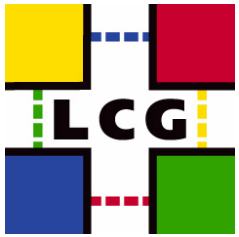


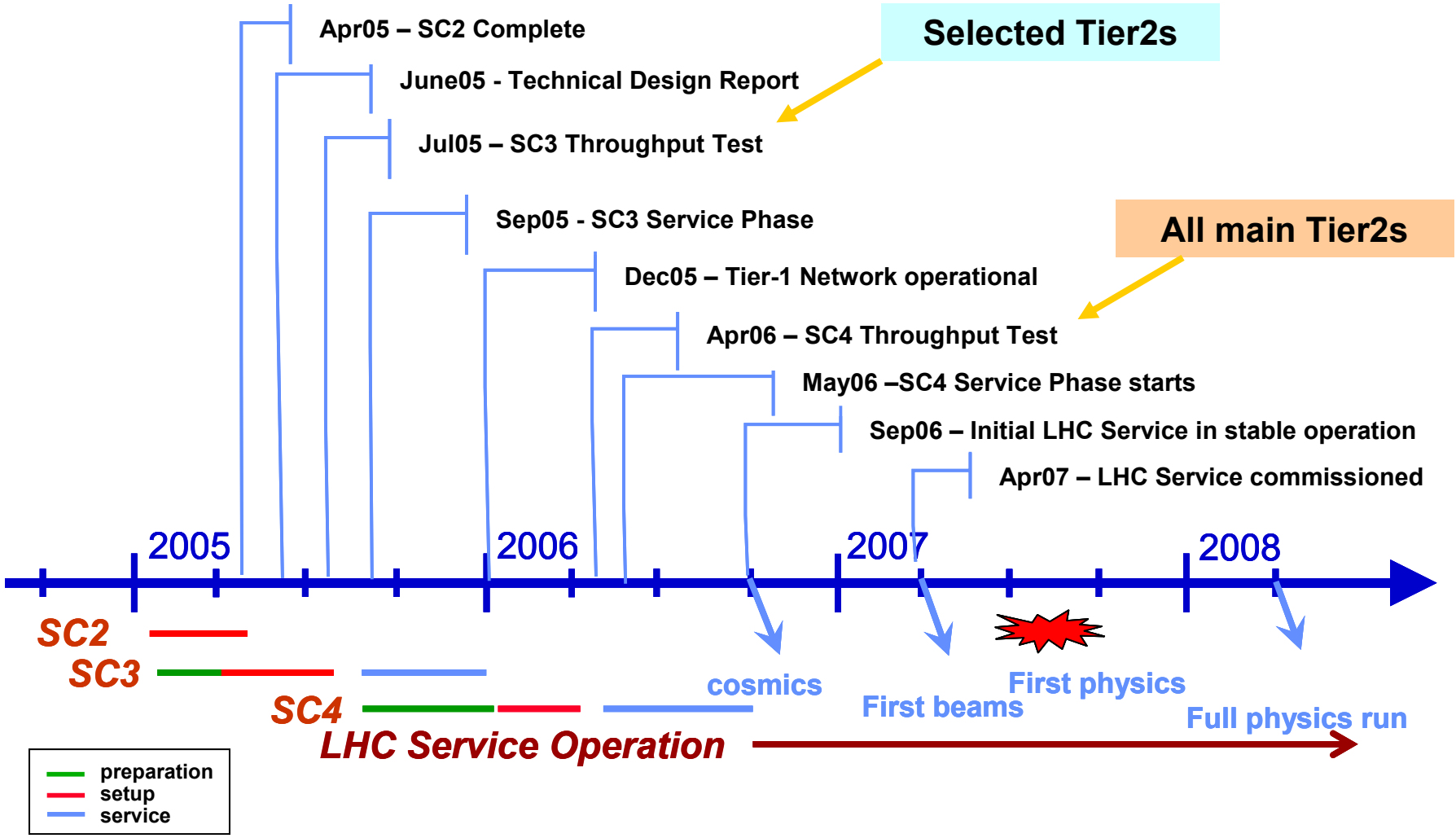
LCG Service Challenges: Planning for Tier2 Sites

Jamie Shiers
IT-GD, CERN

Introduction

- Roles of Tier2 sites
- Services they require (and offer)
- Timescale for involving T2s in LCG SCs
- Simplest (useful) T2 Model
- What is being implemented now…
- Summary

LCG Deployment Schedule



The Problem (or at least part of it…)

- SC1 – December 2004
- SC2 – March 2005

Neither of these involve T2s or even the experiments – just basic infrastructure

- SC3 – from July 2005 involves 2 Tier2s
  - + experiments' software + catalogs + other additional stuff
- SCn – completes at least 6 months prior to LHC data taking. Must involve all Tier1s and ~all Tier2s

- Not clear how many T2s there will be

- Current estimate: **100** – a ***huge*** number to add!
  - ALICE: 15? ,ATLAS: 30, CMS: 25, LHCb: 15; overlap?

Tier2 Roles

- Tier2 roles vary by experiment, but include:
  - **Production of simulated data;**
  - **Production of calibration constants;**
  - Active role in [end-user] analysis

➢ **Must also consider services offered to T2s by T1s**

  - e.g. safe-guarding of simulation output;
  - Delivery of analysis input.

- No fixed dependency between a given T2 and T1
  - But 'infinite flexibility' has a cost…

T2 Functionality

(At least) two distinct cases:

- Simulation output
  - This is relatively straightforward to handle
  - Most simplistic case: associate a T2 with a given T1
    - Can be reconfigured
    - Logical unavailability of a T1 could eventually mean that T2 MC production might stall
  - More complex scenarios possible
    - But why? **Make it as simple as possible, but no simpler…**
- Analysis
  - Much less well understood and likely much harder…

T1/T2 Roles

Tier1

- Keep certain portions of RAW, ESD, sim ESD
- Full copies of AOD + TAG, calibration data
- Official physics group large scale data analysis
- ALICE + LHCb:
  - also contribute to simulation

Tier2

- Keep certain portions of AOD and full copies of TAG for real + simulated data
  - LHCb: sim only at T2s
- Selected ESD samples
- Produce simulated data
- General end-user analysis

Based on "T1 Services for T2 Centres" document

(Just type this into Google)

Analysis

- Certain subsets of the data will be distributed across T0 and T1s
- ➢ **Must allow equal access to all data regardless of users' and its location**
- But this does not imply same physical network connectivity between every T2 and every T1…
- A model whereby data is handed between T1s rather than directly from 'remote' T1 to T2 may be much more affordable and manageable
  - May even be a star configuration
- Analysis not included in SCs until SC4…
  - But we should start to elaborate the models now…

Analysis Cont.

- "The AOD shall be the primary event format made widely available for analysis in CMS"
- "More than 90% of all analysis in CMS can be carried out from AOD samples"
- "It is only in a few, less than 10% … that the physicist has to refer to the full RECON dataset"
  - Navigation from AOD back to ESD and RAW possible via s/w pointers with protection against accidental access.
- The AOD is at all T1s and the TAG is everywhere

> 90% of analyses can be handled by a simple model

  - see later…

Network Requirements (ATLAS)

Source	Inbound from CERN (MB/s)	Outbound to CERN (MB/s)
RAW	30.4	
ESD Versions	20	1.41
AOD versions	18	0.28
TAG Versions	0.18	0
Group DPD		0.81
Total CERN/AverageTier-1	68.58	2.51

T1-T1 Requirements (ATLAS)

Source	Inbound from other Tier-1s (MB/s)	Outbound to other Tier-1s (MB/s)
ESD Versions	12	14
AOD versions	20.8	2.08
TAG Versions	0.21	0.02
Group DPD		4
Total Tier-1 to Tier-1	33	19

T2 Specifications (CMS)

- T2 centres should have WAN connectivity **in the range** of 1GB/s or **more** to satisfy CMS analysis requirements
- T2 centres will require relatively sophisticated **disk cache management** systems, or explicit and enforceable local policy, to ensure sample latency on disk is adequate and to avoid disk/WAN thrashing
- [They do not provide persistent (archival) storage]

CMS Network Requirements

- "We are pushing available networks to their limits in the Tier-1/Tier-2 connections"
  - Tier-0 needs ~2x10Gb/s links **for CMS**
  - Each Tier-1 needs ~10Gb/s links
  - Each Tier-2 needs 1Gb/s for its incoming traffic
  - There will be extreme upward pressure on these numbers as the distributed computing becomes more and more useable and effective

(Presentation to LHCC review of Computing Models)

A Simple T2 Model (1/2)

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

A Simple T2 Model (2/2)

- Each Tier-2 is associated with a Tier-1 that is responsible for getting them set up
- Services at T2 are **managed storage** and **reliable file transfer**
  - FTS: DB component at T1, user agent also at T2; DB for storage at T2
- 1GBit network connectivity – shared (less will suffice to start with, more maybe needed!)
- Tier1 responsibilities:
  - Provide archival storage for (MC) data that is uploaded from T2s
  - Host DB and (gLite) File Transfer Server
  - (Later): also data download (eventually from 3rd party) to T2s
- Tier2 responsibilities:
  - Install / run dCache / DPM (managed storage s/w with agreed SRM i/f)
  - Install gLite FTS client
  - (batch service to generate & process MC data)
  - (batch analysis service – SC4 and beyond)

➢ Tier2s **do not** offer persistent (archival) storage!

Analysis Model

- The 'soft' association of a T2 with a T1 handles the MC case and >90% of the analysis case
- Posit: the <10% left does not necessarily mean generalising the entire model
  - See next slide for a 'simple' example
- LHCb have a simple model for handling this by official 'stripping' phases ~4 times per year
- I would propose addressing now 100% of the MC case and >90% of the analysis cases
  - And further discussion – perhaps joint w/s with ARDA?

Analysis Example (ATLAS)

- A typical analysis scenario might begin with the physicist issuing a query against a **very large tag dataset**, e.g. the latest **reconstruction** of all data **taken to date**. For example, the query might be for events with **three leptons** and **missing transverse energy** above some threshold. The result of this query is used to define a dataset with the AOD information for these events. The analyst could then provide an **Athena algorithm** to make further event selection by refining the **electron** quality or missing **transverse** energy calculations. The new output dataset might be used to create an n-tuple for further analysis or the AOD data for the selected events could be copied into new files. A subset of particularly striking events identified in one of these samples could be used to construct a dataset that includes the ESD and perhaps even RAW data for these events. The physicist might then redo the electron reconstruction for these events and then use it to create a new AOD collection or n-tuple.

Initial Tier-2 sites

- For SC3 we aim for

Site	Tier1	Experiment
Bari, Italy	CNAF, Italy	CMS
Padova, Italy	CNAF, Italy	CMS, (Alice)
Turin, Italy		
DESY, Germany		
Lancaster, UK		
London, UK		
ScotGrid, UK		
US Tier2s		

More sites are appearing!

For CMS, also Legnaro, Rome and Pisa

For Atlas, sites will be discussed next week

Plans for a workshop in Bari end-May advancing well.

Tutorials also foreseen in the UK.

Discussions continuing with FZK / DESY

- Respons
- CERN's r
  - Devel
  - Provic
  - Assist
- Other int

➢ **Addressing larger scale problem via national / regional bodies**

- GridPP, INFN, (HEPiX), US-ATLAS, US-CMS

Experiment Software

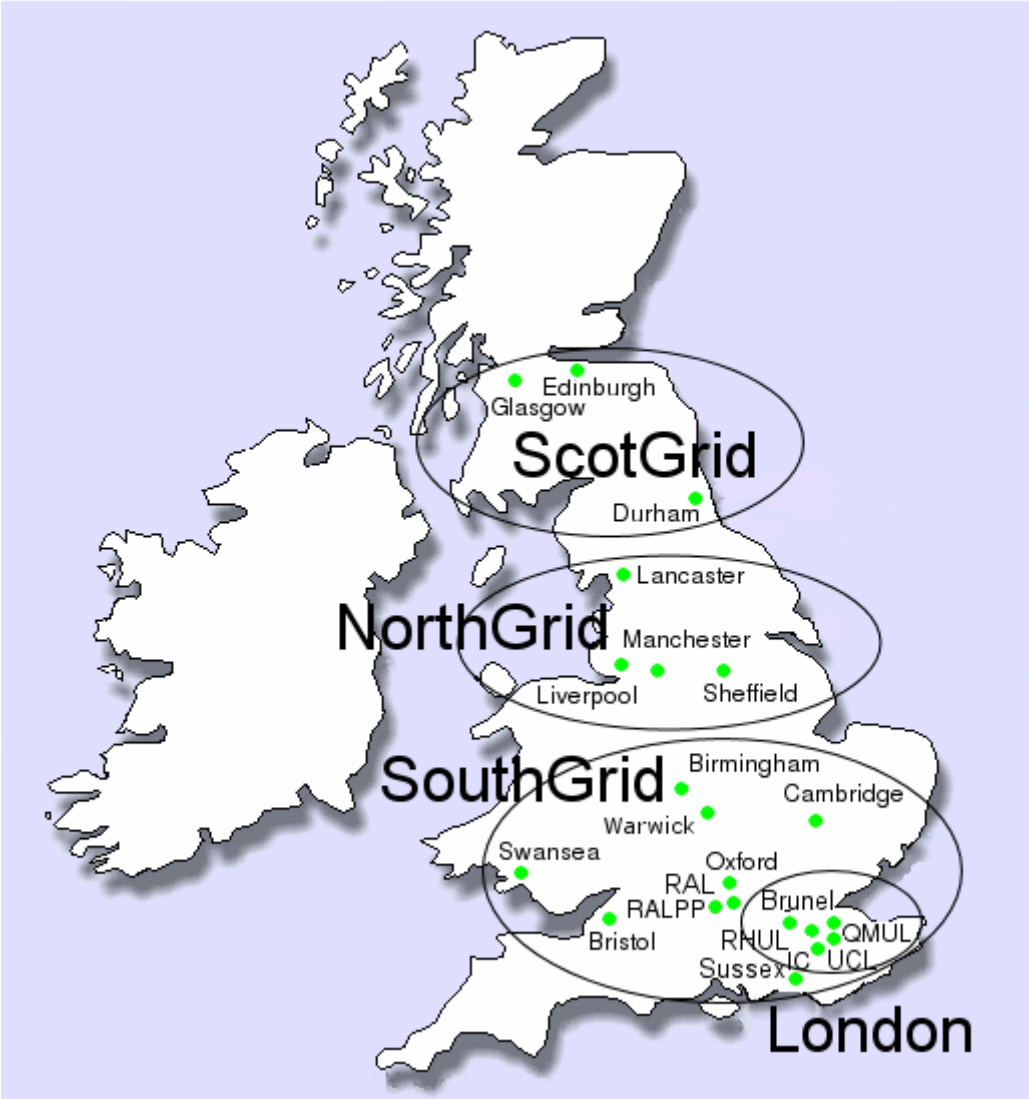
- A complete list of experiment s/w at T2 (and other) sites still has to be established
- However, it is likely to include DB applications (metadata, perhaps also catalogs), as well as "agents"
- More information now being gathered through weekly SC3 preparation phone meetings with T0/T1/T2 partners and experiments

Tier2 Region	Coordinating Body	Comments
Italy	INFN	A workshop is foreseen for May during which hands-on training on the Disk Pool Manager and File Transfer components will be held.
UK	GridPP	A coordinated effort to setup managed storage and File Transfer services is being managed through GridPP and monitored via the GridPP T2 deployment board.
Asia-Pacific	ASCC Taipei	The services offered by and to Tier2 sites will be exposed, together with a basic model for Tier2 sites at the Service Challenge meeting held at ASCC in April 2005.
Europe	HEPiX	A similar activity will take place at HEPiX at FZK in May 2005, together with detailed technical presentations on the relevant software components.
US	US-ATLAS and US-CMS	Tier2 activities in the US are being coordinated through the corresponding experiment bodies.
Canada	Triumf	A Tier2 workshop will be held around the time of the Service Challenge meeting to be held in Triumf in November 2005.
Other sites	CERN	One or more workshops will be held to cover those Tier2 sites with no obvious regional or other coordinating body, most likely end 2005 / early 2006.

Which Candidate T2 Sites?

- Would be useful to have:
  - Good local support from relevant experiment(s)
  - Some experience with disk pool mgr and file transfer s/w
  - 'Sufficient' local CPU and storage resources
  - Manpower available to participate in SC3+
    - And also define relevant objectives?
  - 1Gbit/s network connection to T1 desirable
- First T2 site(s) will no doubt be a learning process
  - On-site training, installation help etc.
- **Need** to (semi-)automate this so that adding new sites can be achieved with low overhead

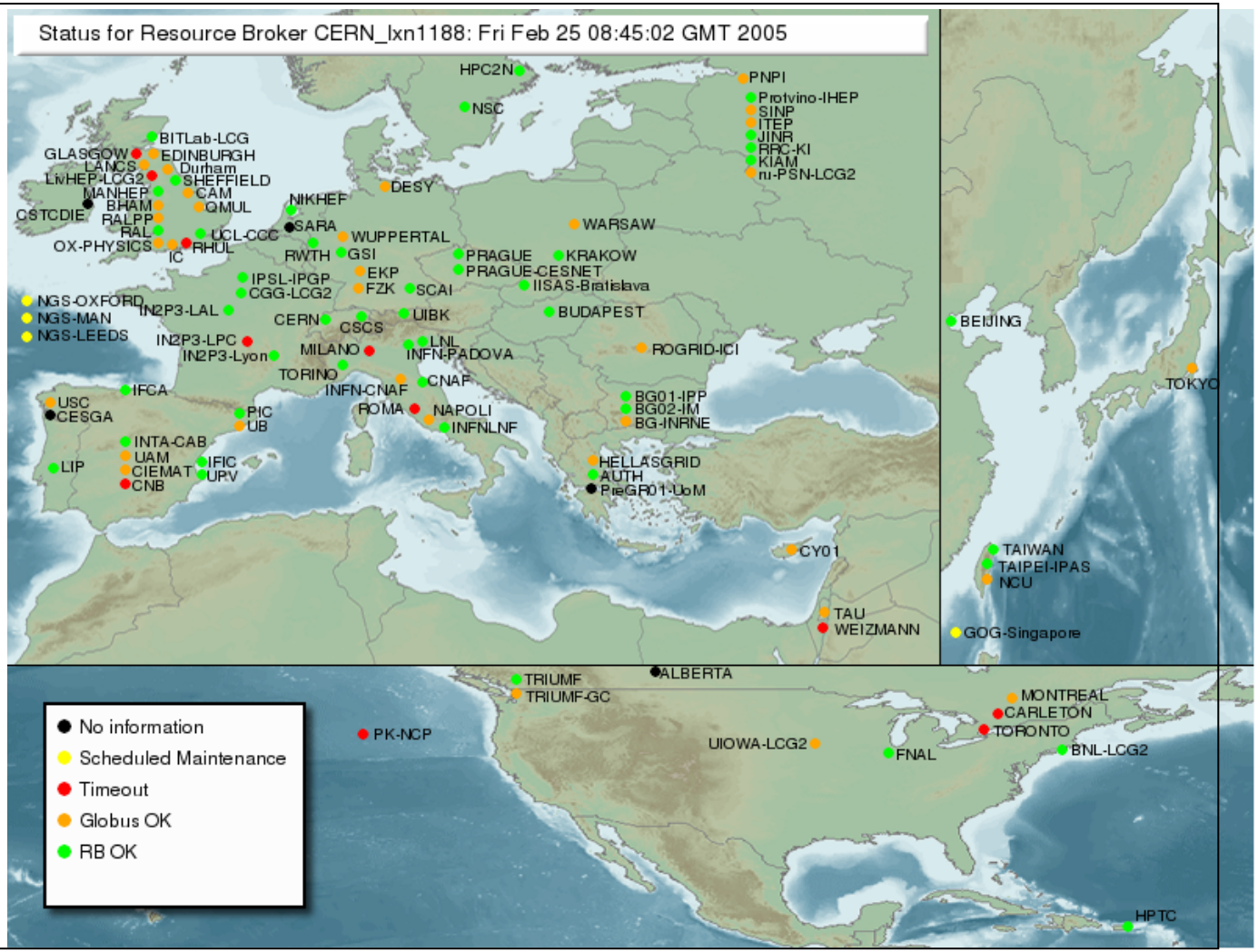
Which T2(s)?


Edinburgh
Glasgow
ScotGrid
Durham
Lancaster
NorthGrid
Manchester
Liverpool
Sheffield
SouthGrid
Birmingham
Cambridge
Warwick
Swansea
Oxford
RAL
RALPP
Brunel
Bristol
RHUL
QMUL
IC
UCL
Sussex
London

Which T2s cont.

Global T2 Planning

- UK
- Italy
- E/P
- D
- …



A Word on Data Rates…

- Summary of LHC Computing Models has resulted in 'agreed' (i.e. not contested) sizes for the different data types / experiment and event rates
- Using these raw numbers and the distribution policies described in the Computing Model documents we can calculate basic data rates
- No 'efficiency factors' for accelerator, load factors for network applied
- Trigger rate 'independent of luminosity' (some beam)

LHC Operation Overview

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	10^{7}	2×10^{33}	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}

(from CMS Computing Model)

Overview of pp running

Expt	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

LHC Schedule

- 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)
- Different ‘seasons’ during a year:
  - Commissioning / startup
  - pp running
  - AA running
  - Shutdown

Rates to T1 Sites During pp running

Centre	*ALICE*	*ATLAS*	*CMS*	*LHCb*	*Rate into T1*	
ASCC, Taipei	0	1	1	0	118.7	
CNAF, Italy	1	1	1	1	205.0	
PIC, Spain	0	1	1	1	179.0	
IN2P3, Lyon	1	1	1	1	205.0	
GridKA, Germany	1	1	1	1	205.0	
RAL, UK	1	1	1	1	205.0	
BNL, USA (takes full ATLAS ESD)	0	1	0	0	72.2	152.2
FNAL, USA	0	0	1	0	46.5	
TRIUMF, Canada	0	1	0	0	72.2	
NIKHEF/SARA, Netherlands	1	1	0	1	158.5	
Nordic Centre	1	1	0	0	98.2	
Totals	6	10	7	6		
Rates per experiment	26	72.2	46.5	60.333		

Rates Continued

- Discussion on whether split / Tier1 should be based on fraction of resources allocated per experiment
- Can (will) calculate raw rates also for other 'seasons'
- Will require some assumptions, e.g. number / location of reprocessing passes
- Resource allocation, e.g. provisioned network bandwidth, I/O to tape, will in any case require a data processing model, efficiency factors etc.

MC Data

	Units	ALICE		ATLAS	CMS	LHCb
		p-p	Pb-Pb	p-p	p-p	
Time to reconstruct 1 event	kSI2k sec	5.4	675	15	25	2.4
Time to simulate 1 event	kSI2k sec	35	15000	100	45	50

Tier2 sites offer 10 – 1000 kSI2K years (??)
ATLAS: 16MSI2K years over ~30 sites in 2008
CMS: 20MSI2K years over ~20 sites in 2008

Parameter	Unit	ALICE		ATLAS	CMS	LHCb
		p-p	Pb-Pb			
Events/year	Giga	1	0.1	2	1.5	20
Events SIM/year	Giga	1	0.01	0.4	1.5	4
Ratio SIM/data	%	100%	10%	20%	100%	20%

GridPP T2 Resources

	CPU (KSI2K)				Disk (TB)			
	ALICE	ATLAS	CMS	LHCb	ALICE	ATLAS	CMS	LHCb
London	0	553	651	217	0	39	58	18
NorthGrid	0	1353	0	144	0	297	0	20
ScotGrid	0	123	0	131	0	4	0	65
SouthGrid	124	269	145	167	5	15	9	11

Summary of Resources

		2008	HR	2008	HR	2008	HR	2009	HR
		ATLAS	ATLAS	CMS	CMS	LHCb	LHCb	ALICE	ALICE
Tier 0 CPU	*MSI2k*	4.1	3.915	4.6				4.5	
CPU at CERN	*MSI2k*	6.3	6.21	7.5	7.38	0.9	2.025		7.416
Tier 0 disk	*PB*	0.35		0.41				0.5	
disk CERN	*PB*	1.95	0.41	1.71	0.796	0.8	0.33		0.53
Tier 0 tape	*PB*	4.2		3.8				2.7	
tape CERN	*PB*	4.6	9	5.6	4.172	1.4	1.22		3.23
Tier 1 cpu	*MSI2k*	18	11.286	12.8	9.18	4.4	6.3	10.6	8.424
Tier 1 disk	*PB*	12.3	2.16	6.7	1.565	2.4	0.75	6.3	1.08
Tier 1 tape	*PB*	6.5	10.8	11.1	5.115	2.1	1.6	8.7	1.48
Tier 2 cpu	*MSI2k*	16.2		19.9	9.675	7.6		10.9	
Tier 2 disk	*PB*	6.9		5.3	1.75	0.02		1.7	
Tier 2 tape	*PB*	0	0	0	1.25	0		0	

T2 Roles (GridPP Summary)

- ALICE - MC Production, Chaotic Analysis
- ATLAS - Simulation, Analysis, Calibration
- CMS - Analysis for 20-100 Physicists, All Simulation Production
- LHCb - MC Production, No analysis
- The resources required are given below (assumed to be ~2008) [1]:

[1] The ATLAS network numbers quoted here have been updated from the Computing Model by Roger Jones.

GridPP Estimates of T2 Networking

	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

Which T2s?

- Currently compiling a list of T2 sites
- Around 100 have been identified, together with the experiments that they serve with priority
- Russian, Ile de France and UK serve all experiments
- CH, PL, Brazilian, Italian serve 2-3
- Remainder largely just 1…
- In some cases, candidate Tier1 is 'obvious'

➢ **In many cases not. How does this get resolved?**

Tier2s

Institution	*Experiments served with priority*			
	ALICE	*ATLAS*	*CMS*	*LHCb*
CSCS, Switzerland		X	X	X
FZU AS, Prague, Czech Rep.	X	X		
Hungarian Tier-2 Federation, Hungary - KFKI, Budapest - SZTAKI, Budapest - Eotvos Univ., Budapest - Debrecen Univ.	X		X	
Helsinki Institute of Physics, Finland			X	
Krakow, Poland	X	X		X
Warszawa, Poland	X		X	X
Russian Tier-2 cluster , Russian Fed	X	X	X	X
HEP-IL Federation, Israel - Technion, Haifa - Weizmann, Rehovot - Tel Aviv Univ.		X		
Pakistan Tier-2 - PAEC - NCP - NUST - COMSATS			X	

Tier2s

Brazilian Tier-2 Federation - *UFRJ* - *UERJ* -.*CBPF*		X	X	X
TIFR, Mumbai, India			X	
VECC/SINP, Kolkata, India	X			
Univ. Blaise. Pascal, Clermont-Ferrand, France	X	X		X
Fédération Ile de France, France - LAL, Orsay - LPNHE-Paris - DAPNIA-Saclay	X	X	X	X
CC-IN2P3 Tier-2, Lyon, France		X	X	

Tier2s

INFN Tier-2s, Italy	X indicates primary experiment			
- INFN-Bari	X			
- INFN-Bologna-CMS			X	
- INFN-Cagliari	X			
- INFN-Catania	X			
- LNF-Frascati		X		
- LNL-Legnaro	X		X	
- INFN-Milano		X		
- INFN-Napoli		X		
- INFN-Padova			X	
- INFN-Pisa			X	
- INFN-Roma1		X	X	
- INFN-Torino	X			

Tier2s

RWTH, Aachen, Germany			*X*	
GSI, Darmstadt, Germany	X			
Freiburg Univ., Germany		*X*		
DESY, Hamburg, Germany		X	X	
Mainz Univ., Germany		*X*		
MPI/LMU, Munich, Germany		*X*		
Univ. Wuppertal, Germany		*X*		
UIBK, Innsbruck, Austria		X		
ATLAS Federation, Spain - IFIC, Valencia - IFAE, Barcelona - UAM, Madrid		X		
LHCb Federation, Spain - UB, Barcelona - USC, Santiago				X
CMS Federation, Spain - CIEMAT, Madrid - IFCA, Santander			X	

Tier2s

Grid London, UK - UCL - ICL - Brunel - RHUL - QMUL	X	X	X	X
NorthGrid, UK - Manchester - Liverpool - Sheffield - Lancaster - *Daresbury Lab.*	X	X	X	X
ScotGrid, UK - Edinburgh - Glasgow - Durham	X	X	X	X
SouthGrid, UK - Birmingham - Cambridge - Warwick - Swansea - Bristol - RAL - Oxford - Sussex	X	X	X	X

Tier2s

South West ATLAS T2, USA - Arlington Univ. - Oklahoma Univ. - Univ. of New Mexico - Univ. Texas (LU)		X		
Mid West ATLAS T2, USA - Univ.of Chicago - Indiana Univ.		X		
Boston/Harvard ATLAS T2, USA - Boston Univ. - Harvard Univ.		X		
CMS T2, USA - MIT - Univ. of Florida - Univ. of Nebraska - Univ. of Wisconsin - Caltech - Purdue Univ. - Univ. of California			X	
Canada East Tier2 federation		X		
Canada West Tier2 federation		X		
IHEP, Beijing, China		X		
Univ. Melbourne, Australia		X		
ICEPP, Tokyo, Japan		X		

Summary

- The first T2 sites need to be actively involved in Service Challenges from Summer 2005
- ~All T2 sites need to be successfully integrated just over one year later
- Adding the T2s and integrating the experiments' software in the SCs will be a massive effort!
- Initial T2s for SC3 have been identified
- A longer term plan is being executed