WLCG Service Report

Maria.Girone@cern.ch
Harry.Renshall@cern.ch
Jamie.Shiers@cern.ch

~~~

WLCG Management Board, 15th December 2009
~~~

Overview

- One year ago I (jds) proposed that by November 2009 the weekly report should be a quick review of the KPIs **and** confirmation of no / few SIRs
- The KPIs – particularly site availability – have improved noticeably over this period with a general rise in the # of ATLAS (in particular) GGUS tickets
- **But** the # of SIRs does not seem to be decreasing

- **First data** brings also new issues, including the realization that we need to be more agile in scheduling interventions
  - e.g. have interventions "ready to go" for a convenient slot in accelerator operation, rather than a fixed slot – which may no longer be convenient when the time comes around…

- **Xmas plans:** here

External References

Q4 2009

Site	Date	Duration	Service	Impact	Report
IN2P3	8 Dec	1.5 hours	Networking	Grid services unavailability caused by load balancing mechanism failure	https://twiki.cern.ch/twiki/pub/LCG/WLCGServiceIncidents/sir_in2p3network_outage_10_12_2009.pdf
CERN	2 Dec	2 hours +	Site wide power cut	Most CC services down	https://twiki.cern.ch/twiki/bin/view/FIOgroup/PostMortem02Dec09
RAL	30 Nov	n/a	Storage	LHCb Data Loss Incident at RAL	http://www.gridpp.ac.uk/wiki/RAL_Tier1_Incident_20091130
IN2P3	12 Nov	n/a	Storage	CMS Data Loss Incident at FR-CCIN2P3	https://twiki.cern.ch/twiki/pub/LCG/WLCGServiceIncidents/2009-11-26_CMS_CCIN2P3_Report.pdf
CERN	20 Nov	1h	SRM/ATLAS	SRM high failure rate and restart after thread exhaustion	https://twiki.cern.ch/twiki/bin/view/FIOgroup/PostMortem20Nov09
CERN	18 Nov	10h	CMS Dashboard	Performance degradation	http://dashboard.cern.ch/reports/CMSmigrationProblem
IN2P3	3 Nov	4h	Many	Many services have been disturbed due to automatic reboot of machines	https://twiki.cern.ch/twiki/pub/LCG/WLCGServiceIncidents/SIR_CCIN2P3_cooling_outage_03nov2009.doc
RAL	9 Oct	n/a	Storage (Castor)	data loss from Castor	http://www.gridpp.ac.uk/wiki/RAL_Tier1_Incident_20091009
IN2P3	14 Oct 2009	13h	batch	only very short jobs able to run	https://twiki.cern.ch/twiki/pub/LCG/WLCGServiceIncidents/sir_BatchIncident_15_10_09.pdf
CERN	13 Oct 2009	1-2h	CASTOR nameserver sick	All CASTOR services dead	https://twiki.cern.ch/twiki/bin/view/FIOgroup/PostMortem20091013
IN2P3	8 & 10 Oct 2009	11h (8 Oct) and 6h (10 Oct)	SRM crashed	SRM service interrupted	https://twiki.cern.ch/twiki/pub/LCG/WLCGServiceIncidents/SIR_CCIN2P3_SRM_incident_08oct2009.doc
RAL	4-9 Oct 2009		disk failures -> Oracle problems	CASTOR, LFC and FTS services down	http://www.gridpp.ac.uk/wiki/RAL_Tier1_Incident_20091004
ASGC	27 Sep - xx Oct	>3 weeks	DBs	down & out	https://twiki.cern.ch/twiki/pub/LCG/WLCGServiceIncidents/ASGC-DB-Sep28.pdf

Q3 2009

Site	Date	Duration	Service	Impact	Report
CERN	21 Sep 2009	08:00 - 18:00	DB Replication	ATLAS Replication Tier0->Tier1 down	https://twiki.cern.ch/twiki/bin/view/PDBService/StreamsPostMortem
RAL	15 - 17 Sept 2009	2 days	CASTOR	Disk to Disk (D2D) transfers started failing during a planned upgrade to the NS	http://www.gridpp.ac.uk/wiki/RAL_Tier1_Incident_20090915
FZK	7 - 16 Sep 2009	10 days	ATLAS RAC	3D Streams replication blocked then degraded	https://twiki.cern.ch/twiki/bin/viewfile/LCG/WLCGServiceIncidents?rev=1;filename=SIR-FZK-20090907.pdf
CERN	5 & 8 Sept 2009	2 * 2 hours	CASTOR LHCb	two Castor Database problems	https://twiki.cern.ch/twiki/bin/view/FIOgroup/PostMortem20090905
CERN	26 Aug 2009	18:40 - 23:30	Batch	Public and production queues closed	https://twiki.cern.ch/twiki/bin/view/FIOgroup/PostMortem20090826
ASGC	17 Jul 2009	6:00 - 10:00	Power cut	Most services went down and restarted	https://twiki.cern.ch/twiki/bin/viewfile/LCG/WLCGServiceIncidents?rev=1;filename=power_cut_ASGC.txt
ATLAS	13 Jul 2009	10:00 - 11:00	Central Catalogs	Degrade of performance	PostMortem13Jul09

Q2 2009

Site	Date	Duration	Service	Impact	Report
NL-T1	STEP09				https://twiki.cern.ch/twiki/pub/Atlas/Step09Feedback/Post_Mortem_STEP09_NL-T1-0.4.pdf
OPN	10 Jun 09	>1 day	LHC OPN	primary circuits to ASGC, CNAF, KIT, NDGF, TRIUMF (incl. backup)	https://twiki.cern.ch/twiki/pub/LCG/WLCGServiceIncidents/Fibre_Cut_June_2009.pdf
FZK	STEP09	many days	storage		https://twiki.cern.ch/twiki/pub/LCG/WLCGServiceIncidents/SIR_storage_FZK_GridKa.pdf

Mind the Gap

- The current WLCG "service model" has been built up and proven over the past 5+ years
  - I am counting from the run-up to the experiment-oriented Service Challenges starting in 2005, followed by CCRC'08 and STEP'09
    - The two latter occurred when "we should have been running"
- This model works – no doubt can be improved / optimized – but needs to survive the multiple transitions we are now facing
  - New group structures, new projects, new SCODs, …
- Important to maintain attendance at WLCG "daily meeting" – reports from ATLAS and ALICE (as generated since a long time by CMS then LHCb) would really help!
- Is the level of incidents leading to SIRs acceptable? (the new baseline? Is there someway we can improve here?

➢ **Include "SIR follow-up" in quarterly GDB operations review**

☝ **Improve "official" information flow from machine out to sites, particularly during early days…**

Introduction

- Covers the weeks 30 November to 13 December. Included period of LHC operation giving 1 million collisions at 450 GEV per beam and 50000 collisions at 1.18 TEV per beam.

- Mixture of problems

- 2 alarm tickets from ATLAS:
  - Test alarm to FZK on 9 Dec after ggus release
  - Alarm to CERN-PROD on 12 Dec:
    - REQUEST_TIMEOUT for ATLASDATADISK

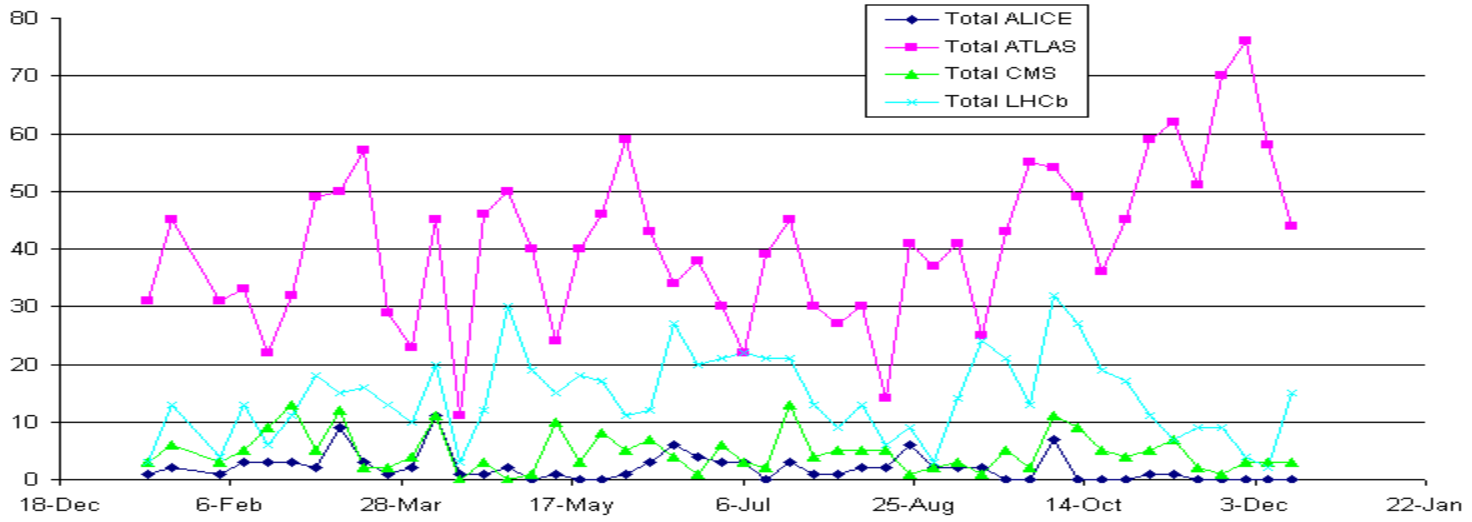
- Incidents leading to (eventual) service incident reports
  - RAL 30 Nov LHCb data loss
  - CERN 2 Dec site wide power cut for just over 2 hours
  - IN2P3 8 Nov DNS load balancing failure affected grid services for 1.5 hours

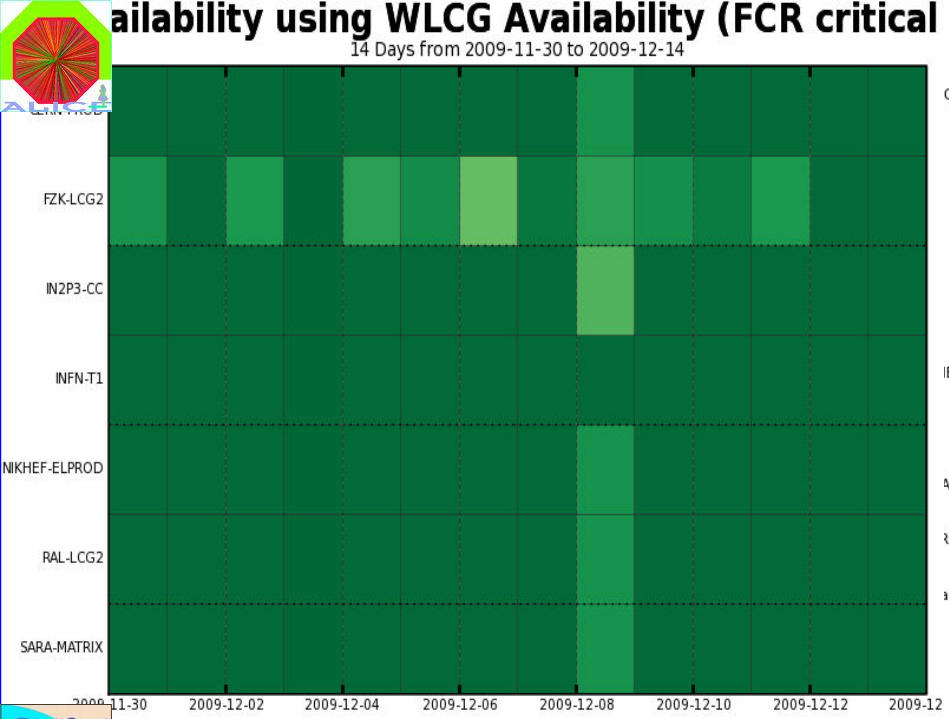
Meeting Attendance Summary (Last week only)

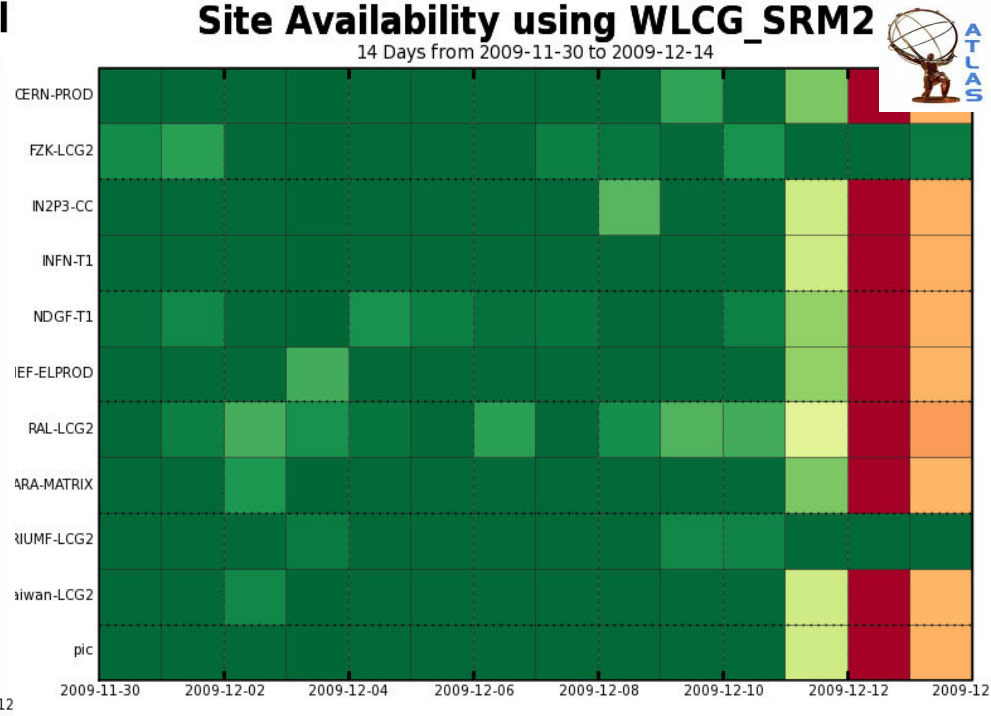
Site	M	T	W	T	F
CERN	Y	Y	Y	Y	Y
ASGC	Y	Y	Y	Y	Y
BNL	Y	Y	Y	Y	Y
CNAF	Y				
FNAL					
FZK	Y	Y		Y	Y
IN2P3	Y	Y	Y	Y	Y
NDGF	Y			Y	Y
NL-T1		Y	Y	Y	
PIC			Y		Y
RAL	Y	Y	Y	Y	Y
TRIUMF					

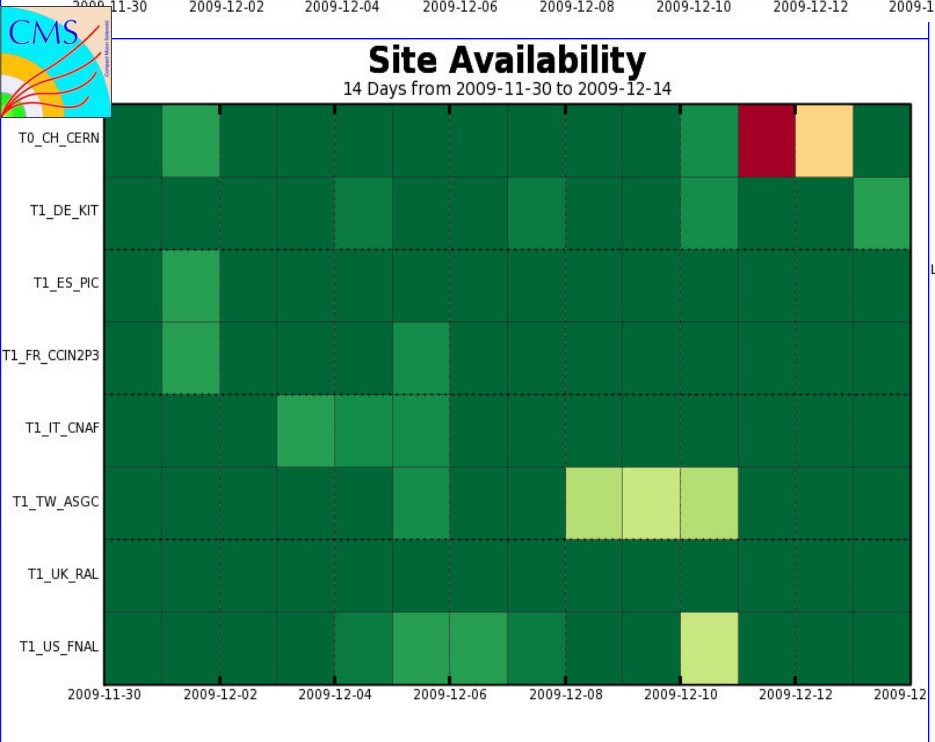
GGUS summary (2 weeks)

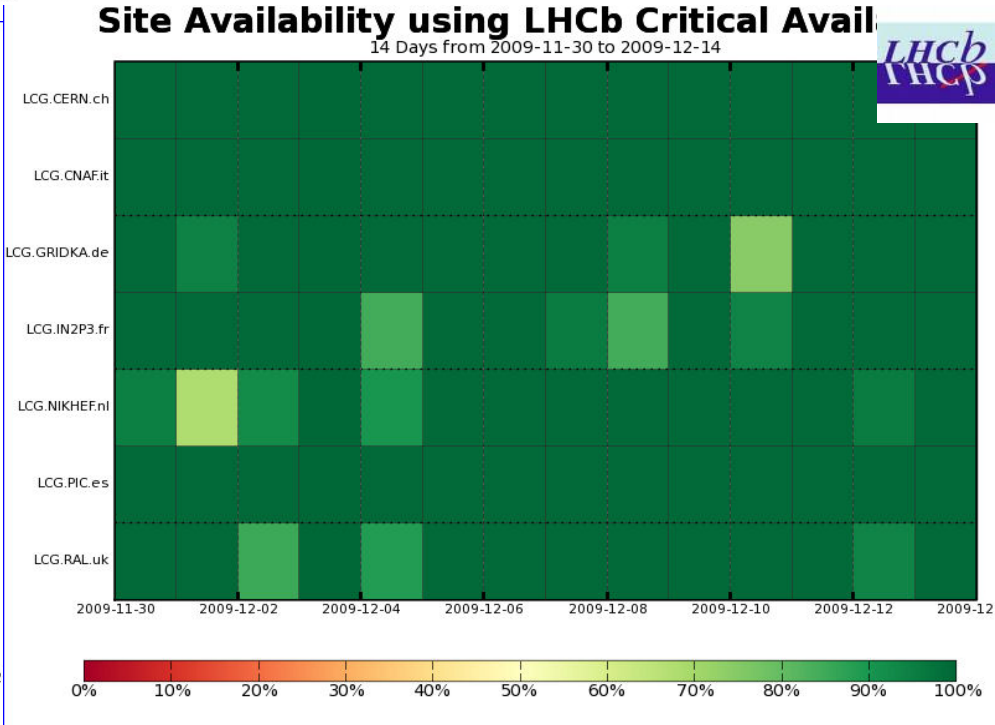
VO	User	Team	Alarm	Total
ALICE	0	0	0	0
ATLAS	27	73	2	102
CMS	5	1	0	6
LHCb	2	15	0	17
Totals	34	89	2	125




ailability using WLCG Availability (FCR critical
14 Days from 2009-11-30 to 2009-12-14
FZK-LCG2
IN2P3-CC
INFN-T1
NIKHEF-ELPROD
RAL-LCG2
SARA-MATRIX
2009-11-30
2009-12-02
2009-12-04
2009-12-06
2009-12-08
2009-12-10
2009-12-12
2009-12


Site Availability using WLCG_SRM2
14 Days from 2009-11-30 to 2009-12-14
CERN-PROD
FZK-LCG2
IN2P3-CC
INFN-T1
NDGF-T1
EF-ELPROD
RAL-LCG2
ARA-MATRIX
RIUMF-LCG2
iwan-LCG2
pic
2009-11-30
2009-12-02
2009-12-04
2009-12-06
2009-12-08
2009-12-10
2009-12-12
2009-12


Site Availability
14 Days from 2009-11-30 to 2009-12-14
T0_CH_CERN
T1_DE_KIT
T1_ES_PIC
T1_FR_CCIN2P3
T1_IT_CNAF
T1_TW_ASGC
T1_UK_RAL
T1_US_FNAL
2009-11-30
2009-12-02
2009-12-04
2009-12-06
2009-12-08
2009-12-10
2009-12-12
2009-12


Site Availability using LHCb Critical Avail
14 Days from 2009-11-30 to 2009-12-14
LCG.CERN.ch
LCG.CNAF.it
LCG.GRIDKA.de
LCG.IN2P3.fr
LCG.NIKHEF.nl
LCG.PIC.es
LCG.RAL.uk
2009-11-30
2009-12-02
2009-12-04
2009-12-06
2009-12-08
2009-12-10
2009-12-12
2009-12
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

SAM Failures and Alarm Tickets

- ATLAS SAM multi-site failures on 12 Dec were due to a User certificate problem.

- CMS SAM failures at CERN on 11 Dec still under investigation.

- ATLAS alarm ticket at 07.30 on 12 Dec: all the transfers to ATLASDATADISK and ATLASMCDISK are failing with error [REQUEST_TIMEOUT] failed to prepare source file in 180 seconds.
  Analysis was: This was a temporary overload that has since corrected itself. The delays in scheduling probably were related to bursts of Ganga jobs in atlasscratchdisk, all trying to access a few files from lxfsrl2008, which is closed in LSF and building up a queue. This does not affect SRM or FTS directly, but since they use the CASTOR-ATLAS stager for the actual transfer, anything that slows down the stager will slow down SRM.

- May not be whole story as a similar incident happened a few days later when there was little scratchdisk activity.

SIRs (1/3)

- **RAL Tier1 Incident 20091130:**
- **Double Disk Failure on Server Led to Data Loss**
- **Site:** RAL-LCG2
- **Incident Date:** 2009-11-30
- **Severity:** Severe
- **Service:** CASTOR, LHCb
- **Impacted:** LHCb
- **Incident Summary:** Two disks failed in gdss138 (lhcbDst, disk 1 tape 0) within 30 minutes of each other, rendering the RAID5 data array inoperable. After attempts to recover the array failed, the data was declared as lost, see the timeline for full details
- **Type of Impact**: Data Loss
- **Incident duration:** 1 day
- Comment: Double disk failures seem to be increasingly likely on the RAID5 machines as they get older. The Tier1 is considering how we can migrate disk 1 tape 0 data from RAID5 systems to RAID6 systems as RAID6 can rebuild an array after two drive failures.

SIRs (2/3)

- **CERN Site wide power cut 01:22 to 03:30 on 2 December**

- All the services that are NOT hosted in the critical area were affected by a ~2 hours Power Cut on 2nd December.
- **Impact was several hours downtime** for the following services:
  - **CASTOR; BATCH**; **LXPLUS**; **Backup**
  - Grid Services:
    - **LFC; FTS; CE** (LCG and CREAM); **WMS**
  - **Tape BACKUP**
- **Limited impact** on the following services:
  - **AFS**
- **Other impact on users**
  - **CASTOR**: "No space left on device" error during the time laps where the CASTOR head nodes had been restarted, but the disk servers were still down.
    - propagated to **SRM**
  - **Computing Elements**: possible mapping problems (no user complaint though) because gridnfs needed to be restarted before the CEs.

SIRs (3/3)

- **IN2P3 Outage of import / export of LHC Data**

- **Duration: 2 hours**
- **Date: December 8th 2009 13:00 to December 8th 2009 14:30**
- **Description** Grid services (bdii, ATLAS LFC) unavailability caused by load balancing mechanism failure.
- **Analysis**
- A problem on the 2 servers called "lbnamed" managing the load balancing service occurred at 13:00.
- Those 2 servers are monitored and when one of them fails, it is automatically restarted.
- The restart process did not work during the incident for a reason which has been identified. It is now corrected.
- The initial reason of this breakdown which occurred simultaneously on the two self supporting servers is not identified. Investigations about the subject are in progress and external factors are suspected.

Miscellaneous Reports

- Most experiments reporting good performance for data export and event reconstruction at Tier-0 and over the grid.
- Xmas Experiment and site plans are available at https://twiki.cern.ch/twiki/bin/view/LCG/WLCGExperimentandSiteplansfortheendof2009holidays
  - Nothing special required for ALICE, CMS and LHCb but ATLAS plan to reconstruct between 30 and 150 million events over Tier-0/1. Should start after validation completed by 22 December and be finished by 1 Jan.

Summary/Conclusions

- After many (many) years of preparation we finally see first real data taking from (accelerated) pp collisions

➢ **Need to remain agile and attentive – now the fun begins!**

- IMHO the grid needs to be **part of the solution** (and not part of the problem)

- Surely there are ways that we can **combine** our efforts, knowledge and experience grid-wide to provide a **better** service at **lower** manpower cost?

- Something to drive through the HEP VRC / HUC in 2010?

2009

season's greetings | meilleurs vœux

European Organization for Nuclear Research - Organisation européenne pour la recherche nucléaire