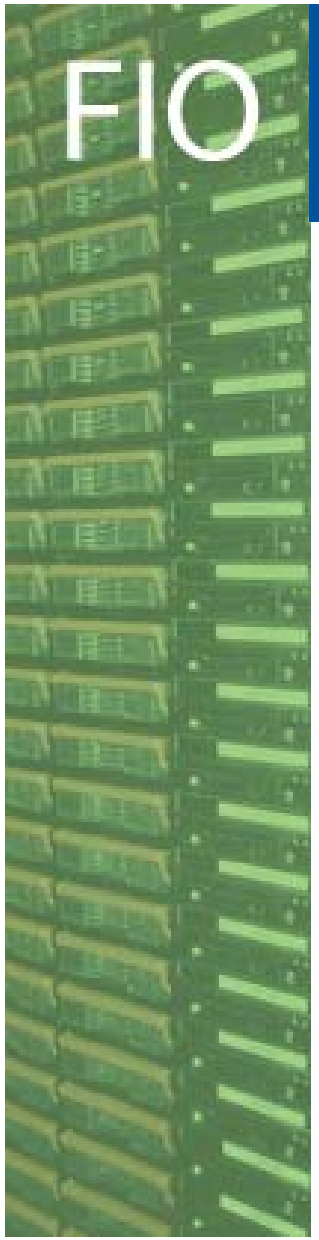


Fabric Infrastructure and Operations

CERN IT Department

WLCG 2009 Data-Taking Readiness Planning Workshop

Tier-0 Experiences

Miguel Coelho dos Santos

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Summary

- Status
- Operations
- Post-Mortems
- Horizon

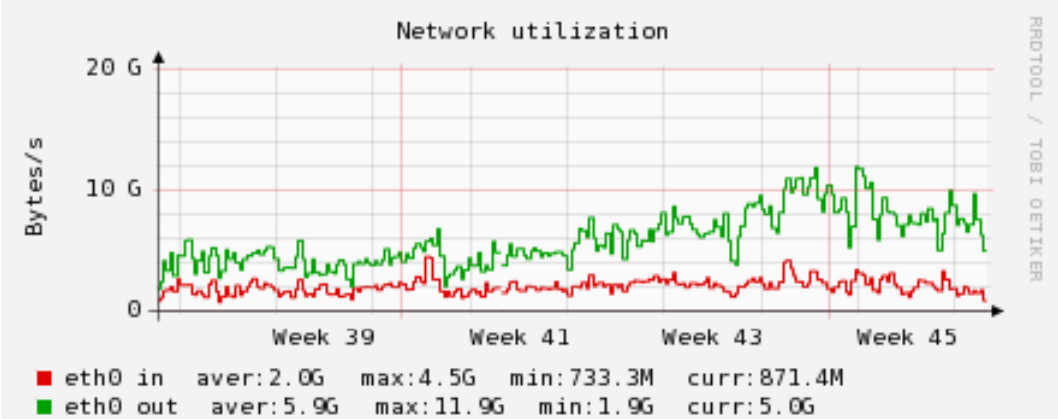
CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Service Status

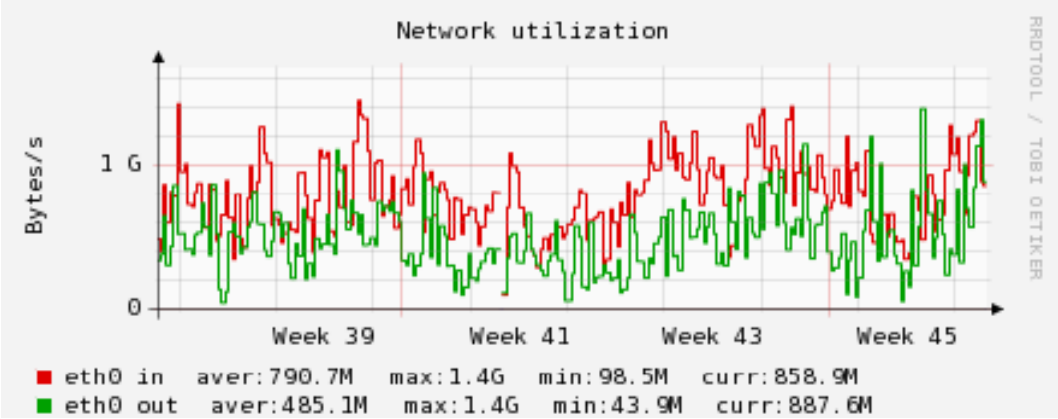
- Generally smooth running on all services. ☺
- Will go over Post Mortems since July in more detail.
- Storage is still most problematic area, most storage problems (volume) have been centred on file access contention (hotspots) and some hardware issues.

Current data activity

- No accelerator... still there is data moving
- Disk Layer

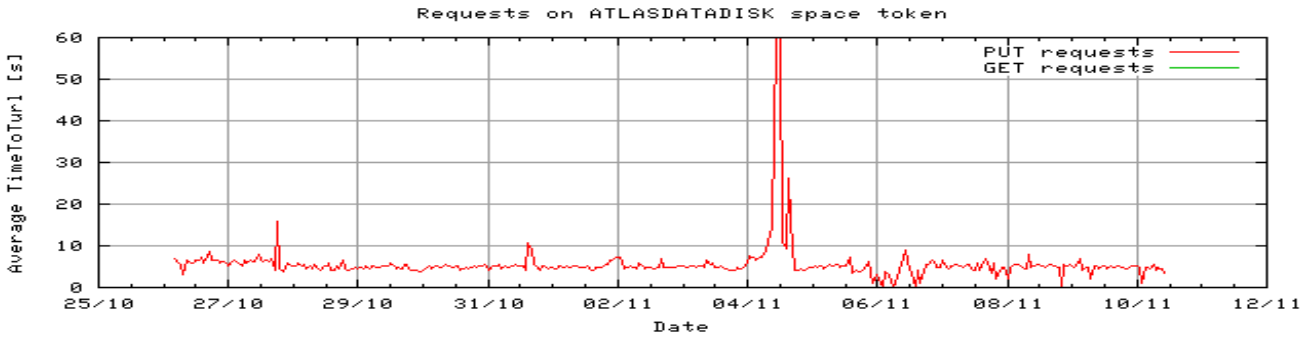


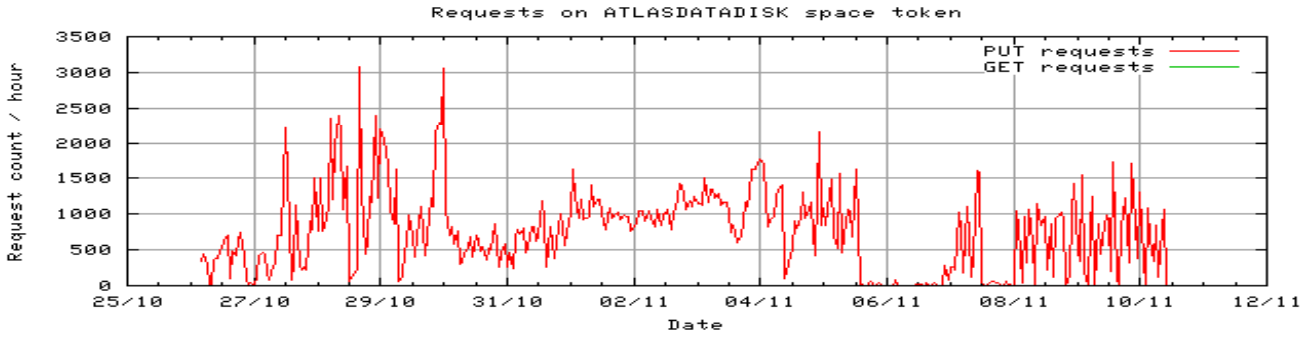
- Tape Layer



CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Current SRM activity





- Production pool, 500 – 1000 requests per hour
- PUT requests take 5 seconds
- Nov 4: scheduled upgrade to Oracle 10.2.0.4

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Operations

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Data Services piquet

- In preparation for data taking:
  - making services more resilient to failures
    - Focus on High Availability deployment using Dynamic DNS
  - Service Manager on Duty rota was extended for 24/7 coverage of critical data services => Data Services Piquet
- Services covered: CASTOR, SRM, FTS and LFC.
- To increase coverage manpower goes up but:
  - The number of admins that can do changes goes up
  - The average service specific experience of admins doing changes goes down

Improve documentation
Make admin Processes more robust (and formal)

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Operations

- Some processes are continuously being improved
- Change management
  - Application upgrades
  - OS upgrade
  - Configuration change or emergencies
- Procedure documentation

DISCLAIMER

ITIL covers extensively some of these topics. The following slides are a summary of a few of our internal processes. They originate from day to day experience, some ITIL guidelines are followed but as ITIL points out, implementation depends heavily on specific environment. Our processes might not apply to other organizations.

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Change Management (I)

- Application Upgrades
  - for example to deploy CASTOR patch 2.1.8-3
- To minimize risk, follow simple rules:
  - Exercise the upgrade on pre-production setup, identical as possible to production setup
  - Document the upgrade procedure thoroughly
  - Reduce concurrent changes
  - Announce with enough lead-time, O(days)
  - "*No upgrades on Friday*", experts need to be available after changes occur
  - Don't upgrade multiple production setups at the same time

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Change Management (II)

- OS Upgrades
- Monthly OS upgrade, 5K servers
  - Test machines are incrementally upgraded every day
  - Test is frozen into Pre-production setup
    - All relevant (specially *critical*) services should (must!) have a pre-prod instance
  - 1 week to verify critical services
    - internal verification
  - 1 week to verify the rest of the services
    - external verification (several *clients,* for example voboxes*)*
  - Changes to production are deployed on a Monday (D-day)
  - D-day is widely announced since D-19 (=7+7+5) days

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Change Management (III)

- Change requests by customer
  - Handled case by case
  - It is generally a documented, exercised, standard change => low risk
- Emergency changes
  - Service already degraded
  - Not change management but incident management!
- Starting to work on measuring change management success (and failures) in order to keep on improving the process

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Procedure documentation

- Document day-to-day operations
- Procedure writer (expert)
- Procedure executer
  - Implementing service changes
  - Handling service incidents
- Executer validates procedure each time it executes it successfully, otherwise the expert is called and the procedure updated
- Simple/Standard incidents are now being handled by recently arrived team member without grid service specific experience

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Post Mortem Review

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Post-mortems (PM)

- According to WLCG a Post Mortem is triggered when a site's MoU commitment has been breached.
- At CERN Post Mortems have been also triggered when interest is high, usually a mixture of:
  - New type of incident
  - Particularly complex incident
  - Out of working hours occurrence
- The following list of Post Mortems include both scenarios and cover the last few months

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

PM History

Date	Service	Time*	Affected	Summary
01/11	Tape	Weekend	All	T10KR1 library down
24/10	FTS	Friday 5pm to 9pm	CERN-ASGC CERN-IN2P3 CERN-RAL NIKHEF-CERN PIC-CERN SARA-CERN	Security Scan
14/10	BATCH	-	LHCb	Resource allocation coupled with activity spike and user error
12/10	CASTOR	Sunday 7am to 2pm	LHCb 	Massive prestaging of files generated high replication traffic
08/10	CASTOR	-	CMS	Silent data corruption on one diskserver
10/09	CASTOR	Wednesday 6:10pm to 8:57pm	ATLAS 	Exports from overloaded server on default pool (**1st real data!)**
28/08	SRM	Thursday 5am to 11am	CMS	ORACLE unable to extend index
25/08	Network (CASTOR disk servers)	Weekend	CMS (cmscaf), LHCb (lhcbdata) 	Intermittent network problems on parts of the network, a subset of disk servers specially affected
29/07	FTS	Monday 6pm to 8pm	All	Oracle sequence corruption after DB export
01/07	Myproxy	Tuesday 7am to 11am	All	STDIN/STDOUT problem when demonizing caused problems when restarts were not manual

PM rate of aprox. 2/month.

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

PM History

Date	Service	Time*	Affected	Summary
01/11	Tape	Weekend	All	Hardware
24/10	FTS	Friday 5pm to 9pm	CERN-ASGC CERN-IN2P3 CERN-RAL NIKHEF-CERN PIC-CERN SARA-CERN	OS
14/10	BATCH	-	LHCb	Configuration
12/10	CASTOR	Sunday 7am to 2pm	LHCb	Application
08/10	CASTOR	-	CMS	Hardware
10/09	CASTOR	Wednesday 6:10pm to 8:57pm	ATLAS	Application
28/08	SRM	Thursday 5am to 11am	CMS	DB
25/08	Network (CASTOR disk servers)	Weekend	CMS (cmscaf), LHCb (lhcbdata)	Network
29/07	FTS	Monday 6pm to 8pm	All	DB
01/07	Myproxy	Tuesday 7am to 11am	All	Application

PM's by configuration item

- By service:
  - FTS, 2
  - CASTOR+SRM, 3+1
  - Tape, 1
  - Network, 1
  - Myproxy,1
  - Batch,1

- By configuration item:
  - Application, 3
  - Hardware, 2
  - DB, 2
  - Network, 1
  - OS, 1
  - Configuration, 1
  - Change, **0**

Missing procedures to, on closure, systematically classify incident reports according to configuration item that was 'involved'

Zero post mortems caused by change!

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Information sources

Information sources

- Currently a lot sources of information:
  - Service status
  - Availability
  - Perceived VO availability
  - Probes
  - Announced interventions
  - Broadcasts
  - Dashboards
  - etc
- Still, it is not trivial to understand (and run) the system we have built!!
- An example follows…

Data Management example

Activity Summary (Last 168 Hours)

Click on the cloud name to view list of sites

Cloud	Transfers: Efficiency	Transfers: Throughput	Transfers: Successes	Registrations: Datasets	Registrations: Files	Errors: Transfer	Errors: Registration	Services: DQ	Services: Grid
+ CERN-PROD_DATADISK	72%	103 MB/s	86027	1183	86016	33545	0	OK	ok

[FTS] FTS State [Failed] FTS Retries [1] Reason [SOURCE error during PREPARATION phase: [CONNECTION] failed to contact on remote SRM [httpg://srm-atlas.cern.ch:8443/srm/managerv2]. Givin' up after 3 tries] Source Host [srm-atlas.cern.ch]	13374
[FTS] FTS State [Failed] FTS Retries [1] Reason [SOURCE error during PREPARATION phase: [REQUEST_TIMEOUT] failed to prepare sour	6649
[FTS] FTS State [Fa ent an error respor urce Host [gridka-d	4845
[FTS] FTS State [Fa] File creation cance	2815
[FTS] FTS State [Fa led to contact on re Host [gridka-dcache.fzk.de]	1310
[FTS] FTS State [Failed] FTS Retries [1] Reason [DESTINATION error during PREPARATION phase: [GENERAL_FAILURE] File creation canceled since diskPool is full] Source Host [gridka-dcache.fzk.de]	1147
[FTS] FTS State [Failed] FTS Retries [1] Reason [SOURCE error during PREPARATION phase: [CONNECTION] [srm2__srmPrepareToGet] failed: SOAP-ENV:Client - CGSI-gSOAP: Could not open connection !] Source Host [srm-atlas.cern.ch]	650

Activity summary.
Error aggregation.
Quantification of “efficiency”.
Good work!

These messages are not dashboard specific, they are from gridftp/SRM/FTS stack.
The dashboard just makes the problem somewhat more evident.
Lets have a closer look at message content...

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Error 1

Activity Summary (Last 168 Hours)

Click on the cloud name to view list of sites

		Transfers			Registrations		Errors		Services	
	Cloud	Efficiency	Throughput	Successes	Datasets	Files	Transfer	Registration	DQ	Grid
+	CERN-PROD_DATADISK	72%	103 MB/s	86027	1183	86016	33545	0	OK	ok

[FTS] FTS State [Failed] FTS Retries [1] Reason [SOURCE error during PREPARATION phase: [CONNECTION] failed to contact on remote SRM [httpg://srm-atlas.cern.ch:8443/srm/managerv2]. Givin' up after 3 tries] Source Host [srm-atlas.cern.ch]

What does this message mean?

Clients can't connect to server.
The service is down!?
Open an incident report.

When can this be normal?

during scheduled downtime

4845

2815

1310

[FTS] FTS State [Failed] FTS Retries [1] Reason [DESTINATION error during PREPARATION phase: [GENERAL_FAILURE] File creation canceled since diskPool is full] Source Host [gridka-dcache.fzk.de]

1147

[FTS] FTS State [Failed] FTS Retries [1] Reason [SOURCE error during PREPARATION phase: [CONNECTION] [srm2__srmPrepareToGet] failed: SOAP-ENV:Client - CGSI-gSOAP: Could not open connection !] Source Host [srm-atlas.cern.ch]

650

Scheduled downtime information should be correlated automatically.

Error 2

Activity Summary (Last 168 Hours)

Click on the cloud name to view list of sites

		Transfers			Registrations		Errors		Services	
	Cloud	Efficiency	Throughput	Successes	Datasets	Files	Transfer	Registration	DQ	Grid
+	CERN-PROD_DATADISK	72%	103 MB/s	86027	1183	86016	33545	0	OK	ok

[FTS] FTS State [Failed] FTS Retries [1] Reason [SOURCE error during PREPARATION phase: [REQUEST_TIMEOUT] failed to prepare source file in 180 seconds] Source Host [srm-atlas.cern.ch]

What does this message mean?

SRM can't get a TURL.
The data is not on disk!?
Open an incident report

When can this be normal?

when data is off-line and needs to be read from tape

	6649
	4845
	2815
led to contact on remote SRM [httpg://srm-atlas.cern.ch:8443/srm/managerv2]. Givin' up after 3 tries] Source Host [gridka-dcache.fzk.de]	1310
[FTS] FTS State [Failed] FTS Retries [1] Reason [DESTINATION error during PREPARATION phase: [GENERAL_FAILURE] File creation canceled since diskPool is full] Source Host [gridka-dcache.fzk.de]	1147
[FTS] FTS State [Failed] FTS Retries [1] Reason [SOURCE error during PREPARATION phase: [CONNECTION] [srm2__srmPrepareToGet] failed: SOAP-ENV:Client - CGSI-gSOAP: Could not open connection !] Source Host [srm-atlas.cern.ch]	650

Error messages should distinguish between on-line and off-line data.

Error 3

Activity Summary (Last 168 Hours)

Click on the cloud name to view list of sites

		Transfers			Registrations		Errors		Services	
	Cloud	Efficiency	Throughput	Successes	Datasets	Files	Transfer	Registration	DQ	Grid
+	CERN-PROD_DATADISK	72%	103 MB/s	86027	1183	86016	33545	0	OK	ok

[FTS] FTS State [Failed] FTS Retries [1] Reason [TRANSFER error during TRANSFER phase: [GRIDFTP] the server sent an error response: 426 426 Transfer aborted (Unexpected Exception : java.io.IOException: Broken pipe)] Source Host [gridka-dcache.fzk.de]

ied to prepare source file in 180 seconds] Source Host [srm-atlas.cern.ch]

What does this message mean?

Transfer died in midair.
Transfers are failing!?
Open an incident report.

When can this be normal?

From the source's perspective, when destination is having problems

	4845
	2815
	1310
[FTS] FTS State [Failed] FTS Retries [1] Reason [DESTINATION error during PREPARATION phase: [GENERAL_FAILURE] File creation canceled since diskPool is full] Source Host [gridka-dcache.fzk.de]	1147
[FTS] FTS State [Failed] FTS Retries [1] Reason [SOURCE error during PREPARATION phase: [CONNECTION] [srm2__srmPrepareToGet] failed: SOAP-ENV:Client - CGSI-gSOAP: Could not open connection !] Source Host [srm-atlas.cern.ch]	650

Logic for determining cause of mid-air failures needs improvement.

Error 4

Activity Summary (Last 168 Hours)

Click on the cloud name to view list of sites

		Transfers			Registrations		Errors		Services	
	Cloud	Efficiency	Throughput	Successes	Datasets	Files	Transfer	Registration	DQ	Grid
+	CERN-PROD_DATADISK	72%	103 MB/s	86027	1183	86016	33545	0	OK	ok

[FTS] FTS State [Failed] FTS Retries [1] Reason [DESTINATION error during PREPARATION phase: [GENERAL_FAILURE] File creation canceled since diskPool is full] Source Host [srm-atlas.cern.ch]

What does this message mean?

Pool is full.
Transfers are failing!?
Open an incident report.

When can this be normal?

when the pool is disk1, not GCed

led to contact on remote SRM [httpg://srm-atlas.cern.ch:8443/srm/managerv2]. Givin' up after 3 tries] Source Host [gridka-dcache.fzk.de]	1310
[FTS] FTS State [Failed] FTS Retries [1] Reason [DESTINATION error during PREPARATION phase: [GENERAL_FAILURE] File creation canceled since diskPool is full] Source Host [gridka-dcache.fzk.de]	1147
[FTS] FTS State [Failed] FTS Retries [1] Reason [SOURCE error during PREPARATION phase: [CONNECTION] [srm2__srmPrepareToGet] failed: SOAP-ENV:Client - CGSI-gSOAP: Could not open connection !] Source Host [srm-atlas.cern.ch]	650

Space token should be evident.
Distinction between system errors and user errors should be introduced.

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Error 5

Activity Summary (Last 168 Hours)

Click on the cloud name to view list of sites

		Transfers			Registrations		Errors		Services	
	Cloud	Efficiency	Throughput	Successes	Datasets	Files	Transfer	Registration	DQ	Grid
+	CERN-PROD_DATADISK	72%	103 MB/s	86027	1183	86016	33545	0	OK	ok

[FTS] FTS State [Failed] FTS Retries [1] Reason [DESTINATION error during PREPARATION phase: [CONNECTION] fai
led to contact on remote SRM [httpg://srm-atlas.cern.ch:8443/srm/managerv2]. Givin' up after 3 tries] Source
Host [gridka-dcache.fzk.de]

What does this message mean?

Destination SRM is down.
The transfers are failing!?
Open an incident report.

When can this be normal?

As seen before, when there is intervention

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Errors and efficiency

Error	Count	% of failures
[FTS] FTS State [Failed] FTS Retries [1] Reason [SOURCE error during PREPARATION phase: [CONNECTION] failed to contact on remote SRM [httpg://srm-atlas.cern.ch:8443/srm/managerv2]. Givin' up after 3 tries] Source Host [srm-atlas.cern.ch]	13374	39.9
[FTS] FTS State [Failed] FTS Retries [1] Reason [SOURCE error during PREPARATION phase: [REQUEST_TIMEOUT] failed to prepare source file in 180 seconds] Source H	6649	19.8
[FTS] FTS State [F TRANSFER phase aborted (Unexpect [gridka-dcache.fzk.	4845	14.4
[FTS] FTS State [F PREPARATION ph full] Source Host [s	2815	8.4
[FTS] FTS State [Failed] FTS Retries [1] Reason [DESTINATION error during PREPARATION phase: [CONNECTION] failed to contact on remote SRM [httpg://srm-atlas.cern.ch:8443/srm/managerv2]. Givin' up after 3 tries] Source Host [gridka-dcache.fzk.de]	1310	3.9
other	4552	13.6
Total number of failures	~~**33545**~~	4552
Total number of successes	**86027**	
Source Efficiency	~~**72%**~~	95%

There is a good chance that, due to known problems, the top5 errors are easily explained and could easily be **avoided**.
Lets go through the mental exercise of removing them...

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Missing information

- To improve operations or troubleshooting more information should to be available across components and error messages need to better reflect the system state.
  - Au… …wntime and err…
  - Distinguish errors on off-line and on-line data
  - Im… …of midair fail…
  - Ma…
  - Distinguish between system and user errors
  - Disambiguate better between source and destination errors

Right now we (experiments+IT) spend time investigating the difference between 72% and 95% efficiency. ☹☹☹

If this information is shared between components *maybe* we will spend our time investigating the difference between 95% and 99.9% efficiency!

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Horizon

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

In preparation

- LFC upgrade
- WMS migration to new hardware and glite release
- LSF minor version upgrade
- RB service shutdown
- Deploy SRM 2.7 in production (PPS available)
  - runs on SLC4, linked against recent Castor version, redundant backend servers, admin tools, improved housekeeping, logging improvements, …
  - Stop SRM 1.3 endpoints
- Stopping FTS on SLC3 (2.0)
- CASTOR upgrade to 2.1.8
  - Important bug fixes, monitoring
  - Improve concurrent read access (xroot)
- Tape drive upgrade (Higher density)
- Oracle upgrades

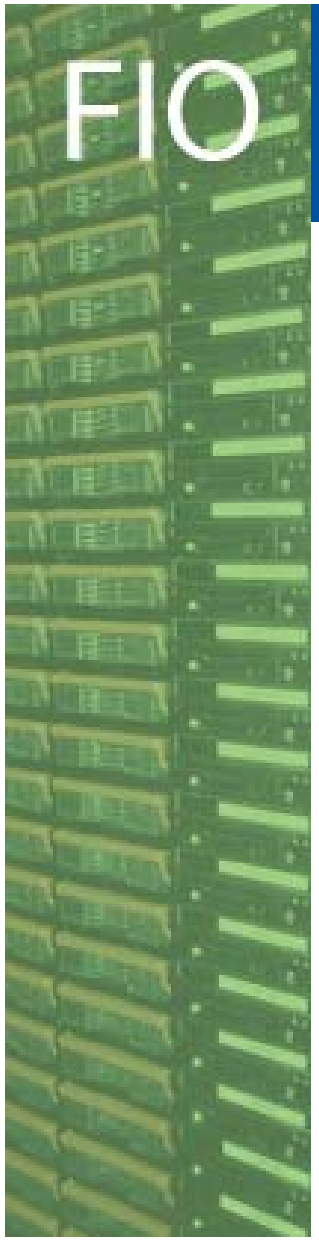
SLC5

- Grid Preproduction
  - first SLC5 WN and lxplus machines in production for testing by the users
  - Starting migration to run the 64bit WN software on SLC5
  - there is a pilot to track the process on this
- In the horizon planning for migration of other services to SLC5, depending on software availability.

Concluding...

- Smooth running
- Focus on operations
  - Streamlining procedures
  - Measuring results
- Storage is still the most sensitive area
  - Information needs to be better shared between components to avoid unnecessary blind retires and improve troubleshooting.
- Important upgrades coming
  - Special interest on those that ease operations and/or make the system more stable

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it



Questions?

CERN IT Department

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it