Regional research exploitation of the LHC: a case-study of the required computing resources

S. Almehed, Ch. Driouichi, P. Eerola, U. Mjörnmark,
O. Smirnova, Ch. Zacharatou Jarlskog, T. Åkesson

*Particle Physics, Institute of Physics, Lund University,
Box 118, 22100 Lund, Sweden*

Abstract

A simulation study to evaluate the required computing resources for a research exploitation of the Large Hadron Collider (LHC) has been performed. The evaluation was done as a case study, assuming existence of a Nordic regional centre and using the requirements for performing a specific physics analysis as a yard-stick. Other input parameters were: assumption for the distribution of researchers at the institutions involved, an analysis model, and two different functional structures of the computing resources.

Key words: MONARC project, regional centers, LHC
PACS: 07.05.T

1 Introduction

1.1 Physics data at the LHC

The Large Hadron Collider (LHC) is the world's biggest accelerator, being built at the European Particle Physics Laboratory, CERN. By the time of its completion in 2006, it will be capable of accelerating and colliding beams of protons at centre of mass energies of 14 TeV. The collider ring will be equipped with several experimental installations, dedicated to studies of various physics phenomena. These detector installations are prepared by international collaborations, involving thousands of researchers from hundreds of institutions, distributed worldwide.

Physics analysis performed at the LHC will involve inclusive and exclusive measurements of various observables, related to the proton-proton collision events. To fulfil such tasks, not only event reconstruction based on the electronic signals should be done, but also a full computer modelling, including event generation, detector simulation, and subsequent analysis and evaluation of acceptance, inefficiency and other corrections should be completed.

LHC experiments expect to have a recorded raw data rate of about 1 PetaByte per year (or 100 MByte/s during running) at the beginning of the LHC operation (see *e.g.* ref. [1]). It is expected that the raw and processed data of the experiments will approach 100 PetaBytes by the year 2010. The geographical spread of the collaboration increases the complexity of the access and analysis of the data. The network bandwidths may constitute a further limitation.

1.2 Analysis of LHC data: physics case-study

One of the important measurements at LHC will be that of the parameter $\sin(2\beta)$, where the angle β is an angle of the CKM unitarity triangle describing quark mixing. It is a central parameter to demonstrate CP violation in the B-meson system [2]. In order to measure this parameter one needs to reconstruct and tag decays $\mathrm{B^0_d} \to \mathrm{J}/\psi\mathrm{K^0_S}$. In addition to the signal events, one also needs to analyze control samples ($\mathrm{J}/\psi\mathrm{K}^*$ and $\mathrm{J}/\psi\mathrm{K}^+$ events). It is assumed here that the measurements will be performed by groups in Copenhagen, Oslo, Bergen and Lund, which will consist of five researchers each. They will be accessing the same data set to do different parts of the analysis: study of signal events, background analysis to define the corresponding dilution factor, and analysis of control samples to define the dilution factor from tagging.

In the ATLAS detector [1], these events are triggered at the low-luminosity running of LHC with the level-1 muon trigger [3] (muons are triggered if their transverse momentum p_{T} exceeds 6 GeV and the pseudorapidity range is within $|\eta| \leq 2.4$). The trigger muons are reprocessed by the level-2 trigger system. It is assumed here that the level-2 single muon output rate is about 4.5 kHz, which requires track element matching between the two detector subsystems: the muon spectrometer and the inner detector. The level-2 trigger then continues the event processing by searching for a second muon ($p_T >$ 5 GeV), or two electrons ($p_T >$ 1 GeV) in addition to the level-1 trigger muon.

The high-level trigger, the event filter, consists of full event processing with off-line type software. The final event filter output rate of direct J/ψ decaying to $\mu^+\mu^-$ was estimated to be 5 Hz [1], whereas the output rate for $\mathrm{B} \to \mathrm{J}/\psi$ decaying to muons was estimated to be 3 Hz. This gives a total of 8 Hz for J/ψ

decaying to muons. The total rate of J/ψ decaying to e^+e^- was estimated to be 28 Hz [3], assuming that the event filter reconstruction included secondary vertex cuts. Thus, in one day of data-taking, there will be approximately $7{\cdot}10^5$ di-muon events and $2.4{\cdot}10^6$ di-electron events, i.e. a total of $3.1{\cdot}10^6$ di-lepton events. In the following, it is assumed that these events will be written out in a separate raw data stream.

To perform the analysis, one needs big sets of computer-generated events, processed through a detailed detector simulation ("fully simulated" data sample), in order to define trigger and reconstruction efficiencies and to correct for detector-dependent effects. It was assumed here that the required simulated sample size is twice the real data size, that is about 6 million events per day. Furthermore, analysis experience shows that events simulated with a parameterized detector simulation ("fast simulation") are also needed in the final analysis for estimating systematic uncertainties. Here it was assumed that the number of fast simulation events was also twice the number of real data events.

1.3 Evaluation of the computing resources

In this paper, a simulation study to evaluate the required computing resources in the Nordic countries for a research exploitation of the LHC has been performed, using the measurement of $\sin(2\beta)$ as one of the several physics cases. It was also assumed that this represents, computing-wise, a typical case not only for the high-energy physics, but also for other sciences. The purpose was to model a regional centre residing in the Nordic countries and serving the local high-energy physics groups, as well as other scientific communities. The paper is organized as follows: in Section 2, the data processing and physics analysis models are described. In Section 3, the simulation of the data processing (both for real and computer-generated data) is described. The processing of the physics data at CERN was considered first (Section 3.1). Then, the analysis of the data at the Nordic regional centre (assumed to be in a so-called Tier2 configuration, see Section 3) was modelled, including also the production of fast-simulated events (Section 3.2). Data transfer between the centres also was taken into account. Finally, the case in which the Nordic regional centre would have a Tier1 status, with more capacity than Tier2, was considered. In this case, the regional centre would also produce full simulation data. The conclusions are drawn in Section 4.

2 Data processing and physics analysis models

2.1 Generic data processing model

In order to cope with the LHC data analysis and storage requirements, a tiered hierarchy of distributed regional centres was proposed by the MONARC project (Models of Networked Analysis at regional centres for LHC Experiments) [4]. In this scheme, the main centre is CERN (Tier0). It is anticipated that the data reconstruction, *i.e.* the processing of the full online data into the analysis object data and those used for event tagging, will be performed at the Tier0. The Tier1 centres are the regional centres with a capacity next largest to CERN. Among the possible activities of the Tier1's, the production and reconstruction of fully-simulated data requires significant resources. Reconstruction of the raw online data could also be partially performed at Tier1 centres. Another function of those centres will be mirroring of the central database of Tier0. The data analysis and fast simulation will be mainly the responsibility of the Tier2 centres of a smaller capacity. Computer farms at institutions and workstations constitute lower tiers. The aim of such a distributed computing structure is to optimize access to the data by end-users, to reduce the workload on the central computing and data handling facilities, and to use efficiently the network bandwidth.

Data recorded directly from the online stream, including the signals from the detector elements and the on-line reconstruction results (called "raw data" or RAW), are expected to reside at CERN, being stored on a mass storage. During the processing at this Tier0 centre, the raw data will first be run through a reconstruction program, which calculates charged particle trajectories and energy depositions in the calorimeters. The reconstruction results are called ESD (Event Summary Data). Further processing algorithms are used at the Tier0 to prepare AOD (Analysis Object Data), which contains reduced information from ESD, and "tag data" or TAG, which is a small set of variables describing the event, such as jet and lepton multiplicities, transverse energy of the most energetic jets and transverse momentum of the most energetic leptons. The information in the TAG data set is meant to be used for initial selection of the AOD data to be analyzed.

The size of these data types per event is expected to be 1 MB for RAW, 0.1 MB for ESD, 0.01 MB for AOD and 0.001 MB for TAG. Data are organized in objects and manipulated by object databases (Objectivity/DB federated databases [5]). The databases are managed by database servers (Objectivity AMS servers). Pointers allow navigation from one object to another across the database.

The MONARC project developed a simulation tool [6] to model various configurations of regional centres. It allows the determination of the optimal resources and strategies needed to achieve the highest efficiency of tasks, performed by users. The MONARC simulation program is built with Java$^{(TM)}$ technology, which has built-in multi-thread support for concurrent processing, particularly suitable for the simulation purposes. The simulation engine provides a dedicated scheduling mechanism, and is designed to be generic for any distributed system. The major components are: the Data Model (based on the Objectivity/DB architecture and the basic object data design used in high-energy physics), the Multitasking Data Processing Model, Network Model and the Arrival Patterns – the mechanism to define the stochastic process of submitting jobs.

2.2 Use-case description

A specific measurement, that of the CP-violation parameter $\sin(2\beta)$ (see Section 1.2), was used as the physics test-case. The simulations addressed the processing and analysis required for one day of data-taking, which includes:

- reconstruction of RAW data: production of ESD, AOD and TAG data at Tier0 (CERN);
- analysis of AOD data at a Nordic Tier2 (or Tier1);
- fast simulation of AOD data at a Nordic Tier2 (or Tier1);
- full simulation of RAW data at a Nordic Tier1;
- reconstruction of the fully simulated RAW data at a Nordic Tier1.

The amount of data was assumed to correspond to one day of data-taking, *i.e.* about 3 million real data events and 6 million fully-simulated events (see Section 1.2).

It was assumed that this specific analysis will be conducted by four experimental groups in the Nordic countries: at the Niels Bohr Institute (NBI) in Copenhagen, at the University of Oslo, at the University of Bergen and at the University of Lund. All the CPU power was placed in NBI, which was considered both in a Tier1 and in a Tier2 configuration (Fig. 1). The other three institutes represent users of the computing power of NBI. When the NBI was considered to be a Tier2 centre, the full simulation was assumed to be performed in three Tier1 centres, producing two million events each. It was assumed here that those Tier1 centres would be located in UK, France and CERN. When the NBI was considered to be a Tier1 centre, the full simulation was assumed to be performed in NBI.

It should be noted that the modelling was independent of the geographical location of the institutes. The given set of institutes was used only to give

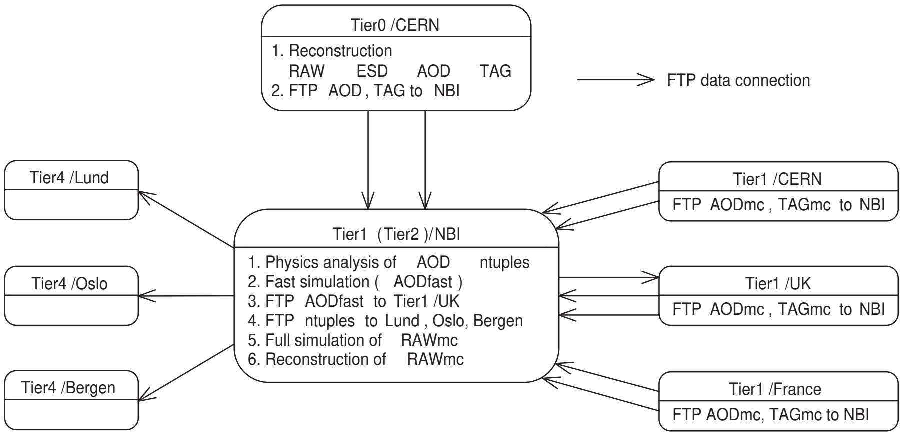

Fig. 1. The example used in the simulation. If NBI is a Tier2, it is assumed that the full simulation results (AOD, TAG) are transferred to the centre from three Tier1 centres (in France, at CERN and in UK)

a concrete example of the data processing and simulation chain, using one specific physics analysis as a test-case. In the Nordic countries there are more than the named four experimental groups, and these groups will wish to analyze different data samples to perform different analyses. Moreover, a Regional Computing Centre is supposed to serve not only high-energy physics, but other researchers as well. Therefore, this study is not meant to be a complete mapping of the required computing resources in the Nordic countries, but rather an example of the user requirements.

3 Modelling of the computing resources

As was stated above, computing resources necessary to process RAW data from the output stream at CERN, and to further analyze them elsewhere, were modelled using the MONARC simulation tool. Modelling was performed for the Tier0 at CERN, and for the regional centre assumed to be at NBI (both in Tier2 and Tier1 configurations).

3.1 Reconstruction at the Tier0 (CERN)

In order to evaluate the time needed to reconstruct 3.1 million of RAW events at CERN, the whole batch was split into sub-jobs (a sub-job being a task running on a single node), and simulation runs were performed by varying the number of sub-jobs for the reconstruction chain. The jobs for the creation of ESD, AOD and TAG were made sequential in the simulation. Table 1 contains the parameters used in the modelling of the Tier0 centre. The size of one data event was assumed as in Section 2.1. Due to the division into sub-jobs, the total number of events processed in the simulation was not always equal to 9.3 million events (where all types of data are taken into account). For this reason, the equivalent execution time was calculated by dividing the 9.3 million events by the processing rate given by each simulation run. This time is shown in Fig. 2.

Table 1
Modelling parameters for the Tier0 study.

Database link speed	100 MB/s
Node link speed	10 MB/s
Database read speed	20 MB/s
Database write speed	15 MB/s
Nr. of nodes (single-processor)	3000
CPU per node	200 SPECint95
Memory per node	512 MB
Processing time for RAW	350 SPECint95·s/event
Processing time for ESD	2.5 SPECint95·s/event
Processing time for AOD	0.5 SPECint95·s/event
Nr. of database servers for RAW data	10
Nr. of database servers for ESD data	5
Nr. of database servers for AOD data	2
Nr. of database servers for TAG data	2

As Fig. 2 suggests, the reconstruction task for the given channel at CERN can be performed well within one day. It is clear from the same plot that the number of sub-jobs can be optimized.

It was assumed that the AOD and TAG data produced at CERN are subsequently transferred to NBI (Tier2 or Tier1). As will be shown in the next section, the issue of whether this transfer is direct or proceeds through another Tier1 centre (in the case of NBI being a Tier2) is not of critical nature, as the time of a transfer via an FTP protocol is expected to be rather small.

3.2 Tier2 study

After the data for the given analysis channel have been processed (reconstructed) at CERN, they are to be used for the analysis by the researchers –

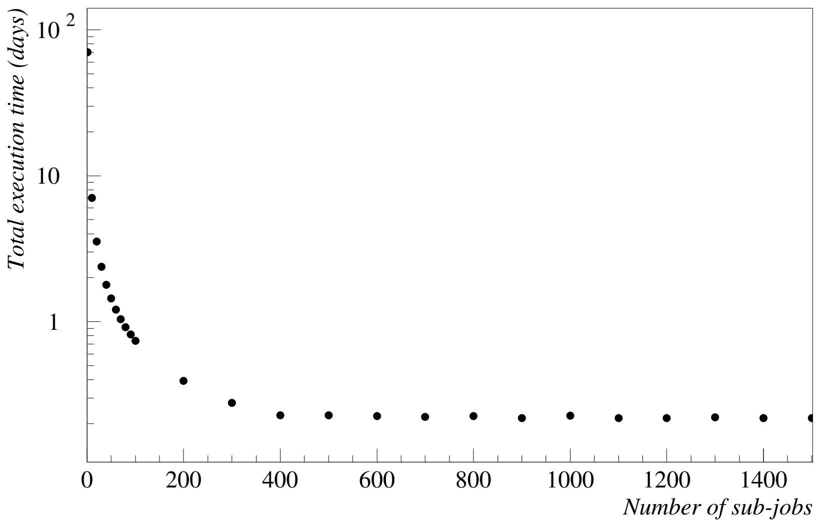

Fig. 2. Total execution time for the reconstruction of the data at CERN.

in this case, by the researchers in the four institutes in the Nordic countries. The data and computing power are concentrated at the regional centre, which in this study was located at NBI.

The regional centre at the NBI was first considered to be a Tier2 centre. The modelled activities in the centre were:

– analysis of 3 million AOD events (coming from CERN) and
– fast simulation of 6 million AOD events.

As specified in the Section 1.2, analysis of AOD data was assumed to be performed by twenty researchers (five per institute), each analyzing the complete sample once in a number of sub-jobs of equal number of events. The number of sub-jobs per person was varied as follows: 1, 5, 10, 15, 20 and 40. The number of nodes at the Tier2 was assumed to be 100, 200, 300, 400 and 800. Nodes were assumed to be single-processor, of 200 SPECint95 each, with memory of 512 MB per node. All the speeds (database link, node link and the database read/write) in the simulation were set to 125 MB/s. Two database servers for the AOD data were used. The output of the simulation was stored also in two database servers. It was assumed that the output for each event was a small number of real or integer values characterizing the event – invariant mass, decay time, flavour tag *etc.* – which were written out into an HBOOK ntuple [8]. Assuming that the output consisted of fifteen real numbers and five integer numbers, the output size was estimated at 140 bytes per event. The time to analyze one event was assumed to be 3 SPECint95·s.

The execution time as a function of the number of sub-jobs and for different numbers of nodes is shown in Fig. 3. It can be seen from the figures that the optimal number of sub-jobs is equal to the number of nodes. For 200 nodes and 10 sub-jobs per researcher (*i.e.*, total of 200 jobs), the execution time is 2 hours 38 minutes. For 300 nodes and 15 jobs per researcher, this time is reduced to 1 hour 52 minutes. The overall conclusion from the Fig. 3 is that the optimum combination is to have as many jobs as nodes and that a centre with 200 nodes would seem to be well suited for the task.

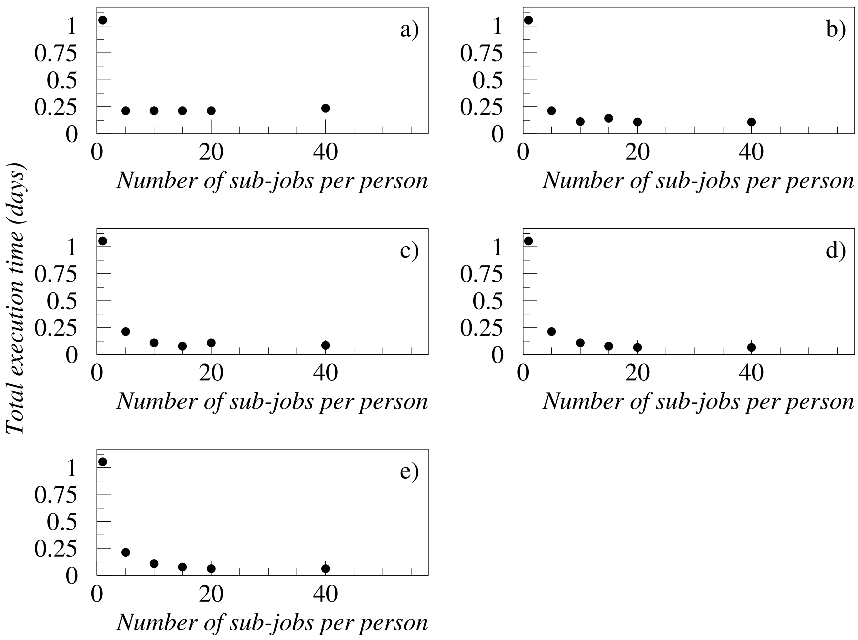

Fig. 3. Total execution time for the analysis of AOD data at Tier2, configured with different amount of nodes: a) 100 nodes, b) 200 nodes, c) 300 nodes, d) 400 nodes and e) 800 nodes.

For the fast simulation, the link and read/write speeds and node parameters (number of nodes, CPU capacity and memory) were the same as in the analysis simulation. The simulation of 6 million events was assumed to be performed by one operator once. The number of sub-jobs was varied as follows: 50, 100, 200, 400, 600 and 800. The size of the simulated data was assumed to be the same as that of the real data. It was calculated that the extra time to write events of twice this size to the databases would be of the order of a few minutes and could therefore be neglected. Four database servers for simulated AOD data were assumed. The time to generate one event was estimated to be 70 SPECint95·s. The execution time as a function of the number of sub-jobs and for different numbers of nodes is given in Fig. 4. The generation time for 200 nodes and 200 jobs was 2 hours 58 minutes, whereas for 400 nodes and jobs, it was 1 hour 30 minutes. This confirms the previous result that the optimal amount of jobs should be equal to the number of nodes, and that a

200-node centre is sufficient to perform the task.

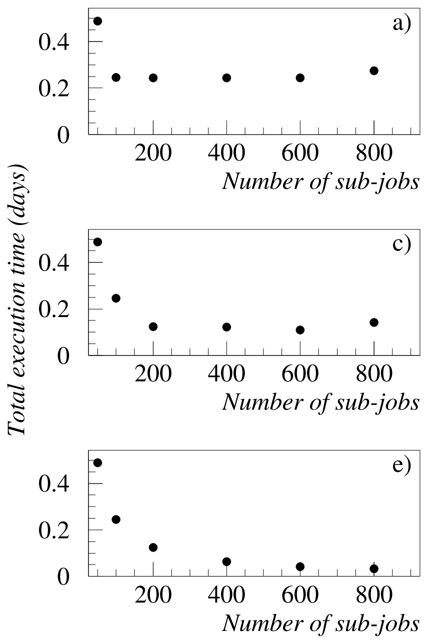

Fig. 4. Total execution time for the fast simulation of AOD data at Tier2, configured with different amount of nodes: a) 100 nodes, b) 200 nodes, c) 300 nodes, d) 400 nodes and e) 800 nodes.

The transfers (via FTP) from and to the Tier2 were estimated as follows. The output files of the AOD analysis were transferred from the NBI Tier2 to Oslo, Bergen and Lund (total of 20 users $\times$ 3 million events $\times$ 140 bytes per event). The transfers into the NBI Tier2 refer to the AOD and TAG data coming from CERN (3 million events at once) and the full simulated AOD and TAG data coming from three Tier1's. It was assumed that each transfer from a Tier1 replicates 4 million events in order to take into account the real size of simulated data (twice larger than the real data). The WAN speed was varied as: 125 MB/s, 12.5 MB/s and 0.04 MB/s. The transfer times are given in Table 2. The table implies that the transfer time can be neglected assuming a WAN speed of 125 MB/s (1 Gbit/s) or more.

Table 2
Data transfer times to and from the Tier2.

WAN speed	Data transfer time		
(MB/s)	from NBI	to NBI from CERN	to NBI from a Tier1
125	50s	4m	5m20 s
12.5	8m 24s	40m	53m 20s
0.04	1d 20h	8d 16h	11d 14h

To summarize, one can observe that for our physics test-case, a centre of 200 nodes with 200 SPECint95 per node would be quite efficient for the analysis of AOD data and the fast simulation of AOD data. The execution time for both tasks would be approximately 5.5 hours (for 200 jobs running simultaneously). The data transfer time can be neglected.

3.3 Tier1 study

The general configuration of the Tier1 regional centre at NBI included the following parameters: link and read/write speeds were set to 160 MB/s and all nodes had a memory of 1024 MB. The number of nodes was varied for the full simulation of RAW data, but was fixed for other activities. The CPU per node was set to either 200 SPECint95 or 500 SPECint95. The number of jobs running was optimized and always taken to be equal to the number of nodes. The following activities were simulated: analysis of AOD data, fast simulation of AOD data, and full simulation and reconstruction of RAW data.

Analysis of AOD data was simulated for the same amount of data and research groups as in the Tier2 study, but with the faster speeds, increased memory per node and increased CPU capacity per node (as described above). Two cases were studied, both for 200 nodes and 200 jobs. In the first case, the nodes had 200 SPECint95 of CPU each, whereas in the second case, the CPU per node was assumed to be 500 SPECint95. In the first case, the improvement from the Tier2 performance was reflected in the memory per node (1024 MB instead of 512 MB) and the various speeds (160 MB/s instead of 125 MB/s). In the second case, the CPU was also increased to 500 SPECint95 per node. The results are shown in Table 3, compared to the Tier2 case. It can be seen that the execution times are almost unaffected by the memory and speeds improvement, but are reduced significantly by increasing CPU speeds.

Table 3
Comparison of execution times at Tier2 and Tier1 in different configurations: case (I) corresponds to 200 SPECint95 per node, and case (II) to 500 SPECint95 per node.

	analysis, 3 mln. events, 200 jobs & nodes	fast simulation, 6 mln. events, 200 jobs & nodes	reconstruction, 6 mln. events, 500 jobs & nodes
Tier2	2h 38m	2h 58m	–
Tier1 (I)	2h 36m	2h 57m	6h 23m
Tier1 (II)	1h 08m	1h 12m	2h 52m

Fast simulation of AOD data was evaluated for the two cases, as for the AOD analysis. In the first case, there were 200 jobs running on 200 nodes of 200 SPECint95 each, and in the second case, the same amount of jobs were executed on 200 nodes of 500 SPECint95 each. The execution times are summarized in Table 3, and exibit the same trend when compared to Tier2 performance: while virtually unaffected by the memory and speeds improvement, the simulation runs much faster when CPU is changed from 200 to 500 SPECint95.

For the evaluation of reconstruction of fully simulated data, 500 jobs on 500 nodes of 200 (500) SPECint95 per node were considered. The reconstruction of 6 million RAW generated events, which were read from 10 database servers, was simulated. The total reconstruction times (for ESD, AOD and TAG creation) are shown in Table 3. It is clear that, although reconstruction requires more time that analysis and fast simulation, it is still well contained in one day.

Full simulation of RAW data is the most demanding task as far as CPU time is concerned: 3600 SPECint95·s per event were required [9]. The number of nodes and jobs assumed the following values: 200, 500, 600, 700, 800 and 900. The CPU per node was assumed to be either 200 SPECint95 or 500 SPECint95. Generation of 6 million events, written to 10 database servers, was simulated. The execution time as a function of the number of jobs (nodes) and the CPU per node is given in Fig. 5. The best estimate for the execution time was 15 hours 18 minutes for a centre with 900 nodes of 500 SPECint95 per node.

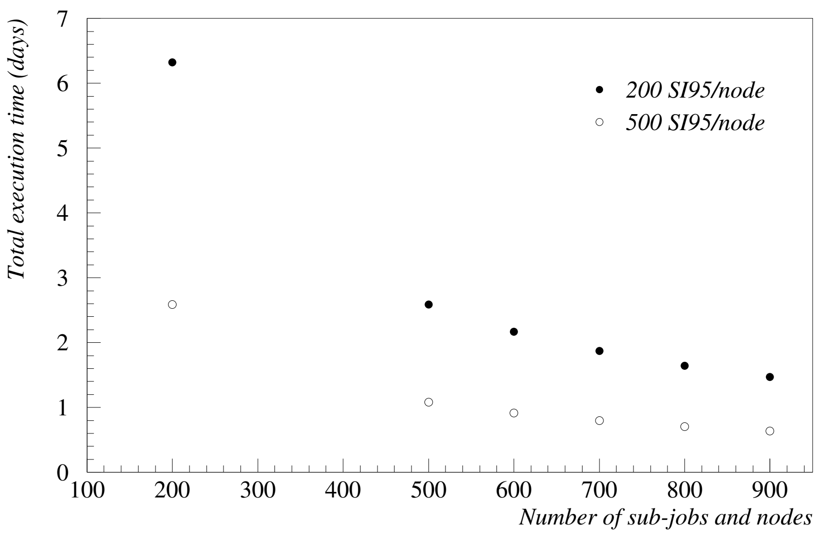

Fig. 5. Execution time for full simulation at the Tier1.

The overall conclusion for the Tier1 study is that all activities can be performed within one day. With 500 SPECint95 per node, one would need:

– 1 hour 8 minutes for AOD analysis (on 200 nodes)

– 1 hour 12 minutes for AOD fast simulation (on 200 nodes)
– 2 hours 52 minutes for reconstruction of fully simulated events (on 500 nodes)
– 15 hours and 18 minutes for generation of fully simulated RAW events (on 900 nodes).

4 Conclusions

The aim of this study was to evaluate the amount of computing resources needed to perform a particular physics analysis task at a future LHC experiment by several groups of researchers in the Nordic countries. The task in question was the measurement of CP violation in decays $\mathrm{B}^0_\mathrm{d} \rightarrow \mathrm{J}/\psi \mathrm{K}^0_\mathrm{S}$. The objective was to perform all the analysis "on-fly", which implied that the data acquired in one day should be immediately processed and analyzed. Processing included not only event reconstruction, but also corresponding modelling using particle generators and full (or simplified) detector simulation.

The required capacities of the Tier0 centre (at CERN) and a Nordic regional centre were investigated. By assuming the CERN capacity as having 3000 single-processor nodes with 200 SPECint95 per node, it was found that it is not only sufficient for performing the task in question, but can accommodate many more jobs. For a more efficient performance, the number of submitted jobs can be optimized. Two configurations of the Nordic regional centre were considered: a Tier2 or a Tier1 centre. It was shown, that after optimizing the number of submitted jobs, a Tier2 centre is sufficient to perform all the necessary tasks, including data analysis and fast simulation of events, with a rate faster than the data production rate at the LHC. In the Tier1 configuration the same tasks can be performed much faster. Assuming that the Nordic Tier1 regional centre will consist of 900 nodes equivalent to 500 SPECint95 each, the full-scale detector simulation would take approximately 16 hours.

The present analysis concerns only one particular high-energy physics task, while a regional centre will serve many other research groups, requiring different data sets and analysis procedures, not only in physics, but also in other sciences, like biology, medicine, Earth observation and so on. Therefore, the results have to be considered as a single typical use-case, one of many at a future Nordic Regional Computing Centre.

Acknowledgements

We would like to thank members of the MONARC team for their assistance, particularly, Krzysztof Sliwa for practical help in running the simulations, and Laura Perini and Iosif Legrand for discussions and suggestions. We also wish to thank Norman McCubbin and Pavel Nevski for valuable help with the execution times, Maria Smizanska for clarifying questions regarding the datacards and Tim Smith for his help with CPU estimations.

References

[1] ATLAS Collaboration, ATLAS Detector and Physics Performance Technical Design Report, CERN/LHCC/99-14. ATLAS TDR 14 (May 1999), Vol. I, p. 399.

[2] For a review, see Y. Nir and H. Quinn in: B Decays, ed. S. Stone (World Scientific, 1994) p. 362, or Ann. Rev. Nucl. and Part. Sci. 42 (1992) 211.

[3] ATLAS Collaboration, ATLAS High-Level Triggers, DAQ and DCS Technical Proposal, CERN/LHCC/2000-17 (March 2000) p. 125.

[4] MONARC Collaboration, Models of Networked Analysis at Regional Centres for LHC experiments (MONARC), Mid-project progress report, LCB 99-5.

[5] Objectivity Database Systems, http://www.objectivity.com.

[6] MONARC Collaboration, Multi-threaded, discrete event simulation of distributed computing systems, presented by I. Legrand in CHEP2000, to be published in CPC Journal special edition CHEP2000.

[7] Standard Performance Evaluation Corporation, http://www.spec.org.

[8] CERN Program Library Long Writeup Y250, HBOOK – Statistical Analysis and Histogramming, CERN, Geneva, 1994.

[9] Ch. Driouichi, P. Eerola, Ch. Zacharatou Jarlskog, Execution times for B-physics simulation, LUNFD6/(NFFL-7206) 2001, ATL-COM-SOFT-2001-009.