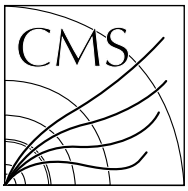

CERN-LHCC-2004-035/G-083
CMS NOTE / 2004-031
December 15, 2004

The CMS Computing Model

The "CMS Computing Model RTAG" [1]

Editors:

Claudio Grandi, *INFN Bologna;*
David Stickland, *Princeton University, and*
Lucas Taylor, *Northeastern University*

This paper describes the top-level requirements and specifications of the CMS Computing Model, and the associated costs. It was prepared primarily for the LHCC review of Computing Resources of January 2005 but will also be used as input to the CMS and LCG Computing TDR's, due for submission in the Summer of 2005. The primary focus is the first year of LHC physics running but approximate estimates of the cost evolution are also given.

[1] **Requirements and Technical Assessment Group members:** Lothar Bauerdick, Nica Colino, Ian Fisk, Claudio Grandi, Vincenzo Innocente Werner Jank, Emilio Meschi, Norbert Neumeister, Dave Newbold, Sasha Nikitenko, Lucia Silvestris, Nick Sinanis, Paris Sphicas, David Stickland, Lucas Taylor, and Avi Yagil.
Internal reviewers: Harvey Newman and Gunter Quast.
External reviewers: Tony Cass, Peter Elmer, Neil Geddes, and John Harvey.
Acknowledgments: We thank the numerous members of CMS and external experts who have contributed through their advice and constructive criticism.

Executive Summary

This document provides a top-level description of the CMS Offline Computing systems, with emphasis on the period immediately following the turn-on of the LHC and CMS. The main features of the CMS Data Model and Physics Analysis Models are described, followed by the corresponding specifications of the CMS Computing Model. Preliminary and approximate cost estimates are provided from 2007 (first collisions) up to 2010 (high luminosity running).

The model and the costs will be reviewed by the LHCC in early 2005. Then the technical and resource issues will be refined further for the CMS Computing TDR due for submission in mid-2005, coincident with the related LCG TDR.

The main cost-drivers of the proposed Computing Model are:

- **The Tier-0 Centre at CERN** is responsible for the safe keeping of the (first copy of the) RAW experiment data; the first reconstruction pass; the distribution of the reconstruction products to Tier-1 centres; and for reprocessing of data during LHC down-times.
- **The Tier-1 Centres** are expected to number about six to ten including one at CERN. They are each responsible for the safe keeping of a share of the (second copy of the) RAW and reconstructed data; for large-scale reprocessing steps and the safe keeping of data products of these steps; for the distribution of data products to Tier-2 centres and for the safe keeping of a share of the simulated data produced at these Tier-2 centres.
- **The Tier-2 Centres** are expected to number about 25. They are each responsible for servicing the analysis requirements of about 20-100 CMS Physicists, depending on size; and each responsible for about 1/25th of the simulated event production (and their reconstruction) requirements of CMS.
- **Networks** are crucial to support large-scale data transfers. Tier-0 to Tier-1 and Tier-2 to Tier-1 needs are reasonably well known and comparatively easily managed compared to the Tier-1 to Tier-2 needs for analysis which is more difficult to predict and more chaotic in nature, both in aggregate and instantaneously. The networking requirements of CMS, particularly for Tier-1's, will be substantial.
- **GRID Middleware and Infrastructure** must make it possible for any suitably-authorised CMS physicist to process data stored at any Tier-1 centre and to move the data products to an appropriate Tier-2 centre. We expect systems for remote job submission, monitoring and data movement to be used via GRID middleware and infrastructure. The computing centres need to deploy interoperable versions and designs of GRID middleware so that variations in local GRID implementations are transparent to CMS Physicists.

This resource estimates in this document focus on the Computing resources. The manpower re-

quired at computing centres and for software development, while both important and significant in scale, is not included. Wide area network costs are generally covered by national or regional bodies and are not explicitly included in the cost estimates.

The estimated cost for this Computing Model to be implemented by 2007 in time for the first substantial LHC run, assumed to be 2008, is **59 MCHF**, comprising the Tier-0 (5 MCHF), about 6 Tier-1's (total of 30 MCHF) and about 25 Tier-2's (total of 23 MCHF).

Incremental costs to handle the increased luminosity, larger data sets, and computing replacements is estimated to be **20MCHF** in 2008-2010, and beyond until major changes in the LHC operation mode require a different profile. of these incremental costs about half are associated to storage media (tape) and the remainder to maintenance and upgrade.

All requirements, specifications and costs described in this document are subject to revision during the preparation of the CMS and LCG Computing TDR's.

Structure of this document

Chapter 1, the Introduction, describes the context of this document.

Chapter 2 describes the main data and physics analysis requirements that the Computing Model must satisfy.

Chapter 3 describes a top-level baseline CMS Computing Model and specifies the needed computing resources.

Chapter 4 converts these technical specifications to estimated costs.

Finally, a glossary of abbreviations, acronyms and terms is provided followed by Appendix A which describes references for further reading and the associated bibliography.

Contents

List of Tables and Figures

List of Physics Requirements

List of Computing Specifications

Chapter 1

Introduction

This document constitutes a first draft of the CMS Computing Model specification which is currently being defined as part of for the CMS Computing TDR due in Summer 2005. It is a snapshot of current thinking and will evolve as the input for the Computing TDR is refined. The writing of the actual Computing TDR, the Physics TDR and the execution of integrated readiness tests such as the foreseen Magnet/Cosmic test in 2005 may each lead to modifications to this computing model. Because of this we stress that all requirements, specifications and costs described in this document will be subject to revision during the preparation of the CMS and LCG Computing TDR's.

This document has been prepared by a CMS "Computing Model RTAG" (Requirements Technical Assessment Group) which followed on, with augmented membership, from the Data Management RTAG [1]. The *modus operandi* of the RTAG was to rapidly identify the top-level requirements, scope, and solutions and seek consensus on them. A number of educated guesses have necessarily been made; some of these will need more rigourous treatment and analysis for the Computing TDR. Wherever possible, our choices are based on operational experience in CMS Data Challenges, production activities, and analysis systems. Care was also taken to confront our choices with the realities of running experiments, through the active participation of experts from CDF, D0, and BaBar in this RTAG. We recognise that while these experiments are the most appropriate running examples we can compare with, neither their trigger nor event flow are directly comparable with CMS and care must be taken in extrapolating to our conditions.

We have tried to establish a consistent ensemble of parameters for the computing specification that together implement for CMS a plausible computing model that is flexible and scalable enough to adapt to realities as they arise. In such estimates of future computing requirements almost every number can be scrutinised and found wanting in precision. While it may be possible to expend many man-years of research and effort on studying some of the quantitative aspects of the model, any real improvement may be largely illusory. At this stage of an experiments life, the computing model should be based on simple parameters and formulae which are themselves founded in real experience and judgments, rather than on excessively fine-tuned and specific solutions.

The main focus of this document is the Computing Model for LHC startup, meaning the first year with a sustained physics run with an assumed luminosity of $\mathcal{L} = 2 \times 10^{33}\text{cm}^{-2}\text{s}^{-1}$. Approximate estimates are also given for the subsequent years in which the luminosity increases to about $\mathcal{L} = 10^{34}\text{cm}^{-2}\text{s}^{-1}$. The possible time evolution of the Computing Model will be elaborated in more detail in the Computing TDR, and indeed some key features of this model may change substantially in the Computing TDR preparation. Clearly all details and strategies of the

computing model will be subject to regular review, particularly when they are confronted with LHC running.

Chapter 2

Requirements – Data, Processing, and Analysis Models

This chapter describes the requirements placed on the offline computing by the needs of the CMS physics program, in particular the event and non-event data management and its processing for the purposes of reconstruction, calibration, analysis and simulation.

2.1 Overview

The first year of running, the main focus of this document, will most likely be characterized by a poorly understood detector, unpredictable machine performance, inadequate computing infrastructure but the potential for significant physics discoveries. We expect to reprocess data often and we have to get that data in its complexity and richness out to the collaboration so their expertise can be brought to bear on detector, software, calibration and physics as effectively as possible. We will need good mechanisms to allow the data to be processed according to the priorities (be they detector understanding or Higgs searches). We will need to use all the Tiers of computing resources as effectively as possible, pre-locating data where they can be most efficiently processed and ensuring that the granularity of job queues at the sites is sufficient to steer the majority of computing resources to the experiments priorities, while ensuring that maverick ideas still have the possibility to be explored.

These principles lead us to a baseline solution that emphasizes:

- Fast reconstruction code (Frequent re-reconstruction)
- Streamed Primary Datasets (Priority driven distribution and processing)
- Distribution of Raw and Reconstructed data together (Easy access to raw detector information)
- Compact data formats (Multiple copies at multiple sites)
- Consistent production reprocessing and bookkeeping (Avoid confusion)

CMS expects to operate a structured analysis environment with analysis groups focusing on the main physics activities. We expect to define priorities on the activities at the Tier-0 and Tier-1 facilities predicated on satisfying the analysis group requirements. Particularly at startup the limited resources must be used carefully and much of this Computing Model is designed to enable this prioritisation to be effectively imposed.

To allow the computing planning in this paper we have used an operations scenario as described in table 2.1. The Computing Model is insensitive to slight changes in the luminosity profile as trigger thresholds will be adjusted up and down to maintain steady data rates as the running conditions vary.

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}

Table 2.1: Scenario of LHC operation assumed for the purposes of this document.

2.2 Event Model

CMS will use a number of event data formats with varying degrees of detail, size, and refinement. Starting from the raw data produced from the online system successive degrees of processing refine this data, apply calibrations and create higher level physics objects.

Table 2.2 describes the various CMS event formats. It is important to note that, in line with the primary focus of this document, this table corresponds to the LHC startup period and assumes a canonical luminosity of $\mathcal{L} = 2 \times 10^{33} cm^{-2}s^{-1}$. At this time the detector performance will not yet be well understood, therefore the event sizes are larger to accommodate looser thresholds and avoid rejection of data before it has been adequately understood. The determinations of the data volume include the effects of re-processing steps with updated calibrations and software and the copying of data for security and performance reasons; the motivation for these factors are described in this and the next chapter.

2.2.1 Raw Data (RAW)

RAW Event Content and Size

Efforts to estimate occupancies for various sub-detectors are an ongoing effort within CMS. They impact not only detector design but obviously also the computing model and its budget. In the following we report numbers derived for the low luminosity running period ($2 \times 10^{33} cm^{-2}s^{-1}$), with a stable well understood detector. There may be multiple ways to measure event size with different formats, packing and compression schemes. The basic format used will be the one generated by the event builder as it assembles the data from the FED's and creates the input to the HLT farm. This will be designated DAQ-RAW.

Event Format	Content	Purpose	Event size (MByte)	Events / year	Data volume (PByte)
DAQ-RAW	Detector data in FED format and the L1 trigger result.	Primary record of physics event. Input to online HLT	1-1.5	1.5×10^9 $= 10^7$ seconds $\times$ 150Hz	–
RAW	Detector data after online formatting, the L1 trigger result, the result of the HLT selections ("HLT trigger bits"), potentially some of the higher-level quantities calculated during HLT processing.	Input to Tier-0 reconstruction. Primary archive of events at CERN.	1.5	3.3×10^9 $= 1.5 \times 10^9$ DAQ events $\times$ 1.1 (dataset overlaps) $\times$ 2 (copies)	5.0
RECO	Reconstructed objects (tracks, vertices, jets, electrons, muons, etc. including reconstructed hits/clusters)	Output of Tier-0 reconstruction and subsequent re-reconstruction passes. Supports re-fitting of tracks, etc.	0.25	8.3×10^9 $= 1.5 \times 10^9$ DAQ events $\times$ 1.1 (dataset overlaps) $\times \Big[$2 (copies of 1st pass)$+$ 3 (reprocessings/year)$\Big]$	2.1
AOD	Reconstructed objects (tracks, vertices, jets, electrons, muons, etc.). Possible small quantities of very localized hit information.	Physics analysis	0.05	53×10^9 $= 1.5 \times 10^9$ DAQ events $\times$ 1.1 (dataset overlaps) $\times$ 4 (versions/year) $\times$ 8 (copies per Tier $-$ 1)	2.6
TAG	Run/event number, high-level physics objects, e.g. used to index events.	Rapid identification of events for further study (event directory).	0.01	–	–
FEVT	Term used to refer to RAW+RECO together (not a distinct format).		–	–	–

Table 2.2: CMS event formats at LHC startup, assuming a luminosity of $\mathcal{L} = 2 \times 10^{33}\text{cm}^{-2}\text{s}^{-1}$. The sample sizes (events per year) allow for event replication (for performance reasons) and multiple versions (from re-reconstruction passes).

Requirement: R-1	The online HLT system must create "RAW" data events containing: the detector data, the L1 trigger result, the result of the HLT selections ("HLT trigger bits"), and some of the higher-level objects created during HLT processing.

The largest contributor is expected to be the silicon strip detector, and its projected size is 130kB/event [2]. This number was derived using the latest tunes for the PYTHIA event generator and the full simulation of the CMS detector, and therefore reflects the current understanding of the experiment. Based on this and similar work in the other sub-detectors, an overall size estimate for the DAQ-RAW format, at an instantaneous luminosity of $2{\times}10^{33}$ cm^{-2}s^{-1}, of 300 kB/event is obtained.

There are various reasons to expect that the event size in reality will be larger than this estimate and we identify the following factors:

- F_{Det} reflects the effects of adverse startup conditions, detector commissioning, not completely effective "zero-suppression";
- F_{HLT} reflects the need to commission and understand the HLT algorithms, must keep all intermediate results;
- F_{MC} reflects the MC being overly optimistic (may be the event generator or the detector simulation or quite likely both).

The first two are the hardest to estimate. The duration of their impact is as hard to predict as their scope. For the CDF experiment at the Tevatron Run II, F_{Det} was as large as 2.5 for a few months and F_{HLT} was 1.25 and lasted about a year. This represents experiment-specific "failure modes" and should be taken as such. As part of the following exercise we use these as our central value estimators for CMS. The third one, F_{MC}, is easier to estimate. Using the CDF data and MC, a comparison between the occupancy predicted by the MC is compared to that observed in data. The MC is underestimating the observed occupancy by a factor of 1.6. There is no obvious reason to expect CMS MC to get better results.

Requirement: R-2	The RAW event size *at startup* is estimated to be $S_{RAW} \simeq 1.5$ MB, assuming a luminosity of $\mathcal{L} = 2 \times 10^{33}\text{cm}^{-2}\text{s}^{-1}$.

This was obtained as follows:

$$\begin{aligned} S_{RAW} &= 300\text{kB} \times F_{Det} \times F_{HLT} \times F_{MC} \simeq 300\text{kB} \times 2.5 \times 1.25 \times 1.6 \\ &\simeq 1.5 \text{ MB/event} \end{aligned}$$

Hopefully these factors will drop quite quickly after a few months (as for CDF) and then steadily asymptote towards unity after some years of experience. It should be noted that at this period one is also trying to time-in the various sub-detectors and overall size is not the main concern. After this initial period, as the detector is more-or-less commissioned, and one is mainly debugging the HLT:

$$\begin{aligned} S_{RAW} &= 300\text{kB} \times F_{HLT} \times F_{MC} \simeq 300\text{kB} \times 1.25 \times 1.6 \\ &\simeq 600 \text{ kB/event} \end{aligned}$$

This period lasted in CDF more than a year. When the HLT is validated one can reach the projected steady-state event size for low luminosity of:

$$\begin{aligned} S_{RAW} &= 300\text{kB} \times F_{MC} \simeq 300\text{kB} \times 1.6 \\ &\simeq 500 \text{ kB/event} \end{aligned}$$

Further reduction in size can be achieved using (lossy) packing and (loss-less) compression of the DAQ-RAW data. Here again we look to CDF who experienced a factor of $F_{pack} = 1.5 - 2.0$ for loss-less packing of the RAW data as it is being written out of the detector. This last step requires a lot of testing and confidence building.

$$\begin{aligned} S_{RAW} &= 300\text{kB} \times F_{MC}/F_{pack} \simeq 300\text{kB} \times 1.6/1.5 \\ &\simeq 300 \text{ kB/event} \end{aligned}$$

It should be clear that in making this estimation we have made a number of best-guesses based on experience at running experiments operating in similar conditions. It is unsafe to predict now when the various safety factors can be decreased; nor can we know now how much worse the actual running conditions may be. This value of 1.5 MB event (entering the offline system) is then a best estimated central value; we cannot exclude that it will in fact be anywhere in the range 1-2 MB for the running in the first sustained LHC data-taking period.

Requirement: R-3	The RAW event size *in the third year of running* is estimated to be $S_{RAW} \simeq$ 1.0 MB, assuming a luminosity of $\mathcal{L} = 10^{34}\text{cm}^{-2}\text{s}^{-1}$.

This asymptotic value accounts for two effects. As the luminosity increases so will the event size, due to the increase in the number of pile-up events. However, as the detector and machine conditions stabilise with time and become better understood the event size, for a given luminosity, will decrease. The error on the quoted value is dominated by the uncertainties in the time evolution of the various F factors described above.

CMS expects the RAW data size to reduce somewhat during the first full year of running, but we are unable at this time to predict when or by how much those reductions will occur. CMS notes that there are possible initial running conditions when RAW data size could be larger than 2MB. (Of course under such conditions we would find ways to reduce the data sizes substantially during the offline processing). CMS therefore plans these computing requirements based on a full year running period in "2008" with event sizes of about 1.5MB. Just-in-time purchasing of some media, but not of operational throughput capacity, could be considered. Actual requirements for the following years will in any case be re-evaluated in time for a sensible purchasing profiles.

RAW Event Rates

Since the early days of the LHC experiments, the rate of events to permanent storage was cited as $\sim 10^2$ Hz. The figure, along with an event size of ~ 1 MB, represented a rough estimate of the rates that could be reasonably sustained through the offline processing stage. As the detector and software designs matured, CMS performed the first early estimates of the rate needed to carry out the main "discovery" physics program. The result is that a minimum of 80 Hz is needed (March 2002, LHCC presentations). When a few calibration samples are included, as was done in a more complete evaluation for the DAQ Technical Design Report, the same figure becomes 105 Hz.

The above figure of 105 Hz assumes that the experiment has been forced to reduce its rate to permanent storage to the bare minimum needed for it to maintain high efficiency for the well-studied Higgs, SUSY and Extra-dimension physics cases. What remains uncovered are the standard-model channels, e.g. jet channels, inclusive missing transverse energy, and lowering of a few thresholds (e.g. the photon thresholds for the Higgs di-photon search so that more of

the standard-model background can be measured directly in the data); as well as a number of topics in top physics. These additional events complete the physics program and guarantee that CMS can effectively study all the physics offered by the LHC machine. A first estimate of these channels results in the addition of another 50 Hz to the rate to storage. Further lowering of the thresholds and loosening of the online requirements (e.g. along lines corresponding to those shown by ATLAS at the same LHCC presentations in 2002) result in the addition of another 50 Hz. In brief, for the same assumed thresholds and physics channels, CMS requires, a total rate to storage of about 200 Hz.

Experience gained from previous experiments at hadron colliders indicates that a lot will be learned with the first collisions at the LHC. And many of the estimates will be firmed up by then.

Requirement: R-4	The RAW event rate from the online system is 150 Hz or 1.5×10^9 events per year.

Given the uncertainties of the rate estimates from the combination of physics generators and the detector simulation, as well as the uncertainties of the machine and experimental backgrounds, we choose to use the figure of 150 Hz for the best estimate of the rate required for the physics program to proceed.

Certainly, CMS plans to record the maximum rate that its resources will accommodate, given that additional rate is simply additional physics reach. There is, a priori, no reason to limit the output of the experiment to any figure, even at 300 Hz, since the physics content is ever richer. The above figures are simply the result of today's estimates on the type of environment that the experiment will encounter as well as an attempt to limit the output to a figure that could be reasonably accommodated by the computing systems that are currently being planned in the context of the LHC Computing Grid (LCG).

2.2.2 Reconstructed (RECO) Data

RECO is the name of the data-tier which contains objects created by the event reconstruction program. It is derived from RAW data and should provide access to reconstructed physics objects for physics analysis in a convenient format. Event reconstruction is structured in several hierarchical steps:

1. Detector-specific processing: Starting from detector data unpacking and decoding, detector calibration constants are applied and cluster or hit objects are reconstructed.
2. Tracking: Hits in the silicon and muon detectors are used to reconstruct global tracks. Pattern recognition in the tracker is the most CPU-intensive task.
3. Vertexing: Reconstruction of primary and secondary vertex candidates.
4. Particle identification: Produces the objects most associated with physics analyses. Using a wide variety of sophisticated algorithms, standard physics object candidates are created (electrons, photons, muons, missing transverse energy and jets; heavy-quarks, tau decay).

Requirement: R-5	Event reconstruction shall generally be performed by a central production team, rather than individual users, in order to make effective use of resources and to provide samples with known provenance and in accordance with CMS priorities.

Requirement: R-6	CMS production must make us of data provenance tools to record the detailed processing of production datasets and these tools must be useable (and used) by all members of the collaboration to allow them also this detailed provenance tracking

Reconstruction is expensive in terms of CPU and is dominated by tracking. The RECO data-tier will provide compact information for analysis to avoid the necessity to access to RAW data for most analysis. Following the hierarchy of event reconstruction, RECO will contain objects from all stages of reconstruction. At the lowest level it will be reconstructed hits, clusters and segments. Based on these objects reconstructed tracks and vertices are stored. At the highest level reconstructed jets, muons, electrons, b-jets, etc. are stored. A direct reference from high-level objects to low-level objects should be possible, to avoid duplication of information. In addition the RECO format will preserve links to the RAW information. Sometimes, in case of fast reconstruction algorithms, it is a trade-off between storing more intermediate objects resulting in bigger event sizes, and accessing RAW data.

Requirement: R-7	The reconstructed event format (RECO) is about 250 KByte/event; it includes quantities required for all the typical analysis usage patterns such as: pattern recognition in the tracker, track re-fitting, calorimeter re-clustering, and jet energy calibration.

The access to all physics objects stored in the RECO format should be provided in a uniform way (interface) which should allow to retrieve the configuration (parameters) used for reconstruction. This event size is in agreement with the size of our current RECO (aka DST) format. Only one RECO format will be supported but the ability to store multiple collections of objects reconstructed with different algorithms (versions) should be possible.

2.2.3 Analysis Object Data (AOD)

AOD are derived from the RECO information to provide data for physics analysis in a convenient, compact format. AOD data are useable directly by physics analyses. The AOD will contain enough information about the event to support all the typical usage patterns of a physics analysis. Thus, it will contain a copy of all the high-level physics objects (such as muons, electrons, taus, etc.), plus a summary of the RECO information sufficient to support typical analysis actions such as track refitting with improved alignment or kinematic constraints, re-evaluation of energy and/or position of ECAL clusters based on analysis-specific corrections etc... The AOD will not support the use of novel pattern recognition techniques, or the application of new calibration constants, which will typically require the use of RECO or RAW information.

Requirement: R-8	The AOD data format at low luminosity shall be approximately 50kB/event and contain physics objects: tracks with corresponding RecHit's, calorimetric clusters with corresponding RecHit's, vertices (compact), and jets.

The AOD size is about 5 times smaller than the next larger (RECO) data format. Historically this factor is about the size reduction at each step that can both give important space and time improvements yet still yield sufficient functionality. New versions of it may be produced very often as the software and physics understanding develops. 50kB is consistent with our current best understanding of the data required in an AOD.

Although the AOD format is expected to evolve in time, with information being added to assist in analysis tasks but also being reduced as the understanding of the detector is improved, this size is not expected to change significantly, especially when the potential use of compression algorithms is taken into account.

2.2.4 Heavy Ion Event Data

The CMS computing requirements are dominated by those required for pp physics; however CMS is also approved for running in heavy ion collisions and has a physics program targeted at this interesting area of study. Heavy ion runs are assumed to follow each major pp operation period as described in table 2.1.

Requirement: R-9	The data rate (MB/s) for Heavy Ion running will be approximately the same as that of pp running however event sizes will be substantially higher, around 5-10MB/event.

To develop the computing requirements in this document we have taken a fairly conservative estimate for the average value of $dN/d\eta = 2000$. Event size estimates then using the same methods as above are estimated to be 7 ± 2.5MB. We have also considered a mix of event types and event processing times per type that give a weighted mean processing time of about 200kSI2k.s (about 10 times that of pp events). If processing power were available, full reconstruction times could be of order 5 times slower than this for a central event - and would yield in turn a richer physics program.

Requirement: R-10	Heavy Ion events will be reconstructed during the (approximately 4 month) period between LHC operations periods; it is not necessary to keep up with data taking as for pp running.

Due to the substantial reconstruction processing requirements for Heavy Ion events it is not foreseen to follow their reconstruction in real time at the Tier-0. Rather a fraction of the events will be reconstructed in real time, while the remainder will be processed after the LHC run is complete. We aim to complete this reconstruction in a time similar to the LHC downtime between major running periods

The estimation of event sizes and processing times are not trivial. There are many unknowns, such as the mean multiplicity and the mix of events in the trigger. CMS expects to use regional reconstruction to keep the reconstruction time as low as possible. However we recognize that the latest results from RHIC show that we can extend the physics reach of the CMS-HI running,

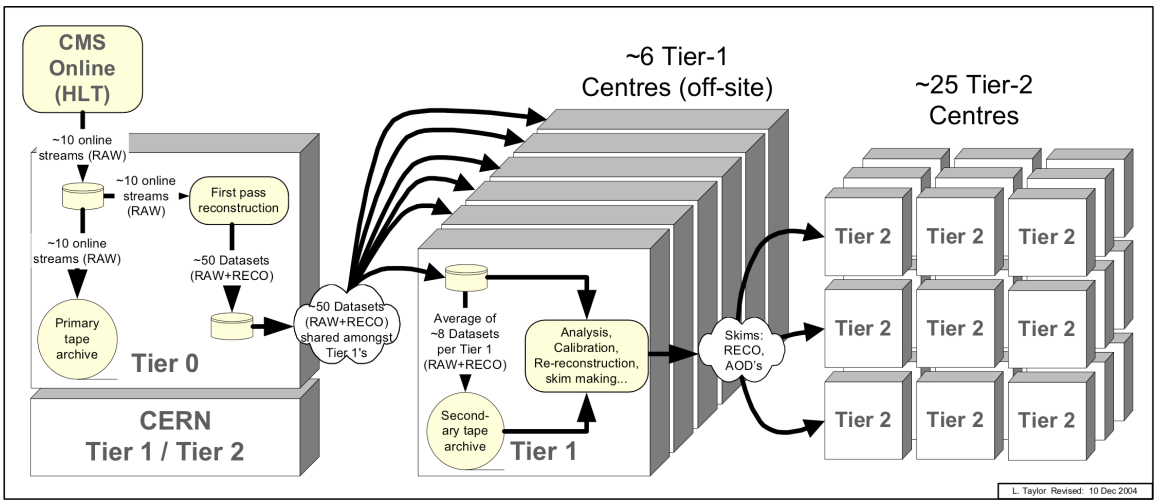

Table 2.3: Schematic flow of bulk (real) event data in the CMS Computing Model. Not all connections are shown - for example flow of MC data from Tier-2's to Tier-1's or peer-to-peer connections between Tier-1's.

by reconstructing more of each event.

Requirement: **R-11**	Heavy Ion reconstruction is costly, 10-50 times that of pp reconstruction. The base Heavy Ion program (as in the CMS Technical Proposal) can be achieved with the lower number, more physics can be reached with the higher.

Increased available computing would allow access to more physics.

2.2.5 Non-Event Conditions and Calibration Data

The are currently no firm estimates for the data volumes of the conditions data produced on the online systems and the calibration data produced offline by calibration programs. These issues will be addressed as part of the Physics TDR studies, which are due for completion at the end of 2005. The total data volume is, however, considered negligible compared to the bulk data needs (RAW, RECO, and AOD). The needs for distributed database services are recognised and will be addressed for the Computing TDR. It is not expected to have a major impact on the hardware resources required.

2.3 Event Data Flow

Figure 2.3 shows the Computing Centres in CMS Computing Model and the schematic flow of the real event data. The CMS online (or HLT) farm processes events from the DAQ system which have successfully passed the L1 trigger criteria. An entire event is distributed to an HLT node which either rejects it forever, or accepts it based on it passing one or more of the HLT selection criteria (the HLT trigger table).

Requirement: R-12	The online system shall temporarily store "RAW" events selected by the HLT, prior to their secure transfer to the offline Tier-0 centre.

This raw event data constitutes the output of the HLT farm. To optimize data handling, raw events are written by the HLT farm into files of a few GB size.

Requirement: R-13	The online system will classify RAW events into $\mathcal{O}(50)$ Primary Datasets based solely on the trigger path (L1+HLT); for consistency, the online HLT software will run to completion for every selected event.

The first attribute of an event that is useful to determine whether it is useful for a given analysis is its trigger path. Analyses rarely make use of more than a well defined number of trigger paths. Thus events will be clustered into a number *Primary Datasets*, as a function of their trigger history. Datasets greatly facilitate prioritisation of first-pass reconstruction, the scheduling of re-calibration and re-reconstruction passes, and the organisation of physics analysis.

Requirement: R-14	For performance reasons, we may choose to group sets of the $\mathcal{O}(50)$ Primary Datasets into $\mathcal{O}(10)$ "Online Streams" with roughly similar rates.

There seems to be no compelling advantages to having a single physics stream (neither online not offline). The proposed strategy is similar to that used by CDF [1]. CMS will use this streaming from the online system as a mechanism for prioritisation of later processing; but the number of such streams cannot be to large. We consider this a technical issue of optimisation - the fundamental physics classifications of interest are the Primary Datasets. The online streams are data management artifacts, they are not visible as such to end users.

The subdivision of events into online streams will allow for example to prioritize processing of a "calibration stream" (One of the Online Streams, one or more of the Primary Datasets) which will result in updated calibration constants to be used for all subsequent processing for that data-taking period. Processing of certain lower-priority online Streams may be deliberately delayed in the event of a partial disruption of service at the Tier-0. Finally, since a given online Stream groups only a subset of the Primary Datasets, handling of production job output will be simplified.

Requirement: R-15	The Primary Dataset classification shall be immutable and only rely on the L1+HLT criteria which are available during the online selection/rejection step.

The Primary Datasets event classifications are immutable; re-classification or rejection of events based on offline reconstruction is not allowed. The reasoning for not rejecting events during re-processing is to allow all events to be consistently classified during later re-processing with improved algorithms, software, and calibrations.

The immutability of Primary Datasets in no way precludes the possibility to form subsets of these Primary datasets for some specific analysis purposes. For example it is expected that subsets of events that further satisfy some more complex offline selection can be made. These subsets

[1] CDF writes about 10 "Online Streams" which are further divided to form about 50 "Primary Datasets", with classification depending only on HLT trigger information. D0 originally classified events only as either "physics events" or "monitor events" but are now introducing streams.

may be genuine secondary event collections (formed by actually copying selected events from the Primary Datasets into new secondary datasets) or may be in the form of *Event-Directories* (list of event numbers/pointers satisfying these selection conditions).

Requirement: R-16	Duplication of events between Primary Datasets shall be supported (within reason - up to about approximately 10%).

The advantage of writing some events into multiple datasets is to reduce the number of datasets to be dealt with for a specific purpose later on (e.g. analysis or re-reconstruction). It facilitates prioritisation of reconstruction, application of re-calibration and re-reconstruction, even if distributed. The total storage requirements should not increase excessively as a result (say not by more than 10%).

In principle therefore different Primary Datasets may contain overlapping events. Provided that it is kept small, it is acceptable to allow event duplication among streams at the price of a more complex book-keeping system. An advantage of keeping the Primary Datasets orthogonal would be to reduce storage and processing needs downstream.

CMS does not need to decide at this time if the 10% duplication will be used or if orthogonality will be forced. By retaining it as a baseline we can ensure that all the tools needed to cope with it are prepared, allowing the best decision to be made at the appropriate time.

Requirement: R-17	The online system will write one or possibly several "Express-Line" stream(s), at a rate of a few % of the total event rate, containing (by definition) any events which require very high priority for the subsequent processing.

As well as being written to a normal online stream, an event may *also* be written into an "express-line" stream (maybe one or several). As the name indicates the sole purpose of this stream is to make certain events available offline with high priority and low latency. The express-line is not intended for final physics analysis but rather to allow for very rapid feedback to the online running and for "hot" and rapidly changing offline analyses. Typical content of the express line could be: events with new physics signatures; generic anomalous event signatures such as high track multiplicities or very energetic jets; or events with anomalously low/high activity in certain detectors (to study dead/noisy channels). All events in the express-line are also written to a normal online stream / Primary Dataset.

Requirement: R-18	The offline system must be able to keep up with a data rate from the online of about 225 MB/s. The integrated data volume that must be handled assumes 10^7 seconds of running.

The numbers above are a baseline that allow the rest of the model to be sized in a coherent way. The HLT farm will write events at the maximum possible data rate, independently of the event size. Trigger thresholds will be adjusted up or down to match the maximum data rate, in order to maintain consistency with the downstream data storage and processing capabilities of the offline systems. All backlogs accumulated during periods of running at peak rates should be absorbed within 24 hours. We assume that in Heavy Ion running periods, CMS writes data from the online farm at the same rate (MB/s).

Requirement: R-19	No TriDAS dead-time can be tolerated due to the system transferring events from the online systems to the Tier-0 centre; the online-offline link must run at the same rate as the HLT acceptance rate.

This will be reflected in specifications of the link and of the Tier-0 mass storage latencies. Even assuming the LHC running time is 50% of each 24 hour period (a number we can only guess at now), we do not apply this factor to the peak rate as we need to transfer events to the Tier-0 in real-time, not in quasi real-time averaged over a full day. This is to allow data to be reconstructed in the Tier-0 farm and monitored in real-time offline to give maximum feedback to the running experiment, even while the LHC fill is still in progress.

Requirement: R-20	The primary data archive (at the Tier-0) must be made within a delay of less than a day so as to allow online buffers to be cleared as rapidly as possible.

2.4 Event Reconstruction

2.4.1 First Pass Reconstruction

Requirement: R-21	CMS requires an offline first-pass full reconstruction of express line and all online streams in quasi-realtime, which produces new reconstructed objects called RECO data.

The Tier-0 offline reconstruction step processes all RAW events from the online system following an adjustable set of priorities (the express-line, by definition has very high priority). This step creates new higher-level physics objects such as tracks, vertices, and jets. These may improve or extend the set produced in the HLT processing step. It must run with minimal delay compared to the online in order to provide rapid feedback to the online operations, for example, identifying detector or trigger problems which can then be rectified dynamically during the same LHC fill.

The offline reconstruction will normally perform the same reconstruction steps for each stream, with the possible exception of specialised calibration streams. In this way we ensure that they are all useful in principle for all analysis groups. We apply this same rule to later re-processings of the data, 2-3 times per year we expect to bring all datasets into consistent status as to applied calibrations and algorithms, as described below.

Requirement: R-22	A crucial data access pattern, particularly at startup will require efficient access to both the RAW and RECO parts of an event

The primary issue here comes from the fact that RAW and RECO parts of the event will most naturally exist in different files. Experience from the CMS data challenges indicates that accessing parts of the data for a single event from two or more files in a single job can put significant demands on the file access/storage/staging systems, often resulting in large numbers of job failures and/or low job throughput. The situation is also complicated by the fact that a large number of local implementations for serving data (CASTOR, dCache, RFIO, xrootd etc)

may be in use at the various sites; all of which must efficiently handle this use-case. While those systems will likely mature over time, it is felt that some alternate solution to accessing multiple files must be available in case problems arise.

The simplest technique to avoid such problems altogether is to rewrite the RAW data in the same file with each new copy of the RECO produced during re-reconstruction. While this might be possible for a few offline streams or at the very beginning, it is not likely to be feasible in the long run due to the storage cost from the extra duplication of the RAW data.

For the Computing Technical Design Report, CMS will examine backup possibilities to provide a single "package" of RAW+RECO (We call this package/union FEVT "Full Event") for data access purposes while minimising the bookkeeping overhead for the user and eliminating needless long-term duplication of data. A trivial example of such a solution might be to provide an additional *temporary* packaging of the latest/active RECO file(s) together with the relevant RAW file(s) in a separate "FEVT" zip archive file. While this results in a second copy of the RAW in the archive, the archive itself can easily be deleted when that version of the RECO is no longer needed/active, leaving only the original, individual RECO/RAW files for long term storage. We would like to have such a solution available to deploy, perhaps even in a site- or stream-dependent way, if necessary.

More importantly, we will ensure that the configurations of our applications are such that the required data files for any given job can be determined a-priori in advance if necessary. This will avoid the problems seen with some of our earlier prototype access implementations (theoretically possible also with POOL [3]) where the full set of necessary files for a given job could only be fully determined by reading the event data itself and finding the relevant pointers.

2.4.2 Re-Reconstruction

Requirement: R-23	The reconstruction program should be fast enough to allow for frequent re-processing of the data.

CMS will need to reprocess data quite often, at least in the beginning when much of the work will involve understanding the detector and triggers, calibration, alignment, algorithm-tuning and bug-fixing.

2.5 Monte Carlo Simulation

Monte Carlo Event samples will be generated to simulate the underlying physics collision. The resulting particles will be tracked through the CMS detector and the electronics and trigger responses will be simulated. Both full and parametrized (fast) simulations will be required. We anticipate using the full simulation package, OSCAR [4, 5], for most of these events. Fully simulated refers to detailed detector simulation based on GEANT4 [6], as opposed to faster parametrized simulations. CMS has developed a fast simulation package, FAMOS [7], that may be used where much larger statistics are required.

Requirement: R-24	Fully simulated Monte Carlo samples of approximately the same total size as the raw data sample (1.5×10^9 events per year) must be generated, fully simulated, reconstructed and passed through HLT selection code. The simulated pp event size is approximately 2 MByte/event.

We currently estimate that we will require the same order of magnitude of simulated events as actual data. If the Monte Carlo requirements greatly exceed this rough real data-sample equality, then more recourse to FAMOS will be necessary. Clearly there are very large uncertainties on the total amount of full and fast Monte Carlo which is required, so ultimately the reality of available resources will constrain the upper limit.

Requirement: R-25	Fully simulated Monte Carlo samples for Heavy Ion physics will be required, although the data volume is expected to be modest compared to the pp samples.

The MC needs for heavy ion physics will need to be estimated in more detail.

In Table 3.4 we note expected Monte Carlo event sizes, processing times etc.

2.6 Analysis Model

Many details of the Analysis Model have still to be confirmed and tested in CMS. Some of this work will be carried out in the analysis of Analysis Scenarios to be carried out in the TDR process. We have based the model we describe here largely speaking on methods that have been tried and tested in recent similar HEP experiments. Work will continue on this and we expect the model to be steadily refined, but it is also possible that there can be significant changes if/as we become convinced that new methods of for example Grid-based analysis are ready for full scale deployment. We anticipate however starting with a traditional model for CMS analysis and introducing changes only as they are tested and validated.

The Primary Datasets are central to our data management and our analysis planning. We call this vertical streaming. We supply to the physicist vertical slices of the data (everything to do with some type of event). In contrast to horizontal streaming where there is a RAW Data Sample, a RECO format, and an AOD format for the whole dataset. The AOD→RECO→RAW data are linked by software pointers allowing full navigation when required from event format Y back to event Format X.

Such navigation should be protected by mechanisms to avoid accidental unwanted requests for large scale data access. There are clearly circumstances, for example event visualisation, where such a capability is very useful.

We have developed this plan based on experience in CMS and other experiments, paying particular attention to the approaches of CDF and D0. [2].

[2]In CDF, most end-user analysis uses common analysis group-wide ROOT-tuples which a derived from processing (RawData+DST) files. Most D0 analysis is based on "thumbnail" (TMB) format event samples of selected events (known as skims). There are typically 25 such skims samples which are produced in conjunction with the "TMB-fixing" step which updates events for new calibrations, bug-fixes, etc.

2.6.1 Analysis of RAW and RECO Event Samples

Requirement: R-26	CMS needs to support significant amounts of expert analysis using RAW and RECO data to ensure that the detector and trigger behaviour can be correctly understood (including calibrations, alignments, backgrounds, etc.).

Requirement: R-27	Physicists will need to perform frequent skims of the Primary Datasets to create sub-samples of selected events.

They may perform this individually or as a group activity. They may select subsets of events. They may run further reconstruction. They may output pure FEVT skims, or they may output new RECO or AOD or Event Directories or they may output specialised data formats. These skims are copied to the Tier-2 centres where they can be studied in detail.

From time to time official reprocessing passes of the FEVT Datasets are carried out. These reprocessed FEVT's can be used for new skims. Having finely segmented FEVT Datasets allows much flexibility in using reprocessing CPU according to the experiment priorities. This places stricter requirements on bookkeeping and mechanisms to ensure that analyses can act coherently across a number of Primary Datasets. Having current, but official, reprocessed FEVT's allows the physics community of CMS to work coherently.

2.6.2 Analysis of RECO and AOD Event Samples

Requirement: R-28	CMS needs to support significant physics analysis using RECO and AOD data to ensure the widest range of physics possibilities are explored.

The AOD and, to a lesser extent, the RECO formats are the primary analysis data formats and should be available (normally on disk) at as many sites as we can afford. The more copies we can have the more flexibly can the computing respond to the analysis requirements.

Requirement: R-29	The AOD data shall be the primary event format made widely available for physics analysis in CMS.

Its contents should (and will) be such that more than 90% of all physics analysis in CMS can be carried out from AOD data samples. It is only in few, less than 10% of the analyses that CMS physicists should have to refer to the full RECO data to perform various detailed studies.

AOD data (Sometimes known as mini-DST) are small (about 50kB/event) and are very useful once the detector and software understanding matures [3].

[3] CDF and D0 are moving towards a rather similar situation in which it requires about ∼100 kB of data per event to do most physics analyses at the Tevatron. This is true even though they approach this from opposite directions - CDF successively prunes and refines the (larger) full event whereas D0 successively adds to the (smaller) thumbnail events.

Requirement: R-30	Access to information stored in AOD format shall occur through the same interfaces as are used to access the corresponding RECO objects.

All missing information shall be conveniently defaulted to indicate its absence, in such a way as to guarantee consistent behaviour when the analysis code is run off full RECO information. If the class interface allows operations on the object data that indirectly require access to RECO information (e.g. a request to add RecHit's to an ECAL cluster), the analysis code invoking one of these operations on AOD data should raise an exception and quit.

2.6.3 Event Directories and TAG's

Requirement: R-31	An "Event directory" system will be implemented for CMS.

In short, an event directory is a means of gaining direct (random) access to a selection of event data representations located sparsely in many files without having to read or scan all of the events in those files. The primary advantage of event directories is that they allow the physicist to track their event selections without requiring him/her to make a full copy of the selected events, thus minimising disk-space usage.

It is expected that event directories will be useful in particular to describe secondary (tertiary, ...) datasets defined by selections applied during analysis. It is therefore not expected that event directories will be produced either by the HLT or by the production offline reconstruction.

In addition to the functionality of pointing to event representations, it is expected that an event directory will contain by value the following keys such as the Event#, Run# (or some Event Id), and HLT trigger bits, possibly generalized to a "TAG" including some number of user defined quantities). At a minimum it should be possible to configure a framework job to use an event directory as input; write an event directory as output; and use some tool to "deep-copy" the event representations pointed to by an event directory, in order to optimise data access and facilitate transport of that event data subset to another location

Open questions regarding event directories include how the "dataset book-keeping" and data management interact with and use event directories, and if and how a user can build an event directory usable by the framework when working with the data from an interactive session, e.g. ROOT [8].

We have not yet made any determination on the technology choice for implementing Event Directories.

Requirement: R-32	Smaller and more specialised TAG/tuple data formats can be developed as required.

Derivative data formats are not expected to be very important in the early running. Many analyses will need access to RAW and RECO data for the first year(s); most will need the complete AOD, which may be reproduced quite frequently. Only as experience develops will it be possible to write significantly reduced data formats, because prior to this the detector and trigger must be very well understood. Candidates include TAG formats which are essentially NTUPLE formats describing the main features of the event for rapid selection of events of

interest from large samples, which may or may not have references/pointers back to the less compact data formats. We do not address these formats in detail here. They are, by definition small. We however expect these data formats to also comply to the Primary Dataset selections.

CMS considers it very important that such specialised data set formats do not proliferate, but that there should be a few well defined and supported ones that can be regularly produced and which all members of the collaboration can make use of, knowing the exact provenance of the data.

2.7 Middleware and Software

Since this document is focused on the top-level view of the CMS Computing Model, we do not describe the Grid middleware nor the applications software in detail in this document.

We assume that the Computing Centres will be, by definition, part of the LHC Computing Grid. We use the term LCG to define the full computing available to the LHC (CMS) rather than to describe one specific middleware implementation and/or one specific deployed GRID. We expect to actually operate in a heterogeneous GRID environment.

Requirement: **R-33**	Multiple GRID implementations are assumed to be a fact of life. They must be supported in a way that renders the details largely invisible to CMS physicists.

Assuring a homogeneous interface to several heterogeneous implementations will be a responsibility of the regions bringing GRID resources to CMS; of coordination organizations such as the LCG itself; and to a limited extent on the CMS computing project itself to build application layers that can operate over (a few at most) well defined grid interfaces.

Requirement: **R-34**	The GRID implementations should support the movement of jobs and their execution at sites hosting the data, as well as the (less usual) movement of data to a job. Mechanisms should exist for appropriate control of the choices according to CMS policies and resources.

We assume that all resources are usable in this way, with finite and reasonable development and support required from CMS itself.

The various Grid projects are described elsewhere, e.g.: LCG-2 Operations [9]; Grid-3 Operations [10]; EGEE [11]; NorduGrid [12]; Open Science Grid [13].

Chapter 3

Specifications – the CMS Computing Model

In this chapter we describe the CMS Computing Model and the specifications that are determined based on the physics requirements described in the previous chapter.

We give the top-level estimate of needs: storage, processing, network. The needs are broken down to the different computing centre levels: Tier-0, Tier-1 and Tier-2 are considered. Profiles for system growth will also be given, starting from the first full LHC run in 2008 to full luminosity, which is assumed for 2010. Wherever appropriate these are based on the PASTA3 report [14] and on LCG reports on estimating the CERN Computing [15] [16].

3.1 Overview

The CMS Computing Model makes use of the hierarchy of computing Tiers as has been proposed in the MONARC [17] working group and in the First Review of LHC Computing [18]. The service agreements for such a hierarchy are being developed now in the LCG Memorandum of Understanding Working Group, and we do not re-discuss them here, although they will form an under-pinning of our Computing Model

We expect this ensemble of resources to form the LHC Computing Grid. We use the term LCG to define the full computing available to the LHC (CMS) rather than to describe one specific middleware implementation and/or one specific deployed GRID. We expect to actually operate in a heterogeneous GRID environment but we require the details of local GRID implementations to be largely invisible to CMS physicists (these are described elsewhere, e.g.: LCG-2 Operations [9]; Grid-3 Operations [10]; EGEE [11]; NorduGrid [12]; Open Science Grid [13].

Assuring this homogeneous interface to heterogeneous implementations will be a responsibility of the regions bringing GRID resources to CMS; of co-ordination organisations such as the LCG itself; and to a limited extent on the CMS computing project itself to build application layers that can operate over (a few at most) well defined grid interfaces. In the following we assume that all resources are usable in this way, with finite and reasonable development and support required from CMS itself. We do not have significant resources to expend on making non-standard environments operational for CMS.

CMS has chosen to adopt a distributed model for all computing including the serving and archiving of the raw and reconstructed data. This assigns to some regional computing centres

some obligations for safeguarding and serving portions of the dataset that would have previously been associated with the host laboratory. The CMS Computing Model includes a Tier-0 centre at CERN, approximately 6-10 Tier-1 centres (including CERN) located at large regional computing centres, and about 25 Tier-2 centres. Figure 2.3 shows the Computing Centres in the CMS Computing Model and the schematic flow of the real event data.

The Tier-0 Centre at CERN is by definition a common facility of CMS. It performs well organised sequential processing of data from the Online and in re-reconstruction passes. We plan to keep one copy of the RAW data at CERN and a second copy distributed over the Tier-1 centres.

Tier-1 Centres have a responsibility to all of CMS, plus additional roles towards their local users (for some suitable definition of "local users", we mean here not necessarily geographically local but also groupings of physicist with common physics interests). Tier-1 centres have a mixture of task types - organised sequential processing and more chaotic analysis activities. Since a Tier-1 centre may have the only available copy of some data samples, they must allow any CMS user to access it. However different users, or groups of users may have different priorities. These relative priorities are set by CMS and by the CMS physics analysis organisation, they are not set by geographical factors. CMS physicists perform selection, skims, reprocessing etc. on the Tier-1 centre computers processing data that has been pre-located at the Tier-1 centres by CMS. Some local users may have long term local storage at the centre, others would move the results to another computer centre (typically a Tier-2 centre). We expect that some analyses would be fully carried out at Tier-1 centres. There are likely to be groups of local users who use the Tier-1 in the usual way but who also have access to local storage and local batch and interactive facilities for their analysis activities. We do not preclude this mode of operation, but expect most CMS users to use a Tier-2 centre as their computing interface to CMS analysis and for them to have local accounts on at least one such centre.

Tier-2 Centres are more responsive to local priorities rather than central CMS requirements. Thus, Tier-2 centres are not generally obliged to respond to computing requests from all CMS users. Most of the Tier-2 facility is used by some groupings of CMS Physicists to perform their iterative analyses. A significant fraction of the resources of a canonical Tier-2 centre is, however, dedicated towards scheduled Monte Carlo Simulation of common CMS samples.[1] CMS does not centrally mandate where skim data and the such-like are kept. However the Tier-2 centres do have to facilitate access to their published data products so that other CMS physicists can analyse them. By not attempting central control of Tier-2 resources we will encourage a certain degree of diversity. We believe this diversity will encourage competition and creativity in analysis of CMS Physics. CMS has the responsibility to supply a solid basis to this analysis activity by assuring that consistently reconstructed datasets are usable by the collaboration.

Tier-3 Centres are modest facilities at institutes for local use. Such computing is not generally available for any coordinated CMS use but is valuable for local physicists. We do not attempt at this time to describe the uses or responsibilities of Tier-3 computing. We nevertheless expect that significant, albeit difficult to predict, resources may be available via this route to CMS.

3.2 The Tier-0 at CERN

The Tier-0 is by definition a purely production facility. It has the following responsibilities:

[1] It is possible that some Tier-2 centres may be well suited to perform the Heavy Ion event reconstruction that is CPU intensive but has modest I/O requirements

- Secure the raw data copy at CERN.
- Perform the real-time reconstruction of the incoming data and stream it into Physics DataSets
- Distribution of that reconstructed data to the collaboration.
- When LHC is not running the Tier-0 will be used to carry out massive re-reconstruction passes. At these times it may also carry out the initial Heavy Ion event reconstruction

It's focus on sequential activities may lead to hardware optimizations directed to such scenarios.

3.2.1 Tier-0 interface with the CMS Online Systems

Specification: S-1	The link from the online to the Tier-0 centre should be sized to keep up with the event flow from the Online farm, with an additional safety margin to permit the clearing of any backlogs caused by downstream throughput problems in the Tier-0.

Data arrives on the Tier-0 input buffer. In case of blockages in this transfer and later recovery, the link speed (and end-to-end components of the link) must be sized so as to recover quickly from the backlogs. Since the Online Buffer space will be of order a few days, one must be able to recover from a two day backlog in say two days thus requiring a two-times safety factor.

There are no large volume data recording facilities at the Online farm. Raw Data files must be transferred for storage at the Tier-0 (and/or Tier-N) centers as they reach a convenient size and are closed. In addition, all data files from an LHC fill must be closed and flushed to the Tier-0 at the end of the fill, regardless of their size. In what follows it is assumed that data from the HLT farm are subdivided into $\mathcal{O}(10)$ Online Streams. With an assumed output rate of the HLT farm at LHC startup of 200 MB/s, a new Raw-Data file of 2GB would be completed every 10 s on the average. Although in reality a more complex time structure will be required to support e.g. low-latency streams for monitoring and calibration feedback, it is a reasonable goal to attain an average job latency of the same order of magnitude as an LHC fill, say 4-8 hours. While larger file sizes are now quite practicable, a 2 GB file would correspond to an event sample which can be reconstructed in a time consistent with this projected latency in a Tier-0 farm with $\mathcal{O}(1000)$ nodes. Ultimately, the exact choice of file size will be a balance between data-management and data-processing issues.

Small files cause considerable difficulties for distribution and storage. At the end of data challenge DC04 CMS demonstrated (reference) that the use of a simple zip archiving (without compression) can allow the construction of optimally sized file concatenations without affecting inter-file references. We anticipate that some such mechanism will be used in the Online farm to build appropriately sized files respecting the required latencies for different Online Streams.[2] There are no large-volume data recording facilities at the Online farm. Data must be transferred for storage at the Tier-0 (and/or Tier-N) centres.

[2] Such concatenation will also be required to meet the performance requirements of network and data storage elements throughout the computing model; we assume that appropriate sizing of files will be a problem that the CMS data management tools must facilitate

3.2.2 Tier-0 First Pass Reconstruction Processing

Specification: S-2	The processing capacity of the Tier-0 centre should be sufficient to keep up reconstructing the real-time event flow from the CMS online system.

We aim to process all the online streams in an initial reconstruction step at the Tier-0. The Tier-0 centre must have sufficient computing power available to reconstruct the data coming from the DAQ system, we propose that the maximum acceptable backlog that may be accumulated is about 24 hours: Furthermore sufficient resources need to be available to meet the time critical nature of the express line.

However, the latency to initial processing can depend on the availability of appropriate calibration data. The reconstruction step will output the events in the $\mathcal{O}(50)$ Primary Datasets which constitute the main output of the reconstruction step. As with the Online farm, we expect to make use of a mechanism, such as the zip-archiving or an explicit file concatenation step, to obtain reasonably sized Primary Dataset file concatenations matched to the latency requirements.

Specification: S-3	The RAW and RECO data components (i.e. the FEVT) of a given set of events are, by default, distributed together. The technical ability to ship them seperately should the need arise shall be maintained.

Technically they may either be in the same physical file or in distinct files which are simply kept close by. There are advantages/disadvantages to both possibilities. We do not make this decision at this time. We have some experience in data challenges that inter-file links can lead to significant numbers of job failures particularly in analysis jobs processing events from a large number of files.

Access to data has been a major problem for all recent HEP startups, we favour erring on the side of simplicity for the initial running and only adopting more complex scenarios when they have been convincingly shown to work at the appropriate scale

Whichever solution we adopt, we cannot afford, nor intend, to keep multiple copies of RAW data. We do not adopt a solution at this time, but specify that we will not keep multiple redundant copies of the RAW data beyond the two currently foreseen; except where access requirements to that data at multiple Tier-1 sites dictates this.

Specification: S-4	The first pass reconstruction step will also produce the AOD data, a copy of which is sent to every single Tier-1 Centre.

As soon as possible in the lifetime of the experiment CMS will produce AOD format datasets aimed at easing the analysis tasks of the majority of the physicists of the experiment.

3.2.3 Tier-0 Data Storage and Buffering

Specification: S-5	Two copies of the CMS RAW Data shall be kept on long term secure storage media (tape): one copy at the Tier 0 and a second copy at the ensemble of Tier-1 centres.

Specification: S-6	The Tier-0 shall store all CMS RAW data on secure storage media (tape) and maintain it long-term.

The main output of the Tier-0 will be the FEVT, which is a union of RAW and Reconstructed data. It would be most economical if this FEVT copy at the Tier-0 were also the CERN copy of the RAW data. However, in the interests of releasing the online buffer space as quickly as possible, we would not like to wait for the reconstruction to be completed before securing a tape copy of the RAW data.

We propose to backup the copy of the RAW data in the Tier-0 input buffer by an explicit copy to a second, temporary, disk buffer. As soon as this has been done the Online output buffer space can be released. Once the Reconstruction step at the Tier-0 is completed and the FEVT is secure on tape at CERN and at a remote Tier-1 centre, then this temporary disk backup can be liberated.

Specification: S-7	The Tier-0 centre shall store a secure copy of all data which it produces as part of its official CMS production passes, including first pass reconstruction (RECO) output, subsequent re-processing steps, and any AOD's produced.

We currently have no estimate of how long such data should be stored. Except in the case of clearly erroneous productions, this will be at least a few years. Some critical experiment data must be kept essentially forever and storage migration for this must be taken into account.

Specification: S-8	Tape Storage at the Tier-0 and Tier-1 centres shall be used as a trusted archive and an active tertiary store

The experience of previous experiments is that tape storage serves as both a (more) trusted archive, and an active tertiary store.[3]

It is likely that tape will play a significant role in CMS computing models for at least the first years of LHC running. This is due to a number of factors: the expected higher cost of disk compared to archival tape; the rate of data and amount of MC; the need to re-reconstruct archived data samples; and the need to access data from previous years. The Tier-0 storage facility must have sufficient read capacity to retrieve data sufficiently quickly to keep the computing resources utilized for re-reconstruction.

The Tier-0 facility will then be responsible for long time secure storage of:

- A copy of all RAW data (initially the unadulterated online format; perhaps later it will be compressed in a lossless or lossy fashion).
- A copy of the first pass RECO data
- The first pass AOD data

[3] In the Tevatron Run2 experiments, of order 3-4 bytes leave the store for every byte that enters the store.

Specification: S-9	The Tier-0 storage and buffer facility shall be optimised for organised and scheduled access during experimental running periods

The Tier-0 centre is not designed to have sufficient resources to serve out of time requests for archived data. The I/O to the Tier-0 mass storage system should be used to write data during running periods and read data during reconstruction running. Non-scheduled data serving requests will need to be processed by Tier-1 regional centres, or at the CERN Tier-1. (In the case of the CERN Tier-1 centre additional mass-storage disk buffers and Tape I/O will be required to service the Tier-1 needs separately from the Tier-0 ones.)

Specification: S-10	The Tier-0 will not support logins from general CMS users, only those carrying out specific production related activities.

3.2.4 Tier-0 Re-Processing

Specification: S-11	The Tier-0 shall support at least one complete re-reconstruction pass of all RAW data, using calibrations and software which are improved compared to the original first-pass processing.

The duty cycle of the LHC means that the Tier-0 centre will not be processing raw data for about 4-6 months of the year. This free capacity and the availability of the complete RAW data sample at CERN means that the Tier-0 will be used to perform large-scale reprocessing of the RAW data sample using updated calibrations and conditions data and improved software.

The exact scheduling and prioritisation of specific datasets will depend on physics priorities, latencies of producing offline calibration results, and which problems need to be resolved. It is assumed that the Tier-0 can re-process all RAW data (at least) once during LHC shut downs.

Likewise the Online Farm, or parts of it, may be available as batch processing farms when the accelerator is not operating. We do not rely on this as experience shows that the Online Farm may be heavily used for other purposes or undergoing maintenance/extension during the LHC off-periods. We can expect to be able to complete one reprocessing step per year at the Tier-0 possibly operating with the Online farm. If so, we must account for the storage at the Tier-0 of the new RECO component and the RECO for each new reprocessing pass. The resultant data sets must also be redistributed to the Tier-1 centres.

Specification: S-12	The re-reconstruction step will also produce the AOD data, a copy of which is sent to every single Tier-1 Centre.

AOD's produced in the initial reconstruction step must also be distributed to the Tier-1 centres and stored at the Tier-0 (this is the baseline). Alternatively its production would be scheduled at the Tier-1 centres on receipt of the Primary Dataset there.

3.2.5 The Tier-0 and Heavy Ion Processing

Heavy Ion event reconstruction is characterized by potentially very high CPU usage, from 10 to 50 times that of pp events. It is not economically practicable to keep up with the HI data acquisition with even pseudo-realtime reconstruction. We expect then to only reconstruct a fraction of the data in the Tier-0 during HI operation. Later using about half of the proposed Tier-0 facility we could complete the regional reconstruction of the HI events during the four months of each year without LHC operation. This is one possible solution to the HI reconstruction challenge, but it could interfere with the re-reconstruction discussed above.

Specification: S-13	About half of the Tier-0 capacity could be used to perform regional reconstruction of Heavy Ion events during LHC downtimes. This time does however eat into that available for re-reconstruction

As discussed in the previous chapter, there would also be physics advantages in more fully reconstructing these events, this does not seem possible at the Tier-0 without compromising the re-reconstruction of pp data or without significantly upgrading its capacity. An alternate, or additional, solution for HI processing would be to use some Tier-2 resources. As noted before, HI reconstruction is an I/O-light and CPU-heavy task; this is the type of task well suited to being carried out on Tier-2 centres, as is Monte Carlo production

Specification: S-14	CPU resources at some Tier-2 centres could be used to carry out the Heavy ion initial reconstruction, or to extend that reconstruction to allow more physics coverage

Given the current uncertainties in event sizes, rates and reconstruction times we cannot choose now the actual Heavy Ion reconstruction scenario. We expect it to be a combination of the outlined options. We will aim to reconstruct the largest sample possible during the actual Heavy ion running, completing that task using some of the Tier-0 in the LHC downtime and some, possibly dedicated, Tier-2 centres. Within the proposed capacity we can achieve our primary Heavy Ion program but note that this program could benefit from increased computing capacity.

3.2.6 Tier-0 Wide Area Network connectivity

Specification: S-15	The Tier-0 shall coordinate the transfer of each Primary Dataset in FEVT format, and all AOD data produced, to a "custodial" Tier 1 centre prior to its deletion from the Tier-0 output buffer.

We call these Tier-1 sites, "Custodial Sites", because they contract with CMS to maintain a secure copy of the initial FEVT (and hence the second RAW data copy). The Primary Datasets can be further distributed to other Tier-1 sites "Discretionary Sites", to make them more readily available for physicists to access. These sites do not have to guarantee a secure copy, they can always make up for lost files by going back to the custodial site. This Tier-1 site can then host reprocessing steps based on these Primary Datasets. There may also be a case for ensuring that all Primary Datasets corresponding to an Online Stream are hosted at a particular Tier-1

centre.

A full specification would include not only bandwidth, but also the quality of service which is also a cost driver. CMS is not planning for real-time (within a couple of hours) feedback from the Tier-1 centres to the DAQ. If T1 centres were seen as a natural extension of the DAQ, this would more stringently define aspects of the network service and redundancy.

In calculation of the WAN requirements we take account the baseline Gb/s, and two additional factors; one (CMS) safety factor to describe additional unaccounted requirements on actual data transfer and a second (Network), headroom, factor to account for usable bandwidth in a given network connection.

3.2.7 Summary of Tier-0 Parameters

Specification: S-16	The Tier-0 centre shall support a range of collaboration services such as: resource allocation and accounting, support for CMS policies; high- and low-level monitoring; data catalogs; conditions and calibration databases; software installation and environment support; virtual organisations and other such services.

From the above considerations and the input parameters specified in Table 3.4 we have used a spreadsheet calculation to estimate the Tier-0 specifications given in Table 3.1.

In this table, and the Tier-1 and tier-2 ones that follow, we have used some efficiency factors that have the effect of increasing the resources. They are extracted from our experience in this and earlier experiments. While it is possible to use CPU in production activities quite efficiently, this never of course reaches 100% due to job startup and ending, data transfer, emptying job queues, system uptime etc., we take this into account as a scheduled CPU efficiency (taken to be 85%). Analysis activities suffer these same problems but much more; for example the analysis programs may need to be moved to the nodes, they may not be locally resident, the staging in of data is typically less well prepared, the job queues more prone to user errors etc., we ascribe this a worse efficiency factor (taken to be 75%). Disk space can never be 100% used: there is always data in movement, file sizes that do not match disk sizes, redundant data that has not yet been purged (or even identified as not being required) etc., we ascribe a disk utilisation efficiency of 70%.

3.3 Tier-1 Centres

Tier-1 regional centres have aspects of custodial data storage, re-reconstruction, data analysis and are also responsible for serving data to Tier-2s for analysis, MC storage and user support. The specification needed for the Tier-1's are the processing for re-reconstruction of custodial data sets, the storage of custodial data sets, networking in from Tier-0, Tier-1 interconnectivity and Tier-1 to all Tier-2s, processing for Analysis and IO from active storage for analysis.

In this paper we define what is needed at a canonical (or average) Tier-1 centre on the assumption that there are 6 such centres for CMS outside CERN plus one at CERN. We expect that some centres may be able to supply more or less than this canonical specification but base our estimates on the assumption that the aggregate resources/services of the N actual centres is equivalent to

The Tier 0 Centre at CERN

The Disk and Tape Storage

ANNUAL TOTALS	# of events	Ev-size MBytes	Tape/disk Tbytes			
			Active	Archive	Disk	
Raw data	1.5E+09	1.5	2250		225	*Note 1*
Heavy Ion Raw Data	5.0E+07	7	350		0	*Note 8*
Calibrat.	1.5E+08	1.5	225		23	*Note 2*
First RECO	1.5E+09	0.25	375		38	*Note 3*
Reprocessed RECO	1.5E+09	0.25	375		0	*Note 4*
HI RECO	5.0E+07	1	50		0	
First AOD	1.5E+09	0.05	75		0	*Note 5*
Reprocssed AOD	1.5E+09	0.05	75			*Note 6*
Total			3775	0	285	

CPU

Data Processing	# of events to Mass Stor.	CPU per event KSI2K/ev.s	CPU total kSI2K	
Reconstruction	1.5E+09	25	3750	
Heavy Ion reconstruction	5.0E+07	200	3858	*Note 9*
Reprocessing	1.5E+09	25	included	*Note 7*
First Pass Calibration	1.5E+08	10	150	
Total			3900	

WAN

	Raw Rates	Safety	Headroom	Totals
	Gb/s	Factor	Factor	Gb/s
From Online	1.8	2	2	7.2
Total Incoming				**7.2**
FEVT Data to Tier-1's	2.1	2	2	8.4
AOD to all Tier-1s	0.4	2	2	1.3
Total Outgoing				**9.7**

Note 1: Input Disk buffer size, 20 days.
Note 2: 10% Fraction of rawdata for detailed calibration analysis
Note 3: Output disk buffer size, 20 days
Note 4: One Re-reconstruction pass when LHC off
Note 5: main Analysis format (when stable)
Note 6: One Re-reco pass
Note 7: Reprocessing assumed to use Tier-0 when LHC off
Note 8: Disk use is not during pp time, so not totalled
Note 9: CPU required to complete Heavy ion Reconstruction 1 month (Real Time)
Note 10: 50% Safety factor to make up for backlogs

Summarized Requirements before efficiency factors are applied

CPU scheduled	3900	kSI2K
Disk	285	Tbytes
Active tape	3775	Tbytes
Tape I/O	300	MB/s

Requirements after application of efficiency factors

			Eff Factors	
CPU scheduled	4588	kSI2K	*85.00%*	
Disk	407	Tbytes	*70.00%*	
Active tape	3775	Tbytes	*100.00%*	
Tape I/O	600	MB/s	*50%*	Note 10

Table 3.1: Parameters of the Tier-0 Centre.

that of sum of these canonical centres. For guidance, we anticipate that Tier-1 centres sized at less than about 1/2 of this canonical value may not be economically or technically practical.

Each Tier-1 centre has the following roles in CMS:

- Securing, and making available to users, a second copy of a share of the RAW data and reconstructed RECO data (The FEVT)
- Receiving and making available a copy of the full CMS AOD data-set.
- Participating, with the Tier-0, to the timely calibration and feedback to the running experiment.
- Running large scale Physics Stream skims and selected reprocessing for analysis groups and individuals of CMS.
- Serving data-sets to the Tier-2 and other regional or institute computing facilities
- Securing and distributing Monte Carlo simulated samples produced in the Tier-2 and other centres. [4]
- Running production reprocessing passes of Primary Datasets and Monte Carlo Samples.

3.3.1 Tier-1 Custodial Data Storage

Specification: S-17	The ensemble of non-CERN Tier-1 centres shall store the second "custodial" copy of the FEVT (= RAW + RECO) data coming from the Tier-0, on secure storage media (tape) and maintain it long-term.

Tier-1 computing centres must be prepared to provide custodial data storage of at least $1/N$ of the raw data set per year, where N is the number of non-CERN Tier-1 centres[5]. Custodial storage can technically be implemented with a variety of solutions. Acceptable data risk is currently based on archival tape systems. Disk based and hybrid systems should demonstrate an acceptable level of data risk.

Tier-1 centres must have the ability to read the data stored on the centre with sufficient performance to efficiently make use of re-reconstruction resources and analysis.

3.3.2 Tier-1 Reconstruction Processing

Specification: S-18	The Tier-1's must have sufficient processing resources to re-reconstruct the RAW data entrusted to that centre twice per year, in addition to the single full reprocessing at the Tier-0 during the LHC shutdowns.

[4] We do not consider it wise at this time to assign custodial data (MC) storage to Tier-2 sites. Many do not have tape systems. Most do not have large staffs that can ensure the accuracy of published data or for example ensure general external access to all their data. Tier-2 centres may not be able to respond quickly to collaboration reprocessing requirements. Finally, if CMS data were to be published by a Tier-2 centre for general collaboration access, we believe this would place unnecessarily high demands on the Grid infrastructure in the early days of LHC operation. This decision can of course be revisited when these factors have been shown to be unimportant, or deemed to be less important than counterbalancing arguments

[5] Clearly the second secure copy away from CERN, cannot be at CERN

Three re-processings per year is an average estimate - some samples of great interest may be processed many more times, while others may receive less attention. The streaming approach CMS has adopted allows very important prioritisation to occur to ensure that key datasets are finished early in such a four-month period, rather than getting each dataset completed only at the end of this period.

Specification: S-19	The Tier-1's must have sufficient processing resources to re-process (reconstruct) twice per year the MC samples which they host .

This reprocessing is required to keep in line with the software and algorithmic improvements, and modeling of detector response in the simulation to match the actual performance of CMS over a given running period (e.g. modeling the time-dependence of dead or noisy channels).

Specification: S-20	Tier-1 centres must store a secure copy of all data they produce as part of official CMS production passes, including RECO and AOD formats.

The exact boundaries of "official" need to be clarified. As noted above we do not yet have a policy for deleting of old versions of data. The definition of custodial site is then currently that such custodial data is maintained "'forever". This lack of ambiguity will protect us in the absence of tools to manage this process. When a deletion policy is developed, tools to ensure that deletion is safely managed will be required.

3.3.3 Tier-1 Analysis Capacity

Specification: S-21	Tier-1 centres shall support limited interactive and batch analysis of data which they host.

An expected use case is individual physicists or groups running on production datasets (RAW, RECO, AOD) and performing studies/calibrations/analyses directly.

Specification: S-22	Tier-1 centres shall support massive selection and skim passes through the data that they host and distribute the product datasets to the requesting Tier-2 centres

Estimating analysis requirements is extremely difficult. We have considered a number of internally consistent ways to estimate the analysis capacity. In the spreadsheet we have developed to estimate these requirements we have used a rather simply expressed model. It is our believe that more complex models are in fact unlikely to shed much more light on the eventual scale of these requirements in the short term. It is however important to fully explore the consistency ramifications of these scenarios during the writing of the computing TDR's. Modeling of some key components may be required, though modeling of the entire problem space is unlikely to add more than is put into the model in the first place.

Tier-1 centres must have sufficient processing resources to meet the analysis needs of the supported community. We make some assumptions on the frequency that selection passes over the

locally cached data are performed. The outputs of these selection passes are skimmed datasets that are typically transferred to a Tier-2 centre for detailed analysis. This calculation also leads to the estimate of the rate of data from the T1 storage facility to the analysis processes; this is NOT the Tape I/O rate as it depends on the actual ratio of disk cache to active dataset size which can be a locally determined decision.

The analysis model we consider is that most physicists run selection/skim passes on T1 centres and move output datasets to their T2 centres for further analysis.

We calculate the CPU power that would be required to make a single selection pass over every event in the current RECO data sample (Data and MC) at a T1 centre processing events at about 0.25kSI2k per event, that is 4Hz on a 2004 variety CPU. We ascribe this to be the Required Selection CPU power. Such a selection pass would require 800MB/s of data serving at the Tier-1.[6] We cannot prove at this time that 0.25kSI2k is required or sufficient for an average event selection, but it seems plausible and the resultant CPU estimate and Disk I/O requirements are compatible with what one can expect of a Tier-1 centre. The argument may be slightly circular, but it demonstrates that it is a consistent solution.

Specification: S-23	Tier-1 centre selection facilities will require high performance (order 800MB/s) data-serving capacity from their local data samples to their selection farms

This data-serving capacity is far from trivial to achieve. The disk cache management must be able to effectively pin frequently accessed data while allowing infrequently used data to be staged in and out. This will place demands on both the local data-management systems but also on tools that can enable CMS to optimise the, actual or logical, partitioning.

Such a rate would permit a complete selection pass over the local data every two days; bearing in mind that we expect of order 10 analysis groups to be active, this would actually correspond to each group having such a capability every three weeks. We consider this to be a reasonable target processing rate.

Specification: S-24	Tier-1 centres must offer sufficiently granular job submission queues to enable CMS to partition priorities arbitrarily between (perhaps different) analysis groups and individuals

3.3.4 Tier-1 Networking

Specification: S-25	Each of the $(N_{T1}-1)$ Tier-1 centres must size its network to: accept its $\sim 1/(N_{T1}-1)$ share of total RAW and RECO data produced at the Tier-0 during running periods; accept MC production data from $\sim N_{T2}/N_{T1}$ of the N_{T2} Tier-2 centres; and export requested datasets to $\sim N_{T2}/N_{T1}$ Tier-2 regional centres.

(The $(N_{T1}-1)$ being used to take into account that the second copy is not at CERN)

[6] This may be serving of disk resident data or of data from a tertiary store, we do not specify here the solution but state the overall data serving requirement

Of these network components, the most demanding is the serving of data to the Tier-2 centres.

3.3.5 Tier-1 Centre at CERN

Specification: S-26	CMS requires Tier-1 functionality at CERN

Because of the presence of the Tier-0 at CERN for massive processing and the tape store associated to this, we can imagine that the CERN-T1 could be differently sized and even play different roles than the offsite Tier-1's. It could be that it is in effect a combined Tier-1 and Tier-2 centre. It has access to the tape copies of all the (CERN produced) data samples. Nevertheless, CMS believes that most of the functionalities described above will be required also at CERN.

Note that the disk buffers of the Tier-0 must be strictly separated from those of the CERN-T1; unexpected access to data on these buffers cannot be permitted or we lose the ability to control the buffer contents and may overload disk and or network components in unexpected ways. The Tier-0 loads must be well defined and controllable subject to agreed policies for the Tier-0 activities.

Specification: S-27	Some portion of the Raw + Reconstructed data will be served from the Tier-1 centre at CERN, but the full second copy of the data will be spread across the regional Tier-1 centres.

Because the primary reconstruction pass has been performed at CERN, the full RECO Dataset is available at CERN. However, if the CERN Tier-1 is sized similarly to a regional Tier-1 it will not have the capacity to serve it all simultaneously to users running at the CERN Tier-1. The existence of the full RECO data at CERN does however give CMS the possibility to dynamically decide which parts of the RECO are available for analysis at CERN in response to possibly changing analysis requirements.

We have not at this time developed the detailed specification for the CERN-T1, we assume it will be similar to those in the regional centres

3.3.6 Summary of Tier-1 Parameters

Specification: S-28	Tier-1 centres shall support a range of collaboration services such as: resource allocation and accounting, support for CMS policies; high- and low-level monitoring; data catalogs; conditions and calibration databases; software installation and environment support; virtual organisations and other such services.

Table 3.2 summarises the parameters of a Tier-1 centre.

Note that the CERN Tier-1 centre can be slightly different, for example it would not need a separate copy of the (first) FEVT data, that has been accounted in the Tier-0.

Each element of a computing system has an efficiency factor which reflects the fact that it cannot run continuously with 100% load and with all disk and tape storage 100% full. The efficiency

Tier 1, both regional and at CERN

The Disk and Tape Storage

	# of events	Ev-size MBytes	Tape/ disk		Tbytes	
			Active	Archive	Disk	
SIM.Out	2.1E+08	2	429		43	Note 1,7
SIM.Rec.	2.1E+08	0.4	86		9	Note 1,7
Raw-sample	2.5E+08	1.5	375		375	Note 2
Calibration	2.5E+07	1.5	38		38	Note 3
Reco	2.5E+08	0.25	63		63	Note 2
Re-proc.Reco	2.5E+08	0.25	125		13	Note 4,7
Re-proc Simu	2.1E+08	0.4	171		17	Note 4,7
General AOD (Data+MC)	3.0E+09	0.05	150		150	Note 5
Revised AOD (Reruns)	3.0E+09	0.05	300		30	Note 4,7
Heavy Ion Sample	8.3E+06	7	58		6	Note 1,7
Analysis Group Space			43		43	Note 6
Total			1837	0	785	

CPU

Data Processing	# of events to Mass Stor.	CPU per event KSI2K/ev.	CPU total kSI2K	
2nd pass Rec-Simulation	2.1E+08	25	510	Note 8
Re-Processing	2.1E+08	25	510	Note 9
Selection	4.6E+08	0.25	672	Note 10
First pass Calibration	2.5E+07	10	25	
Total			1716	

WAN

	Raw Rates	Safety	Headroom	Totals	
	Gb/s	Factor	Factor	Gb/s	
MC Simu/and Reco from Tier	0.1	2	2	0.5	
FEVT/AOD from Tier 0	0.7	2	2	2.2	
AOD Versions from ReReco	1.0	2	2	3.0	Note 11
Total Incoming				**5.7**	
Event Serving to Tier-2s	0.9	2	2	3.5	Note 12
Total Outgoing				**3.5**	

Note 1: 1/NTier1 share of all Sim output
Note 2: 1/(NTier1-1) share of raw data
Note 3: 10% of Rawdata
Note 4: One Re-reconstruction on disk, Nreco/Year on tape
Note 5: Full current AOD on disk
Note 6: Analysis Group Space, 30% of Nphys/NTier1 users with local disk here
Note 7: 10% Samples assumed on Disk
Note 8: Rereconstruction of simulated data made at T1, NRECO times/year (4 months)
Note 9: NRECO re-rereconstructions of local Raw/Reco copy per year (4 months)
Note 10: 10 groups each peform one pass over all local data every 20 days
Note 11: Each full importation to take no more than 1 week
Note 12: Replenish NTier2/NTier1 Analysis samples every two days

Summarized Requirements before efficiency factors are applied

	final	
CPU scheduled	1019	kSI2K
CPU analysis	697	kSI2K
Disk	785	Tbytes
Active tape	1837	Tbytes
Data Serving I/O Rate	800	MB/s

Requirements after application of efficiency factors

			Eff Factors
CPU scheduled	1199	kSI2K	*85.00%*
CPU analysis	929	kSI2K	*75.00%*
Disk	1121	Tbytes	*70.00%*
Active tape	1837	Tbytes	*100.00%*
Data Serving I/O Rate	800	MB/s	

Table 3.2: Parameters of a Tier-1 Centre.

factors shown in the table reflect our experience and are used to convert basic needs into final required capacity.

The disk capacities required for accessing the CMS data are not annual requirements. the active data does not double between year 1 and year 2; however there is of course an increase on the data volume under active analysis. As described in Section 4.3 we look to a process of annual replacement/maintenance to build in the required upgrades in capacity. (Replacing a 3-year old tray of disks by a tray of currently available disks will increase the available volume roughly in line with the expected evolution of requirements and is also required in order to keep operations/maintenance costs under control.)

3.4 Tier-2 Centers

The Tier-2 centres have the following roles in CMS:

- They are each responsible for servicing the analysis requirements of about 20-100 CMS Physicists, depending on size. They will host analysis passes over skimmed data and partial or complete copies of other CMS datasets (Locally resident).
- All Monte Carlo production is carried out at Tier-2 (And Tier-3)
- Quite possibly many detector specific calibration developments and possible operations will be carried out at Tier-2 centres "close" to the relevant detector experts.

3.4.1 Tier-2 Data Processing

Tier-2 responsibilities are to provide analysis computing resources for a geographic region or physics region of interest, as well as the production of the complete simulated event samples for the whole Collaboration. Physicists associated to a Tier-2 centre may have direct login capabilities, will require suitable CMS developer environments, local CMS library installations and the ability to submit jobs locally directly or via Grid interfaces, and ability to submit (Grid) jobs to run at Tier-1 centres and bring results back to their Tier-2 centre. They have local storage facilities for their produced data and by some local mechanism they can cache data products from Tier-1 centres on their Tier-2 centre.

Specification: S-29	Tier-2 centres shall dedicate a significant fraction of their processing capacity to their associated analysis communities.

In a similar vein to the estimate of selection related resources at a Tier-1 centre, we can construct a model of Tier-2 activities.

We make the assumption that the Tier-2 has sufficient local disk-cache to store 5 primary datasets worth of current RECO data. Most re-reconstruction can be run on these samples. This corresponds to 10% of the current RECO data. We do not anticipate that large scale re-reconstruction is being carried out in the Tier-2 centres for analysis purposes; rather specialised studies for calibration or for code development or small sample studies. (This may not really be what they have, more likely they have skimmed the RECO data so they have more primary datasets, without the unwanted events in each one; however the model yields a plausible scale.) We also assume that each Tier-2 has a copy of half of the current (Data and MC) AOD Now

we make the assumption that these physicists/analysis-groups need to flush/replace the local RECO and AOD data sample every three weeks (Compatible with the frequency a group can run major selection passes at the Tier-1 centres). This yields the WAN rate to a Tier-2 to be in the range of 1 Gb/s. (This also of course feeds into the Tier-1 WAN calculation)

Specification: S-30	Tier-2 centres should have WAN connectivity in the range of 1Gb/s or more to satisfy CMS analysis requirements

Specification: S-31	Tier-2 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

This refreshed sample is about 5TB per Tier2 per Day. There are roughly 5 Tier-2 per Tier-1 so on average a Tier-1 is serving 25TB per day to T2 centres. These 25TB are the presumably the result of the selection passes at the T1 centre. Each Tier-1 selection pass has actually processed about 100TB every day, so this is consistent with performing a selection/partial-reconstruction pass and sending about 25% of the volume of data processed to a total of 5 Tier-2 centers. Finally, we presume that these active physicists are also running in a two day period jobs to analyze 1/10 of the events currently in the AOD cache and 1/10 of the events in the reco cache This is quite probably not an actual workflow, but it describes a scale that is about right, and is internally consistent in terms of Tier-1 CPU Power, Tier-1 Data I/O, Tier-1 to Tier-2 WAN, Tier-2 CPU Power and Tier-2 cache size

In addition to these analysis activities, the Tier-2 centres perform the relatively easy to estimate Monte Carlo Simulation and reconstruction activity.

Specification: S-32	Tier-2 centres should provide processing capacity for the production of standard CMS Monte Carlo samples ($\sim 10^9$ events/year summed over all centres), including full detector simulation and the first pass reconstruction.

Finally, as noted in the Heavy Ion discussion, the Heavy Ion event reconstruction may also be an activity that can be efficiently performed at Tier-2 centres.

Specification: S-33	Some Tier-2 centres will provide processing power to allow the Heavy Ion reconstruction to be completed, or extended compared to that available at the Tier-0

3.4.2 Tier-2 facilities at CERN

There will clearly be a significant physicist community operating at CERN, local-staff and visitors, Thus we anticipate a need for Tier-2 capacity also at CERN

Specification: S-34	CMS requires Tier-2 functionality at CERN

We have not yet performed a detailed analysis of this capacity requirement, but any rule-of-thumb would lead to a requirement of a CERN Tier-2 capacity in the region of 2-3 canonical Tier-2 centres. The CERN Tier-2 can act in concert with the CERN Tier-1 to allow a very important analysis activity at CERN.

3.4.3 The Tier-2 Data Storage and Buffering

The Tier-2 centres need enough storage to store the simulated events until they can be safely archived at Tier-1 centres. The Tier-2 centres also need to have enough space to cache and serve the data needed for local analysis. Space must be allocated to locally active Analysis G—roups, or in some cases individual users. The required space and I/O performance of the disk is defined by the number of analysis processing resources available at the site and the networking available to flush the data serving cache.

Specification: S-35	Tier-2 centres are responsible for guaranteeing the transfer of the MC samples they produce to a Tier-1 which takes over custodial responsibility for the data.

Output data rate to archival Tier-1 is fairly modest. The export buffering space is also modest. Even substantial safety factors yield export buffer requirement of no more than 10TB and a maximum network requirement of a few hundred Mb/s.

Specification: S-36	Tier-2 computing centres have no custodial responsibility for any data.

Tier-2 computing centres are unlikely to have tape based mass storage systems.

We have assumed that a Tier-2 center has 5 of the Primary Dataset RECO's. This can actually be 5 complete Primary Dataset RECO's, or perhaps 10% skims of all 50 Primary Dataset RECOs, or 1 complete Primary Dataset FEVT, etc. The local disk requirements are in the range of 30-50TB and thus not considered to be problematic. The more stringent requirement is the refresh rate of this cache. We have assumed this to take place every week; one can imagine scenarios with larger local caches and reduced refresh rates.

3.4.4 Summary of Tier-2 Parameters

Table 3.3 summarises the parameters of a Tier-2 centre.

Each element of a computing system has an efficiency factor which reflects the fact that it cannot run continuously with 100% load and with all disk and tape storage 100% full. The efficiency factors shown in the table reflect real experience and are used to convert basic needs into final required capacity.

As described in section 3.3.6 an maintenance/upgrade path must be considered to keep hardware current and to match evolving requirements. We expect that in many cases Tier-2 centres may not have a steady upgrade path, such as we might expect in the Tier-1 centres, but will actually be upgraded by infrequent and discrete funding requests. Our model will however assume that, integrated over the Tier-2 centres, a steady upgrade evolution can be used at this stage of planning.

Tier2 Centers

Tape and Disk

	# of events	Ev-size MBytes	Tape/ disk Tbytes			
			Active	Archive	Disk	
Local cached reco data (real + simu)	1.5.E+08	0.25			38	Note 1
Local AOD Copies	3.0.E+08	0.05			15	Note 2
Analysis Group Space					40	
Local Privately Simulated Data	3.0.E+07	2			60	
Total			0	0	153	

CPU

Data Processing	# of events	CPU per event KSI2K/ev	CPU total kSI2K	
Simulation	9.0.E+07	45	128	Note 3
Rec-Simulation	9.0.E+07	25	71	Note 4
Heavy Ion reconstruction	2.0.E+06	200	38	Note 8
AOD Analysis	3.0.E+08	0.25	217	Note 5
RECO Analysis (Partial re-reco)	1.5.E+07	2.5	217	Note 7
Total			672	

WAN

	Raw Rates	Safety	Headroom	Totals	
	Gb/s	Factor	Factor	Gb/s	
Event Serving from Tier'1s	0.2	2	2	1.0	Note 6
Total Incoming				**1.0**	
Simu and SimReco data to T1	0.04	2	2	0.1	
Total Outgoing				**0.1**	

Note 1: 5 Primary DataSets of RECO Data
Note 2: 25% the AOD Sample
Note 3: 1/NTier2 Share of all CMS Simulation
Note 4: First reconstruction pass of locally produced Simulated Data
Note 5: Each of 5 groups Analyze AOD data once every 20 days
Note 6: Replenish Local Data every 20 days
Note 7:All locally cached events partially re-reconstructed every twenty days
Note 8: Heavy Ion Reconstruction lasting 4 months

Summarized Requirements before efficiency factors are applied

	final	
CPU scheduled	212	kSI2K
CPU analysis	434	kSI2K
Disk	153	Tbytes

Requirements after application of efficiency factors

			Eff Factors
CPU scheduled	250	kSI2K	*85.00%*
CPU analysis	579	kSI2K	*75.00%*
Disk	218	Tbytes	*70.00%*

Table 3.3: Parameters of a Tier-2 Centre.

3.5 Input Parameters of the Computing Model

For completeness we include in Table 3.4 a list of input parameters that have been used in these calculations

Input Parameters to the CMS Computing Model

Name	Description	Value	Units
L2Rate	pp Rate to Tape	150	Hz
HIRate	Weighted mean HI event Rate	50	Hz
LHCYear	Days of pp Running/year	10000000	sec
HIYear	Seconds of HI Running/year	1.E+06	sec
NRawEvts	Number of pp Raw Events/year	1.5E+09	(derived)
NHIEvts	Number of HI Events/year	5.0E+07	(derived)
RawSize	Raw Data Event Size	1.5	MB
SimSize	Simulated Event Size	2	MB
RecSimSize	Reconstructed Sim Event Size	0.4	MB
RECOSize	Reco Size	0.25	MB
AODSize	AOD Size	0.05	MB
TAGSize	Tag and DPD Size	0.01	MB
HIRawSize	Weighted Mean Heavy Ion Raw Event Size	7	MB
HIRecoSize	Weighted Mean Heavy Ion Reco Size	1	MB
HIAODSize	Weighted Mean Heavy Ion AOD Size	0.2	MB
NPhys	Number of Active Physicists	1000	
NTier1	Number of Tier1 Centers	7	
NTier2	Number of Tier2 Centers	25	
NSimEvt	Number of Simulated Events	1.5.E+09	Evts/Year
FracSimT1	Fraction of NSimEvts done at T1	0%	
NSimPrivate	Number of Private Sim at T2s	8.E+08	Evts/Year
RecCPU	Reconstruction time (Raw)	25	kSI2k.s/ev
SimCPU	Simulation time	45	kSI2k.s/ev
SelCPU	Selection time	0.25	kSI2k.s/ev
AnaCPU	Analysis time	0.25	kSI2k.s/ev
HICPU	Heavy Ion reconstruction time	200	kSI2k.s/ev
NStreamsOFFL	Number of Streams from the off-line	50	
T1RAWCopies	RAW Copies at T1 centers	1	
T0RAWCopy	RAW Copies at CERN	1	
T1RECOCopies	RECO/ESD Copies at T1 Centers	1	
T1AODCopies	AOD Copies	7	
NRECOyear	Reprocessings per year	2	
CalCPU	CPU per Calibration Evt	10	kSI2k.s/ev
CalFrac	Calibration data fraction	10%	
EffSchedCPU	Efficiency factor for Scheduled CPU	85%	
EffAnalCPU	Efficiency factor for Chaotic CPU	75%	
EffDisk	Disk Utilization Efficiency	70%	
EffActiveTape	Active Tape Efficiency	100%	
UserDisk	Group and User Analysis Space	1.0	TB

2007 Cost Estimates			
CHFCPU	CPU	0.55	CHF/SI2k
CHFDisk	Disk	2.18	CHF/GB
CHFTape	Active Tape	0.40	CHF/GB

2007 Performance Estimates			
PerfCPU	Performance per CPU	4	kSI2k
NCPU	Number of CPUs per Box	2	
PerfDisk	GB per Disk	900	GB

Table 3.4: Input Parameters for the computing resource calculations.

3.6 Estimates of additional computing requirements in out-years (2008-10)

Storage media (tape) must be added for each year of LHC operation; not only will there be new data each year but also further re-processings of previous years data. Likewise, data serving/staging disk space must keep track of the new data rate to ensure that a reasonable balance between staging space and total stored volume can be maintained. This spending is somewhat like an Operations component in the spending profile.

With the increase in LHC Luminosity processing times will increase due to the presence of more pileup overlapping the signal events. This can be identified as an upgrade component.

The cost of maintaining "old" hardware and the rapid advances in performance typically encourages steady replacement of many components with time scales of 2-3 years, and this can be identified as a maintenance component.

3.7 Outstanding Issues

As a closing remark it should again be stressed that many issues need further study and that this will indeed be done for the Computing TDR.

In addition to the obviously higher level of detail required for the TDR (and the need to also include software and middleware needs) we note below a number of topics which were identified, in the course of preparing this document, that are especially needful of further consideration. An incomplete list in no particular order:

- Partitioning of Tier-0 and CERN Tier-1 and CERN Tier-2 systems to, on the one hand, ensure Tier-1/2 operations cannot interfere with the Tier-0 and, on the other hand, to make optimal and flexible use of hardware resources.
- Scenarios are needed for physics group and end user analysis. The simplistic estimates of this document need to be replaced with more detailed scenarios/use-cases. What event formats are accessed (RAW, RECO, AOD...)? How many people? How much data each time? How often? How much new data is created? And so on.
- What is the boundary between skim-production and event directories? What are the quantitative issues and trade-offs associated to making deep copies and shallow copies of sub-samples of events?
- What are the data volumes and processing requirements associated to the conditions data and the calibration constants determined offline? What are the associated database requirements for the use of these data?
- A systematic risk analysis is required to: identify potential risks, evaluate their probability and impact, to prioritise their seriousness, and to develop alternative plans and risk avoidance and mitigation actions.,
- Develop a realistic understanding of the issues associated to deletion (or flagging as deletable) of old data (not RAW but derived data such as RECO and AOD). What are the potential savings?

Chapter 4

Summary and Costs

4.1 Overview

To give a cost reference we use the estimates from the Pasta 3 computing cost and performance analysis [14]. We calculate the costs of the required computing on the assumption that the computing required for the 2008 LHC run were purchased entirely in 2007 We do not include the cost of Wide-Area Networking; this is often covered by quite different funding and in any case varies wildly from country to country. We do not include the cost of Local Area Networks. Neither do we include the costs of local computing manpower to run the centres; neither for the generic system installation/management nor for any CMS specific activities (These will be addressed in other documents).

The model of purchasing all computing in 2007 is clearly unrealistic. Experience indicates that much more than a factor of two increase per year can be difficult to manage as new problems are met. However this need to front-load purchasing is counter-balanced by the Moore's law decrease in costs (or increase in capacity that can be purchased for given cost) that favours late purchasing of bulk computing power. We leave these decisions for the computing centres to make based on their perhaps different expertise levels and ability to respond to such issues.

4.2 Costs for 1st year of LHC Running

Table 4.1 summarises the computing requirements and financial estimates for the CMS Computing Model.

As noted above, purchasing all this computing in 2007 is not practicable; some must be in place earlier both for computer centre ramp-up reasons and to service the LHC running in 2007 and for cosmic and the other detector operation. Some storage media costs are best born "just-in-time" rather than a year or even six-months in advance. For the sake of this paper we assume that these effects approximately cancel to yield the above total costs for the 2008 run while actually expecting a spending profile that would share these costs in 2006, 7 and 8.

4.3 Cost Evolution after LHC Startup

We identify two quite different components of out-year costs for LHC computing.

	All (MCHF)	CPU	Disk	Tape	Per Tier	Per Center
T0	4.9	51%	18%	31%	8%	8%
T1	30.4	27%	56%	17%	52%	7%
T2	23.3	49%	51%	0%	40%	2%
Sum MCHF	58.6	22.1	29.9	6.7		
		38%	51%	11%		

Per Center	MSI2k	Disk PB	Tape PB	MCHF	€M	$M
T0	4.6	0.4	3.8	4.9	3.3	4.1
T1	2.1	1.1	1.8	4.3	2.9	3.6
T2	0.8	0.2	0.0	0.9	0.6	0.8

Per Tier	MSI2k	Disk PB	Tape PB	MCHF	€M	$M
T0	4.6	0.4	3.8	4.9	3.3	4.1
T1	14.9	7.8	12.9	30.4	20.3	25.4
T2	20.7	5.5	0.0	23.3	15.5	19.4
Total	40.2	13.7	16.6	58.6	39.1	48.9

Table 4.1: Summary of computing requirements and cost estimates for CMS computing for the first full year of LHC operation (assumed to be 2008).

- Operations, or consumables, costs; that is the annual expenditure on tape and the such like.
- Maintenance and Upgrade costs; that is those associated with hardware replacement to take account of hardware economic lifetime and upgrades in the LHC Luminosity and/or Physics reach of the experiments

Operations costs in the Tier-0 and Tier-1 represent a significant fraction of out-year costs. We see no way to mitigate this, the LHC runs, data must be stored and served.

There are (at least) two ways to treat the Maintenance and Upgrade costs. Using one approach, the annual maintenance spending profits from Moore's law by replacing fixed amount of resources for less and less money. However, in this case an explicit upgrade scenario must take account of Luminosity and other similar effects. Alternately, the maintenance costs can be fixed in annual CHF at one purchases more replacement capacity for a given expenditure each year. We find this second scenario more attractive, the upgrade comes automatically by virtue of Moore's law - we have confirmed that for example investing each year at about 25% of the initial costs (This is the fractional figure used in the table below) satisfies both the hardware replacement requirements and achieves the upgrade path to reach high-luminosity running in 2010

Table 4.2 summarises the assumptions on computing cost evolution that we have used here.).

Cost Evolutions extracted from PASTA3 and fom Bernd Panzer, LHCC presentation

	2006	2007	2008	2009	2010
CHF/SI2K	0.89	0.55	0.37	0.24	0.18
CHF/GB (Disk)	3.49	2.18	1.36	0.85	0.53
CHF/GB (Tape)	0.70	0.40	0.40	0.40	0.40

Table 4.2: Computing Cost Evolution assumed in this document

Table 4.3 summarises the proposed funding profile for post 2008 and beyond (For the LHC operation in the following year in each case).

Annual Expenditures

Per Center	2008	2009	2010
T0 Ops	1.5	1.5	1.5
T0 Maint.	0.9	0.9	0.9
T1 Ops.	0.7	0.7	0.7
T1 Maint.	0.9	0.9	0.9
T2 Maint.	0.2	0.2	0.2
MCHF	20	20	20

Table 4.3: Proposed funding profile for the years following the first major LHC run (2008-2010)

Glossary

AFS	Andrew File System
ANSI	American National Standards Institute
AOD	Analysis Object Data - a compact event format for physics analysis
ATM	Asynchronous Transfer Mode
CAD	Computer-Aided Design
CASE	Computer-Aided Software Engineering
CD	Compact Disk
CDF	Collider Detector Facility experiment at the FNAL Tevatron
CDR	Central Data Recording
CLHEP	Class Library for HEP
CMKIN	CMS Kinematics Package (legacy Fortran)
CMS	Compact Muon Solenoid
CMSIM	CMS Simulation Package (legacy Fortran)
CODEC	Compression/Decompression
CPU	Central Processing Unit
COBRA	Coherent Object-oriented Base for Reconstruction, Analysis and simulation (Framework)
CORBA	Common Object Request Broker Architecture
CVS	Concurrent Versions System
D0	D0 experiment at the FNAL Tevatron
DAQ	Data Acquisition
DBMS	Database Management System
DCS	Detector Control System
DDL	Data Description Language
DFS	Distributed File System
Digi	Digitisation (of detector hit)
DLT	Digital Linear Tape
DST	Data Summary Tape - a compact event format
DVD	Digital Versatile Disk
ECAL	Electromagnetic Calorimeter
EDMS	Engineering Database Management System
EGEE	Enabling Grids for e-science in Europe (a Grid project)
EFU	Event Filter Unit
EPICS	Experimental Physics Industrial Control System
ESNET	Energy Science Network (in the USA)
EVM	Event Manager
Express Line	Online stream for events requiring high priority and low latency offline processing
FDDI	Fibre Distributed Data Interface
FE	Front-End
FED	Front-End Driver
FEVT	Event format comprising the union of RAW and RECO data
FNAL	Fermi National Accelerator Laboratory, USA
GEANT4	Simulation Framework and Toolkit
GIPS	Giga (10^9) Instructions per Second
Gb	Gigabit (10^9 bits)
GB	Gigabyte (10^9 bytes)
GIF	Graphics Interchange Format
GL	Graphics Language (low-level 3D rendering software)
GRID	Infrastructure for Distributed Computing
GUI	Graphical User Interface
HCAL	Hadronic Calorimeter
HEP	High Energy Physics
HEPEVT	HEP Event (generated event format)
HEPiX	HEP Unix environment
HI	Heavy Ion(s)
HLT	Higher Level Trigger (Software)
HTML	Hypertext Mark-up Language

IGUANA	Interactive Graphics for User ANAlysis - used for the CMS Event Display Package
I/O	Input/Output
IP	Internet Protocol
IPC	Interprocess Communication
ISDN	Integrated Services Digital Network
IT	Information Technology
kb	kilobit (10^3 bits)
kB	kilobytes (10^3 bytes)
L1	Level 1 hardware-based trigger
LAN	Local Area Network
LCG	LHC Computing Grid (a common computing project)
LEP	Large Electron Positron Collider
LHC	Large Hadron Collider
LHCC	LHC (review) Committee
Mb	Megabit (10^6 bits)
MC	Monte Carlo simulation program/technique
MBONE	Multicast Backbone
MB	Megabyte (10^6 bytes)
MIPS	Mega (10^6) Instructions per Second
MS	Microsoft (Corporation)
NQS	Network Queueing System
OO	Object Oriented
ODBMS	Object Database Management System
Online Stream	Grouping of events (Primary Datasets) to simplify online data management
OQL	Object Query Language
ORB	Object Request Broker
ORCA	CMS Reconstruction Program
OS	Operating System
OSCAR	CMS GEANT4 Simulation Program
OSF	Open Software Foundation
PAW	Physics Analysis Workstation (legacy interactive analysis application
Pb	Petabit (10^{15} bits)
PB	Petabyte (10^{15} bytes)
POOL	Persistency software from LCG
Primary Dataset	Grouping of events according to physics (trigger) criteria
QA	Quality Assurance
QC	Quality Control
RAID	Redundant Arrays of Independent Disks
RC	Regional Centre / Readout Crate
RAW	Event format from the online containing full detector and trigger data
RECO	Event format for reconstructed objects such as tracks, vertices, jets, etc.
RecHit	Reconstructed hit in a detector element
RHIC	Relativistic Heavy Ion Collider (at Brookhaven, USA)
RISC	Reduced Instruction Set Computer
R/W	Read/Write
SA/SD	Structured Analysis/Structured Design
SFI	Switch Farm Interface
Skim	Subset of events selected from a larger set
SMP	Symmetric Multiprocessor
SNMP	Simple Network Management Protocol
SQA	Software Quality Assurance
SQL	Structured Query Language
STL	Standard Template Library
TAG	Event index information such as run/event number, trigger bits, etc.
Tb	Terabit (10^{12} bits)
TB	Terabyte (10^{12} bytes)
TCP	Transmission Control Protocol
TDR	Technical Design Report
TIPS	Tera (10^{12}) Instructions per Second
VCAL	Very Forward Calorimeter
WAN	Wide Area Network
WWW	World Wide Web
WYSIWYG	What You See Is What You Get (type of GUI)

Appendix A

Further Reading

Technical issues are not addressed in depth in this document (that is the subject of the Computing TDR). In the meantime, the following references may prove to be of interest:

Physics Software:

- OSCAR: An Object-Oriented Simulation Program for CMS [4]
- FAMOS: a FAst MOnte Carlo Simulation for CMS [7]
- Mantis: a Framework and Toolkit for Geant4-Based Simulation in CMS [19]
- CMKIN v3 User's Guide [20]
- ORCA: reconstruction program [21, 22, 23, 24]
- Magnetic field software implementation in CMS [25]
- High Level Trigger software for the CMS experiment [26]
- Monitoring CMS Tracker construction and data quality using a grid/web service based on a visualization too [27]
- Expected Data Rates from the Silicon Strip Tracker [2]

DC04 Data Challenge (computing aspects):

- Distributed Computing Grid Experiences in CMS DC04 [28]
- Role of Tier-0, Tier-1 and Tier-2 Regional Centers during CMS DC04 [29]
- Tier-1 and Tier-2 Real-time Analysis experience in CMS DC04 [30]
- Production Management Software for the CMS Data Challenge [31]
- Planning for the 5% Data Challenge, DC04 [32]
- CMS Distributed Data Analysis Challenges [33]
- Distributed File system Evaluation and Deployment at the US-CMS Tier-1 Center [34]
- Software agents in data and workload management [35]

DC04 Data Challenge (analysis experiences):

- Using the reconstruction software, ORCA, in the CMS data challenge 2004 [36]

- Use of Grid Tools to Support CMS Distributed Analysis [37]
- The CMS User Analysis Farm at Fermilab [38]
- Grid Enabled Analysis for CMS: prototype, status and results [39]
- GROSS: an end user tool for carrying out batch analysis of CMS data on the LCG-2 Grid. [40]
- Clarens Web services [41]

Production systems:

- RefDB (the Reference Database for CMS Monte Carlo Production) [42, 43]
- McRunjob (a High Energy Physics Workflow Planner for Grid Production Processing) [44]
- BOSS (an Object Based System for Batch Job Submission and Monitoring) [45]
- Virtual Data in CMS Production [46]
- Combined Analysis of GRIDICE and BOSS Information Recorded During CMS-LCG0 Production [47]
- Running CMS Software on GRID Testbeds [48]
- Resource Monitoring Tool for CMS production [49]
- The Spring 2002 DAQ TDR Production [50]
- CMS Test of the European DataGrid Testbed [51]
- Use of Condor and GLOW for CMS Simulation Production [52]
- Study and Prototype Implementation of a Distributed System [53]

Core Applications Software:

- Report of the CMS Data Management RTAG [1]
- Status and Perspectives of Detector Databases in the CMS Experiment at the LHC [54]
- Modeling a Hierarchical Data Registry with Relational Databases in a Distributed Environment [55]
- Detector Geometry Database [56]
- Migration of the XML Detector Description Data and Schema to a Relational Database [57]
- De-serializing Object Data while Schemas Evolve [58]
- Evaluation of Oracle9i C++ Call Interface [59]
- 3D Graphics Under Linux [60]
- IGUANA Plan For 2002 [61]
- Evaluation Of Oracle9i To Manage CMS Event Store: Oracle Architecture To Store Petabyte Of Data (PART ONE) [62]
- Composite Framework for CMS User Applications [63]
- Mantis: the Geant4-based simulation specialization of the CMS COBRA framework [5]
- CMS Detector Description: New Developments [64]
- A database perspective on CMS data [65]
- ROOT - An Object Oriented Data Analysis Framework [8]

Software Environment:

- Use Cases and Requirements for Software Installation in Grid and End-User Desktop Environments [66]
- OVAL: The CMS Testing Robot [67]
- Installation/Usage Notes For Oprofile [68]
- CMS Software Quality [69]
- Evaluation Of The CMT And SCRAM Software Configuration, Build And Release Management Tools [70]
- CMS Software Installation [71]
- Parallel compilation of CMS software [72]
- PRS Software Quality Policy [73]
- Software Metrics Report Of CMS Reconstruction Software [74]

General Organisation and Planning:

- CMS Computing and Software Tasks and Manpower for 2003-2007 [75]
- Computing And Core Software (CCS) Schedule And Milestones: Version 33 [76]
- Planning for CTDR [77]
- Proposed Scope And Organization Of CMS-CPT. Computing And Core Software, Physics Reconstruction and Selection, TriDAS (Online Computing) [78]
- CMS Grid Implementation Plan - 2002 [79]
- Plans for the Integration of Grid Tools in the CMS Computing Environment [80]
- Scope and Organization of CMS-CPT [81]

Computing at the Tevatron

- Job and Information Management Deployment for the CDF Experiment [82]
- Monitoring the CDF distributed computing farms [83]
- Testing the CDF Distributed Computing Framework [84]
- Tools for GRID deployment of CDF offline and SAM data handling systems for Summer 2004 computing [85]
- Globally Distributed User Analysis Computing at CDF [86]
- Deployment of SAM for the CDF Experiment [87]
- The Condor based CDF CAF [88]
- Performance of an operating High Energy Physics Data grid, D0SAR-grid [89]
- D0 data processing within EDG/LCG [90]
- Experience using grid tools for CDF physics [91]

References

[1] R.Harris *et al.* "Report of the CMS Data Management RTAG". *CMS IN*, 2004-038, 2004.

[2] I. Tomalin *et al.* "Expected Data Rates from the Silicon Strip Tracker". *CMS NOTE*, 2002-047, 2002.

[3] D. Duellmann *et al.* "POOL development status and plans". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[4] M. Stavrianakou *et al.* "An Object-Oriented Simulation Program for CMS". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[5] M. Stavrianakou *et al.* "Mantis: the Geant4-based simulation specialization of the CMS COBRA framework". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[6] S.Agostinelli *et al.* "Geant4: a simulation toolkit". *NIM*, A 506:250–303, 2003.

[7] F.Beaudette. "FAMOS, a FAst MOnte-Carlo Simulation for CMS". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[8] R. Brun and F.Rademakers. "ROOT - An Object Oriented Data Analysis Framework". In *AIHENP'96 Workshop*, volume Phys. Res. A 389, pages 81–86, Lausanne, Switzerland, September, 1996 1997.

[9] I.Bird *et al.* "Operating the LCG and EGEE Production Grids for HEP". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[10] I.Foster *et al.* "The Grid2003 Production Grid: Principles and Practice". In *Proceedings of the 13^{th} IEEE Intl. Symposium on High Performance Distributed Computing, 2004*, Honolulu, Hawaii, June 4^{th}-6^{th} 2004 2004. To be published.

[11] The EGEE Project. "EGEE Middleware Architecture AND PLANNING (RELEASE 1)". *EGEE-*, DJRA1.1-476451-v1.0, 2004.

[12] P.Eerola *et al.* "Science on NorduGrid". In P.Neittaanmaki *et al.* editors, editor, *Proceedings of The European Congress on Computational Methods in Applied Sciences and Engeneering (ECCOMAS), 2004*, Jyvaskyla, Finland, July 24^{th}-28^{th} 2004 2004.

[13] R.Pordes *et al.* "The Open Science Grid". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[14] "PASTA 3 Final Reports". http://lcg.web.cern.ch/LCG/PEB/PASTAIII/pasta2002Report.htm.

[15] B.Panzer-Steindel. "Price Extraplolation parameters for the CERN LCG Phase II Computing Farm". *CERN-LCG-PEB*, 2004-20, 2004.

[16] B.Panzer-Steindel. "Sizing and Costing of the CERN T0 center". *CERN-LCG-PEB*, 2004-21, 2004.

[17] M.Aderholz (*et al.). "Models of Networked Analysis at Regional Centres for LHC Experiments (MONARC) - PHASE 2 REPORT".* CERN/LCB*, 2000-001, 2000.*

[18] S.Bethke (et al.). "Report of the steering group of the LHC computing review". *CERN/LHCC*, 2001-004, 2001.

[19] M.Stavrianakou *et al.* "Mantis: a Framework and Toolkit for Geant4-Based Simulation in CMS". *CMS NOTE*, 2002-032, 2002.

[20] V.Karimaki *et al.* "CMKIN v3 User's Guide". *CMS IN*, 2004-016, 2004.

[21] N.Neumeister *et al.* "Muon Reconstruction Software in CMS". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[22] T.Speer *et al.* "Kinematic fit package for CMS". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[23] E.Chabanat *et al.* "Deterministic Annealing for Vertex Finding at CMS ". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[24] S.Cucciarelli *et al.* "Pixel Reconstruction in the CMS High-Level Trigger System ". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[25] T.Todorov *et al.* "Magnetic field software implementation in CMS ". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[26] O.van der Aa *et al.* "High Level Trigger software for the CMS experiment". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[27] G.Zito *et al.* "Monitoring CMS Tracker construction and data quality using a grid/web service based on a visualization tool ". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[28] A.Fanfani *et al.* "Distributed Computing Grid Experiences in CMS DC04". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[29] D. Bonacorsi *et al.* "Role of Tier-0, Tier-1 and Tier-2 Regional Centers during CMS DC04 ". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[30] N.De Filippis *et al.* "Tier-1 and Tier-2 Real-time Analysis experience in CMS DC04". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[31] J.Andreeva *et al.* "Production Management Software for the CMS Data Challenge ". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[32] D. Stickland. "Planning for the 5% Data Challenge, DC04". *CMS IN*, 2002-054, 2002.

[33] C.Grandi *et al.* "CMS Distributed Data Analysis Challenges". In *Proceedings of the ACAT'03 Conference*, volume A 534, pages 87–93, Tsukuba, Japan, December 1^{st}-5^{nd}, 2003 2004.

[34] M.Ernst *et al.* "Distributed Filesystem Evaluation and Deployment at the US-CMS Tier-1 Center". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[35] T.Barrass *et al.* "Software agents in data and workload management". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[36] S.Wynhoff *et al.* "Using the reconstruction software, ORCA, in the CMS datachallenge 2004". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[37] A.Fanfani *et al.* "Use of Grid Tools to Support CMS Distributed Analysis". In *Proceedings of the IEEE-NSS'04 Conference*, Rome, Italy, October 16^{th}-22^{nd}, 2004 2004. to be published.

[38] I.Fisk *et al.* "The CMS User Analysis Farm at Fermilab". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[39] F.van Lingen *et al.* "Grid Enabled Analysis for CMS: prototype, status and results ". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[40] H.Tallini *et al.* "GROSS: an end user tool for carrying out batch analysis of CMS data on the LCG-2 Grid.". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[41] C.Steenberg *et al.* "The Clarens Grid-enabled Web Services Framework: Services and Implementation ". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[42] V.Lefebure *et al.* "RefDB". *CMS IN*, 2002-044, 2002.

[43] V.Lefebure *et al.* "RefDB: The Reference Database for CMS Monte Carlo Production". In *Proceedings of the CHEP'03 Conference*, La Jolla, California, March 24^{th}-28^{th}, 2003 2003. Published on eConf.

[44] G.Graham *et al.* "McRunjob: A High Energy PhysicsWorkflow Planner for Grid Production Processing". In *Proceedings of the CHEP'03 Conference*, La Jolla, California, March 24^{th}-28^{th}, 2003 2003. Published on eConf.

[45] C.Grandi *et al.* "Object Based System for Batch Job Submission and Monitoring (BOSS)". *CMS NOTE*, 2003-005, 2003.

[46] R.Cavanaugh *et al.* "Virtual Data in CMS Production". In *Proceedings of the CHEP'03 Conference*, La Jolla, California, March 24^{th}-28^{th}, 2003 2003. Published on eConf.

[47] N.De Filippis *et al.* "Combined Analysis of GRIDICE and BOSS Information Recorded During CMS-LCG0 Production". *CMS NOTE*, 2004-028, 2004.

[48] P.Capiluppi *et al.* "Running CMS Software on GRID Testbeds". In *Proceedings of the CHEP'03 Conference*, La Jolla, California, March 24^{th}-28^{th}, 2003 2003. Published on eConf.

[49] A.Osman *et al.* "Resource Monitoring Tool for CMS production". *CMS NOTE*, 2003-013, 2003.

[50] T.Wildish *et al.* "The Spring 2002 DAQ TDR Production". *CMS NOTE*, 2002-034, 2002.

[51] P.Capiluppi *et al.* "CMS Test of the European DataGrid Testbed". *CMS NOTE*, 2003-014, 2003.

[52] S.Dasu *et al.* " Use of Condor and GLOW for CMS Simulation Production". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[53] S.Schmid. "Study and Prototype Implementation of a Distributed System". *CMS NOTE*, 2004-010, 2004.

[54] I.Vorobiev *et al.* "Status and Perspectives of Detector Databases in the CMS Experiment at the LHC". *CMS NOTE*, 2004-026, 2004.

[55] Z.Xie *et al.* "Modeling a Hierarchical Data Registry with Relational Databases in a Distributed Environment". *CMS IN*, 2004-025, 2004.

[56] M.Liendl *et al.* "Detector Geometry Database". *CMS NOTE*, 2004-011, 2004.

[57] A.J.Muhammad *et al.* "Migration of the XML Detector Description Data and Schema to a Relational Database". *CMS NOTE*, 2003-031, 2003.

[58] R.McClatchey*et al.* "Deserializing Object Data while Schemas Evolve". *CMS NOTE*, 2002-029, 2002.

[59] V.Innocente *et al.* "Evaluation of Oracle9i C++ Call Interface". *CMS NOTE*, 2002-012, 2002.

[60] I.Osborne. "3D Graphics Under Linux". *CMS IN*, 2002-041, 2002.

[61] I.Osborne *et al.* "IGUANA Plan For 2002". *CMS IN*, 2002-018, 2002.

[62] S.Iqbal. "Evaluation Of Oracle9i To Manage CMS Event Store: Oracle Architecture To Store Petabyte Of Data (PART ONE)". *CMS IN*, 2002-002, 2002.

[63] I.Osborne *et al.* "Composite Framework for CMS User Applications". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[64] M.Case *et al.* "CMS Detector Description: New Developments". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[65] A.T.M.Aerts *et al.* "A database perspective on CMS data ". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[66] C.Grandi. "Use Cases and Requirements for Software Installation in Grid and End-User Desktop Environments". *CMS IN*, 2004-047, 2004.

[67] D.Chamont *et al.* "OVAL: The CMS Testing Robot". In *Proceedings of the CHEP'03 Conference*, La Jolla, California, March 24^{th}-28^{th}, 2003 2003. Published on eConf.

[68] G.Eulisse *et al.* "Installation/Usage Notes For Oprofile". *CMS IN*, 2002-053, 2002.

[69] L.Taylor. "CMS Software Quality". *CMS IN*, 2002-050, 2002.

[70] I.Osborne *et al.* "Evaluation Of The CMT And SCRAM Software Configuration, Build And Release Management Tools". *CMS IN*, 2002-046, 2002.

[71] K.Rabbertz *et al.* "CMS Software Installation ". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[72] S.Schmid *et al.* "Parallel compilation of CMS software ". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[73] H.P.Wellish*et al.* "PRS Software Quality Policy". *CMS IN*, 2002-037, 2002.

[74] G.R.Thomson. "Software Metrics Report Of CMS Reconstruction Software". *CMS IN*, 2002-033, 2002.

[75] L.Taylor. "CMS Computing and Software Tasks and Manpower for 2003-2007". *CMS IN*, 2003-038, 2003.

[76] D.Stickland L.Taylor. "Computing And Core Software (CCS) Schedule And Milestones: Version 33". *CMS IN*, 2002-039, 2002.

[77] D. Stickland. "Planning for the Computing TDR". *CMS IN*, 2002-059, 2002.

[78] D.Stickland *et al.* "Proposed Scope And Organization Of CMS-CPT. Computing And Core Software, Physics Reconstruction and Selection, TriDAS (Online Computing)". *CMS IN*, 2002-038, 2002.

[79] c.Grandi *et al.* "CMS Grid Implementation Plan - 2002". *CMS NOTE*, 2002-015, 2002.

[80] C.Grandi *et al.* "Plans for the Integration of Grid Tools in the CMS Computing Environment". In *Proceedings of the CHEP'03 Conference*, La Jolla, California, March 24^{th}-28^{th}, 2003 2003. Published on eConf.

[81] P.Sphicas *et al.* "Scope and Organization of CMS-CPT". *CMS IN*, 2002-068, 2002.

[82] M.Burgon-Lyon *et al.* "JIM Deployment for the CDF Experiment". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[83] I.Sfiligoi *et al.* "Monitoring the CDF distributed computing farms". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[84] V.Bartsch *et al.* "Testing the CDF Distributed Computing Framework". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[85] A.Kreimer *et al.* " Tools for GRID deployment of CDF offline and SAM data handling systems for Summer 2004 computing". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[86] A.Sill *et al.* "Globally Distributed User Analysis Computing at CDF". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[87] S.Stonjek *et al.* " Deployment of SAM for the CDF Experiment". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[88] I.Sfiligoi *et al.* "The Condor based CDF CAF". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[89] B.Quinn *et al.* "Performance of an operating High Energy Physics Data grid, D0SARgrid". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[90] T.Harenberg *et al.* "D0 data processing within EDG/LCG". In *Proceedings of the CHEP'04 Conference*, Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004 2004. Published on InDiCo.

[91] G.Garzoglio *et al.* "Experience using grid tools for CDF physics". In *Proceedings of the ACAT'03 Conference*, volume A 534, pages 38–41, Tsukuba, Japan, December 1^{st}-5^{nd}, 2003 2004.