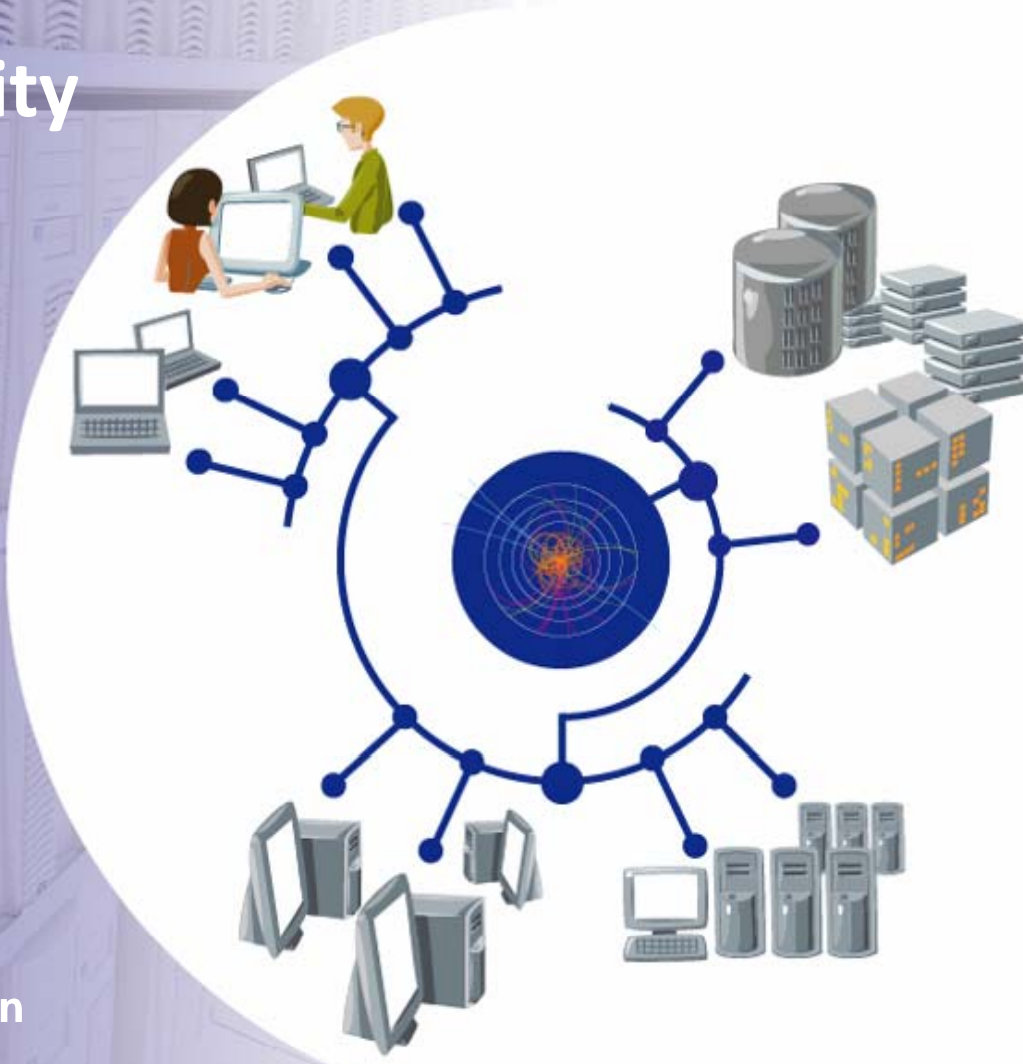




WLCG – Worldwide LHC Computing Grid

Service Reliability

Workshop Summary,
WLCG GDB, 05/12/07



Jamie Shiers
WLCG Service Coordination





Disclaimer

- I cannot cover a 1 week workshop in detail in 30'
- I will not cover every presentation – and those that I do cover will be at different levels of detail
- Just like “Esperanto in 3 months” it is not enough to “buy the book” and leave it on your bookshelf for 3 months



Agenda

- The goal was to understand how to build, deploy and operate robust and reliable services
- The driving force was the list of critical services as viewed by the LHC experiments – plus WLCG MoU
- **💣 Please note – there are constraints! Presenting a request is not a guarantee that it can be met! And there are conditions – “WLCG best practices”**
- Goals: measured improvement in service reliability (sessions at the [April 2008 Collaboration Workshop](#)); “solved” prior to CHEP 2009 (March 2009 in Prague)



Agenda - MB

- There was a high-level summary at yesterday's OB and there will be a more detailed summary at tomorrow's GDB
- I will just highlight some important points here...
 - **Only way to achieve requested level of resilience is by building fault tolerance into the services (including experiment-specific ones)**
- The techniques are simple and well tested
 - industry wide
- We have written a paper summarizing this – it is on the workshop agenda page!



Pros & Cons – Managed Services

☺ **Predictable service level and interventions; fewer interventions, lower stress level and more productivity, good match of expectations with reality, steady and measurable improvements in service quality, more time to work on the physics, more and better science, ...**

☹ **Stress, anger, frustration, burn-out, numerous unpredictable interventions, including additional corrective interventions, unpredictable service level, loss of service, less time to work on physics, less and worse science, loss and / or corruption of data, ...**



The workshop was about the 1st column



Overview

- Roughly 70 people registered – most sessions well attended (09:00 – 09:30 was a bit dead)
- Someone from all Tier1s except FNAL (who presented their strategy wrt reliable services at last [HEPiX](#))
- Loads of good and focussed discussion
 - In this respect (one of) the best WLCG workshops so far(?)
- Experiment participation somewhat patchy, particularly in the sessions on development
- (The tips and techniques apply to their services too!)
- ☺ **Much better (i.e. good) attendance at DB sessions!**
 - Organised together with WLCG 3D by Maria Girone IT-PSS



Main Themes

- Critical services; main deployment techniques; WLCG and experiment case studies
- WLCG operations: what is required to support LHC experiments (also in light of EGI)
- Monitoring: what is required to run reliable services – what is there, what is missing...
- **Robust services: middleware developers' techniques & tips**
- **Database developers' techniques & tips**



CMS Critical Services ([wiki](#))

Rank	Definition	Max. Downtime	Comments
11	CMS Stops Operating	0.5 hours	Not covered yet
10	CMS stops transferring data from Cessy		Cessy output buffer time
9	T0 Production stops		min(T0 input buffer/Cessy output buffer) or defined time to catch up
8	T1/T2 Production/analysis stops		
7	Services critical when needed but not needed all the time (currently includes documentation)	0.5	
6	A service monitoring or documenting a critical service	8	
5	CMS development stops if service unavailable	24	
4	CMS development at CERN stops if service unavailable		
... more ...			



ATLAS Critical Services (PDF)

Tier	Service	Criticality	Consequences of service interruption
0	Oracle database RAC (online, ATONR)	Very high	Possible loss of DCS, Run Control, and Luminosity Block data while running. Run start needs configuration data from the online database. Buffering possibilities being investigated.
0	DDM central services	Very high	No access to data catalogues for production or analysis. All activities stops.
0	Data transfer from Point1 to Castor	High	Short (<1 day): events buffered in SFO disks, backlog transferred as connection is resumed. Long (>1 day): loss of data.
...			
0-1	3D streaming	Moderate	No export of database data. Backlog can be transferred as [soon as] connections are resumed.
... more ...			



LHCb Critical Services ([CCRC08 wiki](#))

Service	Criticality
CERN VO boxes	10=critical=0.5h max downtime
CERN LFC service	10
VOMS proxy service	10
TO SE	7=serious=8h max downtime
T1 VO boxes	7
SE access from WN	7
FTS channel	7
WN misconfig	7
CE access	7
Conditions DB access	7
LHCb Bookkeeping service	7
Oracle streaming from CERN	7
... more ...	



ALICE critical services list

- WLCG WMS (hybrid mode OK)
 - LCG RB
 - gLite WMS (gLite VO-box suite a must)
- FTS for T0->T1 data replications
 - SRM v.2.2 @ T0+T1s
- CASTOR2 + xrootd @ T0
- MSS with xrootd (dCache, CASTOR2) @ T1
- PROOF@CAF @ T0



ATLAS Critical Services ([PDF](#))

Tier	Service	Criticality	Consequences of service interruption
0	Oracle database RAC (online, ATONR)	Very high	Possible loss of DCS, Run Control, and Luminosity Block data while running. Run start needs configuration data from the online database. Buffering possibilities being investigated.
0	DDM central services	Very high	No access to data catalogues for production or analysis. All activities stops.
0	Data transfer from Point1 to Castor	High	Short (<1 day): events buffered in SFO disks, backlog transferred as connection is resumed. Long (>1 day): loss of data.
...			
0-1	3D streaming	Moderate	No export of database data. Backlog can be transferred as [soon as] connections are resumed.

... more ...

CHEP 2007



CMS Critical Services ([wiki](#))

Rank	Definition	Max. Downtime	Comments
11	CMS Stops Operating	0.5 hours	Not covered yet
10	CMS stops transferring data from Cessy		Cessy output buffer time
9	T0 Production stops		min(T0 input buffer/Cessy output buffer) or defined time to catch up
8	T1/T2 Production/analysis stops		
7	Services critical when needed but not needed all the time (currently includes documentation)	0.5	
6	A service monitoring or documenting a critical service	8	
5	CMS development stops if service unavailable	24	
4	CMS development at CERN stops if service unavailable		

... more ...

CHEP 2007



ALICE critical services list

- WLCG WMS (hybrid mode OK)
 - LCGRB
 - gLite WMS (gLite VO-box suite a must)
- FTS for T0->T1 data replications
 - SRM v.2.2 @ T0+T1s
- CASTOR2 + xrootd @ T0
- MSS with xrootd (dCache, CASTOR2) @ T1
- PROOF@CAF @ T0

CHEP 2007



LHCb Critical Services ([CCRC08 wiki](#))

Service	Criticality
CERN VO boxes	10=critical=0.5h max downtime
CERN LFC service	10
VOMS proxy service	10
T0 SE	7=serious=8h max downtime
T1 VO boxes	7
SE access from WN	7
FTS channel	7
WN misconfig	7
CE access	7
Conditions DB access	7
LHCb Bookkeeping service	7
Oracle streaming from CERN	7

... more ...

CHEP 2007




Some First Observations

- Requirements are more stringent for Tier0 than for Tier1s than for Tier2s...
 - Some lower priority services also at Tier0...
- Maximum downtimes of 30' can only be met by robust services, extensive automation and carefully managed services
 - Humans cannot intervene on these timescales if anything beyond restart of daemons / reboot needed (automate...)
- ☛ **Interventions out of working hours are currently "best effort" - there is (so far) no agreement regarding on-call services (CERN)**
- Small number of discrepancies (1?):
 - ATLAS streaming to Tier1s classified as "Moderate" - backlog can be cleared when back online, whereas LHCb classify this as "Serious" - max 8 hours interruption
 - Also, ATLAS AMI database is hosted (exclusively?) at LPSC Grenoble and is rated as "high" (discussions re: IN2P3/CERN)
- Now need to work through all services and understand if "standards" are being followed and if necessary monitoring and alarms are setup...
- Do we have measurable criteria by which to judge all of these services? Do we have the tools? (Again < CCRC'08...)



Definitions of "Critical"

Experiment	Down	Seriously Degraded	Perturbed
ALICE	2 hours	8 hours	12 hours
ATLAS 	As text	As text	As text
CMS	30'	8 hours	24 hours (72)
LHCb	30'	8 hours	24 hours (72)

Quite significant differences in list of services under each heading:

- ATLAS: *only* 2 services are in top category (ATONR, DDM central services)
- CMS: (long) contains also numerous IT services (incl. phones, kerberos, ...)
- LHCb: CERN LFC, VO boxes, VOMS proxy service
- ALICE: CERN VO box, CASTOR + xrootd@T0 (?)



The Techniques

- ☺ DNS load balancing
- ☺ Oracle “Real Application Clusters”
 - H/A Linux (less recommended... because its not really H/A...)
 - 💣 **Murphy’s law of Grid Computing!**
 - Standard operations procedures:
 - Contact name(s); basic monitoring & alarms; procedures; hardware matching requirements;
 - **No free lunch! Work must be done right from the start (design) through to operations (much harder to retrofit...)**
 - Reliable services take less effort(!) to run than unreliable ones!
 - 💣 **At least one WLCG service (VOMS) middleware does not currently meet stated service availability requirements**
 - 💣 **Also, ‘flexibility’ not needed by this community has sometimes led to excessive complexity (complexity is the enemy of reliability) (WMS)**
 - **Need also to work through experiment services using a ‘service dashboard’ as was done for WLCG services (service map??)**

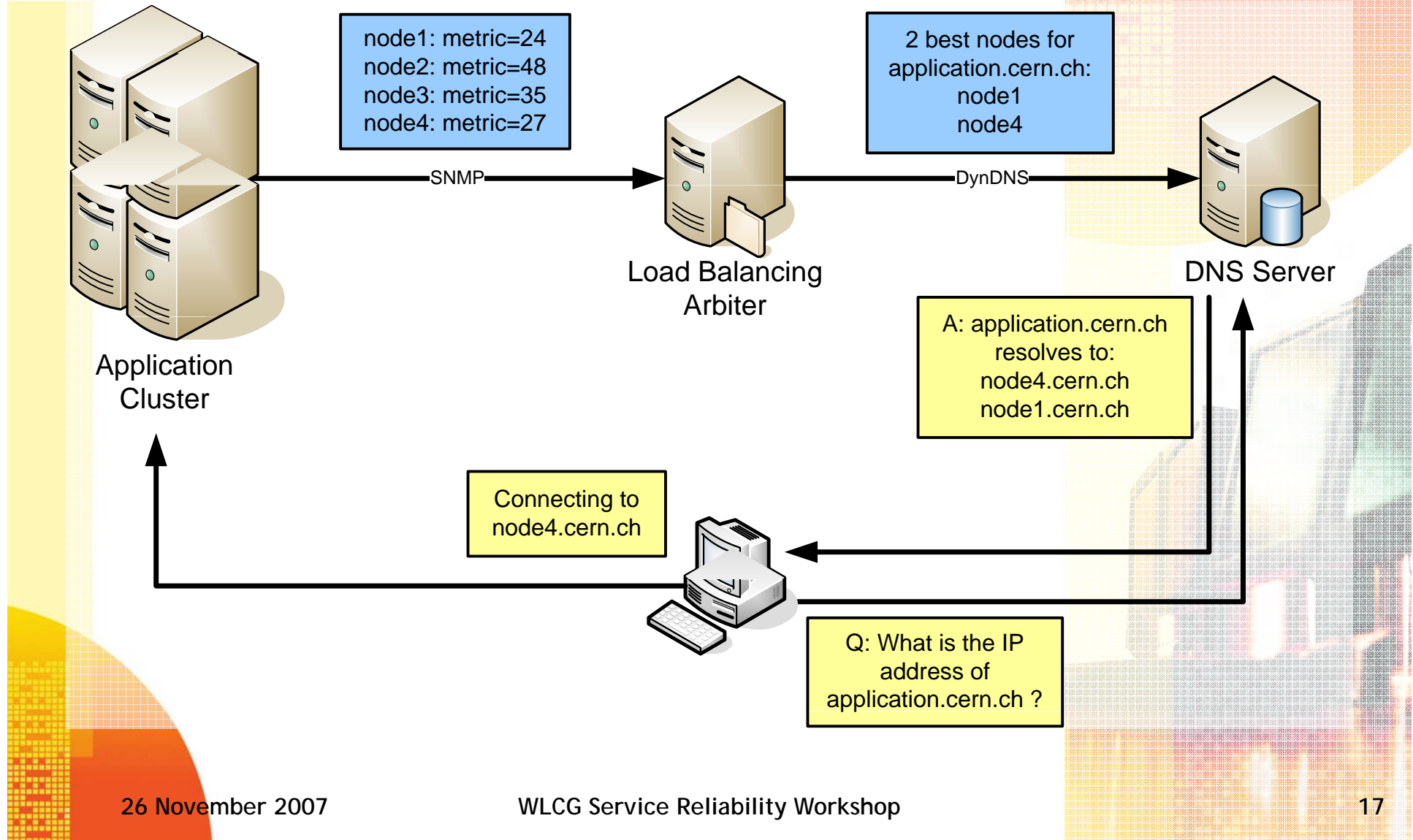
Domain Name System – ideal medium

- 😊 Ubiquitous, standardized and globally accessible database
- 😊 Connections to any service have to contact DNS first
- 😊 Provides a way for rapid updates
- 😊 Offers round robin load distribution (see later)

- 😞 Unaware of the applications
 - Need for an arbitration process to select best nodes
 - Decision process is not going to be affected by the load on the service

➤ **Application load balancing and failover**

Application Load Balancing System

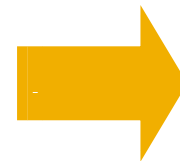


- **Advantage:**

- All LFC software upgrades are transparent for the users
- Except when database schema changes

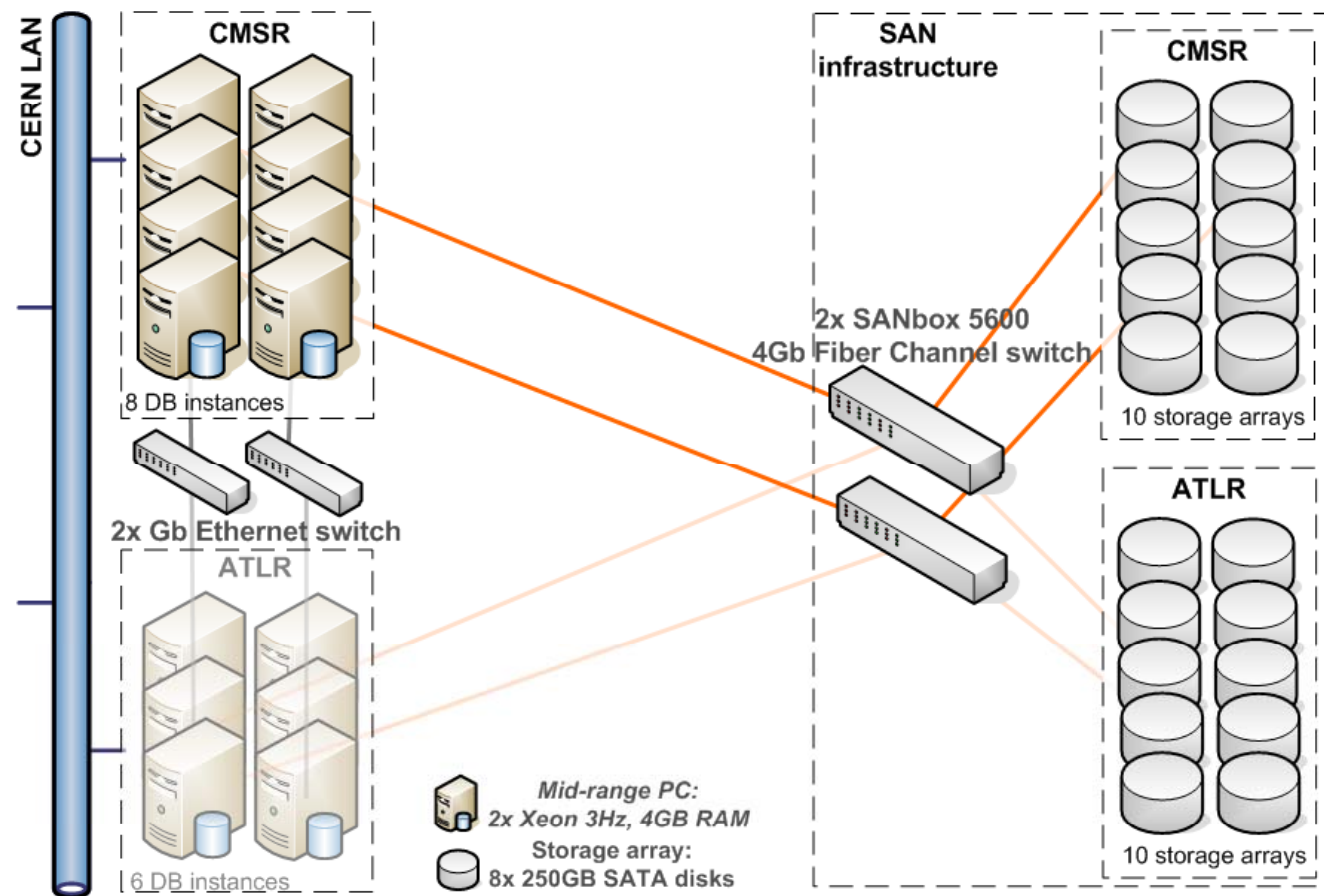
- **Ex: two DNS aliased nodes A and B**

- 1) Put node A in maintenance
 - Wait for node A to be taken out of production by dynamic DNS load balancing
- 2) Stop + upgrade + start LFC on node A
- 3) Take node A out of maintenance
 - Wait for node A to be put back into production by dynamic DNS load balancing
- 4) Start at step 1) with node B



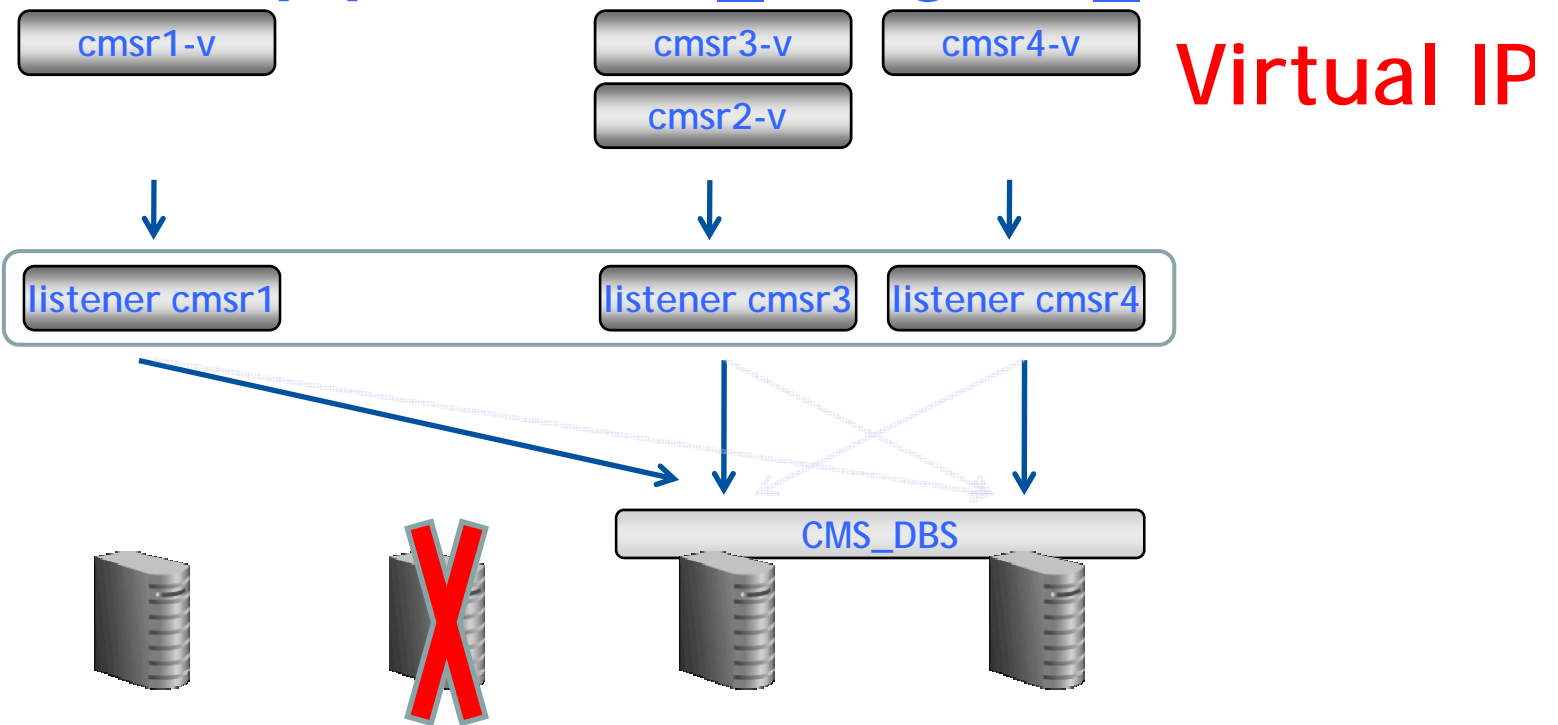
**Transparent
upgrade done!**

- Applications consolidated on large clusters, **per experiment**
- Redundant and homogeneous HW across each RAC



- Used also for rolling upgrades (patch applied node by node)
- 💣 **Small glitches might happen during VIP move**
 - no response / timeout / error
 - applications need to be ready for this → catch errors, retry, not hang

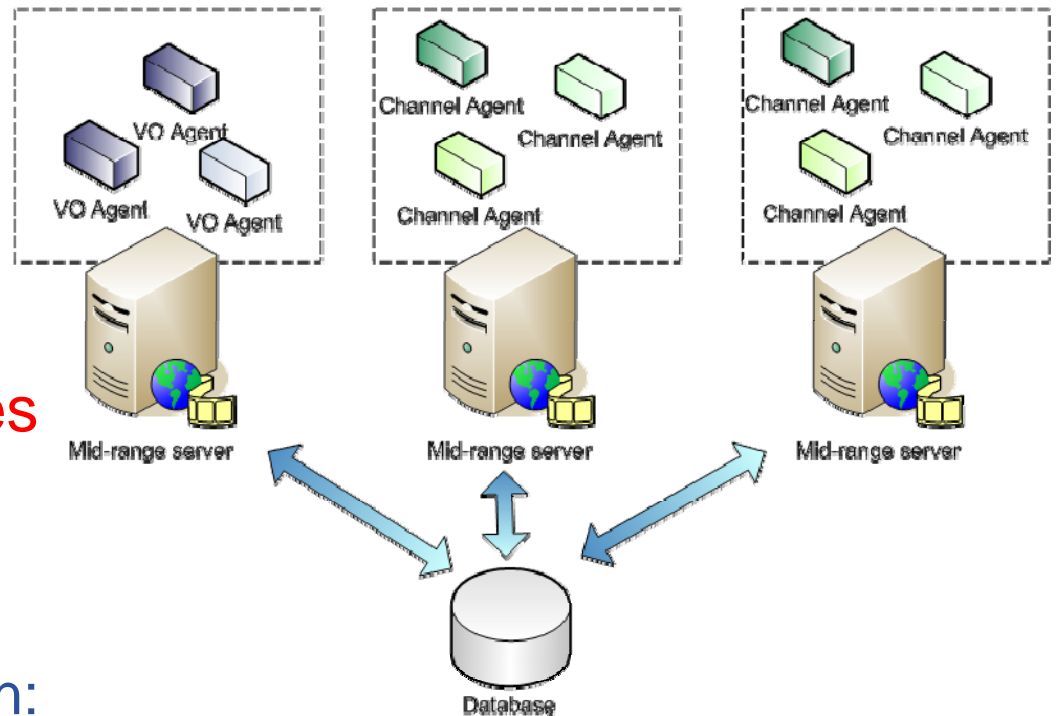
```
$ sqlplus cms_dbs@cms_dbs
```



- The configuration allows the agents daemons to be split arbitrarily over server nodes

– Shares load over multiple nodes

- Limit impact of outages
- One agent down: no impact on the others
 - One node goes down: no impact on channels running on other nodes



- **Procedures are the critical in operations**
- **Different types of procedure**
- **24/7 operator procedures (e.g.)**
 - These react to alarms and have
 - FTA_WRONG: detected one or more failures
 - Procedure: restart them using the service manager
 - Action: if the alarm doesn't go away, open a standard ticket
- **Service manager procedures**
 - Incident response – see previous one (WhatToDoWhen)
 - Planned procedures
 - Scheduled hardware moves, kernel upgrades
 - Procedures should always try to use software / deployment features to minimise the impact to the service

```

NO_CONTACT
SWAP_FULL
TMP_FULL
VAR_FULL
ROOT_FS_FULL

    • Open a standard ticket

GRID_BDII_WRONG

    • Log onto the node as root
    • Restart the BDII daemon: /sbin/service bdii restart
    • If the alarm does not clear in 10 minutes or if the restart fails, open a standard ticket.
    • If the alarm does clear OK, make a log-only entry.

TOMCAT_WRONG

    • Log onto the node as root
    • Restart the Tomcat daemon: /sbin/service tomcat5 restart
    • If the alarm does not clear in 10 minutes or if the restart fails, open a standard ticket.
    • If the alarm does clear OK, make a log-only entry.
    • Regardless of whether the alarm cleared or not, always send mail to grid-cern-prod-dms@cern.ch

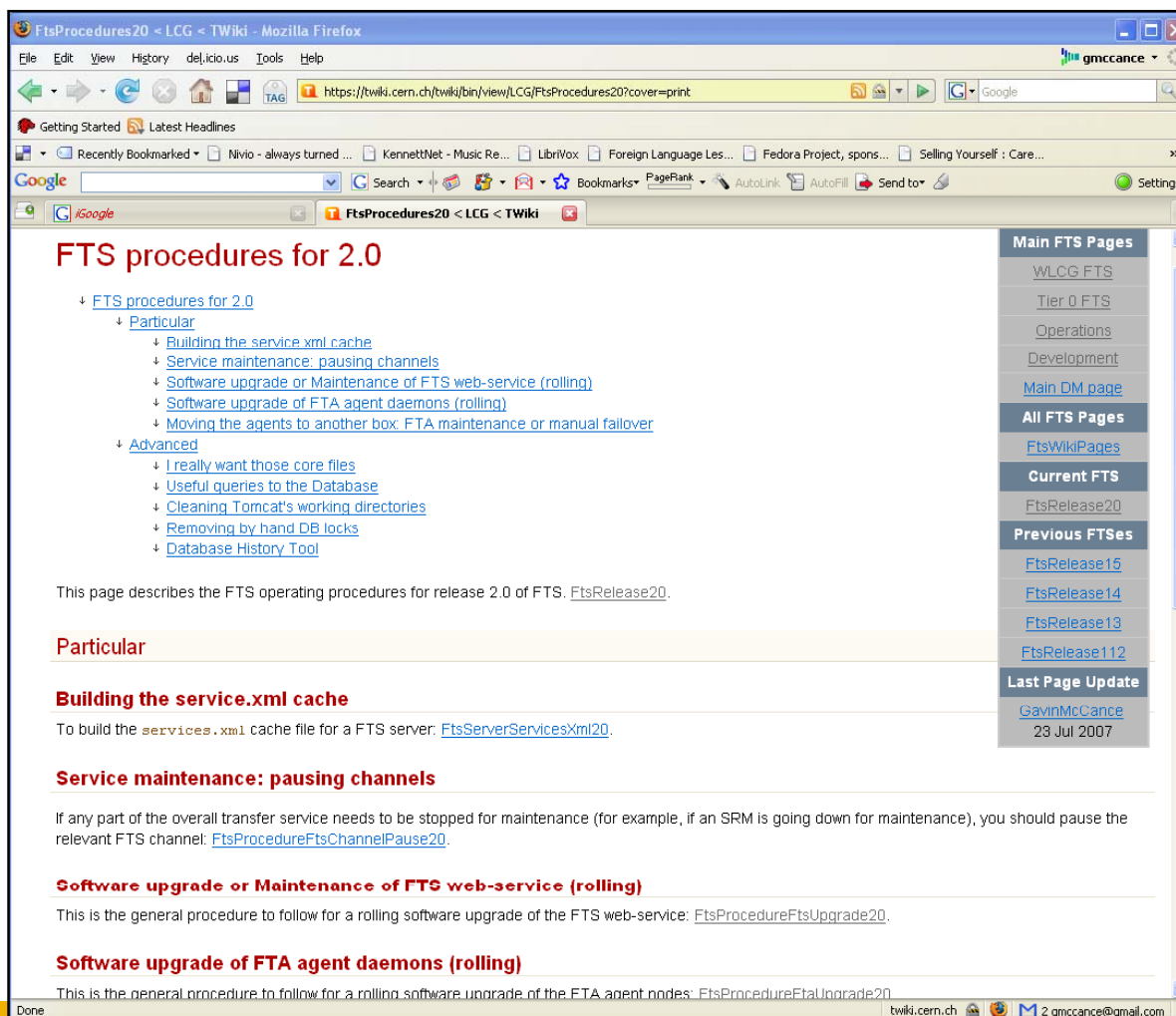
FTS_STUCK

    • Log onto the node as root
    • Restart the Tomcat daemon: /sbin/service tomcat5 restart
    • If the alarm does not clear in 10 minutes or if the restart fails, open a standard ticket.
    • If the alarm does clear OK, make a log-only entry.
    • Regardless of whether the alarm cleared or not, always send mail to grid-cern-prod-dms@cern.ch

FTA_WRONG

    • Log onto the node as root
    • The alarm monitors multiple daemons. Determine which one is wrong:
      o cat /var/lib/fts-agent-alarm/fts-agent.alarm
    • If the file is empty, and the alarm has not cleared, open a standard ticket.
    • Otherwise, you should see a line or lines like "glite-transfer-agent-name BAD" where
      "glite-transfer-agent-name" is the name of the daemon which is down.
    • For every BAD line listed, attempt to restart that daemon, using the full name of it as the instance. The
      exact name depends on the agent type, but will always be as it is printed in the BAD line. For example:
      o service transfer-agents start --instance
  
```

- Service manager procedures
- <https://twiki.cern.ch/twiki/bin/view/LCG/FtsProcedures20>



FTS procedures for 2.0

- ↓ [FTS procedures for 2.0](#)
 - ↓ [Particular](#)
 - ↓ [Building the service.xml cache](#)
 - ↓ [Service maintenance: pausing channels](#)
 - ↓ [Software upgrade or Maintenance of FTS web-service \(rolling\)](#)
 - ↓ [Software upgrade of FTA agent daemons \(rolling\)](#)
 - ↓ [Moving the agents to another box: FTA maintenance or manual failover](#)
 - ↓ [Advanced](#)
 - ↓ [I really want those core files](#)
 - ↓ [Useful queries to the Database](#)
 - ↓ [Cleaning Tomcat's working directories](#)
 - ↓ [Removing by hand DB locks](#)
 - ↓ [Database History Tool](#)

This page describes the FTS operating procedures for release 2.0 of FTS. [FtsRelease20](#).

Particular

Building the service.xml cache

To build the `services.xml` cache file for a FTS server: [FtsServerServicesXml20](#).

Service maintenance: pausing channels

If any part of the overall transfer service needs to be stopped for maintenance (for example, if an SRM is going down for maintenance), you should pause the relevant FTS channel: [FtsProcedureFtsChannelPause20](#).

Software upgrade or Maintenance of FTS web-service (rolling)

This is the general procedure to follow for a rolling software upgrade of the FTS web-service: [FtsProcedureFtsUpgrade20](#).

Software upgrade of FTA agent daemons (rolling)

This is the general procedure to follow for a rolling software upgrade of the FTA agent nodes: [FtsProcedureFtaUpgrade20](#).

Done twiki.cern.ch gmccance@gmail.com

- We always have
- Some examples
 - <https://twiki.cern.ch>
- Need to understand
 - Work out best possible
 - Know people manager, our
 - We appoint a
 - Include the a
 - Document the
 - e.g. [FTS] to your DE so you can

Upgrade of production tier-0 export to FTS 2.0

Scope:

- The production T1 export service and production T2<->T1 service
- The tier-2 production service will not be upgraded at this point.
- The pilot service is already running FTS 2.0.

Intervention announcement...

Scope:

- Production tier-0 export service
- Production tier-2 service

Preparation steps:

- Verify that the FTA agent actuator is disabled when the nodes are in maintenance. **VERIFIED**
- Only two CDB templates need updating `pro_system_gridfts` and `pro_type_gridfts_slc3`. These are now in `~straylen/fts-upgrade` level.
- The primary schema upgrade script is in the `transfer-fts` FTS 2.0 RPM: `/opt/glite/etc/glite-data-transfer-fts/schema/oracle/oracle-upgrade_2.2.1-3.0.0.sql`
- The history schema upgrade script is in `/afs/cern.ch/user/m/mccance/public/fts20-upgrade-intervention/fts_history`

Migration steps:

- Switch all channels to inactive. **DONE**
- Go to coffee while they drain currently running transfers. **DONE**
- Put all production nodes in maintenance. **DONE**
- There are three DBMS user jobs running: stop them (SQL*Plus on `lcg_fts_prod`):
 - `exec fts_stats.stop_hourly_job;` **DONE**
 - `exec fts_history.stop_job;` **DONE**
 - `exec fts_statcount.stop_job;` **DONE**
 - Verify that `select * from user_jobs;` returns no rows. **DONE**
- Stop the web-services (`fts101, fts114, fts115`). **DONE**
- Stop the agent daemons (`fts110, fts111, fts112, fts113`). **DONE**
- Stop the multitude of little scripts running on the FTS monitoring node (`fts102`). **DONE** Move to `/cron.d/`
- Ask DB team (contact Miguel Anjo) to copy the partial schema to the backup account. This should take around 20 minutes. **DONE**
- ... [upgrade software] **DONE**
- ... [upgrade CDB yaim configuration for FTS2.0]. Backup the old one. **DONE**
- BACKOUT 1
- Upgrade the main schema (this should take around 2 minutes) **DONE**
- Upgrade the history schema (this should take around 20 minutes) **DONE**
- Load the delegation schema (YAIM will insist anyway). **DONE**
- Run the writer account script to build new synonyms and make the appropriate grants: [FtsServer20WriterAccount](#). **DONE**
- BACKOUT 2

Cleanup:

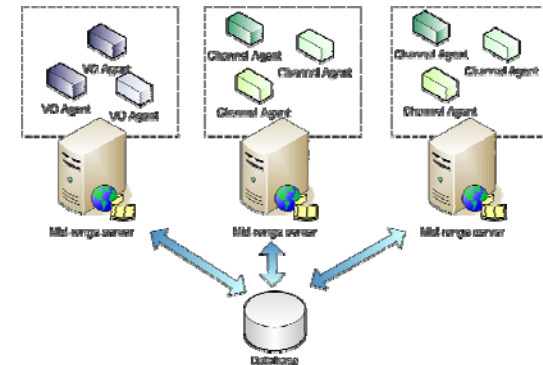
- Restart the web-services (`fts114, fts115`). **DONE**
 - Test a few commands. **DONE**
- Restart the agent daemons (`fts110, fts111, fts112, fts113`). **DONE**
- Restart the monitoring scripts on `fts102`.
- Re-enable jobs:

plete
ho

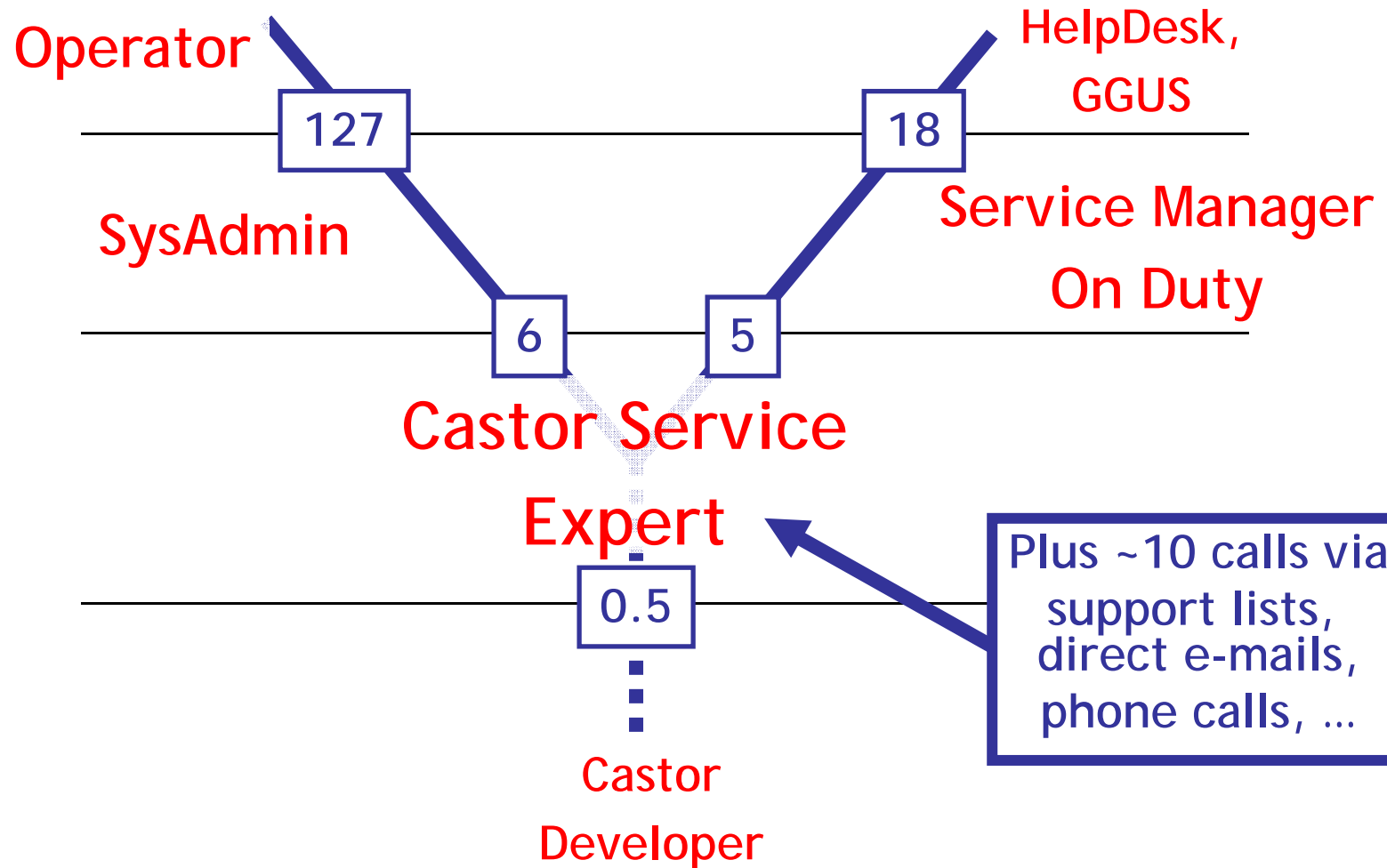
rice

Talk
de

- We're working on our plan for SL4 upgrade on CERN-PROD
 - [Pre-certification software being piloted (btw)]
- **The plan is to do this with zero user-visible service downtime and minimal downtime on the FTS channels**
- Schema upgrade shouldn't be needed (we already did it in FTS2.0)
- Pre-prepare (and test!) elfms Quattor configuration
- Take each FTS web-service node out of DNS in turn and upgrade it
 - No user-visible downtime to the service
- For each agent node, in turn:
 - Swap template, pause the channels on that node, take node down and rebuild it
 - The other nodes and channels will continue to run
 - The new node should come back up running
 - Restart channels
 - Each channel will experience a ~15 minute interruption



- Finally, we always try to follow-up any interventions and incidents
- We do this on CERN-PROD within our regular operations meetings
 - We look for things that could have gone better – there's usually something. e.g.
 - We sometimes pick up channel configuration problems
 - Forgotten 'workarounds' (aka hacks) that bit you during the upgrade
 - The (open) issue of schema fragmentation was found during one of our interventions because it badly affected it
 - We look for feedback to the developers
 - "It'd be so much better if you..."



VOBOX: the CERN-IT-FIO definition:

- A box dedicated to a VO, running one (or more) VO service(s)
- IT-FIO “VOBOX Service” handles:
 - Choice of hardware according to user specifications
 - Base OS installation & software upgrades
 - Hardware monitoring & maintenance
 - Installation & monitoring of common services
 - Eg: apache
 - SLA document in preparation
- User-specific Service installation & configuration managed by the VO
 - in compliance with the SLA

- CMS “Cessy->T0 transfer system”: Criticality level '10' (lxgate39)
 - ☹ Importance = “45” → **NO** Piquet Call if needed
 - 🚫 Only **ONE** machine
 - ? Monitoring (xrootd monitored by LEMON)
- CMS considerations
 - *machine essential for us, somehow part of the online system*
 - 🚫 **software can't be load-balanced**
 - **why? What if the machine breaks? Would a spare and test machine be useful ?**
 - *once real data operations start, machine needs to be up whenever there is detector activity (beam, cosmics, calibration).*
 - *We have buffer spaces to bridge downtime of component and machines and there are provisions to shutdown and restart our software.*
 - *But we design for steady-state operations and everything that gets us out of steady-state is a very big deal as it causes ripple effects through the rest of the system.*

Inter-site issues

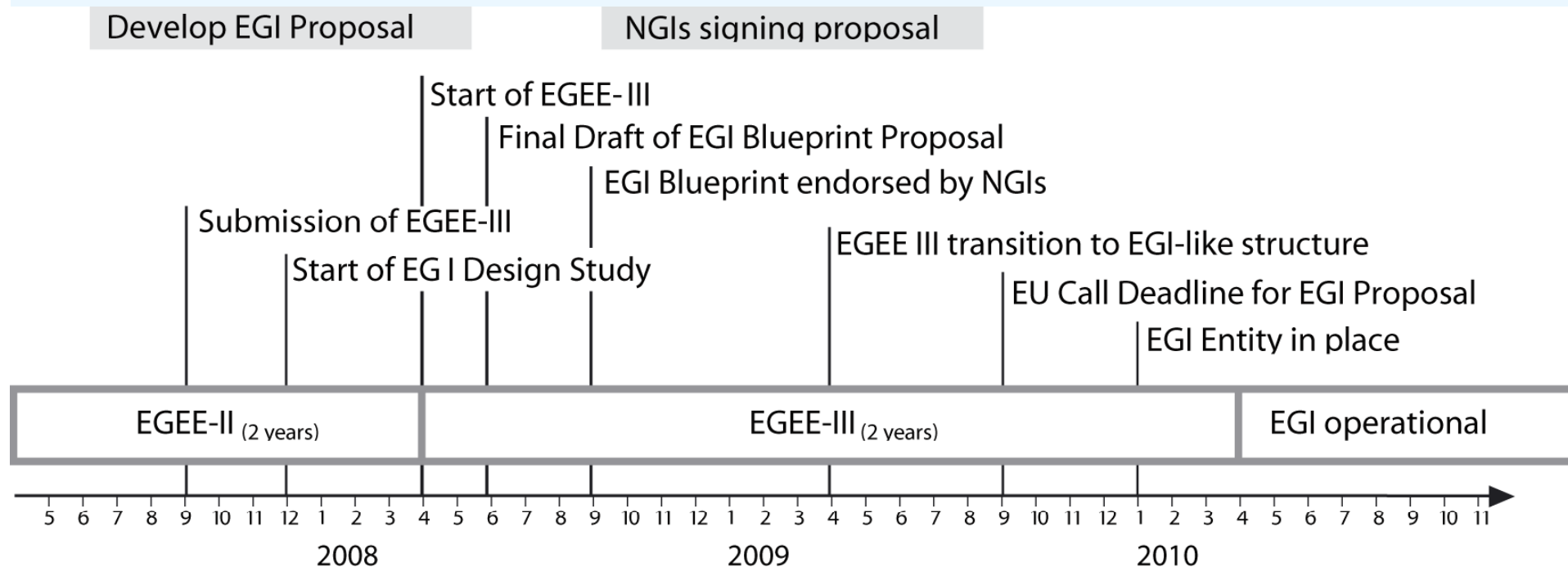
- Issue involves more than one site
- WAN network problem
- Service problems (file transfer, db synchronization...)
 - Performance bottlenecks
 - Errors
 - Overload
- Ownership: unknown/none
- Motivation to fix: underused or wasted resources
- Contract: none
- Workaround: avoid sites?



What is Grid Operations? Well...

- Infrastructures
 - Production service
 - Pre-production service (PPS)
- Processes
 - Middleware release process
 - Site registration
 - VO registration
- Communications
 - Weekly, monthly, bi-annual meetings for all stakeholders
- Interoperations with other grids (OSG)
- Grid security
- User + Operations support
- Operations tools
 - CIC Portal
 - Broadcast tool
 - VO ID cards
 - GOC database
 - Monitoring
 - Trouble ticketing system (GGUS)
- ... among other things!

A schedule



- **March 2008 (M7) 13-14 in Rome:**
2nd EGI Workshop (Responsible partner: INFN)
- *Before that at CERN*
 - WP5 on Jan.11
 - WP3 on Jan 29-31

Defining a first function schema



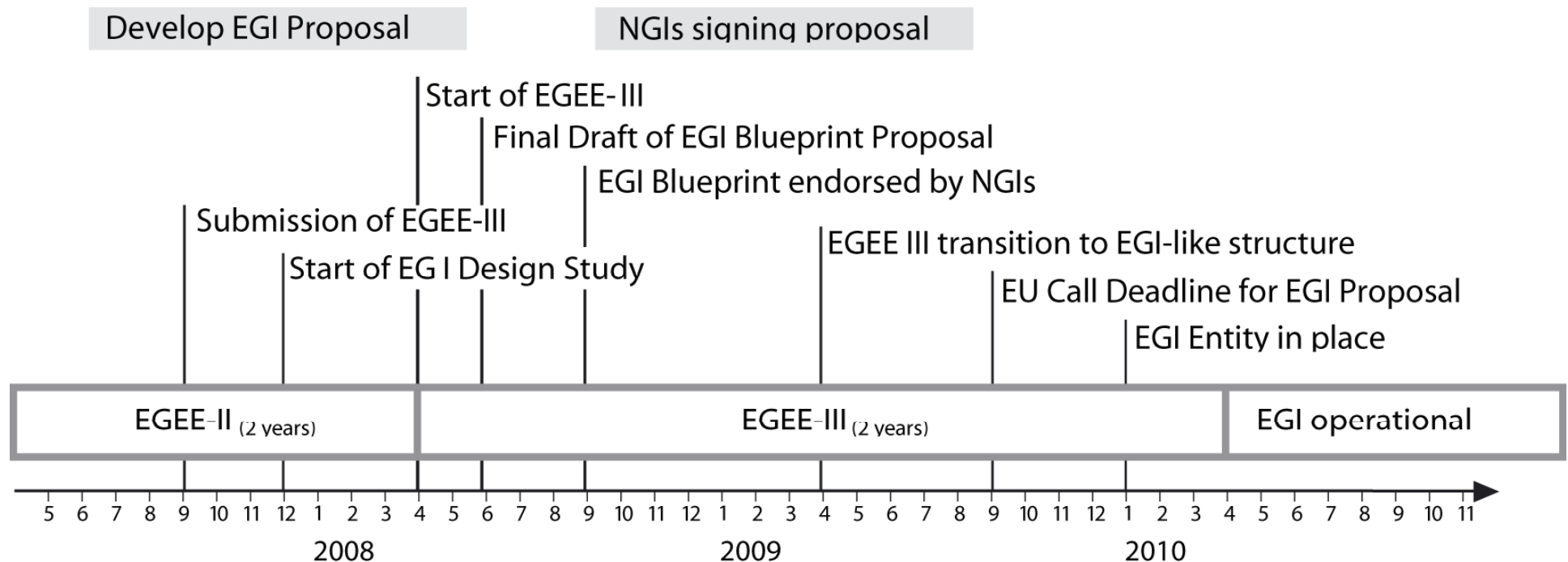
1. Operation of a reliable Grid infrastructure **CERN**
2. Coordination of middleware development and standardization **INFN**
3. Development and operation of build and test systems **CERN**
4. Components selection, validation, integration and deployment **CERN**
5. Mechanisms for resource provisioning to Virtual Organisations **GRNET**
6. Application support **CERN**
7. Training efforts **STFC**
8. Outreach and dissemination **INFN**
9. Industry take-up **INFN**
10. Contribution to the Open Grid Forum (OGF) and other standardisation bodies **INFN**
11. Policy, Strategy, e-IRG **STFC**
12. Representation of European Grid efforts, international cooperation, and ESFRI **GRNET**
13. Security **STFC**
14. Management **DFN**

EGI_DS Timeline

- In 2010, the LHC will reach design luminosity
- In 2010, EGEE III will terminate
- It is inconceivable that we:
 - a. Don't run the LHC machine
 - b. Run the LHC machine without a computing infrastructure (Grid)
 - c. Run the computing infrastructure without Grid operations
- **This is required for other mission critical applications that are dependant on this infrastructure**
- The transition to the new scenario must be
 - a. On time
 - b. Non-disruptive
- **This is a fundamental requirement – it is not an issue for discussion**

From the DoW...

- The establishment of EGI is guided by two basic principles:
 - 1. Build on the experience and successful operation of EGEE and related projects**
 - 2. Make EGI operational before EGEE III ends**

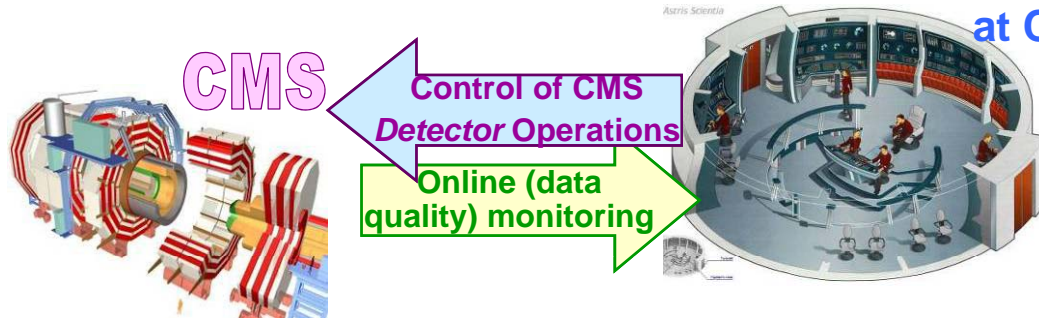


CMS Centres

❖ CMS Control Room

- Slow control, safety
- Operates detectors
 - Calibrations
 - Data-quality monitoring
- Data acquisition
- Data transfer to Tier0

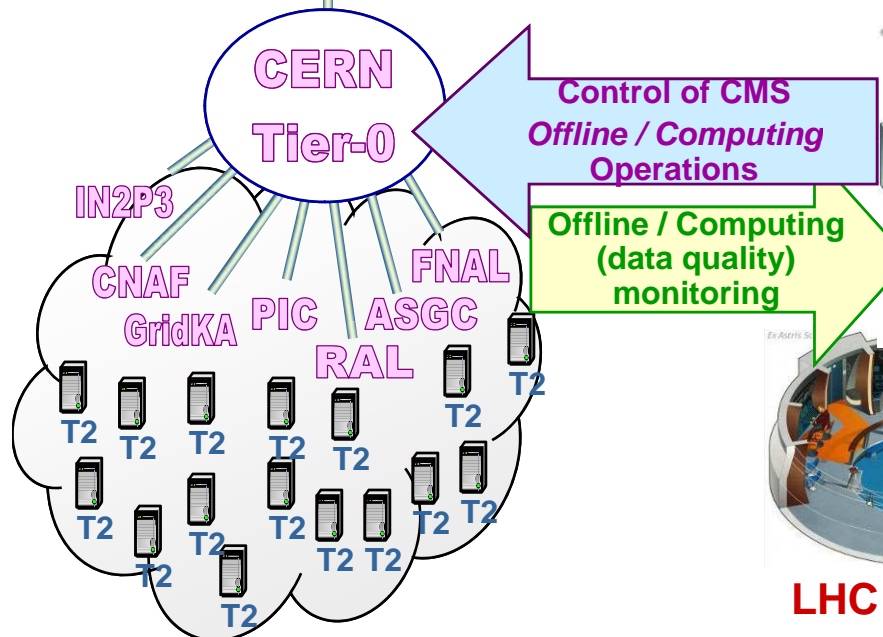
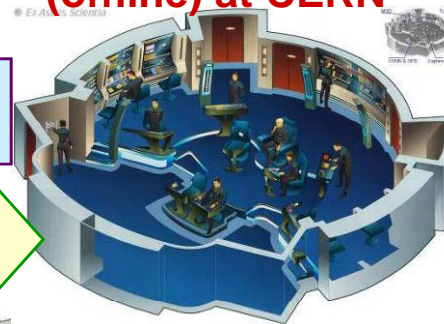
CMS Control Room (Online) at Cessy



❖ CMS Centre and LHC@FNAL

- Link to Control room
 - Mirrors displays
 - Communications
- Computing
 - Operations
 - Tier-0 production
 - Data storage / transfer
- Sub detectors
 - Data quality monitoring (also post Tier-0)
 - Calibration
 - Good/bad runs
 - Software fixes
- Express analysis

CMS Centre (offline) at CERN



LHC@FNAL

Other Centres (?)



Monitoring (Wednesday)

- Has its own summary...



Middleware Developers' Tips

- *"The key point about designing middleware for robustness and resilience is to incorporate these aspects into the initial design.*
- *This is because many of the deployment and operational features already discussed have an impact on the basic architecture and design of the software; it is typically much more expensive to retrofit high-availability features onto a software product after the design and implementation (although it is possible).*
- *In general, the service scaling and high-availability needs typically mandate a more decoupled architecture."*
- **See presentation for more details plus paper for EELA 3**
- **(It's hard to summarize a workshop whilst its going on - let alone write 3 + 1 paper!)**

Top 5 things done 'wrong' (ok, well 6)

Top 5 things done wrong

1. Not using Bind variables
2. Not using Bind variables
3. Not using Bind variables
4. Not using Bind variables
5. Not using Bind variables

Only kidding, but it is my #1 thing “done wrong”

Top ~~5~~ 6 things done wrong

1. Not using Bind variables
2. Not having a test environment
3. Not having any configuration management
4. Database Independence
5. DBA vs Developer
6. Not building to Scale, not building secure

Performance

- Would you compile a subroutine, run it and then throw away the object code for every set of inputs?
 - Hard Parse
- So, why do you do that to SQL...

- Would you compile a subroutine every time to run it, regardless of the inputs?
 - Soft Parse
- So, why do you do that in SQL.....

- Design, test, design, test ...
- Try to prepare a testbed system – workload generators, etc.
- Do not test changes on a live production system.
- IT-PSS provides test and integration system (preproduction) with the same Oracle setup as on production clusters
 - contact PhyDB.Support to obtain accounts and ask for imports/exports.

- More and more data centers run Oracle databases on commodity hardware relying on:
 - Software solutions for high availability (RAC, ASM)
 - Hardware redundancy
- Using commodity hardware may impose relatively frequent hardware changes due to:
 - Short hardware lifetime
 - Short support period

Replacing database hardware without significantly compromising service availability, becomes a challenge as database systems grow larger and larger.

- The Data Guard based migration procedure has been used this year at CERN:
 - we migrated all production and validation databases ~15 systems in total
 - we moved from RHEL 3 to RHEL 4 at the same time
 - we also enlarged all production clusters
 - downtime associated with the migration did not exceed 1 hour per database



When to apply updates / upgrades?

- An issue that we have still not concluded on is when to apply needed updates / upgrades
- I assume that we agree that major changes, machine room configurations etc are done **outside** the period of LHC operation
 - And carefully planned / scheduled / tested...
- But priority bug / security fixes are a fact of life!

Options:

1. Schedule during machine stop / technical development
2. Schedule when necessary - sufficient buffering / redundancy must be built in so no loss of data occurs in short downtimes and active processing of the data will **definitely** occur even with beam off
3. Are there any others?

Where are we?

- We have the lists of prioritized requirements from all LHC experiments
- Work is required to consolidate these:
 - Only 1(?) clear mismatch of service level between VOs
 - But different definition of service level requirements
 - **Are they complete? Are service level requests achievable?**

- DOWN; SERIOUSLY DEGRADATED

- Need to be realistic about 'background' - cannot be avoided
- Can only achieve highest level of service with:
 - RESOURCES
 - WORK
 - "BEST PRACTICES"

Problem description

- User expectations of IT services:
 - 100% availability
 - Response time converging to zero
- Several approaches:
 - Bigger and better hardware (= increasing MTBF)
 - Redundant architecture
 - Load balancing + Failover
- Situation at CERN:
 - Has to provide uninterrupted services
 - Transparently migrate nodes in and out of production
 - Caused either by scheduled intervention or a high load
 - Very large and complex network infrastructure



Thanks

- To all the people who participated, remotely or locally...
- To all who 'volunteered' for various roles...
- In particular, for the co-organisers & co-chairs



Workshop on Resiliency In High-Performance Computing

[Resilience 2008]

19-22 May 2008 @ Lyon, France

In conjunction with 8th IEEE International Symposium on Cluster Computing and the Grid

- The 2008 International workshop on Resiliency in High Performance Computing (Resilience 2008)
<http://xcr.cenit.latech.edu/resilience2008/>
- In conjunction with the 8th IEEE Intentional Symposium on Cluster Computing and Grid (CCGRID 2008), May 18-22, 2008, Lyon, France.
- Important Dates:
 - Paper Submission Deadline extended: December 9, 2007



Summary

- Measured improvements in service quality: April workshop
- Monitor progress using a 'Service Map'
- Size of box = criticality; colour = status wrt "checklist"
- CHEP 2009: all main services at required service level
- Database(-dependent) and data / storage management services appear (naturally) very high in the list + experiment services!
- 24x7 stand-by rota should be put in place at CERN for these services, at least for initial running of the LHC