LCG-France Tier-1 & AF

Réunion de Coordination

Fabio Hernandez
fabio@in2p3.fr

Lyon, 16 avril 2009

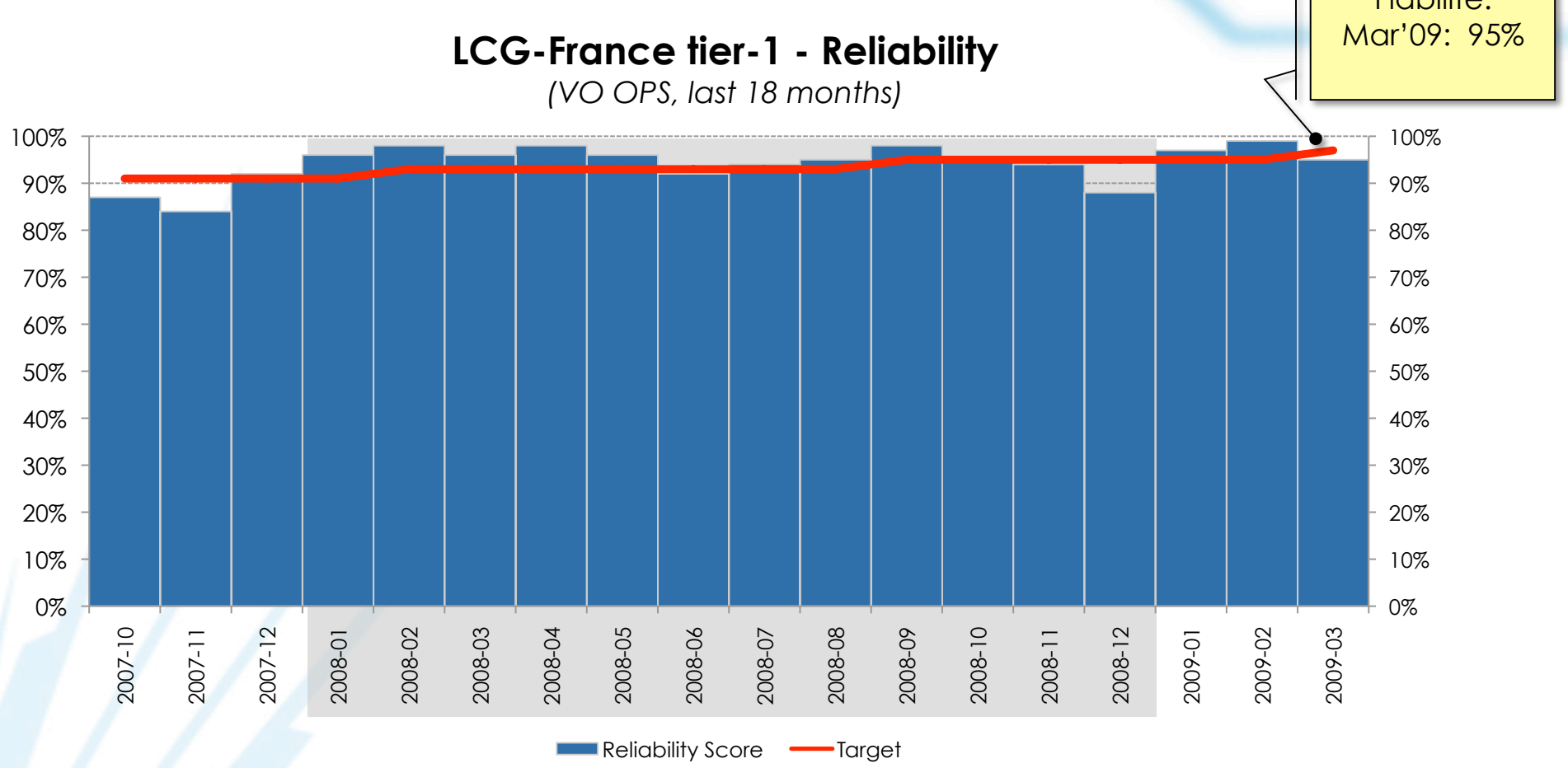
Table des Matières

- Réunions GDB, MB et WLCG workshop de mars 2009
- Disponibilité, fiabilité, efficacité des sites
- Chantiers en cours
- Thème(s) du jour

GDBs & MBs

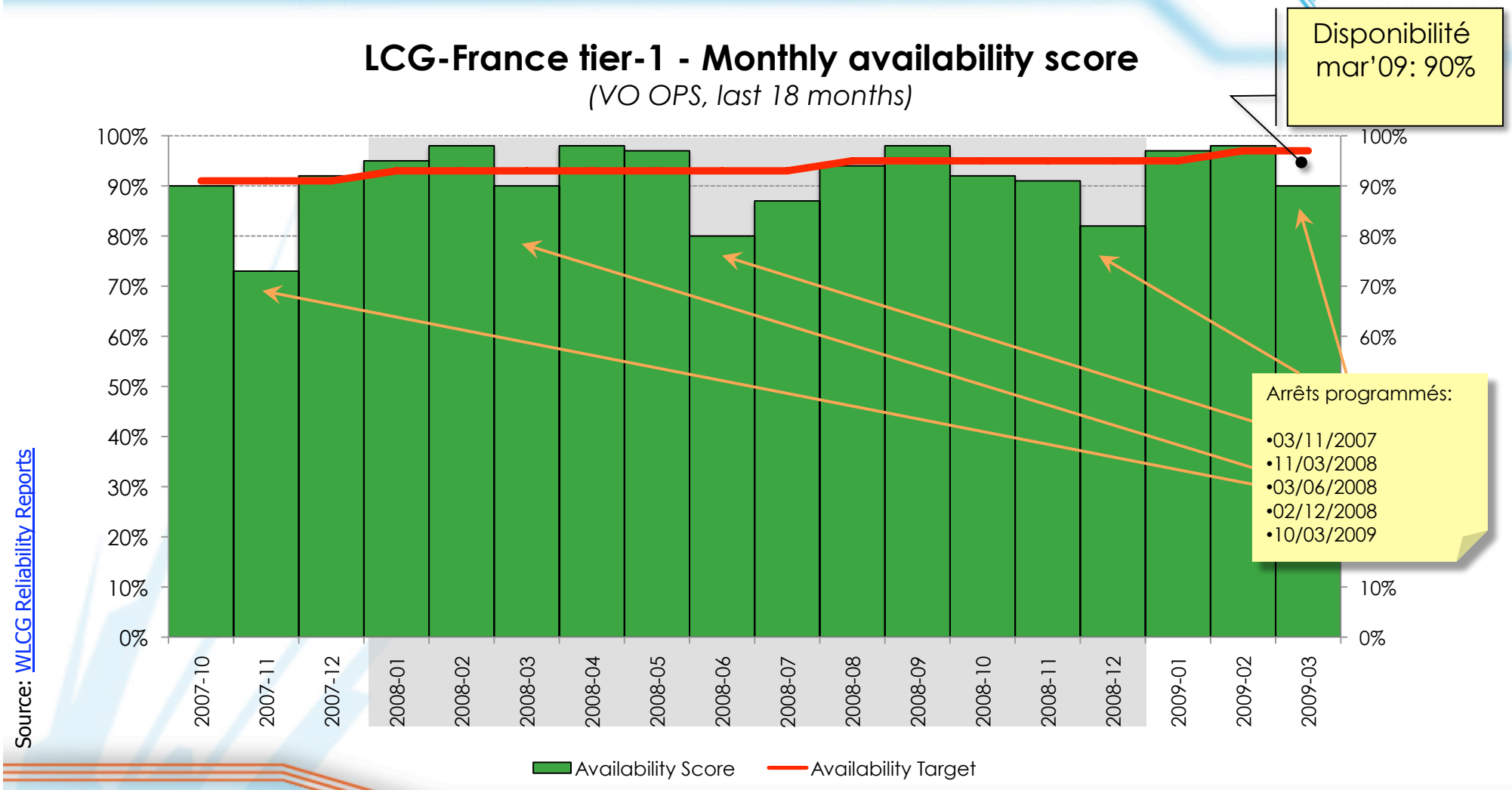
- Agendas:
  - GDB: http://indico.cern.ch/categoryDisplay.py?categId=3l181
  - MB: http://indico.cern.ch/categoryDisplay.py?categId=666
- Principaux sujets traités
  - Mises à jour middleware
  - Calendrier démarrage du LHC
  - STEP'09

Fiabilité du tier-1



Source: WLCG Reliability Reports

Disponibilité du tier-1

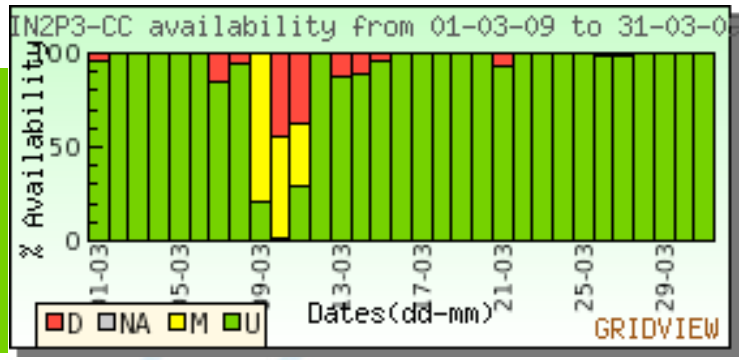


Source: WLCG Reliability Reports

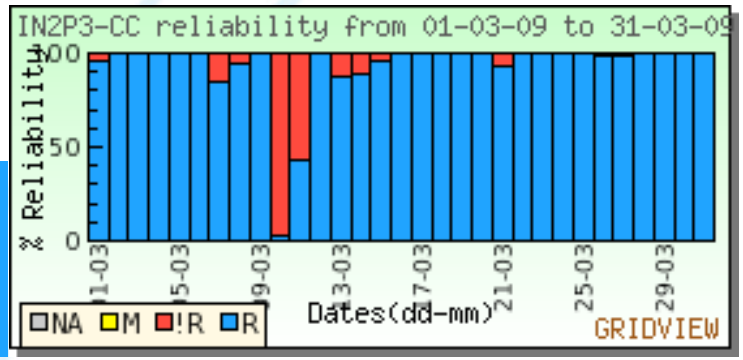
Disponibilité & Fiabilité tier-1

Mars 2009

Disponibilité



Fiabilité

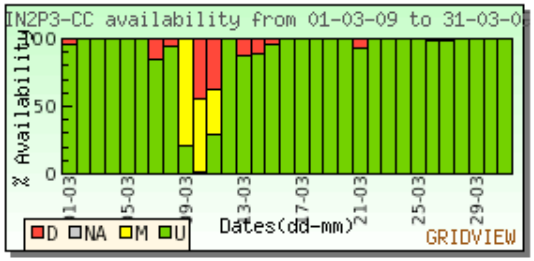


Source: http://gridview.cern.ch

Disponibilité tier-1

- Mars 2009

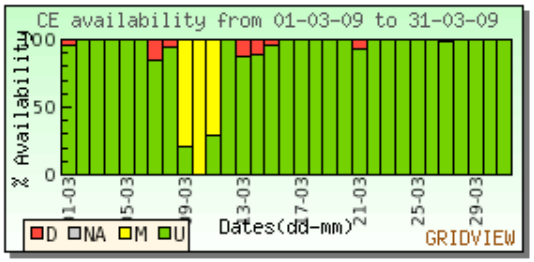
Overall Service Availability for Site:IN2P3-CC VO:OPS (Daily Report)



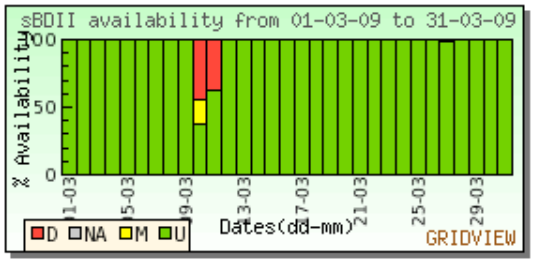
Score: 90%

Individual Service Availability for site:IN2P3-CC VO:OPS (Daily Report)

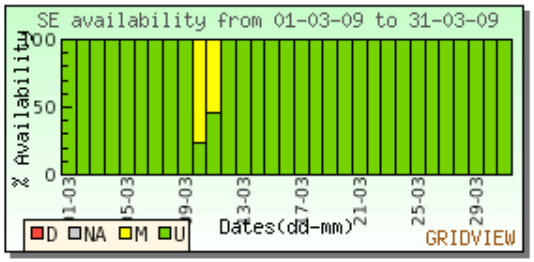
CE



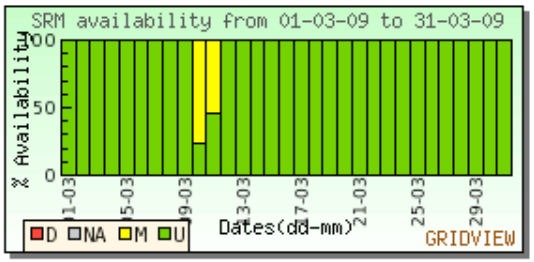
sBDII



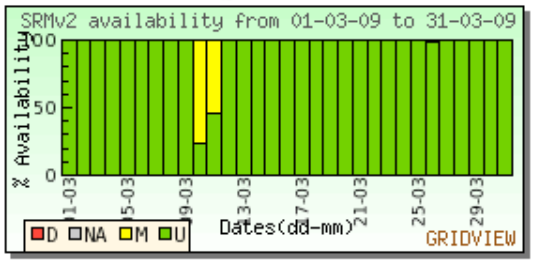
SE



SRM



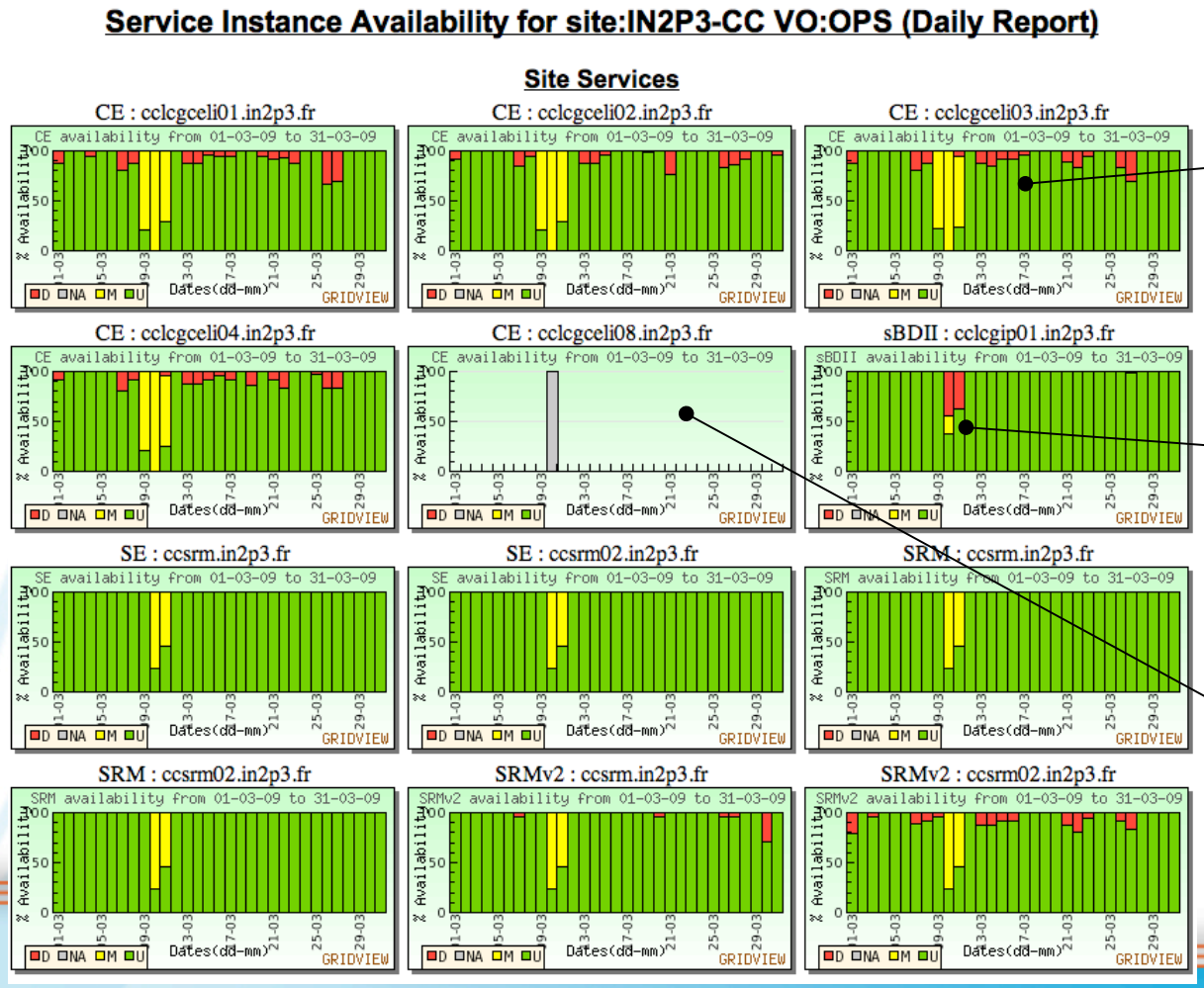
SRMv2



Source: http://gridview.cern.ch

Disponibilité tier-1 (suite)

- Mars 2009 (suite)



Ces échecs attribués aux CEs sont-ils liés aux tests SRM?

SiteBDII indisponible pendant l'arrêt?

Un CE transient?

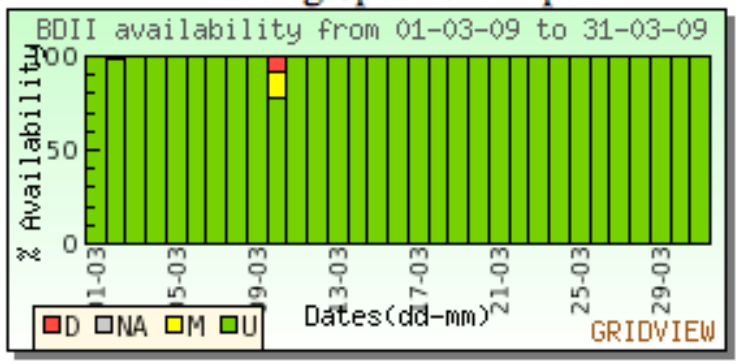
Source: http://gridview.cern.ch

Disponibilité tier-1 (suite)

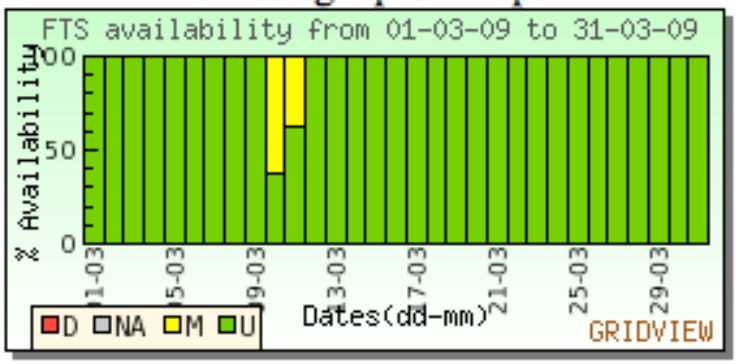
- Mars 2009 (suite)

Central Services

BDII : cclcgtopbdii01.in2p3.fr



FTS : cclcgftsprod.in2p3.fr

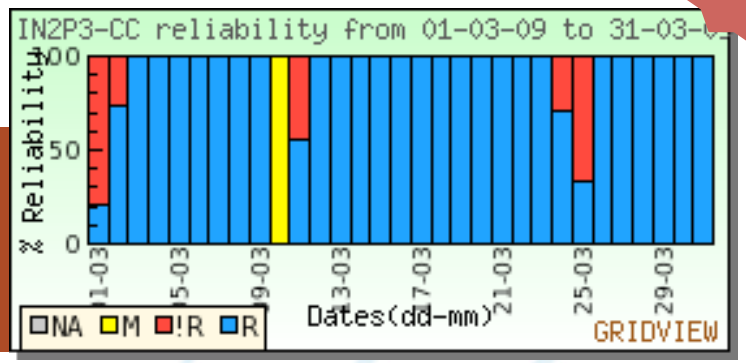


Source: http://gridview.cern.ch

Fiabilité vue par les VOs: Mars 2009

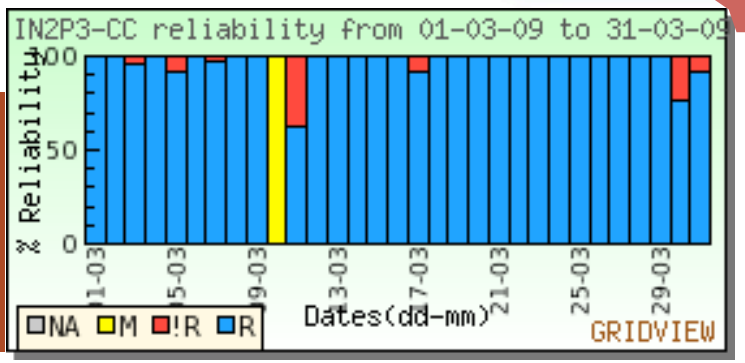
ALICE

??%



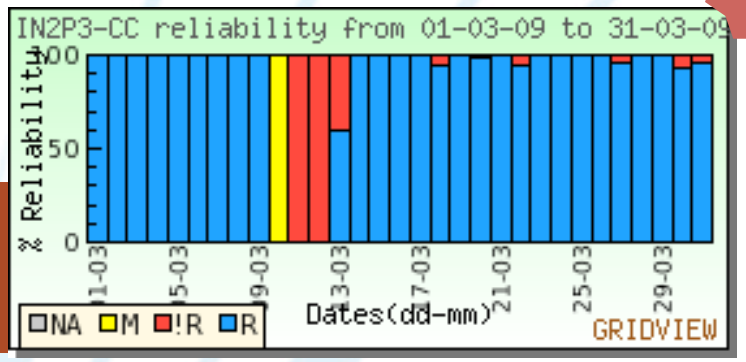
ATLAS

??%



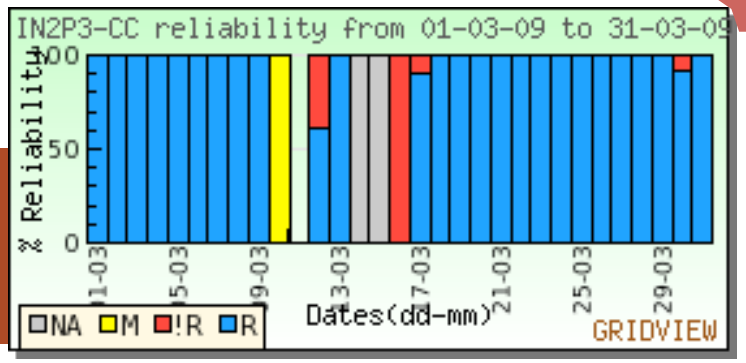
CMS

??%



LHCb

??%



Source: http://gridview.cern.ch

EGEE France – disponibilité & fiabilité

Mars 09

Region	Avail-ability	Reli-ability
UKI	94 %	95 %
CentralEurope	93 %	94 %
SouthWesternEurope	92 %	96 %
Russia	92 %	93 %
France	92 %	96 %
CERN	91 %	93 %
GermanySwitzerland	89 %	92 %
SouthEasternEurope	88 %	91 %
NorthernEurope	87 %	89 %
Italy	74 %	82 %
AsiaPacific	27 %	55 %

Source: https://edms.cern.ch/document/963325

EGEE France – disponibilité & fiabilité

Mars 09

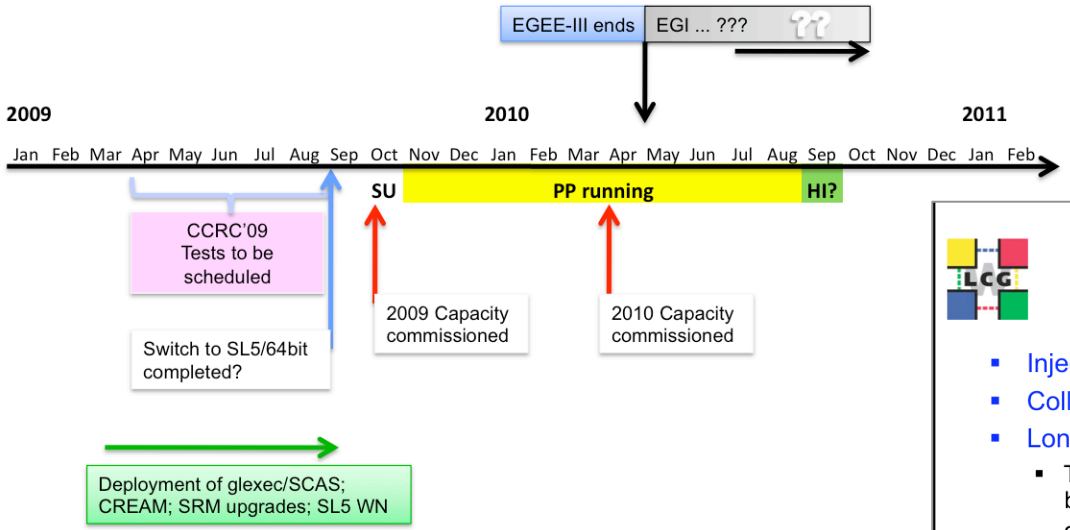
Region	Site	Phy. CPU	Log. CPU	Avail-ability	Reli-ability	Availability History Dec-08	Jan-09	Feb-09
France (France)								
	AUVERGRID	2	2	98 %	98 %	100 %	98 %	100 %
	CGG-LCG2	80	1	96 %	96 %	74 %	46 %	73 %
	ESRF	16	16	98 %	99 %	N/A	84 %	89 %
	GRIF	3,356	2,664	92 %	95 %	97 %	99 %	100 %
	IBCP-GBIO	32	10	94 %	95 %	21 %	60 %	67 %
	IN2P3-CC	3,563	0	90 %	95 %	82 %	97 %	98 %
	IN2P3-CC-T2	3,563	0	89 %	97 %	82 %	96 %	97 %
	IN2P3-CPPM	360	344	95 %	96 %	97 %	99 %	97 %
	IN2P3-IPNL	452	120	98 %	99 %	96 %	98 %	96 %
	IN2P3-IRES	664	652	93 %	93 %	95 %	95 %	96 %
	IN2P3-LAPP	512	512	98 %	99 %	73 %	94 %	100 %
	IN2P3-LPC	448	448	99 %	99 %	97 %	94 %	93 %
	IN2P3-LPSC	120	112	94 %	94 %	96 %	97 %	99 %
	IN2P3-SUBATECH	276	276	99 %	99 %	94 %	99 %	96 %
	IPSL-IPGP-LCG2	34	34	96 %	99 %	66 %	94 %	100 %

Ces chiffres, correspondent-ils à la réalité de ce qui est installé?

Source: https://edms.cern.ch/document/963325

Calendrier

WLCG timeline 2009-2010



Likely scenario

- Injection: end September 2009
- Collisions: end October 2009
- Long run from ~November 2009 for ~44 weeks
  - This is equivalent to the full 2009 + 2010 running as planned with 2010 being a nominal year
  - Short stop (2 weeks) over Christmas/New Year
- Energy will be limited to 5 TeV
- Heavy Ion run at the end of 2010
  - No detailed planning yet
- 6 month shutdown between 2010/2011 (?) – restart in May ?
- Now understand the effective amount of data taking in 2009+2010 will be ~6.1 x 10^6 seconds (cf 2 x 10^7 anticipated in original planning)

rd@cern.ch 7

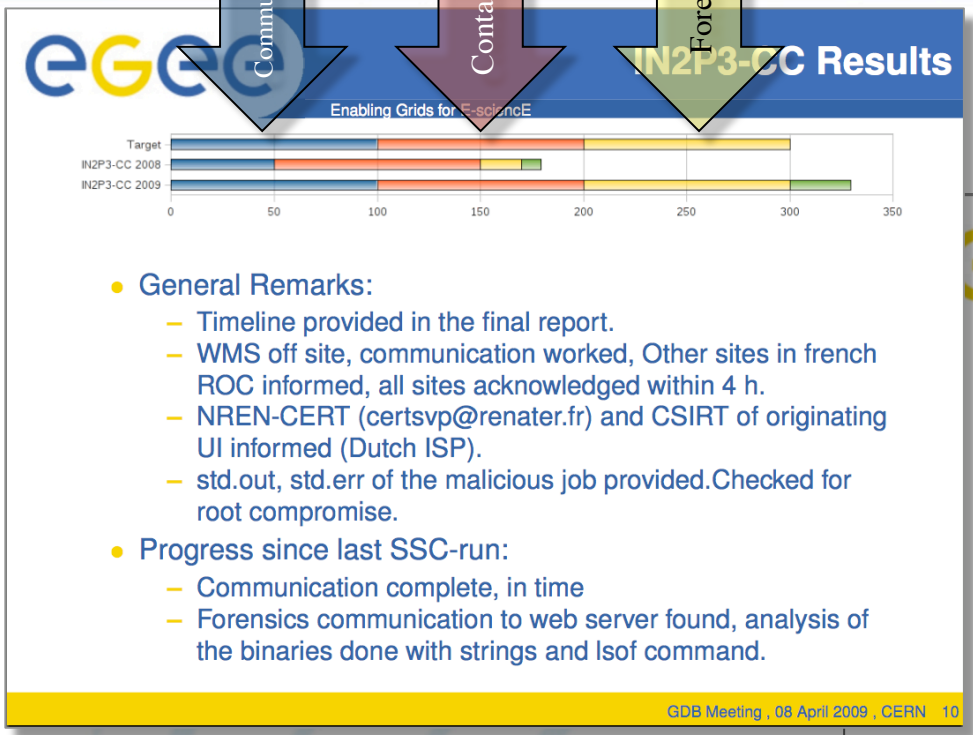
Source: Ian Bird, WLCG Workshop, Prague, 21-22/03/2009

Security Service Challenge

Communication

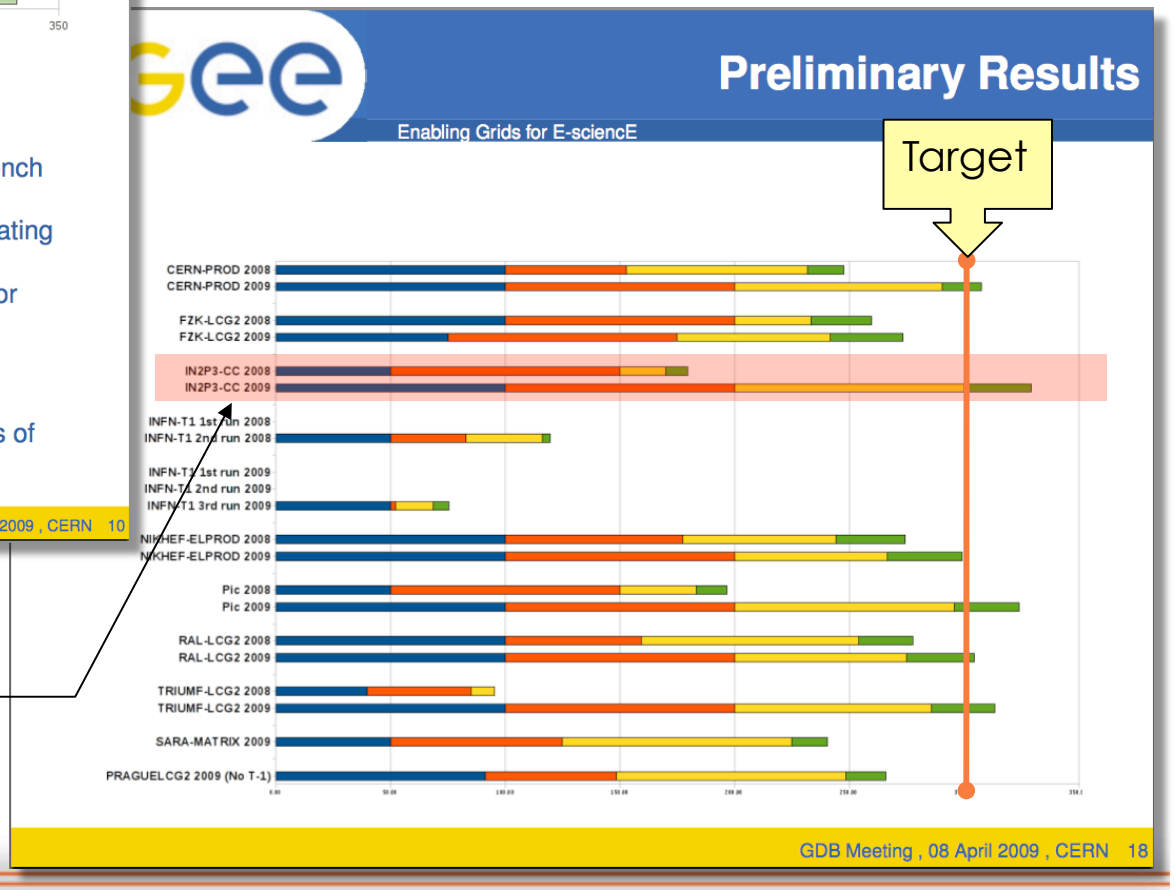
Containment

Forensics



Source: Sven Gabriel, GDB, 08/04/2009

Challenge a eu lieu le 26/02/2009



- Amélioration du résultat par rapport à l'exercice précédent, en particulier sur le volet commnication et analyse de l'attaque
- Obtention du meilleur score parmi les sites participants

SCAS tests

Enabling Grids for E-sciencE

IN2P3

- IN2P3 is currently deploying glexec. Thanks to their peculiar deployment model (multiple WN installation shared on AFS) they can allow the resources to be published as 'Production'. This enables LHCb to run there. This installation is taking longer then expected because of several issues observed in deploying glexec over the AFS nodes.
- There is an interesting thread going on between Maarten, the glexec developers and Pierre Girard. The deployment scenario of IN2P3 is however not very frequent.
- Both Atlas and LHCb are waiting to connect their submission frameworks to Lyon.

Source: Antonio Retico, GDB, 08/04/2009

More details: https://twiki.cern.ch/twiki/bin/view/LCG/PPIslandFollowUp2009x04x08

Middleware pour SL5

gLite 3.2/SL5

Enabling Grids for E-sciencE

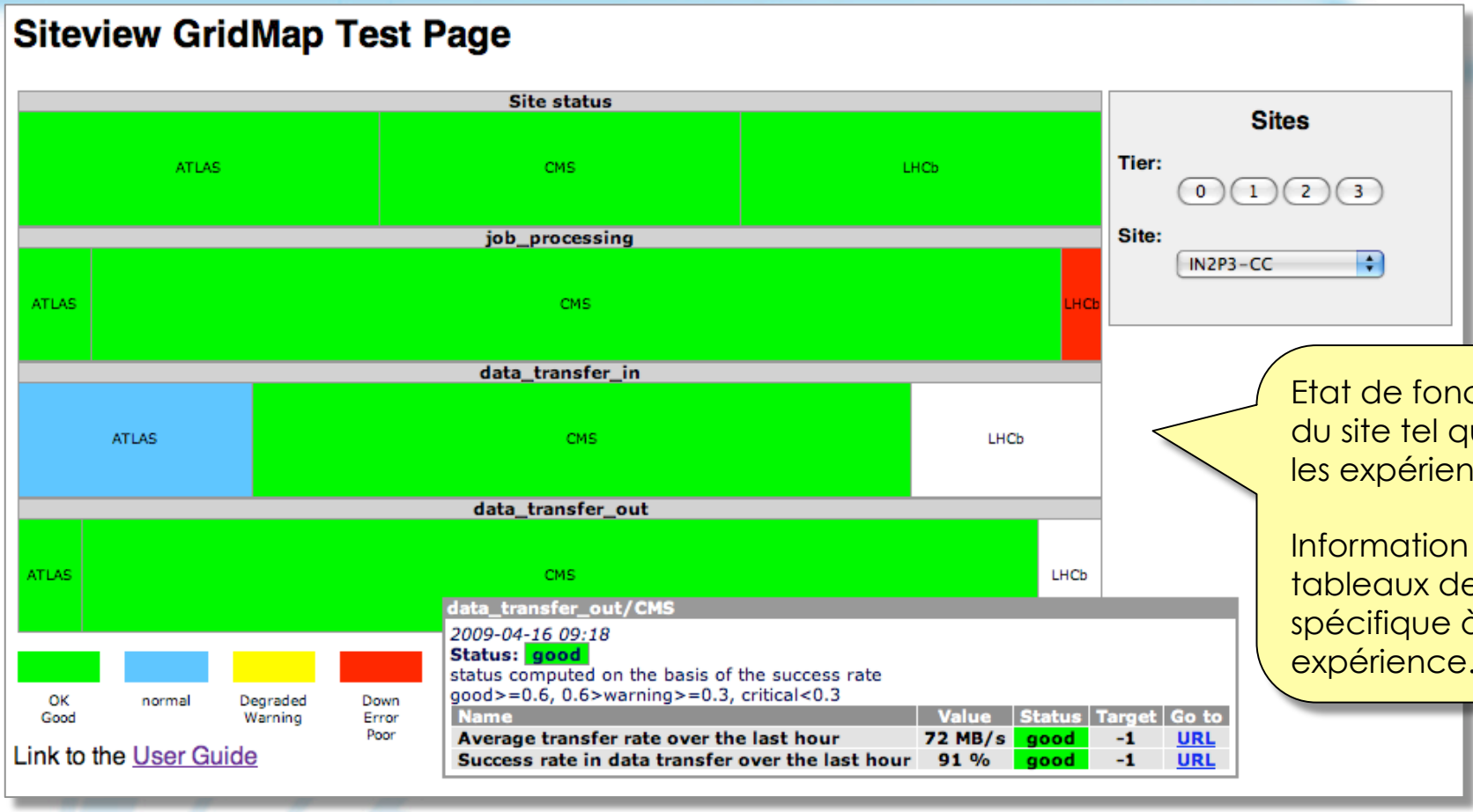
- **WN released 23. 03. 09**
  - Didn't receive any complaints so far (but not yet installed widely)
  - We know that lcg-ManageVOTag and lcg-tag don't work. Fixes already provided; patches in "Ready for Certification"
- **LCG AA area for SL5**
  - New scripts to also update on SL5 are currently being tested
- **SL5 UI**
  - Pre-certification testing during the past 2 weeks
  - Found some issues
  - Received fix for last blocker (wms-ui) yesterday
  - Patch can go to certification now
  - If no problems occur it can be released to production end of April

Enabling Grids for E-sciencE

- **DPM and LFC**
  - SL5 release tag provided by developers this week
  - DPM and LFC can now be released independently (thanks to Akos Frohner)
  - Needs an updated VDT rpm (globus gsi threading bug). Provided by VDT this week
  - We can now construct the SL5 node types
  - Release to production: starting first half of May
- **FTS/FTA/FTM**
  - Release tag not yet received
  - Needs an updated gLite trustmanager (not yet available)

Source: Andreas Unterkircher, GDB, 08/04/2009

GridMap site view



Etat de fonctionnement du site tel que perçu par les expériences.

Information extraite des tableaux de bord spécifique à chaque expérience.

Tier-1: http://dashb-siteview.cern.ch/clients/site-monitoring/test.html?site=IN2P3-CC
Tier-2: http://dashb-siteview.cern.ch/clients/site-monitoring/test.html?site=IN2P3-CC-T2

More information: Elisa Lanciotti, WLCG Workshop, Prague, 21-22/03/2009

Service Incident Reports (1/2)

Service Incident Reports: When? CERN IT Department

- Degradation goes beyond some MoU target for any service classified as critical for at least one of the VOs !
- SCOD asks for it !
- When it's useful for your own purpose
  - Tracking of incidents and the restoration → your knowledgebase for ***when*** it happens next time

Service Incident Reports: Why? CERN IT Department

- Any noticeable service outage deserves an explanation
- Choices
  - Wait for and answer the questions when they come
    - Even a well-explained event will be distorted down the line as the information is spread
    - Your mailbox is the knowledgebase
  - Upfront detailed explanation somewhere where everybody can see it
    - Including yourself when you have the same incident in the future

Source: Olof Barring, WLCG Workshop, Prague, 21-22/03/2009

Service Incident Reports (2/2)

Service Incident Report: How?

CERN IT Department

- Written report with an appropriate level of details
- Focus on:
  - What went wrong
  - Who was affected (impact)
  - How and when you noticed
  - How and when you announced
  - Main steps of the service restoration
  - When the service was restored and announced as such
  - Follow-ups/actions
- A timeline is useful
- Avoid names or other details of CIs (machines, people, …)
- Be honest
- Attempt to classify the cause
  - Change
  - Bug
  - DB
  - Human
  - The network
- In ITIL terms SIR writing activity probably closer to Problem Management

Exemples de rapports CERN/FIO: https://twiki.cern.ch/twiki/bin/view/FIOgroup/PostMortems
Exemples de rapports GridPP: http://www.gridpp.ac.uk/wiki/Incidents_TemplateV2.0
Exemples de rapports internes du CC-IN2P3: https://cctools2.in2p3.fr/operations/wiki/doku.php?id=incidents:start

Source: Olof Barring, WLCG Workshop, Prague, 21-22/03/2009

STEP'09

- STEP: **S**cale **T**esting for the **E**xperiment **P**rogram
- Expériences considèrent les sites en production
  - L'objectif de l'exercice est de valider leur capacité à tenir la charge en accord avec l'échelle de prise de données
  - Au moins ATLAS et CMS simultanément

STEP'09 -- ATLAS

What ATLAS would like to test

- Full computing model
  - Data distribution
  - Detector data re-processing
  - Monte Carlo Simulation production
  - Simulation data re-reconstruction
  - Analysis
- Tape writing and reading simultaneously in T1's and T0
- Processing priorities and shares in T1- and -2's
- Monitoring of all those activities
- Simultaneously with other experiments (test shares)
- All at nominal rates for 2 weeks
- Full shift schedule in place like for cosmics data taking
- As little disruptive as possible for detector comissioning

Source: Kors Bos, GDB, 08/04/2009

STEP'09 -- ATLAS

Data Distribution in STEP09

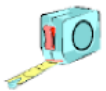
1. DDM Functional test (Simone Campana, Stephane Jezequel)
   - tests the full ATLAS data placement model including tape (RAW) writing:
     T0→ tape, T0→T1 (disk), T0→T1(tape), T1→T1 (disk),T1→T2(disk)
   - We can create "nominal" load and file sizes of all data types
   - Run at "average rate" 940 MB/sec out of CERN but for 24 hrs/day
   - Will calculate the T1 rates and volumes based on new (May C-RRB) shares
   - Also run calibration data distribution 4x T0→T2 (disk)

- Metrics
  - All T1's (including Taipei) must participate
  - No outage allowed for >12 hours
  - Allow 48 hours grace period at the end
  - 100% data must have been moved then, otherwise counted as inefficiency
  - All T2's must participate unless they sign out (before May 15)
  - No outage allowed for >24 hours

Re-processing in STEP09

1. Repeat Cosmic Ray Data Re-processing (Douglas Smith, Pavel Nevski)
   - RAW pre-staging from Tape and data access from the WN's
   - Uses conditions data tar balls
   - Output ESD's and AOD's merged and archived to tape again
   - Merged ESD's and AOD's distributed to other clouds (T1—T1's and T1→ T2's)

- Metrics
  - Site perticipation as above
  - One full re-processing during the 2 weeks exclusively in T1's
  - >99.9% files processed after grace period of 2 days
  - All successful data merged and distributed



Source: Kors Bos, GDB, 08/04/2009

STEP'09 -- ATLAS

ATLAS timing for STEP09

- Week 0 preparation May 25-31
  - Setting up all tests at lower rates
- Week 1 June 1-8
  - Real CCRC09, run all tests at full rate
- Week 2 June 8-12
  - Real STEP09, run all tests at full rate
  - Grace period June 13-14
- Week 3 June 15-19
  - Contingency
  - Reporting

- May shift 1 week up or down for other experiments

Source: Kors Bos, GDB, 08/04/2009

STEP'09 -- CMS

T1

Necessary CMS tests on T1 level:

- stress test the MSS systems at scale with all concurrent T1 workflows
- CMS workflows at T1's:
  - archive RAW+RECO+AOD
  - archive MC production from T2 level (~10s MB/s)
  - serve data to T2's (bursty)
  - re-reco
  - skimming

Tapes are involved in:

- write :: archiving custodial data (from T0, T1 or T2 level)
- read :: {RAW must be staged from tapes for re-reco}
- read :: {MC and data transfers to T2 must be staged from tape}

Daniele Bonacorsi, Oliver Gutsche 5

Calendrier provisoire:
- Phase I: 11-31 mai 2009
- Phase II: 8-21 juin 2009

T1 tape stress tests in STEP'09

Scope: find an optimal operation point for the T1 tape systems

- schedule incremental tests to be able to conduct a complete test in June:
  - test continuous import from T0/T2 *to tapes* (NOTE: already demonstrated in CCRC'08)
  - rolling re-reco while recalling data *from tape* (requires pre-staging and rolling processing tests [see next])
  - burst transfers to T2 level including possibly necessary tape recalls [see later]

Metric definition:

- a mixture of {participation + throughput + efficiency } metrics best fits here
  - transfer quality, job success rate, comparing achieved performance with reference figures, …
- measure concurrency among CMS activities
- measure overlap with other VO's
- in general, heartbeat the tape systems and how CMS uses them
  - monitor CMS tape writing backlog at T1 sites (e.g. should not go above X TB)
  - compare "pre-staging NO" and "pre-staging YES" scenarios
    - in "pre-staging YES" scenario, check how much data was accessed from disk, how much nevertheless from tape
  - monitor pre-staging success daily

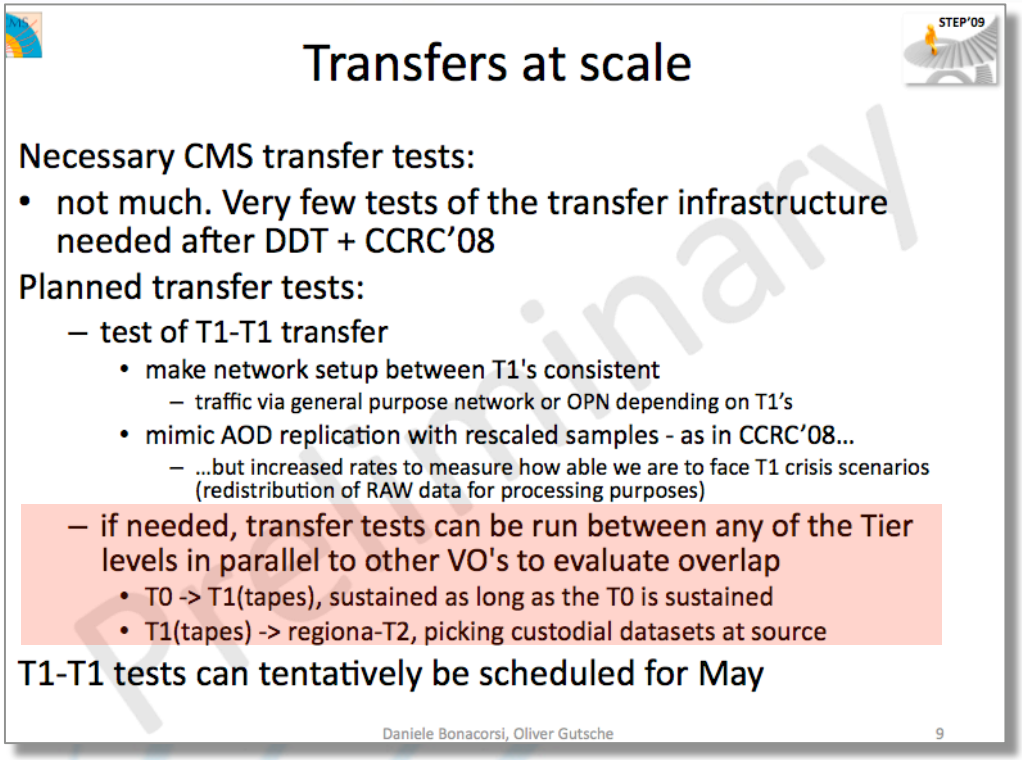
Timeline:

- Prepare in April. Tune / pre-run in May. Run in June in multi-VO mode
  - e.g. 2 weeks, from June 8th to June 21st, and can be multi-VO flexible here

Daniele Bonacorsi, Oliver Gutsche 6

Source: Daniele Bonacorsi, Oliver Gutsche, GDB, 08/04/2009

STEP'09 -- CMS



Calendrier provisoire:

- Phase I: 11-31 mai 2009 (T1⇄ T1)
- Phase II: 1-28 juin 2009

Source: Daniele Bonacorsi, Oliver Gutsche, GDB, 08/04/2009

STEP'09 -- CMS

Analysis at scale at T2's in STEP'09

Scope: increase scale of analysis at T2's to 150k-200k jobs/d

- test user stage-out scenarios at scale
- test Frontier system behavior

Metric definition:

- determine job success rates depending on number of users submitting jobs
  - emphasis on stage-out success
- monitor Frontier system for bottle necks

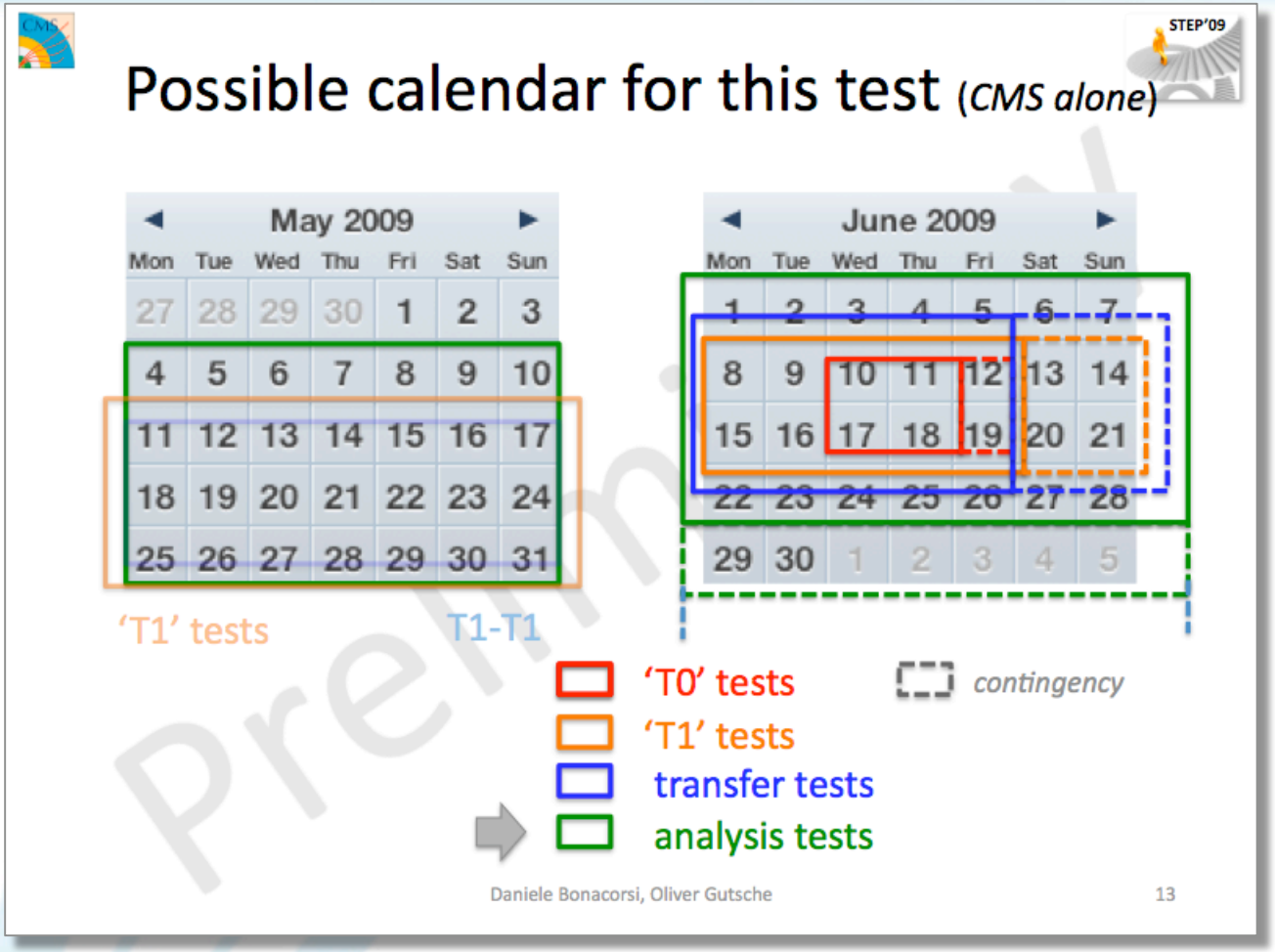
Timeline:

- Multi-VO overlap may not be critical on this particular test
  - prepare in April/May, run in May/June, may extend beyond STEP'09 as needed

Calendrier provisoire:
4 mai – 28 juin 2009

Source: Daniele Bonacorsi, Oliver Gutsche, GDB, 08/04/2009

STEP'09 -- CMS



Simultanéité avec les activités ATLAS pendant la période
25 mai – 19 juin 2009

Source: Daniele Bonacorsi, Oliver Gutsche, GDB, 08/04/2009

Chantiers récents et en cours

- Réinstallation de l'ensemble des VO boxes sur nouveau matériel
- Début de migration vers SL5 64 bits (avec compatibilité 32 bits) pour les worker nodes
  - Philippe Olivero (assisté de Nadia Lajili) coordonne l'opération
  - Détails: https://cctools2.in2p3.fr/operations/wiki/doku.php?id=migrations:migration-sl4-sl5
- Utilisation d'une nouvelle unité de puissance CPU appelée SPEC HEP 06
  - Basée sur le benchmark SPEC CPU 2006
  - Facteur de conversion: 1000 SI2000 = 4 SPEC HEP 06
  - Impact sur les engagements, données et rapports d'accounting
- Embryon d'une infrastructure pour l'analyse utilisateur finale en place
  - Détails: http://cctools2.in2p3.fr/elog/Analyse/

Chantiers récents et en cours (suite)

- Déploiement en production de FTS v2.1 sur SL4 64bits programmé pour la fin avril
  - Validation de la publication dans le système d'information, intégration avec FTS monitor et tests initiaux avec ATLAS faits
  - David Bouvet coordonne
- dCache
  - Introduction de nouveaux serveurs de disques Thors et campagne de sortie de production progressive des premiers Thumpers (acquis en 2006)
  - Disques plus capacitifs (0.5 TB ⇸1 TB): meilleur rapport Watt/GB et GB/m^2
- Ménage dans les espaces de stockage (dCache et HPSS) des expériences LHC
  - Campagne de suppression de fichiers non utiles
  - Pierre-Emmanuel coordonne

RAPPEL: axes de travail prioritaires

- Tier-1
  - Améliorer l'efficacité de la copie bande → disque, via l'ordonnancement des requêtes passées à HPSS par dCache
- Séparation claire des données tier-1 et tier-2
  - Portes d'accès différentes pour chacun des sites, sans pour autant dupliquer les données
- Déploiement d'un prototype de ferme d'analyse interactif basée sur PROOF
- Finaliser le mécanisme d'installation du logiciel des expériences dans la zone AFS
  - Utilisation de la réplication de volumes AFS pour l'équilibrage de charge
  - Documentation et généralisation pour les 4 expériences LHC
- Monitoring !!!

Evénnements passés et à venir

- WLCG Collaboration Workshop
  - Prague, 21-22/03/2009
  - http://indico.cern.ch/conferenceDisplay.py?confId=16861
- CHEP 2009
  - Prague, 23-27/03/2009
  - http://www.particle.cz/conferences/chep2009/
- Réunion semestrielle des sites LCG-France
  - Annecy, 18-19 mai 2009
  - http://indico.in2p3.fr/conferenceDisplay.py?confId=1660
- Hepix Spring 2009
  - Umeå (Suède), 25-29 mai 2009
  - http://www.hpc2n.umu.se/events/workshops/09/hepix/

Aujourd'hui et à venir

- Aujourd'hui
  - Avancement du développement de l'ordonnanceur de requêtes de tape staging
  - Bilan partiel de l'exercice de reprocessing ATLAS
  - Incidents opérationnels impactant les expériences LHC
- Prochaine réunion
  - Jeudi 14 mai, 13h30-15h30, salle 202
  - Sujets: état des CEs, déploiement de SL5 sur les workers, intégration de BQS et CREAM, précisions sur STEP'09
- Toutes les réunions à venir
  - http://indico.in2p3.fr/categoryDisplay.py?categId=102

Questions/Commentaires