How to setup a cluster

What's a cluster???

There are two types of "clusters":

- Thightly coupled cluster ("real" cluster)
  - Latency of the interconnect is of main importance
  - Connected via InfiniBand, Myrinet or Quadrics
  - Mostly used for a single, common task ("massively parallel")
  - Identical setup of (almost) all machines in the cluster
- Loosely coupled cluster ("farm")
  - Latency is not important -> any interconnect possible
  - Applications on separate machines usually independent
  - Heterogeneous cluster very likely -> different setups

What do we have at CERN

The obvious answer is:

- A very heterogeneous farm
  - Off-the-shelf PCs (Dual-CPU)
  - Several generations of machines
  - Soon different architectures: x86, em64t (amd64)
  - Different usages for a particular generation possible
  - A few thousand boxes in the CC
- Special machines
  - Database servers (Oracle, Oracle RAC)
  - Opencluster: Itanium architecture
  - Standard boxes for special purposes

How do we setup/install the farm?

Very carefully ;-)

- Using standard tools
  - Kickstart and Netboot (TFTP/PXE)
  - RPMS reside on two servers (load balancing and failover)
- Custom tools -> QUATTOR
  - Template based configuration database
  - Automatic generation of kickstart-files
  - Central control of configuration(s)
- Private tools
  - Mainly scripts

How does it work??

With QUATTOR

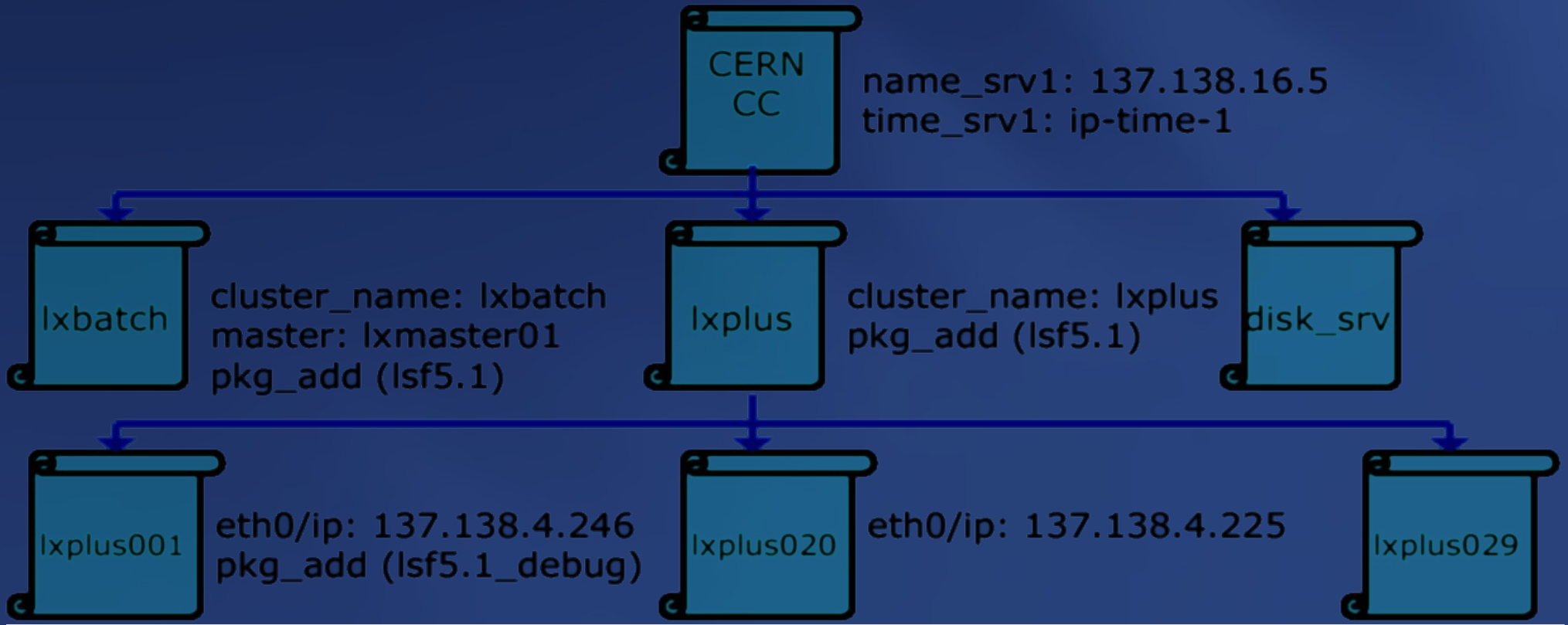
- Create the templates
  - One per hardware configuration (rel. large)
  - One per box (small, mainly IP info)
  - Special info included as required
- Create the Kickstart file
- Install the box via kickstart/netboot
- Run the QUATTOR daemons
  - Upgrades; installation of additional software, etc.

Key Concepts of QUATTOR

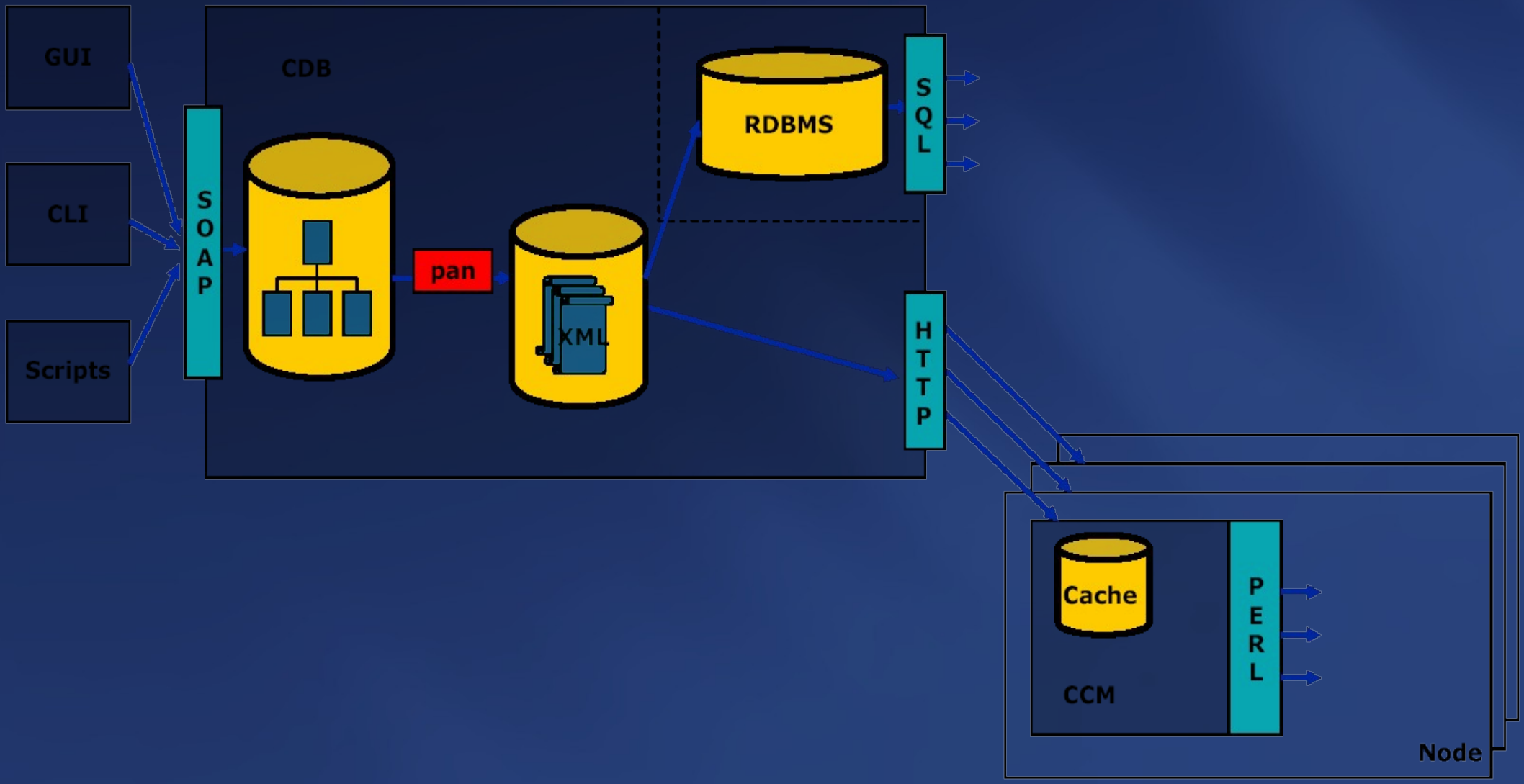
- Autonomous Nodes
  - Local config files; No remote mgmt. scripts; Indep. of global FS (AFS, NFS)
- Central Control
  - Primary config file is kept centrally (and replicated on the nodes)
  - Single source for all config files
- Reproducibility
- Scalability
  - Load balanced servers, Scalable protocols, etc.
- Use of public Standards
  - http(s), XML, rpm/pkg, SysV init scripts, etc.
- Portability
  - Any Linux distro, Solaris, other Unices, etc.

The Configuration

- Information is arranged in templates
- Common properties are stored only once
- Hierarchy of templates possible



Configuration Management



Configuration Database (CDB)

- Keeps complete configuration information
- Describes the *desired* state of the machines
- Transaction mechanism enforces data consistency
- Configurations are validated and under version control
- Roll-back is possible!
- Conflicts are detected

Examples of information in CDB

- Hardware
  - CPU, Hard disk, Network Card, Memory Size
  - Location of the node in the CC
- System
  - Partition Table, Load Balancing information
- Cluster information
  - Cluster name and type, Batch master
- Software
  - Repository information, Service definition (group of rpms)
- Audit information
  - Contract type and number, Purchase date, etc.

Install Manager

- On top of standard vendor installer
  - OS version to install, Network and partition information
  - Which core packages, Custom post-install instructions
- Automated generation of Kickstart file
- Takes care of DHCP (TFTP/PXE) entries
- Available for RedHat Linux
  - Plugins for SuSE, Debian, Solaris, etc. possible

How does openlab do it??

Pre-QUATTOR

- Create Kickstart file by Hand
  - Special platform: supports two architectures (x86 + ia64)
    - Workarounds might be required
  - Relatively complex files
  - Assisted by a number of scripts
  - Current install procedure fits only this particular setup!!
    - ... but for this it's almost "perfect" ;-)
    - It doesn't scale beyond $O(10^2)$ boxes
    - It doesn't scale beyond single admin

What about real clusters??

- Specialized Linux distributions/tools
  - Linux distribution for clusters: i.e. ROCKS
  - Cluster tools on top of standard distros: i.e. OSCAR
  - “Private” or vendor specific OS/Linux distro
- Mostly very simple setup procedure
  - Identical setup of all cluster machines
  - Complications due to large number of boxes ($\mathcal{O}(10^3)$)
  - Possibility to use features of the interconnect (i.e. IB)