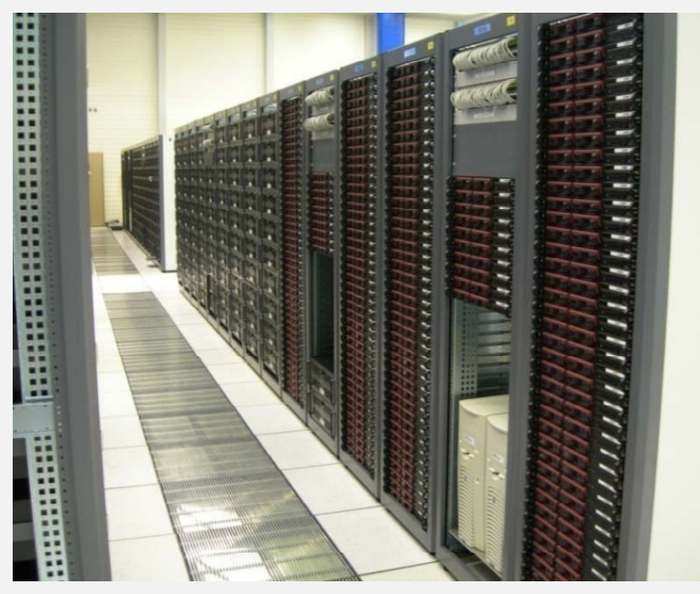
The Tier-0 Road to LHC Data Taking

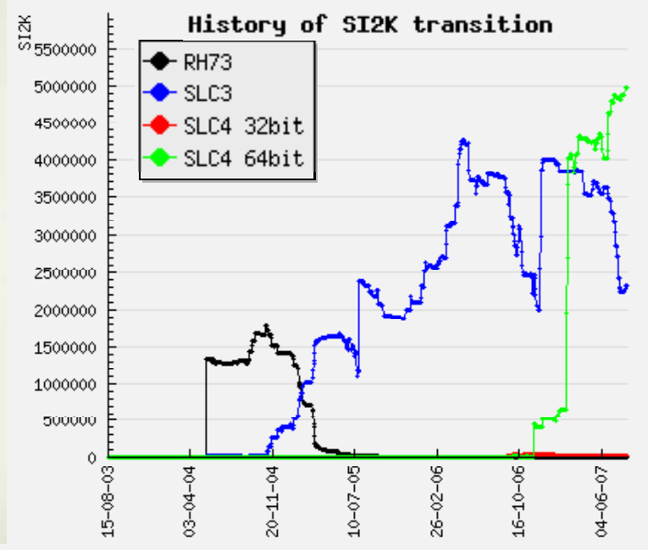
CERN IT Department

Commodity Hardware

except for

CPU Servers





Batch Farm
- 2,750 dual-cpu servers
- 7.5MSI2K capacity
- 150,000 jobs/week

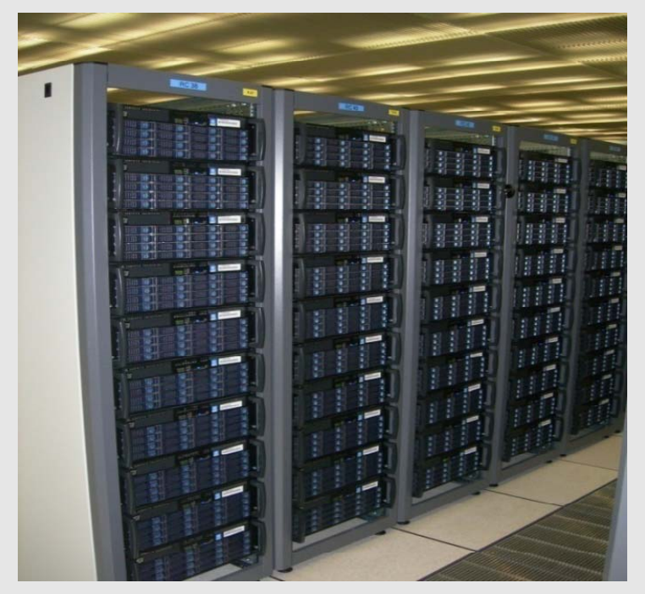
Other functions
- 1,600 dual-cpu servers (generally older hardware)

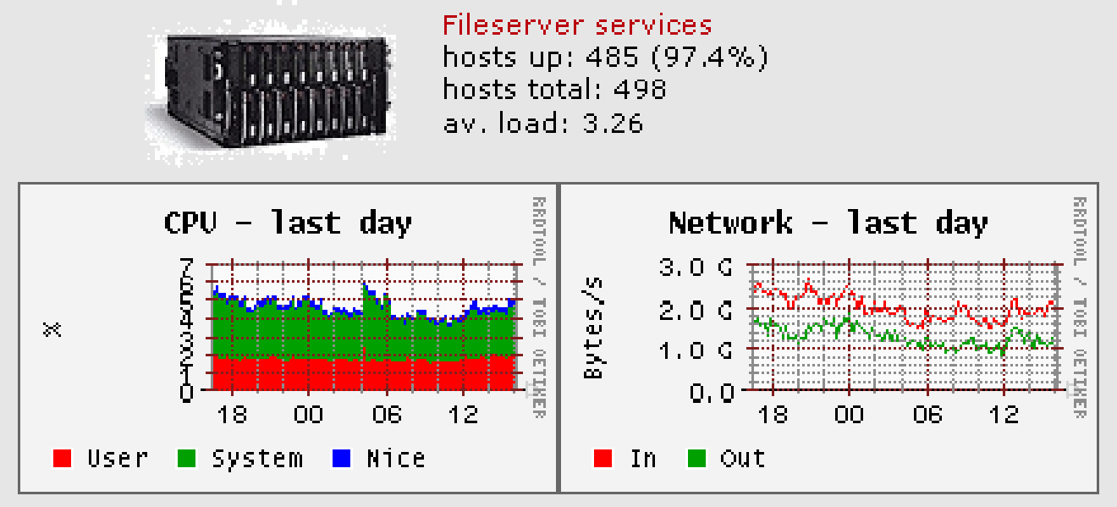
Tenders issued for kSI2K capacity to be measured using SPEC benchmark run as specified by CERN. “White box” suppliers offer best price/performance. On-site maintenance required (3 day response) and 3-year warranty.

No concerns about hardware reliability; the biggest problem is power consumption: 40 dual-motherboard servers in a 19” rack gives a heat load of 30kW/m^2!

Power consumption penalty now included in tenders but 1U boxes are still favoured over blade servers.

Disk Servers





- 700 servers (~500 in CASTOR2); 3PB total capacity

Tenders issued for disk capacity plus I/O performance requirements. “White box” suppliers offer best price/performance. On-site maintenance required (4 hour response) and 3-year warranty.

We *do* experience hardware problems but these are almost always disk-related and tier-1 vendors (IBM, HP, …) use the same disk suppliers!

Commodity approach provides greatest capacity with no significant performance or reliability penalty.

- ❖At least two suppliers chosen for each set of servers to be installed to reduce the risk that capacity upgrades are delayed due to problems with one supplier or one system component.
- ❖Rigorous “burn-in” tests prior to moving servers into production catch hardware problems early.
- ❖Given past experience with hardware problems, especially with disk servers, we favour ordering early to ensure capacity is delivered on schedule rather than delaying purchases as long as possible.

Network Fabric

Ethernet everywhere:
- 8 Force10 10GbE routers
- 300 10GbE ports
- 100 HP Switches
- 2.4Tb/s switching capacity

Using Ethernet standards enables provision of a high performance core network (2.4Tb/s switching capacity and designed for fully non-blocking performance) whilst keeping per-port costs low. Using Ethernet as the unique interconnection technology (as opposed to FiberChannel or Infiniband) avoids costly and inefficient translation gateways to connect to the rest of CERN, the experiments and the T1s.

However, we use switches with custom ASICS rather than true commodity devices as ability to cope with congestion is an important feature in the CERN environment; significant effort is required for testing during the hardware selection process.

Tape Drives & Robotics

3 Sun/STK SL8500 Libraries
- 23,200 slots
- 11.6PB capacity

50 Sun/STK T10K tape drives

3 IBM 3584 Libraries
- 8,650 slots
- 6PB capacity

60 IBM 3592 tape drives

CERN tape system costs are dominated by the media:
@ €100/600GB, 15PB/year => €2.5M/year!
LTO cartridge costs are lower, but media cannot be reformatted at higher density when new drives are introduced. IBM and STK media can be reformatted; e.g. next generation T10K drives will be able to write 1TB on current media. The costs saved by reusing media more than offset the higher costs of non-commodity drives.

Twin vendor solution preserves flexibility, e.g. if one vendor has a delay releasing a new drive.

Media repack to enable volume reuse is expected to be an ongoing operation; a dedicated CASTOR2 component has been created to facilitate this task.

Innovative Software

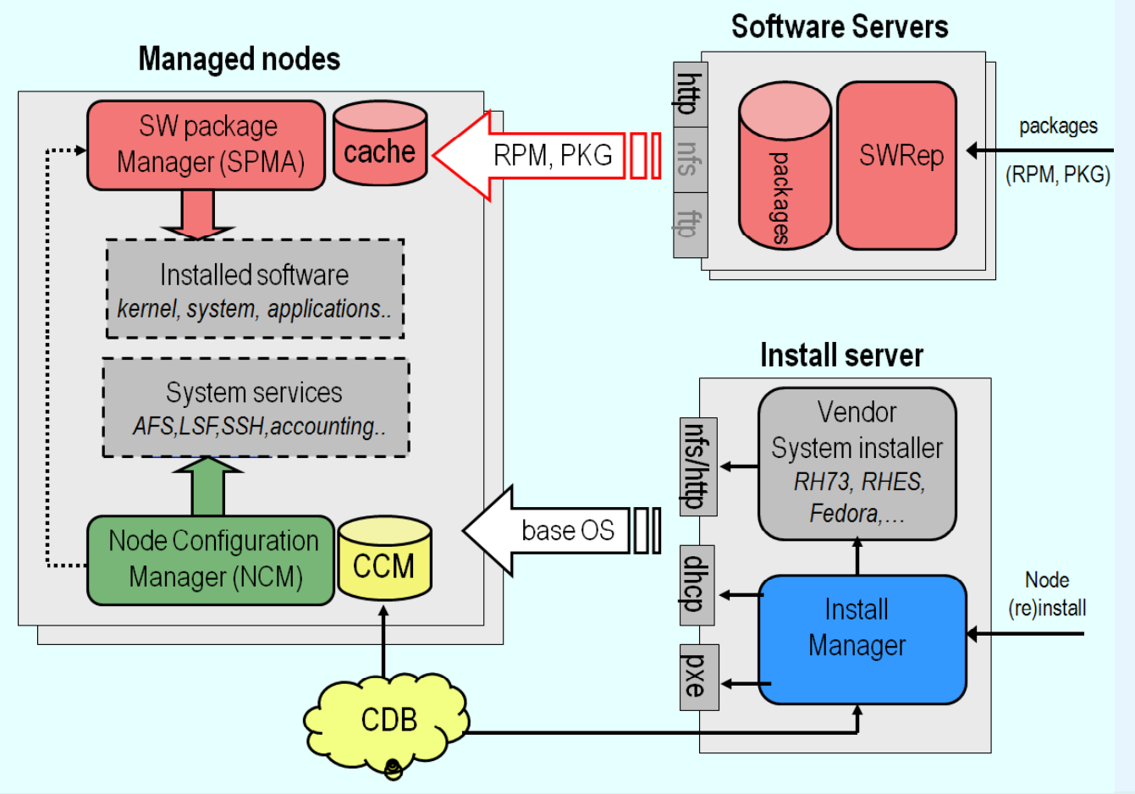
ELFms
Extremely Large Fabric management system

quattor

http://cern.ch/quattor

quattor is a system administration toolkit for the automated installation, configuration and management of clusters and farms.

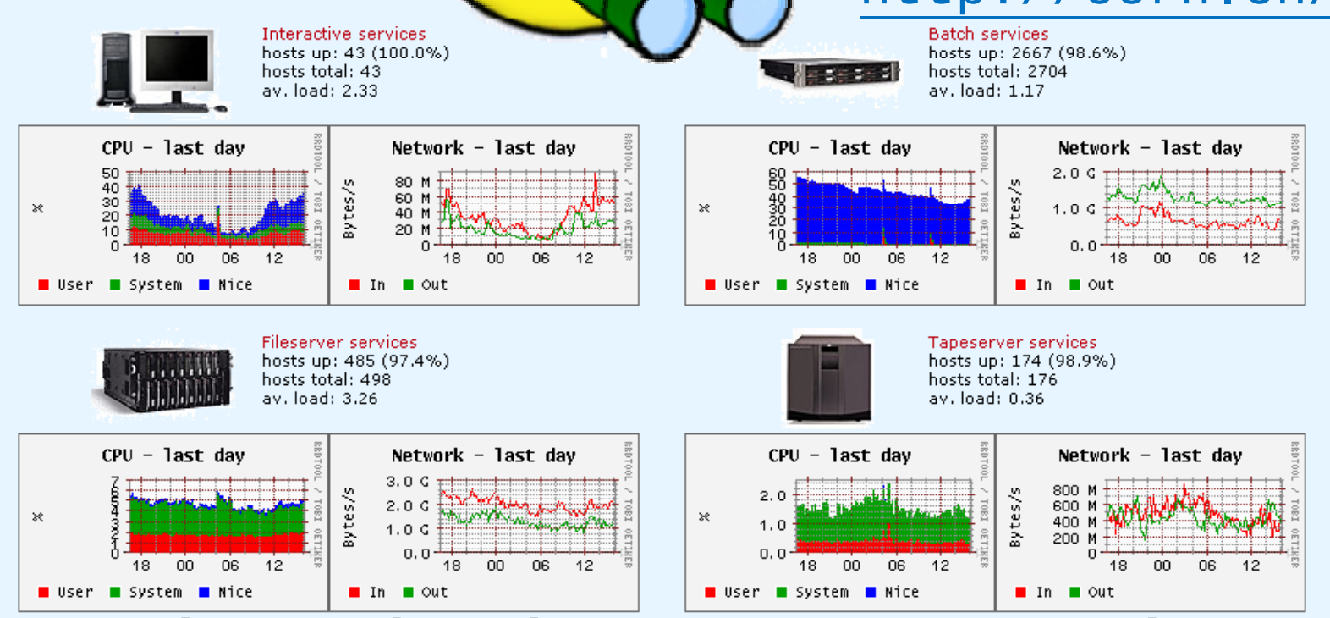
Key quattor components are
- CDB, the fabric wide configuration database, which holds the desired state for all nodes
- SPMA, the Software Package Management Agent, which handles local software installations using the system packager (e.g. RPM)
- NCM, the Node Configuration Manager, configures/reconfigures local system and grid services using a plug-in component framework.

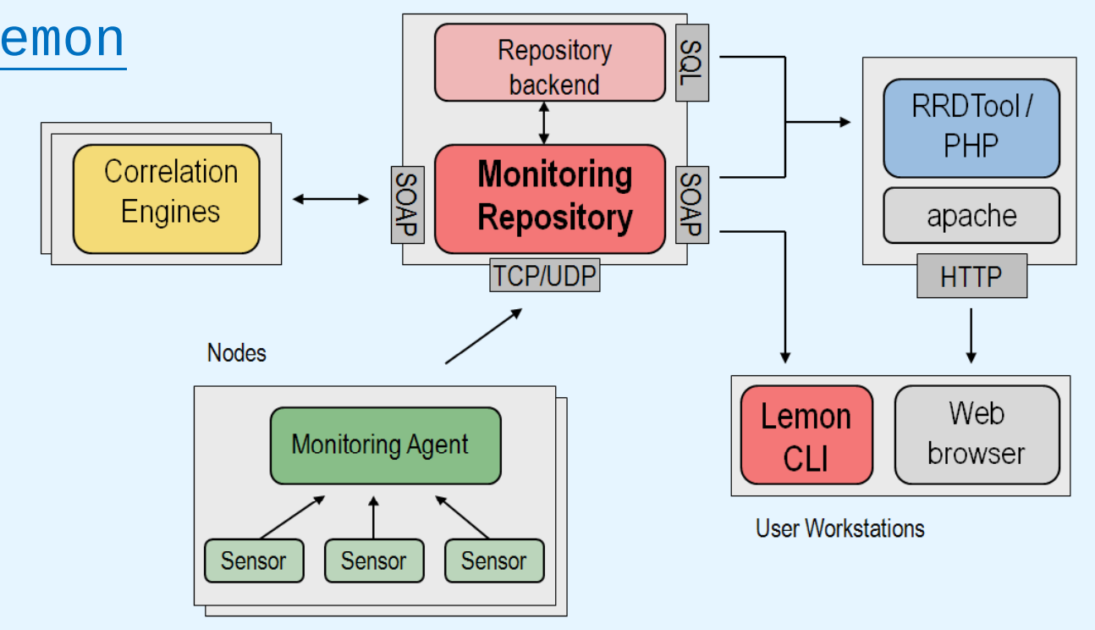


quattor development started in the scope of the EDG project (2001-2003). Development and maintenance is coordinated by CERN in collaboration with other partner institutes (including BARC, BEGrid, IN2P3/LAL, INFN/CNAF, NIKHEF, Trinity College Dublin, UAM Madrid and others)

Lemon

http://cern.ch/lemon





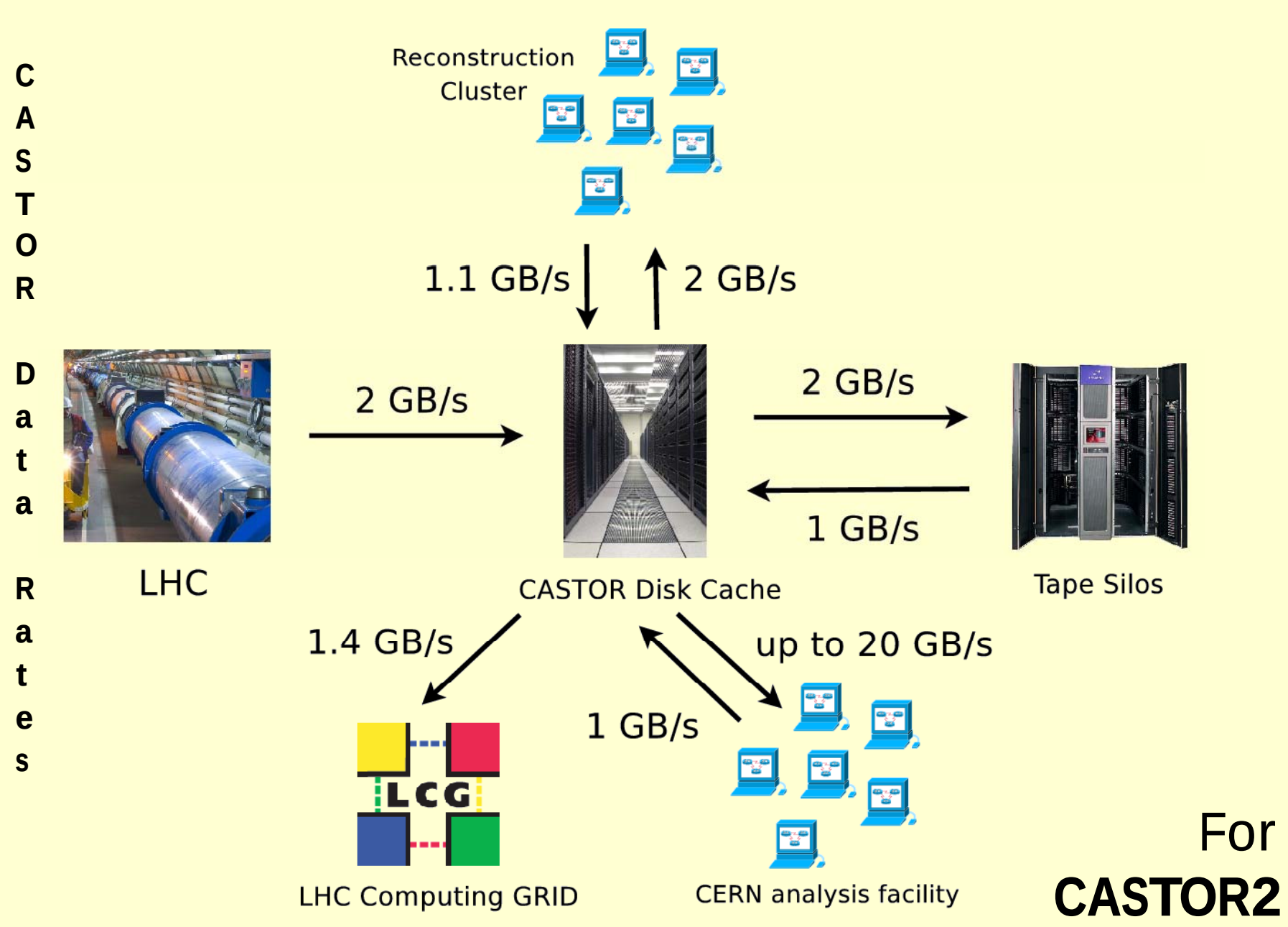
- Distributed monitoring system scalable to 10k+ nodes
- Provides active monitoring of
  - software and hardware
  - the Computer Center environment (power usage, temperature, …)
- Facilitates early error detection and problem prevention
- Provides persistent storage of the monitoring data
- Executes corrective actions and sends notifications
- Offers a framework for further creation of sensors for monitoring
- Most of the functionality is site independent

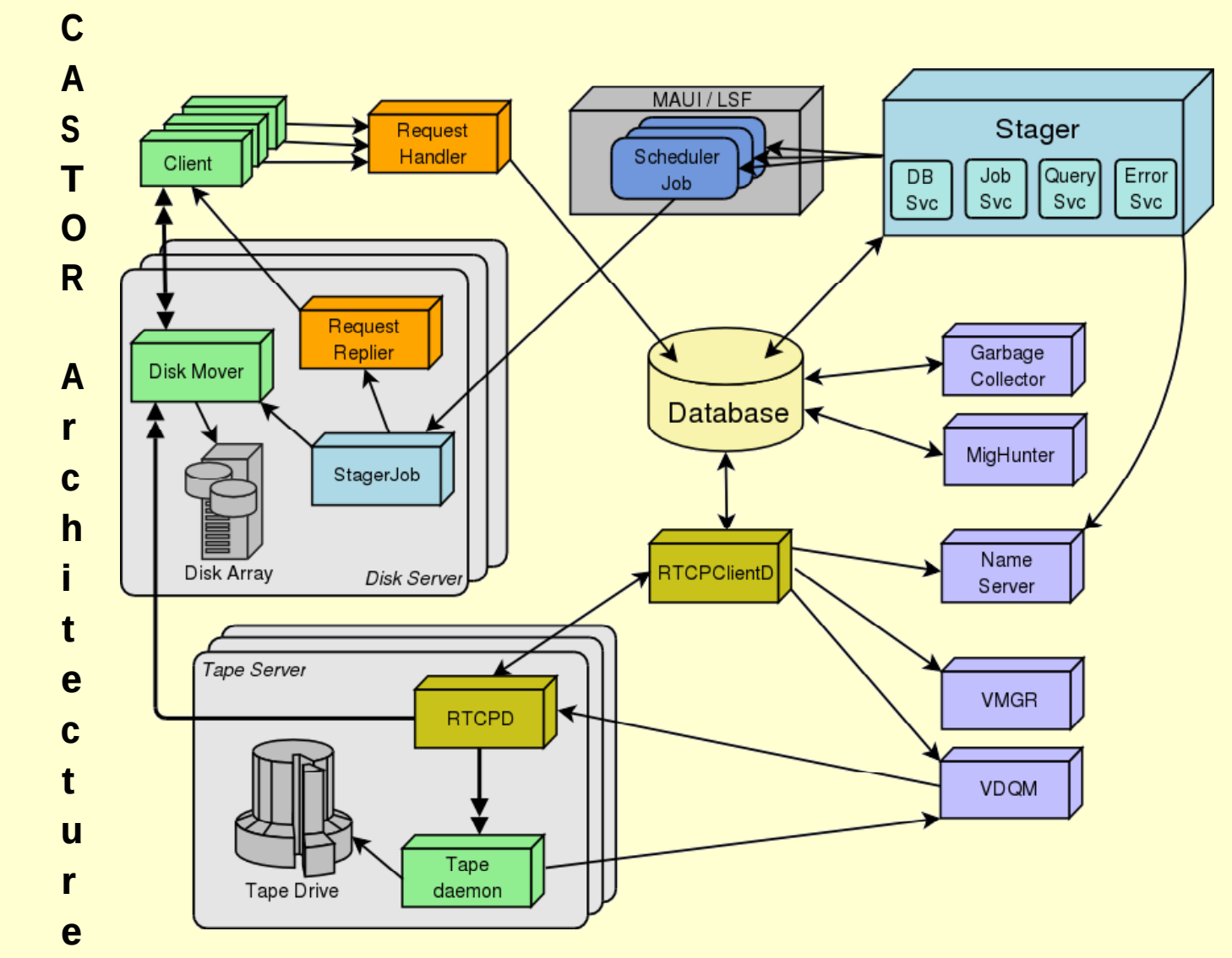
CASTOR

http://cern.ch/castor

Database centric architecture for increased reliability & performance
- core services isolated from user queries
- stateless daemons can be replicated for performance and restarted easily after (hardware) failure

Access request scheduling
- allows prioritisation of requests according to user role/privilege
- guarantees I/O performance for users and avoids h/w overload



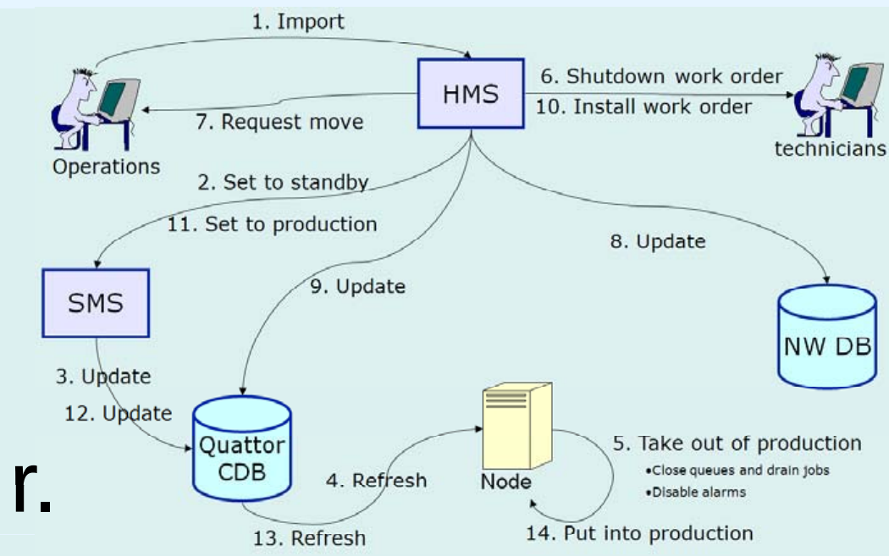


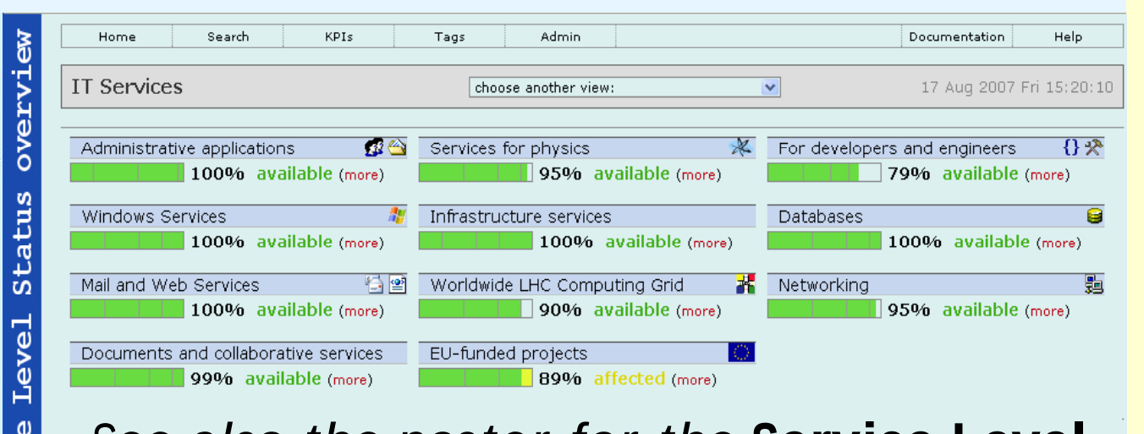
For more information, see contribution #62:
CASTOR2: Design and Development of a scalable architecture for a hierarchical storage system at CERN
(Monday @16:50, Carson hall B)

LEAF

http://cern.ch/leaf

Workflow tracking for box management (hardware and software):
- install/relocate systems and track vendor maintenance actions
- oversee software upgrade across 3,000+ node cluster
- Integrates monitoring, services and configuration database: e.g. on disk server failure, reconfigure CASTOR cluster in CDB and call vendor for repair.





IT Services — 17 Aug 2007 Fri 15:20:10

Administrative applications — 100% available (more)
Services for physics — 95% available (more)
For developers and engineers — 79% available (more)
Windows Services — 100% available (more)
Infrastructure services — 100% available (more)
Databases — 100% available (more)
Mail and Web Services — 100% available (more)
Worldwide LHC Computing Grid — 90% available (more)
Networking — 95% available (more)
Documents and collaborative services — 99% available (more)
EU-funded projects — 89% affected (more)

See also the poster for the **Service Level Status Overview** *(#91; this session)*

CERN - IT Department
CH-1211 Genève 23
Switzerland

www.cern.ch/it-fio