Managing Large Linux Farms at CERN

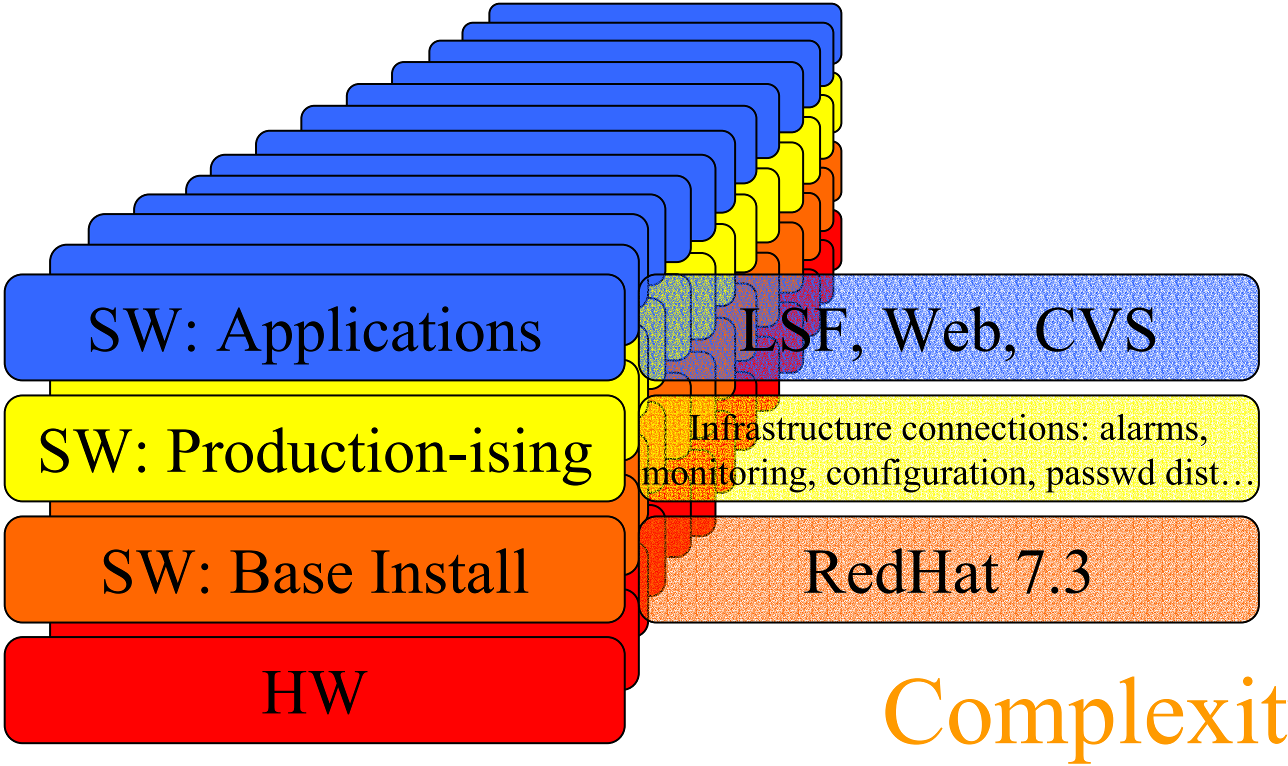
OpenLab:
Fabric Management Workshop

Tim Smith CERN/IT

Contents

- Our Challenge (non-solutions)
  - Scale
  - Complexity
  - Dynamics
- Our Solution
  - Architecture
  - Current Status

Simple Question of Scale?



SW: Applications

SW: Production-ising

SW: Base Install

HW

LSF, Web, CVS

Infrastructure connections: alarms, monitoring, configuration, passwd dist…

RedHat 7.3

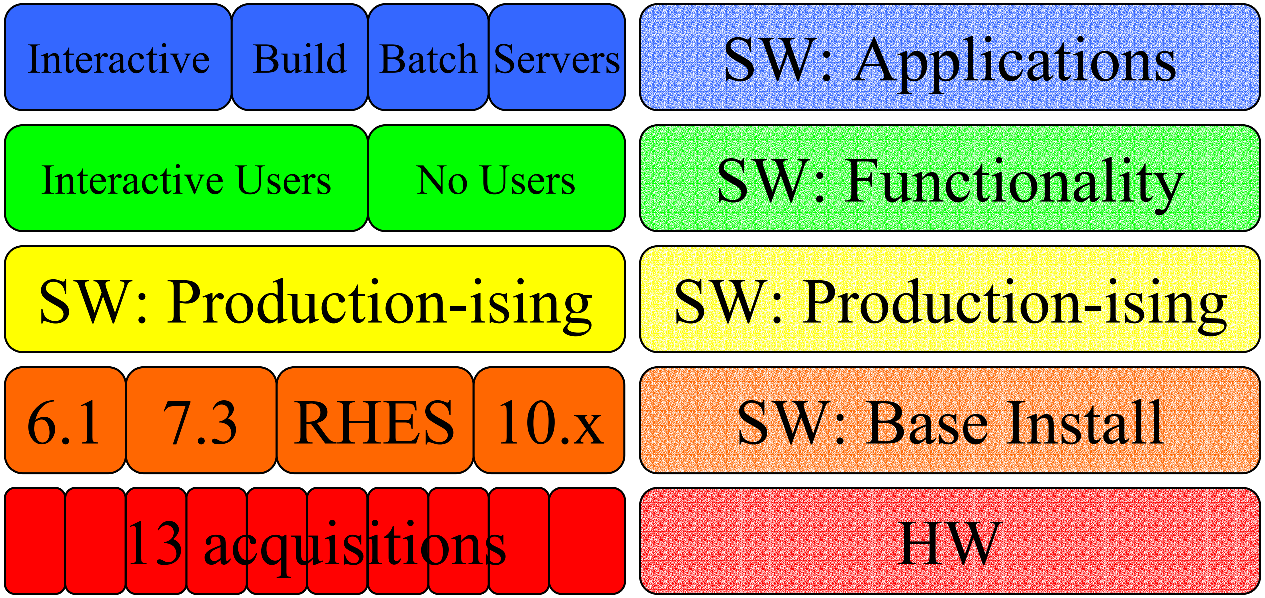
Complexity!

Scale is important!

[~]

- 1000 boxes
- 800,000 Si2k
- 140,000 jobs/wk
- 12,000 uids
- 30 user communities
- 150 simul. indep. applics
- Public network
- 20 root priv

Complexity: Static Configuration



Interactive | Build | Batch | Servers — SW: Applications

Interactive Users | No Users — SW: Functionality

SW: Production-ising — SW: Production-ising

6.1 | 7.3 | RHES | 10.x — SW: Base Install

13 acquisitions — HW

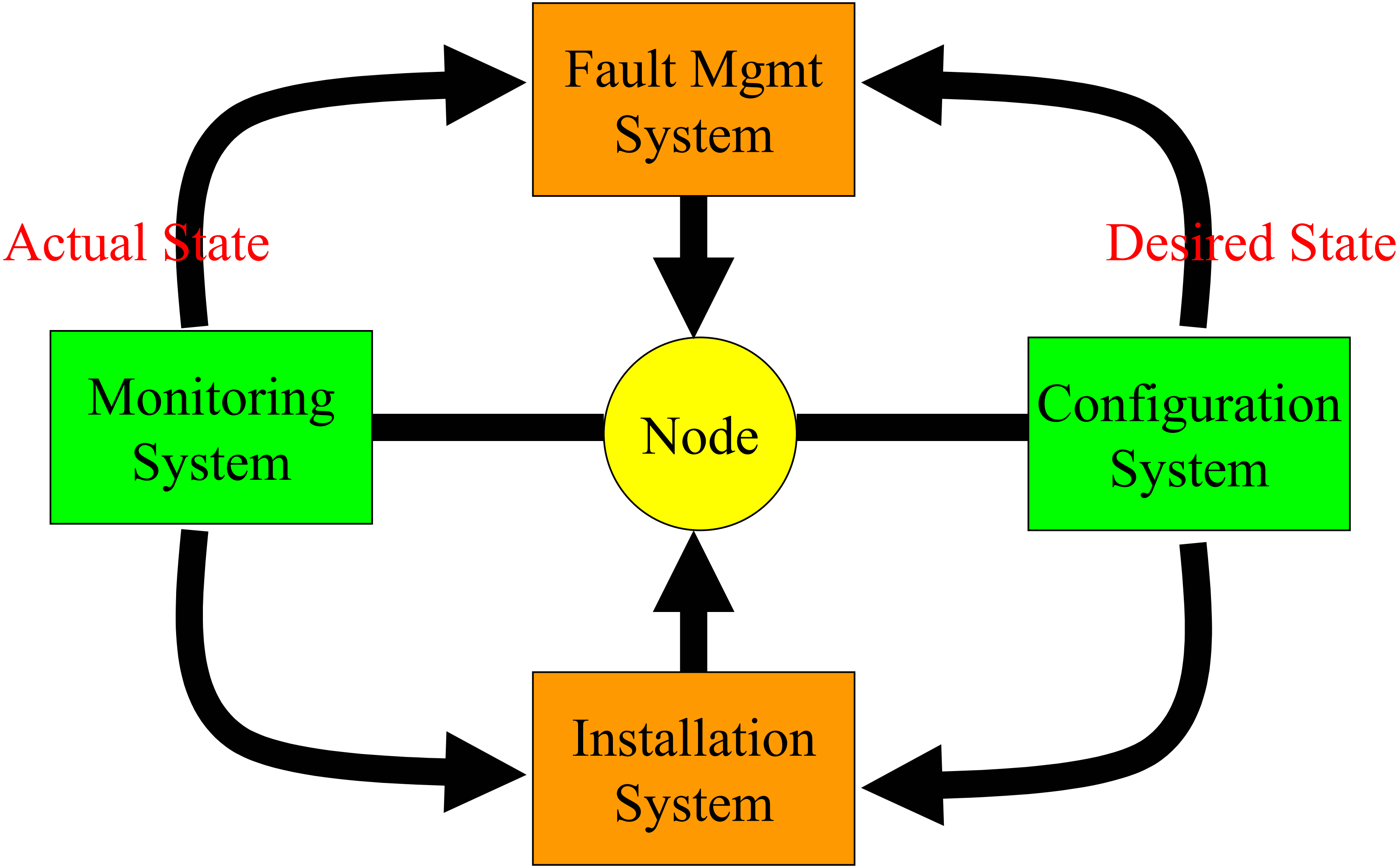
Complexity: Dynamics

- Volatile configurations

Fast	passwd files (every couple of hrs)
Med	Access lists
Med	SW security updates
Slow	OS upgrades

- Proliferation
  - Hardware Failures
- Asymptotic configuration changes
  - Node quiescence
  - Hardware down / at vendor
  - User community constraints

A Clean Restart



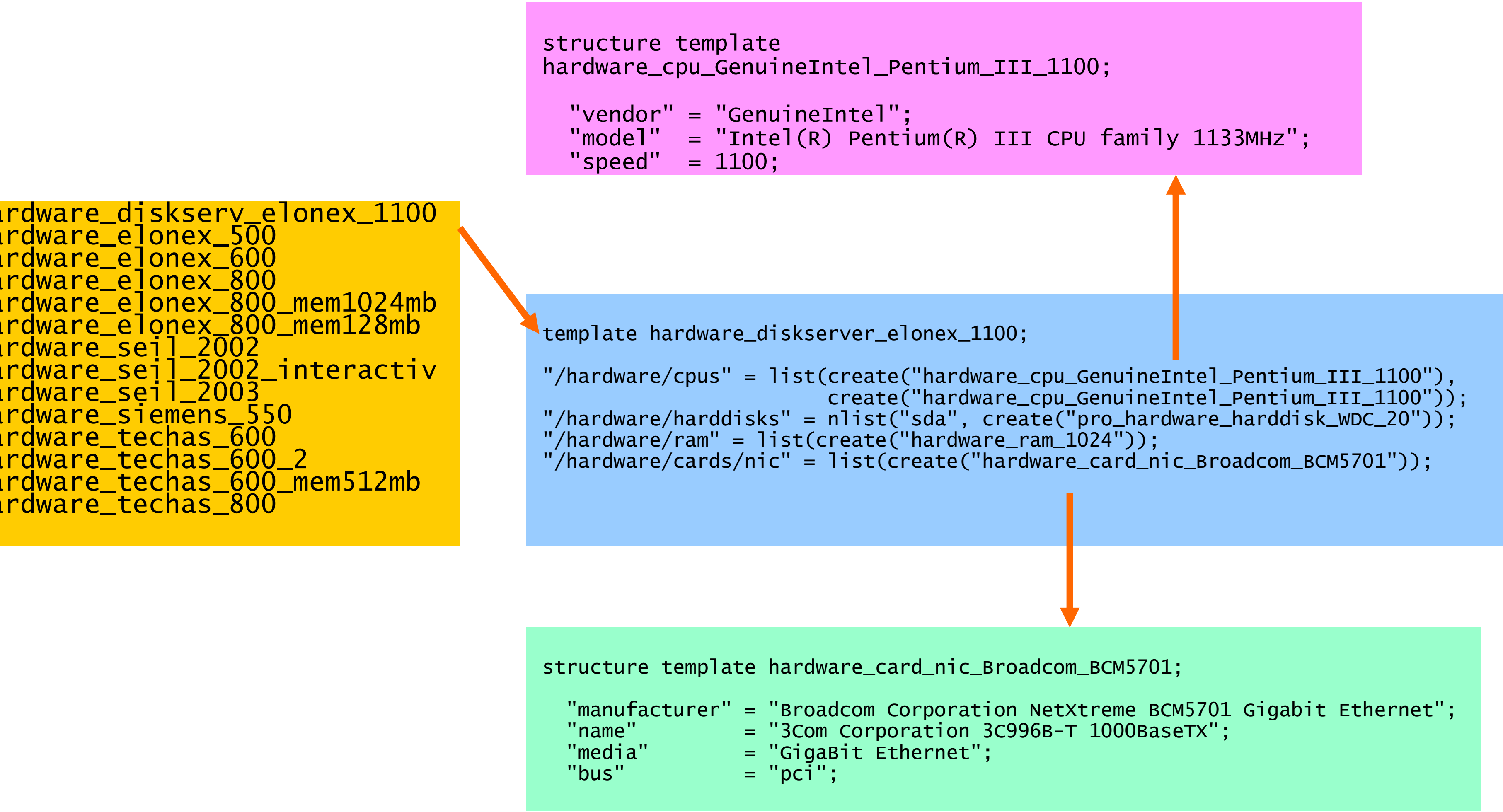
Framework Considerations

- Lightweight/modular/coupling/protocols/interfaces…
  - Decoupling
    - Local config files
    - Local programs do all work
  - Avoid inherent drift
    - No external crontabs or remote mgmt scripts
    - No unregistered application provider triggered updates
    - No reliance on mgmt tools for parallel cmd
  - Reproducible in time and space ☺
  - Staggered replacement of existing tools
- Scalable
  - Load balanced servers
  - Time smeared transactions
  - Pre-deployment caches
  - *Head-nodes ?*

Config/SW Considerations

- Hierarchical configuration specification
  - Graph rather than tree structure
  - Common properties set only once
- Node profiles
  - Complete specification in one XML file
  - Local cache
  - Transactions / Notifications
- Externally specified, Versioned: CVS repos.
- Clean Initial State
  - Linux Standards Base, RPM
- One tool to manage all SW: SPMA
  - System and application
- Update verification nodes + release cycle
- Procedures and Workflows

Hardware variety



```
rdware_diskserv_elonex_1100
rdware_elonex_500
rdware_elonex_600
rdware_elonex_800
rdware_elonex_800_mem1024mb
rdware_elonex_800_mem128mb
rdware_seil_2002
rdware_seil_2002_interactiv
rdware_seil_2003
rdware_siemens_550
rdware_techas_600
rdware_techas_600_2
rdware_techas_600_mem512mb
rdware_techas_800
```

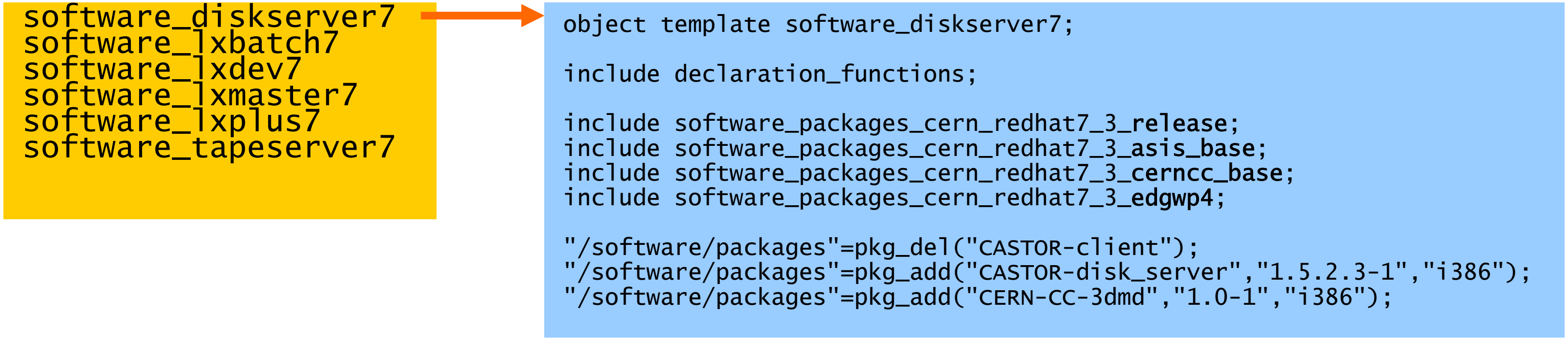
```
structure template
hardware_cpu_GenuineIntel_Pentium_III_1100;

  "vendor" = "GenuineIntel";
  "model"  = "Intel(R) Pentium(R) III CPU family 1133MHz";
  "speed"  = 1100;
```

```
template hardware_diskserver_elonex_1100;

"/hardware/cpus" = list(create("hardware_cpu_GenuineIntel_Pentium_III_1100"),
                        create("hardware_cpu_GenuineIntel_Pentium_III_1100"));
"/hardware/harddisks" = nlist("sda", create("pro_hardware_harddisk_WDC_20"));
"/hardware/ram" = list(create("hardware_ram_1024"));
"/hardware/cards/nic" = list(create("hardware_card_nic_Broadcom_BCM5701"));
```

```
structure template hardware_card_nic_Broadcom_BCM5701;

  "manufacturer" = "Broadcom Corporation NetXtreme BCM5701 Gigabit Ethernet";
  "name"         = "3Com Corporation 3C996B-T 1000BaseTX";
  "media"        = "GigaBit Ethernet";
  "bus"          = "pci";
```

Software variety

- CERN RedHat Linux 7.3.2
  - ~ 2400 packages declared in CDB



```
software_diskserver7
software_lxbatch7
software_lxdev7
software_lxmaster7
software_lxplus7
software_tapeserver7
```

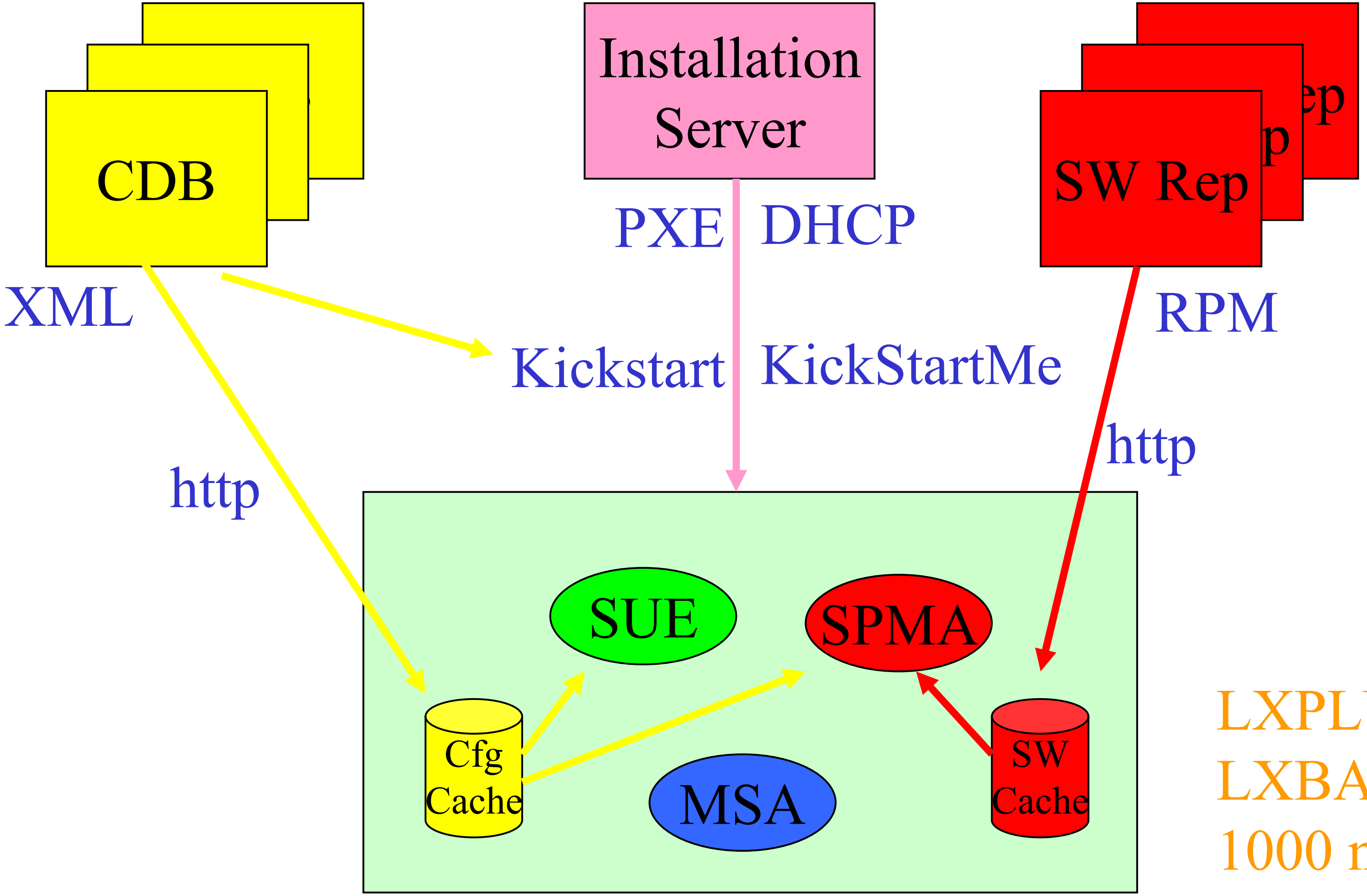
```
object template software_diskserver7;

include declaration_functions;

include software_packages_cern_redhat7_3_release;
include software_packages_cern_redhat7_3_asis_base;
include software_packages_cern_redhat7_3_cerncc_base;
include software_packages_cern_redhat7_3_edgwp4;

"/software/packages"=pkg_del("CASTOR-client");
"/software/packages"=pkg_add("CASTOR-disk_server","1.5.2.3-1","i386");
"/software/packages"=pkg_add("CERN-CC-3dmd","1.0-1","i386");
```

- RedHat Enterprise Server 2.1
  - ~ 1300 packages declared in CDB
- Diff subsets are selected for diff services
- Complete control over installed software

What is in CDB ?

- Hardware
  - CPU
  - Hard disk
  - Network card
  - Memory size
  - Location
- Software
  - Repository definitions
  - Service definitions = groups of packages (RPMs)
- System
  - Partition table
  - Cluster name and type
  - CCDB name
  - Site release
  - Load balancing information

Current Implementation



LXPLUS
LXBATCH
1000 nodes

Conclusions

- Maturity brings…
  - Degradation of initial state definition
    - HW + SW
  - Accumulation of innocuous temporary procedures
- Scale brings…
  - Marginal activities become full time
    - Many hands on the systems

- Combat with strong management automation