Proceedings of the Third Conference of the EELA Project
R. Gavela, B. Marechal, R. Barbera et al. (Eds.)
CIEMAT 2007

Experience with Large Scale Simulations on the EGEE Grid for the AUGER collaboration

Jaroslava Schovancová[1,*], Jiří Chudoba[1,2], František Dvořák[1], Jiří Filipovič[1], Jan Kmuníček[1], Aleš Křenek[1], Luděk Matyska[1], Miloš Mulač[1], Miroslav Ruda[1], Zdeněk Salvet[1], Jiří Sitera[1], Zdeněk Šustr[1], Petr Trávníček[2]

[1]*CESNET, z.s.p.o., Prague, Czech Republic*
[2]*Institute of Physics of the Academy of Sciences, Prague, Czech Republic*
[*] *E-mail: schovan@sirrah.troja.mff.cuni.cz*

Abstract

We share our experience with the Large Scale Monte Carlo Simulations using the CORSIKA simulation program performed by the VO AUGER users on the EGEE Grid environment. We report on the AUGER CPU Challenge performed in April 2007 as a test of availability of the VO AUGER dedicated resources. We developed a set of scripts for an easy handling of a Large Scale Simulations by a very small number of users. We show status of the AUGER Offline Production ran with the CORSIKA simulation program, where these scripts were used. We report our preliminary results with testing the Job Provenance as the long-term information storage.

1. Introduction

The Pierre Auger Cosmic Ray Observatory [1] is studying ultra-high energy cosmic rays, the most energetic and rarest of particles in the Universe. These highly energetic particles initialize extensive air showers while crossing the Earth atmosphere. CPU intensive Monte Carlo (MC) simulations are needed to compare predictions of different models with observed data. We run the MC Simulations using the EGEE Grid resources [2] accessible to the members of the Virtual Organization AUGER.

The Virtual Organization AUGER (VO AUGER) was created in 2006 by the Czech group [3] in cooperation with CESNET. CESNET provides and maintains central resources, such as LCG Resource Broker (LCG RB), gLite Workload Manager Service (WMS), Logging and Bookkeeping (LB), User Interface (UI), LCG File Catalog (LFC), registration portal and the Virtual Organization Membership Service server (VOMS server). At the present time, there are few tens of users registered to the VO AUGER.

We summarize results of the AUGER CPU Challenge in Section 2. We share experience with Large Scale MC simulations using the EGEE Grid environment in Section 3. We report preliminary test results of the Job Provenance for the AUGER experiment in Section 4.

2. AUGER CPU Challenge

In order to test the stability of sites and the reliability of the infrastructure and to get realistic numbers of CPUs available at the VO AUGER disposal, we ran the AUGER CPU Challenge in April 2007. We split this CPU Challenge into two phases: during the Phase 1 (April 13th to April 15th 2007) there was no notification to the involved sites, but we announced the Phase 2 (April 20th to April 22nd 2007) in advance. We used a program with many floating-point operations and almost no I/O operations as an etalon. This program consumes approximately 1 hour of CPU time, normalised to the 2 GHz CPU. During both phases of the CPU Challenge we submitted enough jobs to keep the queues full. All the jobs were submitted by a single user to common queues for the VO AUGER (as seen from the to common user's prospective), we did not use special queues dedicated just for production. Each Phase took roughly 58 hours (1 weekend per phase).

2.1. Phase 1 of the AUGER CPU Challenge, April 13th – 15th, 2007

There were 4 Computing Elements (CE) with AUGER dedicated queue engaged in the first phase of the CPU Challenge:

- golias25.farm.particle.cz:2119/jobmanager-lcgpbs-gridauger,
- grid10.lal.in2p3.fr:2119/jobmanager-pbs-auger,
- skurut17.cesnet.cz:2119/jobmanager-lcgpbs-auger, and
- lcgce.ijs.si:2119/jobmanager-pbs-auger.

We were able to run on average 90 concurrent jobs at the same time, the maximal count of concurrent running jobs was 160 – see Figure 1 for the temporal distribution of running jobs on involved CEs. There were 2152 Done jobs (80 % of all submitted jobs) at the end of the Phase 1.

However, we faced a VO-specific misconfiguration issue at the IJS site and with the WMS stability during this Phase. Fortunately, the misconfiguration was corrected and the IJS site was working properly during the second phase of CPU Challenge. Issue with the stability of the WMS was solved before the second Phase as well.

Despite the fact we have not announced the ongoing first phase of the CPU Challenge to the CEs in advance we were able to consume more than 2100 normalised CPU hours over one weekend. Thus, we have shown that the EGEE Grid environment is ready-to-use for the VO AUGER members.

2.2. Phase 2 of the AUGER CPU Challenge, April 20th – 22nd, 2007

The second phase of the AUGER CPU Challenge was announced to the CE maintainers after the successful end of the first Phase. We desired to fix VO-specific misconfiguration in order to perform as smooth run of the second Phase as possible. There were the same four sites involved in the Phase 2 of the AUGER CPU Challenge. We were able to run on average 100 concurrent jobs (220 at maximum) at the same time – see Figure 2 for the temporal distribution of running jobs. There were 2250 Done jobs (51 % of submitted jobs) at the end of the Phase 2.

In the present days (end of November 2007), there are resources of 8 queues at 7 sites in 5 countries at a disposal to the VO AUGER members:

- apcpc79.in2p3.fr:2119/jobmanager-pbs-auger,
- grid10.lal.in2p3.fr:2119/jobmanager-pbs-auger,

- ce02.lip.pt:2119/jobmanager-lcgsge-augergrid,
- tbn20.nikhef.nl:2119/jobmanager-pbs-qlong,
- tbn20.nikhef.nl:2119/jobmanager-pbs-qshort,
- skurut17.cesnet.cz:2119/jobmanager-lcgpbs-auger,
- golias25.farm.particle.cz:2119/jobmanager-lcgpbs-gridauger, and
- lcgce.ijs.si:2119/jobmanager-pbs-auger.

We are able to run around 120 concurrent jobs at the same time. New resources contributors are expected soon.

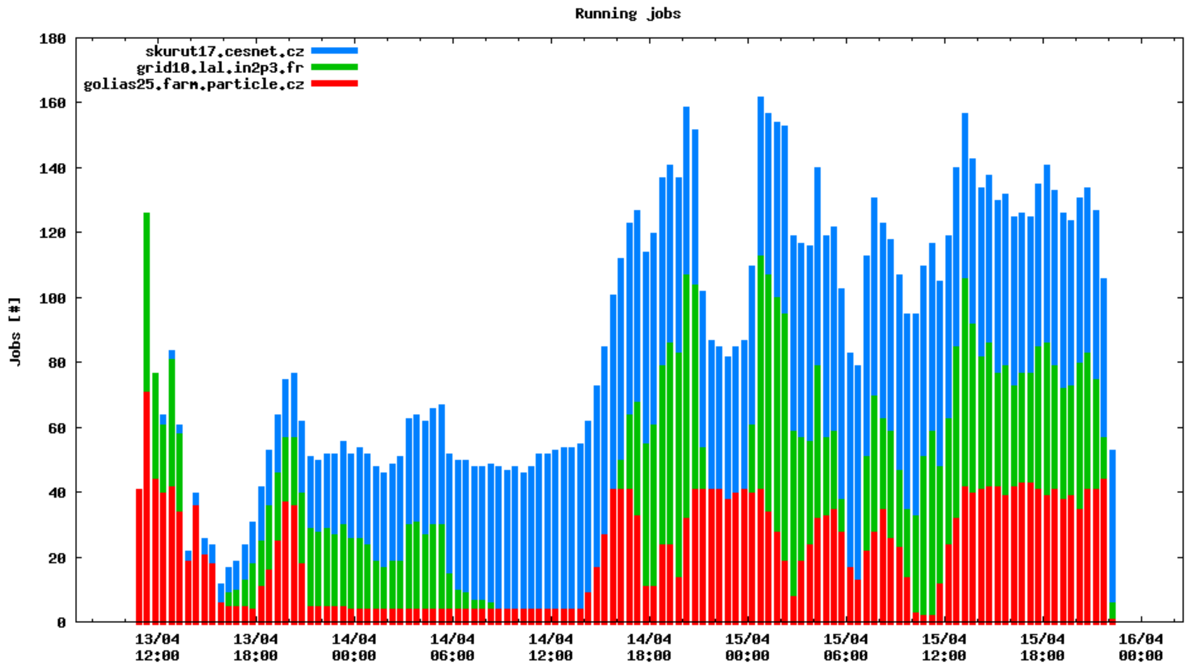


Figure 1: Number of Running jobs during the 1st phase of the AUGER CPU Challenge. Sum the CE CPU number contribution to get the number of the running jobs.

3. Monte Carlo Simulations

After the successful mapping of the reliability of the VO AUGER dedicated resources we managed to run the Offline Production using the CORSIKA simulation program (COsmic Ray SImulations for Kascade), [3]. We used the VO AUGER UI, where the LCG 2.7 and gLite 3.1 middlewares were available.

We developed a set of bash scripts for easy handling the Large Scale Simulations. We use these scripts for the job submission, resubmission of Aborted jobs (or jobs which seem to be Waiting for a long time due to the LB fallout) and OutputSandbox retrieval in a user-friendly way. The scripts can work with both mentioned middlewares, i.e. using the glite- (or glite-wms-) or edg- commands.

When a user running the CORSIKA production wants to submit a bunch of jobs, the only things he/she has to do are

1. obtain the copy of the scripts,
2. put the inputs for CORSIKA into a selected $WORKDIR,
3. initialise proxy (voms-proxy-init or grid-proxy-init),
4. type and confirm the following command

   ./Submit_corsika.6617.sh gliteWMS auger

For resubmission, user moves to the $WORKDIR, where the inputs of affected jobs are stored, and types and confirms the command

Resubmit_corsika.6617.sh gliteWMS auger

The OutputSandbox retrieval (or an attempt to retrieve) can be performed from the $WORKDIR at any time after the successful (re)submission of the jobs simply typing

GET_OUTPUT.sh gliteWMS

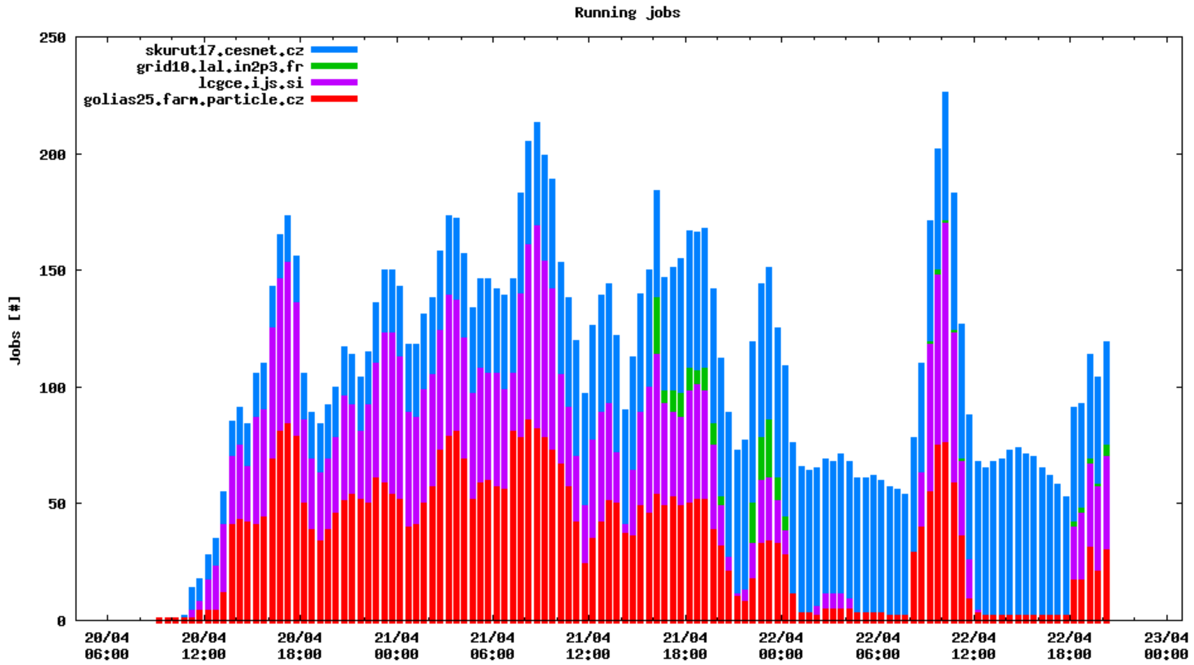


Figure 2: Number of Running jobs during the 2nd phase of the AUGER CPU Challenge.

3.1. Test Production (epos_gr01)

In order to get an impression about the duration of the CORSIKA Offline Production using the VO AUGER resources we ran the test production, epos_gr01, from August 28th to September 14th 2007. For the test production we used CORSIKA v. 6.611 with the EPOS and FLUKA (version Fluka2006.3b, [5, 6]) models. To be consistent with previous CORSIKA offline production done at a local farm at Leeds, we used a proton as a primary particle. We probed cosmic ray particles with 8 fixed energy bins of energy: log(E/eV)=17.5, 18.0, 18.5, 19.0, 19.5, 20., 20.5, 21.0. The zenith angle was explored in 7 fixed bins: theta=0, 18, 26, 38, 45, 53, 60 degrees. For each [energy, zenith angle] pair we ran a bunch of 50 jobs, each job had a different random number seed.

The generic values used in the CORSIKA input files are shown in the Table 1. For explanation of the various parameters please refer to the CORSIKA manual. The random numbers required for the Monte Carlo simulation were generated before the submission using the *srand(time(NULL))* and *rand()* functions implemented in the standard math.h library.

We used our scripts for the job submission, resubmission and OutputSandbox retrieval. The job outputs were uploaded to the Storage Element (SE). For this purpose we used SE golias100.farm.particle.cz. We set the VO AUGER policy for the SE usage. Every output file of the production should reside in the directory

/grid/auger/prod/PROD_ID/ENERGY/THETA,

Where the *PROD_ID* is the production identificator, e.g. epos_gr01, *ENERGY* is the energy tag, e.g. en21.000, and *THETA* is the zenith angle tag, e.g. th60.000. During the test phase we stored the whole gzipped and tarred job-working directory. However, we did not store the DATxxxxxx output file during the test. The whole packed directory took only about 250 MB.

The summary of job counts and CPU time consumption during the test phase of the production is shown in the Table 2. Our scripts enabled to perform such a Large Scale Production by a single user within 2 weeks. During these 2 weeks there have been submitted 2800 jobs and their OutputSandboxes have been retrieved. The test production consumed more than 28,000 hours of CPU walltime.

We have met several issues concerning the WMS stability. The WMS stability issue was resolved in matter of a few days. During these few days we used the LCG middleware instead of the gLite middleware, which was used in most cases during the testing phase.

3.2. Proton and Iron CORSIKA Production (epos_gr03, epos_gr04)

The experience gained during the testing phase of the CORSIKA Production turned to be very useful during its non-testing phase. We consider a proton (PRMPAR=14) and an iron atom (PRMPAR=5626) as the primary projectile particles. The differences in the CORSIKA input files are also shown in the Table 1. We use the AUGER SE policy for the data outputs storage. We store the binary data file produced by CORSIKA, the DATxxxxxx file (size up to 1 GB), the longitudes file DATxxxxxx.long (less than 50 kB), the MD5 checksum file (DATxxxxxx.md5.sum, order of kB), and the tarball DATxxxxxx_small.tar containing the CORSIKA input file DATxxxxxx.input, CORSIKA logs DATxxxxxx.lst and DATxxxxxx.tab, and the job stdout and stderr logs, the tarball has roughly 0.5 MB.

For the epos_gr03 and epos_gr04 phase we use CORSIKA v. 6.617 with EPOS and FLUKA (version Fluka 2006.3b.7, [5,6]) models. We use the same 8 energy bins and 7 zenith angle bins. For each multiple [primary projectile particle, energy, zenith angle] we submit a bunch of 50 jobs. Thus, both non-testing production phases will have 2800 jobs in total each. Summary of job distribution on CE is also shown in Table 2.

During this phase of the CORSIKA Production we have met several issues with the WMS instability, VOMS functionality, gLite/LB bugs and a FLUKA random seed-specific bug.

The WMS instabilities are usually solved “on demand” in a matter of few days. According to the VOMS functionality issue, we have faced the VOMS extension validity limit, which is set to 24 hours for the VO AUGER members. Thus, user cannot initialise voms-proxy for a period longer than 24 hours. There is a known gLite bug which prevents job purge from a LB database after successful retrieval of the OutputSandbox.

There is a reported bug in the FLUKA model. We have found that 2.5 % of all Done jobs fails because of specific random seed numbers. These affected jobs have the DATxxxxxx.long file of zero length.

We expect to finish the epos_gr03 and epos_gr04 Offline Production phases in the first quarter of 2008. We show the daily number of submitted and finished jobs for the first month of the Offline Production in the Figure 3.

4. Job Provenance for the AUGER collaboration

Grid middleware stacks, including gLite, matured into the state of being able to process up to millions of jobs per day. Logging and Bookkeeping, the gLite job-tracking service, keeps pace with this rate; however, it is not designed to provide a long-term archive of information on executed jobs. Job Provenance (JP) is a generic gLite service designed for long-term archiving of information on executed jobs focusing on scalability, extensibility, uniform data view, and configurability, allows more specialized catalogues to be easily built [7, 8].

Table 1. CORSIKA input file template.

Phase	**epos_gr01**	**epos_gr03 / epos_gr04**
Parameter	**Value**	**Value (if differs from gr01)**
RUNNR	1 (different for each job)	
NSHOW	1	
PRMPAR	14	14 (gr03), 5626 (gr04)
ESLOPE	-2.7	
ERANGE	1.0E+12 1.0E+12	
THIN	1.000000E-06 1.000000E+06 1.000000E+04	
THINH		**1.000E+00 1.000E+02**
THETAP	6.000000E+01 6.000000E+01	
PHIP	-180. 180.	
SEED	21618436 0 0 (different for each job)	
SEED	129718521 0 0 (different for each job)	
OBSLEV	0.	**1.452E+05**
FIXCHI	0.	
MAGNET	20.0 42.8	**2.010E+01 -1.420E+01**
HADFLG	0 0 0 0 0 2	
ECUTS	1.000E-01 1.000E-01 2.500E-04 2.500E-04	
MUADDI	T	
MUMULT	T	
ELMFLG	T T	
STEPFC	1.0	
RADNKG	200.E2	**5.0E+05**
ARRANG	0.	
EPOPAR	input ../epos/epos.param	
EPOPAR	fname inics ../epos/epos.inics	
EPOPAR	fname iniev ../epos/epos.iniev	
EPOPAR	fname initl ../epos/epos.initl	
EPOPAR	fname inirj ../epos/epos.inirj	
EPOPAR	fname inihy ../epos/epos.ini1b	
EPOPAR	fname check none	
EPOPAR	fname histo none	
EPOPAR	fname data none	
EPOPAR	fname copy none	
LONGI	T 10. F F	**T 5. T T**
ECTMAP	1.E2	**2.5E+5**
MAXPRT	0	**1**
DIRECT	./	
DATBAS	T	
USER	yourname	
PAROUT	F T	**T T**
DEBUG	F 6 F 1000000	
EXIT		

Table 2. Overall CE job statistics for the test and ongoing productions.

CE	epos_gr01			epos_gr03			epos_gr04		
	Jobs [#]	CPU time [hr]	Avg. CPU time [hr]	Jobs [#]	CPU time [hr]	Avg. CPU time [hr]	Jobs [#]	CPU time [hr]	Avg. CPU time [hr]
IJS, Ljubljana	311	4092	13	617	12390	20	96	1413	15
FzU, Prague	515	7976	15	365	9073	25	154	2659	17
LAL, Paris	254	2768	11	91	876	10	27	335	12
LIP, Lisboa	345	3351	10	45	697	15	12	132	11
NIKHEF, Amsterdam	1361	9723	7	1486	15808	11	444	5440	12
CESNET, Prague	14	699	50	141	6133	43	3	157	52
TOTAL	**2800**	**28609**	**10**	**2745**	**44976**	**16**	**736**	**10136**	**14**

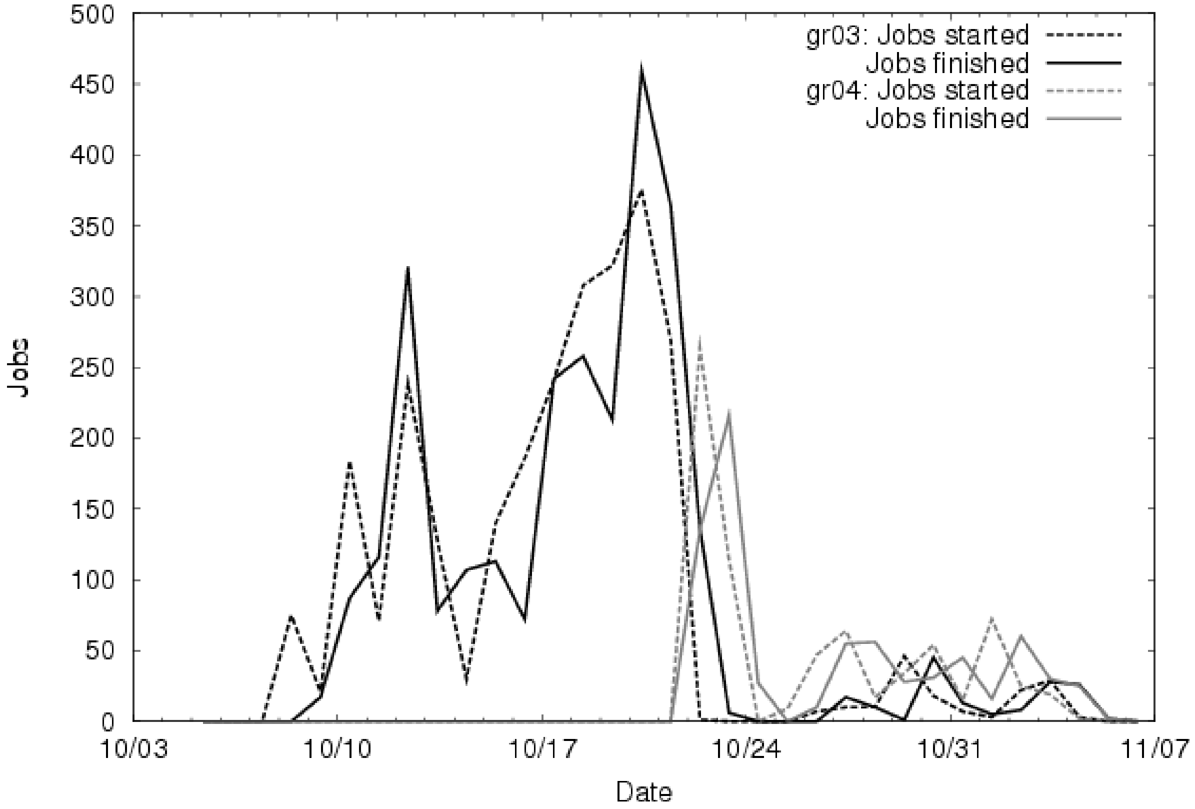


Figure 3: Number of Submitted and Done jobs during the epos_gr03 and epos_gr04 production phase.

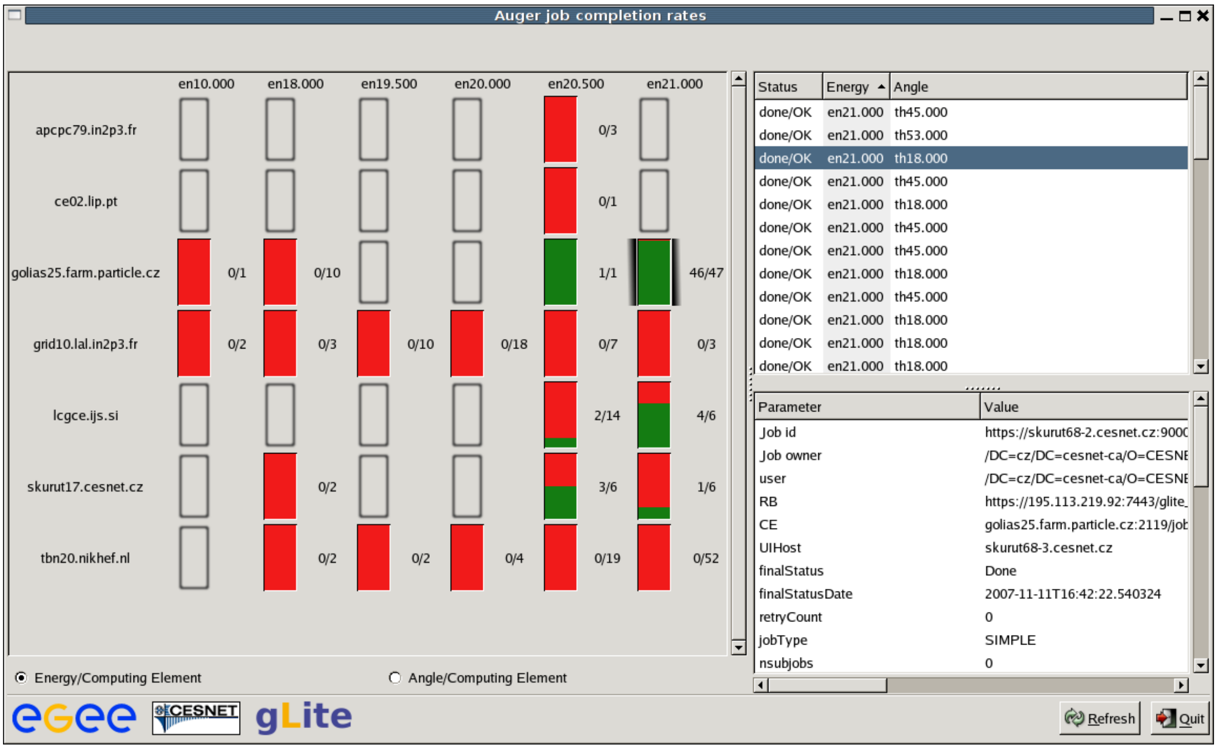
We present the first results of an experimental JP deployment for the AUGER Offline production infrastructure where a JP installation was fed with a part of AUGER jobs. The main outcome of this work is a demonstration that JP can serve the purpose of application-specific job catalogues, which had been developed for large experiments like ATLAS [nejakou citaci na ProdDB]. With JP such catalogue functionality, fullfilling requirements of different application, can be achieved with minimal development effort. In order to make use of JP we define the following attributes of AUGER Offline Production jobs:

- auger_host (hostname of the Worker Node),
- auger_type (type of jobs, e.g. AUGER production jobs),

- auger_CORSIKA_energy (energy of the cosmic ray particle),
- auger_CORSIKA_theta (zenith angle of the cosmic ray particle wavefront),
- auger_CORSIKA_primaryparticle (primary particle, i.e. proton or an iron atom),
- auger_CORSIKA_cpu_vendor_id (CPU information),
- auger_CORSIKA_cpu_model_name,
- auger_CORSIKA_cpu_frequency_MHz,
- auger_CORSIKA_cpu_cache_size,
- auger_CORSIKA_cpu_bogomips,
- auger_CORSIKA_output_long (DATxxxxxx.long file length in Bytes),
- auger_CORSIKA_output_dat (DATxxxxxx file length in Bytes),
- auger_CORSIKA_cpu_output_cputime (CPU time spent, using /usr/bin/time),
- auger_CORSIKA_output_walltime (physical time spent – in seconds), and
- auger_CORSIKA_coredump (indicator of a core dump failure).

The atributes enter the system in the form of L&B User Tags, either via JDL (those known upon job submission) or via *glite-lb-logevent* command invoked by the job (those known only at job runtime).

The information stored in the JP database can be accessed via a GUI. This GUI – see Figure 4 – enables an user to show summaries of jobs distribution on CEs with respect to the energy bin of the cosmic ray particle or to the zenith angle bin. It also shows various information about the job.


Auger job completion rates

	en10.000	en18.000	en19.500	en20.000	en20.500	en21.000
apcpc79.in2p3.fr					0/3	
ce02.lip.pt					0/1	
golias25.farm.particle.cz	0/1	0/10			1/1	46/47
grid10.lal.in2p3.fr	0/2	0/3	0/10	0/18	0/7	0/3
lcgce.ijs.si					2/14	4/6
skurut17.cesnet.cz		0/2			3/6	1/6
tbn20.nikhef.nl		0/2	0/2	0/4	0/19	0/52

● Energy/Computing Element ○ Angle/Computing Element

Status	Energy ▲	Angle
done/OK	en21.000	th45.000
done/OK	en21.000	th53.000
done/OK	en21.000	th18.000
done/OK	en21.000	th45.000
done/OK	en21.000	th18.000
done/OK	en21.000	th45.000
done/OK	en21.000	th45.000
done/OK	en21.000	th18.000
done/OK	en21.000	th45.000
done/OK	en21.000	th18.000
done/OK	en21.000	th18.000
done/OK	en21.000	th18.000

Parameter	Value
Job id	https://skurut68-2.cesnet.cz:900(
Job owner	/DC=cz/DC=cesnet-ca/O=CESNE
user	/DC=cz/DC=cesnet-ca/O=CESNE
RB	https://195.113.219.92:7443/glite_
CE	golias25.farm.particle.cz:2119/jol
UIHost	skurut68-3.cesnet.cz
finalStatus	Done
finalStatusDate	2007-11-11T16:42:22.540324
retryCount	0
jobType	SIMPLE
nsubjobs	0

Figure 4: Left pane shows summaries of job success (green) vs. failure (red) rate of particular Computing element vs. energy combination. Right pane shows details on jobs falling into a selected cell on the left.

5. Conclusions

We are able to handle Large Scale AUGER Offline Production even with the limited manpower. We have shown that the EGEE Grid is very helpful in the Large Scale Production execution.

Acknowledgments

This work was supported by the European Union within the EGEE-II project, INFSO-RI-031688.

References

[1] Pierre Auger Cosmic Ray Observatory, *http://www.auger.org/*

[2] EGEE-II: Enabling Grids for E-SciencE, *http://www.eu-egee.org/*

[3] J. Chudoba et al., VO AUGER – Preparation and First Applications, Poster presented at the EGEE'06 International Conference, Geneva, Switzerland, September 25-29, 2006

[4] CORSIKA: D. Heck et al., Report FZKA 6019 (1998), Forschungszentrum Karlsruhe; *http://www-ik.fzk.de/corsika/physics_description/corsika_phys.html*

[5] A. Fasso', A. Ferrari, J. Ranft, and P.R. Sala, "FLUKA: a multi-particle transport code", CERN 2005-10 (2005), INFN/TC_05/11, SLAC-R-773

[6] A. Fasso', A. Ferrari, S. Roesler, P.R. Sala, G. Battistoni, F. Cerutti, E. Gadioli, M.V. Garzelli, F. Ballarini, A. Ottolenghi, A. Empl and J. Ranft, "The physics models of FLUKA: status and recent developments", Computing in High Energy and Nuclear Physics 2003 Conference (CHEP2003), La Jolla, CA, USA, March 24-28, 2003, (paper MOMT005) eConf C0303241 (2003), arXiv:hep-ph/0306267

[7] L. Matyska et al., Job tracking on a grid – the Logging and Bookkeeping and Job Provenance services, Tech. rep. CESNET, 2007, *http://www.cesnet.cz/doc/techzpravy*

[8] F. Dvořák et al., Proc. IPAW'06, LNCS 4145, pp. 246–253, 2006