LCG Service Challenges: Status and Plans

Jamie.Shiers@cern.ch – http://cern.ch/jamie/

July 2005

Service Challenges 3 & 4

- Service Challenge 3:
  - Where do we stand?
  - What remains to be done?
- Service Challenge 4 – time to look ahead…
  - Component delivery end Jan 2006;
  - Throughput phase April, Service from May…
- Service Challenge 4 planning:
  - Possible workshop in September
    - We need to actively review SC3 progress as well as make concrete steps in SC4 planning even without workshop!
  - Probable workshop in February immediately before CHEP
- Important that Tier1 representatives, larger Tier2s and all experiments are adequately represented!
  - Network issues?

LHC Computing Grid

Distrubuted Production Environment for Physics Data Processing

Known Knowns

- Model for 'Production' much debated and now well understood
  - All stages from data taking leading into to end-user analysis
  - Has been exercised through experiment data challenges
  - Will also be covered during Service Phase of SC3
    - Main goal is to thoroughly stress-test the service infrastructure
- Data types, rates, flows that correspond to above all 'clear'
  - Processing, re-processing, stripping, AOD / TAG Production etc
- Roles played by different tiers, services that they offer, services that they require etc also understood
  - Services still not fully setup; in some cases software maybe…
  - Still a large number of Tier2s with no clear Tier1
  - Expect to continue to make good progress on this prior to SC4
- Current plan is for 50 days of data taking in 2007 @ $x.10^{32}$ $cm^{-2}s^{-1}$
- No change in Service Challenge schedule or delivery of production system

Known Unknowns

- End-user analysis still a mystery
  - Can easily result in significant network bandwidth / support load
  - What is the model for Analysis Facilities?
    - Dedicated PROOF farms? 100+ nodes, 50+TB disk
    - Batch mode? Single stream? Parallel?
- Startup phase of LHC unknown
  - It will certainly not be like steady-state
  - Strong pressure to exploit **(needed)** distributed resources
  - There will be a strong presence at CERN, but nevertheless fundamental need to allow detector / physics groups outside have rapid / peer access to the data
- How to provide reliable distributed services, '24 x 7' …
  - More on this later…

LCG Service Hierarchy

Tier-0 – the accelerator centre

- Data acquisition & initial processing
  - Close to 2GB/s during AA running
- Long-term data curation
- Distribution of data → Tier-1 centres
  - ~200MB/s per site; ~12 sites

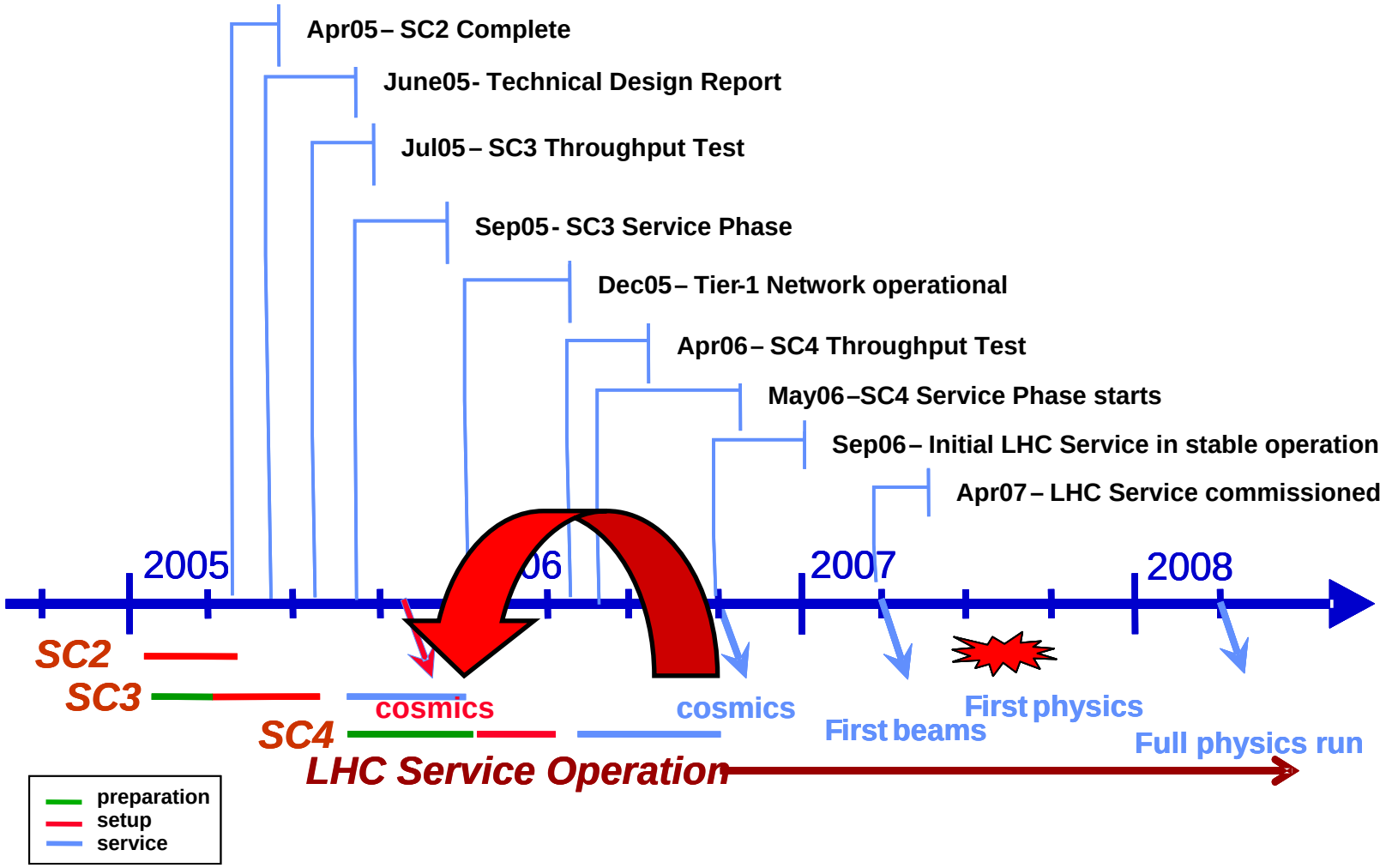
Tier-1 – "online" to the data acquisition process → high availability

- Managed Mass Storage – → grid-enabled data service
- Data intensive analysis
- National, regional support
- 10Gbit/s dedicated links to T0
- (+ significant inter-T1 traffic)

Tier-2 – ~100 centres in ~40 countries

- Simulation
- End-user analysis – batch and interactive
- 1Gbit/s networks

LCG Deployment Schedule



SC3 – Future Milestones (TDR)

Date	Description
31 July 05	**Service Challenge 3 Set-up: Set-up complete and basic service demonstrated. Performance and throughput tests complete. See Section** 6.2.4 for detailed goals.
1 Sept 05	**Service Challenge 3: start of stable service phase, including at least 9 Tier-1 and 10 Tier-2 centres.**
31 Dec 05	**Tier-0/1 high-performance network operational at CERN and 8 Tier-1s.**
31 Dec 05	**750 MB/s data recording demonstration at CERN: Data generator**→ disk → tape sustaining 750 MB/s for one week using the CASTOR mass storage system.
~~28 Feb 06~~ **31 Jan 06**	**All required software for baseline services deployed and operational at all Tier-1s and at least 20 Tier-2 sites.**

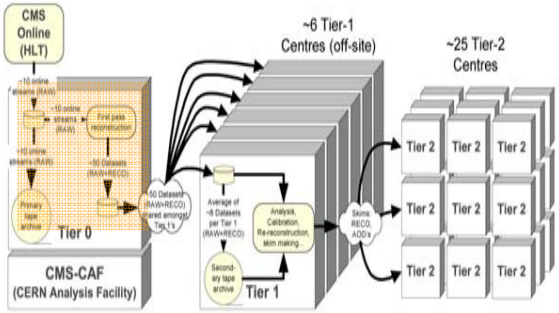
ATLAS Computing Model

- Tier-0:
  - Copy RAW data to Castor tape for archival
  - Copy RAW data to Tier-1s for storage and reprocessing
  - Run first-pass calibration/alignment (within 24 hrs)
  - Run first-pass reconstruction (within 48 hrs)
  - Distribute reconstruction output (ESDs, AODs & TAGS) to Tier-1s
- Tier-1s:
  - Store and take care of a fraction of RAW data
  - Run "slow" calibration/alignment procedures
  - Rerun reconstruction with better calib/align and/or algorithms
  - Distribute reconstruction output to Tier-2s **(and also inter-T1s)**
  - Keep current versions of ESDs and AODs on disk for analysis
- Tier-2s:
  - Run simulation
  - Keep current versions of AODs on disk for analysis

ATLAS event data flow from EF

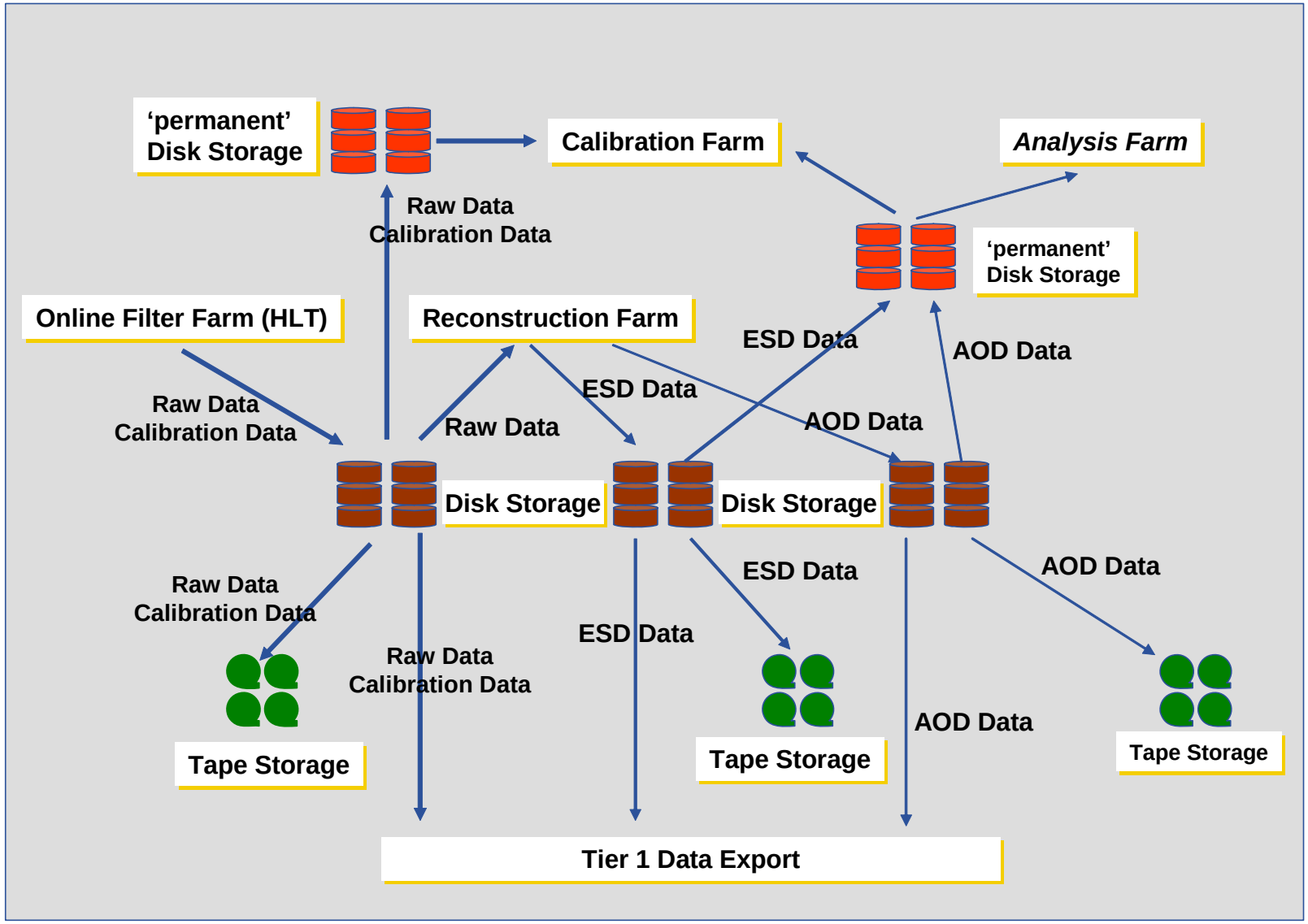
- Events are written in "ByteStream" format by the Event Filter farm in 2 GB files
  - ~1000 events/file (nominal size is 1.6 MB/event)
  - 200 Hz trigger rate (independent of luminosity)
  - Currently 4 streams are foreseen:
    - Express stream with "most interesting" events
    - Calibration events (including some physics streams, such as inclusive leptons)
    - "Trouble maker" events (for debugging)
    - Full (undivided) event stream
  - One 2-GB file every 5 seconds will be available from the Event Filter
  - Data will be transferred to the Tier-0 input buffer at 320 MB/s (average)
- The Tier-0 input buffer will have to hold raw data waiting for processing
  - And also cope with possible backlogs
  - ~125 TB will be sufficient to hold 5 days of raw data on disk

Tier-0 Center (CMS)



- Functionality
  - Prompt first-pass reconstruction
    - NB: Not all HI reco can take place at Tier-0
  - Secure storage of RAW&RECO, distribution of second copy to Tier-1
- Responsibility
  - CERN IT Division provides guaranteed service to CMS
    - Cast iron 24/7
  - Covered by formal Service Level Agreement
- Use by CMS
  - Purely scheduled reconstruction use; no 'user' access
- Resources
  - CPU 4.6MSI2K; Disk 0.4PB; MSS 4.9PB; WAN 5Gb/s

Tier0 Data Flows



Summary of Tier0/1/2 Roles

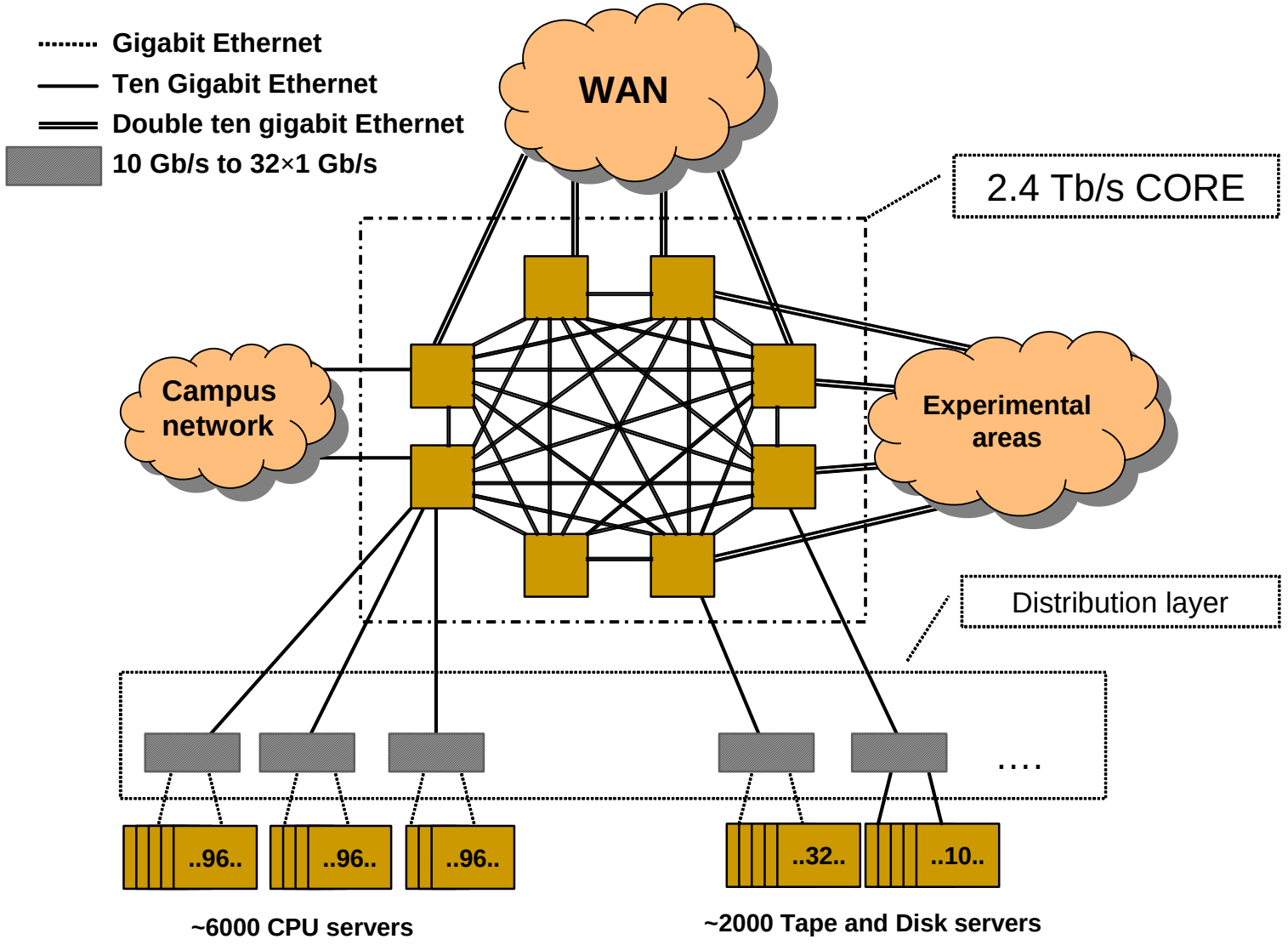
- Tier0 (CERN): safe keeping of RAW data (first copy); first pass reconstruction, distribution of RAW data and reconstruction output to Tier1; reprocessing of data during LHC down-times;
- Tier1: safe keeping of a proportional share of RAW and reconstructed data; large scale reprocessing and safe keeping 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!

Analysis Use Cases (HEPCAL II)

- Production Analysis (PA)
  - **Goals in Context** *Create AOD/TAG data from input for physics analysis groups*
  - **Actors** *Experiment production manager*
  - **Triggers** *Need input for "individual" analysis*
- (Sub-)Group Level Analysis (GLA)
  - **Goals in Context** *Refine AOD/TAG data from a previous analysis step*
  - **Actors** *Analysis-group production manager*
  - **Triggers** *Need input for refined "individual" analysis*
- End User Analysis (EA)
  - **Goals in Context** *Find "the" physics signal*
  - **Actors** *End User*
  - **Triggers** *Publish data and get the Nobel Prize :-)*

Tier0 Network



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

Experiment	T0	T1	T2	Total (PB)
ALICE	2.3	7.5	-	9.8
ATLAS	4.7	6.5	-	11.2
CMS	3.8	12.9	-	16.6
LHCb	1.359	2.074	-	3.433
Total (2008)	**12.2**			**41**

2008 requirements: ~linear increase with time (plus reprocessing)

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.

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, NL	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…

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

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

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…

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!)

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 (CMS figure, including overhead = every 3 weeks(!) – upper limit?)
- 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

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

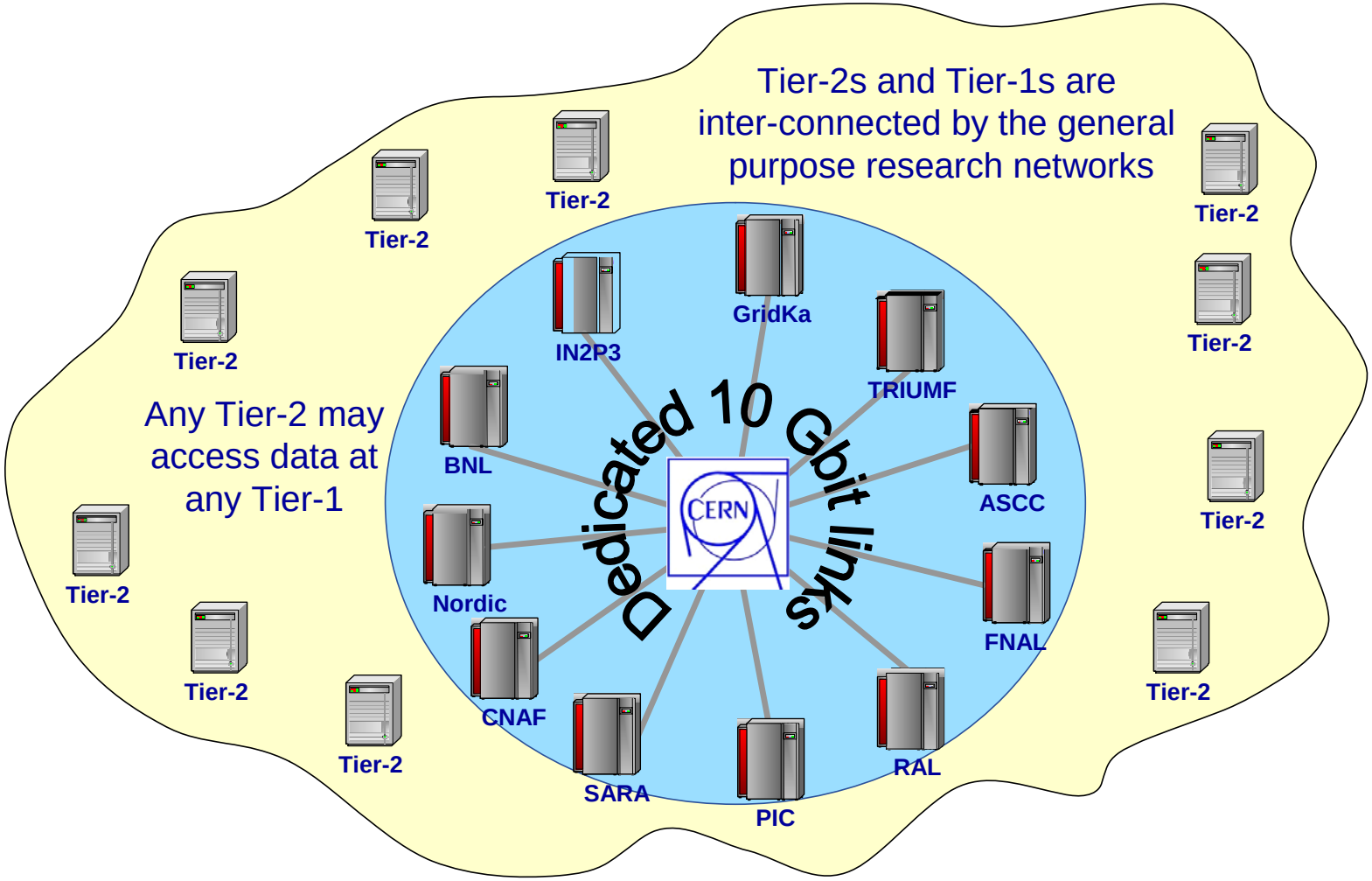
Tier1	(Approx) Status mid-June
ASCC, Taipei	Yes; preparing for T2 support in Asia - Pacific
CNAF, Italy	Yes; workshop held last week in Bari
PIC, Spain	Yes; no Oracle service for FTS; CMS transfers with PhEDEx
IN2P3, Lyon	Yes; LAL + IN2P3
GridKA, Germany	Yes – study with DESY
RAL, UK	Yes – plan in place 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	Yes; preparing T1 / T2s in Nordic region

- Virtually all Tier1s preparing for Tier2 support.
- Much interest from Tier2 side: hope to debug process rapidly
- Some Tier2s still need to identify their Tier1 centre

Tier1 / Tier2 Bandwidth Needs

	ALICE	ATLAS	CMS	LHCb
Parameters:				
Number of Tier-1s	6	10	7	6
Number of Tier-2s	21	30	25	14
Real data 'in-Tier-2':				
TB/yr	120	124	257	0
Mb/s (rough)	31.9	32.9	68.5	0.0
Mb/s (w. safety factors)	**95.8**	**98.6**	**205.5**	**0.0**
MC 'out-Tier-2':				
TB/yr	14	13	136	19
Mb/s (rough)	3.7	3.4	36.3	5.1
Mb/s (w. safety factors)	**11.2**	**10.2**	**108.9**	**15.3**
MC 'in-Tier-2':				
TB/yr	28	18	0	0
Mb/s (rough)	7.5	4.9	0	0.0
Mb/s (w. safety factors)	**22.5**	**14.7**	**0.0**	**0.0**

Tier0 / Tier1 / Tier2 Networks



Tier-2s

~100 identified – number still growing

Services (all 24 x 7, now – 2020)

- Managed storage: SRM 1.1 interface, moving to 2.1 (2.2)
  - No assumption / requirement for tertiary storage at T2s
  - Monte Carlo generation: write-through cache to T1s
  - Analysis data: read-only (?) cache from T1s; ~30 day lifetime(?)
- Reliable network links: 1Gbit/s at T2s, 10Gbit/s at T1s, support full data rate to all T1s out of T0
  - If network link goes down, data must be re-routed to an alternate site; pro-longed outage a major problem; need correspondingly large data buffers at T0 / T1s
- Reliable File Transfer services
  - Gridftp, srmcopy + higher level functionality - SERVICE
- File catalogs, data management tools, database services
- Basic Services: workload management, VO management, monitoring etc.

➢ **Multiple levels of experiment specific software and corresponding additional complexity**

More on Services

- 24 x 7 services do not mean that people have to be chained to the computer 24 x 7
- Services must be designed / deployed to be as reliable and recoverable as possible
  - Monitor to check that this is so – including end to end monitoring
- Cannot tolerate failure of a major component Friday evening not looked at until Monday morning… after coffee…
  - Eventually run in degraded mode?
- Need to use existing experience and technology…
  - Monitoring, alarms, operators, SMS to 2nd / 3rd level support…
- Now is the time to get these procedures in place
  - Must be able to arrange that suitable experts can have network access within reasonable time
  - Even from the beach / on the plane …

Services at CERN

- Building on standard service model
- First level support: operations team
  - Box-level monitoring, reboot, alarms, procedures etc
- Second level support team: Grid Deployment group
  - Alerted by operators and/or alarms
  - Follow 'smoke-tests' for applications
  - Identify appropriate 3rd level support team to call
  - Responsible for maintaining and improving procedures
  - Two people per week: complementary to System Manager on Duty
  - Provide daily report to SC meeting (09:00); interact with experiments
  - Members: IT-GD-EIS, IT-GD-SC (including me)
  - Phone numbers: 164111; 164222
- Third level support teams: by service
  - Notified through operators and / or 2nd level (by agreement)
  - Should be called (very) rarely… (Definition of a service?)

Where are we now?

- Roughly mid-point in activity (first proposal to completion)
- Demonstrated sustained disk – disk data rates of **100MB/s** to multiple Tier1 sites, **>500MB/s** out of CERN for some 10 days; **800MB/s** to a single site (FNAL)
- Now (July): demonstrate **150MB/s** to Tier1s; **1GB/s** out of CERN (disk – disk) plus **60MB/s** to tape at Tier1s
- In terms of data rate alone, have to double data rates, plus whatever is necessary in terms of 'safety factors' ,including recovering backlogs from outages etc.

➢ **But so far, these tests have just been with dummy files, with the bare minimum software involved**

- In particular, none of the experiment software has been included!
- Huge additional work: add major complexity whilst doubling rates and providing high quality services

➢ (BTW, neither of first two challenges fully met their goals)

Baseline services

- Storage management services
  - Based on SRM as the interface
- Basic transfer services
  - gridFTP, srmCopy
- Reliable file transfer service
- Grid catalogue services
- Catalogue and data management tools
- Database services
  - Required at Tier1,2
- Compute Resource Services
- Workload management

- VO management services
  - Clear need for VOMS: roles, groups, subgroups
- POSIX-like I/O service
  - local files, and include links to catalogues
- Grid monitoring tools and services
  - Focussed on job monitoring
- VO agent framework
- Applications software installation service
- Reliable messaging service
- Information system

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
- File Catalog, which will act as a site-local catalog for ATLAS/CMS and a central catalog with >1 R/O replicas for LHCb
- (Database services behind all of these but not for experiment data)

SRM – Requirements Beyond 1.1

1. Pin/Unpin
2. Relative paths in SURLS ($VO_HOME)
3. Permission functions
4. Direction functions (except mv)
5. Global Space reservation
6. srmGetProtocols
7. AbortRequest etc

This list and schedule for delivery agreed at PEB 28 June

Core Site Services

- CERN
  - Storage: Castor/SRM
  - File catalogue: POOL LFC Oracle
- FNAL
  - Storage: dCache/SRM
  - File catalogue: POOL Globus RLS
- CNAF
  - Storage: Castor/SRM
  - File catalogue: POOL LFC Oracle
- RAL
  - Storage: dCache/SRM
  - File catalogue: POOL LFC Oracle?
- IN2P3
  - Storage: dCache/SRM
  - File catalogue: POOL LFC Oracle
- SARA/NIKHEF
  - Storage:dCache/SRM
  - File catalogue: POOL LFC MySQL(?)
- PIC
  - Storage: Castor/SRM
  - File catalogue: POOL LFC MySQL
- FZK
  - Storage: dCache/SRM
  - File catalogue: POOL LFC Oracle
- ASCC
  - Storage: Castor/SRM
  - File catalogue: POOL LFC Oracle
- BNL
  - Storage: dCache/SRM
  - File catalogue: POOL LFC Oracle
- TRIUMF
  - Storage: dCache/SRM
  - File catalogue: POOL LRC MySQL(?)
- NDGF
  - Storage:
  - File catalogue:

Running FTS service for T2s

Service Challenge 3 - Phases

High level view:

- Setup phase
  - Finishes with 2 weeks sustained throughput test in July 2005
  - 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 – must be run as *the* real production service
  - 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 – Deadlines and Deliverables

- May 31st 2005: basic components delivered and in place
- June 2005: integration testing
- June 13 – 15: SC3 planning workshop – experiment issues
- June 30th 2005: integration testing successfully completed
- July 1 – 10: start disk – disk throughput tests
  - Assume a number of false starts / difficulties
  - Bug-fix release of FTS Friday 1st July fixes critical issues
  - Site testing continues this week (CMS transfer in parallel)
- July 11 – 20: disk tests
- July 21 – 27: tape tests
- July 28 – 31: T2 tests

Service Challenge Workshop

- Three-day meeting (13-15 June)
  - First two days with presentations from Experiments. 1/2 day per experiment to cover:
    - Summary of Grid Data Challenges to date
    - Goals for SC3
    - Plans for usage of SC3 infrastructure
  - Third day focused on issues for the Tier-1 sites
    - Discussion of issues raised during previous two days
    - SRM requirements presentations from experiments and developers
- Approximately 40 people for first two days and 60 for last day
  - Many CERN IT people appearing for last day
  - Not all sites present during first two days (??) – if present, very quiet!

SC3 – Experiment Goals

All 4 LHC Experiments have concrete plans to:

- Test out the infrastructure
  - core services: storage, file transfer, catalogs, …
- Run a prolonged production across T0/T1/T2 sites
  - (ALICE / LHCb represent the two extremes; CMS / ATLAS between)

➢ **Expect long-term services to be delivered as an output of SC3**

- These services required from October 2005 / January 2006
  - Variation by experiment based on detector installation schedule
- These services (with upgrades) run until end of LHC – circa 2020

Experiment Goals and Plans

- All four experiments plan to be involved in SC3
- Brief "one-line" summary
  - LHCb will evaluate the new tools via the pilot and do a data management challenge in September. Assuming ok will want to use a service from October
  - ALICE will also evaluate the new tools but want to run a full data challenge based on this infrastructure asap
  - CMS will use the resources to run two challenges in September and November, but with modest throughput. These includes T0-T1-T2 data movement and T2-T1 movement for MC Data
  - ATLAS plan to run a Tier-0 exercise in October along with MC production at T2 and reprocessing at Tier-1. They will use their new DDM software stack

ALICE 2005 Physics Data Challenge

- Physics Data Challenge
  - Until September 2005, simulate MC events on available resources
  - Register them in the ALICE File Catalogue and store them at CERN-CASTOR (for SC3)
- Coordinate with SC3 to run our Physics Data Challenge in the SC3 framework
- Use case 1: RECONSTRUCTION
  - (Get "RAW" events stored at T0 from our Catalogue)
  - First Reconstruct pass at T0
  - 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
  - Transfer Data to supporting T1's

ATLAS and SC3

- > Mid-October; we intend to run:
  - Tier-0 exercise
    - Reconstruction at Tier-0 et production of ESD; AOD; Event collections
    - Data distributed from Tier-0 to Tier-1s then Tier-2s
  - Distributed Monte Carlo production
    - Data generated on Tier-2s, Tier-1s and are stored on Tier-1s for permanent storage
  - Use of conditions database will be part of the "game"
  - Reprocessing
    - Run at Tier-1s, "where the data is".
  - But this will be done in the last months of 2005
  - For DC3 we need to produce "few" 10 Million events
- We don't forget analysis!

CMS - Schedule Overview

- 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

LHCb - goals

Phase (I)

a) Moving of 8 TB of digitised data from CERN/Tier-0 to LHCb participating Tier1 centers in a 2-week period.

- The necessary amount of data is already accumulated at CERN
- The data are moved to Tier1 centres in parallel.
- The goal is to demonstrate automatic tools for data moving and bookkeeping and to achieve a reasonable performance of the transfer operations.

b) Removal of replicas (via LFN) from all Tier-1 centres

c) Moving data from Tier1 centre(s) to Tier0 and to other participating Tier1 centers.

- The goal is to demonstrate that the data can be redistributed in real time in order to meet the stripping processing.

d) Moving stripped DST data from CERN to all Tier1's

- The goal is demonstrate the tools with files of different sizes
  - Necessary precursor activity to eventual distributed analysis

Phase (II)

- MC production in Tier2 and Tier1 centers with DST data collected in Tier1 centers in real time followed by Stripping in Tier1 centers
  - MC events will be produced and reconstructed. These data will be stripped as they become available
- Data analysis of the stripped data in Tier1 centers.

Interpretation of Experiments' Goals

At high-level, strong commonality across experiments:

- First try, then test, then stress-test core data management services
- 'Trying' can be done in parallel to SC3 production activities (pre-production, pilot, …)
- 'Testing' requires coordination of resources and clear goals / metrics agreed up-front
- 'Stress-testing' (simulating / exercising primary offline use cases except EU analysis) requires further coordination of resources + clear goals / metrics agreed up-front
- We have to be clear that these are the goals and work together to achieve them
- We also have to be realistic and explicit about the level of functionality and service that can be provided
- (The summer will be long and hot)

Experiment Goals and Plans

- Concern that the experiment timelines all overlap
  - Creating a unified timeline from the detailed presentations
  - We need to respond with what is possible
- Pilot services for FTS and LFC are of great interest to experiments.
  - They'd like Fireman as well for testing
- Long discussions about "VO Boxes" at all sites – neither sites, experiments or middleware providers have worked through full implications of this
  - First we need to list exactly what the expt requirements are
  - Plan is to provide an interim solution for evaluation during SC3

Overall Schedule (Raw-ish)

Sep	Sep	Oct	Oct	Nov	Nov	Dec	Dec
ALICE	ALICE						
			ATLAS	ATLAS			
CMS	CMS			CMS	CMS		
LHCb		LHCb					

Sep	Sep	Oct	Oct	Nov	Nov	Dec	Dec
ALICE	ALICE						
				ATLAS	ATLAS		
	CMS	CMS			CMS	CMS	
		LHCb	LHCb				

Resource Requirements

- Clear storage requirements from ALICE, ATLAS and LHCb
- Explicit CPU request from ALICE
  - 300 nodes at T0 / 600 summed over T1s
  - Former possible but needs to be scheduled; Latter ok?
  - (Storage requirements less explicit…)
- And from LHCb…
  - 400 nodes at 'not T1' during phase II
  - 2(!) / site at T0/CERN
- Other experiments CPU requirements fit into existing allocation
  - +200 for ALICE at T0 out of 'pool' for 2-3 weeks?
- Time allocation in previous tables should not be taken as definitive – shows that minimal(?) overlap between experiments should be possible
- This has to be summarised in a plan by the time of July 20 GDB

Tier-1 Plans and Goals

- Clear message from June workshop that some sites did not understand what SC3 mean in terms of compute resources
  - "more than a transfer test"
- We need to resolve how to integrate SC3 resources into the production grid environment
  - "there can only be one production environment" – discussed in June GDB:
  - http://agenda.cern.ch/fullAgenda.php?ida=a045323
- Service levels provided will be "best-effort"
  - We should be able to live with a site being down for a while
  - But we must measure site uptime/availability/response during the challenge.

Software at Tier-1s

- Many SRM services are late – deadline was for end May
  - Many sites still haven't got services ready for SC3
    - Some need to upgrade versions (BNL)
    - Some need to debug LAN network connections (RAL)
    - Some are finalizing installs (FZK, ASCC, …)
  - And we're still mostly at the level of debugging SRM transfers
    - Many errors and retries detected at FTS level
    - (Timeout problem in FTS? Fast-track bug fix 1/7)
- Still need to rerun iperf tests to measure expected network throughput for all sites
- Activity required from Tier-1s to run the network measurement tests and more SRM level tests
  - Sites need to be more proactive in testing and publishing the information

Sample "Use Case" Jobs

Action on the experiments:

- Provide example jobs that demonstrate sample Use Cases.
- To be useful this has to be done (including the delivery of the jobs) by the middle of July if we are to be able to conclude on the setup phase of SC 3

Services Required – LHCb

CERN:

- Dedicated LFC (separated from LHCb production one).
- FireMan FC (for FC stress testing).
- Central FTS DB and Server.
- SRM v1.1 to MSS with gridFTP.
- LFC, FTS Client Tools.
- Hosting Machine for VO Agents (could be based on jobmanger-fork component of LCG CE) with managed file system.
- gLite CE

Tier1:

- Read only LFC (>1 Tier1).
- SRM v1.1 to MSS with grifFTP.
- LFC, FTS Client Tools.
- Hosting Machine for VO Agents (could be based on jobmanger-fork component of LCG CE) with managed file system.
- gLite CE

Tier2:

- SE with SRM or gridFTP access.
- LFC, FTS Client Tools.

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

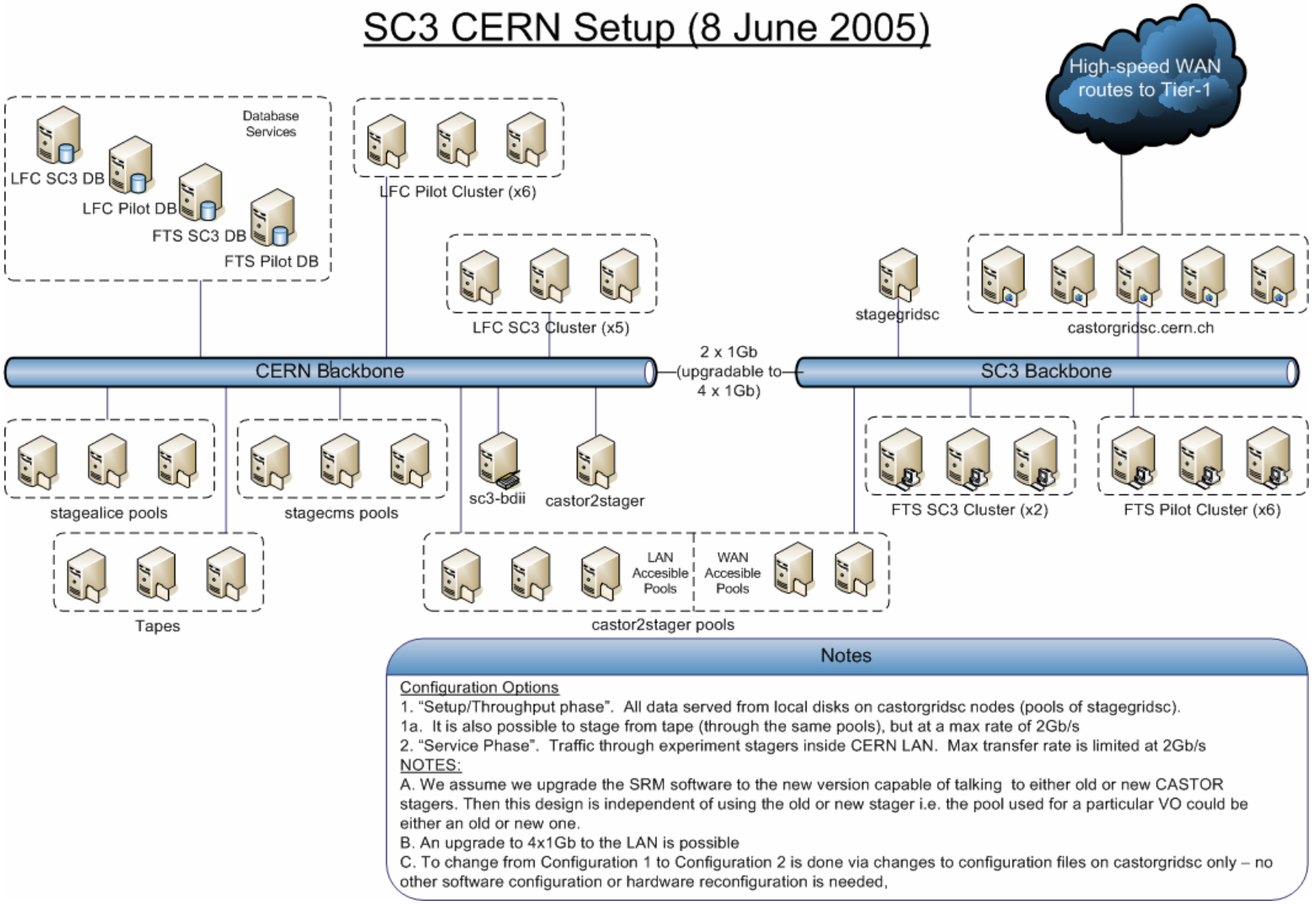
- Training in UK May 13th and in Italy May 26-27th. Training at CERN June 16th.
- Other interested parties: Prague, Warsaw, Moscow, ..
- **Addressing larger scale problem via national / regional bodies**
  - GridPP, INFN, HEPiX, US-ATLAS, US-CMS, Triumf (Canada)
- Cannot handle more for July tests, but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

CERN Services



SC3 Services Status

- FTS
  - SC3 service installed and configured. Limited testing undergone with Tier-1s. Many Tier-1's still upgrading to dCache and it's not all stable yet
  - BNL have a version of the FTS Server for their T1-T2 traffic
    - seeing many problems in getting it installed and configured
    - working with gLite team to try and solve these
  - Pilot services not ready yet
    - Installed but not configured yet
  - Experienced long delays for new software through gLite build+test process
    - but we now have a tag that will be ok for setup/throughput
    - This is part of LCG-2_5_0
  - Will need new version of FTS for service phase
    - Current version does not do inter-VO scheduling
    - This presents a risk since it will be a major rewrite

SC3 Services Status – T0

- LFC
  - Pilot and SC3 services are installed, configured and announced to experiments
  - POOL interface now available (POOL 2.1.0)
  - Not much usage yet by experiments
- CASTORGRIDSC SRM
  - 20TB setup running using old stager and old SRM code
  - Plan is to migrate to new CASTOR stager
    - 2TB migrated and being tested now
  - fallback solution is to use old stager for setup phase
- Migration of SC setup to new Castor stager is in progress

SC3 Services Status – T0 cont.

- Starting to put in place the service teams for SC3
  - First level support at CERN from operators
  - Second line support at CERN from GD SC and EIS teams
  - Third line support from software experts
    - LFC, FTS, Castor-SRM, …
    - Wan-Data.Operations@cern.ch (castor)
    - See also https://uimon.cern.ch/twiki/bin/view/LCG/LCGServiceChallenges
  - Site support through site specific service challenge mailing lists
    - What is the level of support we will get?
- Operator procedures and problem escalation steps still not clear
  - Reporting of problems through e-mail – tied into problem tracking system

SC Communication

- Service Challenge Wiki
  - Takes over from service-radiant wiki/web-site used in SC1 & 2

  https://uimon.cern.ch/twiki/bin/view/LCG/LCGServiceChallenges
  - Contains Tier-0 and Tier-1 contact/configuration information and work logs for SC teams
- Weekly phonecons on-going
  - Dial-in number: +41227676000
  - Access code: 0164222
- Daily service meetings for CERN teams from 27th June
- Technical communication through service-challenge-tech@cern.ch list

- What else is required by Tier-1s?
  - Daily (or frequent) meetings during SC?

SC3 Summary

- Good understanding and agreement on goals of SC3
  - What services need to run where
  - Proposed metrics to define success
  - Detailed schedule
  - Detailed discussion of experiment goals/plans in June 13 – 15 workshop
- Concerns about readiness of many sites to run production-level services
  - Preparations are late, but lots of pressure and effort
  - Are enough resources available to run *services*?
    - Backups, single points of failure, vacations, …
- Experiments expect that SC3 leads to real production service by end of year
  - Must continue to run during preparations for SC4
- This is the build up to the LHC service – must ensure that appropriate resources are behind it

Service Challenge 4 – SC4

- SC4 starts April 2006
- SC4 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 experiments
  - All Tier2s (and Tier1s…) at full service level
  - Anything that dropped off list for SC3…
  - Services oriented at analysis and end-user
  - What implications for the sites?
- Analysis farms:
  - Batch-like analysis at some sites (no major impact on sites)
  - Large-scale parallel interactive analysis farms and major sites
  - (100 PCs + 10TB storage) x N
- User community:
  - No longer small (<5) team of production users
  - 20-30 work groups of 15-25 people
  - Large (100s – 1000s) numbers of users worldwide

Analysis Use Cases (HEPCAL II)

- Production Analysis (PA)
  - **Goals in Context** *Create AOD/TAG data from input for physics analysis groups*
  - **Actors** *Experiment production manager*
  - **Triggers** *Need input for "individual" analysis*
- (Sub-)Group Level Analysis (GLA)
  - **Goals in Context** *Refine AOD/TAG data from a previous analysis step*
  - **Actors** *Analysis-group production manager*
  - **Triggers** *Need input for refined "individual" analysis*
- End User Analysis (EA)
  - **Goals in Context** *Find "the" physics signal*
  - **Actors** *End User*
  - **Triggers** *Publish data and get the Nobel Prize :-)*

SC4 Timeline

- September 2005: first SC4 workshop(?) – 3rd week September proposed
- January 31st 2006: basic components delivered and in place
- February / March: integration testing
- February: SC4 planning workshop at CHEP (w/e before)
- March 31st 2006: integration testing successfully completed
- April 2006: throughput tests
- May 1st 2006: Service Phase starts (note compressed schedule!)
- September 1st 2006: Initial LHC Service in stable operation
- Summer 2007: first LHC event data

SC4 Milestones

Date	Description
31 Jan 06	**All required software for baseline services deployed and operational at all Tier-1s and at least 20 Tier-2 sites.**
30 Apr 06	**Service Challenge 4 Set-up: Set-up complete and basic service demonstrated. Performance and throughput tests complete: Performance goal for each Tier-1 is the nominal data rate that the centre must sustain during LHC operation (see** Table 7.2 below) CERN-disk → network → Tier-1-tape. Throughput test goal is to maintain for three weeks an average throughput of 1.6 GB/s from disk at CERN to tape at the Tier-1 sites. All Tier-1 sites must participate. The service must be able to support the full computing model of each experiment, including simulation and end-user batch analysis at Tier-2 centres.
31 May 06	**Service Challenge 4: Start of stable service phase, including all Tier-1s and 40 Tier-2 centres.**
30 Sept 06	**1.6 GB/s data recording demonstration at CERN: Data generator→** disk → tape sustaining 1.6 GB/s for one week using the CASTOR mass storage system.
30 Sept 06	**Initial LHC Service in operation:** Capable of handling the full target data rate between CERN and Tier-1s (see Table 7.2). The service will be used for extended testing of the computing systems of the four experiments, for simulation and for processing of cosmic-ray data. During the following six months each site will build up to the full throughput needed for LHC operation, which is twice the nominal data rate.
1 Apr 07	**LHC Service Commissioned: A series of performance, throughput and reliability tests completed to show readiness to operate continuously at the target data rate and at twice this data rate for sustained periods.**

SC4 Use Cases

Not covered so far in Service Challenges:

- T0 recording to tape (and then out)
- Reprocessing at T1s
- HEPCAL II Use Cases
- Individual (mini-) productions

Additional services to be included:

- Full VOMS integration
- PROOF? xrootd? (analysis services in general…)

September SC3.5 workshop

- SC3 experience
  - Sites
  - experiments
  - outlook for remainder of service phase
  - Requirements gathering from site + experiment view points + report (by two rapporteurs from above sessions)

- SC4 preparation
  - (recent) experiment goals / plans in terms of HEPCAL use cases
  - proof / xrootd / roles / plans
  - LCG SRM status
  - targets for SC4
  - T1 plans for incorporating T2s
  - T2 plans

Remaining Challenges

- Bring core services up to robust 24 x 7 standard required
- Bring remaining Tier2 centres into the process
- Identify the remaining Use Cases and functionality for SC4
- Build a cohesive service out of distributed community
- Clarity; simplicity; ease-of-use; functionality

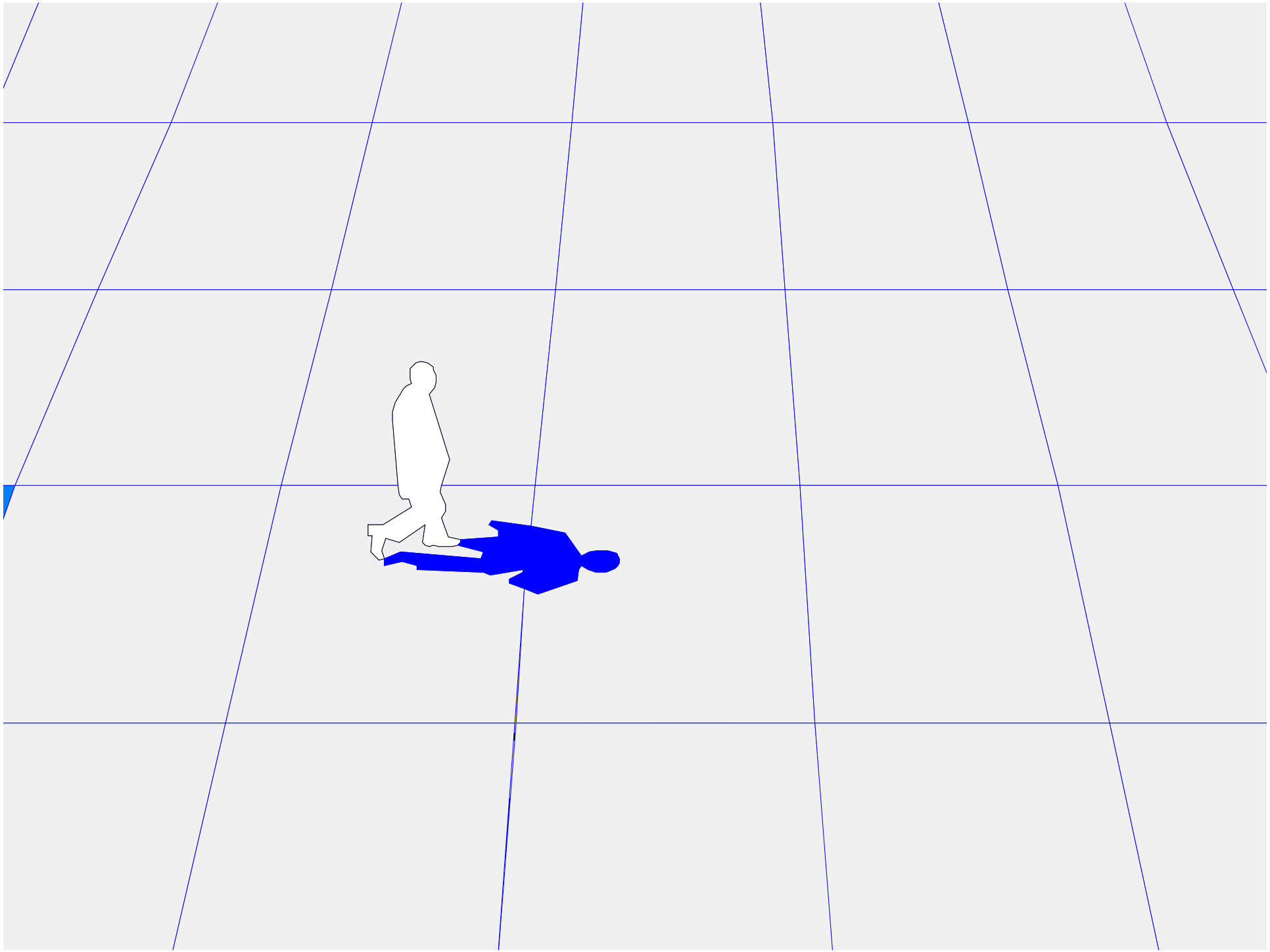
Summary and Conclusions

- Mid-way (time wise) in aggressive programme to deploy world-wide production Grid
- Services must be provided no later than year end…
  … and then run until career end
- Deploying production Grids is hard…
  … and requires far too much effort at the various sites
- We need to reduce this effort…
  … as well dramatically increase the ease of use
- Today hard for users to 'see' their science…
  … main effort is overcome the complexity of the Grid

Summary and Conclusions

- Mid-way (time wise) in aggressive programme to deploy world-wide production Grid
- Services must be provided no later than year end…
  … and then run until career end
- Deploying production Grids is hard…
  … and requires far too much effort at the various sites
- We need to reduce this effort…
  … as well as the dramatically increase ease of use
- Today hard for users to 'see' their science…
  … main effort is overcome the complexity of the Grid

The Service is the Challenge



GLIF Intrastructure