Computing Model

Baseline model of LHCb's distributed computing facilities

Document Version:	0.9
Document Date:	22 March 2000
Document Created:	6 March 2000
Document Status:	Draft
Document Editor:	

Abstract

This document describes the dataflow model for all stages in the processing of real and simulated LHCb events from the production of the raw data to the final physics results. Estimates of the cpu and data storage requirements have been made using measurements from existing software used in the ongoing detector optimisation studies. The estimates are being continuosly revised as the software evolves and our understanding of the issues matures. We also provide a first ideas on how we intend to distribute the processing load between the various computing facilities available to LHCb, both at CERN and at regional computing centres. This model depends very much on the evolution of the computing infrastructure (networks) and on the detailed planning of the various institutes and funding agencies. Many of these plans are still tentative and we therefore expect that the current model will evolve with time.

Status of the document

This is a first draft that has been prepared for the 2000 LHC Computing Review meeting of March 23rd 2000. There has not been time to consult widely in the collaboration before this meeting. A first discussion of the document in an open meeting of the collaboration will take place during the next LHCb Software Week (April 5-7). The document will then be revised to take into account all feedback received.

1 Logical Dataflow and Workflow Model

There are several phases in the processing of event data and here we describe the terminology used to define each processing step and the data sets that are produced. The various stages normally follow eachother in a sequential manner, but some stages may be repeated a number of times. This workflow will be described as far as it can be anticipated. The terminology follows that in common usage by all four LHC experiments and has been documented in various reports - see for example [1].

Raw data production is made in the Event Filter farm of the online system, for real data, and in the compute facilities of the offline system, for simulated events (Figure 1).In the experiment itself, the first step is to collect data, triggering on events of interest. This procedure involves processing data coming from the detectors using sophisticated and highly optimised algorithms (L2/L3 Triggers). The trigger software incorporates pieces that apply calibration corrections, that reconstruct physical properties of the particles and that apply selection based on physics criteria. The results of this step are the **RAW Data** and **RAW Tag** objects that are a classification of the events determined by the trigger code. The Trigger step can never be repeated; data not selected for permanent storage by the trigger are lost forever. Small samples of rejected events are kept for monitoring and efficiency studies.

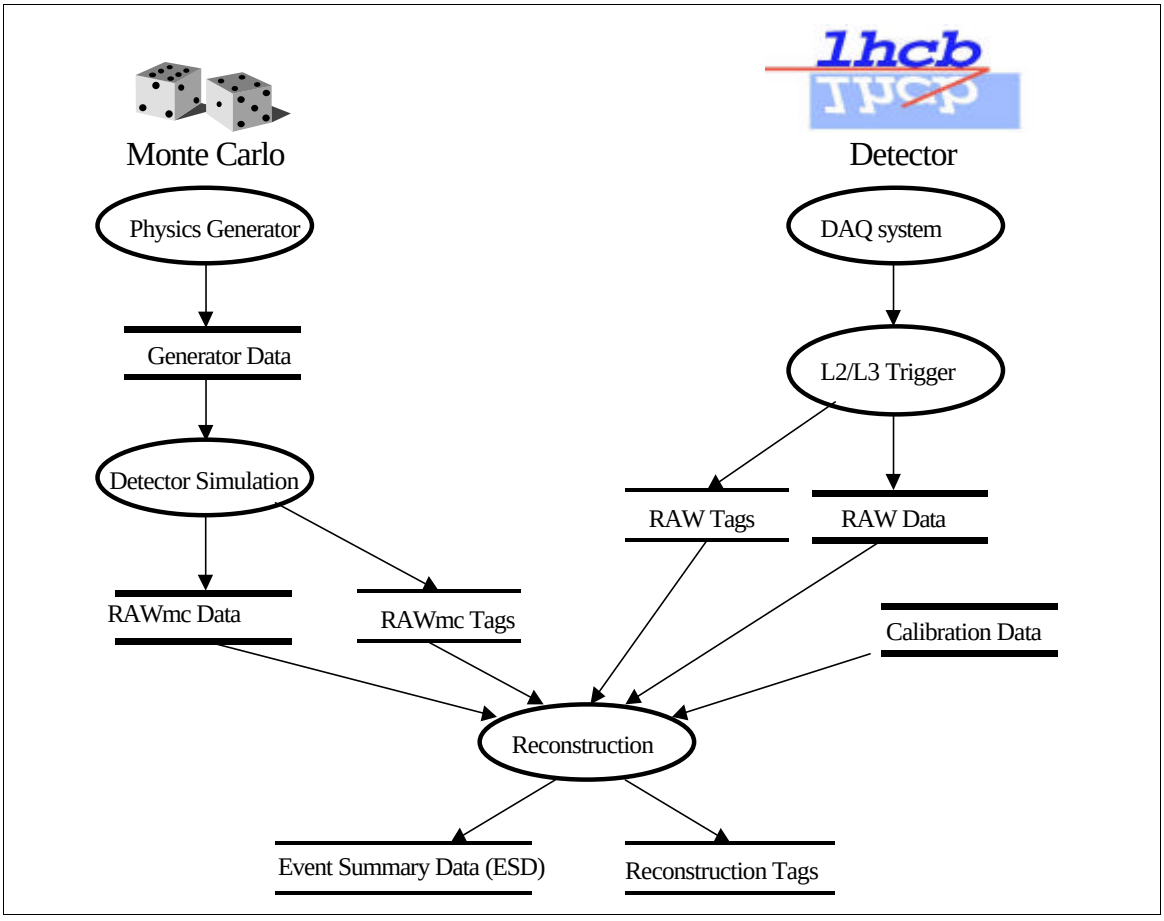

Figure 1 LHCb Computing Logical Dataflow Model

Simulation studies result in the generation of **RAWmc Data** sets(Figure 1). These data sets contain simulated hit information and extra 'truth' information. The truth information is used to record the physics history of the event and the relationships of hits to incident particles. This history is carried through to subsequent steps in the processing (AOD) so that it can be used during analysis. Simulated raw data sets are therfore larger than real raw data. Otherwise the format of the simulated raw data is the same as for real data and they are processed using the same reconstruction software.

These raw data must then be reconstructed such that raw physical quantities such as energy in calorimeter cells and hits are assigned to tracks and particles. Event reconstruction results in the generation of new data, namely the first version of **Event Summary Data (ESD)** and **Reconstruction Tag** objects(Figure 1). The pattern recognition algorithms in the reconstruction program make use of calibration and alignment constants to correct for temporal changes in the response of the detector and its electronics, and in its movement. Each subdetector has associated with it a special set of procedures for calibrating its response to environmental conditions (pressure, temperature etc), for calibrating the response of its readout electronics, for measuring displacements from its nominal position in the experimental hall i.e. alignment etc. (Figure 2).

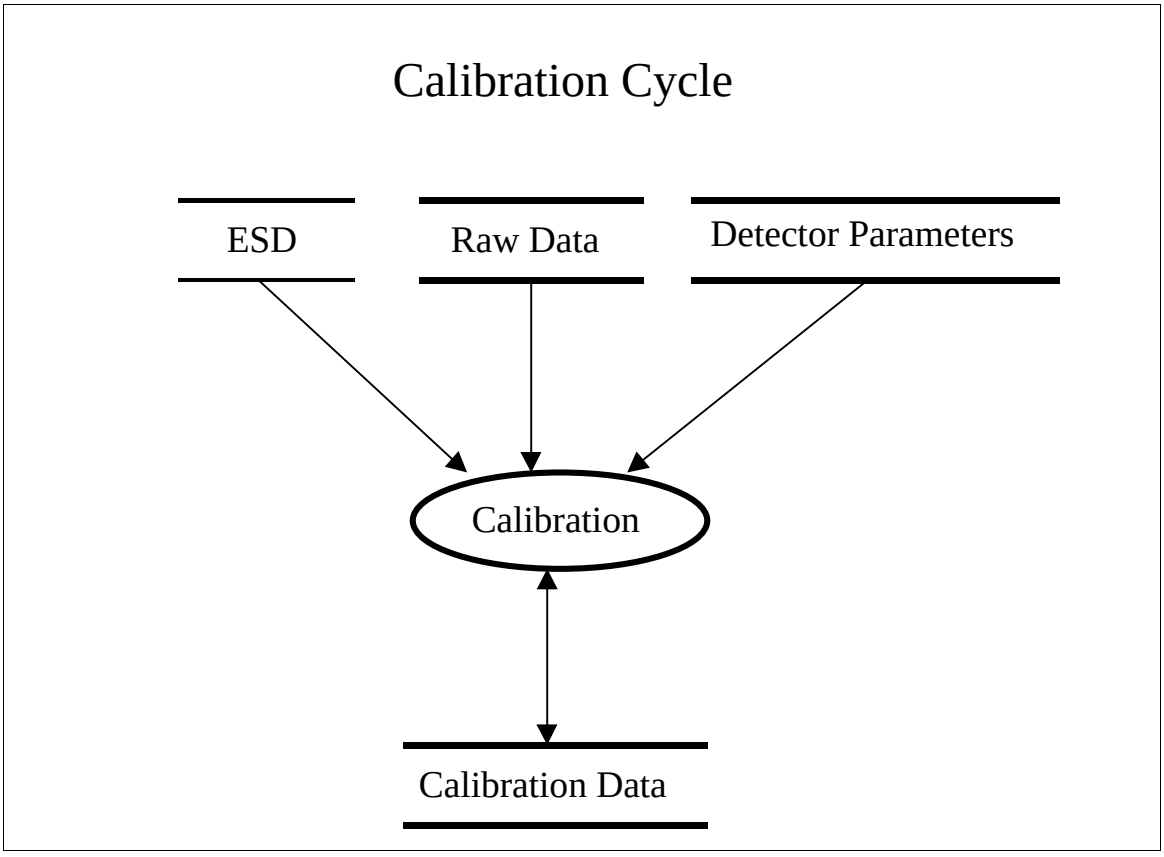

Figure 2 LHCb Calibration Cycle

In practice, the reconstruction step has to be repeated a number of times to accommodate improvements in the algorithms and also to make use of improved measurements of the calibration and alignment of the detector to regenerate new improved ESD information. As the first data are collected we expect the number of reprocessings to be numerous as during this phase we will be learning how the detector behaves. We imagine that the extra load this generates may be compensated by a shorter duty cycle of

the LHC machine, and all periods without datataking the computing resources can be used for reprocessing of data already taken. With time we assume that all data taken in each year will normally need to be re-processed two or three times every year and possibly one final re-processing after a particular phase of operation of the machine.

Following event reconstruction the ESD data (tracks, energy clusters, particle id) are analysed to determine the momentum four vectors corresponding to the measured particle tracks, to locate vertices, to reconstruct invariant masses and to run tagging algorithms to identify candidates for composite particles (e.g. J/Ψ, π^0,..). Since these algorithms are common to many different physics analyses they are run in production as a first step in the analysis as the data are collected.This step makes use of the Reconstruction Tag information to optimise the selection procedure. However the algorithms can be quite time consuming as they have to deal with combinatorics and more than one algorithm could have to be run on any single event. It is planned to run one production job executing all the physics tag algoritms for the experiment and this will be repeated several times(3-4 times per year) as the selection cuts and analysis algorithms evolve. The results of the analysis are stored as **Analysis Object Data (AOD)** and the **Analysis Tag** information will be stored in the **Tag database** (Figure 3).

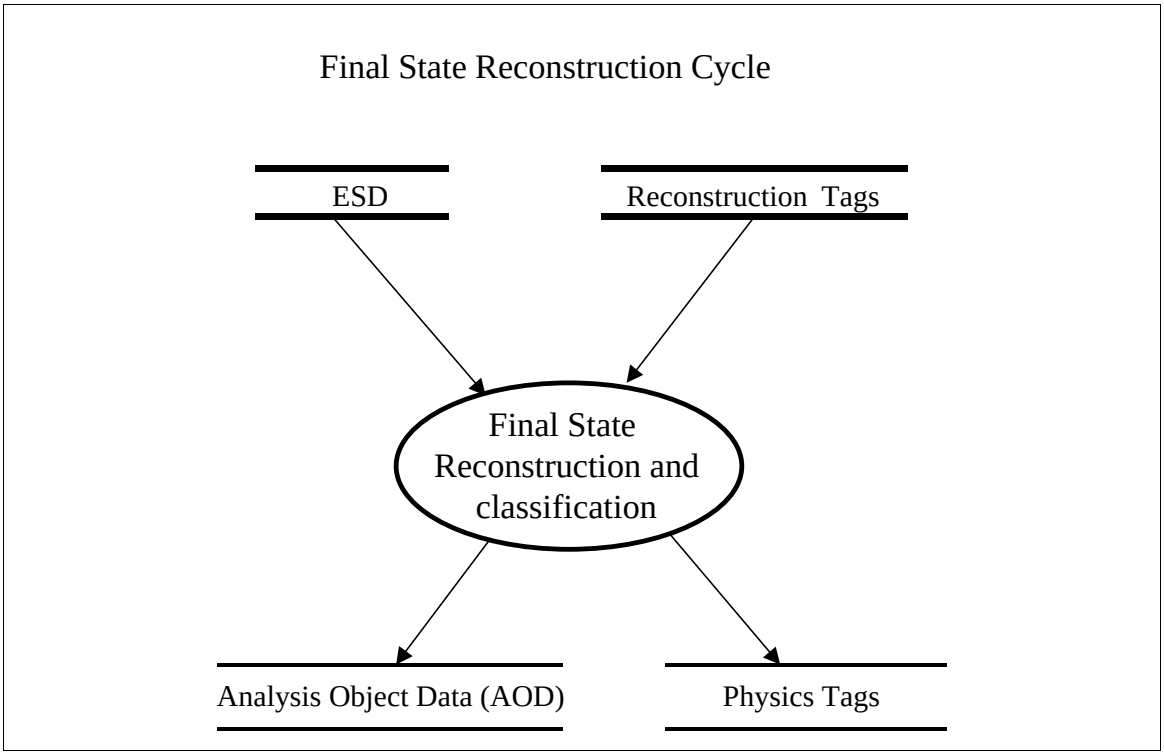

Figure 3 LHCb Decay Reconstruction Cycle

Finally physicists will run their **Physics Analysis** jobs (Figure 4). They process AOD corresponding to events with interesting physics analysis tags and run algorithms to reconstruct the B decay channel being studied. Since the number of channels to be studied is very large, we can assume that each physicist is performing a separate analysis on a specific channel. This analysis step generates private data (e.g. **Ntuples**), which is interrogated interactively to produce the final physics results. They may also run jobs that require access to ESD, but this typically involves small event samples. In addition raw data and calibration data may be accessed in order to study individual events in detail, for example with the event display, but this will involve processing only very small event samples. Our goal would be to

give each physicist transparent access to all AOD data, whereas access to raw and ESD data will be on a 'need to know' basis. An implied goal is therefore to make the AOD data as complete as possible in order to minimise access to ESD data in the analysis step.

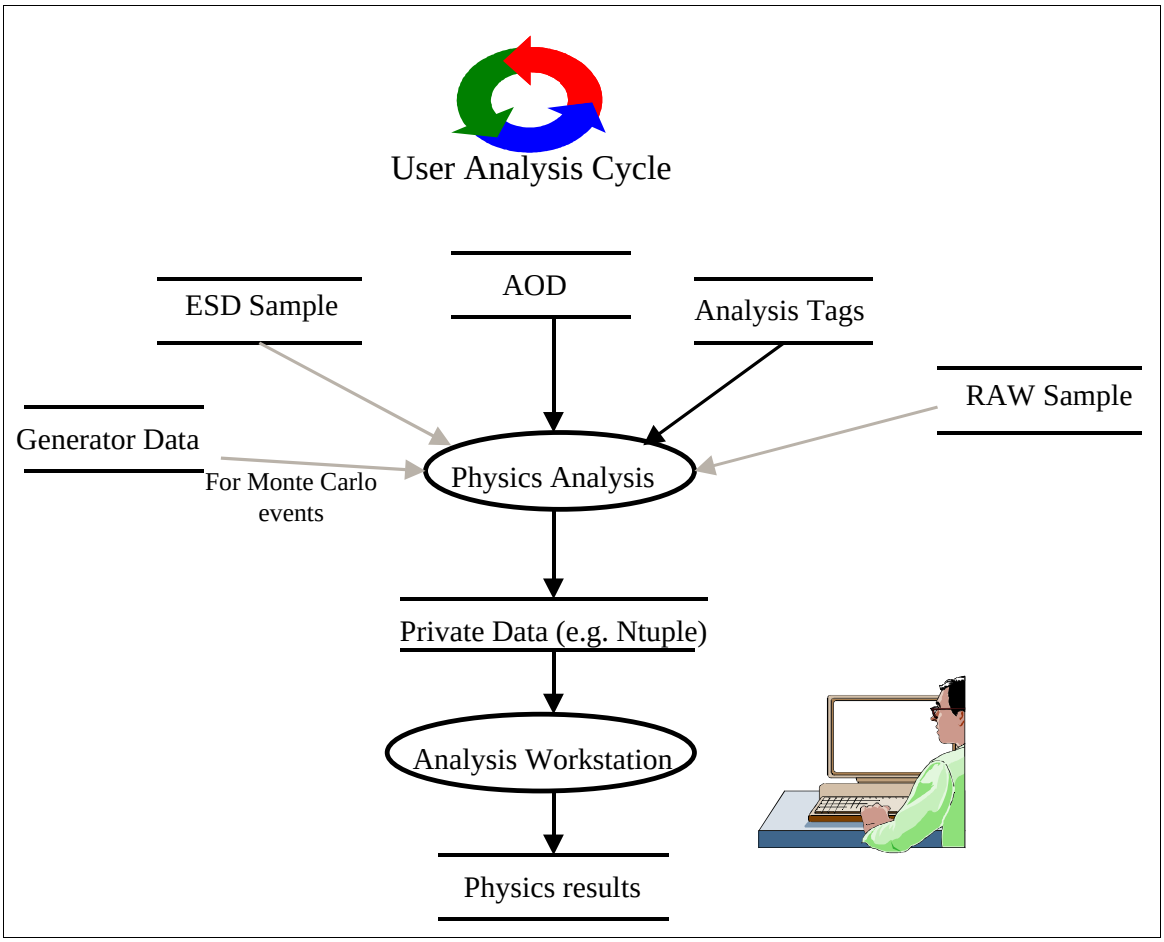

Figure 4 LHCb physicist analysis Cycle

2 Data processing and storage requirements

The frequency of each of the data processing operations, the volume of input and output data, and the amount of computing hardware resources needed to accomplish the tasks must be quantified in order to specify the computing model precisely. A detailed breakdown of the processing and data requirements has been made in terms of individual subdetectors for each processing stage. The spreadsheets containing this information are included in Appendix A. In the following we comment in some detail on the way these numbers were derived in order to give some insight into how precisely they are known.

2.1 Online requirements

A summary of the processing requirements, in terms of data volume, rates and CPU needs are summarised in Table 1.

The high level triggers receive data corresponding to the full event after each positive Level 1 decision. The high level trigger can then be applied in a series of steps of increasing refinement until the event can be either positively accepted or rejected. Broadly speaking we distinguish between two basic steps in this procedure. The first, which we have called the Level 2 trigger, is designed to match vertex information provided by the silicon detector with the momentum information provided by the tracking system. This identifies and rejects L1 triggers with fake displaced secondary vertices. Most of the Level -2 cpu requirement comes from the momentum measurement, and existing algorithms have been benchmarked at about 0.15 SI95 sec / track. This has not yet been optimised. Our goal for the L2 processing is 0.25SI95 sec / event. From these figures we estimate the total installed cpu capacity as follows :

$$(0.25 \text{ SI95 sec})\ (40 \text{ kHz}) = 10{,}000 \text{ SI95}$$

The second step, which we call Level 3, uses refined and optimised reconstruction algorithms to select B decays with different event topologies (charged two body, dilepton, low multiplicity with neutrals, D mesons, non- b-physics channels). Our goal is 5 SI95 sec/event.

$$(5 \text{ SI95sec})(5 \text{ kHz}) = 25{,}000 \text{ SI95}$$

The size of the raw event has been estimated from simulation studies to be on average 70 kB. We add a 50% contingency and assume 100 kB. The trigger rate to storage we have assumed (200 Hz) has been estimated from what we can reasonably afford to store and process in the initial phase of understanding in detail the behaviour of our detector. N.B. The expected rate of interesting physics events is estimated to be only a few Hz. The trigger software will be adapted as this understanding evolves and the rate to storage may therefore be expected to decrease with time. Assuming a running period of 120 days and a duty cycle of the LHC machine of 50%, we therefore expect to accumulate raw data at a rate of ~1 TB a day, or ~100 TB a year.

2.2 Reconstruction requirements

The cpu power needed for reconstruction and the size of the ESD have been more difficult to estimate as work is still very much in progress to develop our pattern recognition algorithms. Measurements taken from the existing FORTRAN-based codes have been used as a basis of the figures quoted in Table 1. For certain subdetectors, such as the RICH, the development of OO based algorithms is quite advanced and performance measurements on these codes, which are written in C++, indicate that similar performance can be achieved without major efforts in optimisation. Using all this information, and assuming that a significant improvement can be achieved (factor 2) once the optimisation has been done, we have set a target of 250 SI95 sec per event. (N.B. similar significant improvements were seen by both HERA-B and BaBar following optimisation.) From these figures we estimate the total installed cpu capacity required to keep up with datataking as follows :

$$(250 \text{ SI95 sec})(200 \text{ Hz}) = 50{,}000 \text{ SI95}$$

The size of the ESD data is estimated conservatively as 100 kB per event i.e. comparable to the raw data. Thus we expect to generate 1 TB of ESD data a day, and 100 TB per year.

Reprocessing of the complete year's data sample will need to be performed at least once and possibly twice. This reprocessing can be performed on the same online farm during non-datataking periods, such as the shutdown. All the cpu capacity will be available, including processors normally used for the high level triggers. The time available would normally allow at least two full reprocessings of the complete data sample taken during the previous year.

Table 1 Data volumes and cpu requirements for processing and storage of real data

Length of data taking period per year	120 days ~ 10^7 secs
Duty cycle of the LHC machine	50%
Rate of events to storage	200 Hz
Total number of events per day	$(0.5)(8.6\ 10^4)(200) \sim 10^7$
Total number of events per year	$(0.5)\ (10^7)\ (200) \sim 10^9$
The raw data size per event	100 kB
Total raw data per day	(100kB) (10^7) = 1 TB
Total raw data per year	(100kB) (10^9) = 100 TB
ESD size per event	100 kB
Total ESD size per day	(100kB) (10^7) = 1 TB
Total ESD data per year	(100kB) (10^9) = 100 TB
AOD size per event	20 kB
Total AOD data per day	(20kB) (10^7) = 0.2 TB
Total AOD data per year	(20kB) (10^9) = 20 TB
TAG size per event.	1 kB
Total TAG data per day	(1kB) (10^7) = 0.01 TB
Total TAGdata per year	(1kB) (10^9) = 1 TB
CPU power for L2 processing	10,000 SI95
CPU power for L3 processing	25,000 SI95
CPU power for reconstruction	50,000 SI95
CPU power for production analysis	2,000 SI95
CPU power for user analysis at Regional Centre	10,000 SI95
CPU power for user analysis at CERN	20,000 SI95

2.3 Production analysis requirements

The production analysis is typically repeated a number of times as the selection algorithms are refined. The total amount of cpu power required for this phase is therefore constrained by the time to process each event and the requirement of following a realistic time schedule. In order to estimate processing requirements for this analysis, we envisage a possible schedule as follows:

The final state reconstruction is performed soon after the data are taken. Different interesting final states are reconstructed and a rough tag of the event produced. A first selection step will identify the appropriate analysis to perform. Different analyses could run on the same event because the selection criteria for many could be satisfied (e.g. 2μ, 2hadrons). We expect each analysis will run on a subset of the 10^9 events (~10^8 depending on the analysis algorithm) and produce a positive tag for ~10^7 events. We estimate from measurements of existing algorithms that this step will take ~5 SI95 sec per event to complete.

We expect that these algorithms will change and that the production analysis will need to be repeated on the full event sample ~4 times a year. A reasonable requirement would be for the whole production to be completed in ~ 3 days.Thus the total installed CPU power required is estimated to be ~2000 SI95.

The size of the AOD dataset has been estimated to be 20 kB for real data. This allows for inclusion of some data generated by the reconstruction step (ESD) which is needed during the user analysis phase and which is included for convenience.

The design of the event tag information has not been made, but it is expected that after reconstruction the tagging dataset would take ~100 bytes/event and will grow to a few hundred bytes after analysis stages. The main function of the 'event tags' is to optimise analysis process access to interesting events according to the physics channel. It is foreseen that the selection processing will follow a tree-like logic.

2.4 User analysis requirements

The user analysis is performed by each physicist in a semi-interactive mode i.e. within a short response time of a few hours (say 4 hours). The user analysis job focuses on one particular analysis channel and runs on AOD data that have been appropriately tagged. The steps involved are as follows:

The ~10^7 candidates tagged by the production analysis are scanned and events of interest for the user physics analysis are selected. In the worst case ($B > D^*\pi$) all 10^7 events are processed, whereas for analyses studying other channels only 10^6 events need to be processed by the analysis algorithm. The selection step requires ~0.25 SI95 sec / event. The physics analysis is performed on all selected events and requires on average ~20 SI95/event.

The total installed cpu power required for user analysis depends on the number of active physicists. We make the following assumptions

- there are 140 physicists actively doing analysis
- each physicist submits on average one production job per weekday. This may be exaggerated, but allows for the fact that they will also use the production facility for very many short jobs on which cuts and algorithms are refined.
- the analysis is distributed over a number of regional facilities and it is assumed that that on each facility there are ~20 physicists (typical for a Regional Centre) to ~40 physicists (at CERN) submitting jobs on any one day.
- that on average one fifth of analysis jobs analyse all 10^7 events whilst four fifths of jobs analyse 10^6 events.

With these assumptions we estimate that the total installed CPU power required for user analysis ranges from 10,000 SI95 at a regional centre to 20,000 SI95 at CERN.

2.5 Simulation requirements

Simulation studies are made in order to measure the acceptance of the detector and the efficiency of the full reconstruction and analysis of the B decay channel. The number of simulated events that need to be generated is determined from the number of signal and background events taken during real datataking.

We assume that 10 times as many simulated signal events are required as are found in the real data sample.The signal event sample is dominated by B-> $D^*\pi$ decays for which we expect ~10^6 events per year, including background, in the real event sample. We therefore need to produce ~10^7 simulated B-> $D^*\pi$ events per year.

The simulation involves a number of steps:

- physics generation (e.g. using PYTHIA), cuts are applied to take only those events that go into the detector
- the tracking through the detector using GEANT to produce detector hit information
- digitisation to simulate the response of the detector and produce digitisings
- triggering to select those events that would cause the LHCb trigger to fire.
- full reconstruction of the triggered event sample
- apply analysis cuts to reconstructed event sample and do CP analysis

The cpu power required to produce the B-> $D^*\pi$ event sample is shown in Table 2. The total cpu power required is 3x 10^{12} SI95 sec, corresponding to an installed cpu capacity of 100,000 SI95.

Table 2 CPU power required to simulate 10^7 B-> $D^*\pi$ events in one year

Step	Number of events	cpu time / event	total cpu power
physics generator	10^{10}	200 SI95 sec	2 x 10^{12} SI95 sec
GEANT tracking	10^9	1000 SI95 sec	10^{12} SI95 sec
Digitisation	10^9	100 SI95 sec	10^{11} SI95 sec
Trigger	10^9	100 SI95 sec	10^{11} SI95 sec
Reconstruction	10^8	250 SI95 sec	2.5x10^{10} SI95 sec
Final State Reconstruction	10^7	20 SI95 sec	2x10^8 SI95 sec

Concerning background, we know that ~100 k bb inclusive events are produced in the detector every sec, of which ~100 events are logged. Thus the "efficiency" of the LHCb detector for triggering on these events is 10^{-3}. Thus if we need as many simulated events as those found in the real data then 10^{12} bb inclusive events will need to be generated, tracked with GEANT, digitised and triggered, and 10^9

events will need to be reconstructed. This would correspond to a staggering 3x 10^{14} SI95 sec per year, or a factor of 100 more than for the simulated signal sample.

In view of this we are studying ways of optimising background simulation so as to reduce these requirements. Clearly some savings can be made by storing data produced at the event generator level and reusing them in subsequent simulations. This is at the expense of extra data storage requirements.

Further improvements in performance will require optimisation of the generator itself, in such a way that the physics is not biased. Background that is particularly dangerous for a specific physics channel will be identified and generated in the amount necessary for the specific physics channel studies. Cuts on this background will be made as early as possible in the simulation sequence. We expect that this would reduce our cpu requirements by 1-2 orders of magnitude.

It is clear that the amount of background we can simulate will be limited by the installed cpu capacity available. Currently we assume 400,000 SI95 but this estimate will be revised as our understanding of this issue evolves.

Estimates of data volumes and cpu requirements for processing and storage of simulated data are given in Table 3. The size of the raw data for each event is larger than for real data due to the presence of "truth information". This truth information consists of the physics history of the event and the relationships between hits and incident particles. This history information is carried through to the AOD data set and is used during the user analysis phase.

Table 3 Summary of processing and data storage requirements for simulated data

CPU power for signal events	100,000 SI95
CPU power for background events	400,000 SI95
Raw data size per event	200 kB
Total raw data per production	(10^9) (200 kB) = 200 TB
Generator data size per event	12 kB
Total generator data	12TB
ESD data size per event	100 kB
Total ESD data per production	(10^9) (100 kB) = 100 TB
AOD data size per event	20 kB
Total AOD data per production	(10^9) (30 kB) = 20 TB
TAG data size per event	1 kB
Total TAG data per production	(10^9) (1 kB) = 1 TB

3 Baseline Computing Model

3.1 Architecture of Baseline Computing Model

The baseline LHCb computing model is based on a distributed multi-tier regional centre model. In the following we use the MONARC [1] terminology for components of the model and describe how we intend to adapt the general architectural model for our experiment (Figure 5).

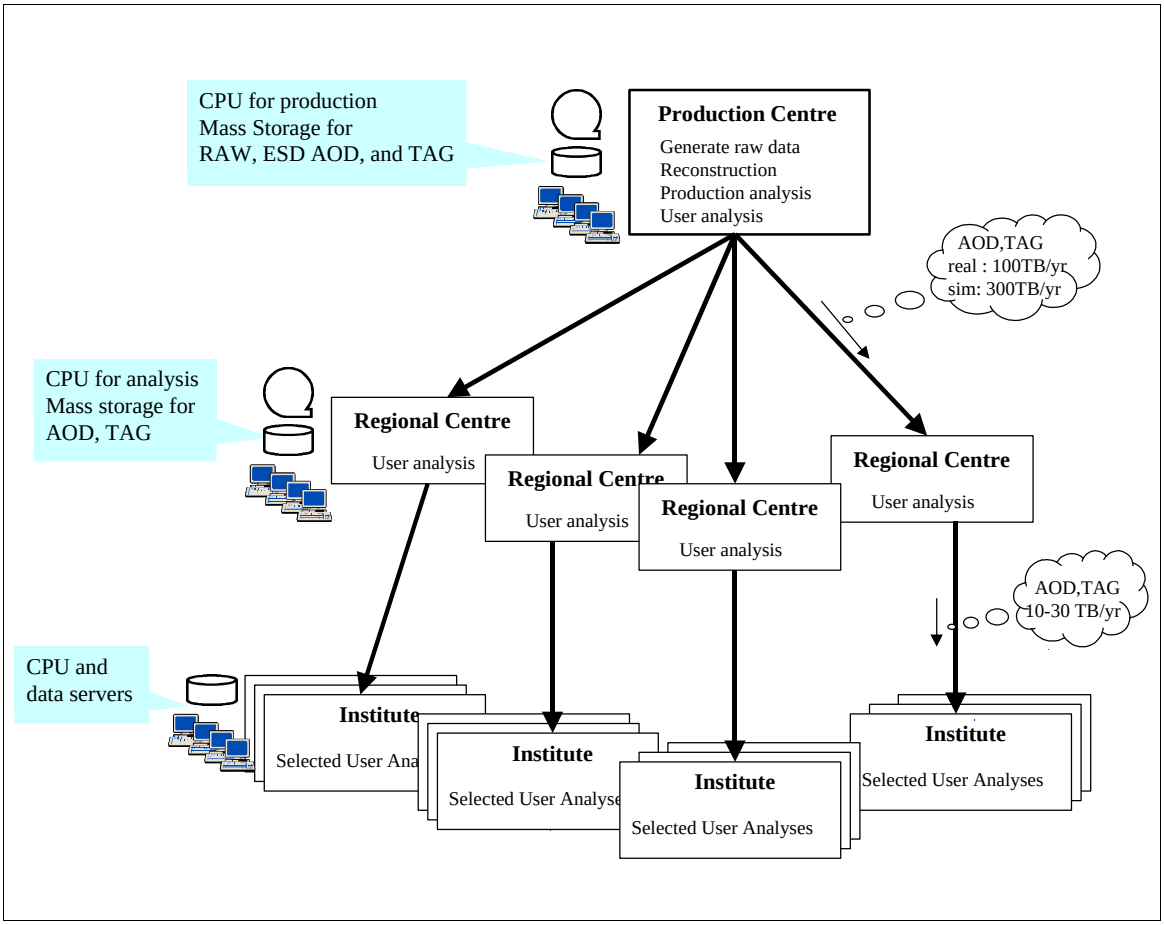

Figure 5 The LHCb multi-tier distributed computing model

The facility at which data are produced is called the *production centre*. At present we assume that the production centre will be responsible for all production processing phases i.e. for generation of data, and for the reconstruction and production analysis of these data. The production centre will store all data generated by the entire production, namely RAW, ESD, AOD and TAG data. Furthemore we assume that physicists will do the bulk of their analysis using the AOD and TAG data only and therefore only the AOD and TAG data will be shipped to other regional centres on an automatic basis.

After 2005, the role of CERN will be to be the production centre for real data. All production processing of real data up to the generation of AOD data sets will be done at CERN. The AOD and

TAG datasets will be shipped to the LHCb regional centres, which will serve these data to physicists running analysis jobs in production there. The AOD and TAG datasets will be distributed to the regional centres each time they are regenerated i.e. after each cycle of the analaysis production step. Assuming four analysis cycles per year, the total amount of data to be shipped corresponds to ~80 TB per year. The user analysis will be performed using these data on the analysis production facility at the regional centre. The private data generated by the user analysis (ntuples) will be shipped to the physicists desktop at his institute. In addition AOD and TAG data corresponding to specific event samples may also be shipped to the institute.

In order to produce an equitable sharing of the computing load we are considering a model in which, after 2005, all simulation production will be done on facilities outside CERN, at regional centres with mass storage capability that will be required for archiving all data generated (RAW, ESD, AOD and TAG). Those institutes not having mass storage but having cpu capacity will archive their data at the nearest regional centre. Thus for simulation the regional centre is filling the role of the production facility with respect to the distribution of these data to the rest of the collaboration. As in the case for real data, the AOD and TAG data for simulation productions will be shipped from the production centre to other LHCb regional centres, including CERN, to serve the analysis jobs of remote physicists.The total amount of data to be distributed to each centre corresponds to ~120 TB per year.

Figure 6 illustrates this role of computing centres for the two types of data. Note that in our current thinking, unlike the MONARC model, we do not distinguish between Tier 1 and Tier 2 centres.

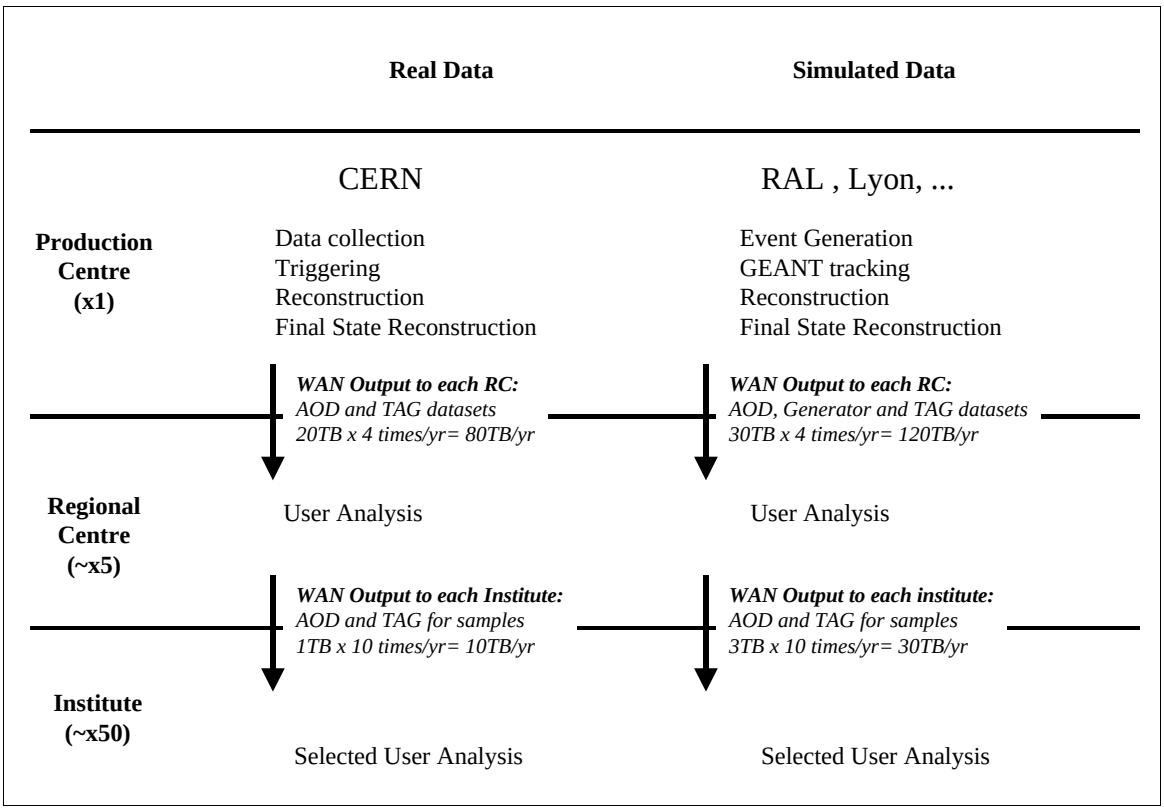

Figure 6 LHCb computing model showing centres involved in real datataking and simulation and requirements on data transfer between centres

3.2 Computing Model Scenarios

The way in which the baseline model would be applied in practice can be understood through specific examples. In here we described three scenarios for the analysis of different data samples by a physicist situated remotely from the data production centre.

3.2.1 Scenario 1 - User analysis on real data

A physicist at Imperial College London wishes to perform a $B > J\Psi/K_S$ physics analysis on all real data taken during the first year of datataking.

The physics algorithm must first scan TAG data corresponding to all 10^9 events. The selection criteria will identify 10^7 candidates with physics features of interest during the production analysis phase.The physicist's analysis job will be run on the AOD data corresponding to these events to select the particular $B > J\Psi/K_S$ candidates ($\sim 10^6$). The number of real fully reconstructed $B > J\Psi/K_S$ events is expected to be $\sim 10^5$. The analysis job outputs ntuple and statistical information and these data are interrogated interactively many times. In some cases the decay selection algorithm may need to be changed in which case the whole procedure will need to be repeated on the full 10^9 events. Otherwise the user physics analysis algorithm may change but this will be run only on the 10^7 selected events. In addition systematic studies will be performed to look at the influence of particle identification, tracking etc. on the algorithms that require access to ESD information. The data sample required is approximately the same as for the real signal i.e. $\sim 10^5$ events. Physicists may look in detail, for example using the event display, at complete event information including the raw data for very small samples of ~100 events.

This analysis would be realised as follows:

1. The AOD and TAG data for all 10^9 events would be available at RAL, having been distributed automatically from CERN to RAL (the UK's Tier 1 centre) as they are produced.
2. The scanning procedure would be run on the RAL computer facility to identify the 10^7 candidates of interest.
3. The AOD and TAG information corresponding to the 10^7 selected events would be copied from RAL to Imperial College. The total amount of data moved corresponds to 200 GB (AOD) and 10 GB (TAG). This amount of data could be conveniently shipped over the WAN in a few hours.
4. The analysis job would typically be run at Imperial College many times on these selected events.
5. If a change in the selection algorithm is needed the above procedure would be repeated from step 2.
6. Systematic studies would involve copying ESD data from the production centre (in this case CERN) for 10^5 events i.e. 10 GB of data.
7. A small sample of events would be checked interactively, for example with the event display. For these events, RAW and ESD data for ~100 events would be copied from CERN to IC i.e. 100 MB of data.

N.B. This scenario applies to most signal analyses except for $B \to D^*\pi$. In this case the number of events is ten times larger and requires a significantly larger cpu facility and WAN bandwidth. This

analysis would presumeably be performed at the regional centre. Note also that the copying of data sets between RAL and IC only has to be done once and should not be repeated for each IC physicist doing analysis. A tag database is needed to keep track of which data is available locally. The data caching and replication software (grid software) may help to ensure this happens transparently.

3.2.2 Scenario 2- User analysis of simulated background events

A physicist in Glasgow wishes to run his B > JΨ/K_s analysis on background bbar events generated at a simulation production facility located at the Lyon/IN2P3 regional centre.

The sample of events to be analysed corresponds to the total real data sample i.e. 10^9events. The total amount of data generated is RAW - 200TB, ESD -100 TB, AOD and corresponding generator data - 30 TB and TAG 1 TB.

1. The RAW and ESD data are archived at the production centre i.e. Lyon.
2. The AOD and TAG data are automatically distributed to other Tier 1 centres, in this case RAL, and are archived there. This involves transfer of 31 TB of data from Lyon to RAL.
3. The physicist will run a job at the RAL Tier 1 centre that scans the TAG dataset to select interesting events (~10^5events).
4. The AOD and TAG data corresponding to these 10^5events are then copied to Glasgow. The total volume of data transfered is very small i.e. 22 GB.

3.2.3 Scenario 3 - User analysis of simulated signal events

A physicist at Orsay wishes to run a B > JΨ/K_s physics analysis on simulated data that are produced at the Liverpool computing facility (MAP). This facility has large cpu resources but no data archiving capability.

The total sample of B > JΨ/K_s simulated events needed for this analysis should be ~10 times the number produced in the real data. In one year of datataking we expect to collect and fully reconstruct 10^5 events and therefore the number of simulated B > JΨ/K_s events to be produced is 10^6. The number of events that have to be generated, stored and reconstructed to produce this sample is 10^7.

This analysis exactly as described in scenario 1 and would be realised as follows :

1. The production of the events would be made at Liverpool. The total amount of data generated corresponding to 10^7events is 2 TB of RAW, 1TB of ESD, 0.3 TB of AOD and 10 GB of TAG.
2. All data generated (~4TB) would be transfered to RAL, the Tier 1 centre for archive. In this case RAL will fulfill the role of production centre and will distribute the AOD and TAG datasets to other LHCb Tier 1 facilities, including Lyon.
3. The physicist would either run his analysis on the Lyon analysis facility or copy the AOD and TAG data from Lyon to Orsay and do his analysis there.
4. He would also copy the 10% of the ESD data for systematic studies (~100 GB).

N.B. This production is particularly suited to a facility, such as Liverpool, that has significant cpu capacity but somewhat limited storage capability.

3.3 Differences with MONARC Model

Although the basic architecture of the LHCb computing model corresponds well with the MONARC model there are a number of details that distinguish LHCb from the larger LHC experiments:

- The first stage in the analysis is performed in common for all the analyses that subsequently follow. We do not explicitly identify group analyses. The number of different analyses is very large due to the large number of decay channels that are studied. In this sense physicists are working largely independently on different decay channels.
- We intend to run all data processing from the production of the RAW data through to the generation of the AOD data at the production centre. We believe that it will be necessary to only ship AOD and TAG data to outside facilities. The data loads are such that this should be realised in nearly all cases by using the wide area network infrastructure.
- Although no discussions have taken place to reach a formal decision, it seems natural to devote CERN resources to the processing of real data produced at the experiment and to produce simulated events exclusively in the facilities outside CERN. This will depend on sufficient resources being identified in the computing facilities of LHCb institutes.
- Our data processing requirements are such that at present we do not see a clear need to distinguish Tier 1 and Tier 2 centres.

4 Event Processing at the Collaboration Centre (CERN)

Several activities proceed in real-time at the collaboration centre (CERN) as the data are collected. Figure 7 is a schematic view showing a possible implementation scenario of the compute facility installed at CERN. In this scenario high level triggering and reconstruction run together on the cpu farm close to the detector in the LHCb pit. From Table 1 we see that the total installed cpu capacity required to do this is estimated at ~100,000 SI95 units. The raw data and ESD data sets resulting from the reconstruction are sent over a Central Data Recording link directly to the computer centre where the data sets are archived on data recording media. Thus the load on the link from the pit to the centre will be ~ 40 MB/s during datataking (20 MB/s raw data and 20 MB/s for ESD data).

Sufficient storage capacity must be installed at the pit to be able to stage raw data for processing by reconstruction and to accommodate interruptions in the connection to CERN for realistic periods. We estimate these data to accumulate at the rate of 2TB per day (see Table 1). For example, 10TB of storage can accommodate at least 5 days of data-taking. In normal operation priority is obviously given to trigger processing but any spare capacity can be used for re-reconstruction.

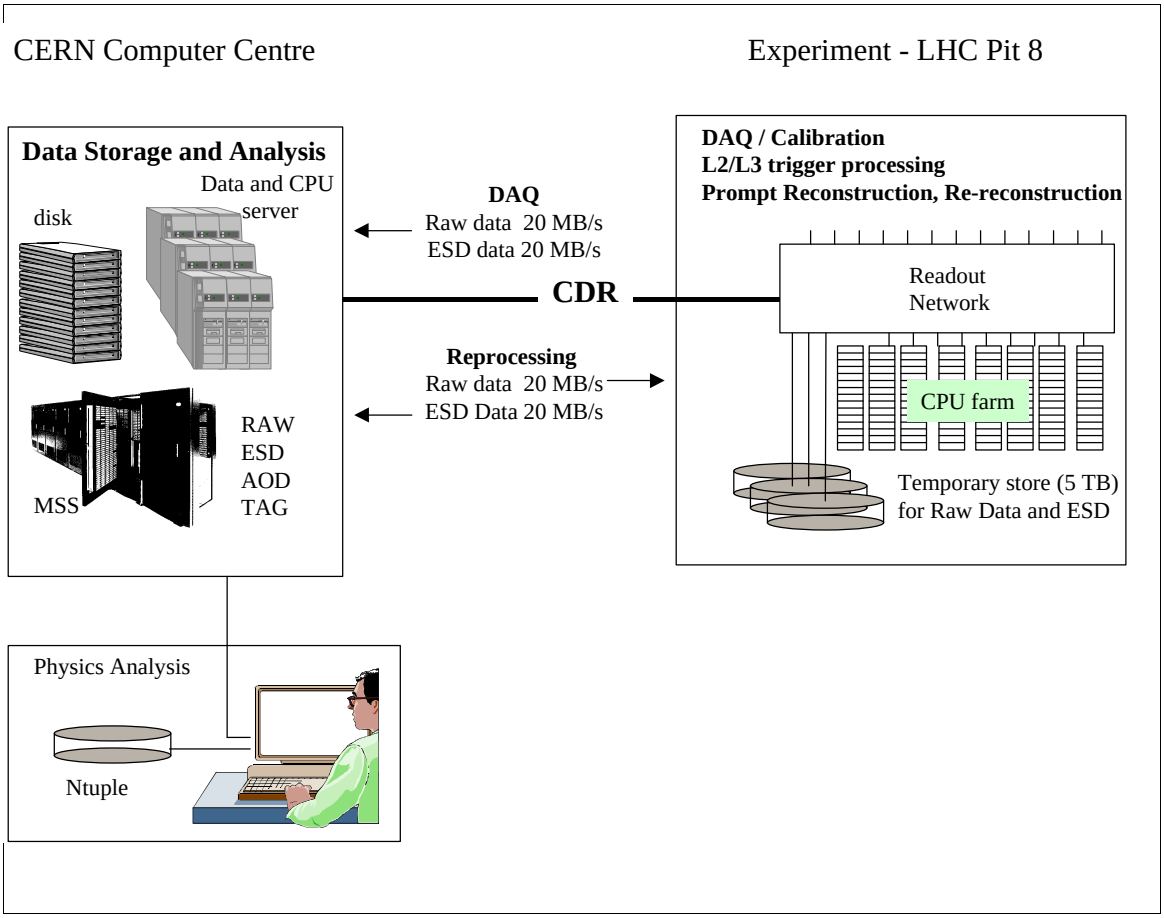

Figure 7 Schematic of CERN based CPU and Storage Facilities

Table 4 Parameters describing the resources in the compute facility at the LHCb pit

CPU facility	~100,000 SI95
disk storage for event related data	>10 TB
disk storage for calibration and other secondary data generated by online	5 TB
Capacity of data link from pit to Computer centre	80 Gbps

The advantage of this scheme is that any spare cpu capacity can be used for re-processing data already taken. For example during periods between fills and when the LHC machine is not operational the entire farm can be used for the re-processing of raw data already taken. In addition, soon after startup the duty cycle of the machine can be expected to be low and to slowly improve with time. All the cpu capacity will be available, including processors normally used for the high level triggers. Therefore the time available would normally allow at least two full reprocessings of the complete data sample taken during the previous year.This gives optimal use of all installed cpu capacity. Re-processing will involve reading data from the centre back to the pit, processing them there and then sending the results (ESD datasets) back to the centre for storage. If the reprocessing proceeds twice as fast, then the total load on the link will be ~ 80 MB/s, 40 MB/s to read the raw data and 40 MB/s to copy the ESD data back again. The bandwidth of the installed link is planned to be ~80 Gbps, and so these rates can be easily accommodated. Table 4 contains a summary of the requirements for an installed compute facility at the experimental area.

The archiving of all raw and ESD data resulting from datataking will amount to ~200 TB per year of data being stored at the computer centre. In addition a backup of the raw data will be made generating another 200 TB per year. Production physics analysis will run at the CERN computer centre requiring access to ESD data and calibration data. The installed CPU capacity needed to run the physics production is estimated to be 1000 SI95 units (see Section 2.3). The analysis will require direct access to all ESD data i.e. 100 TB.

In addition to the raw, ESD and AOD data resulting from real datataking, we will also import AOD data sets resulting from the production of simulation data. These simulated data will be produced in regional facilities external to CERN and will be imported via WAN connections. Likewise AOD data sets from real dataking will be exported to outside regional centres to allow physics analysis by physicists working remotely from CERN. Parameters reflecting LHCb requirements on the size of the CERN facility required to store and access data, to run the analysis productions on real and simulated events, and to import and export data over WAN connections are shown in Table 5.

The total installed CPU power required for user analysis on real and on simulated data has been estimated to be 30,000 SI95 at CERN (see Section 2.4). It is assumed that further interactive data analysis takes place on the desktop using the CPU power available there. Physicists will store TAG data at their desktops and possibly private collections of events in various data formats. Estimates of the resources required are also summarised in Table 5.

.

Table 5 LHCb's requirements on compute resources in the CERN Computer Centre

Raw data storage	100 TB/year
Copy of raw data	100 TB/year
ESD data storage	100 TB/year
AOD data storage	20 TB/year
Tag data storage	1 TB/year
AOD & generator simulated data storage	30 TB imported 4 times / year
Tag simulated data storage	1 TB imported 4 times / year
Total data storage	~ 250 TB / year
Installed CPU for AOD real data production	5000 SI95
Installed CPU for production user analysis jobs	20,000 SI95
Data storage on the desktop	100 GB
WAN requirements AOD and TAG export	80 TB/year
WAN requirements AOD and TAG import	124 TB/year

5 LHCb Facilities and Resources

Currently the LHCb collaboration comprises 45 institutes distributed across Europe, North and South America and China (Figure 8).In this chapter we summarise the situation for existing computing facilities, and the planning for their enhancement in the lead-up to 2005.

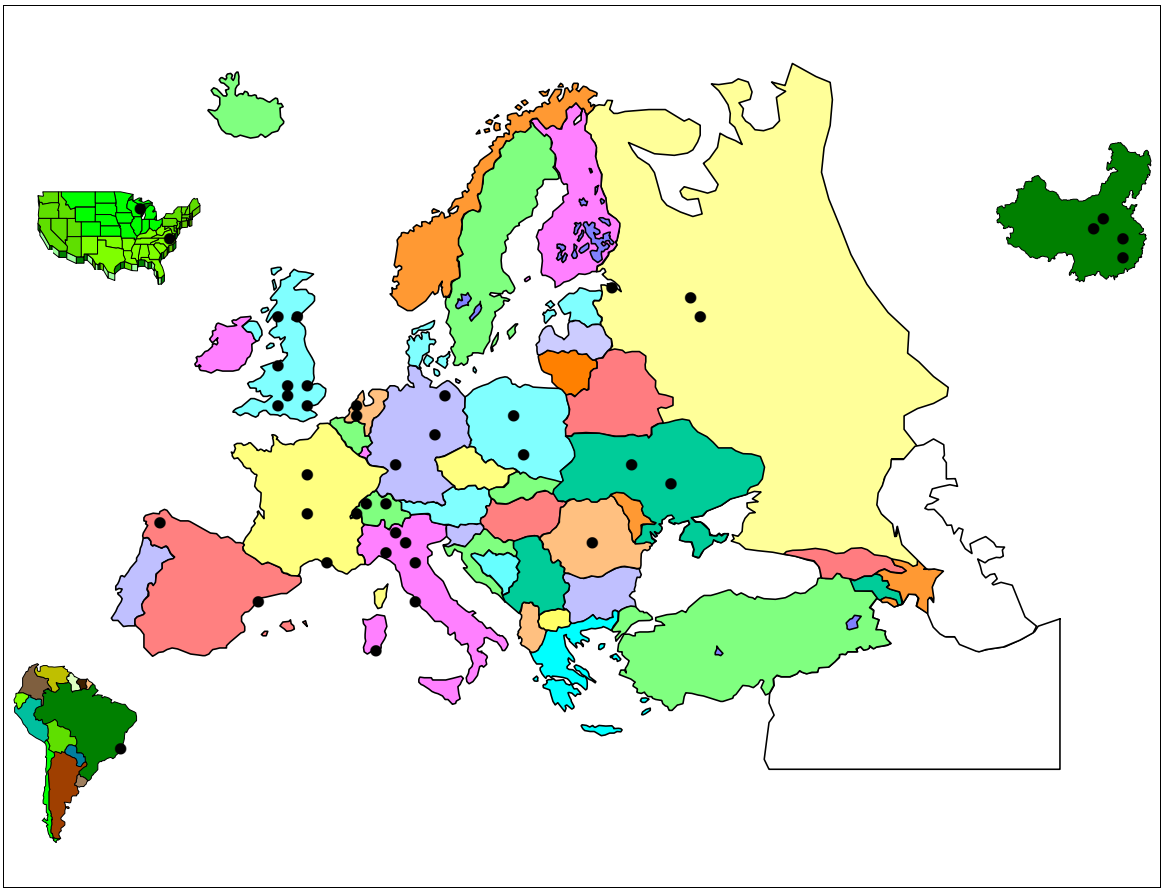

Figure 8 Schematic figure showing the world-wide distribution of LHCb Institutes

Planning for dealing with the computing needs of the LHC experiments is under discussion in all countries represented in our collaboration. Our current understanding for the role each of our centres is likely to play can be summarised as follows :

- Today LHCb makes use of computing centres in France ($CCIN_2P_3$/Lyon) and in the UK (RAL) for satisfying a significant fraction of the demand for CPU to produce simulated events. It is rather clear that these will become Regional Centres for LHCb, as well as for the other LHC experiments. They will be used for production of simulation events and also to support physics analysis.
- There are discussions in Italy, Holland, Germany and Switzerland for the setting up of regional centres with significant cpu and mass storage capability. These would be used for supporting physics analysis, but we would also expect them to share the simulation load.

- There are plans for a regional centre in Moscow and already a cpu farm has been assembled and is producing simulated events for CMS. A crucial issue will be the bandwidth of the link between Moscow and CERN.
- Currently we are not aware of national initiatives in Poland and Spain.
- A 300 node cpu farm has been operating at Liverpool University for ~1 year and has been producing large quantities of simulated events for LHCb. This facility is dedicated to simulation. Facilities also exist at other institutes (e.g. ETH Zurich, Lausanne,..) and may be brought into operation in the future.

More detailed information is given in the following sections of the current status of computing in the various countried represented in the LHCb collaboration.

5.1 BRAZIL

There is one institute involved, Rio, which participates in the design and construction of the muon detector. In addition this group is very strong in physics analysis and computing sytems support.

There are currently 5 physicists involved, this growing to ~10 by 2005. Currently the group has access to a Linux Farm comprising 14 Pentium II machines (233 MHz) and 40 GB storage.

5.2 FRANCE

There are 3 institutes involved , LAL/Orsay, LPC/Clermont and CPPM/Marseille. There are currently 20 physicists involved, this number expanding by 2005. They are working now on the design and prototyping of the electromagnetic calorimeter, and the calorimeter and muon triggers. They will enter the construction phase in 2003.

The computing needs in 2000-2002 are based on Monte Carlo needs associated with the above activities. Following on from that physics analysis activities will build up based on increased volumes of MC data.

It is foreseen that at Lyon there will be a Tier1 centre servicing all the 4 LHC experiments, together will other French and international experiments. France is also planning a substantial participation in EU-GRID. One would expect that LHCb/France might be involved in the Testbed part of this project

At Marseille and at LAL there is an on-going migration to Linux PCs as desktops with general services (backup, licensed software, etc) running on on a UNIX server (Sun). At the Lyon centre MC production is currently running on AIX and HP-UX and also a migration to Linux based systems.

Data Storage is required for test beam data (2000-2002 ~10TB/year, 2003-2005 ~50TB/year) and for simulation (2000-2002 ~2 TB/year, 2003-2005 ~7-10 TB/year).

CPU requirements at the Lyon centre are ~150-300 SI95 (2000-2002) rising to ~500-2000 SI95 (2003-2005)

5.3 GERMANY

The 6 German institutes (Dresden, Freiburg, Humboldt, KIP Heidelberg, MPI Heidelberg, Univ Heidelberg) are involved in the Vertex Detector, the Vertex Trigger and the Inner/Outer Trackers. Current computing effort concerns the design and construction of these, with more effort moving in 2002 to more general physics studies.

The long term strategy for computing for LHC in Germany is under discussion. It is intended to apply for funds for a regional centre to be established after 2004. Although no effort is currently available, the EU-GRID initiative is being followed closely.

At MPI-Heidelberg there is a Linux farm with 15 CPUs, 10-20 GB disk. There is also access to a central facility with 300 GB disk and 10 TB tape library.

5.4 ITALY

There are 8 Italian LHCb institutes (Bologna, Cagliari, Ferrara, Firenze, Genoa, Milan, Rome-1, Rome-2) involved in the design and construction of the ECAL, Muon, RICH detectors and the Level-0 calorimeter trigger.

LHCb /Italy are currently writing a proposal defining their computing requirements from now up to the start-up. This will be submitted to INFN in April 2000. This will be based on the MC and analysis needs of the Italian groups during the design phase, followed by intensive physics studies leading up to the start-up.

This will include some data-challenge aspect according to the overall strategy of LHCb. Also it may well relate to the overall Italian GRID effort in a way to be defined.

We present below the current situation in Bologna as being typical of the Italian situation (more details on other institutes to follow). They use a facility dedicated to LHCb and Hera-B comprising

- 1 HP 9000/800/K250 Server with 4 160 MHz CPUs
- 60 GB disk
- 2 DLT units
- 18 DLT 7000 tape library
- 4 dual 400 MHZ P II CPUs running Linux

5.5 Netherlands

There are 4 Dutch institutes (Nikhef-FOM,Univ of Amsterdam,Vrije Univ. and Univ. of Utrecht) involved in the design and construction of the Outer Tracker. There will be a substantial MC requirement in the next 2-3 years for this, being followed by a build up to large scale physics studies.

NIKHEF is combining with SARA(Academic Computing Services,Amsterdam) and KNMI(Dutch Meteorological Institute) to participate in the proposed European GRID. Included in the grid would be

a large CPU farm to be used for Monte Carlo production and data analysis. Thus Nikhef is foreseen to be a Tier1/2 centre for LHC computing . Its scale remains to be defined, but it will include a substantial farm facility. Indeed the Dutch grid will include large farms and storage facilities in several centres.

The current facilities being used by Nikhef are ~ 4 high end HP Unix servers, and a growing Linux PC farm. Nikhef is currently planning to install a Farm of an equivalent a 50 dual 600 MHz PC's, a disk space of 1.2 TB, and a dedicated tape writing facility. A 10% version of this farm is ordered now, and should become available by summer, the full farm is expected functional in a year or so.

5.6 Poland

There are 2 Polish institutes in LHCb, Cracow and Warsaw, involved in the Outer Tracker and Muon detectors repectively.

There is as yet no national initiative for computing for the lHC experiments. However all the Polish groups are interested in such an initiative, as well as the proposed GRID project.

It is hoped that, at least, there will be a small centre offering facilities for the final stages of physics analysis with local copies of AOD/Ntuples. It is foreseen that the main problems will be manpower and communications to CERN.

The physics group at Cracow has access to a small Linux farm of 5 CPUs. Two are dual-CPUs with RAID controllers and 4*20 GB disk. The central computing group hopes to set up, in addition, a farm of 10 CPUs.

5.7 Russia

There are 5 Russian institutes in LHCb (INR Moscow, ITEP Moscow, LPI Moscow, IHEP Serpukhov, NPI Petersburg), working on the design and construction of the hadron calorimeter and muon detectors.

Nearly all Russian Institutes support the case for a Regional Centre in Russia, and have signed a memorandum. The project "A Russian Regional Centre for LHC" is supported by the Russian Ministry of Science and Technology. However funding remains to be agreed.

It is proposed to have a TIER-1 centre in Moscow, and TIER-2 centres in Dubna (JINR), Gatchina(PNPI), Serpikhov (IHEP) and Novosibirsk (Budker Institute). A vital part of this project is the connection with CERN. It will be necessary to have a 622 Mbit/s connection to CERN to be a full-scale TIER-1 centre.

There is significant Russian interest in the EU-GRID proposal.

Current activities include the development of PC (Linux) farms in local institutes (ITEP, MSU INR, Dubna , Serpukhov), studying their connectivity, working with distributed resources, and production MC.

A prototype of a PC-farm has been constructed at ITEP. This is a common project between the LHCb-ITEP group and the CMS-ITEP and CMS-MSU INR groups. It has been tested tested and is now producing MC data for CMS. It is planned to double or triple the production power of this farm by the end of the year. They are ready to produce some MC samples for Russian LHCb physicists.

5.8 Spain

There are 2 Spanish institutes in LHCb (Barcelona, Santiago), working on the Calorimeter and Inner Tracker respectively.

Current facilities available to LHCb at Santiago include 2 Sun Sparc-10 machines, a Sun Ultra60 and 20 GB disk

In general the LHCb Spanish institutes are just beginning their planning for computing.

5.9 Switzerland

The Swiss institutes (Lausanne,Zurich) are working on the Vertex Detector/Trigger and the Inner Tracker.

Discussions have begun on the possibility of starting a farm-project to provide a facility to be shared between Geneva, Lausanne and EPFL, with the farm situated at EPFL. Also Swiss physicists are currently discussing their position regarding regional centres for LHC, and also the EU-GRID project.

At Lausanne-IPHE there is a dual processor DEC 4100 server with 100 GB of disk, which is shared between LHCb and NOMAD and 4 Linux PCs. Currently the connection to university backbone and to CERN is at 155 Mbps.

Zurich currently has a cluster of 36 Linux machines and 80 GB disk scattered over machines.

5.10 UK

There are 8 UK institutes involved in LHCb,- Bristol, Cambridge, RAL, Edinburgh, Glasgow, Liverpool, Imperial College and Oxford). They have reponsibilitities for the design and construction of the RICH1, RICH2 and VELO detectors. In addition they are working on optimisation of the high level triggers which involve both detectors.

The UK has been developing a policy for LHC computing for some while. It is foreseen that there will be a national Tier1 centre (probably at RAL), and a few Tier-2 centres (Liverppol, Glasgow/Edinburgh). LHCb people have been very active in these matters, with Liverpool already running a 300 node PC farm, and Glasgow/Edinburgh proposing a joint effort, with a MC farm at Glasgow and a large datastore at Edinburgh. These will be part of the UK GRID infrastructure. LHCb/UK are also active in MONARC, with Oxford providing 2 LHCb representatives.

A proposal has been made on behalf of all the 4 experiments for funds for developing the infrastructure of a national centre in the period 2001-2003. This would gradually build up facilities according to the following pattern.

Table 6 Proposed evolution of a UK regional centre for LHC experiments from 2001 - 2003

	2001	**2002**	**2003**
Processors(SI95)	16000	32000	64000
Disk (TB)	25	50	125
Tape (TB)	67	130	330

LHCb UK recently estimated their computing needs for the period 2001-3. These will be primarily in detector and trigger optimisation, with physics background studies starting in earnest in 2003. This estimate was:

Table 7 Estimated LHCb/UK computing needs 2001-2003

	CPU SI95)	**Storage (TB)**
2001	4000-8000	5-10
2002	4000-8000	5-10
2003	8000-1200	10-20

It should be noted that LHCb/UK made a strong case for upgrades of network connectivity to the Tier2 centres and the institutes. Also in specifying the storage needs it also should be noted that Liverpool have been developing 1 TB analysis engines, and strongly favour moving away from tape to largely disk-based sytems.

LHCb institutes are prominent in the UK-GRID initiative, and also in the UK effort for the EU-GRID proposal. Certainly at least Liverpool, Glasgow, Edinburgh and RAL should be integrated early on into a UK HEP grid.

5.10.0.1 Current facilities used by UK/LHCb

Liverpool have been operating a 300 node (400 MHz P II) farm since late 99, and have so far produced 10^7 MC events for LHCb . The farm produces about 10^6 B events/week. This will continue to be used by LHCb, but will also be used by Atlas.

A 15 node NT farm, based at RAL, has been used for MC generation of interesting B-decay and minimum bias events. It has so far produced some 10^6 events.

All the LHCb/UK institutes are comnnected to the Janet academic backbone, and as such have good effective connectivity to RAL and CERN (~500 kB/s). This will require updating for the needs of LHC.

All institutes have powerful Unix servers, and typically many desktop NT or Linux systems. For example Oxford has 120 NT desktops in the physics department. However CPU-intensive work is performed on high-end Unix servers (Oxford has of the order of 12 PC99 CPU power spread over 6 servers). All institutes are looking to upgrade their local server capabilities once the facilities at the Tier1 and 2 centres become better defined.

As can be seen from the above, the LHCb collaboration is already making extensive use of computer facilities at regional centres in France (IN2P3/Lyon) and in the UK (RAL) for production of simulation events. We assume that these will become regional centres for LHCb institutes in France and the UK respectively. Several institutes have significant computing resources. Liverpool University has assembled a 300 node cpu farm (MAP project) for making detailed simulation studies, which are used for optimising the design of the detector.

5.11 Networking

The current situation is as follows :

CERN is connected to TEN-155 public network via a 10 Mbps connection.

Lyon is connected to CERN by a 6 Mbps link, soon to be upgraded to 34 Mbps.

The capacity of the link from Rio to CERN is only ~5 kB/s.

The performance of the network between Heidelberg and CERN and to other centres in Germany is currently ~ 500 kB/s.

In Italy the CERN-CNAF/Bologna connection is 155 Mbps. Current plans are to upgrade the link to 2.4 Gbps in 2003.

Nikhef has a 155 Mbps connection to the Dutch national academic backbone (622 Mbps), and to the other European academic networks. The Dutch backbone is to be upgraded in 2000 to 10 Gbps.

For Moscow the requirement is to have a 622 Mbps connection to CERN to be a full-scale TIER-1 centre.

At Lausanne-IPHE the connection to the university backbone and to CERN is at 155 Mbps.

In the UK the planning for SuperJANET 4 is to provide an initial capacity of 2.4 Gbps by Jan 2001 and 10 Gbps by June 2002 to most UK institutes. The LHC community is aiming to have at least a 622 Mbps link from RAL to CERN in place by 2005. It is aimed to finance by 2002 a dedicated 50 Mbps link as a stepping stone. This link would be shared by the 4 experiments.

5.12 Storage Technology

This matter is the subject of technical investigation within LHCb at Liverpool University who are developing analysis stations with up to 1 TB of disk store attached to a single PC. The questions of all-disk systems are being addressed to avoid the manpower and hardware costs associated with tape

robots. Before addressing how such systems could be used in the LHC context it is relevant to summarise some cost and technical factors.

Concerning costs, currently IDE disks cost 22CHF/GB. If we assume a 35% price improvement / year for 5 years. This goes down to 2.6 CHF/GB. Folding in the use of RAID gives 3.2 CHF/GB. For tape systems media costs dominate with large robotics. This is currently 2 CHF/GB, and will drop (L Robertson predicts to 0.5 CHF/GB).

For overall convenience one would aim to have 'active' data on disk. In the LHCb model this is AOD and TAG data, giving a total/year of ~20-40 TB. The production RAW+ESD data is produced at the rate of ~2 TB/day, so 40 TB would hold 3 weeks of data production.

It has been pointed out that as disk volumes go up the I/O capacity of disk farms goes down. Thus the I/O rate for a 36 GB system is 40 mb/s. However as the capacity goes to 72 GB the rate MB/sec/GB halves.

Thus extrapolating to to 2006, the rate will fall from 800 MB/sec/TB to 200 MB/sec/TB. Thus it is probably better, for I/O rate reasons, to buy more lower capacity disks.

5.12.1 Moving jobs to data

In the model where one has say 1TB attached to a high performance PC then one could plan to move the job to the CPU associated with the required data. However this requires an appropriate resource allocation strategy.

5.12.2 Moving volume data around

Moving tapes is well understood. As is moving data over suitably high performance network links. It has been suggested that one might consider moving disks. However the labour costs of extracting and inserting disks from/to servers has to be evaluated. If possible one would rather move the jobs to the data. However this will not be feasible with the model of copying the AOD+TAG data from the source of the data to other centres. It is hoped to handle that with high performance networking .

5.12.3 Summary of current LHCb thinking

We plan to store the active AOD+TAG data on disks. Also ideally we would store the RAW and ESD data for the current year online on disk storage. The remainder would be on an archiving medium.

We hope to have sufficiently high performance networking to support moving (AOD+TAG) and selected (RAW+ESD) between centres .

6 Plans for deployment of Computing Model

6.1 Plans for Production of simulated events between now and 2005

In 2000 and 2001 we plan to produce ~10^7 simulated events in each year in order to continue detector optimization studies. This work will be needed to finalise the designs ready for the production of the TDR's in 2001. Between 2001 and 2003 efforts will concentrate of studies of the high level trigger algorithms for which we will require ~2x 10^7 events/year. We would expect to begin assembling and commissioning large scale compute facilities in 2004 and 2005 and we will take advantage of these to produce large samples of background events (~ 10^8 events/year). For example, before datataking starts we will aim to use the facility being installed at the pit for background simulation studies.

6.2 Plans for large scale tests

From 2001-2004 we will make tests to validate our computing model. This will include deploying software for operating compute facilities and for supporting distributed processing (e.g. grid middleware). We plan to participate in the HEP Application WP of the EU Grid Proposal, for example by running all our data processing codes over a properly configured subset of the LHCb distributed computing infrastructure. The example given in Section 3.2.3, which involves a user analysis of simulated signal events that makes use of several different facilities, could be a typical example of the form these tests would take. Wherever possible we will make use of ongoing simulation activities to generate the cpu and data storage loads.

Operational experience with large farms by making use of the test-bed to be setup by IT division. This exercise will also be useful for large scale testing testing of our software.

6.3 Rough sizing estimate for an LHCb Regional Centre

The sizing of the computing facility required at CERN was given in Table 5.

A regional centre should be capable of archiving all AOD and TAG data for real data and have sufficient cpu capacity to support user analysis. The requirements give an indication of the cpu and storage capacity that will be needed (Table 8).

Table 8 Regional Centre sizing stimates for supporting user analysis

	2000-2001	2002-2003	2004-2005	> 2005
archive of real AOD and TAG data	-	-	-	80 TB/year
archive of simulated AOD , Generator and TAG data	2 TB	5 TB	20TB	120 TB/year
CPU for analysis of real and simulated data	3000 SI(%	5000 SI95	10000 SI95	10000 SI95

In addition the centre will contribute to simulated event production. The CPU requirements are significant and we assume that the total load will be shared between 5 centres (Table 9).

Table 9 Regional Centre Sizing estimates for supporting simulation

	2000-2001	2002-2003	2004-2005	> 2005
archive of RAW, ESD, AOD, TAG data	5 TB	10 TB	33 TB	333 TB /year
CPU for generation of data-sets	20000 SI95	40,000 SI95	60,000 SI95	100,000 SI95

The basic model assumes that AOD data sets are shipped to other regional centres as they are produced. The size of the AOD is 20TB and is produced in a few (~3) days. This corresponds to a production rate of 100 MB/s. These data can either be distributed over the network or can be shipped by tape, according to what is most convenient and affordable.

6.4 Manpower and Costings

LHCb has formed a working group (Chair F Harris) with national representatives for computing .

- Brazil P Colrain (Rio)
- CERN J Harvey
- France A Tsaregorodtsev (Marseille)
- Germany M Schmelling (MPI Heidelberg)
- Italy D Galli, U Marconi (Bologna)
- Holland M Merk (NIKHEF)
- Poland M Witek (Cracow)
- Russia I Belyaev (ITEP)
- Spain B.Adeva (Santiago)
- Switzerland P Bartalini (Lausanne)

- UK AHalley (Glasgow), TBowcock (Liverpool)

To be added....

A Spreadsheets showing Computing Requirements

to be added

References

1 Models of Networked Analysis at Regional Centres for LHC Experiments, Phase 2 Report, MONARC collaboration