IMPLEMENTING HIGH AVAILABILITY WITH COTS COMPONENTS AND OPEN-SOURCE SOFTWARE

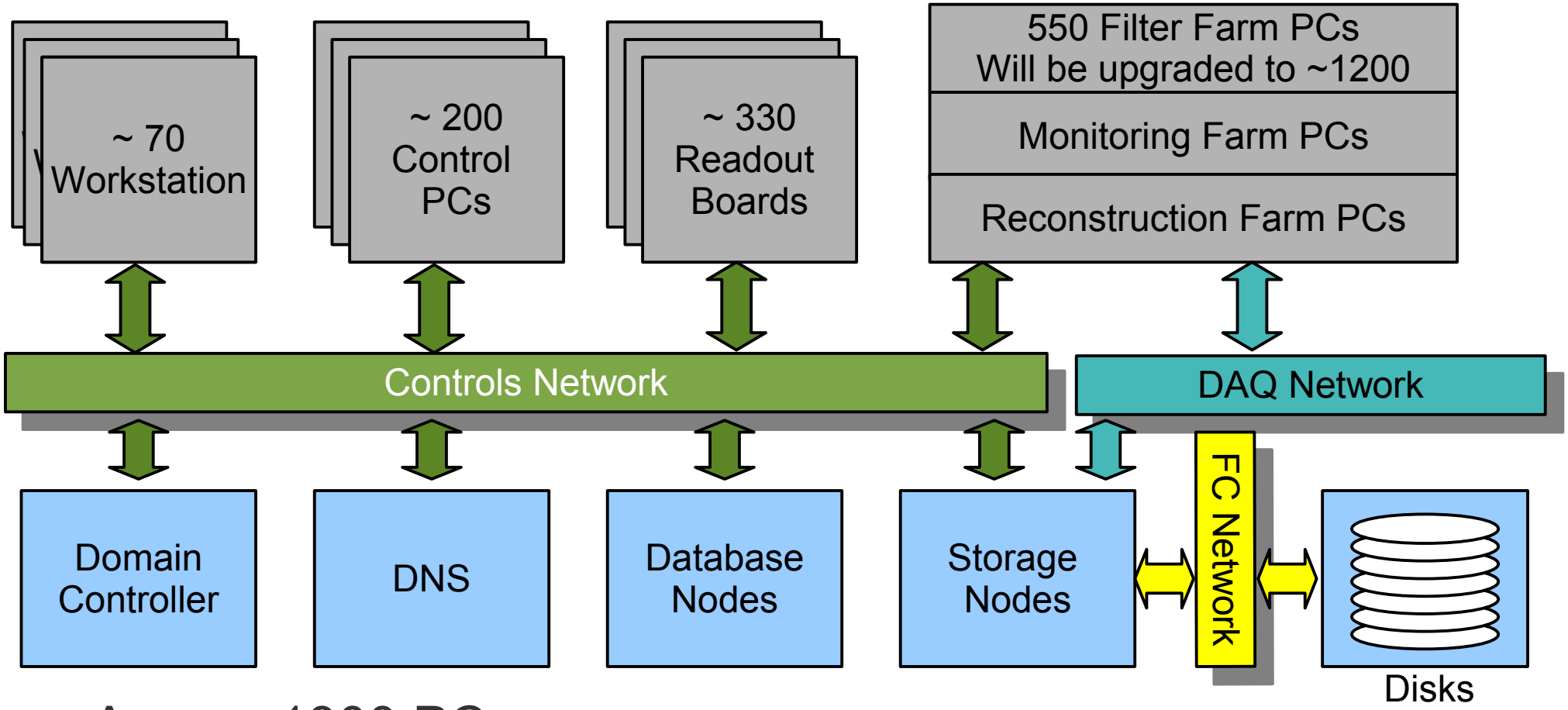
Rainer Schwemmer – CERN
On behalf of the LHCb Online Administrators

ICALEPCS 2009

Our motivation for HA

- Beam Time is VERY expensive. We have an obligation to use it as efficiently as possible.
- LHCb is a precision experiment. Need as much statistics as possible.
- Need to stay in control of the sensitive Hardware. Especially HV systems.
- Man power is expensive. People need to be able to work.
- Don't underestimate people's moral and their trust in the system.
- Upgrading/Fixing system components without taking it down

LHCb Online System



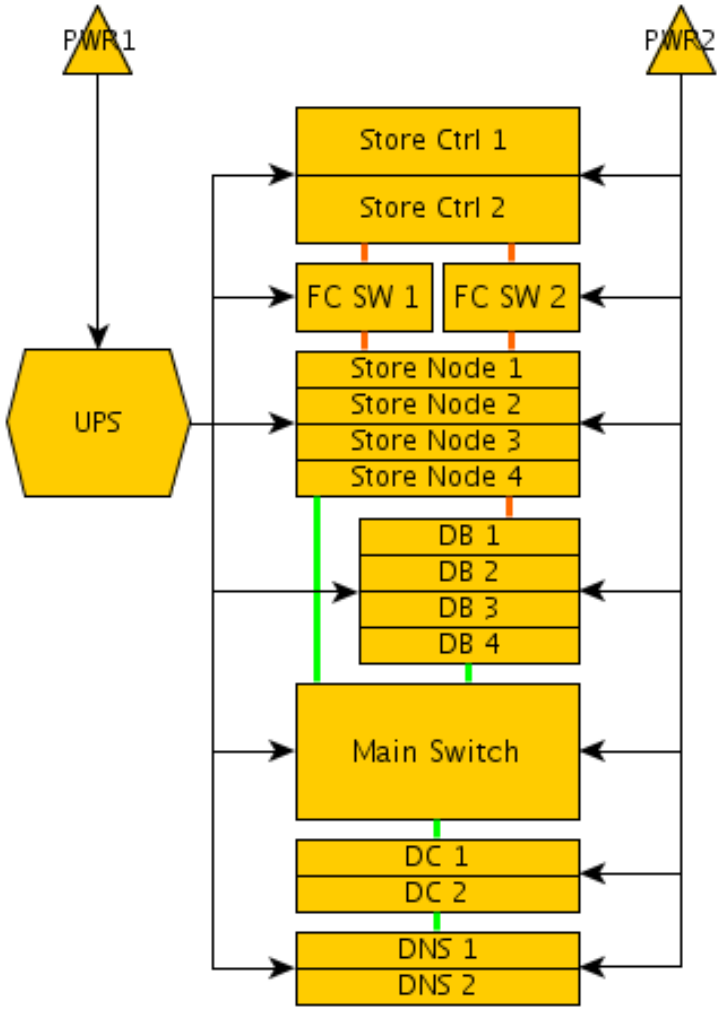
- Approx 1300 PCs
- Approx 400 Users

Services that need to be HA

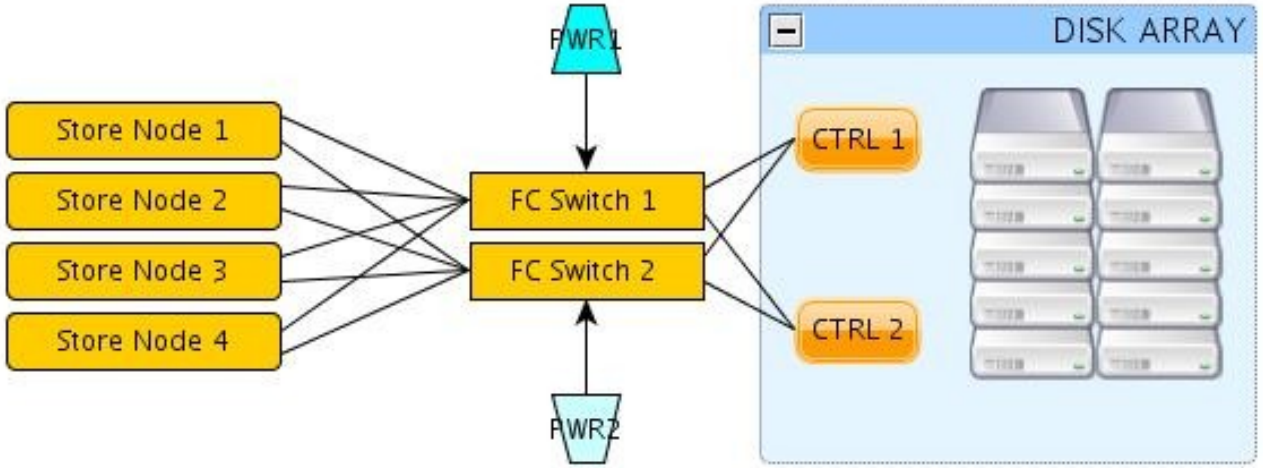
- Not all services are created equal
  - Can keep running without control of a bunch of readout boards.
  - System is completely useless without shared file systems
- Most critical services (Core System)
  - Databases
  - Domain Controller
  - Domain Name Service
  - Central File System services
  - Some experiment specific services (Event Writers, Data Movers, etc)
- DBs, DC and DNS come with their own HA scheme
- FS too (if you can afford/want to pay for it)

Core System Architecture (Power)

- Primary cause for failure:
  Unexpected Power Cuts ...
- ... Due to safety system mis-triggers
- Speedy recovery is about as important as fail prevention
- Unclean shut down of core system means:
  - Possible FS corruption
  - Several hours in best case
  - O(Days) in worst case
- UPS monitoring guarantees graceful shut down in case of long power failure.



Core System Architecture (SAN & FC Net)



- Banyan like FC network. Each component at least twice
- Cluster FS that can use multiple paths => Fully symmetric Active-Active system …
- …Almost: Tier (Raidset) of Disks is owned by certain controller. Access of Tier through wrong controller => performance hit
- Solution: Each FS has LUNs on at least two tiers. FS software writes to both tiers in parallel => no penalty any more
- Fail-over: Other controller takes ownership of Tier => NP

Heartbeat/Pacemaker 101

- Heartbeat => Detection and Execution Layer
- Pacemaker => Decision Layer
- Resources and Nodes
- Normal Linux services as HA resources
  - Works with standard Linux start/stop scripts.
    Better: use OCF compliant script
  - Stateles services are best
  - Stateful services need to store state on shared storage
- Not just limited to programs, also IP addresses, Disk, etc
- Fencing - STONITH (Shoot The Other Node In The Head)
  - Makes sure that presumably dead = dead
  - Protects resources that need exclusive access
  - IPMI in our case via separate network

Our HA services

- Interfaces to the Run Database
  => No new file names or run numbers without this service
- Data Movers and Book Keeping
  => Data should be moved to CASTOR asap
- Monitoring Daemon for the UPS
  => Emergency shut down in case of long power outage
- SNMP trap daemon
  => Critical messages forwarded to cell phones
- Several IP addresses
  =>Best to group specific service with IP address
  =>For NFS/Samba server

Active-Active NFS/Samba Server

- How it works:
  - N Servers with NFS/Samba server
  - 1 Virtual IP per NFS/Samba server instance
  - IPs are put into DNS round robin
  - In case of failure, IP is migrated to different node by Heartbeat; Clients reconnect immediately
- Advantages
  - Cheap
  - Don't need to rely on proprietary code
  - Scales well (up to a certain point)
- Disadvantages
  - Tech-support is tricky
  - File locking features will not really work
  - Write back caching is dangerous

NFS/Samba Pitfalls

- NFS:
  - All NFS server programs have to use same Ports on all Store Nodes (Mountd, lockd, etc)
  - Need to make sure, all files have identical NFS handle on all Store Nodes (Need Cluster FS, use fsid directive in /etc/exports
  - We are using NFS v3, because of bug in TCP connection fail-over (This is for SLC4, have not re-tested on SLC5)
- Samba:
  - Fail-over works out of the box without any further configuration of Samba
  - Running inside a domain with ADS authentication is tricky
    DC is too paranoid when fail-over happens and denies any further authentication
    => use method rpc instead of ads when joining domain

Conclusion

- Heartbeat has protected us from a lot of downtime.
- Fail-over is almost transparent to the user.
- Active-Active configuration has increased our total system performance significantly.
- One of the biggest advantages is that we can update/upgrade software/hardware while the system is running.

That's all folks

Thanks for listening

rainer.schwemmer@cern.ch