The LCG Service Challenges: Rolling out the LCG Service

Jamie Shiers, CERN-IT-GD-SC

http://agenda.cern.ch/fullAgenda.php?ida=a053365

June 2005

Agenda

- Service Challenges: What and Why?
- Overview of LCG Tier Model
- Overview of LHC Experiments' Computing Models
- Service Challenges 1 & 2: What did we achieve?
- Service Challenge 3: July – December 2005
- Service Challenge 4: April – August 2006
- The Full Production Service
- Concerns and Issues for CERN / IT

Executive Summary

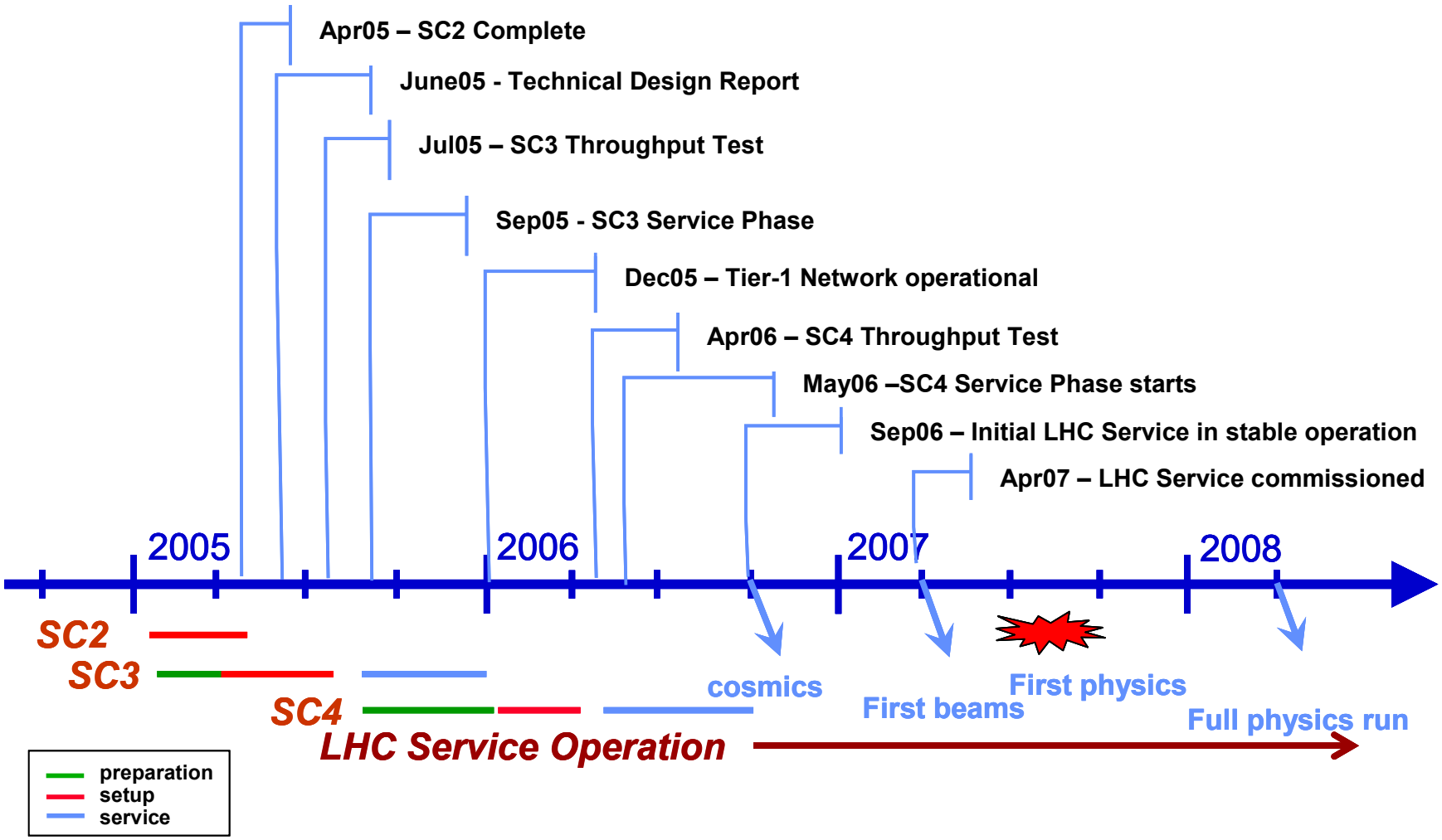
LCG Service Challenges - Overview

- LHC will enter production (physics) in summer 2007
  - Will generate an enormous volume of data
  - Will require huge amount of processing power
- LCG 'solution' is a world-wide Grid
  - Many components understood, deployed, tested..
- But…
  - Unprecedented scale
  - **Humungous challenge of getting large numbers of institutes and individuals, all with existing, sometimes conflicting commitments, to work together**
- LCG must be ready at full production capacity, functionality and reliability in **little more than 1 year** from now
  - Issues include h/w acquisition, personnel hiring and training, vendor rollout schedules etc.

➢ **Should not limit ability of physicist to exploit performance of detectors nor LHC's physics potential**

- Whilst being stable, reliable and easy to use

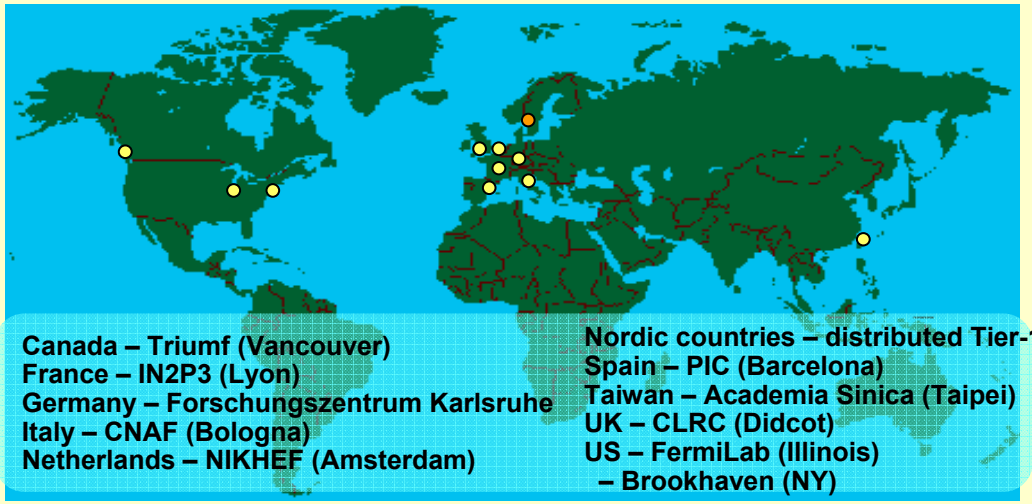
LCG Deployment Schedule



LCG Service Hierarchy

Tier-0 – the accelerator centre

- Data acquisition & initial processing
- Long-term data curation
- Distribution of data → Tier-1 centres



Tier-1 – "online" to the data acquisition process → high availability

- Managed Mass Storage – → grid-enabled data service
- Data intensive analysis
- National, regional support

Tier-2 – ~100 centres in ~40 countries

- Simulation
- End-user analysis – batch and interactive

Networking

- Latest estimates are that Tier-1s will need connectivity at ~10 Gbps with ~70 Gbps at CERN
- There is no real problem for the technology
  as has been demonstrated by a succession
  of Land Speed Records
- But LHC will be one of the few applications needing –
  - this level of performance as a service on a global scale
- We have to ensure that there will be an effective international backbone –
  that reaches through the national research networks
  to the Tier-1s
- LCG has to be pro-active in working with service providers
  - Pressing our requirements and our timetable
  - Exercising pilot services

Mandatory Services: for -

- Data acquisition & initial processing
- Long-term data curation
- Distribution of data to Tier-1 centres

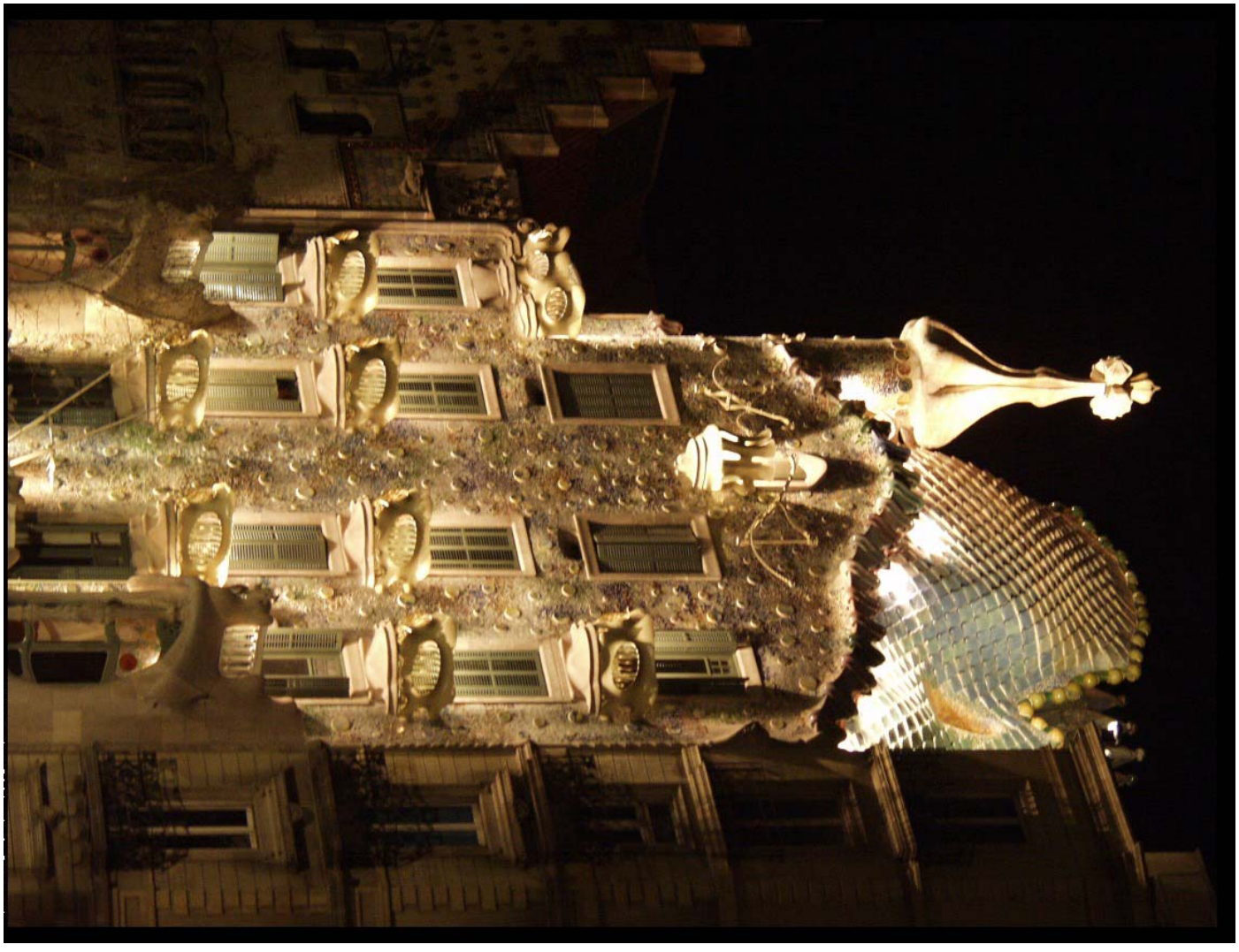
Tier-1 Centres

				ALICE	ATLAS	CMS	LHCb	
1	GridKa	Karlsruhe	Germany	X	X	X	X	4
2	CCIN2P3	Lyon	France	X	X	X	X	4
3	CNAF	Bologna	Italy	X	X	X	X	4
4	NIKHEF/SARA	Amsterdam	Netherlands	X	X		X	3
5	NDGF	Distributed	Dk, No, Fi, Se	X	X			1
6	PIC	Barcelona	Spain		X	X	X	3
7	RAL	Didcot	UK	X	X	X	X	4
8	Triumf	Vancouver	Canada		X			1
9	BNL	Brookhaven	US		X			1
10	FNAL	Batavia, Ill.	US			X		1
11	ASCC	Taipei	Taiwan		X	X		2
				6	10	7	6	

A US Tier1 for ALICE is also expected.

LCG Service Challenges – Deploying the Service

LCG Service Challenges – Deploying the Service

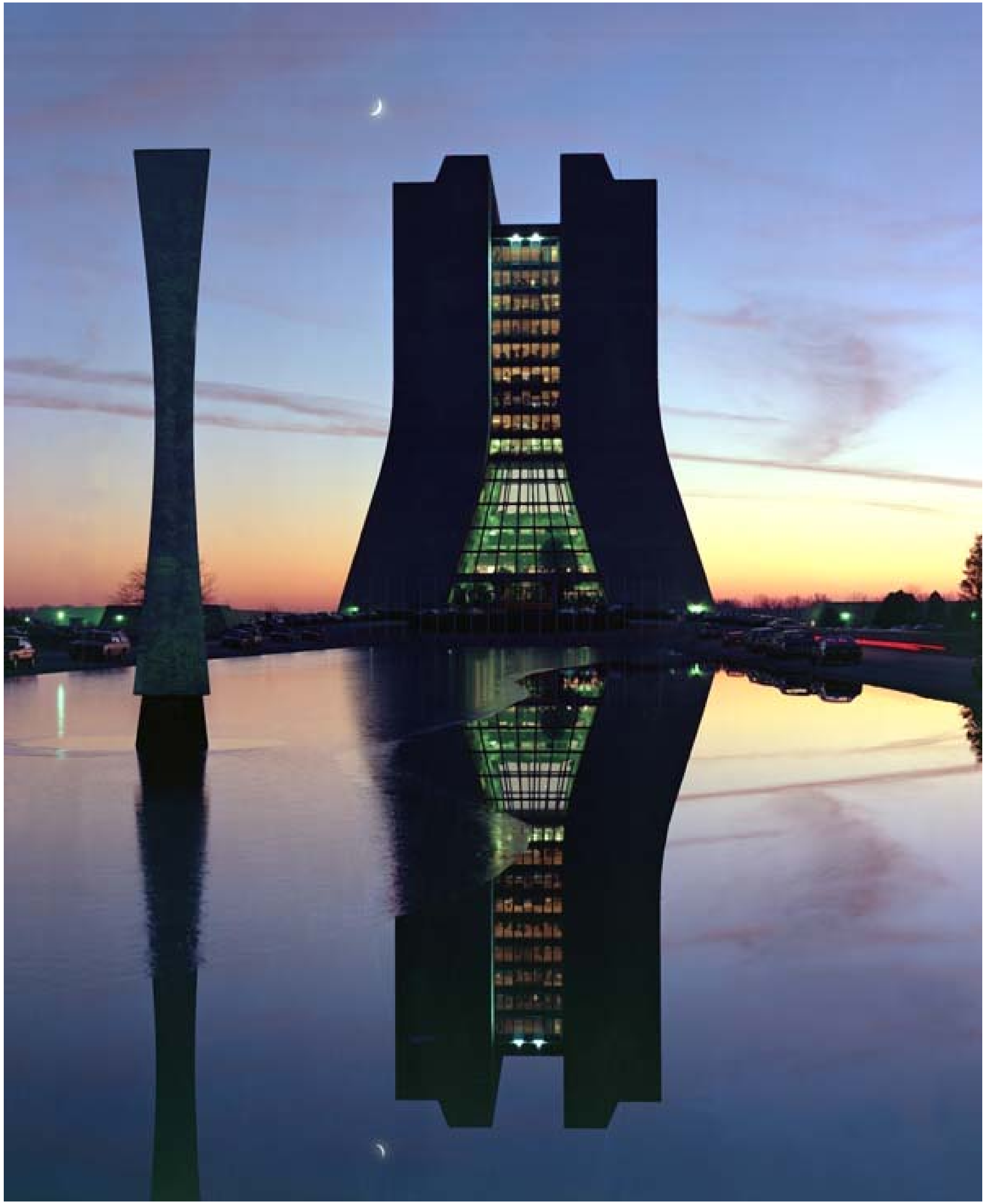


LCG Service Challenges – Deploying the Service

LCG Service Challenges – Deploying the Service


寶島鐘錶
金興發生活館
士林店
新榮旅社
HOTEL
28812531
28827372
奇奇鐘錶
隱形眼鏡
寶島眼鏡
皇家眼鏡
台北士林店
上越國際髮型
朋友
思沛克咖啡
手機維修回收中心
好鄰居


Jerome Favre / AP



LCG Service Challenges – Deploying the Service

LCG Service Challenges – Deploying the Service

LCG Service Challenges – Deploying the Service

Summary of Tier0/1/2 Roles

- Tier0 (CERN): safe keeping of RAW data (first copy); first pass reconstruction, distribution of RAW data and reconstruction output to Tier1; reprocessing of data during LHC down-times;
- Tier1: safe keeping of a proportional share of RAW and reconstructed data; large scale reprocessing and safe keeping of corresponding output; distribution of data products to Tier2s and safe keeping of a share of simulated data produced at these Tier2s;
- Tier2: Handling analysis requirements and proportional share of simulated event production and reconstruction.

N.B. there are differences in roles by experiment
Essential to test using complete production chain of each!

Service Challenges

- Purpose
  - Understand what it takes to operate a **real grid service** – run for days/weeks at a time (outside of experiment Data Challenges)
  - Trigger/encourage the Tier1 & large Tier-2 planning – move towards real resource planning – based on realistic usage patterns
  - Get the essential grid services ramped up to target levels of reliability, availability, scalability, end-to-end performance
  - Set out milestones needed to achieve goals during the service challenges
- NB: This is focussed on Tier 0 – Tier 1/large Tier 2
  - Data management, batch production and analysis
- Short term goal – **by end 2004** –
  have in place a robust and reliable data management service and support infrastructure and robust batch job submission

From early proposal, May 2004

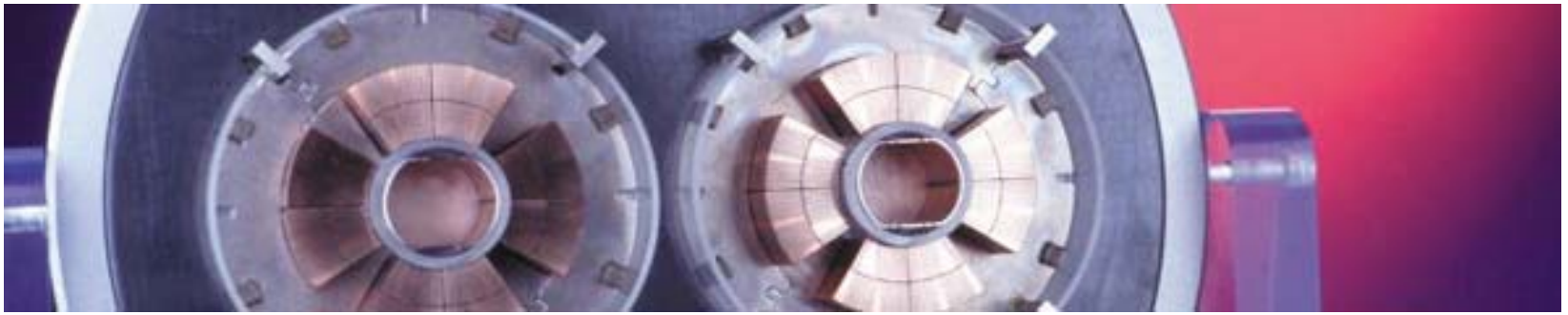
Overview of LHC Operation Schedule and Experiments' Computing Models

LHC Parameters (Computing Models)

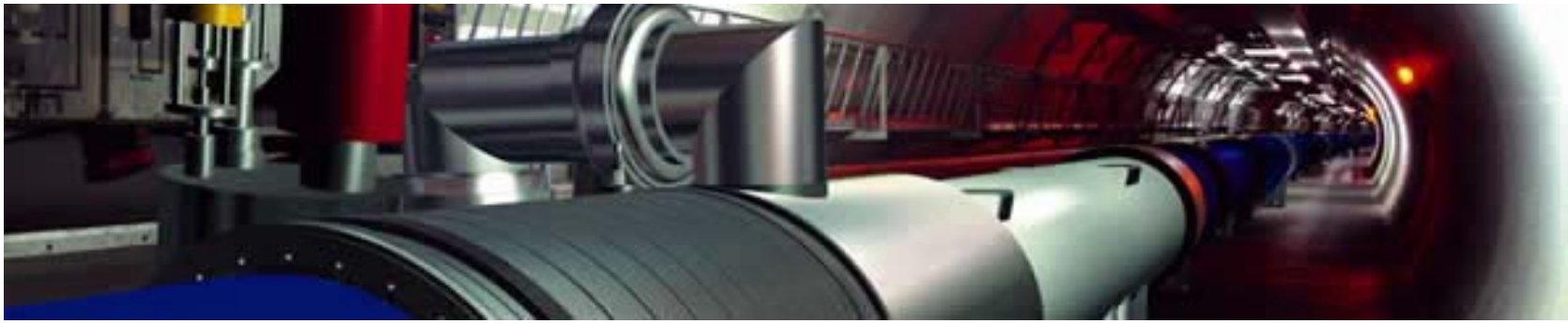
Year	pp operations		Heavy Ion operations	
	Beam time (seconds/year)	Luminosity ($cm^{-2}s^{-1}$)	Beam time (seconds/year)	Luminosity ($cm^{-2}s^{-1}$)
2007	5×10^{6}	5×10^{32}	-	-
2008	$(1.8 \times) 10^{7}$	2×10^{33}	$(2.6 \times) 10^{6}$	5×10^{26}
2009	10^{7}	2×10^{33}	10^{6}	5×10^{26}
2010	10^{7}	10^{34}	10^{6}	5×10^{26}

(Real time given in brackets above)

LHC Schedule – "Chamonix" workshop



- First collisions: two months after first turn on in August 2007
- 32 weeks of operation, 16 weeks of shutdown, 4 weeks commissioning = 140 days physics / year (5 lunar months)



Overview of pp running

Experiment	SIM	SIMESD	RAW	Trigger	RECO	AOD	TAG
ALICE	400KB	40KB	1MB	100Hz	200KB	50KB	10KB
ATLAS	2MB	500KB	1.6MB	200Hz	500KB	100KB	1KB
CMS	2MB	400KB	1.5MB	150Hz	250KB	50KB	10KB
LHCb		400KB	25KB	2KHz	75KB	25KB	1KB

Storage (2008 only!)

Experiment	T0 (AF)	T1	T2	Total (PB)
ALICE	2.3	7.5	-	9.8
ATLAS	4.7 (0.5)	6.5	-	11.2
CMS	3.8	12.9	-	16.6
LHCb	1.359	2.074	-	3.433
	12.2			

Cumulative storage needs should be considered…

Streaming

- All experiments foresee RAW data streaming, but with different approaches:
- CMS: O(50) streams based on trigger path
  - Classification is immutable, defined by L1+HLT
- Atlas: 4 streams based on event types
  - Primary physics, Express line, Calibration, Debugging and diagnostic
- LHCb: >4 streams based on trigger category
  - B-exclusive, Di-muon, D* Sample, B-inclusive
  - Streams are not created in the first pass, but during the "stripping" process

→ Not clear what is the best/right solution. Probably bound to evolve in time.

Reprocessing

- Data need to be reprocessed several times because of:
  - Improved software
  - More accurate calibration and alignment
- Reprocessing mainly at T1 centers
  - LHCb is planning on using the T0 during the shutdown – not obvious it is available
- Number of passes per year

Alice	Atlas	CMS	LHCb
3	2	2	4

➢ But experience shows the reprocessing requires huge effort!

➢ Use these numbers in the calculation but 2 / year will be good going!

pp / AA data rates (equal split)

Centre	ALICE	ATLAS	CMS	LHCb	Rate into T1 (pp)	Rate into T1 (AA)
ASCC, Taipei	0	1	1	0	118.7	28.2
CNAF, Italy	1	1	1	1	205.0	97.2
PIC, Spain	0	1	1	1	179.0	28.2
IN2P3, Lyon	1	1	1	1	205.0	97.2
GridKA, Germany	1	1	1	1	205.0	97.2
RAL, UK	1	1	1	1	205.0	97.2
BNL, USA	0	1	0	0	72.2	11.3
FNAL, USA	0	0	1	0	46.5	16.9
TRIUMF, Canada	0	1	0	0	72.2	11.3
NIKHEF/SARA, Netherlands	1	1	0	1	158.5	80.3
Nordic Data Grid Facility	1	1	0	0	98.2	80.3
Totals	6	10	7	6		

N.B. these calculations assume equal split as in Computing Model documents. It is clear that this is not the 'final' answer…

Data Rates Per Site

- Nominal rates per site expected to converge on 150 – 200MB/s during proton running
  - Balance of data vs resources and community served at various Tier1s
- In terms of number of tape drives provisioned at a Tier1, this is essentially the same number
  - Slight variation depending on assumed efficiency and technology
  - But drives are quantised…
- 5 drives per site for archiving share of raw data?

➢ **For now, planning for 10Gbit links to all Tier1s**

- Including overhead, efficiency and recovery factors…

Nominal	These are the raw figures produced by multiplying e.g. event size x trigger rate.
Headroom	A factor of 1.5 that is applied to cater for peak rates.
Efficiency	A factor of 2 to ensure networks run at less than 50% load.
Recovery	A factor of 2 to ensure that backlogs can be cleared within 24 - 48 hours and to allow the load from a failed Tier1 to be switched over to others.
Total Requirement	**A factor of 6 must be applied to the nominal values to obtain the bandwidth that must be provisioned.** **Arguably this is an over-estimate, as "Recovery" and "Peak load" conditions are presumably relatively infrequent, and can also be smoothed out using appropriately sized transfer buffers.** **But as there may be under-estimates elsewhere…**

Base Requirements for T1s

- **Provisioned bandwidth comes in units of 10Gbits/sec although this is an evolving parameter**
  - *From* Reply to Questions from Computing MoU Task Force…
- Since then, some parameters of the Computing Models have changed
- Given the above quantisation, relatively insensitive to small-ish changes
- Important to understand implications of multiple-10Gbit links, particularly for sites with Heavy Ion programme
  - Spread of AA distribution during shutdown probably means 1 link sufficient…
- **For now, planning for 10Gbit links to all Tier1s**

Service Challenges: Key Principles

- Service challenges result in a **series** of services that exist in **parallel** with **baseline production** service
- Rapidly and successively approach production needs of LHC
- Initial focus: core (data management) services
- Swiftly expand out to cover **full spectrum** of production and analysis chain
- Must be as realistic as possible, including end-end testing of key experiment **use-cases** over extended periods with recovery from **glitches** and **longer-term** outages

➢ **Necessary resources and commitment pre-requisite to success!**

- Effort should not be under-estimated!

Why Service Challenges?

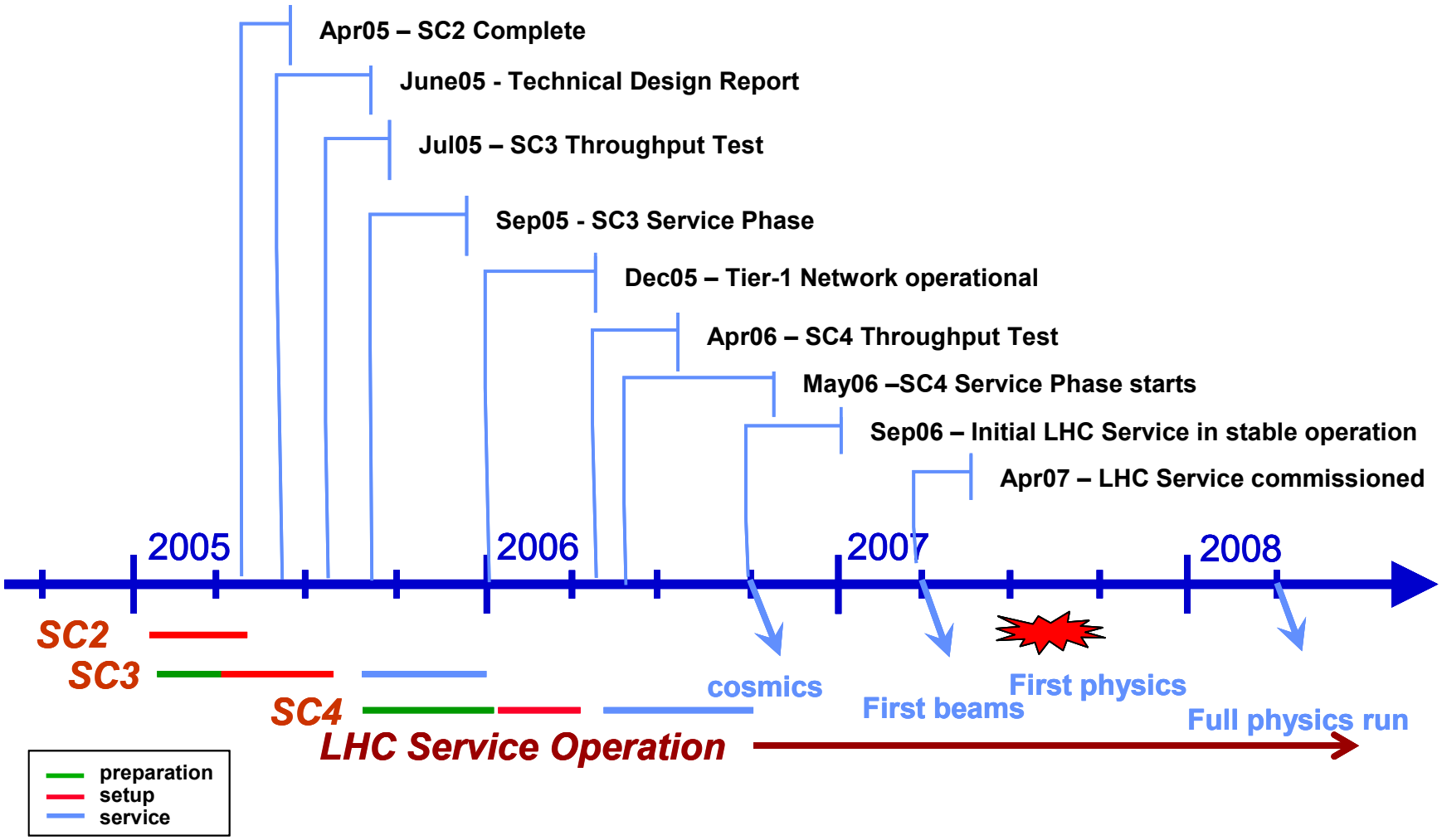
To test Tier-0 ←→ Tier-1 ←→Tier-2 services

- Network service
  - Sufficient bandwidth: ~10 Gbit/sec
  - Backup path
  - Quality of service: security, help desk, error reporting, bug fixing, ..
- Robust file transfer service
  - File servers
  - File Transfer Software (GridFTP)
  - Data Management software (SRM, dCache)
  - Archiving service: tapeservers,taperobots, tapes, tapedrives, ..
- Sustainability
  - Weeks in a row un-interrupted 24/7 operation
  - Manpower implications: ~7 fte/site
  - Quality of service: helpdesk, error reporting, bug fixing, ..

➢ **Towards a stable production environment for experiments**

Kors Bos – Presentation to LHCC, March 7 2005

LCG Deployment Schedule



SC1 Review

☹ SC1 **did not** complete its goals successfully

- Dec04 - Service Challenge I complete
  - mass store (disk) - mass store (disk)
  - 3 T1s (Lyon, Amsterdam, Chicago) (others also participated…)
  - 500 MB/sec (individually and aggregate)
  - 2 weeks sustained
  - Software; GridFTP plus some scripts

☠ **We did not meet the milestone of 500MB/s for 2 weeks**

- We need to do these challenges to see what actually goes wrong
  - A lot of things do, and did, go wrong
- We need better test plans for validating the infrastructure before the challenges (network throughput, disk speeds, etc…)

- OK, so we're off to a great start with the Service Challenges…

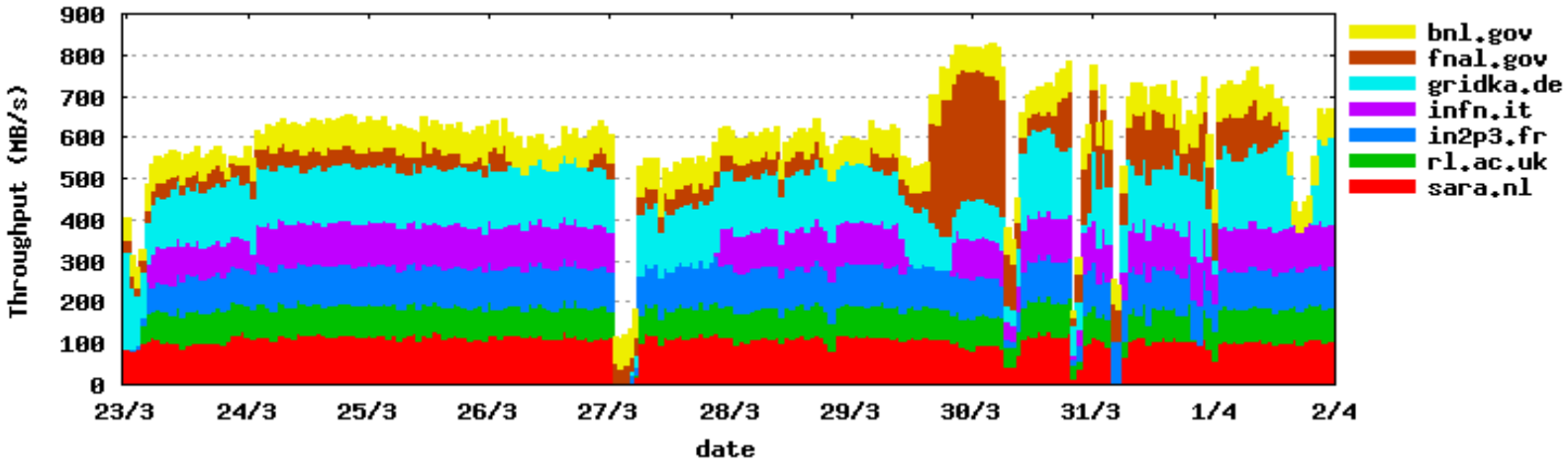
SC2 - Overview

- "Service Challenge 2"
  - Throughput test from Tier-0 to Tier-1 sites
  - Started 14th March
- Set up Infrastructure to 7 Sites
  - NIKHEF/SARA, IN2P3, FNAL, BNL, FZK, INFN, RAL
- 100MB/s to each site
  - 500MB/s combined to all sites at same time
  - 500MB/s to a few sites individually
- Goal : by end March, sustained 500 MB/s at CERN

SC2 met its Throughput Targets

- ☺ >600MB/s daily average for 10 days achieved
  - goal 500MB/s
- ☺ 7 sites participated (more than foreseen):
  - NIKHEF/SARA, IN2P3, FNAL, BNL, FZK, INFN, RAL
- ☹ Not without outages, but system showed it could recover

➢ **But we still don't have anything we could call a service…**

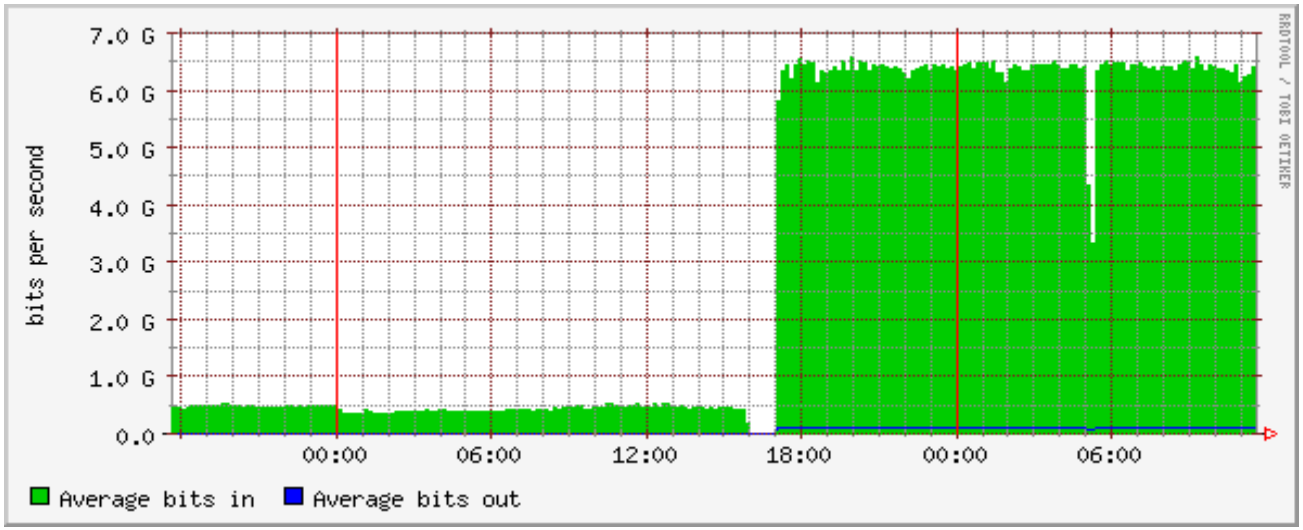


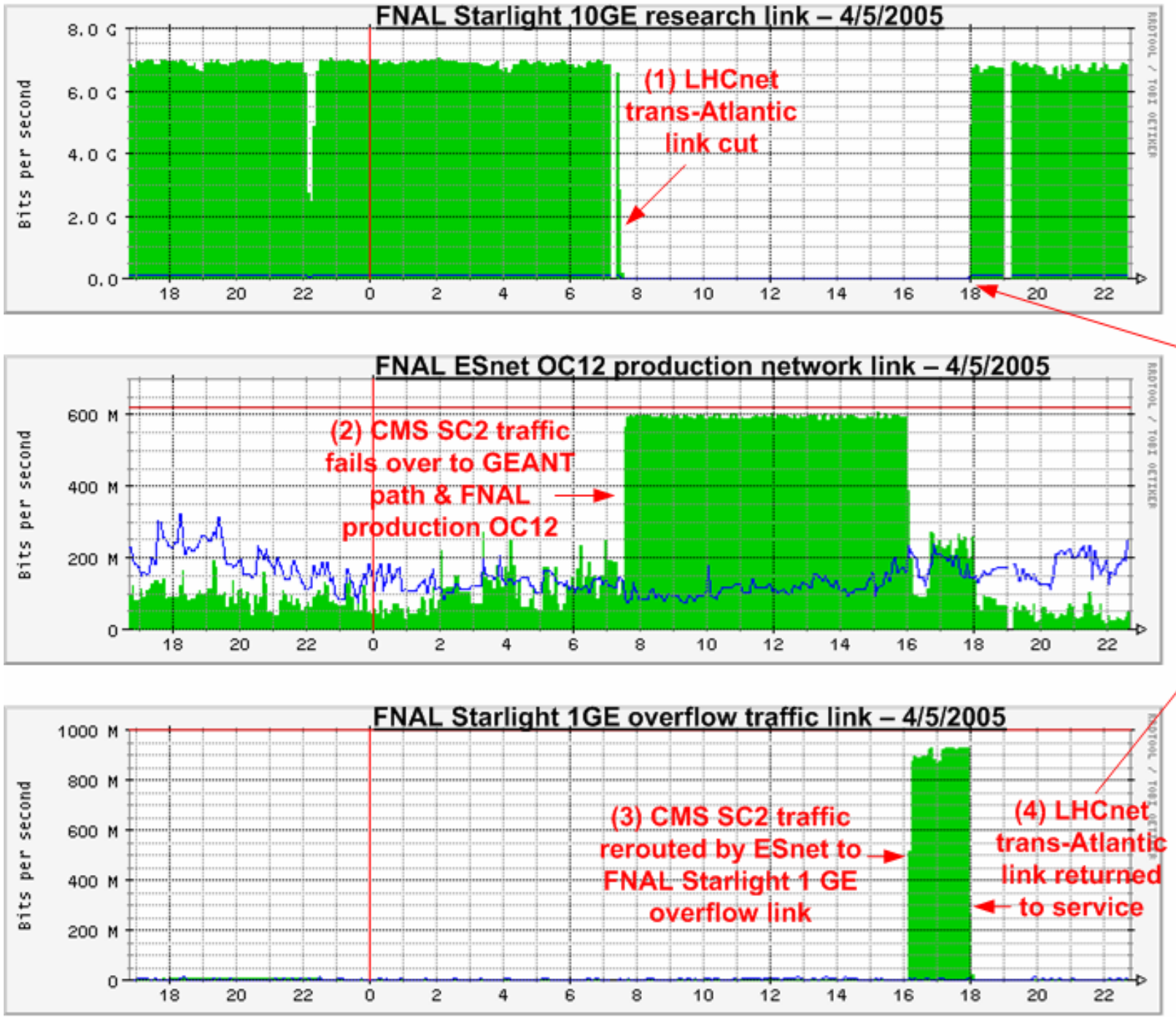
Division of Data between sites

Site	Average throughput (MB/s)	Data Moved (TB)
BNL	61	51
FNAL	61	51
GridKA	133	109
IN2P3	91	75
INFN	81	67
RAL	72	58
SARA	106	88
TOTAL	600	500

Individual site tests

- Overlapped with LCG Storage Management Workshop
  - Sites can pick days in next two weeks when they have the capacity
  - 500MB/s to disk
  - 60MB/s to tape
- FNAL was running 500MB/s disk tests at the time…




FNAL Starlight 10GE research link – 4/5/2005
Bits per second
8.0 G
6.0 G
4.0 G
2.0 G
0.0
(1) LHCnet trans-Atlantic link cut
FNAL ESnet OC12 production network link – 4/5/2005
600 M
400 M
200 M
0
(2) CMS SC2 traffic fails over to GEANT path & FNAL production OC12
FNAL Starlight 1GE overflow traffic link – 4/5/2005
1000 M
800 M
600 M
400 M
200 M
0
(3) CMS SC2 traffic rerouted by ESnet to FNAL Starlight 1 GE overflow link
(4) LHCnet trans-Atlantic link returned to service
18 20 22 0 2 4 6 8 10 12 14 16 18 20 22

SC2 Summary

- SC2 met it's **throughput** goals – and with **more** sites than originally planned!
  - A big improvement from SC1
- ☹ **But we still don't have something we can call a service**
  - Monitoring is better
  - We see outages when they happen, and we understood why they happen
    - First step towards operations guides
- Some advances in infrastructure and software will happen before SC3
  - gLite file transfer software
  - SRM service more widely deployed
- We have to understand how to incorporate these elements

SC1/2 - Conclusions

- Setting up the infrastructure and achieving reliable transfers, even at much lower data rates than needed for LHC, is complex and requires a lot of technical work + coordination
- Even within one site – people are working very hard & are stressed. Stressed people do not work at their best. Far from clear how this scales to SC3/SC4, let alone to LHC production phase
- Compound this with the multi-site / multi-partner issue, together with time zones etc and you have a large "non-technical" component to an already tough problem
- But… the end point is fixed (time + functionality)
- We should be careful not to over-complicate the problem or potential solutions
- And not forget there is still a humungous amount to do…
- (much much more than we've done…)

Service Challenge 3

Goals and Timeline for Service Challenge 3

(so far we have the "challenge", but not the "service"…)

2005 Q1 - SC3 preparation

Prepare for the next service challenge (SC3)
-- in parallel with SC2 (reliable file transfer) –

Build up 1 GByte/s *challenge* facility at CERN

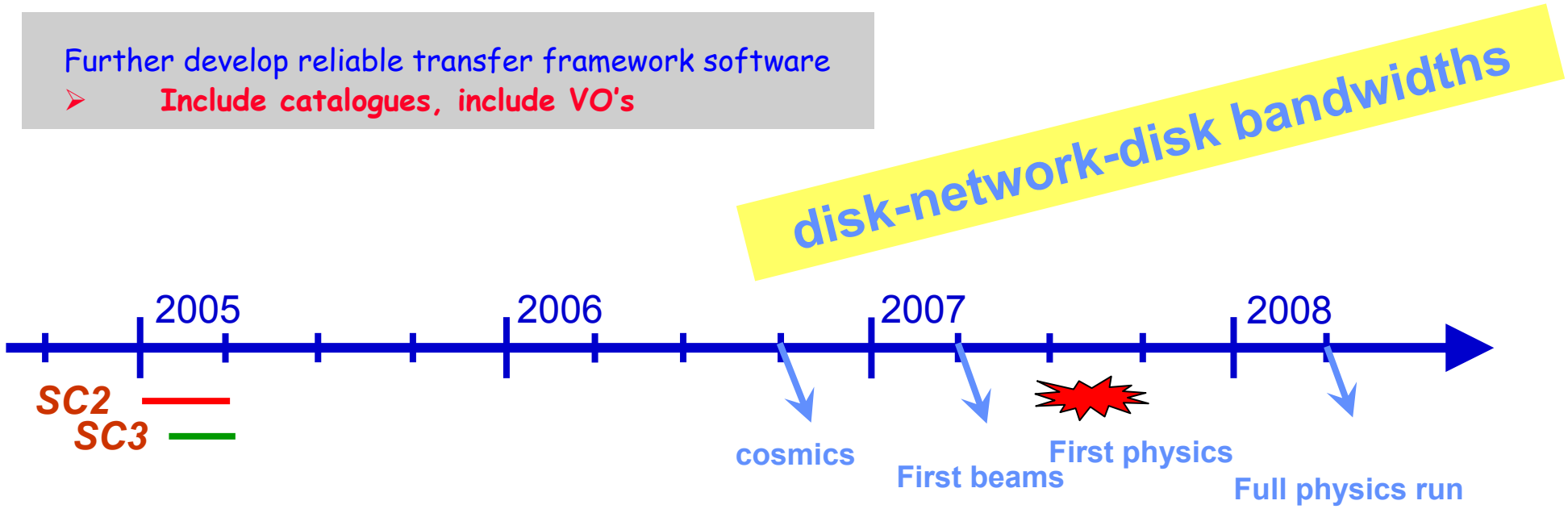
- **The current 500 MByte/s facility used for SC2 will become the *testbed* from April**

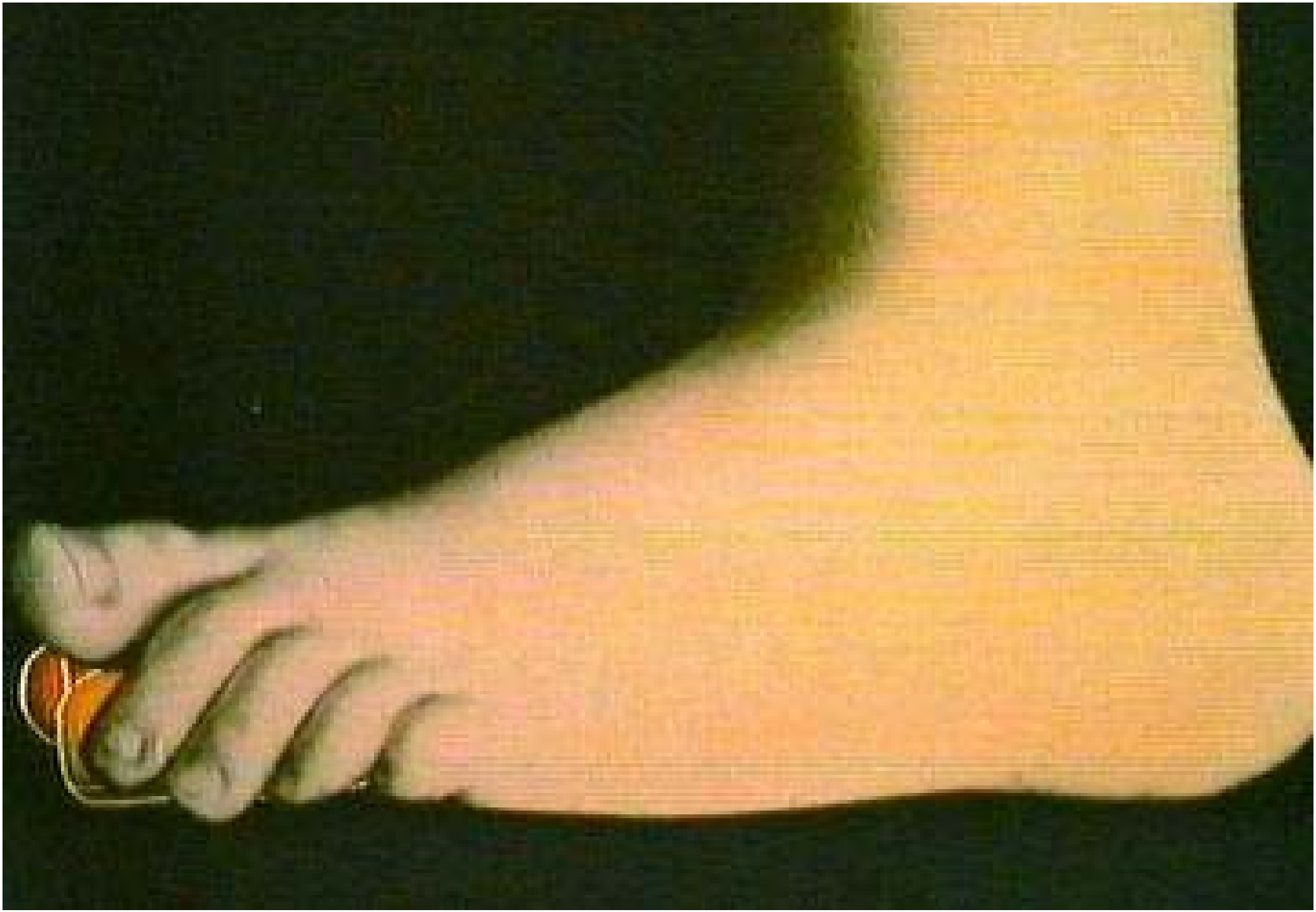
Build up infrastructure at each external centre

- **Average *capability* ~150 MB/sec at a Tier-1 (to be agreed with each T-1)**

Further develop reliable transfer framework software

- **Include catalogues, include VO's**





SC3 – Milestone Decomposition

- File transfer goals
- Tier1 goals
- Tier2 goals
- Experiment goals
- Service goals

LC 3
“ ”

Agenda

- Reminder of high-level time-line of SC3
- Deadlines; decision points; detailed time-line
- On-going procedure for monitoring progress:
  - Weekly meetings at CERN (Wed 11:00);
  - Weekly con-calls (Wed 17:30);
  - Daily (hourly?) informal follow-up…
- **Need higher level monitoring, particularly July 1st on…**
- This is one of the weak areas that still needs to be addressed in order to deliver a **SERVICE**

Service Challenge 3 - Phases

High level view (there is much more structure than this…):

- Setup Phase (including throughput tests…)
  - 2 weeks sustained in July 2005
    - "Obvious target" – GDB of July 20th
  - Primary goals:
    - 150MB/s disk – disk to Tier1s;
    - 60MB/s disk (T0) – tape (T1s)
  - Secondary goals:
    - Include a few named T2 sites (T2 -> T1 transfers)
    - **Encourage remaining T1s to start disk – disk transfers**
- Service phase
  - September – end 2005
    - Start with ALICE & CMS, add ATLAS and LHCb October/November
    - All offline use cases except for analysis
    - More components: WMS, VOMS, catalogs, experiment-specific solutions
  - Implies production setup (CE, SE, …)

SC3 – Milestone Decomposition

- File transfer goals:
  - Build up disk – disk transfer speeds to 150MB/s
  - Include tape – transfer speeds of 60MB/s
  - T2-T1 transfers: ~3 x 1GB files / hour over several days

☺ Tier1 goals:
  - Bring in additional Tier1 sites wrt SC2
  - **All currently named T1s will perform transfers, although not necessarily full rate**

☺ Tier2 goals:
  - Start to bring Tier2 sites into challenge
    - Agree services T2s offer / require
    - On-going plan (more later) to address this via GridPP, INFN, HEPiX, FNAL, BNL etc.

- Experiment goals:
  - Address main offline use cases ***except*** those related to analysis
    - i.e. real data flow out of T0-T1-T2; simulation in from T2-T1

- Service goals:
  - Include CPU (to generate files) and storage
  - Start to add additional components
    - Catalogs, VOs, experiment-specific solutions etc, 3D involvement, …
    - Choice of software components, validation, fallback, …

SC3 – Deadlines and Deliverables

- May 31st 2005: basic components delivered and in place
- June 2005: integration testing
- June 13 – 15: SC3 planning workshop at CERN – experiment issues
- June 30th 2005: integration testing successfully completed
- July 1 – 10: start disk – disk throughput tests
  - Assume a number of false starts / difficulties
- July 11 – 20: disk tests
- July 21 – 27: tape tests
- July 28 – 31: T2 tests

Courtesy of I. Bird, LCG GDB, May 2005

Baseline services

- Storage management services
  - Based on SRM as the interface
- gridftp
- Reliable file transfer service
- X File placement service – perhaps later
- Grid catalogue services
- Workload management
  - CE and batch systems seen as essential baseline services,
  - ? WMS not necessarily by all
- Grid monitoring tools and services
  - Focussed on job monitoring – basic level in common, WLM dependent part

- VO management services
  - Clear need for VOMS – limited set of roles, subgroups
- Applications software installation service
- From discussions added:
  - Posix-like I/O service → local files, and include links to catalogues
  - VO agent framework

Basic Components For SC3 Setup Phase

- **Each T1 to provide 10Gb network link to CERN**
- **Each T1 + T0 to provide SRM 1.1 interface to managed storage**
  - This goes for the named T2s for the T2-T1 transfer tests too
- T0 to provide File Transfer Service
- Also at named T1s for T2-T1 transfer tests
  - BNL, CNAF, FZK, RAL using FTS
  - FNAL and PIC will do T1<->T2 transfers for CMS using PhEDEx
- Baseline Services Working Group, Storage Management Workshop and SC3 Preparation Discussions have identified one additional data management service **for SC3**, namely the LFC
  - Not all experiments (ALICE) intend to use this
  - Nor will it be deployed for all experiments at each site
- However, as many sites support multiple experiments, and will (presumably) prefer to offer common services, this can be considered a basic component

Dedicated connections for SCs

Tier1	Location	NRENs	Status dedicated link
ASCC	Taipei, Taiwan	ASnet, SURFnet	1 Gb via SURFnet, testing
BNL	Upton, NY, USA	ESnet, LHCnet	622 Mbit shared
CNAF	Bologna, Italy	Geant2, GARR	1 Gb now, 10 Gb in Sept
FNAL	Batavia, ILL, USA	ESnet, LHCnet	10 Gb, tested
IN2P3	Lyon, France	Renater	1 Gb now, 10 Gb in Sept
GridKa	Karlsruhe, Germany	Geant2, DFN	10 Gb, tested
SARA	Amsterdam, NL	Geant2, SURFnet	10 Gb, testing
NorduGrid	Scandinavia	Geant2, Nordunet	Would like to start performing transfers
PIC	Barcelona, Spain	RedIris, Geant2	Will participate in SC3 but not full rate
RAL	Didcot, UK	Geant2, Ukerna	2 x 1 Gb via SURFnet soon
Triumf	Vancouver, Canada	Canet, LHCnet	1 Gb via SURFnet, testing

Agreement on SRM functionality

- Basic understanding
  - SRM V1.1 is not sufficient
  - Full functional set of SRM V2.1 and later is not required
  - "LCG-required" functionality agreed – Baseline services group and Storage Management workshop
- For SC3
  - V1.1 is sufficient
- For SC4
  - LCG-set is required
  - Workshop put in place a group (developers and experiments) to plan and monitor progress
  - Summary of prioritised requirements and key use cases (N. Brook)
  - Implementation and production deployment concerns (M. Ernst)
  - ➢ Decision(?)

LCG-required SRM functions

- SRM v1.1 insufficient – mainly lack of pinning
- SRM v3 not required – and timescale too late
- Require Volatile, Permanent space; Durable not practical
- Global space reservation: reserve, release, update (mandatory LHCb, useful ATLAS,ALICE). Compactspace NN
- Permissions on directories mandatory
  - Prefer based on roles and not DN (SRM integrated with VOMS desirable but timescale?)
- Directory functions (except mv) should be implemented asap
- Pin/unpin high priority
- srmGetProtocols useful but not mandatory
- Abort, suspend, resume request : all low priority
- Relative paths in SURL important for ATLAS, LHCb, not for ALICE

SRM status at Tier 1 sites

- CERN
  - Castor – in production; update to new Castor for SC3
- FNAL
  - In production; dCache
- BNL
  - dCache – in production, used in SC2. ??? status ???
- CC-IN2P3
  - dCache – under test; planned for SC3
- PIC
  - Castor – in production; update to new Castor as CERN
- CNAF
  - Castor – in production; update to new Castor as CERN
- FZK
  - dCache – testing; not yet in production – planned for SC3
- ASCC
  - Castor – not yet in production – install this month on new hardware.
- RAL
  - dCache used in production (disk only). Tape backend used successfully in SC2. This should be production for SC3.
- NIKHEF/SARA
  - dCache/SRM on test cluster. Expect production in SC3.
- NDGF
  - No information

Summary of SRM Discussions

1. All 3 experiments consider SRM v1.1 insufficient, especially because of lack of pinning. They do not require any method from V3 (for example streaming mode is seen as useful but not mandatory) and they consider V3 as available too late.

2. They all require spaces Volatile and Permanent. Durable space requires too much intervention.

3. Global Space reservation ReserveSpace, ReleaseSpace and UpdateSpace are seen as useful by Atlas and Alice, mandatory by LHCb. CompactSpace is not needed.

4. Permission functions: All 3 experiments see the permissions on directories as mandatory, would like them to be based on roles and not DNs. They would be happy to get SRM integrated with VOMS but are sceptical about timescale.

5. Directory functions - should be implemented with high priority (except mv).

6. Pin/UnPin get high priority

7. There are a lot of concern about the duplication between FTS and srmCopy. All 3 experiments would like to see one of them as reliable and not 2 overlapping methods.

8. srmGetProtocols is seen as useful but not really mandatory

9. AbortRequest, SuspendRequest and ResumeRequest get a very low ranking.

10. Relative paths in SURLs very important by for Atlas and LHCb, but not Alice.

CMS had not provided written feedback at time of writing…

Status of SRM Con-Calls

- Despite a well attended and useful call during the Storage Management workshop, it has not been possible to keep this up since.
- Service Challenge 4 milestones require production deployments of components latest end January 2006
- Given the time required for implementation and production deployment, this requires agreement before the summer!
- The following proposal has been sent to the PEB (and SRM con-call mailing list…):

SRM Proposal

- Ask all experiments to prioritize the above list giving key use cases
  - Nick Brook will report on these on Wednesday June 15th during SC3 planning workshop
- Ask SRM developers to respond to this list in terms of implementation and deployment issues / concerns
  - An expert – not an SRM developer – will report on these following Nick's talk

➢ **The assumption is that at least the top priority enhancements will be formally requested by the experiments for production deployment latest SC4**

Summary of SRM Discussions

1. …
2. They all require spaces Volatile and Permanent. Durable space requires too much intervention.
3. Global Space reservation ReserveSpace, ReleaseSpace and UpdateSpace are seen as useful by Atlas and Alice, mandatory by LHCb. CompactSpace is not needed.
4. Permission functions: All 3 experiments see the permissions on directories as mandatory, would like them to be based on roles and not DNs. They would be happy to get SRM integrated with VOMS but are sceptical about timescale.
5. Directory functions - should be implemented with high priority (except mv).
6. Pin/UnPin get high priority
7. …
8. …
9. …
10. Relative paths in SURLs very important by for Atlas and LHCb, but not Alice.

If you remove low priority items and statements, the list is shortened to the above. At least some of these are relatively straightforward to implement.

Additional Components

- Whilst the above presents a 'simple' model, there are clearly additional applications / 'services' that need to be provided
- These include "Agents and Daemons" (next);
- Applications Area-related services;
  - E.g. COOL for ATLAS; LHCb, …
- Applications Area-software;
  - E.g. GEANT4, ROOT, POOL, SEAL, …
- Experiment-specific software and services… **(many using / requiring MySQL…)**
  - E.g. ATLAS HVS, book-keeping systems (and a lot more mentioned during Rome production…)
  - CMS will use PhEDEx for all transfers

➢ **We have to document all of these (and any others on the critical path for SC3)…**

- Cannot afford for production to stop due to a lack of documentation / procedures…
- Many of these also have implications on Tier1s and Tier2s…

File Catalog

Status end-May 2005:

- ALICE will use their own catalog
- ATLAS and CMS require local catalogs at all sites
- LHCb requires a central catalog and would like 1-2 R/O copies
- For the time being we are setting up an LFC system at CERN
  - Actually 2 (or more): a "pilot" to expose functionality; an SC3 service
- Need more clarification on catalog deployment model(s)
- LFC on MySQL or Oracle likely at most / many sites(?)

Agents and Daemons

Agents and Daemons

- This is something clearly identified during BSWG discussions
- And corresponds to what is running on lxgate machines at CERN today…
- Current assumption: Each experiment requires **ONE** such box at **EACH** site
  - T0, T1, T2 (some expt ppts suggest even more…)
- Must agree on minimal service level:
- Standard system; installation; box-level monitoring; intervention procedures, etc.
- These are – by definition – critical components of the production system and hence must be treated as such
- I believe that we need separate instances of these for SC3
  - And not mix with on-going production / other work
- (I also doubt that lxbatch style machines are the right answer for "service machines" but this can perhaps be deferred…)

LCG Service Challenges: Planning for Tier2 Sites

Update for HEPiX meeting

Jamie Shiers
IT-GD, CERN

T2 Executive Summary

- Tier2 issues have been discussed extensively since early this year
- The role of Tier2s, the services they offer – and require – has been clarified
- The data rates for MC data are expected to be rather low (limited by available CPU resources)
- The data rates for analysis data depend heavily on analysis model (and feasibility of producing new analysis datasets IMHO)
- LCG needs to provide:
  - Installation guide / tutorials for DPM, FTS, LFC
- **Tier1s need to assist Tier2s in establishing services**

	Number of T1s	Number of T2s	Total T2 CPU	Total T2 Disk	Average T2 CPU	Average T2 Disk	Network In	Network Out
			KSI2K	TB	KSI2K	TB	Gb/s	Gb/s
ALICE	6	21	13700	2600	652	124	0.010	0.600
ATLAS	10	30	16200	6900	540	230	0.140	0.034
CMS	6 to 10	25	20725	5450	829	218	1.000	0.100
LHCb	6	14	7600	23	543	2	0.008	0.008

A Simple T2 Model

N.B. this may vary from region to region

- Each T2 is configured to upload MC data *to* and download data *via* a given T1
- In case the T1 is logical unavailable, **wait and retry**
  - MC production might eventually stall
- For data download, **retrieve** via **alternate** route / T1
  - Which may well be at lower speed, but hopefully rare
- Data residing at a T1 other than 'preferred' T1 is transparently delivered through appropriate network route
  - T1s are expected to have at least as good interconnectivity as to T0
- Each Tier-2 is associated with a Tier-1 who is responsible for getting them set up
- Services at T2 are **managed storage** and **reliable file transfer**
  - DB component at T1; user agent also at T2
- 1GBit network connectivity – shared (less will suffice to start with, more maybe needed!)

Tier2 and Base S/W Components

1) Disk Pool Manager (of some flavour…) with SRM 1.1 i/f
   - e.g. dCache, DPM, …
2) gLite FTS client (and T1 services)
3) Possibly / Probably also local catalog (ATLAS, CMS)
   - e.g. LFC…
4) **Experiment-specific s/w and services ('agents')**

Must be run as SERVICES!

1 – 3 will be bundled with LCG release.

Experiment-specific s/w will not…

[N.B. we are talking **interfaces** and not implementation]

→ We are still focussing on the infrastructure layer; the experiment-specific requirements for the Service Phase are still being collected

Tier2s and SC3

- Initial goal is for a small number of Tier2-Tier1 partnerships to setup agreed services and gain experience
- This will be input to a wider deployment model
- Need to test transfers in both directions:
  - MC upload
  - Analysis data download
- Focus is on **service** rather than "throughput tests"
- As initial goal, would propose running transfers over at least several days
  - e.g. using 1GB files, show sustained rates of ~3 files / hour T2->T1
- More concrete goals for the Service Phase will be defined together with experiments in the coming weeks
  - Definitely no later than June 13-15 workshop

➢ Experiment-specific goals for SC3 Service Phase still to be identified…

Initial Tier-2 sites

- For SC3 we aim for (updated from input at May 17 GDB):

Site	Tier1	Experiment
Legnaro, Italy	CNAF, Italy	CMS
Milan, Italy	CNAF, Italy	ATLAS
Turin, Italy	CNAF, Italy	Alice
DESY, Germany	FZK, Germany	ATLAS, CMS
Lancaster, UK	RAL, UK	ATLAS
Imperial, UK	RAL, UK	CMS
Edinburgh, UK	RAL, UK	LHCb
US Tier2s	BNL / FNAL	ATLAS / CMS

- Training in UK May 13th and in Italy May 26-27th. Training at CERN June 16th.
- Other interested parties: Prague, Warsaw, Moscow, ..
- **Addressing larger scale problem via national / regional bodies**
  - GridPP, INFN, HEPiX, US-ATLAS, US-CMS, Triumf (Canada)
- Cannot handle more for July tests, but please let us know if you are interested! (T1+T2 partnerships)

T2s – Concrete Target

- We need a **small number** of **well identified** T2/T1 partners for SC3 as listed above
- Initial target of end-May is not realistic, but not strictly necessary either…
- Need prototype service in at least two countries by end-June
- Do not plan to strongly couple T2-T1 transfers to T0-T1 throughput goals of SC3 setup phase
- **Nevertheless, target one week of reliable transfers T2->T1 involving at least two T1 sites each with at least two T2s by end July 2005**

Tier2 participation by Tier1

ASCC, Taipei	No known plans
CNAF, Italy	Yes; workshop held last week in Bari
PIC, Spain	Yes; no Oracle service for FTS; CMS transfers with PhEDEx
IN2P3, Lyon	Yes; LAL + IN2P3
GridKA, Germany	Yes – study with DESY
RAL, UK	Yes – plan in place for several Tier2s
BNL, USA	Yes – named ATLAS Tier2s
FNAL, USA	Yes – CMS transfers with PhEDEx; already performing transfers
TRIUMF, Canada	Yes – planning to install FTS and identify T2s for tests
NIKHEF/SARA, Netherlands	No known plans
Nordic Centre	N/A

Significantly further advanced than foreseen at beginning of year (or May GDB for that matter…)

SC3 – Setup Phase Recap

- Discussions over the past months have led to clarification on the services required for the 'infrastructure'
- All of these need to be in place for the Setup Phase – July
- The basic requirements for most sites – 10Gb network connectivity to CERN + production-quality SRM 1.1 service – are not new
- LFC and FTS for those sites concerned require Oracle or MySQL backend service. A MySQL port of FTS is clearly a post-SC3 priority
  - Unless someone could actually help now…
- Deployment model at CERN:
  - Dedicated disk server for the Oracle database service for each
  - Farm node for middle tier per VO plus one spare
  - How this deployment model works can be considered part of the 'Service' challenge…

➢ **There is still an awful lot to do just for these services…**

➢ **And this is the basic infrastructure, on top of which we need to consider the experiments' needs…**

Status of Core Components

- Basic infrastructure and core services more or less in place at CERN and some Tier1s
- But it took a lot of effort to get here…
- Goal is to announce services, including problem reporting procedures and internal problem resolution procedures, prior to June SC3 workshop (13 – 15)
- SRM 1.1 for new CASTOR on target: PEB review June 7 of CASTOR deployment plans
- Main outstanding issues: understanding of experiment requirements and deployment model
- Target: prior to June workshop with full discussion at workshop
- The model of involving Tier1s and Tier2s through workshops, phone meetings, site visits etc is working!

SC3 – Experiment Goals

- Meetings on-going to discuss goals of SC3 and experiment involvement
- Focus on:
  - First demonstrate robust infrastructure;
  - Add 'simulated' experiment-specific usage patterns;
  - Add experiment-specific components;
  - Run experiments offline frameworks but don't preserve data;
    - Exercise primary Use Cases ***except*** analysis (SC4)
  - Service phase: data is preserved…

➢ **Has significant implications on resources beyond file transfer services**
- Storage; CPU; Network… Both at CERN and participating sites (T1/T2)
- May have different partners for experiment-specific tests (e.g. not all T1s)

➢ **In effect, experiments' usage of SC during service phase = data challenge**

- Must be **exceedingly clear** on goals / responsibilities during each phase!

SC3 Preparation Workshop

- This workshop will focus on very detailed technical planning for the whole SC3 exercise.
- **It is intended to be as interactive as possible, i.e. not presentations to an audience largely in a different (wireless) world.**
- There will be sessions devoted to specific experiment issues, Tier1 issues, Tier2 issues as well as the general service infrastructure.
- This is an opportunity to get together to iron out concerns and issues that cannot easily be solved by e-mail, phone conferences and/or other meetings prior to the workshop.

➢ **Dates: June 13 – 15: B160 1-009 for first 2 days then 513 1-024**

  ➢ **4 x ½ days on Experiment-specific issues (ALICE, ATLAS, CMS, LHCb)**

http://agenda.cern.ch/fullAgenda.php?ida=a051784

SC3 Workshop Agenda (June 13-15)

- Experience from recent Grid data challenges and productions:
  - What worked, what didn't?
  - What are the key issues to be addressed?
- Detailed experiment goals for Service Challenge 3:
  - What are the primary / secondary / optional objectives?
  - How will do you define success / partial success / failure?
- Detailed experiment plans:
  - What are the services you need in place? (core infrastructure, experiment-specific agents and so forth, ...),
  - What environment do you depend on? (platforms, O/S, ...),
  - What applications area packages and specific releases,
  - What other external packages and releases, experiment-specific packages and release?

http://agenda.cern.ch/fullAgenda.php?ida=a051784

SC3 Experiment Status

- Experiment goals for the Service Phase are becoming clearer
- ➢ Numerous experiment-specific services dependent on MySQL
- Need to finalise plan in coming weeks, including overall schedule and confirmation of resources at all sites
  - Assumed to come from existing pledges
- Need to be very clear on responsibilities
- Setup for service phase overlaps the July SC3 setup phase
  - And has in some cases already started…
- Once again, the goal of the SCs is to deliver THE SERVICE

ALICE & LCG Service Challenge 3

- Goal:
  - **test of data transfer and storage services (SC3)**
  - test of distributed reconstruction and data model (ALICE)
- Use case 1: RECONSTRUCTION
  - Get "RAW" events stored at T0 from ALICE Catalogue
  - Reconstruct at T0 (at least partially)
  - Ship from T0 to T1's (goal: 500 MB/S out of T0)
  - Reconstruct at T1 with calibration data
  - Store/Catalogue the output
- Use Case 2: SIMULATION
  - Simulate events at T2's
  - Same as previous replacing T0 with T2

ATLAS & SC3

- April-July: Preparation phase
  - Test of FTS
  - Integration of FTS with DDM
- July: Scalability tests (commissioning data; Rome Physics workshop data)
- September: test of new components and preparation for real use of the service
  - Intensive debugging of COOL and DDM
  - Prepare for "scalability" running
- Mid-October
  - Use of the Service
  - Scalability tests of all components (DDM)
  - Production of real data (MonteCarlo; Tier-0; …)
- Later
  - "continuous" production mode (data needed for ATLAS DC3)
  - Re-processing
  - Analysis

CMS SC3 Schedule

- July: throughput phase
  - Optional leading site-only tuning phase, may use middleware only
  - T0/T1/T2 simultaneous import/export using CMS data placement and transfer system (PhEDEx) to coordinate the transfers
  - Overlaps setup phase for other components on testbed; will not distract transfers – setting up e.g. software installation, job submission etc.
- September: service phase 1 — modest throughput
  - Seed transfers to get initial data to the sites
  - Demonstrate bulk data processing, simulation at T1, T2s
    - Requires software, job submission, output harvesting, monitoring, …
    - Not everything everywhere, something reasonable at each site
- November: service phase 2 — modest throughput
  - Phase 1 + continuous data movement
  - Any improvements to CMS production (as in MC production) system
    - Already in September if available then

Post-SC3 Service

- Experiments will start taking data with cosmics from late 2005 / early 2006

- **REQUIRE a reliable, long-term service from that time for primary use cases**
  - Data taking, archiving, processing, distribution to Tier1s etc

- And at the same time, perform final preparations for SC4…

SC4

- SC4 starts April 2006
- It ends with the deployment of the FULL PRODUCTION SERVICE

➢ **Deadline for component (production) delivery: end January 2006**

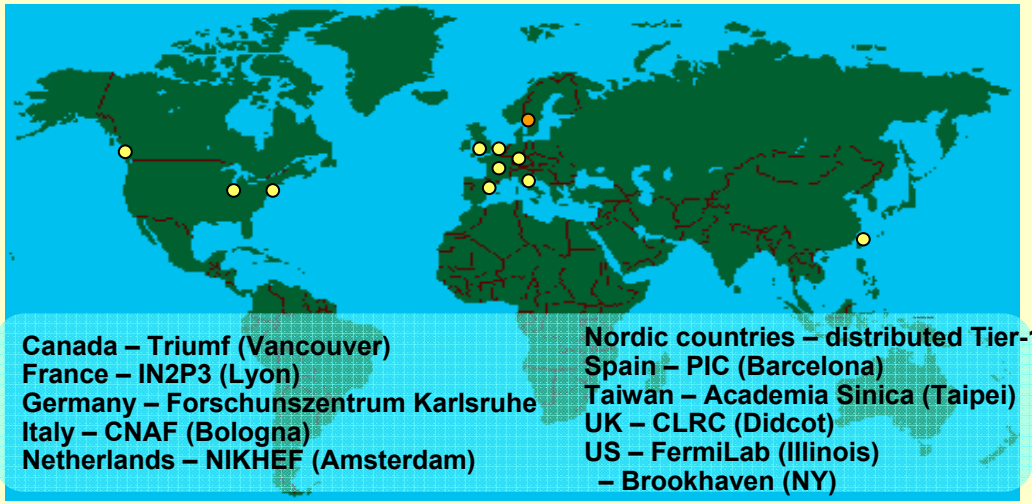
- Adds further complexity over SC3
  - Additional components and services
  - Analysis Use Cases
  - SRM 2.1 features required by LHC expts
  - All Tier2s (and Tier1s…) at full service level
  - Anything that dropped off list for SC3…
  - xrootd?? proof??

- "Analysis" clearly needs significant thought and work
  - Step 1: define and agree common terms for primary Use Cases
  - And never say 'analysis' again… (without qualifying it…)
  - e.g. "AOD production", "interactive ROOT analysis", …

LCG Service Hierarchy

Tier-0 – the accelerator centre

- Data acquisition & initial processing
- Long-term data curation
- Distribution of data → Tier-1 centres



Tier-1 – "online" to the data acquisition process → high availability

- Managed Mass Storage – → grid-enabled data service
- Data-heavy analysis
- National, regional support

Tier-2 – ~100 centres in ~40 countries

- Simulation
- End-user analysis – batch and interactive

Mandatory Services

- The (DM) services LCG absolutely must provide (IMHO) are:

1. **Management of the primary copies of the RAW data, ESD etc.**
2. **Reliable file transfer service to Tier1s (including networking)…**

- **Baseline services: as per Baseline Services Working Group**
- Additional services include file catalog, conditions database etc.
  - We managed these for LEP and LEP-era experiments

➢ **The focus must be on the mandatory services which simply cannot be called 'rock solid' today**

Summary

- We are just about on target with delivering the service components required for SC3
  - Possible slippage is in days.. but we will try to get there on time…
- Further clarification between now and June workshop on precise goals of experiments, additional experiment software and services, resource allocation and schedule
- Expect detailed presentations at June workshop, including experience with current services
  - e.g. ATLAS experience from ROME production
- We need to catalog a list of issues to address, rather than just 'a feeling of frustration…'

LCG Service Challenges – Deploying the Service

Concerns

- All come down to one word – SERVICE
- Far from clear that we will be ready to offer an acceptable service level
- Far from clear that people see this as a key priority for IT
- Significant internal inefficiencies experienced in attempting to setup base services – do we have the best structures and procedures for this task?
- Admittedly compounded by unclear messages / uncertainty from experiments
- But we've lost MONTHS – not days or weeks!

Conclusions

- To be ready to fully exploit LHC, significant resources need to be allocated to a series of **Service Challenges** by all concerned parties
- These challenges should be seen as an **essential** on-going and **long-term** commitment to achieving production LCG
- The countdown has started – we are already in (pre-)production mode
- Next stop: 2020

The Service is the Challenge