LCG File Transfer Service Challenge Requirements

Version	Date	Major Changes
0.1	7 February 05	Uses latest CM numbers and section on T1-T2 sites. To be reviewed at February GDB.
0.0	14 January 05	Initial Version.

This document summarises the requirements of the LHC experiments in terms of File Transfer services between the sites involved (Tier0, Tier1, Tier2). It is based on the Computing Models submitted to the LHCC in December 2004 and reviewed in January 2005, which should be consulted for further details.

The overall capacity that must be provided by the primary tiers is presented. These numbers are compared with those prepared by the LCG MoU task force. The intent is that these requirements provide a clear baseline for defining the detailed Data Management Service Challenges that need to be scheduled, as well as the work-plans of the groups involved at the various sites (e.g. the "physics" groups in IT, namely ADC, FIO, GD and CS groups at CERN) and the LHC experiments themselves.

It is assumed that this capacity must be provided by all involved sites at least 6 months prior to first physics data taking at the LHC (scheduled for April 2007), i.e. by end Q3 2006 and in planned incremental steps between now and then.

A series of Service Challenge Milestones based on these requirements will be defined in detail.

Unless explicitly stated, all numbers quoted below should be considered 'nominal', without any additional factors applied. We define the key factors relevant for network transfers below.

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.

Figure 1 - Network Uplift Factors

1. Overview of the Computing Models

All four LHC experiments assume a Grid-based solution – i.e. the LCG – and have Computing Models that can be viewed as that proposed by the MONARC project with Grid extensions. All define largely similar functions for the Tier0, Tier1 and Tier2 sites. This document is primarily concerned with the Tier0 and Tier1 sites, although the needs of the Tier2 sites are also included.

At the highest level, all experiments have a requirement for exporting a copy of the raw data across the Tier1 sites for that experiment in close to real-time for p-p running of the LHC. During heavy ion running it is assumed that the distribution of the data be carried out over a somewhat longer period (e.g. 4 months in the case of ALICE), as listed below.

The distribution of a copy of the raw data provides a significant, if not necessarily the main[1], input to the networking requirements between the sites and also as regards sizing the infrastructure required for sending / receiving these data volumes. We therefore first concentrate on this requirement, before describing the needs in terms of data flow into CERN, between different T1 sites and between T1s and T2s.

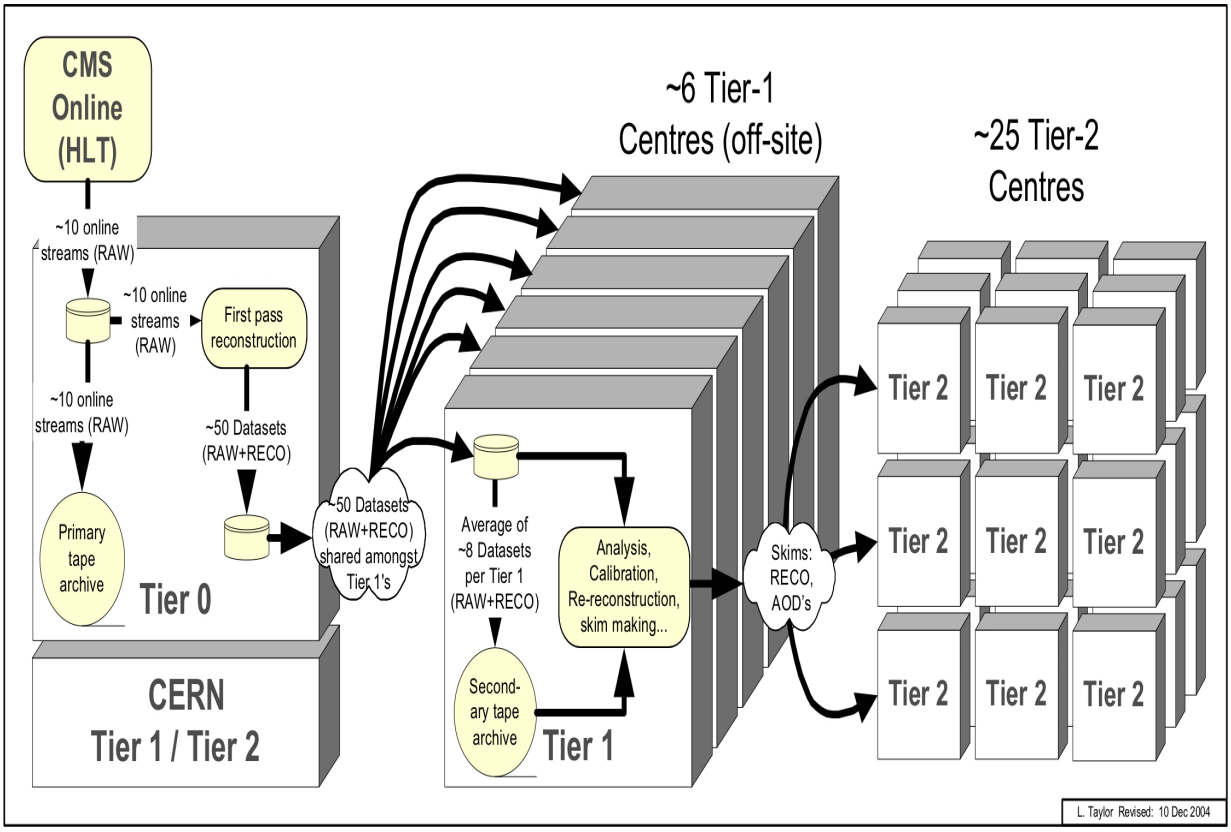

Figure 2 - Overview of the CMS Computing Model

[1] End-user analysis – unless well controlled – may result in a significant load and will need to be addressed shortly.

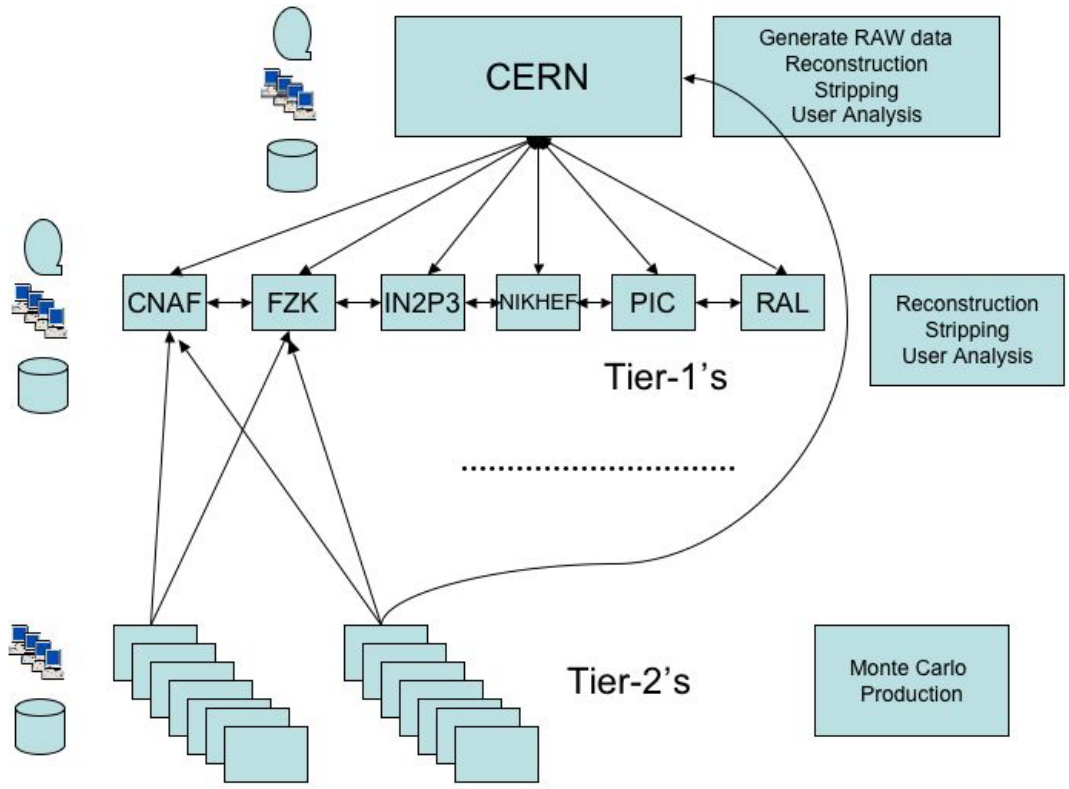

Figure 3 - Overview of the LHCb Computing Model

2. Summary of Tier0/1/2 Roles

Whilst there are differences between the roles assigned to the tiers for the various experiments, the primary functions are as follows:

- 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.

3. Tier-1 Centres

The following table gives the Tier-1 centres that have been identified at present, with an indication of the experiments that will be served by each centre. Many of these sites offer services for multiple LHC experiments and will hence have to satisfy the integrated rather than individual needs of the experiments concerned.

CENTRE	*LOCATION*	*ALICE*	*ATLAS*	*CMS*	*LHCB*
CNAF	Bologna, Italy	X	X	X	X
PIC	Barcelona, Spain		X	X	X
CCIN2P3	Lyon, France	X	X	X	X
GridKA	Karlsruhe, Germany	X	X	X	X
RAL	Didcot, UK	X	X	X	X
NIKHEF/SARA	Amsterdam, Netherlands	X	X		X
Nordic	Scandinavia	X	X		
BNL	Long Island, NY, USA		X		
FNAL	Batavia, IL, USA			X	
Triumf	Vancouver, Canada		X		
ASCC	Taipei, Taiwan		X	X	

Figure 4 - List of Tier1 Centres (Updated at January 2005 GDB)

4. Tier-2 Centres

The roles of the Tier-2 centres as well as the services that must be provided for them are discussed in a document produced by Gonzalo Merino et al. The numbers listed in this document need to be updated to correspond to the latest version of the LHC experiments' Computing Models, but globally speaking the current conclusion is that 1GBit links between a given Tier-2 and 'its' Tier-1 will be sufficient. Whilst robust file transfer services are also needed from Tier-2 sites to Tier-1s – to upload simulated data – and from Tier-1s to Tier-2s – to distributed relevant portions of TAG, AOD and ESD data – the same level of redundancy and recovery may not be necessary. For example, if a Tier-2 is unable to upload a Monte Carlo sample for a period of a day or so, the best strategy may simply be to wait and retry. Similarly, if 'only' 95% of the Tier-2 facilities are available at a given time for distributed analysis for whatever reason, the impact may be acceptable and indeed preferable to the additional complexity that might be required to avoid such situations.

For the time being, we simply use the same table as in the above document, although the number of Tier-2 sites may well have increased since then – at least for ATLAS.

	ALICE	ATLAS	CMS	LHCb
Parameters:				
Number of Tier-1s	4	6	6	5
Number of Tier-2s	20	24	25	15
Real data "in-T2":				
TB/yr	120	124	257	0
Mbit/sec (rough)	31.9	32.9	68.5	0.0
Mbit/sec (w. safety factors)	95.8	98.6	205.5	0.0
MC "out-T2":				
TB/yr	14	13	136	19
Mbit/sec (rough)	3.7	3.4	36.3	5.1
Mbit/sec (w. safety factors)	11.2	10.2	108.9	15.3
MC "in-T2":				
TB/yr	28	18	0	0
Mbit/sec (rough)	7.5	4.9	0	0.0
Mbit/sec (w. safety factors)	22.5	14.7	0.0	0.0

Table 1 - Bandwidth estimation for the T1 to T2 network links

5. LCG MoU Task Force Numbers

The summary table from the LCG Phase II Task Force is given below. The full spreadsheet can be found at:

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

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

Figure 5 - Summary of Bandwidth Required per Experiment

6. Summary of Data Transfer Requirements (p-p)

The data transfer requirements of the different experiments and their Tier0/1 site(s) are listed below. These are based on the latest numbers from the Computing Model documents and the same spreadsheet as was used in the Phase II planning process. The LHC schedule is assumed to be as follows:

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}

Figure 6 - Scenario of LHC Operation (from CMS Computing Model document)

The current figures for p-p running are given below. Both ATLAS and CMS explicitly state that the trigger rate can be assumed to be independent of luminosity. CMS further argue that their quoted RAW event size includes a factor to cater for not fully optimised zero-suppression or HLT (CMS Computing Model p 22).

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

Figure 7 - Summary of Data Sizes by Data Type and Experiment (pp)

The current figures for Heavy Ion data are given below. It is currently assumed that Heavy Ion running starts in late 2008 and, at least in the case of ATLAS, the Computing Model for these data is not yet completed established.

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

Figure 8 - Summary of Data Sizes and Data Type (Heavy Ions)

Using the above data sizes, the summary table per Tier0/1 site is as below. We note that the CMS RAW event has grown from 1MB to 1.5MB whilst the ESD has decreased from 500KB to 250KB since the previous calculation. Furthermore, the final factor of 2 to allow for recovery from outages was not previously included. This means that in some cases a single 10GBit link from the Tier0 to Tier1 is insufficient (e.g. for FZK, IN2P3 and CNAF), although it was already arguably too tight to cater for the possibly uncertainties that remain in the models.

Furthermore, for these centres there is relatively little remaining headroom. Increasing, for example, the RAW and ESD event sizes of ATLAS and CMS to 2MB and 1MB each also causes the requirement without the ‘recovery factor’ to exceed the capacity of a single 10GBit link.

MB/Sec	RAL	FNAL	BNL	FZN	IN2P3	CNAF	PIC	T0 Total
ATLAS	106.87	0.00	173.53	106.87	106.87	106.87	106.87	707.87
CMS	103.93	103.93	0.00	103.93	103.93	103.93	103.93	623.57
ALICE	0.00	0.00	0.00	144.10	144.10	144.10	0.00	432.29
LHCb	24.00	0.00	0.00	24.00	24.00	24.00	24.00	120.00
T1 Totals MB/sec	234.80	103.93	173.53	378.89	378.89	378.89	234.80	1883.73
T1 Totals Gb/sec	1.88	0.83	1.39	3.03	3.03	3.03	1.88	15.07
Estimated T1 Bandwidth Needed (Totals * 1.5*(headroom))*2(capacity)*	5.64	2.49	4.16	9.09	9.09	9.09	5.64	45.21
Assumed Bandwidth Provisioned	10.00	10.00	10.00	10.00	10.00	10.00	10.00	70.00

Figure 9 - Bandwidth Required per Tier1 Using January 2005 Computing Model Numbers

7. Conclusions

The overall requirements in terms of Data Transfer between Tier0 and Tier1s, together with the associated uncertainties, are relatively well understood and are presented above.

A single 10Gbit line between the Tier0 and Tier1s will in some cases be insufficient. There is limited flexibility to cater for increases in event sizes before a third (or second in some cases) line is required.

An aggressive and demanding programme of work is required to ensure not only that the required infrastructure and support personnel are put in place in a timely manner but also that the services provided can support experiment use-cases over extended periods of time – compatible with that of the LHC running period – including seamless handling of failures of individual components, including complete sites, even the Tier0(!)