WLCG

CCRC2008

Outcome of CCRC'08 Post-Mortem

Jamie.Shiers@cern.ch

~ ~ ~

LCG-LHCC Mini Review

1st July 2008

http://indico.cern.ch/conferenceDisplay.py?confId=23563

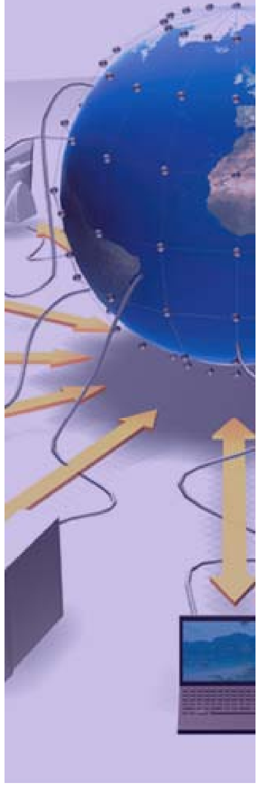
Agenda

- Reminder of goals of CCRC'08
- Summary of main (high-level) achievements
- Some details with respect to the (service) metrics
- Readiness for LHC data taking
- Outlook

CCRC08 phase I & II (ATLAS)

- **The aim of CCRC08 is to test all experiments' activities together**

- **CCRC08 Phase I:**
  - Mostly a test of SRMv2 installation/configuration
    - (functionality)
  - For ATLAS, very short exercise
    - Concurrent with FDR in week I and II
- **CCRC08 Phase II:**
  - Tests carried along for the all month
    - No overlap with FDR (1st week of June)
    - **CCRC08 ONLY during week days**
    - Cosmic data during the weekend (commissioning and M7)
  - **Focused on data distribution**
    - **T0->T1, T1->T1, T1->T2**
  - **Very demanding metrics**
    - **More than you will need to do during 2008 data taking**



How we monitor & report progress

- For CCRC'08, we have the following three sets of metrics:

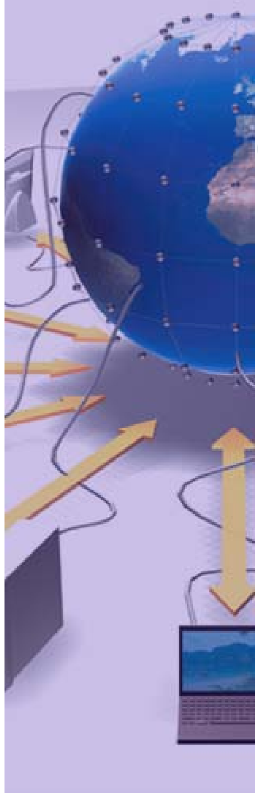
1. The **scaling factors** published by the experiments for the various **functional blocks** that will be tested. These are monitored continuously by the experiments and reported on at least weekly;

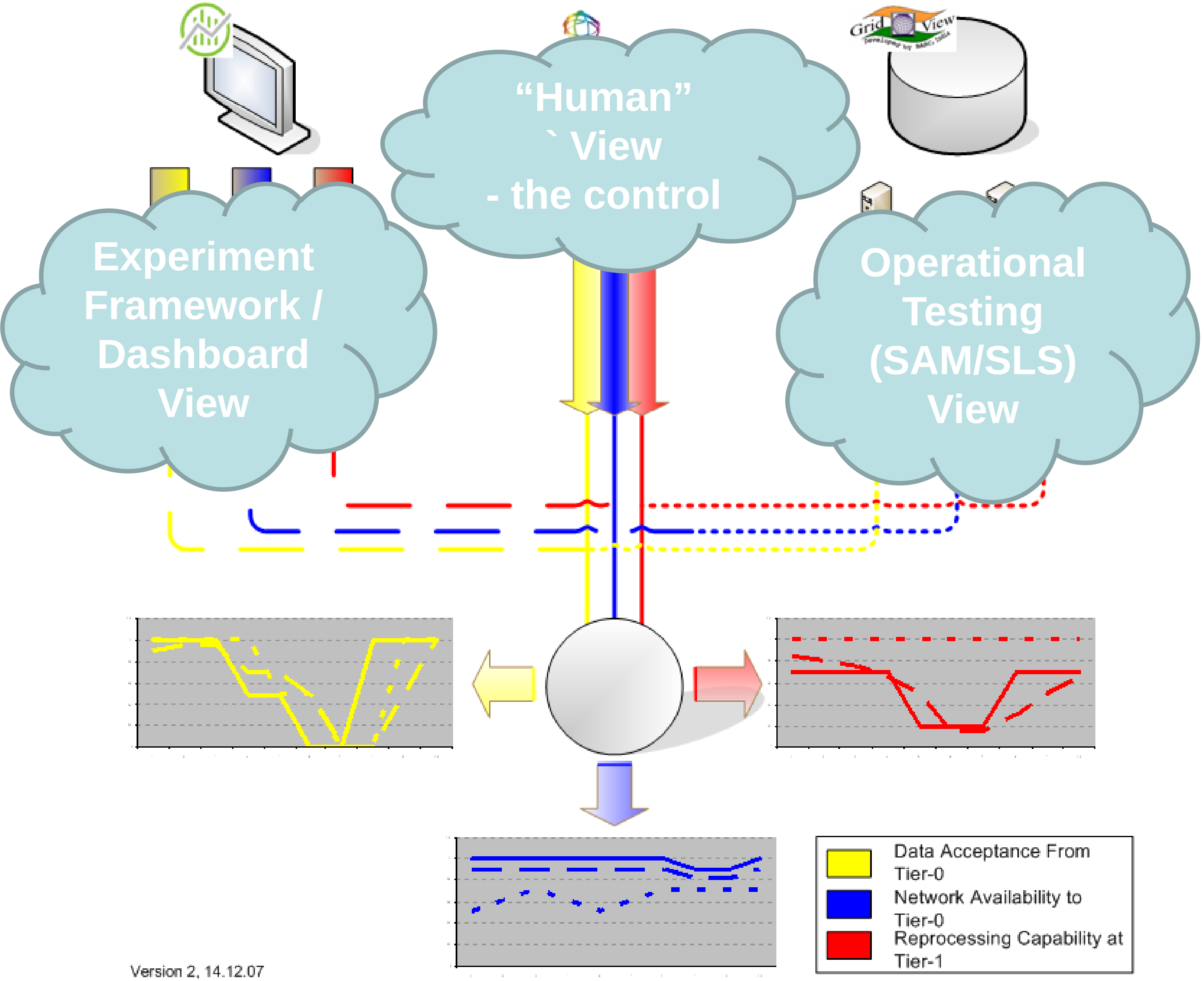
2. The lists of **Critical Services**, also defined by the experiments. These are complementary to the above and provide additional detail as well as service targets. It is a goal that all such services are handled in a standard fashion – i.e. as for other IT-supported services – with appropriate monitoring, procedures, alarms and so forth. Whilst there is no commitment to the problem-resolution targets – as short as 30 minutes in some cases – the follow-up on these services will be through the daily and weekly operations meetings;

3. The services that a site must offer and the corresponding availability targets based on the **WLCG MoU**. These will also be tracked by the operations meetings.

Measuring MoU availability

Comparing Metrics from Dashboard and SAM/Gridview against the User Experience





Experiment Framework / Dashboard View

“Human” ` View - the control

Operational Testing (SAM/SLS) View

Grid View
🟨 Data Acceptance From Tier-0
🟦 Network Availability to Tier-0
🟥 Reprocessing Capability at Tier-1

Version 2, 14.12.07

MoU Targets: Post-Mortems

- RAL power micro-cut (8.5 h downtime of CASTOR)
  - See next slide [hidden]

- NIKHEF cooling problems (4 day downtime of WNs)

- CERN CASTOR + SRM problems
  - The postmortem of the CERN-PROD SRM problems on the Saturday 24/5/2008 (morning) can be found at https://twiki.cern.ch/twiki/bin/view/FIOgroup/PostMortemMay24 . The problem affected all endpoints.
  - Problems on June 5th (5 hour downtime): https://prod-grid-logger.cern.ch/elog/Data+Operations/13

RAL Power Cut

- We lost power on one phase at about 07:00, but by the time pagers went off on-call staff were already in transit to RAL and were not able to respond until normal start of working day (which is within our 2 hour target out of hours).
- We suffered a very short (I am told milliseconds) loss of power/spike that took out one whole phase. As we have no UPS at RAL (will have in new machine room) this caused crash/reboot of over 1/3 of our disk servers.

• Restart commenced about	09:00
• CASTOR Databases Ready	12:00
• Disk Servers Ready	13:45
• Last CASTOR Instance Restarted	16:44

- So - about 2 hours because we were out of hours and had to respond and assess.
- 3 hours for ORACLE concurrent with 4:45 for clean disk server restart
- 3 hours for CASTOR restart and testing before release

➢ **Additionally, experience highlights the potential gap in the on-call system when people are in transit**

Power & Cooling – Post CCRC’08

Site	Comments
IN2P3	Had a serious problem this w/e with A/C. Had to stop about 300 WNs - waiting for action this week to repair A/C machine. Keep info posted on website.
INFN	CNAF - suffered serious problem. UPS too heavy & floor collapsed!

- IMHO, a “light-weight” post-mortem, as prepared by various sites for events during the May run of CCRC’08, should be circulated for both of these cases.
- I believe that this was in fact agreed (by the WLCG MB) during the February run of CCRC’08
- ☝ **I think we should assume that power & cooling problems are part of life and plan accordingly**

Jumping to the conclusions

- The main monitoring sources for the challenge were experiment specific monitoring tools.
- For activities at CERN (Tier0, CAF) Lemon was widely used.
- SAM and SLS were used by all experiments for monitoring of the status of the services and sites
- In general worked quite well and provided enough information to follow the challenge, to see whether the targets are met, to spot the problem rather quickly
- In most cases the problems were triggered by people on shifts using the monitoring UIs, alarms are not yet common practice.
- We do not yet have a straight forward way to show what is going on in the experiments for people external to the VO and even for users inside the VO (non-experts).
- For performance measurements except Lemon for CERN related activities and T0-T1 transfer display in GridView, nothing else was provided to show the combined picture of experiments metrics sharing the same resources.
- Sites are still a bit disoriented. They do not have clear idea how to to understand their own role/performance and whether they are serving the VOs well

Work is ongoing to address the last points

Baseline Versions for May CCRC'08

Storage-ware – CCRC'08 Versions by Implementation
CASTOR: SRM: v 1.3-21, b/e: 2.1.6-12
dCache: 1.8.0-15, p1, p2, **p3** (cumulative)
DPM: (see below)
StoRM 1.3.20

M/W component	Patch #	Status
LCG CE	Patch #1752	Released gLite 3.1 Update 20
FTS (T0)	Patch #1740	Released gLite 3.0 Update 42
FTS (T1)	Patch #1671	Released gLite 3.0 Update 41
FTM	Patch #1458	Released gLite 3.1. Update 10
gFAL/lcg_utils	Patch #1738	Released gLite 3.1 Update 20
DPM 1.6.7-4	Patch #1706	Released gLite 3.1 Update 18

The Storage Solution WG

- The **goal of the SSWG** is
  Address issues uncovered through the challenges and provide timely solutions
- **This is achieved with**:

Management Board of June 17th concluded that priority is production & short-term fixes – work on longer-term features in SRM v2.2 MoU addendum delayed until after experience from 2008 data taking.

- Detailed discussions on experiences in CCRC'08 will take place during the workshop.
  - This includes release / patch handling, dependencies between different components etc.

(Achilles') Heel # 1 – Storage-ware

- The storage services are still somewhat unstable and there are repeated complaints that it is not clear exactly which versions, patch levels, configurations etc are required

- This information exists and is discussed regularly but is probably not well summarized / easily accessible

- My proposal is that the necessary information is summarized on a weekly basis on the joint EGEE – OSG – WLCG operations meeting in a table, e.g.

Implementation	**Version (Release/Patch)**	**Comments (or URL)**
dCache	**1.8.0-15 p6**	http://trac.dcache.org/trac.cgi/report/18

Storage Versions – Present & Future

Component	Version	Comments
CASTOR core	2.1.7-10	will be released this week Tier1s are recommended to upgrade faranno l'upgrade verso meta' Luglio
	2.1.8	will be released the first week of August - Tier0 will upgrade before the end of August - Tier1 will follow
CASTOR SRM	1.3-27 on SLC3	2.7-1 on SLC4 as soon as released
dCache	1.8.0-15p6	fixes a bug with caching credential produced through grid-proxy-init
	1.8.0-15p7	is about to come out. It fixes a problem with checksum verification when copy a file in push mode between 2 dCache sites
StoRM	1.3.20 on SLC4	
DPM	1.6.10 on SLC4	

Middleware Summary

- The software process operated **well!**
  - No special treatment for CCRC
  - Priorities are updated twice a week in the EMT
- 4 Updates to gLite 3.1 on 32bit
  - About 20 Patches
- 2 Updates to gLite 3.1 on 64bit
  - About 4 Patches
- 1 Update to gLite 3.0 on SL3

- During CCRC we
  - Introduced new services
  - Handled security issues
  - Produced the regular stream of updates
  - Responded to CCRC specific issues

Summary of DBs in CCRC’08

- Distributed database infrastructure is ready for accelerator turn-on
  - Smooth running during CCRC’08
  - Experiments are ramping-up to full use of the Tier1 infrastructure
  - Minor issues found and all being followed-up
- Oracle Data Guard for “critical DBs” at Tier0 during CCRC’08 worked well
  - Need a more defined plan if this becomes a request from experiments and WLCG
- “DB dashboard” is a key tool for the application developers and DB resource coordinators
  - Well appreciated by our users
  - Would like to extend it to the Tier1 sites, picking up the recent developments from ATLAS
- Reminder: 24x7 on “best effort”

Physics DB Conclusions **Worked well!**

nent

- Recognizing the importance DB services to the experiments' activities, we have built up robust, scalable and flexible solutions

- These solutions successfully address a wide-range of use cases

- Testing and validation – hardware, DB versions, applications – proven key to smooth production

- Many years of close cooperation between application developers and database administrators have resulted in reliable, manageable services

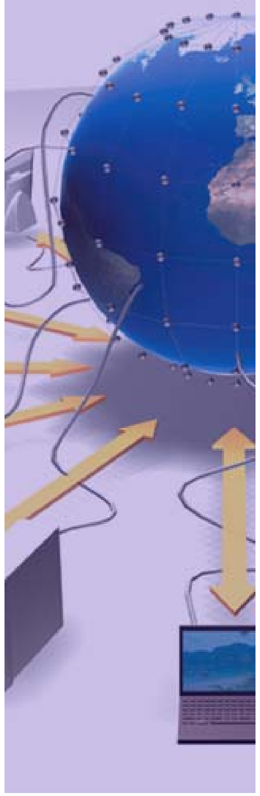
CCRC '08 – Areas of Opportunity

- Tier2s: MC well run in, distributed analysis still to be scaled up to (much) larger numbers of users
- Tier1s: data transfers (T0-T1, T1-T1, T1-T2, T2-T1) now well debugged and working sufficiently well (most of the time...); reprocessing still needs to be fully demonstrated for ATLAS (includes conditions!!!)

➢ **Tier0: best reviewed in terms of the experiments' "Critical Services" lists**

  - These **strongly emphasize** data/storage management and database services!
  - **We know how to run stable, reliable services**
  - IMHO – these take **less** effort to run than 'unreliable' ones...
  - ➢ **But they require some minimum amount of discipline...**

ATLAS Conclusions



- **The data distribution scenario has been tested well beyond the use case for 2008 data taking**
- **The WLCG infrastructure met the experiments' requirements for the CCRC08 test cases**
- **Human attention will always be needed**

- Activity should not stop
  - ATLAS from now on will run continuous "heartbeat" transfer exercise to keep the system alive

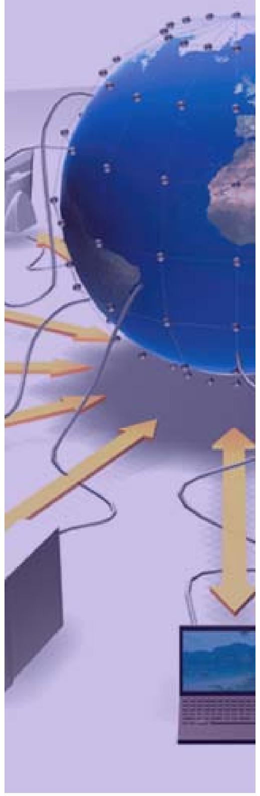
Summary

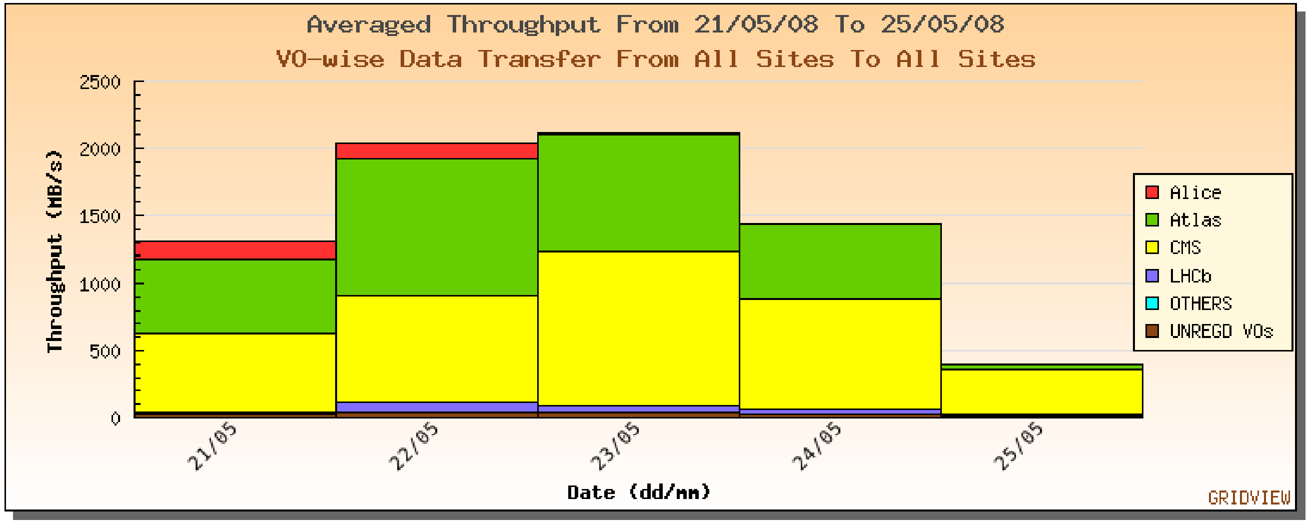
- Data transfer of CCRC08 using FTS was successful
- Still plagued with many issues associated data access
  - Issues improved since Feb CCRC08 but…
  - 2 sites problematic for large chunks of CCRC08 - 50% of LHCb resources!!
  - Problems mainly associated with access with dCache
  - Commencing tests with xrootd
- DIRAC3 tools improved significantly from Feb
  - Still need improved reporting of problems
- LHCb bk-keeping remains a major concern
  - New version due prior to data taking
- LHCb need to implement a better interrogation of log files

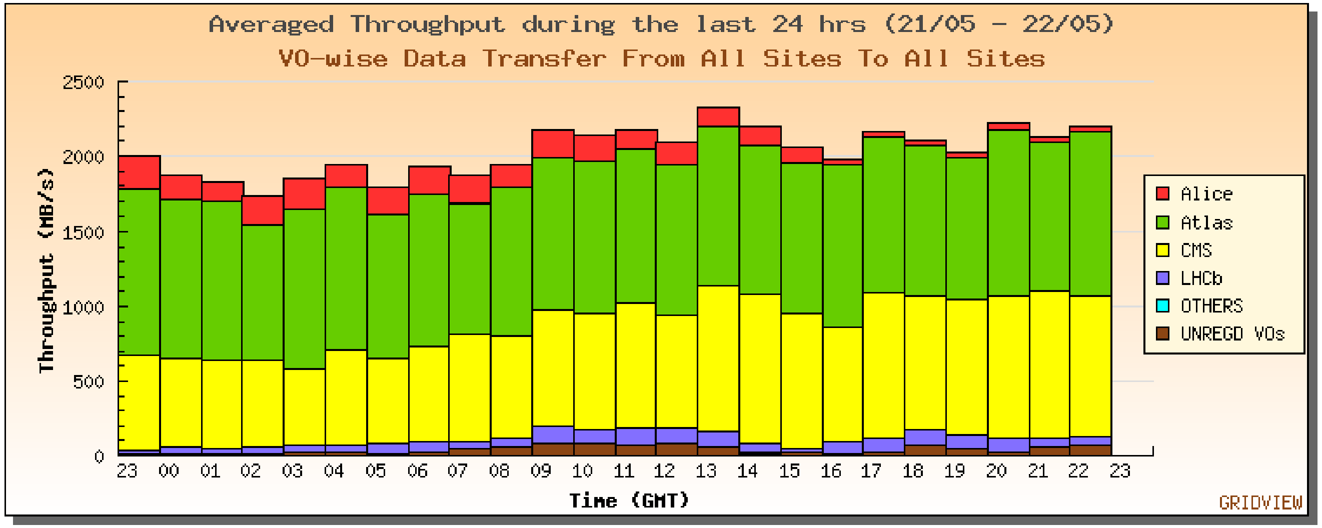
GS

Week-3: all experiments

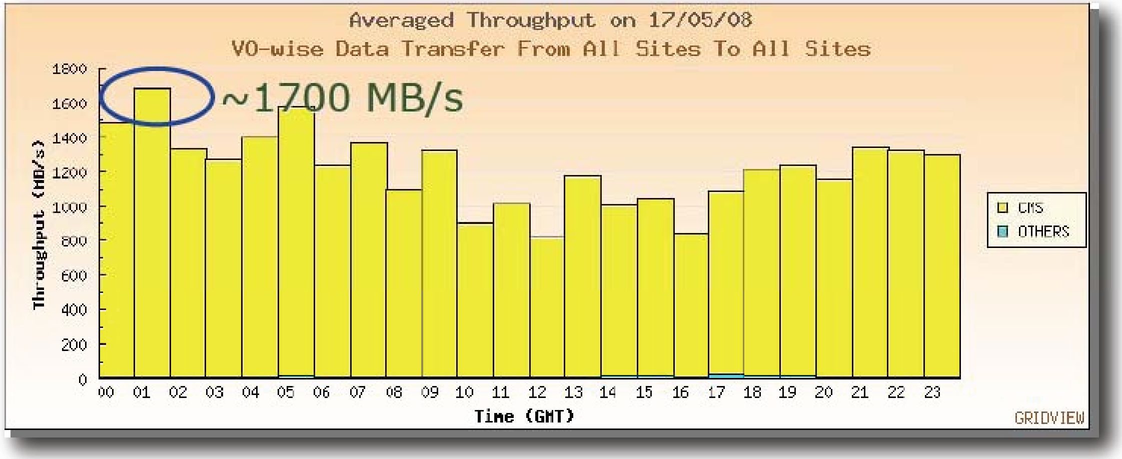
CERN IT Department



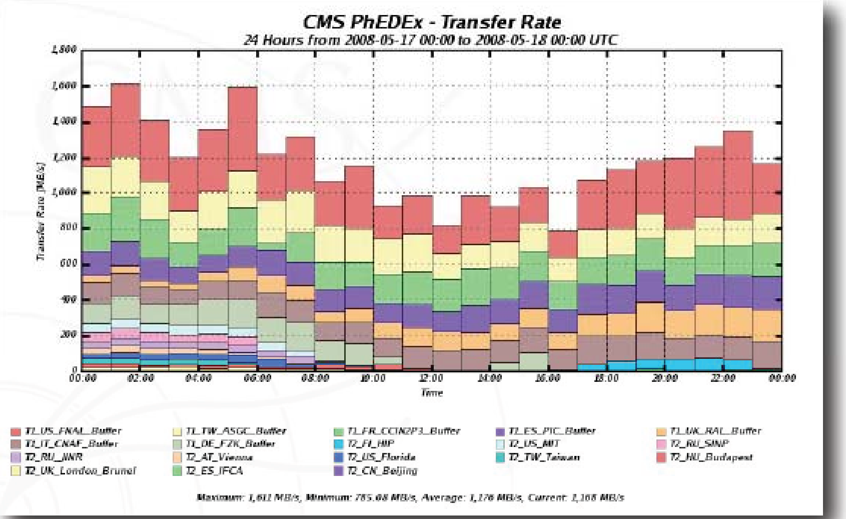




T0→T1: hourly zoom

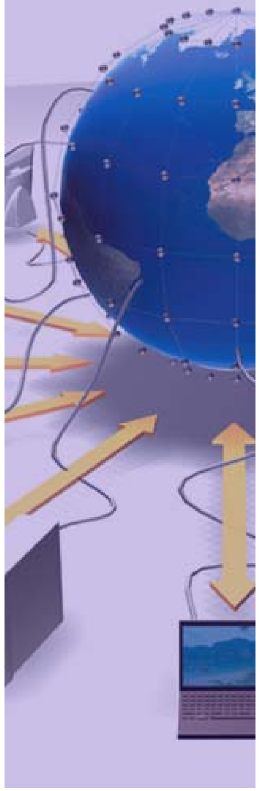


Remarkably high (hourly) rate out of CERN



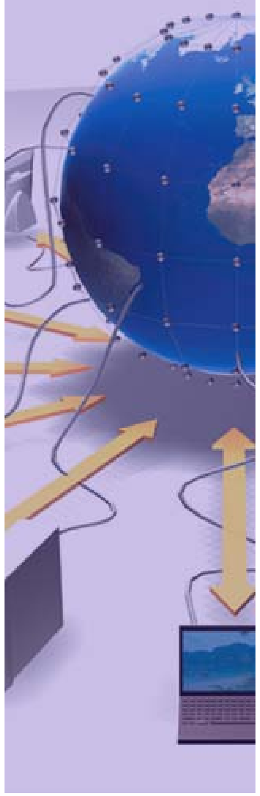
GS

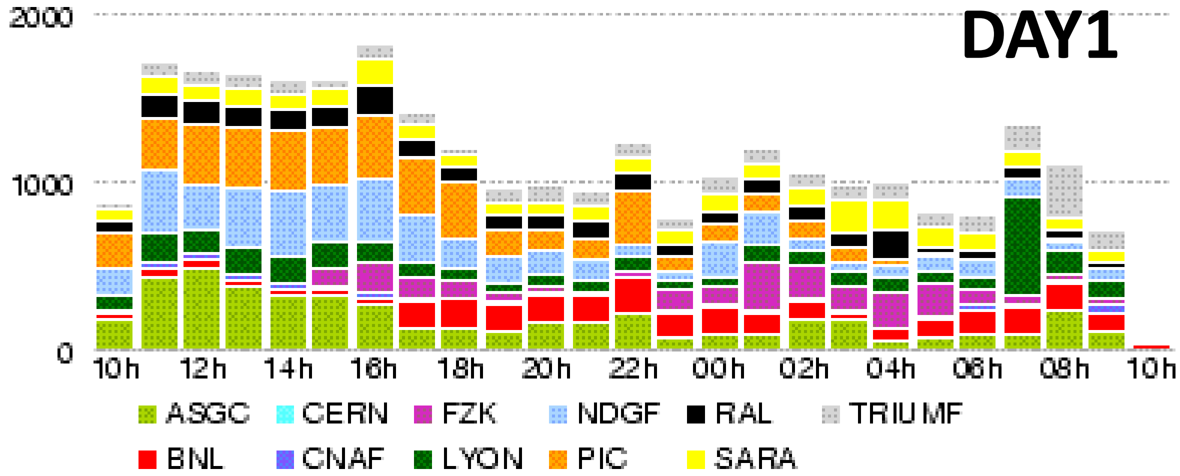
Week-2: T1-T1 test



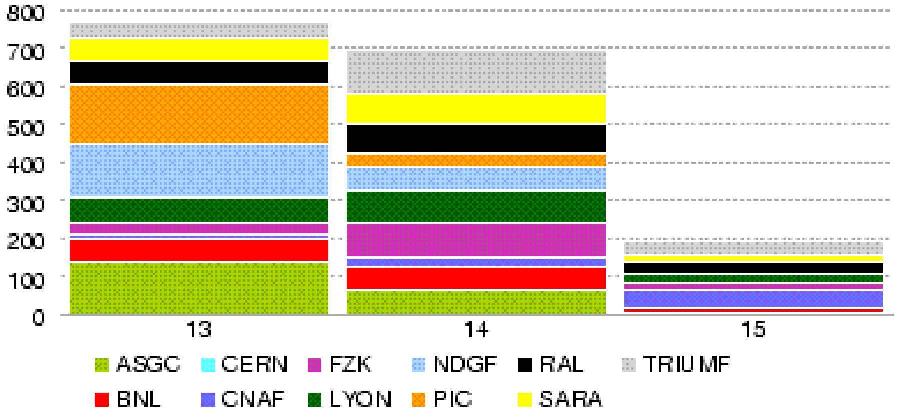
- Replicate ESD of week 1 from “hosting T1” to all other T1s.
  - Test of the full T1-T1 transfer matrix
  - FTS at destination site schedules the transfer
  - Source site is always specified/imposed
    - No chaotic T1-T1 replication … not in the ATLAS model.
- Concurrent T1-T1 exercise from CMS
  - Agreed in advance

Week-2: Results

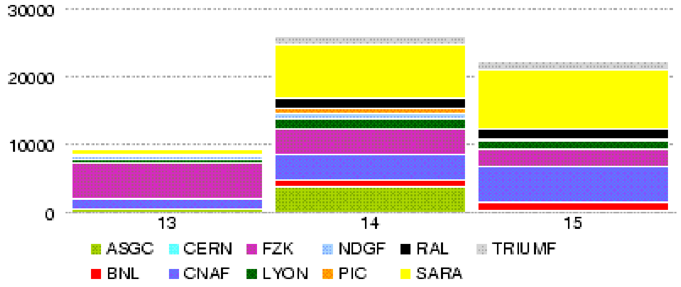




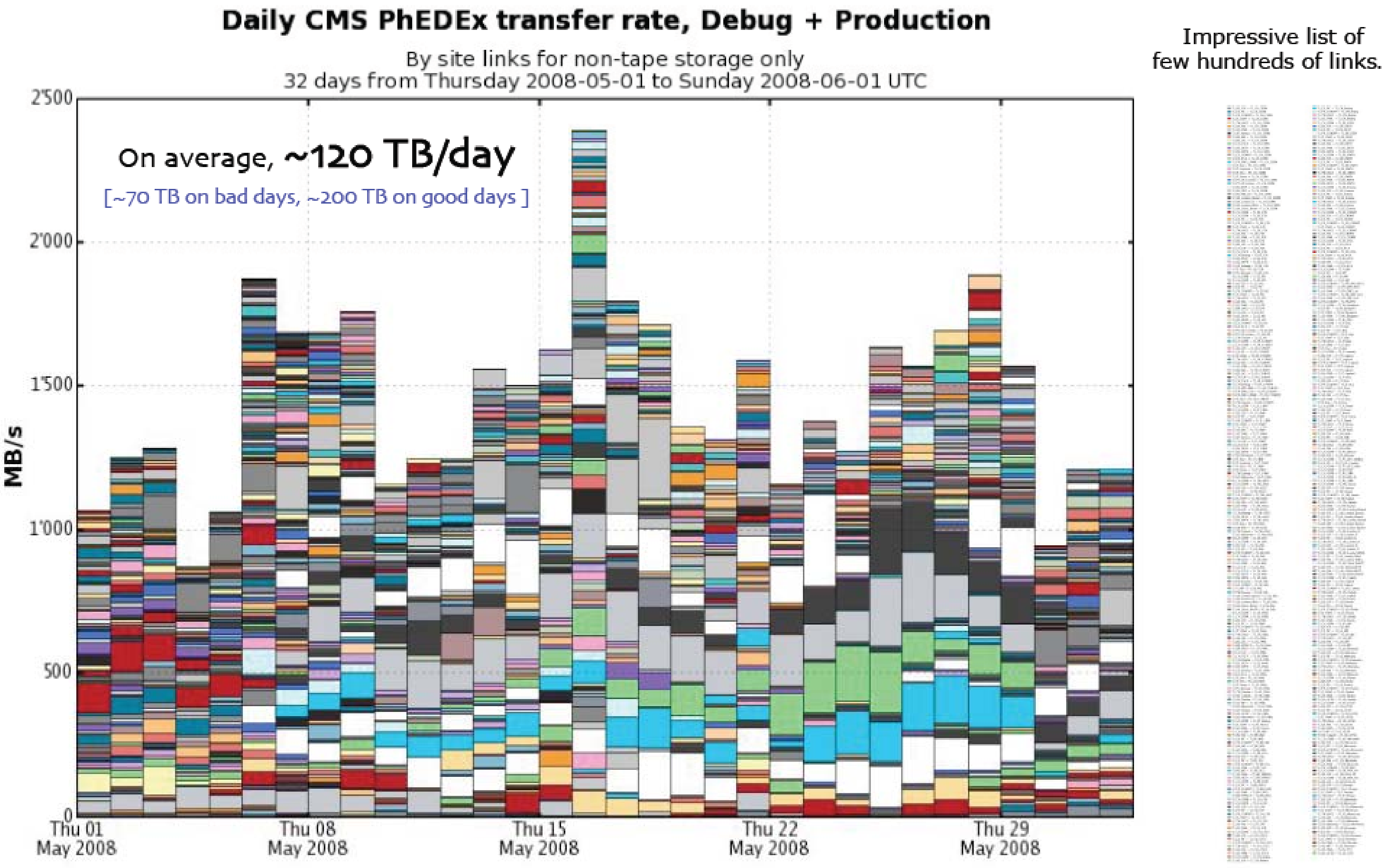
All days (throughput)



All days (errors)



Tier-x to Tier-x in CCRC'08/phase-2



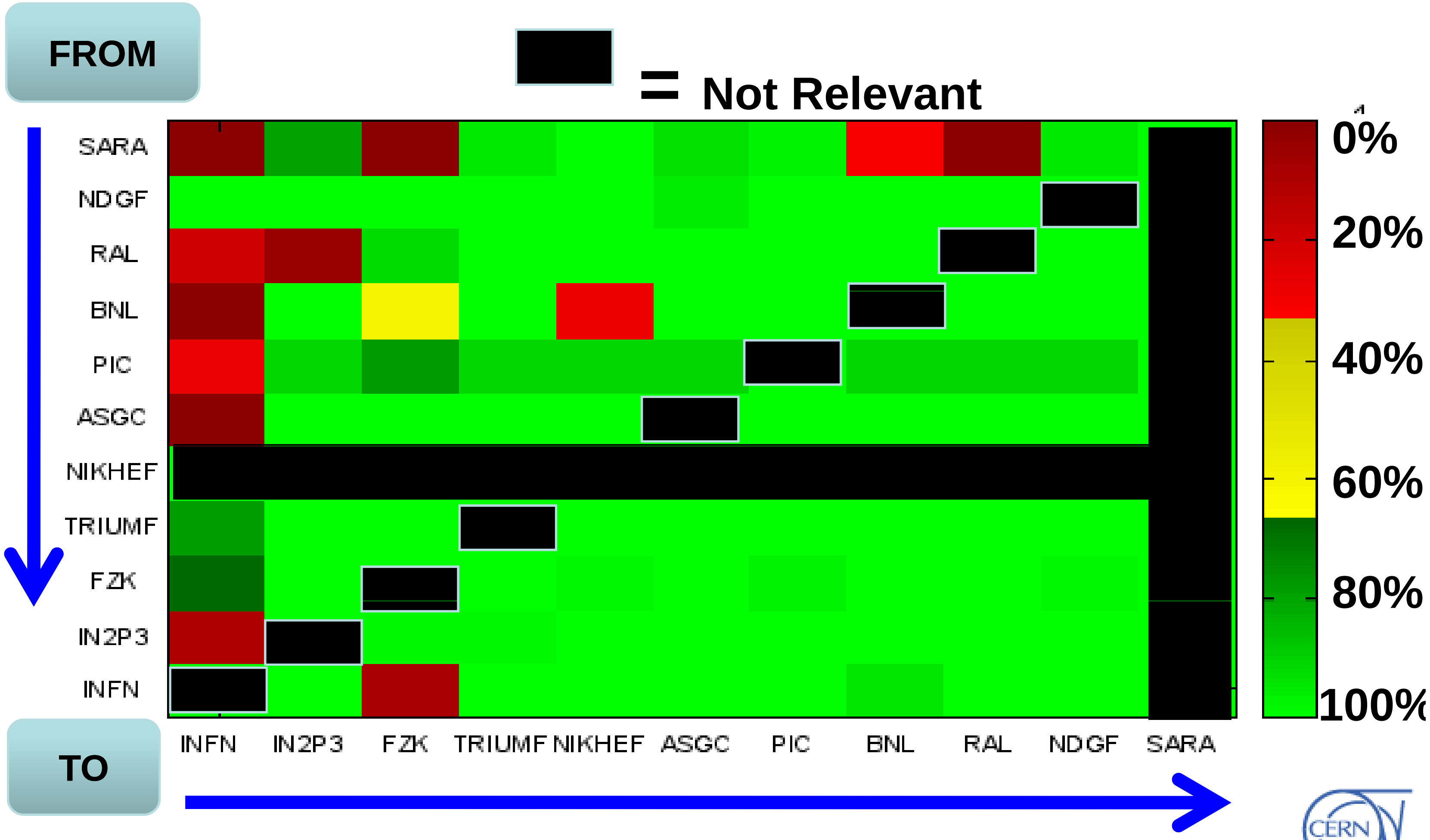
On average, **~120 TB/day**

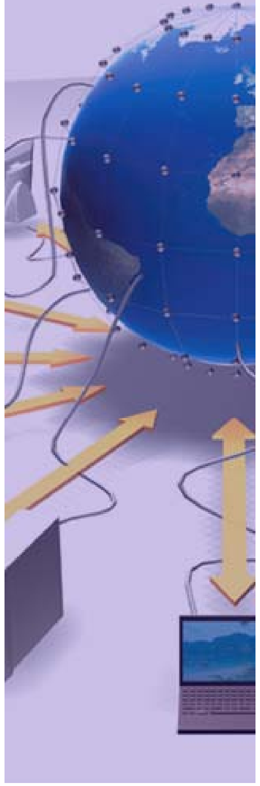
[~70 TB on bad days, ~200 TB on good days]

Impressive list of few hundreds of links...

Week-2: Results

Fraction of completed dataset

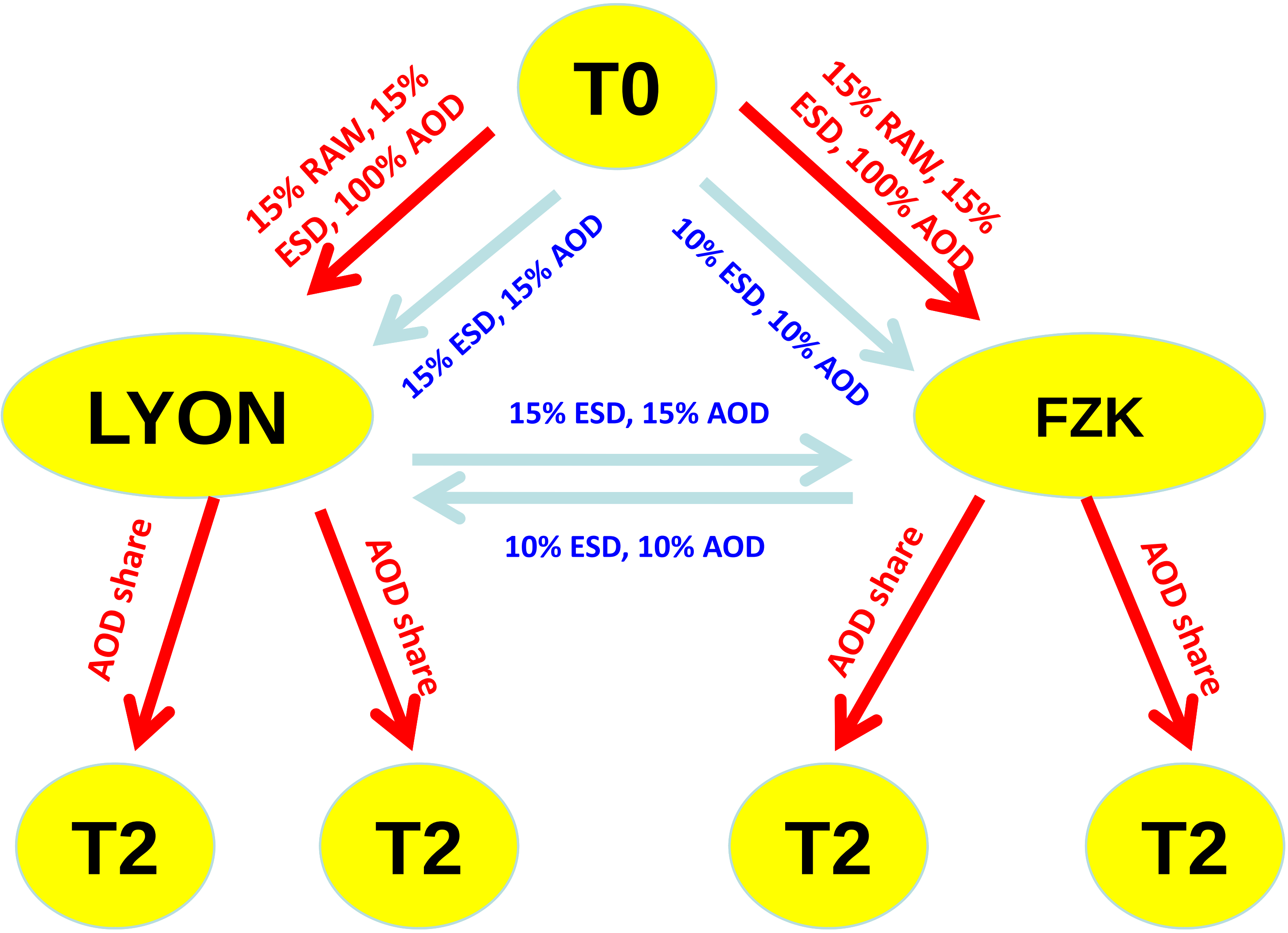




Week-4: setup

Load Generator at 100%, NO RAW

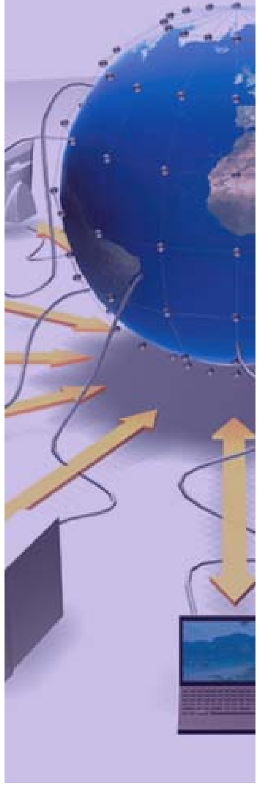
Load Generator at 100%

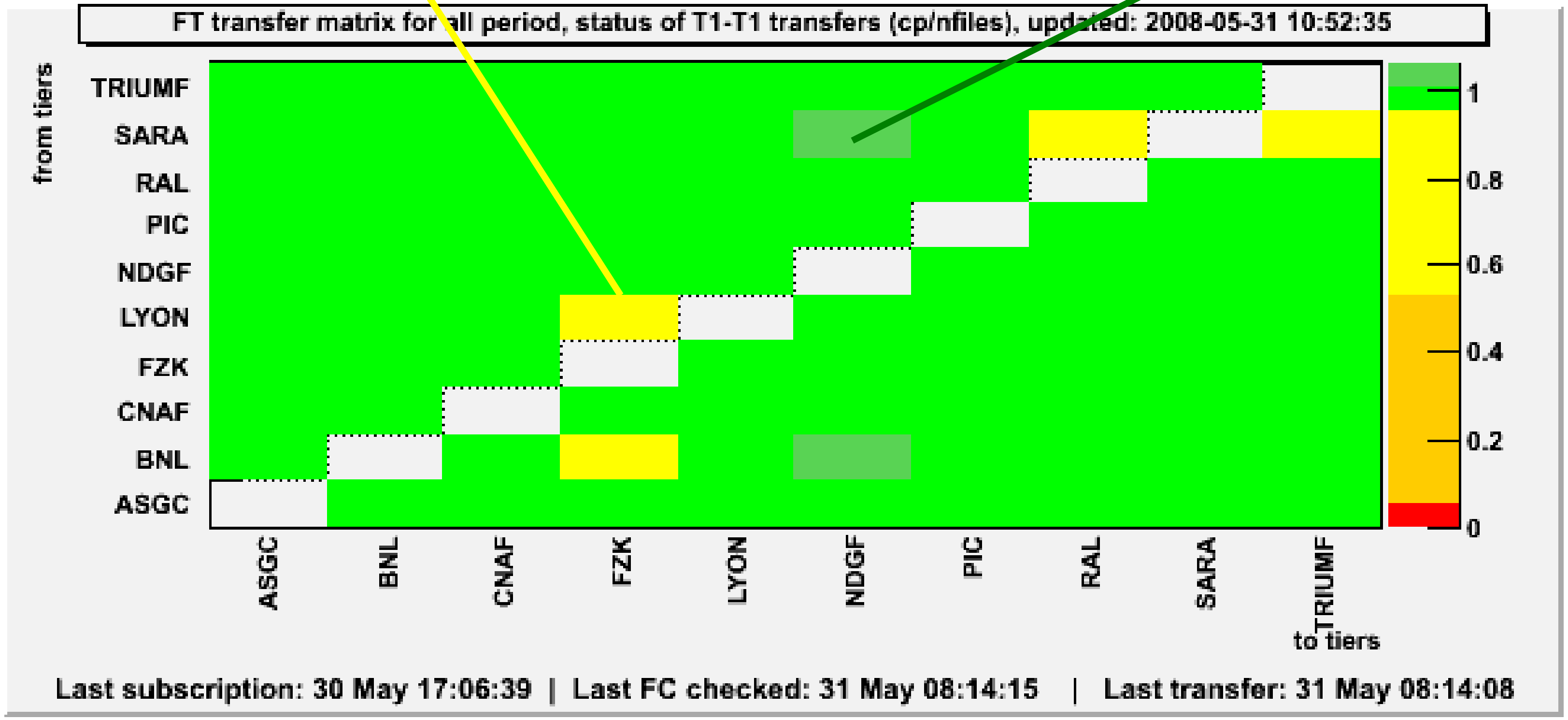


Week-4: T1-T1 transfer matrix

YELLOW boxes
Effect of the power-cut

DARK GREEN boxes
Double Registration problem

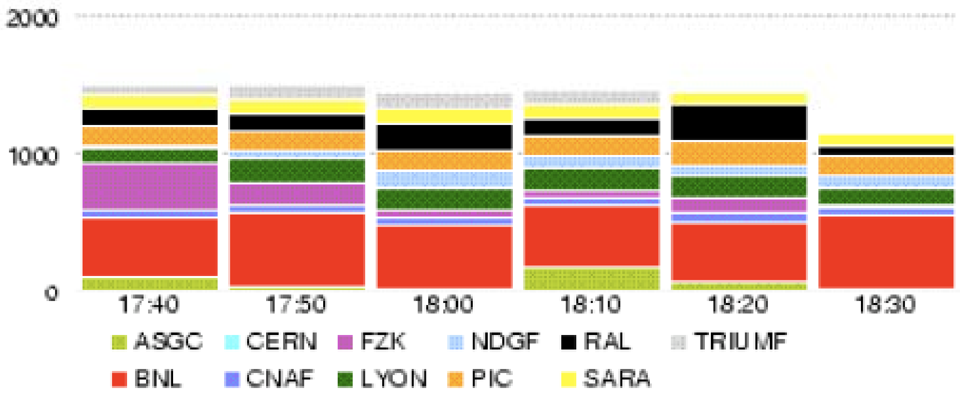




Last subscription: 30 May 17:06:39 | Last FC checked: 31 May 08:14:15 | Last transfer: 31 May 08:14:08

Compare with week-2 (3 problematic sites)
Very good improvement

Throughput (MB/s)

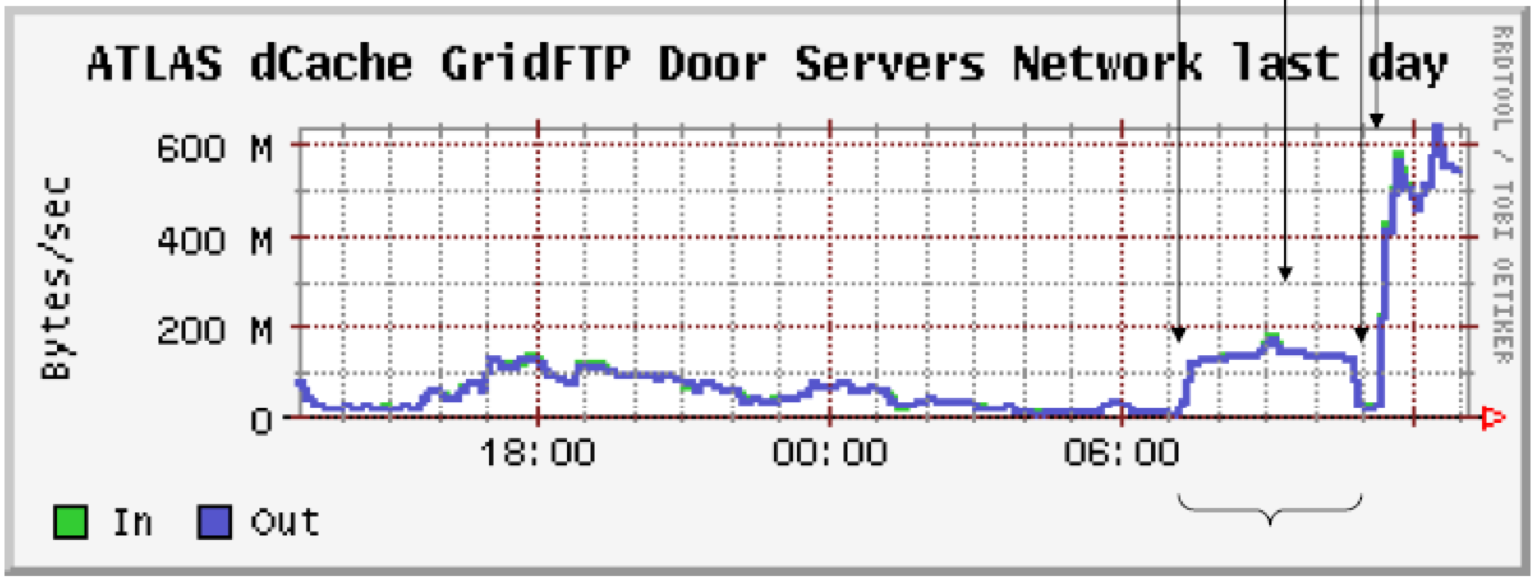


Start of ATLAS CCRC Transfers to BNL

Castor outage at CERN

Problem solved by ESnet

Problem reported to ESnet



Bandwidth limited to 1 Gbps
Due to configuration problem at ESnet

CCRC ’08 – How Does it Work?

- Experiment “shifters” use Dashboards, experiment-specific SAM-tests (+ other monitoring, e.g. PhEDEx) to monitor the various production activities
- Problems spotted are reported through the agreed channels (ticket + elog entry)
- Response is usually rather rapid – many problems are fixed in (<)< 1 hour!
- A small number of problems are raised at the daily (15:00) WLCG operations meeting

☞ **Basically, this works!**

- We review on a weekly basis if problems were not spotted by the above → fix **[+ MB report]**

➢ **With time, increase automation, decrease eye-balling**

Are We Ready?

- Based on the experience in the February and May runs of the Common Computing Readiness Challenge, an obvious question is

¿ "Are we ready for LHC Data Taking?"

  - **any time mid-July 2008 on…**

- The honest answer is probably:

- "We are ready to **face** LHC Data Taking"

- There are subtle, but important, differences between the two…

On WLCG Readiness

- The service runs smoothly – **most of the time**
- Problems are typically handled rather rapidly, with a decreasing number that require escalation
- We have a well-proven “**Service Model**” that allows us to handle anything from “**Steady State**” to “**Crisis**” situations
- We have repeatedly proven that we can – typically rather rapidly – work through even the most challenging “**Crisis Situation**”
- Typically, this involves short-term work-arounds with longer term solutions

☝ **It is essential that we all follow the “rules” (rather soft…) of this service model which has proven so effective…**

Heel # 2 – The Service Itself

- The “WLCG Service” is (highly) complex and there are many inter-dependencies and couplings

- The number of (major) service interventions per week and their scheduling is limited by human resources – and our ability to communicate needed information about the various dependencies

- The number of interventions in June was IMHO too high – O(1) per day (more?).
  - **1 – or perhaps 2 – per week is manageable**

- As prior to the May run of CCRC’08, some of these interventions have not been fully discussed beforehand
  - e.g. no LCG SCM due to lack of time / clashes with F2F meetings and workshops

- Some components – e.g. VOMS & friends & GridView – are still not able to handle some of the basic recovery required for a **SERVICE**
  - e.g. Gracefully recovering when DB comes back (scheduled or unscheduled)

Common Misconceptions –
Even amongst Grid Experts!

- Talking of a middleware component / release / distribution as if it were as **service**

- Believing / assuming that service reliability can be added ***post facto***

- Thinking that robust and reliable services take more **manpower** than crummy ones – or are more expensive

- That more ***monitoring*** (alone) makes services more reliable...

- “The Grid” certainly adds significant extra complexity, as per Ian Foster’s 3 laws of **Gridability...**

Ticklist for a new service

- User support procedures (GGUS)
  - Troubleshooting guides + FAQs
  - User guides
- Operations Team Training
  - Site admins
  - CIC personnel
  - GGUS personnel
- Monitoring
  - Service status reporting
  - Performance data
- Accounting
  - Usage data
- Service Parameters
  - Scope - Global/Local/Regional
  - SLAs
  - Impact of service outage
  - Security implications
- Contact Info
  - Developers
  - Support Contact
  - Escalation procedure to developers
- Interoperation
  - ???

- **First level support procedures**
  - **How to start/stop/restart service**
  - **How to check it's up**
  - **Which logs are useful to send to CIC/Developers**
    - **and where they are**
- **SFT Tests**
  - **Client validation;**
  - **Server validation**
  - **Procedure to analyse these**
    - **error messages and likely causes**
- **Tools for CIC to spot problems**
  - **GIIS monitor validation rules (e.g. only one "global" component)**
  - **Definition of normal behaviour**
    - **Metrics**
- **CIC Dashboard**
  - **Alarms**
- **Deployment Info**
  - **RPM list**
  - **Configuration details (for yaim)**
  - **Security audit**

In a Nutshell…

Services	
ALL	WLCG / “Grid” standards
KEY PRODUCTION SERVICES	+ Expert call-out by operator
CASTOR/Physics DBs/Grid Data Management	+ 24 x 7 on-call

- **Escalation almost never needed…**
- **Therefore, it is to be “call 911” simple…**

☺ At CERN it is – call x5011!

Overall Conclusions

- All of the things tested in CCRC’08 went as well – ***or better*** – than expected

☺ **Some even went better than planned**

- Some key elements not tested – e.g. ATLAS reprocessing
- Storage is still a key issue – and likely to remain so, at least at the Tier0 and Tier1s – for some time

➢ **Priorities are bound to change when data arrives**

- We have shown that we are ready under controlled test conditions

☛ **Can we handle the stress of data from collisions?**

CCRC ’09 - Outlook

- SL(C)5
- CREAM
- Oracle 11g
- SRM v2.2++
- **Other DM fixes…**
- SCAS
- [new authorization framework]
- …

- 2009 resources
- 2009 hardware
- Revisions to Computing Models
- EGEE III transitioning to more distributed operations
- Continued commissioning, 7+7 TeV, transitioning to normal(?) data-taking (albeit low luminosity?)
- New DG, …

CCRC’08 – Conclusions (LHCC referees)

- The **WLCG** service is running (reasonably) **smoothly**

- The functionality **matches:** what has been tested so far – and what is (known to be) required

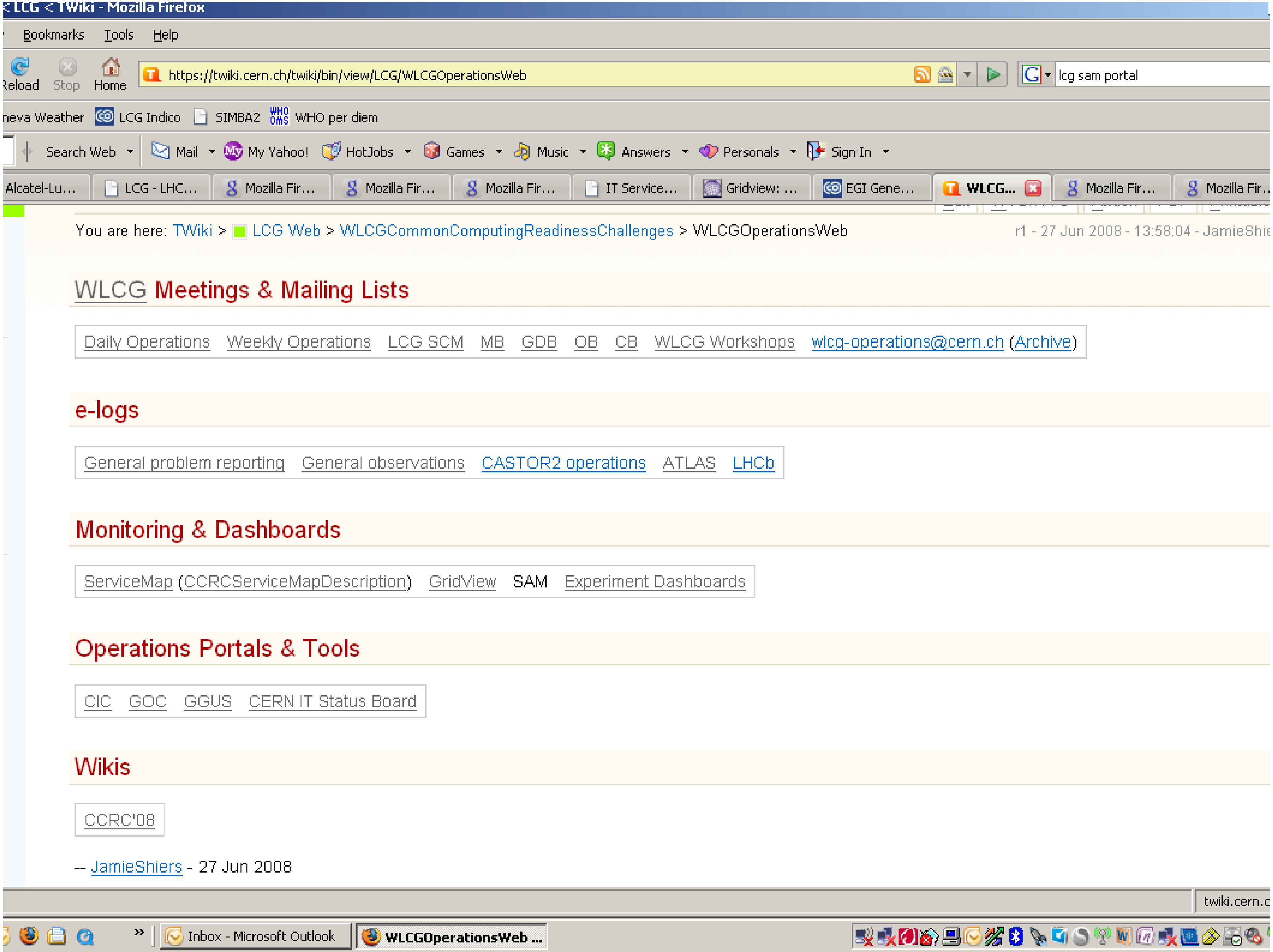
☞ **We have a good baseline on which to build**

- (Big) **improvements** over the past year are a good indication of what can be expected over the next!

- (Very) detailed **analysis** of results compared to up-front metrics – in particular from experiments!

CCRC’08 – July Phase

- A “July Phase” of CCRC’08 has been mentioned a number of times – but was not part of the original proposal in September
- Goal: test any remaining Use Cases not exercised fully in Feb / May and / or demonstrate resolution of any problems encountered – e.g. reprocessing(!)
- Additional motivation – keep exercising system in the case of no “real” (i.e. from pp collisions...) data
- ***Many*** changes have taken place since May – service upgrades, on-going hardware replacements, deployment of additional resources, ...
- No formal plan has (yet) been proposed for July, other that continued running against existing metrics with as much overlap of all experiments as possible!


LCG < TWiki - Mozilla Firefox
Bookmarks Tools Help
Reload Stop Home
https://twiki.cern.ch/twiki/bin/view/LCG/WLCGOperationsWeb
lcg sam portal
neva Weather LCG Indico SIMBA2 WHO per diem
Search Web Mail My Yahoo! HotJobs Games Music Answers Personals Sign In
Alcatel-Lu... LCG - LHC... Mozilla Fir... Mozilla Fir... Mozilla Fir... IT Service... Gridview: ... EGI Gene... WLCG... Mozilla Fir... Mozilla Fir...
twiki.cern.c
Inbox - Microsoft Outlook
WLCGOperationsWeb ...

You are here: TWiki > LCG Web > WLCGCommonComputingReadinessChallenges > WLCGOperationsWeb r1 - 27 Jun 2008 - 13:58:04 - JamieShie

WLCG Meetings & Mailing Lists

Daily Operations Weekly Operations LCG SCM MB GDB OB CB WLCG Workshops wlcg-operations@cern.ch (Archive)

e-logs

General problem reporting General observations CASTOR2 operations ATLAS LHCb

Monitoring & Dashboards

ServiceMap (CCRCServiceMapDescription) GridView SAM Experiment Dashboards

Operations Portals & Tools

CIC GOC GGUS CERN IT Status Board

Wikis

CCRC'08

-- JamieShiers - 27 Jun 2008

Summary

- Both official phases of CCRC'08 (Feb & May) have been largely successful in achieving their goals
- However, the overlap between the experiments was less than optimal and some important aspects of the Computing Models were not fully tested even in May
- The targets set were – largely speaking – well above what is likely during 2008 data taking
- Service is working relatively smoothly – continued attention and improvements are still needed, particularly in the key area of storage
- Production activities continue – and even the name!

Post Script

The service is still the challenge…