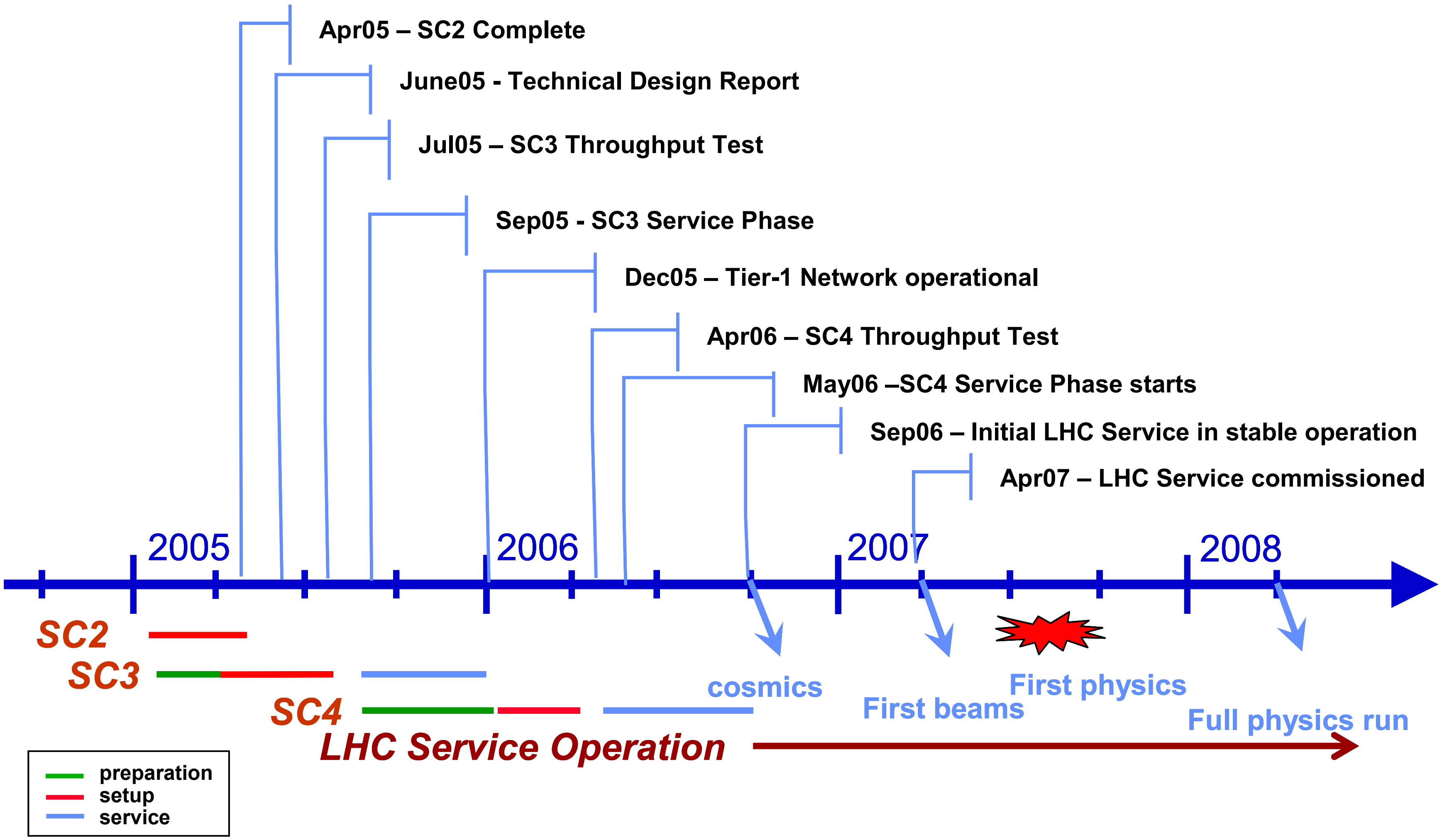
The LCG Service Challenges: Summary of Computing Models

Jamie Shiers, CERN-IT-GD-SC

April 2005

LCG Service Challenges – Deploying the Service

LCG Deployment Schedule



Apr05 – SC2 Complete

June05 - Technical Design Report

Jul05 – SC3 Throughput Test

Sep05 - SC3 Service Phase

Dec05 – Tier-1 Network operational

Apr06 – SC4 Throughput Test

May06 –SC4 Service Phase starts

Sep06 – Initial LHC Service in stable operation

Apr07 – LHC Service commissioned

🟩 preparation
🟥 setup
🟦 service

Computing Model Summary - Goals

- Present key features of LHC experiments' Computing Models in a consistent manner
- High-light the commonality
- Emphasize the key differences
- Define these 'parameters' in a central place (LCG web)
  - Update with change-log as required
- Use these parameters as input to requirements for Service Challenges
- To enable partners (T0/T1 sites, experiments, network providers) to have a clear understanding of what is required of them

➢ **Define precise terms and 'factors'**

Where do these numbers come from?

- Obtained from LHC Computing Models as reviewed in January
- Part of plan is to understand how sensitive overall model is to variations in key parameters
- Iteration with experiments is on-going
  - i.e. I have tried to clarify any questions that I have had
- Any mis-representation or mis-interpretation is entirely my responsibility
- Sanity check: compare with numbers from MoU Task Force
- (Actually the following LCG document now uses these numbers!)

http://cern.ch/LCG/documents/LHC_Computing_Resources_report.pdf

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…**

All numbers presented will be **nominal** unless explicitly specified

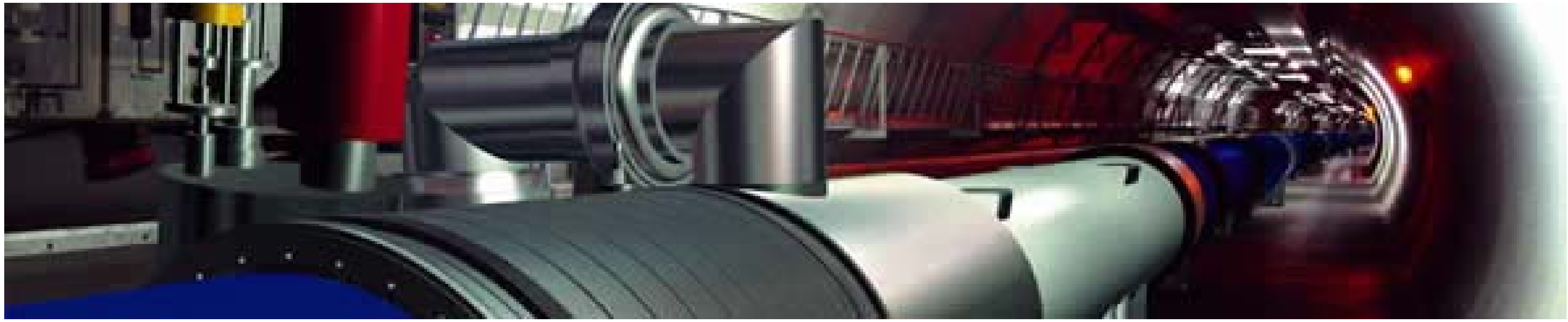
LHC Parameters (Computing Models)

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)
Based on 7 months pp, 1 month AA, 4 months shutdown (next)

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

pp questions / uncertainties

- Trigger rates essentially independent of luminosity
  - Explicitly stated in both ATLAS and CMS CM docs

- Uncertainty (at least in my mind) on issues such as zero suppression, compaction etc of raw data sizes
  - Discussion of these factors in CMS CM doc p22:

- RAW data size ~300kB (Estimated from MC)
  - Multiplicative factors drawn from CDF experience
    - MC Underestimation factor 1.6
    - HLT Inflation of RAW Data, factor 1.25
    - Startup, thresholds, zero suppression,…. Factor 2.5
  - Real initial event size more like **1.5MB**
    - Could be anywhere between 1 and 2 MB
  - Hard to deduce when the even size will fall and how that will be compensated by increasing Luminosity

➢ i.e. total factor = 5 for CMS raw data

- N.B. must consider not only Data Type (e.g. result of Reconstruction) but also how it is used
  - e.g. compare how Data Types are used in LHCb compared to CMS

- All this must be plugged into the meta-model!

Overview of Heavy Ion running

Experiment	SIM	SIMESD	RAW	Trigger	RECO	AOD	TAG
ALICE	300MB	2.1MB	12.5MB	100Hz	2.5MB	250KB	10KB
ATLAS			5MB	50Hz			
CMS			7MB	50Hz	1MB	200KB	TBD
LHCb	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Heavy Ion Questions / Uncertainties

- Heavy Ion computing models less well established than for pp running
- I *was* concerned about model for 1st/2nd/3rd pass reconstruction and data distribution
- *"We therefore require that these data (Pb-Pb) are reconstructed at the CERN T0 and exported over a four-month period after data taking. This should leave enough time for a second and third reconstruction pass at the Tier 1's"* (ALICE)
- Heavy Ion model has major impact on those Tier1's supporting these experiments
  - All bar LHCb!
- These issues have since been clarified:
  - Raw data export will be spread over shutdown;
  - First pass reconstruction should complete at least 6 months prior to following year's AA data taking (ALICE) or during shutdown (CMS);
  - 2nd pass (ALICE) will not involve the full data sample;
  - 3rd pass is a complete pass which should complete prior to following year's AA run
- Implies data out of CERN roughly constant; will vary per T1 (pp/AA/shutdown) depending on experiments supported

Data Rates from MoU Task Force

MB/Sec	RAL	FNAL	BNL	FZK	IN2P3	CNAF	PIC	T0 Total
ATLAS	106.87	0.00	173.53	106.87	106.87	106.87	106.87	707.87
CMS	69.29	69.29	0.00	69.29	69.29	69.29	69.29	415.71
ALICE	0.00	0.00	0.00	135.21	135.21	135.21	0.00	405.63
LHCb	6.33	0.00	0.00	6.33	6.33	6.33	6.33	31.67
T1 Totals MB/sec	182.49	69.29	173.53	317.69	317.69	317.69	182.49	1560.87
T1 Totals Gb/sec	1.46	0.55	1.39	2.54	2.54	2.54	1.46	12.49
Estimated T1 Bandwidth Needed								
(Totals * 1.5*(headroom)*)*2*(capacity)*	4.38	1.66	4.16	7.62	7.62	7.62	4.38	37.46
Assumed Bandwidth Provisioned	**10.00**	**10.00**	**10.00**	**10.00**	**10.00**	**10.00**	**10.00**	**70.00**

http://cern.ch/LCG/MoU%20meeting%20March%2010/Report_to_the_MoU_Task_Force.doc

LCG Service Challenges – Deploying the Service

Tier1 Sites

Centre	*ALICE*	*ATLAS*	*CMS*	*LHCb*	*Target Data Rate MBytes/sec*
ASCC		X	X		110
CNAF	X	X	X	X	220
PIC		X	X	X	200
CC-IN2P3	X	X	X	X	220
FZK	X	X	X	X	220
RAL	X	X	X	X	220
BNL		X			154
FNAL			X		50
TRIUMF		X			65
NIKHEF	X	X		X	175
NORDIC DATA GRID FACILITY	X	X		X	90
Target data rate at CERN					1,600

Target data rates calculated from raw computing model numbers during pp running

- No (in)efficiency factors, no overhead, etc.
- Assume each T1 takes equal fraction (except BNL)
- Balanced by fraction of resources allocated per experiment?

Heavy Ion Data Rates

- ATLAS / CMS data rates limited by online system
- LHCb does not participate in Heavy Ion programme
- Current model is that data is distributed during shutdown, rather than inline with data taking

pp / AA data rates (equal split)

Centre	*ALICE*	*ATLAS*	*CMS*	*LHCb*	*Rate into T1*	*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 Centre	1	1	0	0	98.2	80.3
Totals	6	10	7	6		

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.

Francesco Forti, Pisa

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!

Francesco Forti, Pisa

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

Data Rate / Site - Conclusions

- It is clear that these are only estimates
- Experiments will almost certainly split data into streams which will not be of equal size
- The share of resources that each T1 provides to a given experiment will also influence the amount of data that is sent there
- But it is impossible (and not relevant?) to do a precise calculation
- Which in any case becomes further clouded when the reprocessing and analysis is folded in…

Initial Tier2 Sites for SC3

Site	*Tier1*	*Experiment*
Bari, Italy	CNAF, Italy	Alice, CMS
Turin, Italy	CNAF, Italy	Alice
DESY, Germany	FZK, Germany	ATLAS, CMS
Lancaster, UK	RAL, UK	ATLAS
London, UK	RAL, UK	CMS
ScotGrid, UK	RAL, UK	LHCb
US Tier2s	BNL / FNAL	ATLAS / CMS

Prime Tier-2 sites

- For SC3 we aim for

Site	Tier1	Experiment
Bari, Italy	CNAF, Italy	CMS
Turin, Italy	CNAF, Italy	Alice
DESY, Germany	FZK, Germany	ATLAS, CMS
Lancaster, UK	RAL, UK	ATLAS
London, UK	RAL, UK	CMS
ScotGrid, UK	RAL, UK	LHCb
US Tier2s	BNL / FNAL	ATLAS / CMS

- Responsibility between T1 and T2 (+ experiments)
- CERN's role limited
  - Develop a manual "how to connect as a T2"
  - Provide relevant s/w + installation guides
  - Assist in workshops, training etc.
- Other interested parties: Prague, Warsaw, Moscow, ..
- ➢ Also attacking larger scale problem through national / regional bodies
  - GridPP, INFN, HEPiX, US-ATLAS, US-CMS

Tier2 and Base S/W Components

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

1 – 3 will be bundled with LCG release.
Experiment-specific s/w will not…

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