WLCG Production Services using EGEE Infrastructure

Jamie.Shiers@cern.ch

Grid Operations Workshop

Stockholm, June 2007

Agenda

- WLCG service operation & MoU targets
- WLCG Service Coordination roles
- S.W.O.T. analysis of WLCG service
- LHC startup challenges – we're not there yet!

Agenda

- **WLCG service operation & MoU targets**
- WLCG Service Coordination roles
- S.W.O.T. analysis of WLCG service
- LHC startup challenges – we're not there yet!

Background

- 100% of my **Grid** experience relates to the deployment and delivery of **Production Services**
- This started already in the days of EDG with the Replica Location Service and its deployment at CERN and some key (WLCG) Tier1 sites
- In the then-current WLCG Computing Model, the EDG-RLS was a critical component which, if unavailable, meant:

➢ **Running jobs could not access existing data**

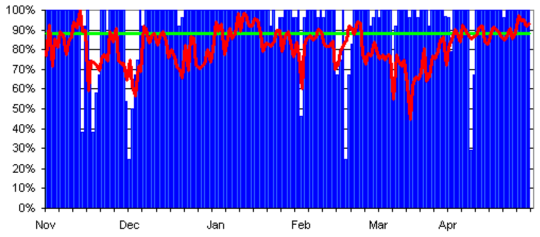
➢ **Scheduling of jobs at sites where the needed data was located was not possible**

☹ **The Grid – if not down – was at least seriously impaired…**

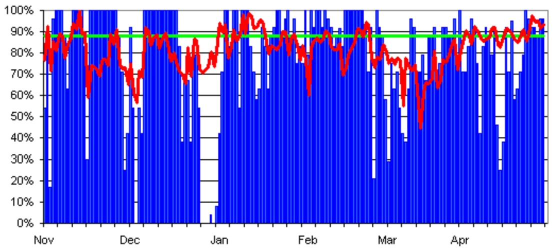
- **This was taken into account when designing the service deployment strategy & procedures – a taste of things to come!**

Problem Response Time and Availability targets
Tier-1 Centres

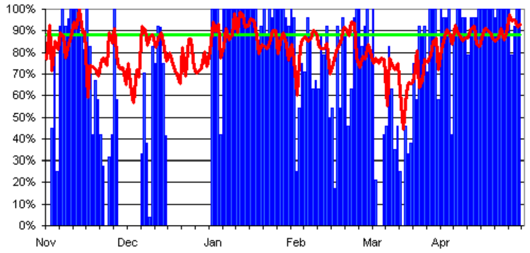
Service	*Maximum delay in responding to operational problems (hours)*			*Availability*
	Service interruption	*Degradation of the service*		
		> 50%	*> 20%*	
Acceptance of data from the Tier-0 Centre during accelerator operation	**12**	**12**	**24**	**99%**
Other essential services – prime service hours	**2**	**2**	**4**	**98%**
Other essential services – outside prime service hours	**24**	**48**	**48**	**97%**



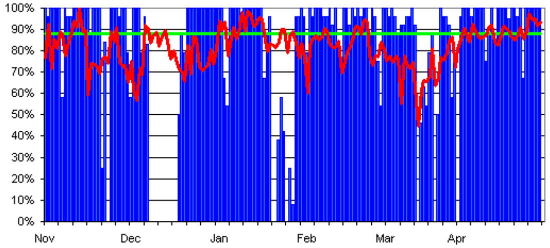
CERN-PROD av.reliability last 3 mths 94%



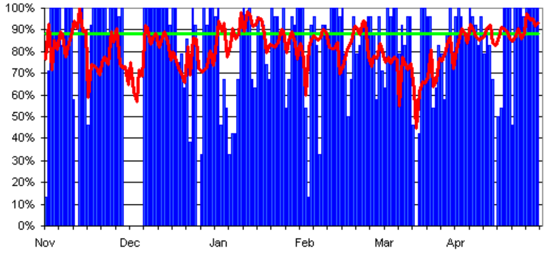
FZK-LCG2 av.reliability last 3 mths 81%



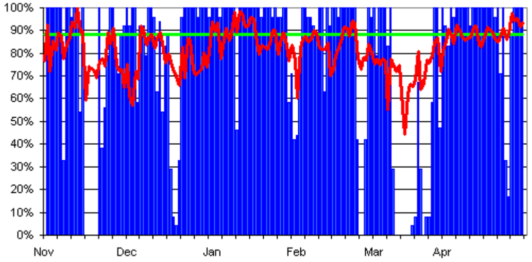
IN2P3-CC av.reliability last 3 mths 76%



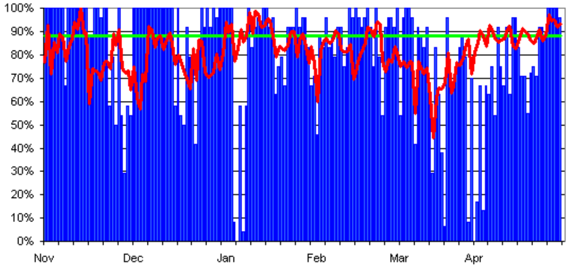
INFN-T1 av.reliability last 3 mths 89%



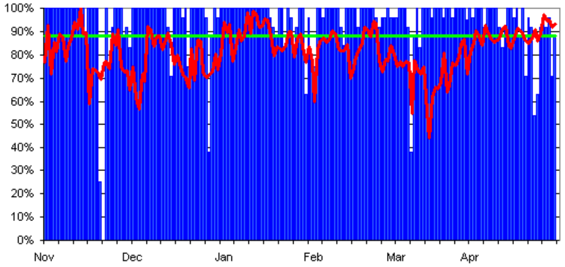
RAL-LCG2 av.reliability last 3 mths 83%



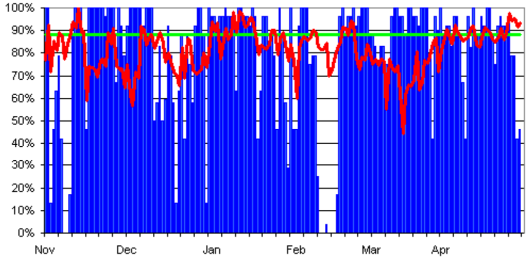
SARA-MATRIX av.reliability last 3 mths 75%



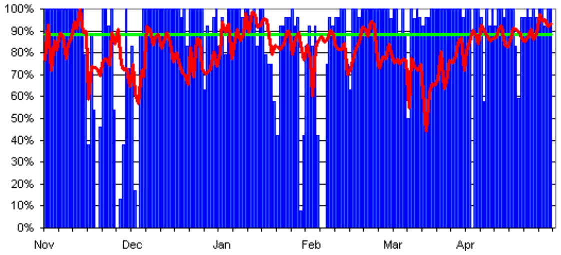
TRIUMF-LCG2 av.reliability last 3 mths 77%



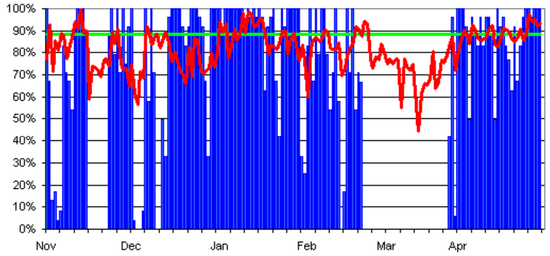
Taiwan-LCG2 av.reliability last 3 mths 94%



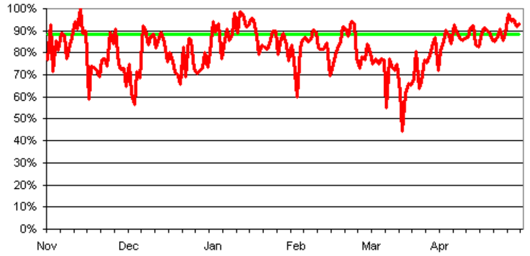
USCMS-FNAL-WC1 av.reliability last 3 mths 80%



PIC av.reliability last 3 mths 92%



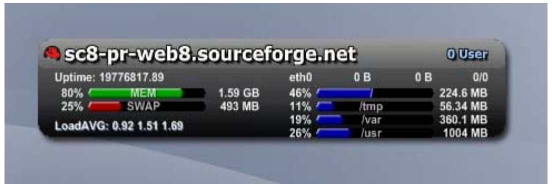
BNL-LCG2 av.reliability last 3 mths 53%



NDGF av.reliability last 3 mths n/a

The WLCG Dashboard(s)

- Sounds like a conventional problem for a 'dashboard'
- But there is not one single viewpoint…
  - Funding agency – how well are the resources provided being used?
  - VO manager – how well is my production proceeding?
  - Site administrator – are my services up and running? MoU targets?
  - Operations team – are there any alarms?
  - LHCC referee – how is the overall preparation progressing? Areas of concern?
  - …
- Nevertheless, much of the information that would need to be collected is common…
- So separate the collection from presentation (views…)
- As well as the discussion on metrics…



```
sc8-pr-web8.sourceforge.net                          0 User
Uptime: 19776817.89            eth0     0 B      0 B    0/0
 80%     MEM       1.59 GB     46%                 224.6 MB
 25%     SWAP      493 MB      11%    /tmp         56.34 MB
                               19%    /var         360.1 MB
LoadAVG: 0.92 1.51 1.69        26%    /usr         1004 MB
```

Invited talk "State of Readiness of LHC Computing" – CHEP '06

Monitoring - Status

- Since CHEP ’06 the number of monitoring / logging / reporting / dashboard efforts has increased

- There is a tendency for at least **some** of these efforts to attempt to cover the entire space

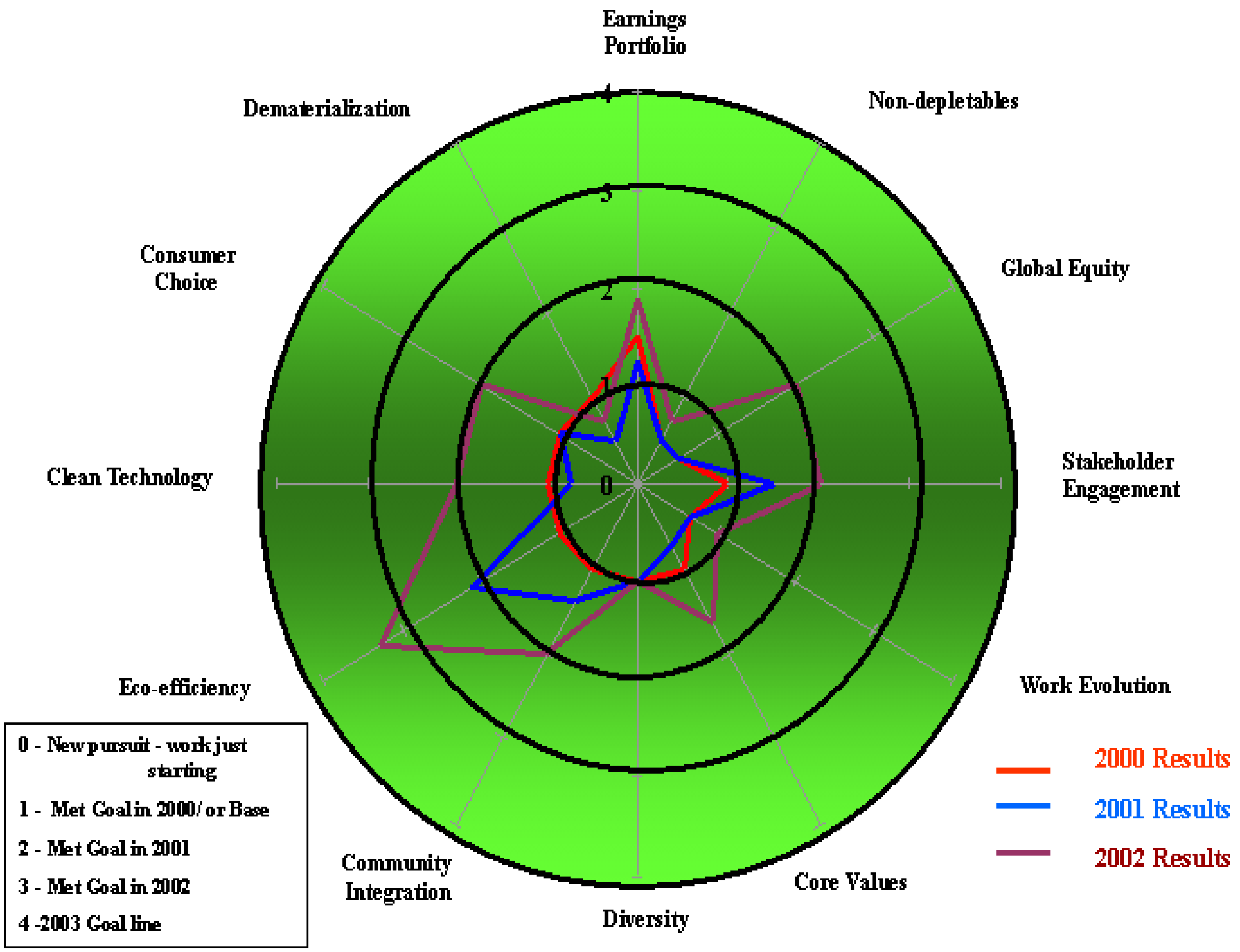
➢ **But this conflicts with the basic requirements!**

- Sites / VOs necessarily have their own monitoring tools and / or need for specific views

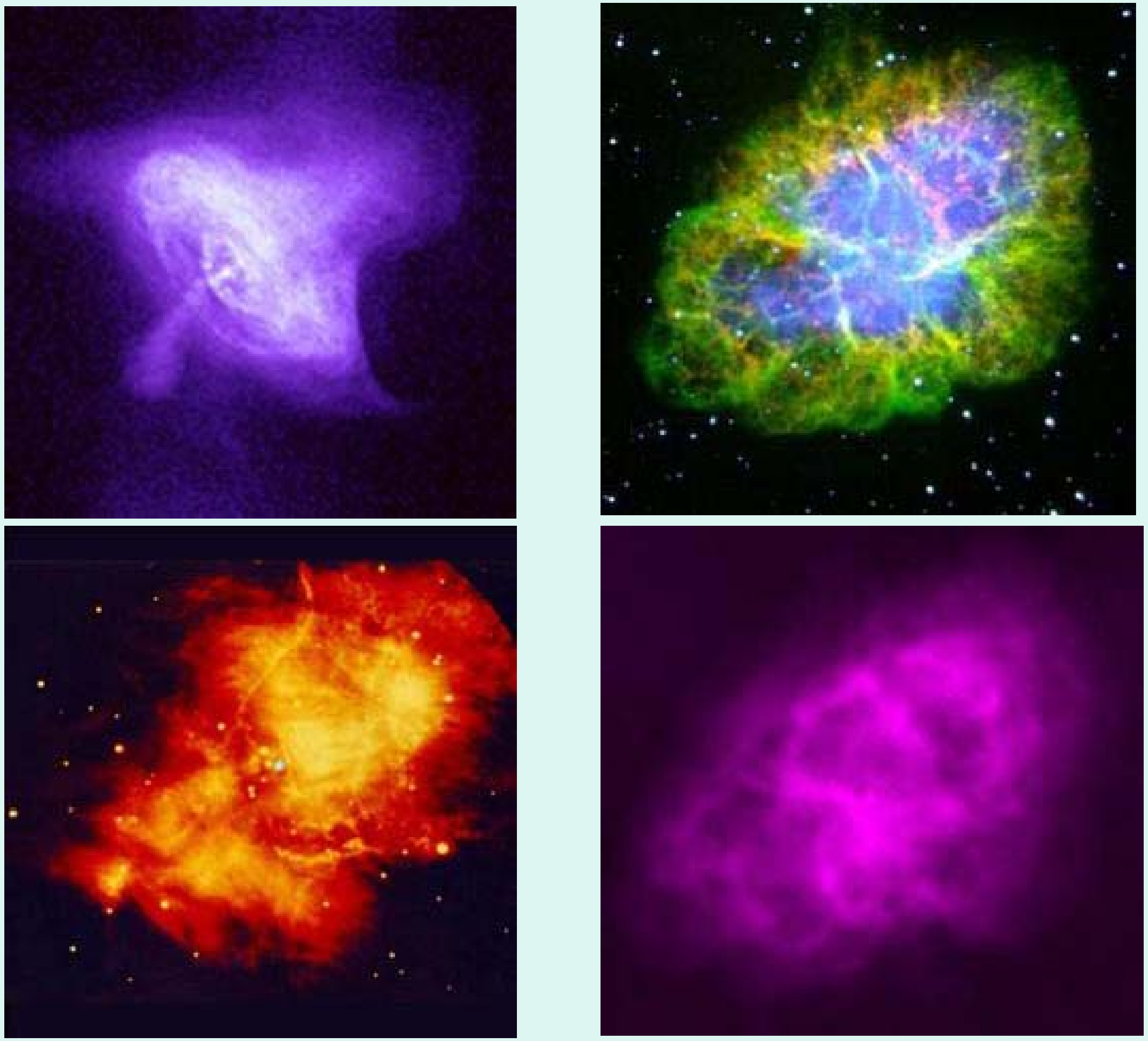
➢ **The ability to select the specific bits of information and correlate views from different sources is fundamental**

- c.f. ‘screen proliferation’ in the early days of LEP...

The Dashboard Again…



The Importance of Different ‘Views’

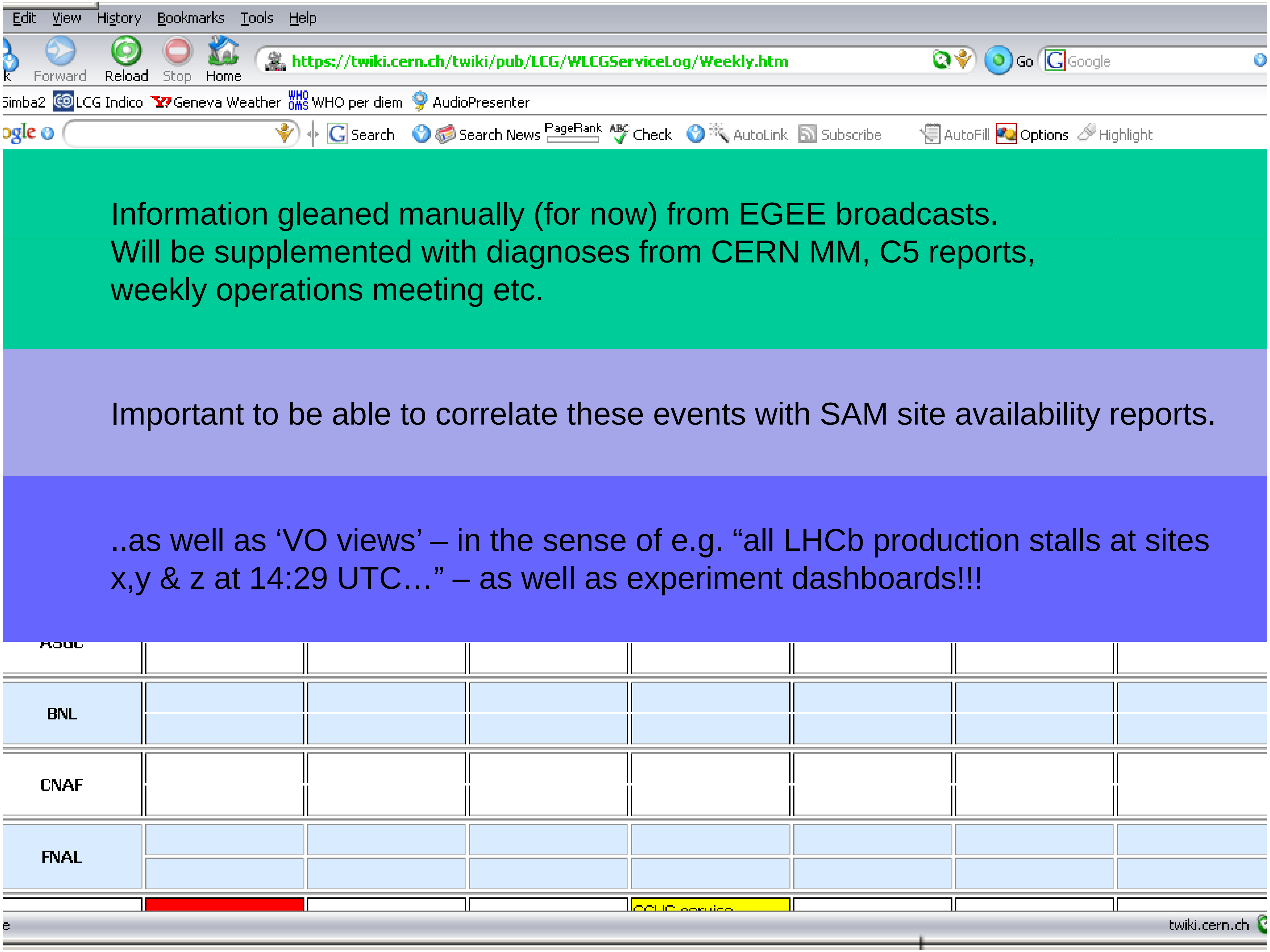


Nova first sighted 1054 A.D. by Chinese Astronomers

Now: Crab Nebula
X-ray,
optical,
infrared, and
radio

Slide courtesy of Robert Brunner @ CalTech.

http://research.microsoft.com/~gray/talks/cern_2001.ppt



Information gleaned manually (for now) from EGEE broadcasts. Will be supplemented with diagnoses from CERN MM, C5 reports, weekly operations meeting etc.

Important to be able to correlate these events with SAM site availability reports.

..as well as ‘VO views’ – in the sense of e.g. “all LHCb production stalls at sites x,y & z at 14:29 UTC…” – as well as experiment dashboards!!!

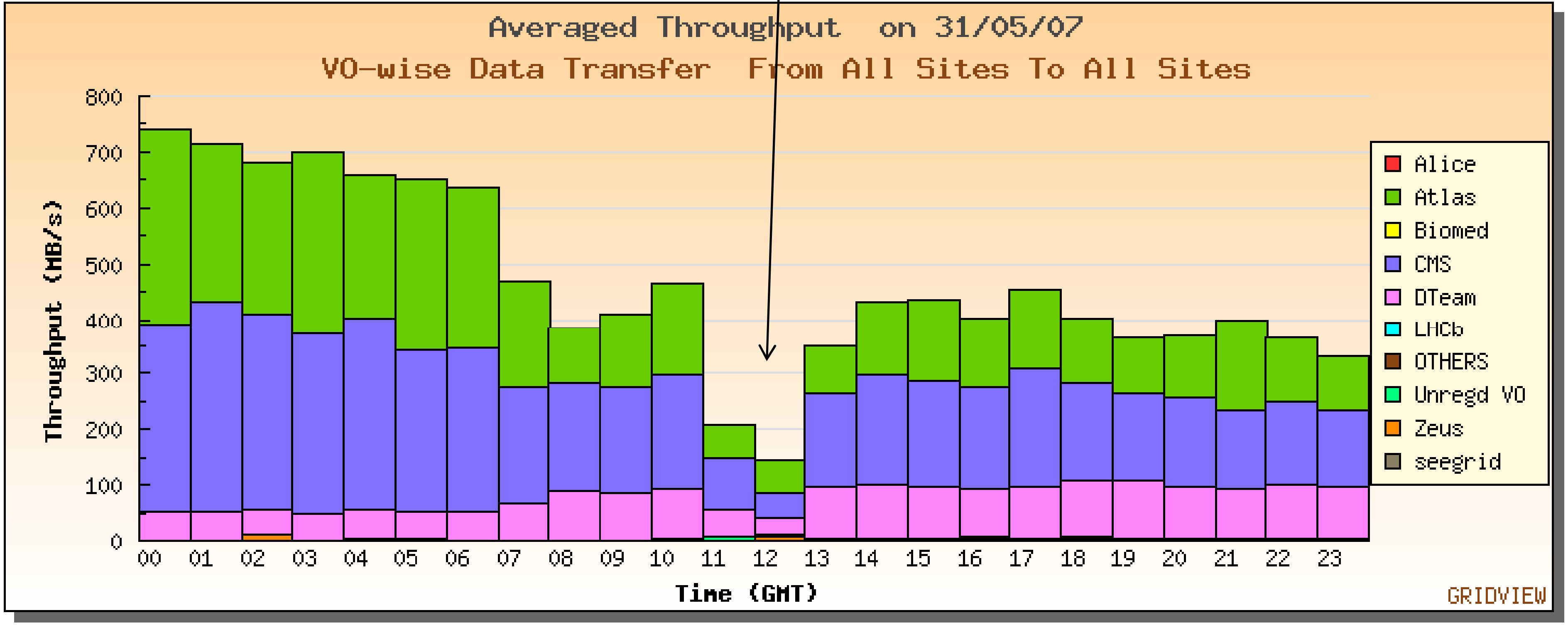
ASGC							
BNL							
CNAF							
FNAL							
	🟥			CCIIC service			

WLCG File Transfer Service

~All SRM V1.1 nodes at CERN down; all in same rack / on same switch.

Need to systematically record reasons for such problems – and their resolution

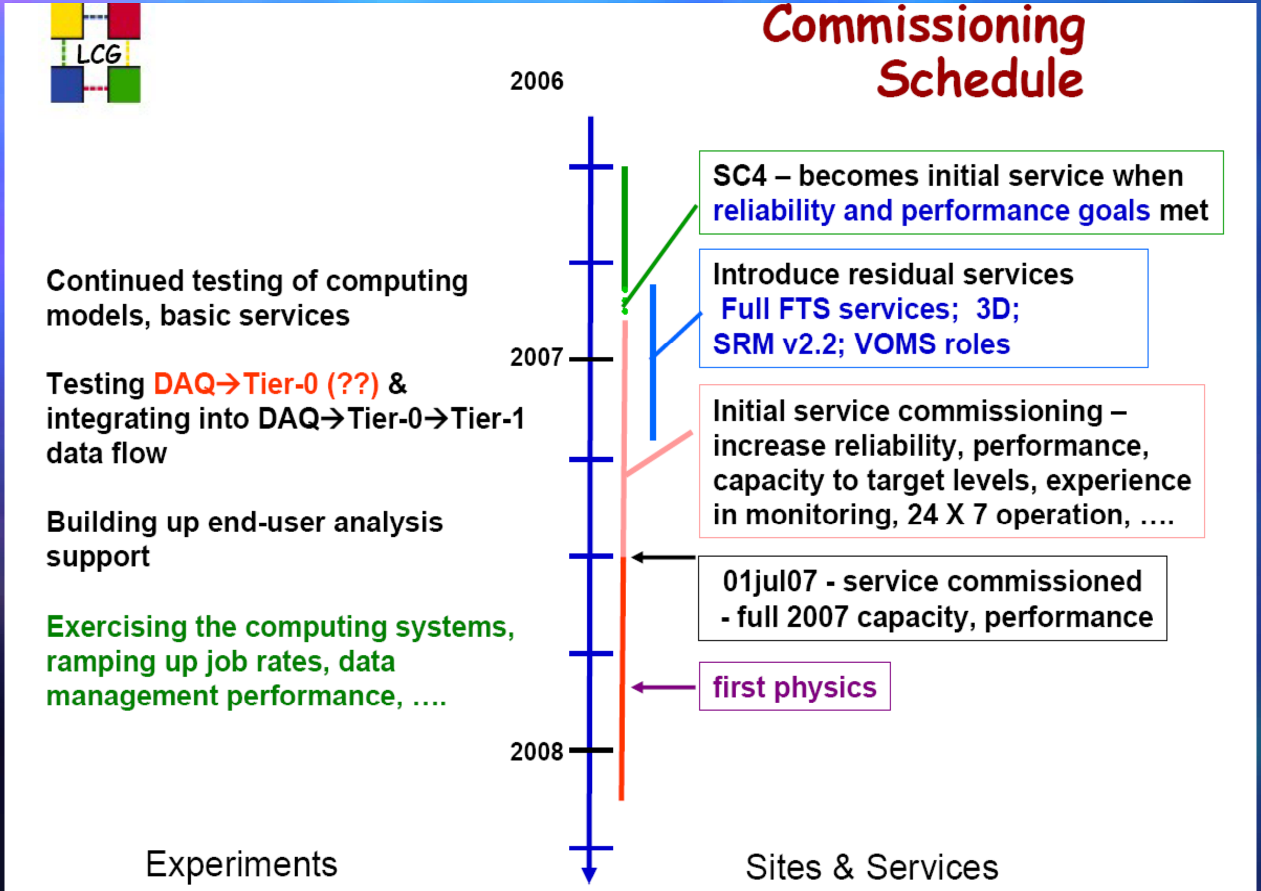
This is something we have been working on since end 2006, but it needs help from particularly the WLCG Tier1 sites in diagnosing and resolving problems.



WLCG Service: S / M / L vision

- Short-term: ready for Full Dress Rehearsals – now expected to fully ramp-up ~mid-September (>CHEP)
  - The only thing I see as realistic on this time-frame is FTS 2.0 services at WLCG Tier0 & Tier1s
  - Schedule: June 18th at CERN; mid-July at Tier1s
- Medium-term: what is needed & possible for 2008 LHC data taking & processing
  - The remaining 'residual services' **must** be in full production mode early Q1 2008
  - Significant improvements in monitoring, reporting, logging → more timely error response → service improvements
- Long-term: anything else
  - The famous 'sustainable e-Infrastructure'... ?

WLCG Commissioning Schedule



Commissioning Schedule

2006

2007

2008

Continued testing of computing models, basic services

Testing DAQ→Tier-0 (??) & integrating into DAQ→Tier-0→Tier-1 data flow

Building up end-user analysis support

Exercising the computing systems, ramping up job rates, data management performance, ….

SC4 – becomes initial service when reliability and performance goals met

Introduce residual services
Full FTS services; 3D;
SRM v2.2; VOMS roles

Initial service commissioning – increase reliability, performance, capacity to target levels, experience in monitoring, 24 X 7 operation, ….

01jul07 - service commissioned
- full 2007 capacity, performance

first physics

Experiments

Sites & Services

- **Still an ambitious programme ahead**
- Timely testing of full data chain from DAQ to T-2 chain was major item from last CR
  - DAQ→ T-0 still largely untested

Service Progress Summary

Component	Summary – updates presented at June GDB
LFC	Bulk queries deployed in February, Secondary groups deployed in April. ATLAS and LHCb are currently giving new specifications for other bulk operations that are scheduled for deployment this Autumn. Matching GFAL and lcg-utils changes.
DPM	SRM 2.2 support released in November. Secondary groups deployed in April. Support for ACLs on disk pools has just passed certification. SL4 32 and 64-bit versions certified apart from vdt (gridftp) dependencies.
FTS 2.0	Has been through integration and testing including certificate delegation, SRM v2.2 support and service enhancements – now being validated in PPS and pilot service (already completed by ATLAS and LHCb); will then be used in CERN production for 1 month (from June 18th) before release to Tier-1. Ongoing (less critical) developments to improve monitoring piece by piece continue.
3D	All Tier 1 sites in production mode and validated with respect to ATLAS conditions DB requirements. 3D monitoring integrated into GGUS problem reporting system. Testing to confirm streams failover procedures in next few weeks then will exercise coordinated DB recovery with all sites. Also starting Tier 1 scalability tests with many ATLAS and LHCb clients to have correct DB server resources in place by the Autumn.
VOMS roles	Mapping to job scheduling priorities has been implemented at Tier 0 and most Tier 1 but behavior is not as expected (ATLAS report that production role jobs map to both production and normal queues) so this is being re-discussed.

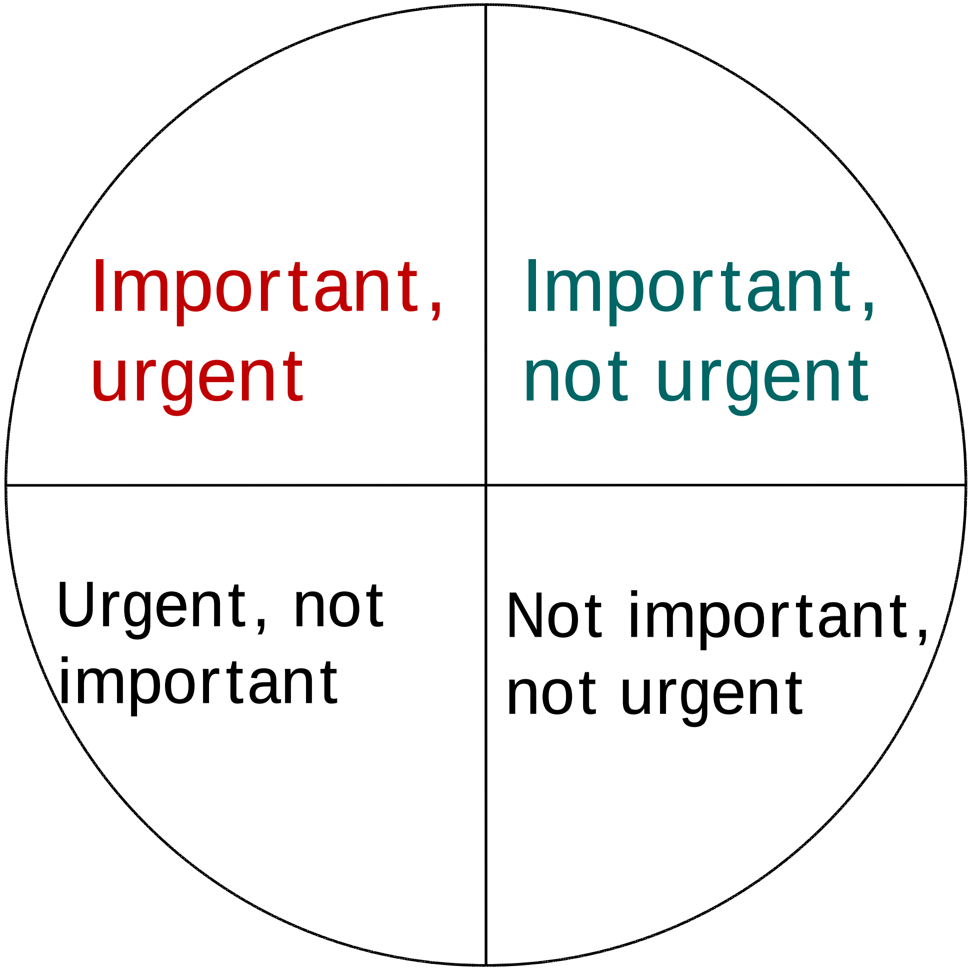
Service Progress Summary

Component	Summary – updates presented at June GDB
gLite 3.1 WMS	WMS passed certification and is now in integration. It is being used for validation work at CERN by ATLAS and CMS with LHCb to follow. Developers at CNAF fix any bugs then run 2 weeks of local testing before giving patches back to CERN.
gLite 3.1 CE	CE still under test with no clear date for 'completion'. Backup solution is to keep the existing 3.0 CE which will require SLC3 systems. Also discussing alternative solutions.
SL4	SL3 built SL4 compatibility mode UI and WN released but decision to deploy left to sites. Native SL4 32 WN in PPS now and UI ready to go in. Will not be released to production until after experiment testing is completed. SL4 DPM (needs vdt) important for sites that buy new hardware.
SRM 2.2	CASTOR2 work is coupled to the ongoing performance enhancements; dCache 1.8 beta has test installations at FNAL, DESY, BNL, FZK, Edinburgh, IN2P3 and NDGF, most of which also are in the PPS.
DAQ-Tier-0 Integration	Integration of ALICE with the Tier-0 has been tested with a throughput of 1 GByte/sec. LHCb testing planned for June then ATLAS and CMS from September.
Operations	**Many improvements are under way for increasing the reliability of all services. See this workshop & also WLCG Collaboration w/s @CHEP**

Agenda

- WLCG service operation & MoU targets
- **WLCG Service Coordination roles**
- S.W.O.T. analysis of WLCG service
- LHC startup challenges – we're not there yet!

Activities	Results
1. Crisis & problems	**Stress, burn-out, fire fighting, crisis management**
2. Planning, new opportunities	**Vision, balance, control, discipline**
3. Interruptions, e-mail, …	Out of control, victimised
4. Trivia, time wasting	Irresponsible, …



Service Availability Targets

- The WLCG Memorandum of Understanding defines:
  - **The services that a given site must provide (Tier0, Tier1, Tier2);**
  - **The availability of these services (measured on an annual basis);**
  - **The maximum time to intervene in case of problems.**

- Taken together, these service availability targets are somewhat aggressive and range from 95% to 99% for **compound** services, e.g.
  - **Acceptance of raw data from Tier0**
  - **Data-intensive analysis services, including networking to Tier0**

- Such ‘services’ involve many sub-services, e.g. storage services, catalog and metadata services, DB services, experiment-specific services etc.

- Major concerns include both **scheduled** and unscheduled interventions – must design all elements of the service correspondingly
  - **Hardware configuration; procedures & documentation; middleware**

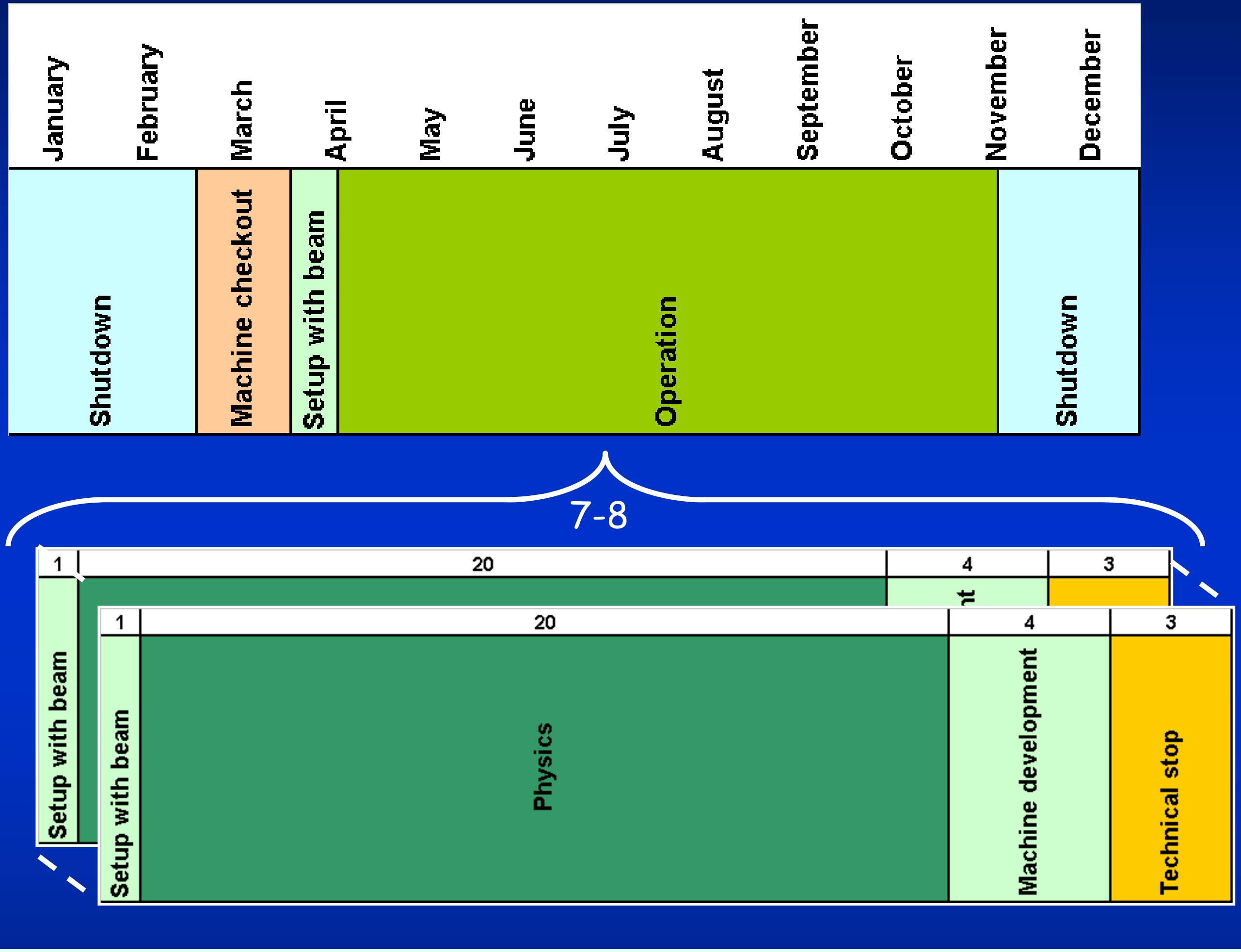
Service Availability - Experience

- Experience to date is that **scheduled** interventions account for far more downtime than unscheduled ones
    - Non-scheduled ‘transparent’ interventions can be highly pernicious…
- The worst interventions of all so far have been extended downtimes at numerous sites for cooling / power work
- **The “WLCG Tier0” service is so complex that there are interventions every week – often concurrently**
- Further pressure will be generated from the LHC running schedule (next) – effectively **reducing** the time slots when such necessary & essential work can take place
- **But** – and it’s a big but – apart from ‘pathological cases’, most interventions **could** be made ‘transparently’

Breakdown of a normal year

- From Chamonix XIV -

Service upgrade slots?



January February March April May June July August September October November December

Shutdown | Machine checkout | Setup with beam | Operation | Shutdown

7-8

1 | 20 | 4 | 3

Setup with beam | Physics | Machine development | Technical stop

~ 140-160 days for physics per year

Not forgetting ion and TOTEM operation

Leaves ~ 100-120 days for proton luminosity running

? Efficiency for physics 50% ?

~ 50 days ~ 1200 h ~ 4 10^6 s of proton luminosity running / year

Scheduled Service Interventions

Intervention type	CERN Daily OPS	Weekly OPS	EGEE Broadcast
“Transparent”	✓	Recommended	✓
Up to 4 hours	✓	Recommended	✓
Up to 12 hours	✓	Week of intervention, at least one working day in advance	✓
Over 12 hours	✓	Week prior to intervention	✓

EGEE Broadcast	Both prior to (at least one working day in advance) and after the intervention.
GOCDB	All scheduled downtimes must be entered in GOCDB – essential for SAM / GridView tests and service availability reports
CERN Daily OPS meeting (Tier0 interventions)	Announce the day **before**, reminder the day **of**, with follow-up the day **after** the intervention.
LCG ECM	It is recommended that all interventions are announced / discussed as far in advance as possible.
LCG SCM	Internal Tier0 intervention planning

Unscheduled Service Interventions

EGEE Broadcast	Immediately, with best prognosis of when the (full) service will be back. If not known or uncertain, date / time when further news will be provided. Further announcement when the service is restored.
SMOD / GMOD	Should be kept informed of progress
CERN Daily OPS meeting (Tier0 interventions)	Follow-up on any unscheduled interventions of the previous day.
Post-mortems	Covered in report to weekly operations meeting. Incidents resulting in prolonged down-time or service degradation should be covered in an explicit agenda item.

These guidelines should also be used for scheduled interventions in the case of problems.

Transparent Interventions - Definition

- Have reached agreement with the **LCG VOs** that the combination of hardware / middleware / experiment-ware **should** be resilient to service “glitches”

➢ **A glitch is defined as a short interruption of (one component of) the service that can be hidden – at least to batch – behind some retry mechanism(s)**

➢ **How long is a glitch?**

- All central CERN services are covered for power ‘glitches’ of up to 10 minutes
  - **Some are also covered for longer by diesel UPS but any non-trivial service seen by the users is only covered for 10’**
- Can we implement the services so that ~all interventions are ‘transparent’?

☺ **YES** – with some provisos

to be continued...

Advantages

- The advantages of such an approach are simply huge!

☺ **The users see a greatly improved service**

☺ **Service providers have significantly more flexibility in scheduling interventions**

☺ **The service provider – user relationship is enhanced**

☺ **Everyone's stress levels plummet!**

➢ **But it must be supported by the middleware…**

More Transparent Interventions

- *I am preparing to restart our SRM server here at IN2P3-CC so I have closed the IN2P3 channel on prod-fts-ws in order to drain current transfer queues.*
- *I will open them in 1 hour or 2.*

- Is this a transparent intervention or an unscheduled one?

- A: technically unscheduled, since it's SRM downtime.

☺ An EGEE broadcast was made, but this is just an example…

- But if the channel was first paused – which would mean that no files will fail – it becomes instead **transparent** – at least to the FTS – which is explicitly listed as a separate service in the WLCG MoU, both for T0 & T1!

- i.e. if we can trivially limit the impact of an intervention, we **should** **(c.f. WLCG MoU services at Tier0/Tier1s/Tier2s)**

Interventions – Moving Ahead

- At CHEP ’07 I will talk more about transparent interventions and how they could be implemented

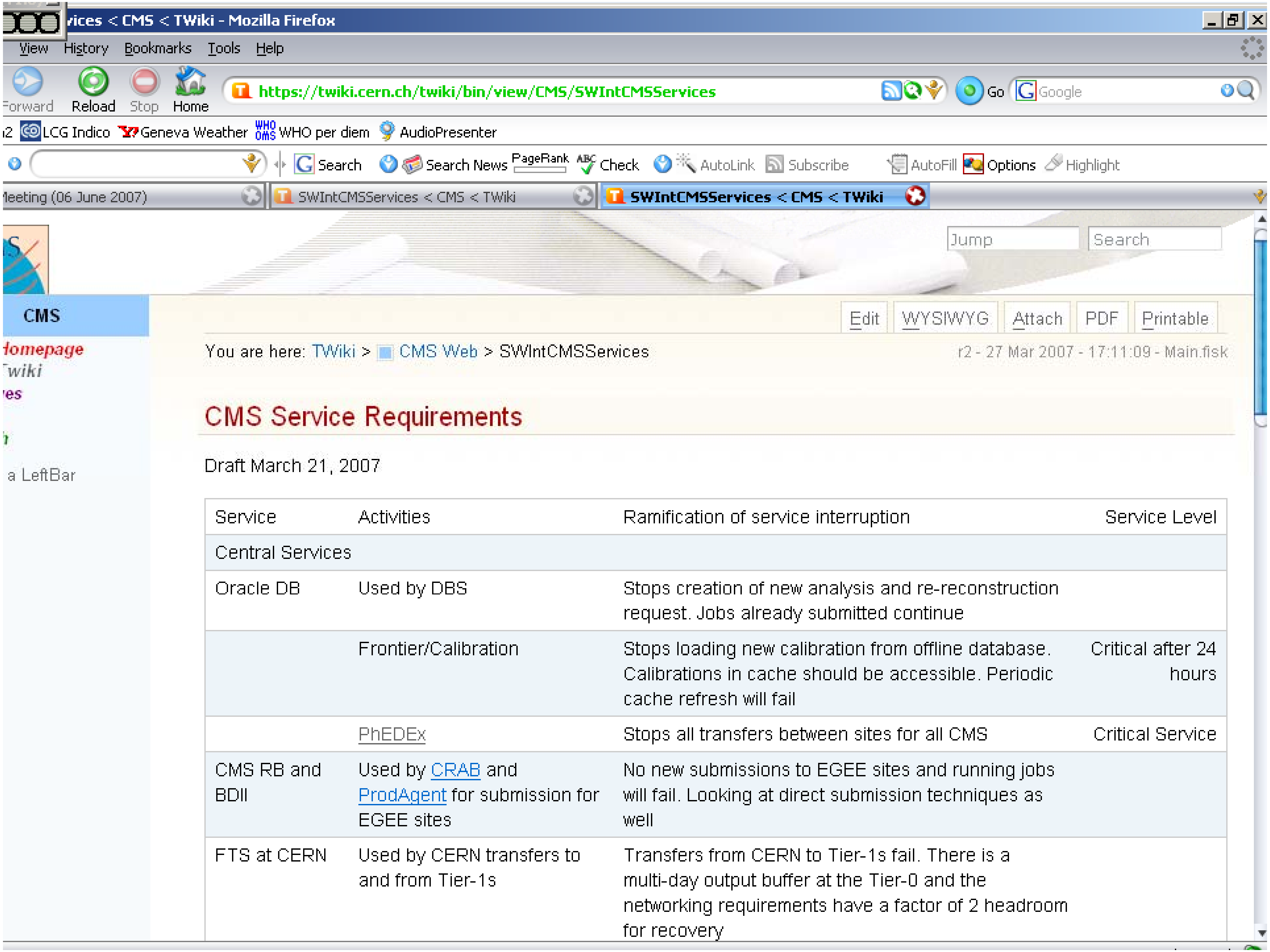
- However, it is clear that this has a much larger scope that originally foreseen (just e.g. upgrading services)

- And we need to develop a clear plan for its deployment

- Fits in the wider WLCG discussion to “improve reliability”, particularly in terms of best practices and service availability
  - Discussions so far in (Tier0) LCG Service Coordination Meetings, as well as with CMS, but our focus this year has been Residual Services (still)

- The focus should be given by the priorities of the LHC VOs – see next slide on critical services from CMS viewpoint…

- Although clearly some of the services – and hence techniques – will be generic and therefore of interest to other VOs



You are here: TWiki > CMS Web > SWIntCMSServices

r2 - 27 Mar 2007 - 17:11:09 - Main.fisk

CMS Service Requirements

Draft March 21, 2007

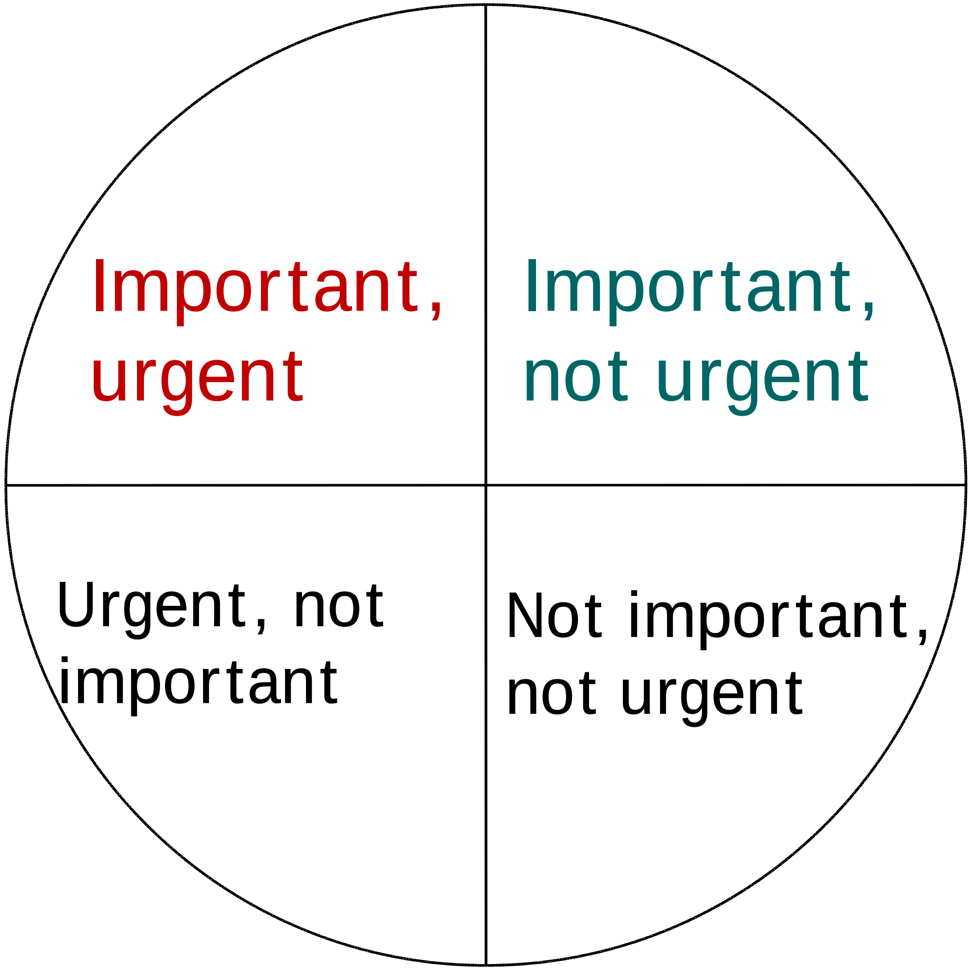
Service	Activities	Ramification of service interruption	Service Level
Central Services			
Oracle DB	Used by DBS	Stops creation of new analysis and re-reconstruction request. Jobs already submitted continue	
	Frontier/Calibration	Stops loading new calibration from offline database. Calibrations in cache should be accessible. Periodic cache refresh will fail	Critical after 24 hours
	PhEDEx	Stops all transfers between sites for all CMS	Critical Service
CMS RB and BDII	Used by CRAB and ProdAgent for submission for EGEE sites	No new submissions to EGEE sites and running jobs will fail. Looking at direct submission techniques as well	
FTS at CERN	Used by CERN transfers to and from Tier-1s	Transfers from CERN to Tier-1s fail. There is a multi-day output buffer at the Tier-0 and the networking requirements have a factor of 2 headroom for recovery	

CMS Control Centre at CERN

SC is regular visitor (at ***least*** daily) in control centres of all 4 VOs

WLCG Control Centre of the Future

Activities	Results
1. Crisis & problems	Stress, burn-out, fire fighting, crisis management
2. Planning, new opportunities	Vision, balance, control, discipline
3. Interruptions, e-mail, …	Out of control, victimised
4. Trivia, time wasting	Irresponsible, …



Important, urgent

Important, not urgent

Urgent, not important

Not important, not urgent

Agenda

- WLCG service operation & MoU targets

- WLCG Service Coordination roles

- **S.W.O.T. analysis of WLCG service**

- LHC startup challenges – we’re not there yet!

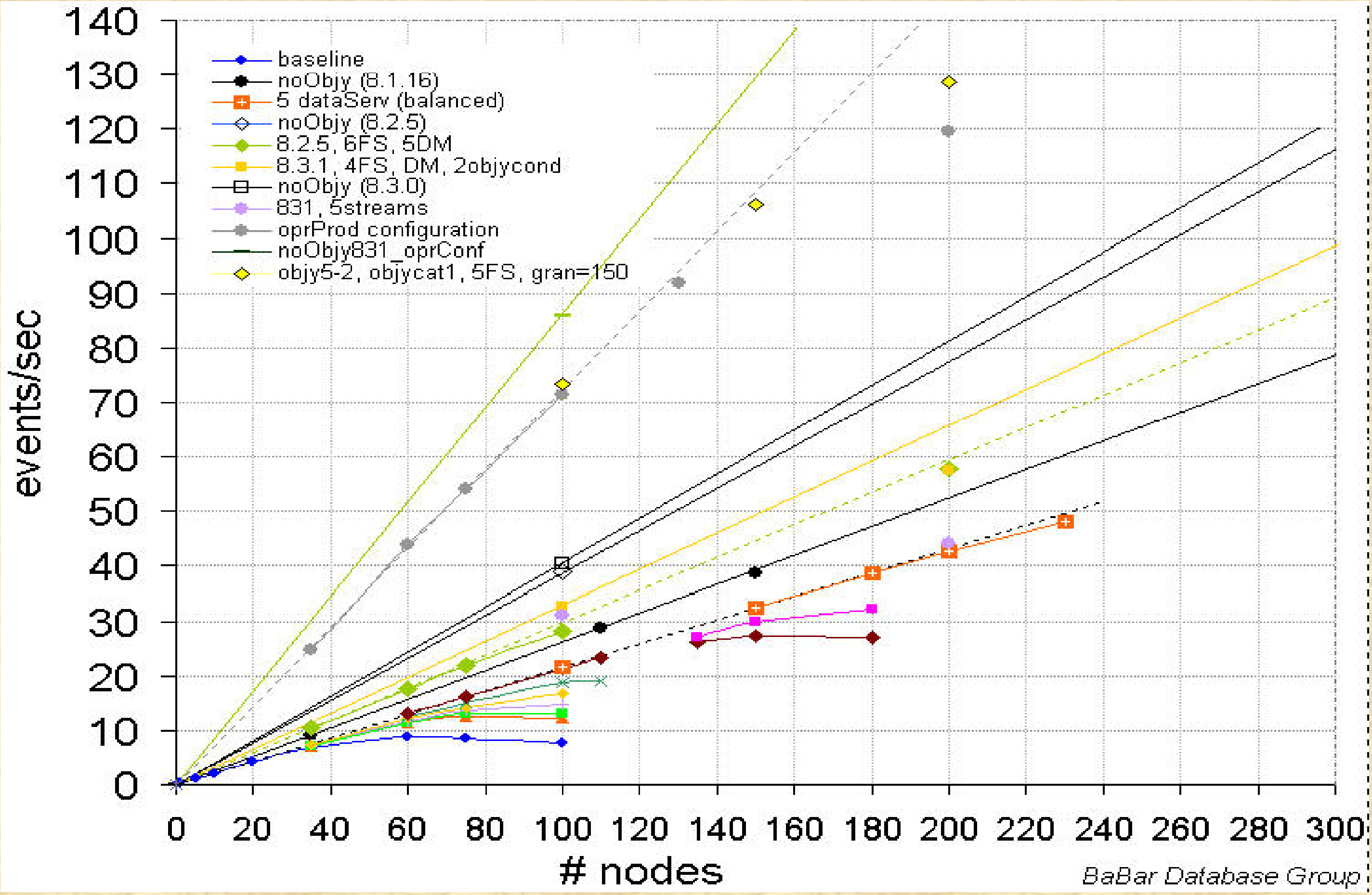
S.W.O.T. Analysis of WLCG Services

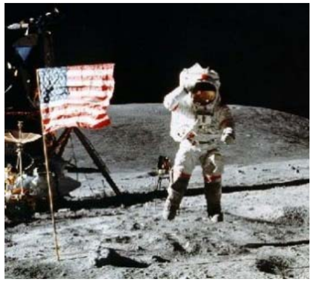
Strengths	**We do have a service that is used, albeit with a small number of well known and documented deficiencies (with work-arounds)**
Weaknesses	Continued service instabilities; holes in operational tools & procedures; ramp-up will take at least several (many?) months more…
Threats	**Hints of possible delays could re-ignite discussions on new features**
Opportunities	Maximise time remaining until high-energy running to: **1.) Ensure all remaining residual services are deployed as rapidly as possible, but only when sufficiently tested & robust;** **2.) Focus on smooth service delivery, with emphasis on improving all operation, service and support activities.** **All services (including 'residual') should be in place no later than Q1 2008, by which time a marked improvement in the measurable service level should also be achievable.**

Agenda

- WLCG service operation & MoU targets
- WLCG Service Coordination roles
- S.W.O.T. analysis of WLCG service
- **LHC startup challenges – we're not there yet!**

Startup woes – BaBar experience





“Conventional wisdom” - 2000

- “Either you have been there or you have not”

➢ **Translation: you need to test everything both separately and together under full production conditions before you can be sure that you are really ready. For the expected.**

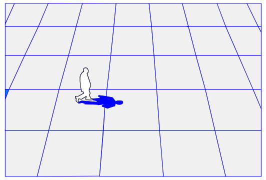
- There are still significant things that have not been tested by a single VO, let alone by all VOs together

- CMS CSA06 preparations: careful preparation and testing of all components over several months - basically everything broke first time (but was then fixed)... **This is a technique that has been proven (repeatedly) to work...**

💣 **This is simply “Murphy’s law for the Grid”...**

? How did we manage to forget this so quickly?

The Planetary System in Gliese 581
(Artist's Impression)
ESO Press Photo 22a/07 (25 April 2007)



The 1st Law Of (Grid) Computing

- **Murphy's law** (also known as **Finagle's law** or **Sod's law**) is a popular adage in Western culture, which broadly states that things will go wrong in any given situation. "If there's more than one way to do a job, and one of those ways will result in disaster, then somebody will do it that way." It is most commonly formulated as **"Anything that can go wrong will go wrong."** In American culture the law was named after Major Edward A. Murphy, Jr., a development engineer working for a brief time on rocket sled experiments done by the United States Air Force in 1949.

- … first received public attention during a press conference … it was that nobody had been severely injured during the rocket sled [of testing the human tolerance for g-forces during rapid deceleration.]. Stapp replied that it was because **they took Murphy's Law under consideration**.

- **"Expect the unexpected" – Bandits (Bruce Willis)**

Expecting the un-expected

- The Expected:
  - When services / servers don't respond or return an invalid status / message;
  - When users use a new client against an old server;
  - When the air-conditioning / power fails (again & again & again);
  - When 1000 batch jobs start up simultaneously and clobber the system;
  - **A disruptive and urgent security incident… (again, we've forgotten…)**
- The Un-expected:
  - When disks fail and you have to recover from backup – and the tapes have been overwritten;
  - When a 'transparent' intervention results in long-term service instability and (significantly) degraded performance;
  - When a service engineer puts a Coke into a machine to 'warm it up'…
- The Truly Un-expected:
  - When a fishing trawler cuts a trans-Atlantic network cable;
  - When a Tsunami does the equivalent in Asia Pacific;
  - When Oracle returns you someone else's data…
  - When mozzarella is declared a weapon of mass destruction…

➢ **All of these (and more) have happened!**

Summary

- **Robustness & operability are key issues to focus on for 2007 / 2008**
- Experience confirms that adding / upgrading services is a lengthy process – nothing extra beyond what is already agreed & underway!
- Production deployment of all residual services, coupled with significant improvements in service level, required for early Q1 2008

➢ Transparent interventions medium term goal – need plan for achieving this…

BACKUP SLIDES

WLCG & EGEE – An Analogy

- I have a company that has a contract with Swedish rail to move things about in trains
- We have an SLA – there are penalties if >x% of the deliveries are more than y hours late per month
- Swedish rail monitors the network – looking for bottlenecks, delays, mis-routings, break-downs etc.
- We have a monthly review meeting
- ➢ **But my business is not trains – its delivering snails from Burgundy to luxury French restaurants!**
- Do Swedish rail operators know this? Do they care?
- Obviously, EGEE & WLCG have a much more symbiotic (look it up) relationship than this…
- One level up this is the WLCG FTS (built using EGEE m/w) – for all we know, CMS could be transferring ‘digital snails’, which are re-constituted by the re-construction jobs at the Tier1s!

Scalability

Some targets for scalability and real life
experience in implementing them

Scalability – File Transfer Example

- LHC Experiments use a file size ~1GB
- Based on expected data rates & number of sites, the number of files to be transferred Tier0→Tier1 is 10^5 - 10^6 per day
  - Correspondingly higher if Tier2s also included in the game
- ‘Manual intervention’ to resolve file transfer problems is very time consuming, i.e. expensive and non-scalable
- Target: maximum 1 such problem per site per day

➢ **Service has to be reliable to 1 in $10^{5/6}$**

Scalability – Operations Example

- Current operations model is very ‘eye-ball intensive’
- ***And its not 24 x 7…***
- ***Don’t even mention public holidays…***
- **How will /can this scale to:**
  - Many more users?
  - A production Grid infrastructure?

➢ **It won’t.** Service reliability will of course help, but much more automation is clearly needed...

Scalability – User Support Example

- The story is the same...
- How many Ticket Processing Managers (TPMs) can we afford?
- How many users do we / will we have?
- ***How do we get the service to be so reliable and so well documented that we can survive?***

➢ **Need to think of the cost of each ticket**

- One that takes 1 hour of TPM-time costs €10-30 – possibly much more if user / VO costs also included!
  - Whilst TPMs probably rarely spend 1 hour / ticket, 3rd level support often spend considerably longer! Some unscheduled ‘transparent’ interventions have cost several weeks of expert time and caused extreme user dissatisfaction!
- **This is why call centres charge you per call!**
  - And why they are where they are...

Scalability - Conclusions

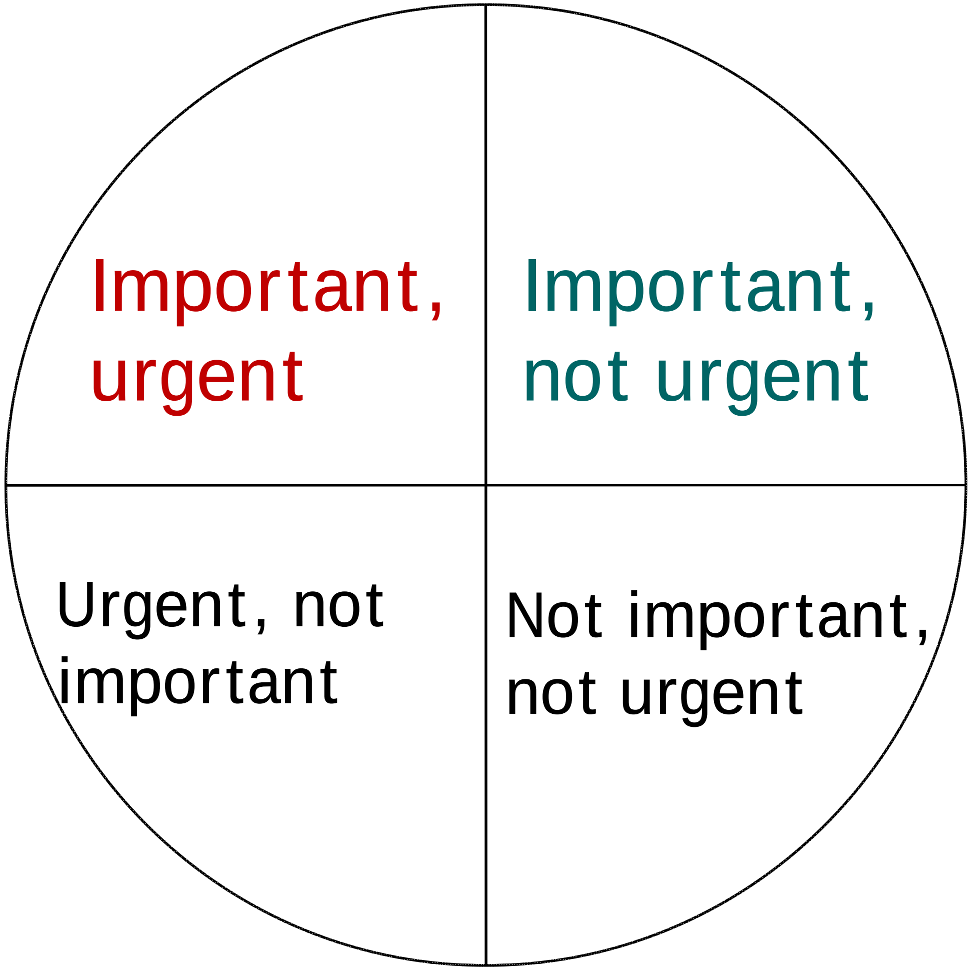
- If solutions are to cope with very large numbers of users (or whatever), **great care** must be taken to ensure that the solutions really scale

➢ **The critical issue (in most cases) is the available / required manpower to provide a solution**

- Computers are much better at doing repetitive tasks (rapidly) than humans!

- **If you can write a procedure to be followed, you can also write a script / programme / tool**

Activities	Results
1. Crisis & problems	Stress, burn-out, fire fighting, crisis management
2. Planning, new opportunities	**Vision, balance, control, discipline**
3. Interruptions, e-mail, …	Out of control, victimised
4. Trivia, time wasting	Irresponsible, …



Important, urgent

Important, not urgent

Urgent, not important

Not important, not urgent