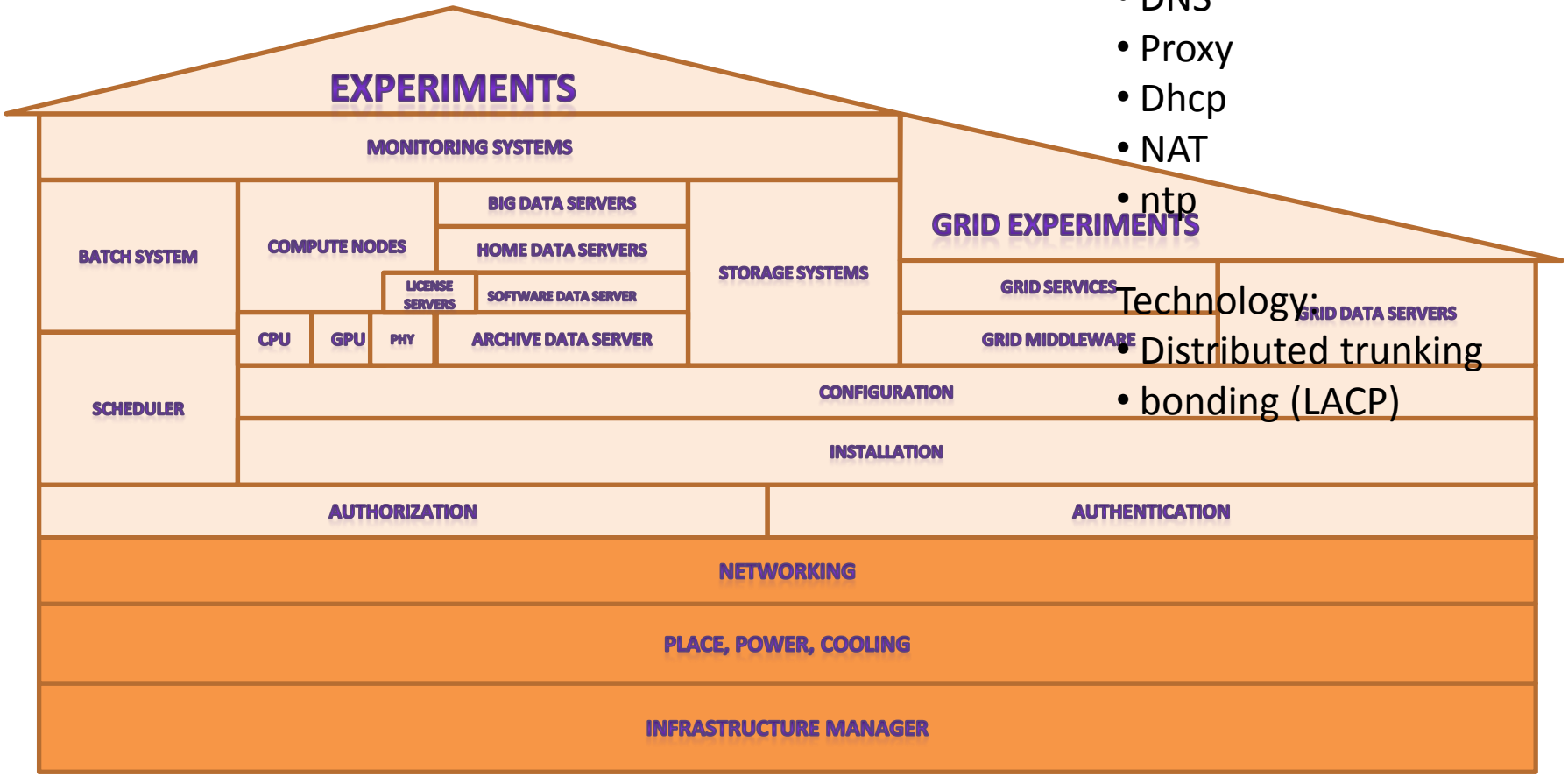
Cluster HOW
Data Center Infrastructure Manager (opendcim)
- Site Management
- Daily Operations
- Auditing
- Reporting



Cluster software step by step

Cluster network HOW:

- Ssh configuration
- DNS
- Proxy
- Dhcp
- NAT
- ntp



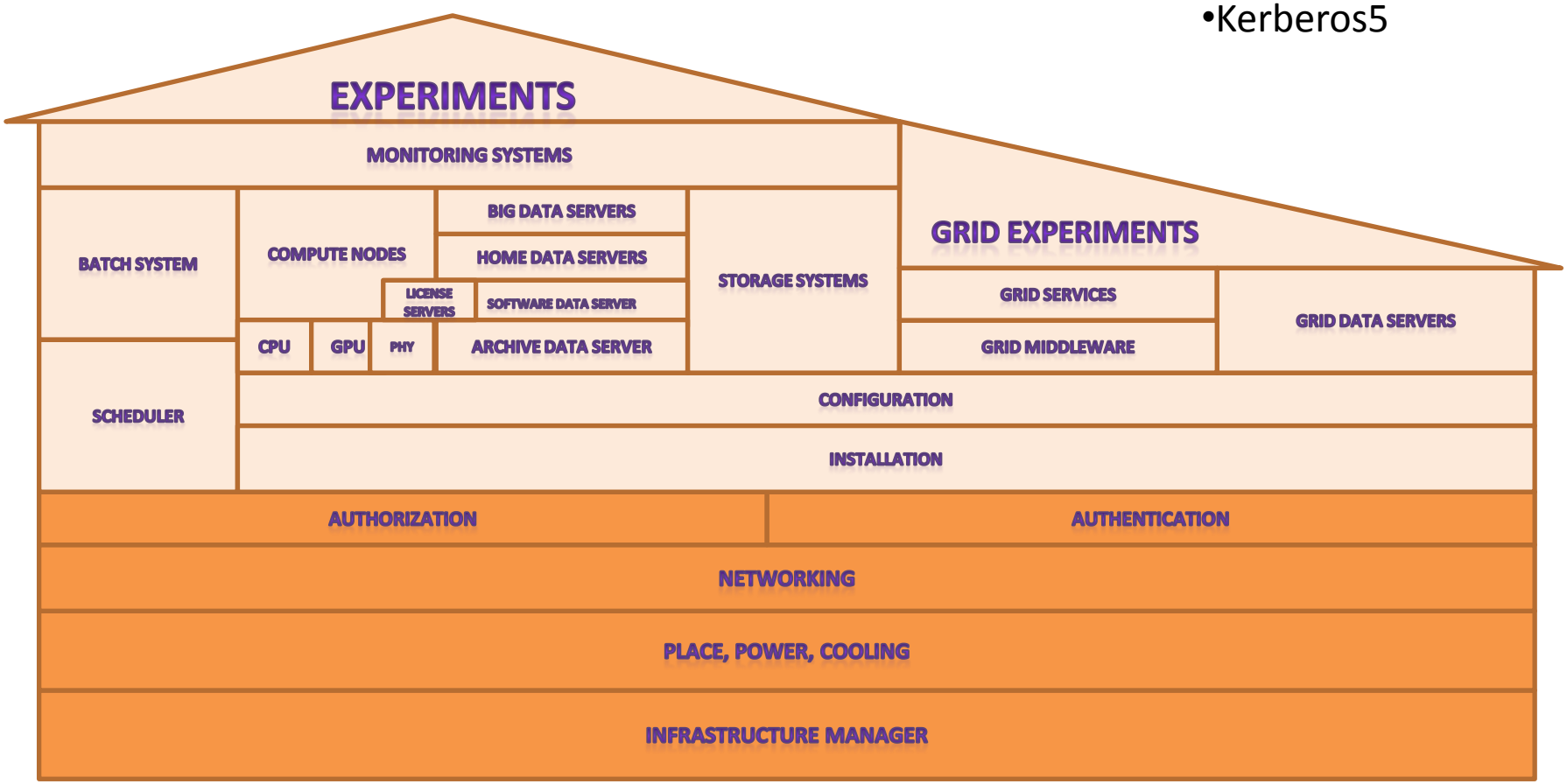
Technology:

- Distributed trunking
- bonding (LACP)

Cluster software step by step

Cluster 2A HOW:

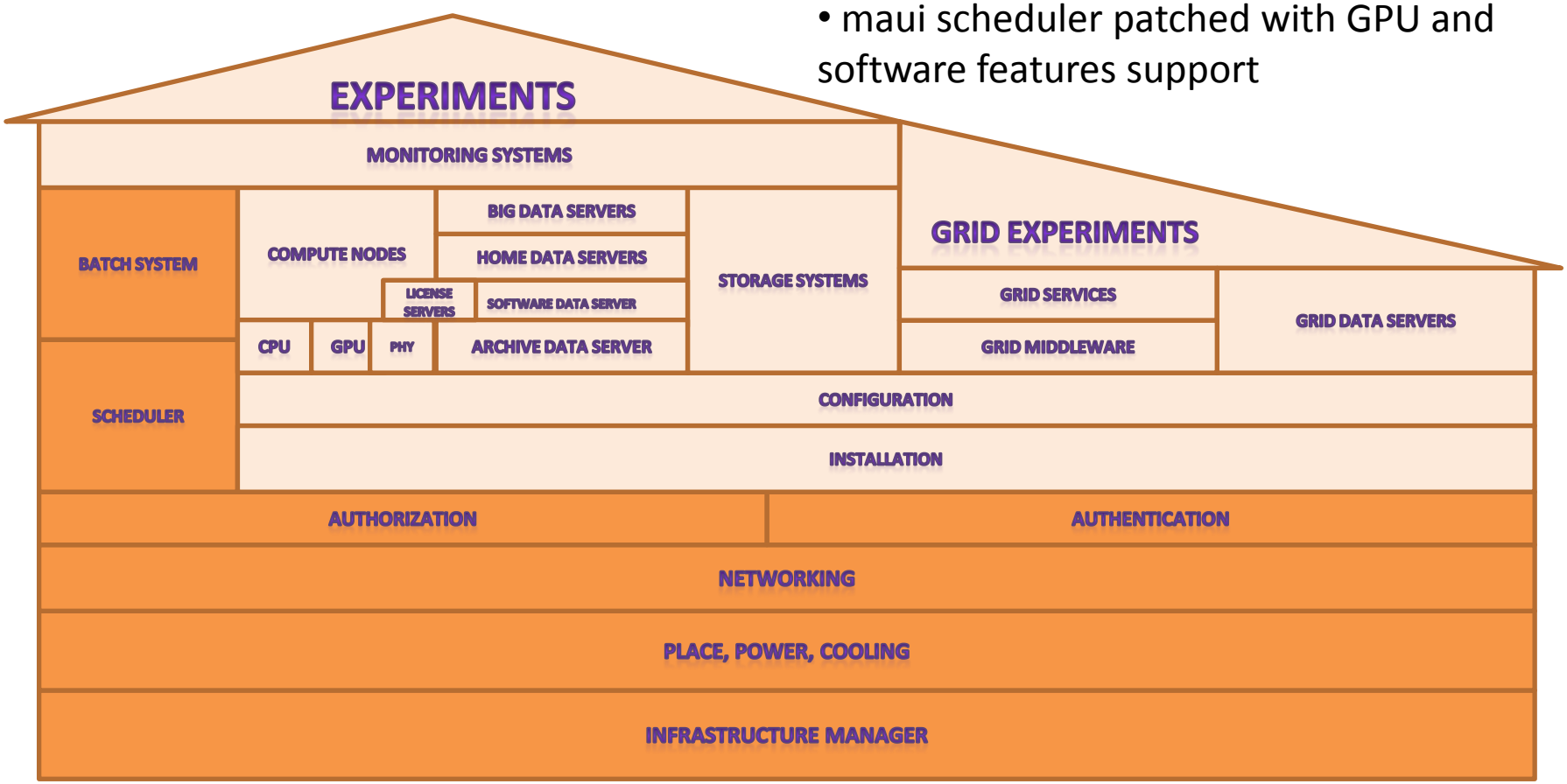
- OpenLDAP
- Kerberos5



Cluster software step by step

Cluster Batch system HOW:

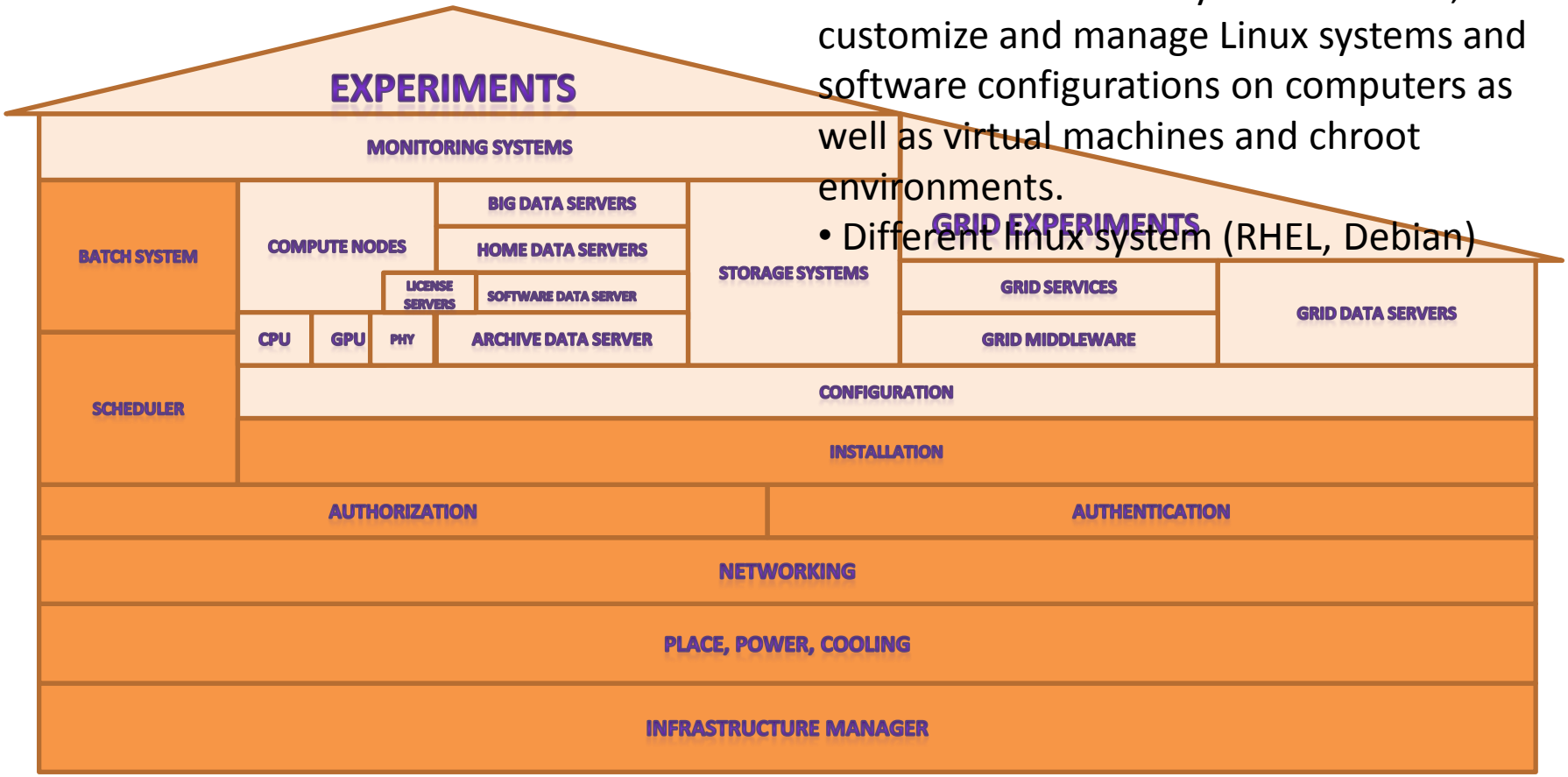
- TorquePBS patched with Kerberos5 support
- maui scheduler patched with GPU and software features support



Cluster software step by step

Cluster installation HOW:
Fully Automatic Installation (FAI)

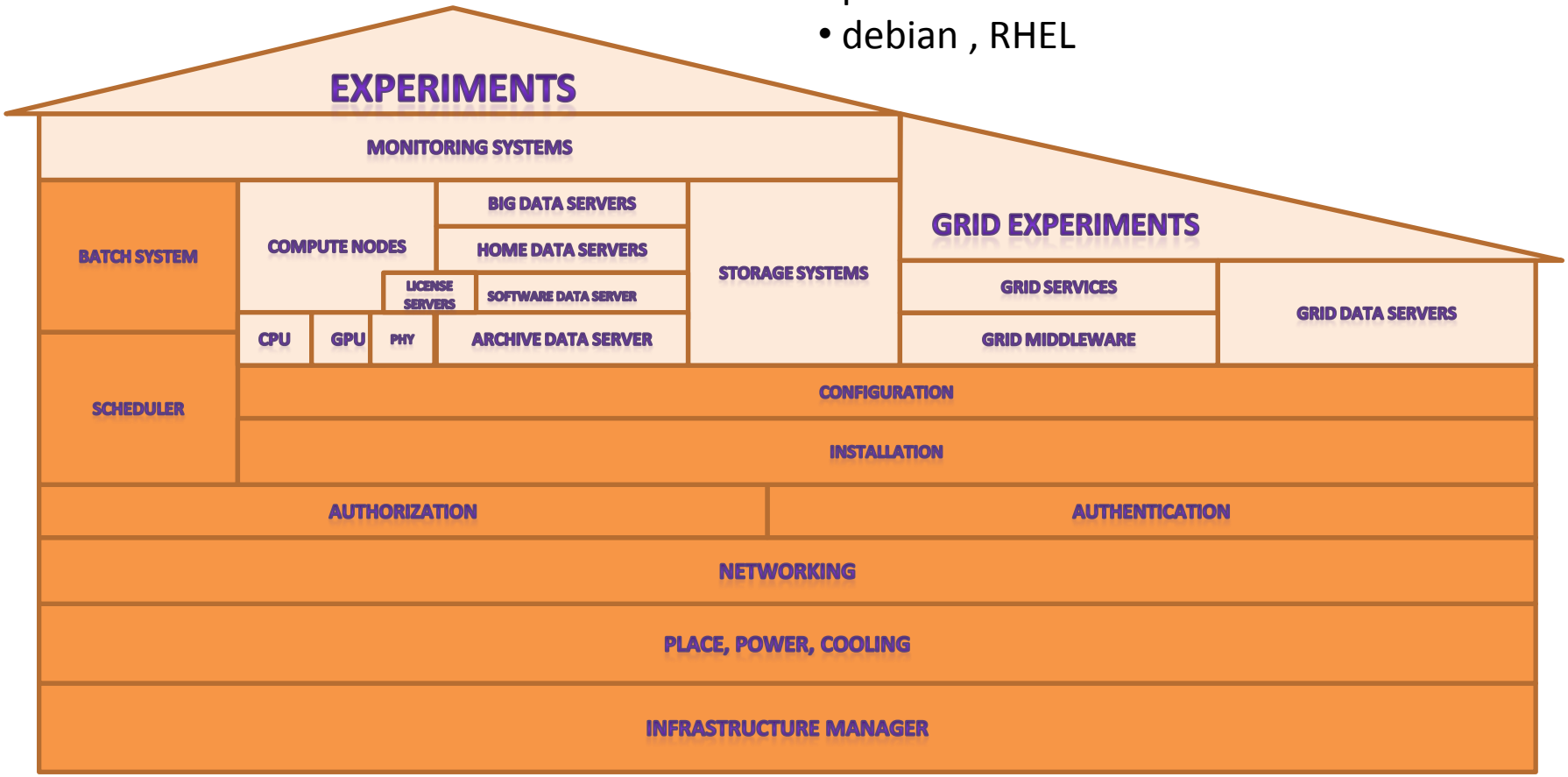
- is a non-interactive system to install, customize and manage Linux systems and software configurations on computers as well as virtual machines and chroot environments.
- Different linux system (RHEL, Debian)



Cluster software step by step

Cluster configuration HOW:

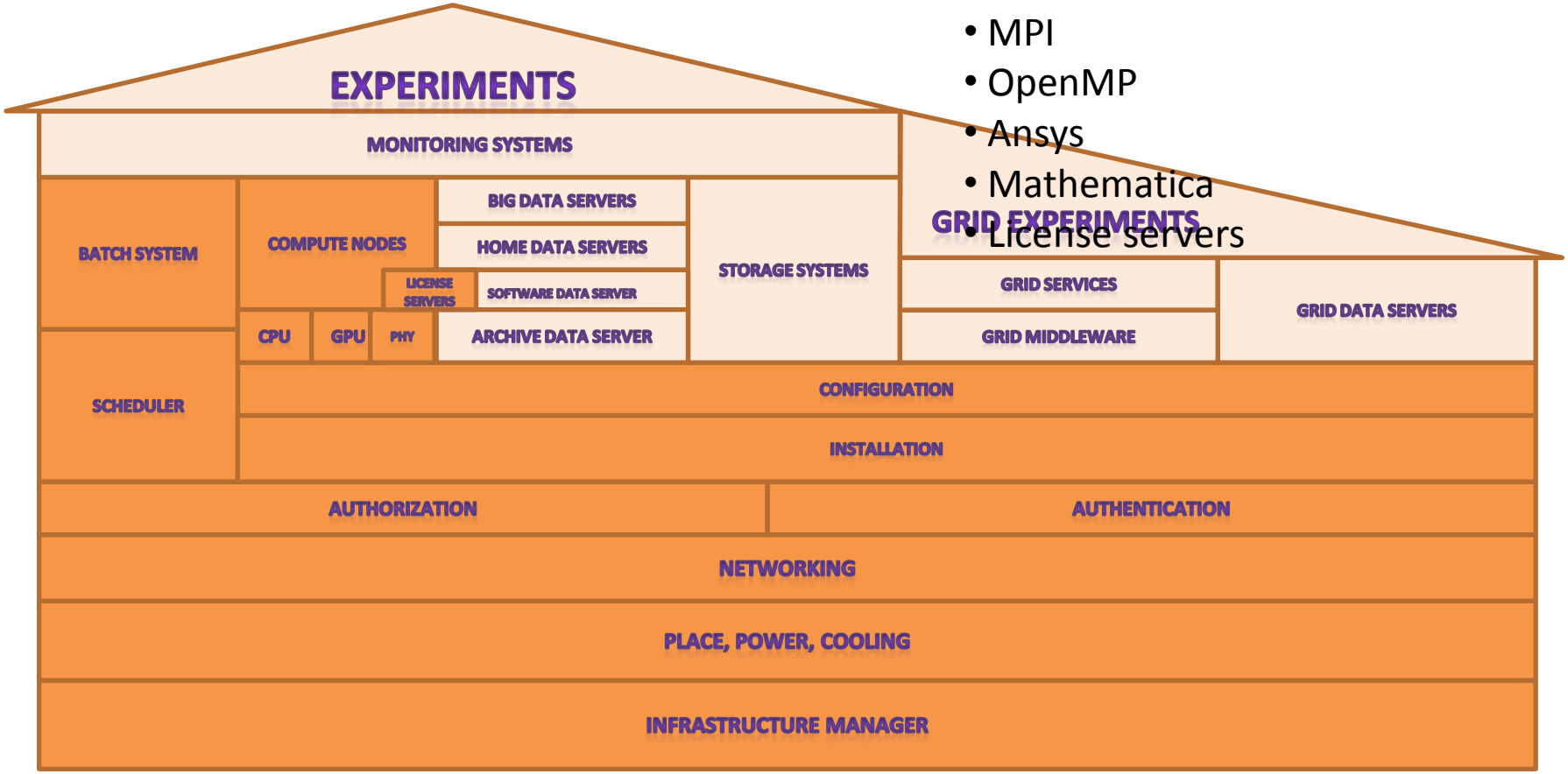
- Puppet (git)
- pdsh
- debian , RHEL



Cluster software step by step

Cluster compute nodes HOW:

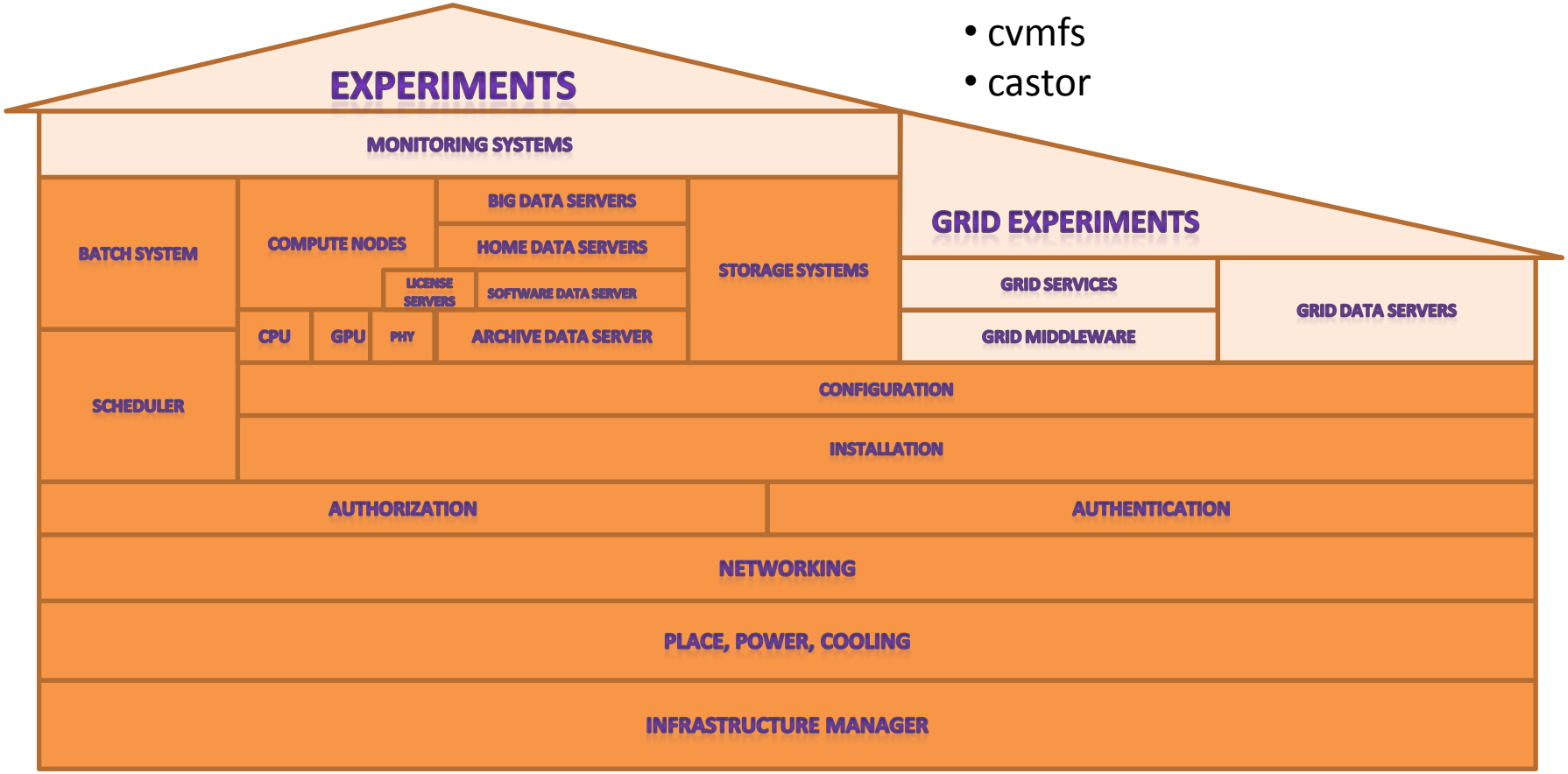
- Scientific Linux
- Cuda
- MPI
- OpenMP
- Ansys
- Mathematica
- License servers



Cluster software step by step

Cluster storage systems HOW:

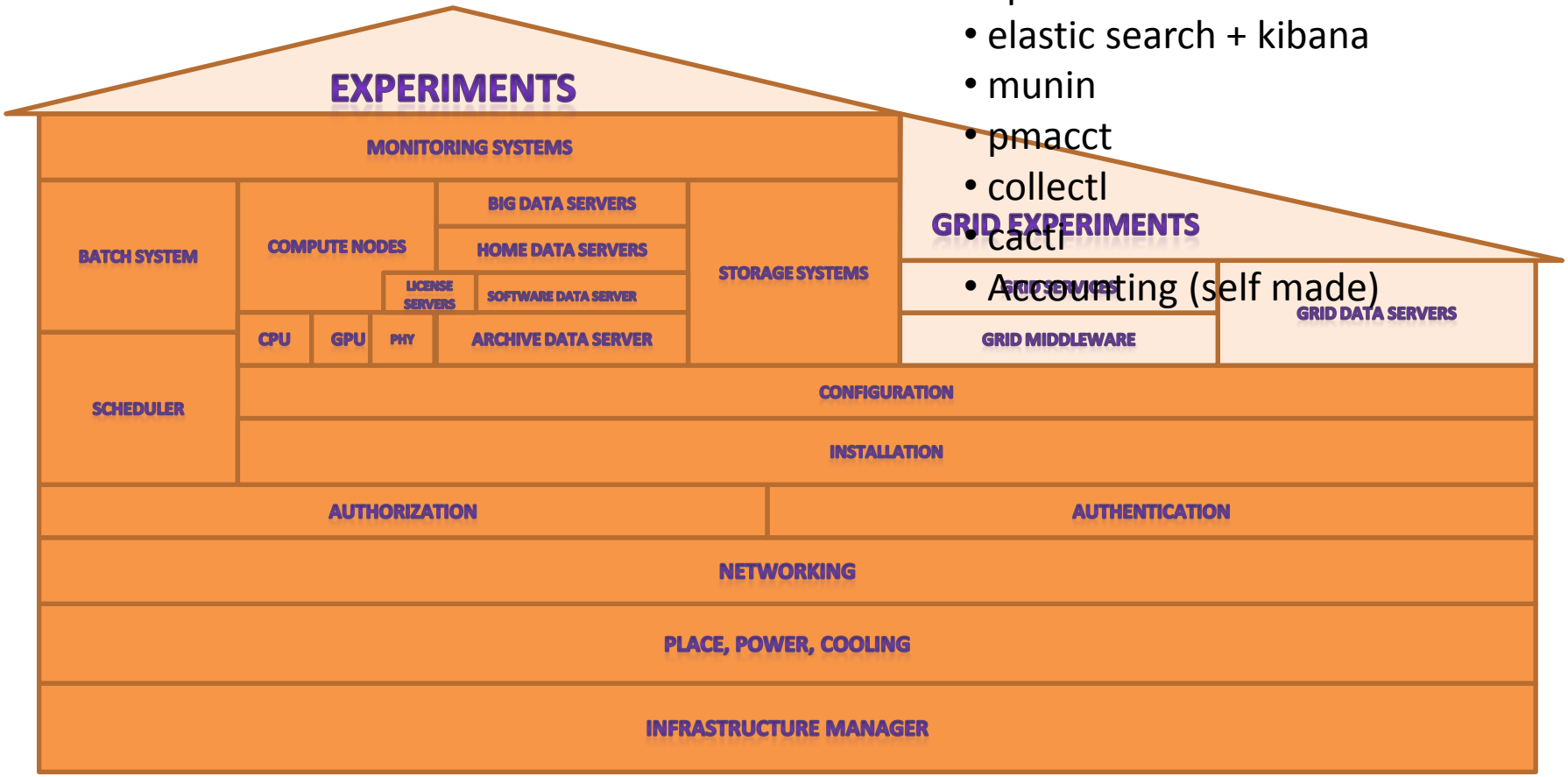
- Lustre
- AFS
- cvmfs
- castor



Cluster software step by step

Cluster monitoring systems HOW:

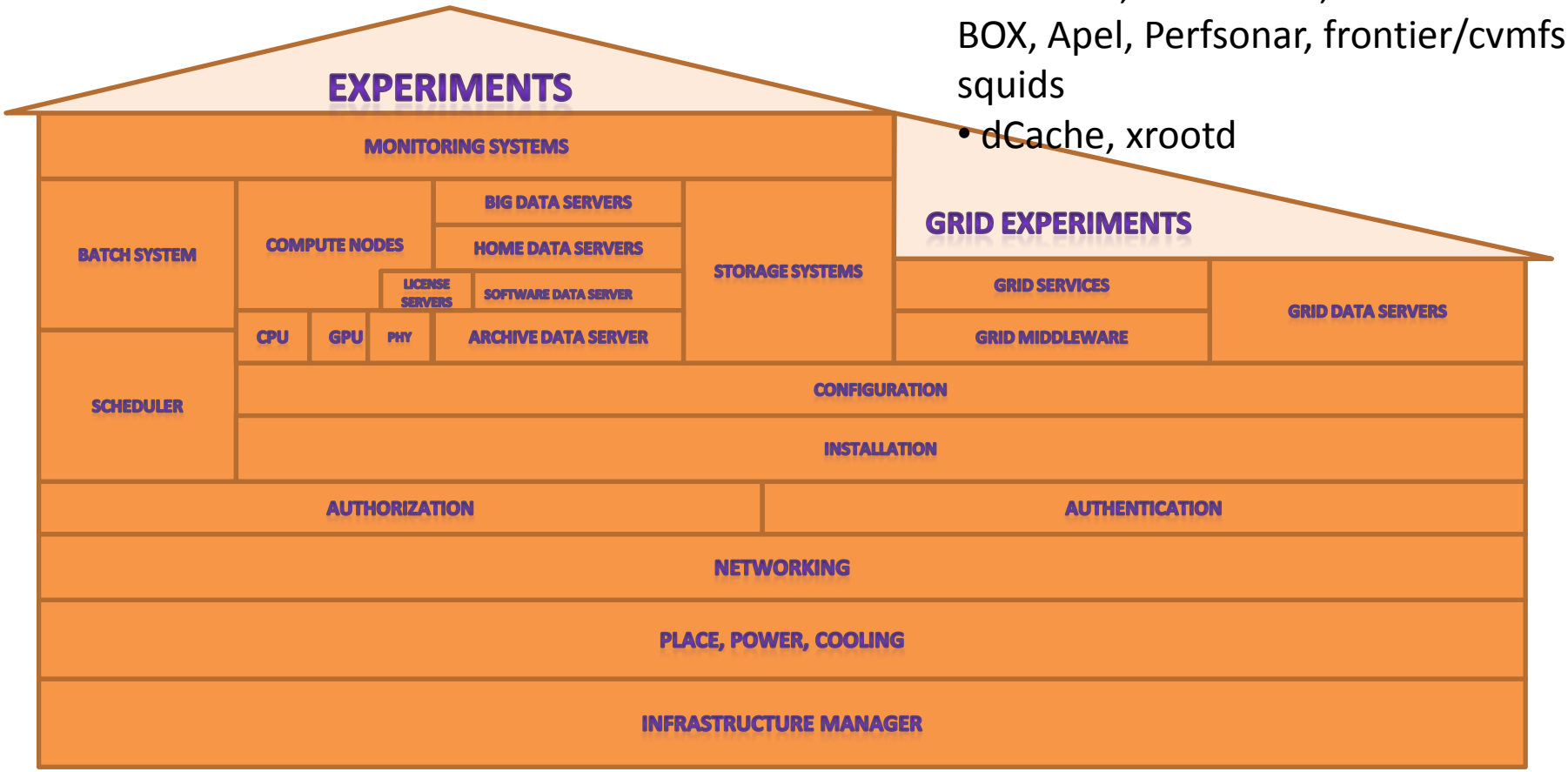
- nagios
- Splunk
- elastic search + kibana
- munin
- pmacct
- collectl
- cacti
- Accounting (self made)



Cluster software step by step

Cluster Grid software HOW:

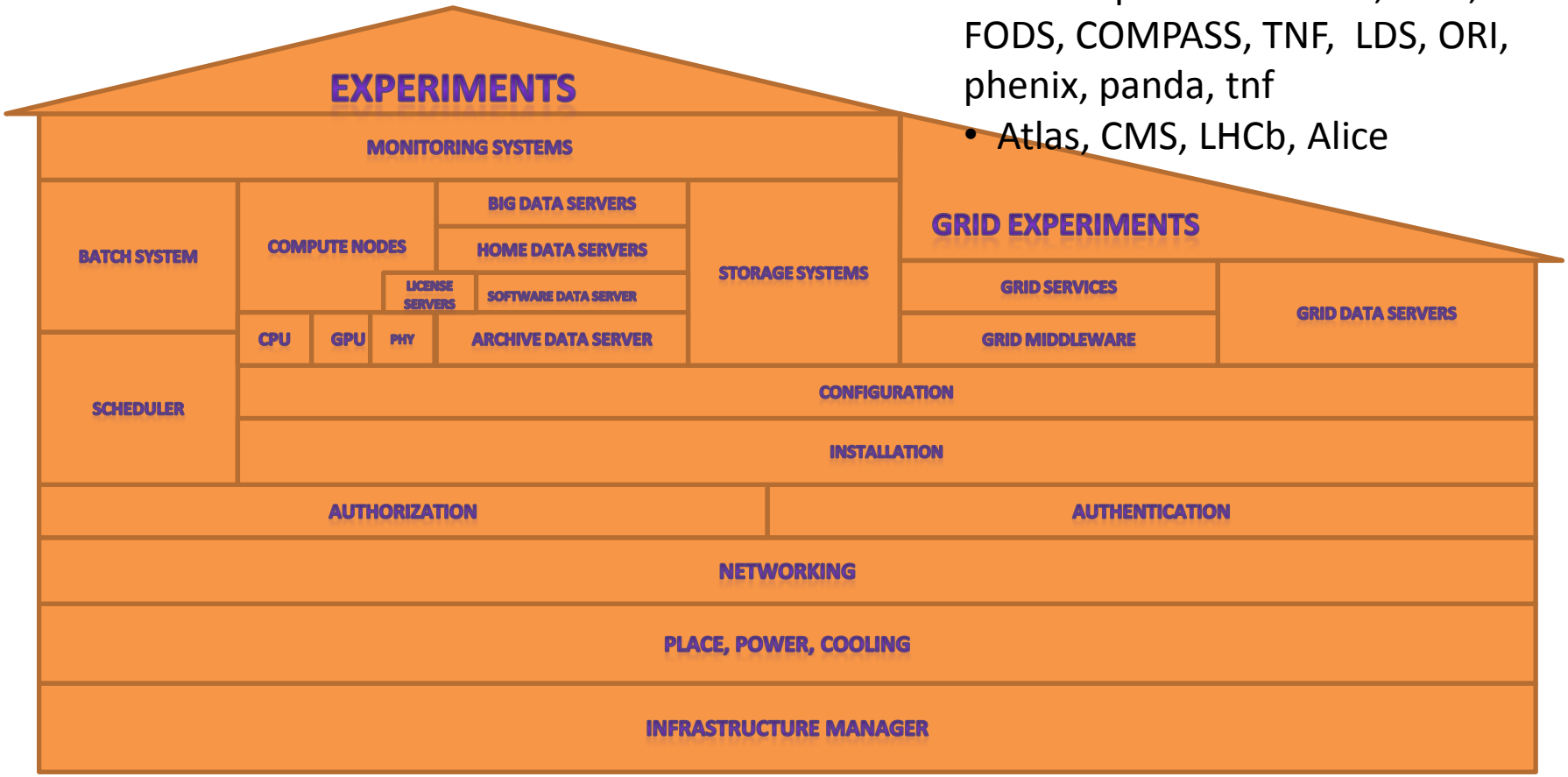
- EMI+UMD
- site BDII, CREAM-CE, WLCG VO-BOX, Apel, Perfsonar, frontier/cvmfs squids
- dCache, xrootd



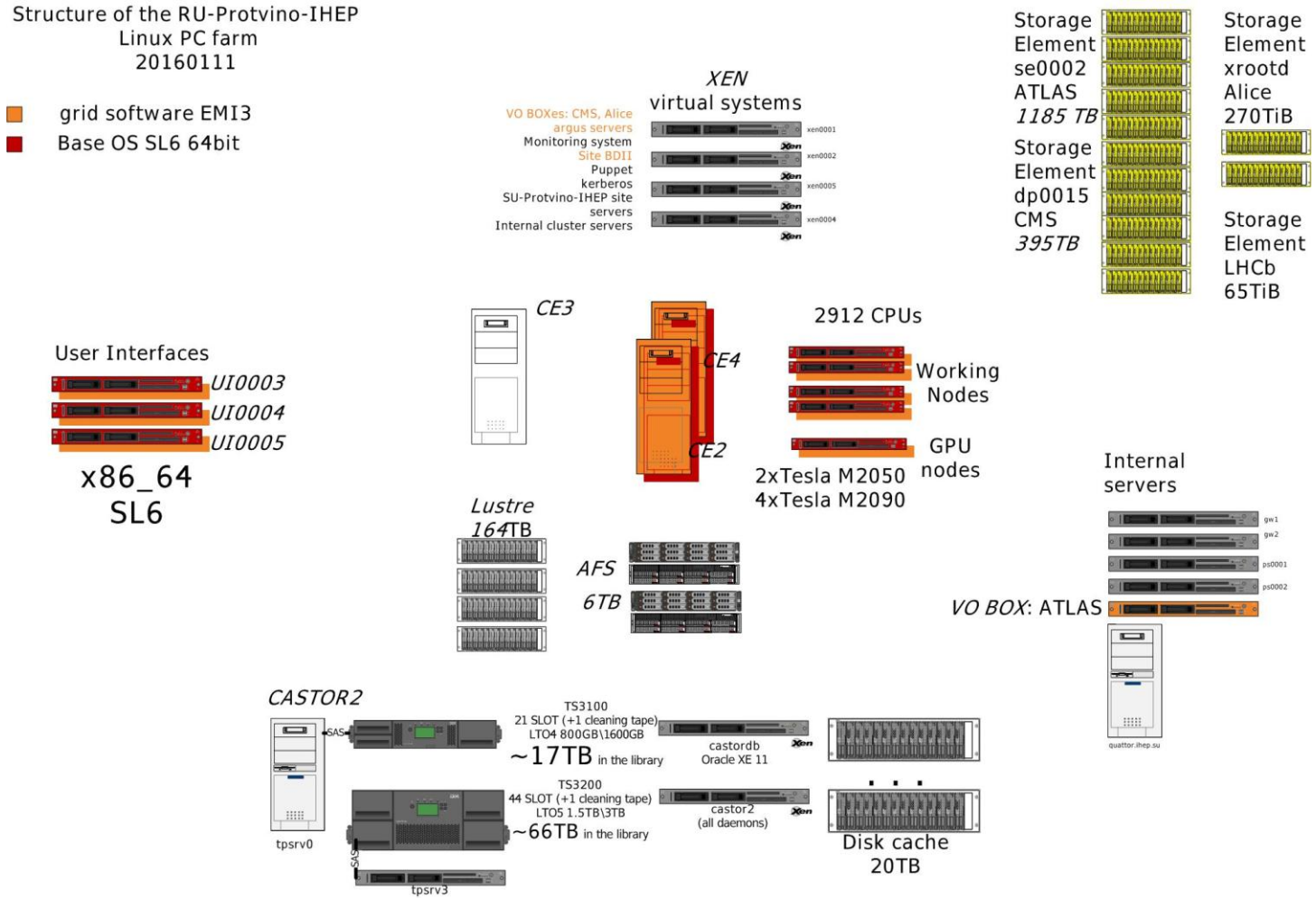
Cluster software step by step

Cluster for WHOM:

- IHEP physicist
- IHEP experiments: BEC, OKA, FODS, COMPASS, TNF, LDS, ORI, phenix, panda, tnf
- Atlas, CMS, LHCb, Alice



Cluster structure



Thank you!

Any questions?