LONG-TERM EXPERIENCE WITH GRID-BASED MONTE CARLO MASS PRODUCTION FOR THE ZEUS EXPERIMENT

H. Stadie*, M. Ernst, R. Mankel, K. Wrona, DESY, Hamburg, Germany
J. Ferrando, University of Glasgow, United Kingdom

Abstract

The detector and collider upgrades for HERA-II have considerably increased the demand on computing resources for Monte Carlo production for the ZEUS experiment. In order to close the gap, the existing production system was extended to be able to use Grid resources as well. This integrated system has been successfully in production since November 2004. A total of 400 million events have been simulated by forty different Grid sites within 15 months exceeding the capacity of the old system. We present the production setup and and its implementation which is based on the ZEUS Grid-toolkit. Our setup includes an elaborate system to monitor the participating sites and every submitted Grid job.

INTRODUCTION

The ZEUS experiment is a multi-purpose high-energy physics detector at the HERA collider. HERA is an electron-proton collider located at the DESY laboratory in Hamburg, Germany. The ZEUS detector recorded its first collisions in 1992. After the luminosity enhancement of the collider, subsequently named HERA-II, the considerably upgraded ZEUS detector started data taking in 2002.

A large amount of simulated data is needed to analyze the events taken by the ZEUS experiment. This Monte Carlo simulation is very compute-intensive.

HERA-I Monte Carlo production system

A distributed Monte Carlo production system had been developed by the ZEUS collaboration already in the early 1990s to ensure adequate production capacity. This system, called *Funnel* [1], manages all Monte Carlo requests from the physicists centrally and distributes them to compute farms at the collaborating institutes. A "simulation request" consists of the input file for the detector simulation, which has been produced using a high-energy physics generator, and the selected simulation version. Three executables are run to complete a request: the detector simulation, the trigger simulation, and the reconstruction program. Its output files are centrally stored in the DESY tape systems. Based on the production in 2004, this production system reached a capacity of 330 million HERA-I events per year using around 200 CPUs worldwide.

* Hartmut.Stadie@desy.de

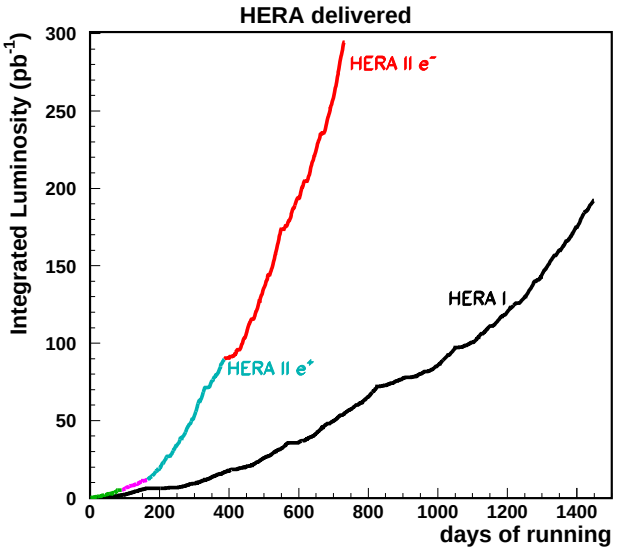

Figure 1: Integrated luminosity over the number of running days for the HERA-I and the HERA-II running. The curve for HERA-II has a much steeper slope due to the luminosity upgrade.

Monte Carlo demand for the HERA-II running period

The computing needs for Monte Carlo production are largely increased by the upgrades of the detector and the luminosity upgrade of the collider. Figure 1 compares the HERA-I and HERA-II running and shows the much higher luminosity for HERA-II, which increases the number of Monte Carlo events needed. According to an evaluation based on the expected integrated luminosity, the demand for Monte Carlo might reach as high as 700 million HERA-II events per year in 2007.

Furthermore, the CPU time for simulation and reconstruction of a HERA-II event is larger compared to HERA-I by approximately a factor of two due to the additional detector components. This effectively halves the capacity of the HERA-I-style production system. Therefore, a more powerful Monte Carlo processing system is vital to meet the requirements of ZEUS physics analyses.

THE INTEGRATED MONTE CARLO PRODUCTION SYSTEM

We have developed a new production system to fulfill the Monte Carlo production requirements of the ZEUS col-

laboration by integration of Grid resources [2]. Through its member institutes, the ZEUS collaboration has access to many evolving computing resources that are accessible via Grid technology. The access is based on the middleware used for either the LCG/EGEE grid [3] or the Grid3/OSG [4] infrastructure.

In addition, the DESY IT group have set up Grid infrastructure [5, 6] which serves as the hub of the ZEUS Grid activities [7]. This includes the hosting of the ZEUS virtual organization, operation of a resource broker, a powerful storage element, and local computing resources. Of similar importance for this project has been the acquirement of Grid expertise and the support for our project by the DESY IT Grid group.

Design goals

The following criteria were set for the design of the integrated production system:

- The existing resources outside of the Grid should still be usable.
- The inclusion of Grid resources should be completely transparent to the physicists requesting a Monte Carlo sample. A request is formulated in the same way as before and the same bookkeeping information about the requests is returned. Output files created with the new system can be accessed in the same way as for the files produced with the old system.
- The production system should run almost fully automatically.
- ZEUS Grid jobs should run without pre-installed software. This enables us to add new sites quickly and use resources opportunistically.
- The Grid jobs should be able to run on different middleware flavors. The Grid middleware is evolving rapidly. In fact, completely new projects are emerging and even the behavior of existing client tools changes from release to release. Therefore, our system should allow us to add the support for a new middleware version easily.

The ZEUS Grid-Toolkit

The ZEUS Grid-toolkit [8] is the basic toolkit for the implementation of the new production system. It is written in object-oriented Perl and consists of a set of classes for basic data structures, job submission, data transfer, and output logging and validation. A main advantage of Perl is that the same toolkit can be used for both the production system and for the production scripts running on the worker nodes. The toolkit itself is completely independent of other ZEUS software.

A variety of client tools to access Grid services exist on different sites and new projects are being developed. To support different projects simultaneously, the ZEUS Grid-toolkit has been designed using the *Strategy pattern* and adds an additional layer of abstraction to the usage of Grid client tools. Abstract interfaces are defined for data handling and job operations, and the appropriate middleware implementation is chosen at run time based on a configuration file and the installed middleware packages. These implementations use fault tolerant methods, e.g. automatic retrial of failed data transfers, and have largely increased the overall job efficiency.

Layout of the integrated Monte Carlo production system

To include Grid resources in the Monte Carlo production system completely transparently for the users, the existing mature interfaces for the submission of a Monte Carlo request and querying its state are reused. As can be seen in Figure 2, a central scheduler distributes incoming Monte Carlo requests either to the traditional Monte Carlo sites or to a gateway to the Grid resources. This setup allows us to preserve the resources of the HERA-I production system and, in fact, to reuse most of the existing scheduler.

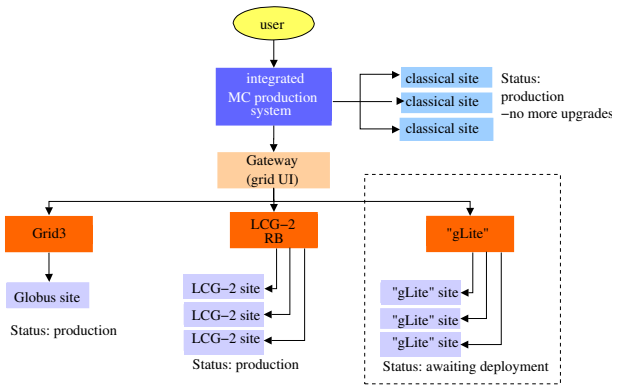

Figure 2: Layout of the integrated Monte Carlo production system. The “gLite” depicts a placeholder for any new middleware project.

The “gateway” node is a LCG-2 user interface and acts as a bridge between the production system and the Grid world. Cron jobs process the incoming requests and keep track of the individual Grid jobs. A database is used to store the state of the Monte Carlo requests and their associated Grid jobs. All the code is written in object-oriented Perl and the ZEUS Grid-toolkit is used to copy files to Grid and submit jobs.

For any new request, a cron job translates this request into a set of Grid jobs and copies the input file to the storage element at DESY. The jobs assigned to the LCG/EGEE sites are submitted using the resource broker at DESY. Since the ZEUS Grid-toolkit is able to simultaneously support different middleware projects, we were able to establish submission to a Grid3/OSG site using an implementation of our job submission interface based on the

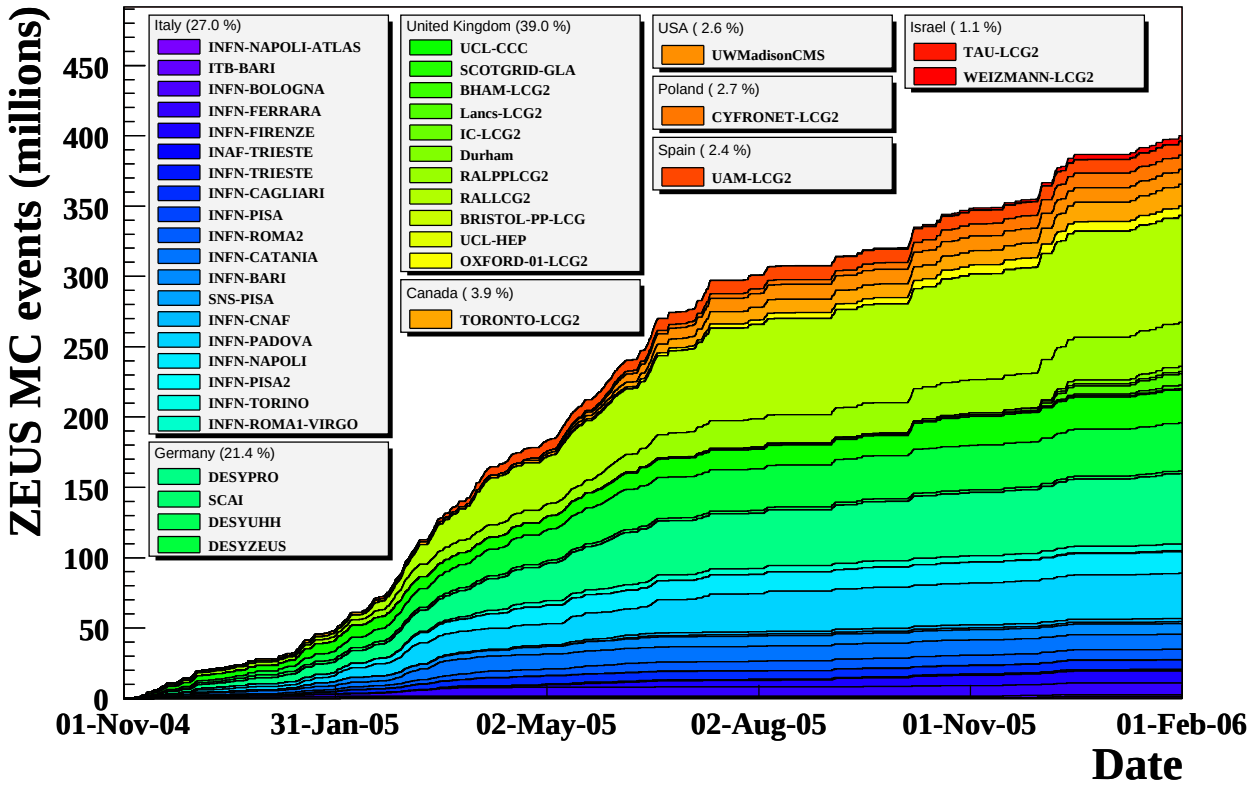

Figure 3: Integrated overall production of ZEUS Monte Carlo events on the Grid by sites.

Globus toolkit [9]. Every Grid job processes between 1000 and 2000 events which corresponds to a run time of around three hours. The status of the jobs is updated regularly. When a Grid job has finished, its output sandbox is retrieved and the log files are checked for error conditions. The result of these checks is stored in an additional database table. This database is used by our monitoring system and enables us to identify problems quickly. If the job has passed the check, the Monte Carlo output file is transferred from the DESY storage element to the final tape storage pool. As both systems use the dCache mass storage subsystem [10, 11], this a very fast operation. Grid jobs that encountered an error are automatically resubmitted. This guarantees an efficiency, defined as the number of produced events divided by the number of requested events, of very close to 100%.

When all Grid jobs for one request have finished, the gateway returns the request to the production system with all necessary bookkeeping files and an archive containing all log files of the individual jobs.

The Grid jobs

The Grid Jobs that simulate the Monte Carlo events get all the run scripts and the calibration constants from two archives that are copied from the local storage element to the worker node. As the three executables belonging to the requested Monte Carlo version are also copied from Grid storage, no pre-installed software is needed to run the production jobs. This allows us to add new Grid sites quickly to our production.

All commands on the worker node are run with an enforced time-out using a special class of the ZEUS Grid-toolkit to avoid run-away jobs. Run-away jobs would be killed by the local batch system and would return no output to the user making it impossible to identify the cause of the problem. All errors can be identified easily with this measure and the logging of all commands and data handling operations to standard output.

OPERATION OF THE GRID PRODUCTION

Figure 3 shows the integrated Grid production of ZEUS Monte Carlo events over time. As can be seen, the presented production system has been in continuous operation since November 2004 and used forty different Grid sites in eight different countries. Time periods with a small number of produced events are usually related to a temporary absence of Monte Carlo requests. This indicates that the overall capacity of the production system exceeds the demands.

Monitoring

Good monitoring is needed to use the Grid resources efficiently. Our monitoring system regularly contacts the information services of the participating sites. For every

queue, the number of free and used CPUs and the number of waiting and running jobs are stored with RRDtool, a system to store and display time-series data [12]. Furthermore, it queries our production database and stores for each queue the number of waiting and running jobs known to our production system.

A site with temporary problems might e.g. accept Grid jobs, but erroneously never submit them to its local batch system. This causes the site to always publish available resources and attract more and more jobs like a "black hole". This problem can be detected by comparing the number of scheduled and running jobs as reported by the site with the numbers from our production database that are obtained from the resource broker. The monitoring system offers this information via a web interface. A monitoring plot for a site that acted like a black hole is shown in Figure 4. Not a single job was running or being scheduled at this site although the resource broker had sent more than six hundreds jobs to this site. Although our production system cancels jobs that are not running after a certain amount of time, this problem could only be cured by manually removing the site temporarily from our production.

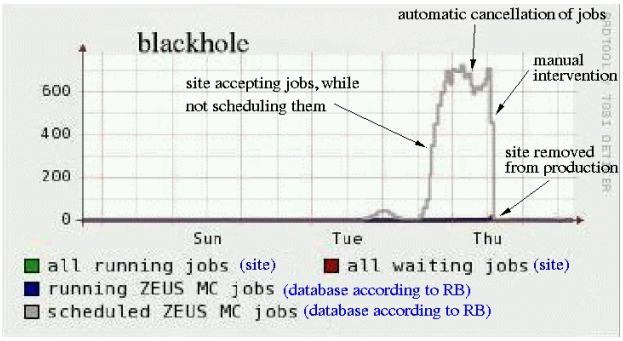

Figure 4: Example of a monitoring plot showing a site acting as a "black hole". The site is publishing free resources and is receiving Grid jobs without submitting them to its local batch system.

A further source of information for the monitoring system are the database entries for every finished Grid job. They contain, for example, the site to which the job was sent, the number of produced events, and the error message, in case of a failure. The failure modes and the production rate for every site can be studied separately with this information. In addition, one can view the Grid job efficiency for every site in real-time. A web page lists all errors that occurred in the last 24 hours.

STATUS AND CONCLUSIONS

The integrated ZEUS Monte Carlo production system that utilizes both the traditional Monte Carlo production resources and Grid resources started production in November 2004. Since then, 40 different LCG sites in eight different countries have contributed to the ZEUS Monte Carlo production. Biggest individual contributor has been the computing center at the Rutherford Appleton Laboratory in the United Kingdom, followed by the clusters at DESY in Hamburg, Germany, and at Padua in Italy. In May 2005, the Grid3 site of the University of Wisconsin in the USA has been added using the Globus implementation in the ZEUS Grid-toolkit, which highlights the flexibility of our design.

A total of 400 million Monte Carlo events have been simulated and reconstructed on the Grid until February 2006 and are now being used in the ongoing ZEUS physics analyses. The Grid production by now exceeds the number of events produced by traditional (non-Grid) sites. In fact, daily Grid production rates four times above the HERA-I Monte Carlo production rates were attained. This illustrates that the inclusion of Grid resources largely increased the Monte Carlo production capacity for ZEUS.

ACKNOWLEDGMENT

We would like to thank the administrators of all participating sites for their support.

REFERENCES

[1] B. Burow, "Funnel: Towards Comfortable Event Processing", proc. International Conference on Computing in High Energy Physics 1995, DESY-95-236D (1995).

[2] H. Stadie et al., "Monte Carlo Mass Production for the ZEUS Experiment on the Grid", Nucl. Instrum. Meth. A, article in press.

[3] LCG: `http://lcg.web.cern.ch/LCG/`.

[4] Open Science Grid: `http://www.opensciencegrid.org`.

[5] A. Gellrich et al., "Grid Technology in Production at DESY", Nucl. Instrum. Meth. A, article in press.

[6] A. Gellrich, "Deploying an LCG-2 Grid Infrastructure at DESY", these proceedings.

[7] M. Ernst, A. Gellrich and R. Mankel, "DESY becomes hub for Grid-based HERA events,", CERN Cour. 45N3 (2005) 19-20.

[8] K. Wrona et al., "The ZEUS Grid-Toolkit - an experiment independent layer to access Grid services", these proceedings.

[9] I. Foster, C. Kesselmann, "Globus: A Metacomputing Infrastructure Toolkit" Intl. J. Supercomputer Applications 11(2) (1997) 115-128.

[10] P. Fuhrmann, "dCache, LCG Storage Element and enhanced use cases", proc. International Conference on Computing in High Energy Physics 2004.

[11] dCache: `http://www.dcache.org/`.

[12] RRDtool: `http://people.ee.ethz.ch/~oetiker/webtools/rrdtool/`.