Experience with on-demand physics simulations on Sun Grid at Network.com

Maxim Potekhin, Jérôme Lauret

Brookhaven National Laboratory

Gabriele Carcassi, Ari Shamash, Rohit Valia

Sun MicroSystem

CHEP 2007

Highlights of the STAR simulations activity in 2007

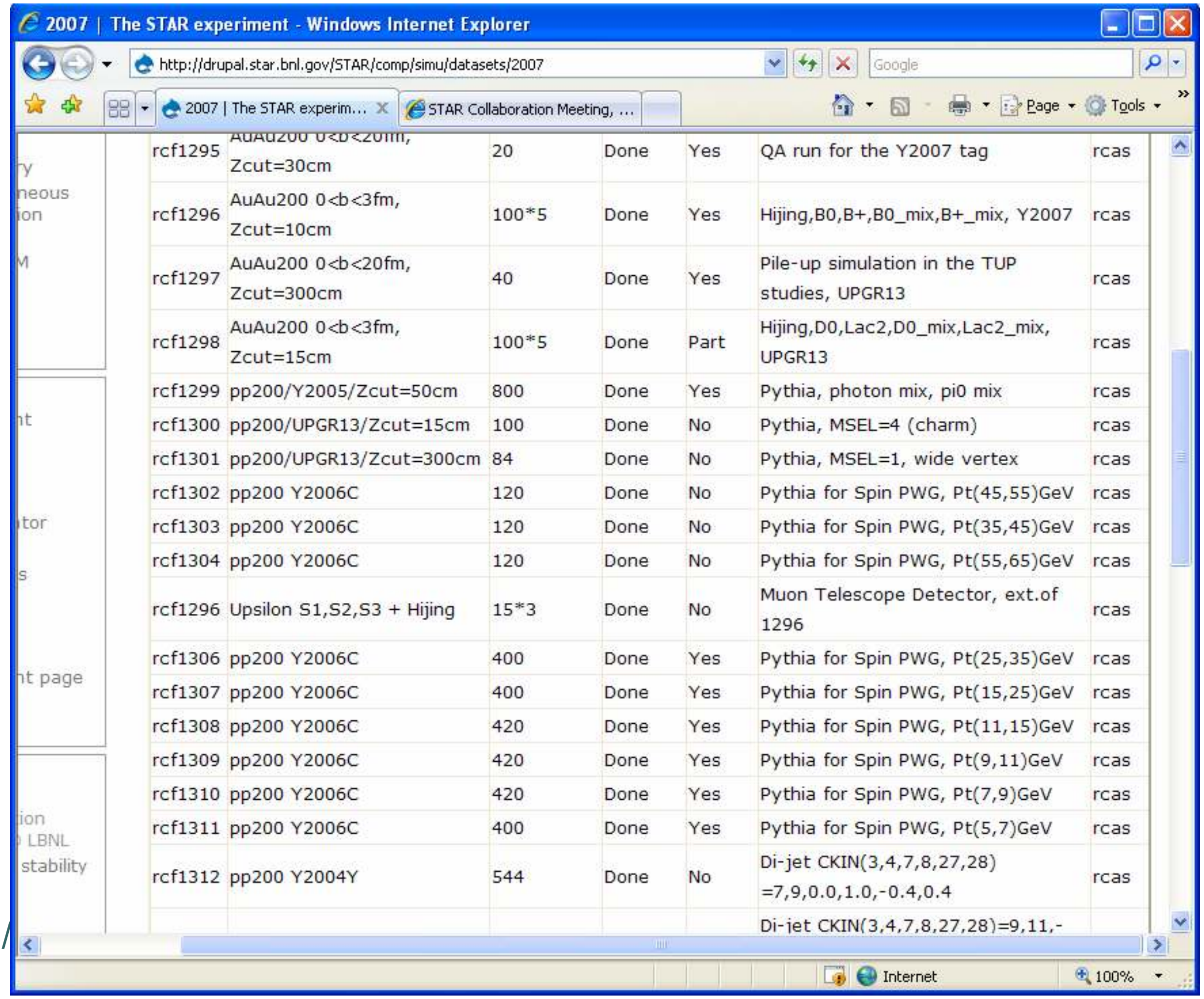
The STAR simulation program remains an ever active area of work, catering to the needs of the Physics Working Groups as well as detector R&D projects, such as

- p-p collisions: Herwig, Pythia in multiple configurations
  - cross sections and trigger rate studies
  - detailed, fine-grained requests covering multiple configurations of input parameters
- Heavy Flavor
  - B0/B+ and Hijing event mixing run (event enrichment technique)
- Tracking Upgrade Project (high-resolution pixel detector)
- Muon Telescope Detector Project
- Misc.: a study of non-photonic electrons

We generate a few dozen datasets per year, which contain ~10^7 events. Of these, there are typically ~10^6 heavy-ion events, which require considerable resources for simulation and reconstruction.

Highlights of the STAR simulations activity in 2007

- The status of individual requests is tracked on the Web:


2007 | The STAR experiment - Windows Internet Explorer
http://drupal.star.bnl.gov/STAR/comp/simu/datasets/2007
Google
2007 | The STAR experim...
STAR Collaboration Meeting, ...
Page
Tools
y
neous
ion
M
nt
tor
s
nt page
ion
LBNL
stability
Internet
100%

rcf1295	AuAu200 0<b<20fm, Zcut=30cm	20	Done	Yes	QA run for the Y2007 tag	rcas
rcf1296	AuAu200 0<b<3fm, Zcut=10cm	100*5	Done	Yes	Hijing,B0,B+,B0_mix,B+_mix, Y2007	rcas
rcf1297	AuAu200 0<b<20fm, Zcut=300cm	40	Done	Yes	Pile-up simulation in the TUP studies, UPGR13	rcas
rcf1298	AuAu200 0<b<3fm, Zcut=15cm	100*5	Done	Part	Hijing,D0,Lac2,D0_mix,Lac2_mix, UPGR13	rcas
rcf1299	pp200/Y2005/Zcut=50cm	800	Done	Yes	Pythia, photon mix, pi0 mix	rcas
rcf1300	pp200/UPGR13/Zcut=15cm	100	Done	No	Pythia, MSEL=4 (charm)	rcas
rcf1301	pp200/UPGR13/Zcut=300cm	84	Done	No	Pythia, MSEL=1, wide vertex	rcas
rcf1302	pp200 Y2006C	120	Done	No	Pythia for Spin PWG, Pt(45,55)GeV	rcas
rcf1303	pp200 Y2006C	120	Done	No	Pythia for Spin PWG, Pt(35,45)GeV	rcas
rcf1304	pp200 Y2006C	120	Done	No	Pythia for Spin PWG, Pt(55,65)GeV	rcas
rcf1296	Upsilon S1,S2,S3 + Hijing	15*3	Done	No	Muon Telescope Detector, ext.of 1296	rcas
rcf1306	pp200 Y2006C	400	Done	Yes	Pythia for Spin PWG, Pt(25,35)GeV	rcas
rcf1307	pp200 Y2006C	400	Done	Yes	Pythia for Spin PWG, Pt(15,25)GeV	rcas
rcf1308	pp200 Y2006C	420	Done	Yes	Pythia for Spin PWG, Pt(11,15)GeV	rcas
rcf1309	pp200 Y2006C	420	Done	Yes	Pythia for Spin PWG, Pt(9,11)GeV	rcas
rcf1310	pp200 Y2006C	420	Done	Yes	Pythia for Spin PWG, Pt(7,9)GeV	rcas
rcf1311	pp200 Y2006C	400	Done	Yes	Pythia for Spin PWG, Pt(5,7)GeV	rcas
rcf1312	pp200 Y2004Y	544	Done	No	Di-jet CKIN(3,4,7,8,27,28) =7,9,0.0,1.0,-0.4,0.4	rcas
					Di-jet CKIN(3,4,7,8,27,28)=9,11,-	

Operational environment of the STAR simulation program

- To state the obvious, local computing resources available to STAR are becoming increasingly scarce in relation to the amount of data that needs to be processed.
  - This applies to both CPU power and storage.

- Old mode of running simulations in STAR:
  - **Access to dedicated centralized** local **disk space** (2TB) to store the simulated data
  - **Load sharing logic** distributed between the binary executable and the driver script, using the local disk system to maintain a set of tokens that directed the workflow to the running processes
  - **Jobs submitted mostly via LSF** under a production account, which increased the load on the farm and in certain cases impeded users’ analyses
  - **Many jobs writing to a centralized disk** simultaneously are causing I/O bottlenecks for themselves as well as others working on the farm

- Conclusion:
  - **The old mode of operation does not scale due to saturation of storage, bandwidth and CPU slots**

Operational environment of the STAR simulation program – the Grid solution

- A solution – to migrate to the Grid technology, in one (or a few) of its incarnations
- STAR is an active and important user & contributor to the OSG project

Open Science Grid

- Technology has been available for some time, and now appears to be mature enough to be used for production (97% efficiency or more on OSG)
- Running simulations on the Grid is easier than reconstruction jobs
  - They are less sensitive to less than ideal efficiency – hence the Monte Carlo can have a head start
- Data transfer is a critical part of a working production system, and, as our experience shows… It works.
- In the Grid environment, jobs are run independently of each other.
  - This forces us to simplify the workflow and in fact allows for significant simplification of the driver script used in simulations due to elimination of tokens.

Operational environment of the STAR simulation program – the Grid solution

Approach:

- We have packaged all the necessary components, such the binary executable, a few shared libraries and configuration files into a self-contained unit (a tarball).
  - Any dependence on shared components such as ones previously stored in AFS is eliminated
- Necessarily, the executable must be built for the target platform

Platforms:

- We have tested this approach to run the STAR Monte Carlo production jobs on the following two Grid systems
  - Sun Grid (www.network.com) – the Sun Grid Compute Utility
  - Open Science Grid (OSG)

Why do we want to talk about Sun Grid ?

- The main thrust of STAR expansion into the Grid technology is with the OSG
- Nevertheless, we need to anticipate and analyze the trends which are leading to the emergence of industrial grids and understand issues with interfacing and utilization of such novel type of resource
- In future we will be likely facing a “grid-of-grids”
- Ex: China Grid resembles more SunGrid than a typical OSG approach …
- It provides us with an additional, and welcome, experience in packaging the payload for execution on remote clusters

The following slides will illustrate our experience with using Sun Grid

Operational environment of the STAR simulation program – the Sun Grid solution

- What is Sun Grid?
  - a commercial corporate Grid built by Sun Microsystems
  - comprised of 10^3 AMD CPUs located in datacenters owned by Sun, running the ***Solaris 10*** operating system
  - The Web interface remains the only officially supported interface of the Sun Grid, available through their portal (www.network.com) - API interface was provided as Beta (on demand)
  - users are billed at a flat rate of $1 per CPU hr *(comment: due the nascent nature of the open market of farmed CPU, this pricing is probably determined by mnemonics rather than economic realities)*
  - users are allotted dedicated storage space on Sun Grid, with pricing scheme being still worked out
  - STAR was granted a block of promotional CPU hrs to be able to complete the feasibility study, as well as half a TB of storage

Network.com Sun Grid Usage Models

On-demand Infrastructure

Sun Grid Compute Utility @ Network.com

Bring your own applications to Sun Grid Compute Utility

Powerful compute resource for the enterprise

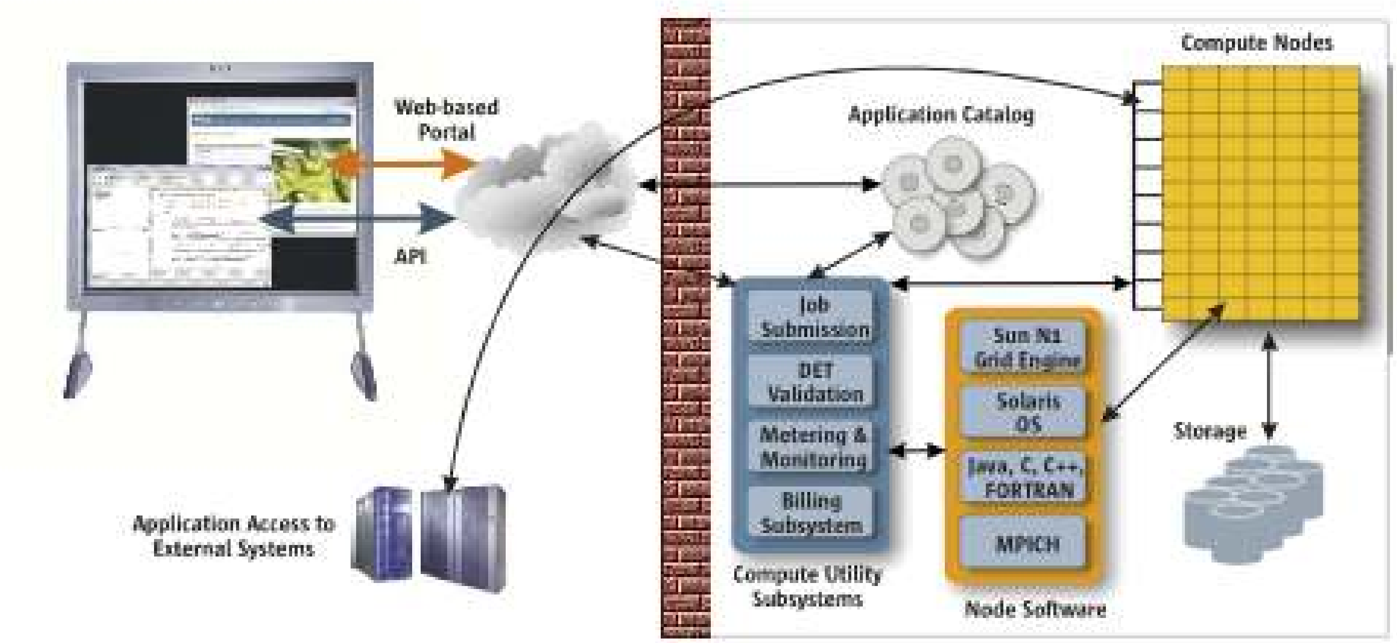
On-demand Applications

Network.com Application Catalog

Use ISV applications published on Network.com

Immediate access to on-demand HPC applications

Sun Grid
Architecture Overview



Network.com Application Catalog

Job Catalog

- Allows users to search and check out applications that have been shared via “Job Template” mechanism
- Easy search mechanism for applications

Digital Entitlement Tokens

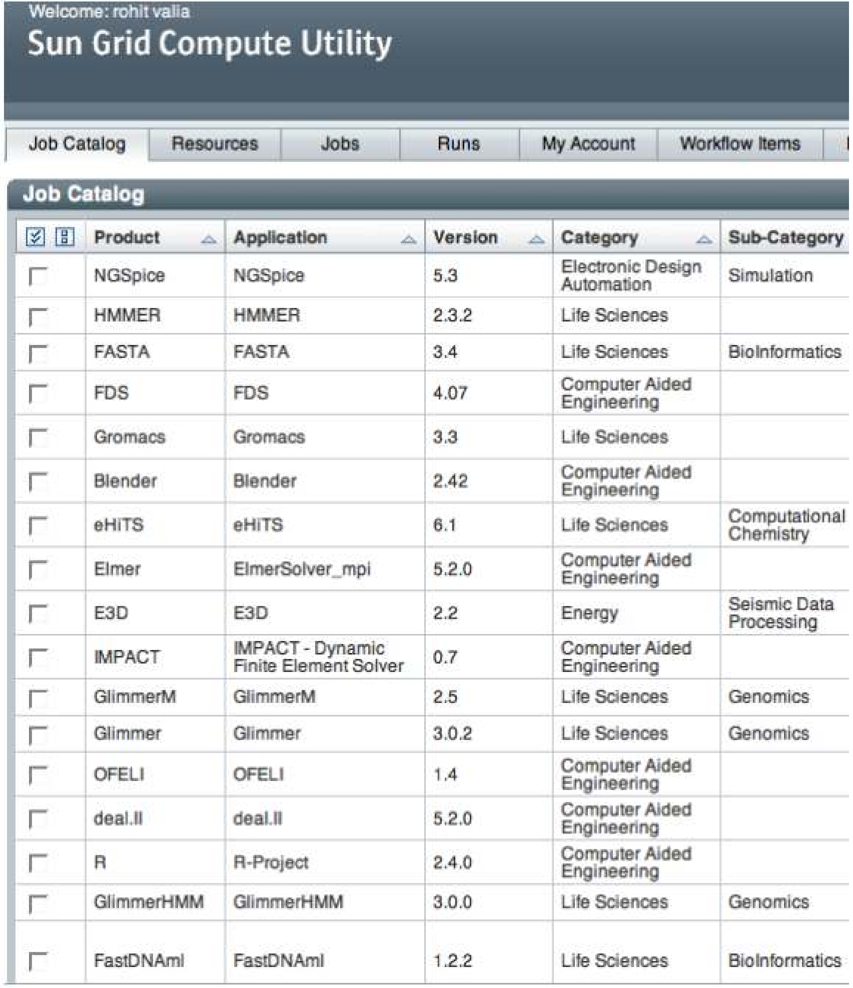
- Allows publishers to protect their software via fine-grained access controls
- Publishers may elect to require use of Digital Entitlement Tokes to run their software
- End users can easily acquire tokens via website link in the catalog
- Use readily available X.509 certificates and signed jar files that simplify creating, maintaining, and using the application

Network.com Application Catalog

Enabling On-Demand Delivery of HPC Applications

- EHITS - SimBio Sys
- Rational Numbers – MathSpec
- Gromacs
- Readseq
- T-Coffee
- GlimmerM
- Glimmer
- GlimmerHMM
- FastDNAml
- BLAST
- ClustalW

- E3D
- namd
- Calculix
- FreeFEM
- IMPACT
- OFELI
- deal.II
- R-Project
- FDS
- Blender
- ElmerSolver_mpi



Welcome: rohit valia

Sun Grid Compute Utility

Job Catalog

☑	Product	Application	Version	Category	Sub-Category
☐	NGSpice	NGSpice	5.3	Electronic Design Automation	Simulation
☐	HMMER	HMMER	2.3.2	Life Sciences	
☐	FASTA	FASTA	3.4	Life Sciences	BioInformatics
☐	FDS	FDS	4.07	Computer Aided Engineering	
☐	Gromacs	Gromacs	3.3	Life Sciences	
☐	Blender	Blender	2.42	Computer Aided Engineering	
☐	eHiTS	eHiTS	6.1	Life Sciences	Computational Chemistry
☐	Elmer	ElmerSolver_mpi	5.2.0	Computer Aided Engineering	
☐	E3D	E3D	2.2	Energy	Seismic Data Processing
☐	IMPACT	IMPACT - Dynamic Finite Element Solver	0.7	Computer Aided Engineering	
☐	GlimmerM	GlimmerM	2.5	Life Sciences	Genomics
☐	Glimmer	Glimmer	3.0.2	Life Sciences	Genomics
☐	OFELI	OFELI	1.4	Computer Aided Engineering	
☐	deal.II	deal.II	5.2.0	Computer Aided Engineering	
☐	R	R-Project	2.4.0	Computer Aided Engineering	
☐	GlimmerHMM	GlimmerHMM	3.0.0	Life Sciences	Genomics
☐	FastDNAml	FastDNAml	1.2.2	Life Sciences	BioInformatics

Network.com
New Features

- Internet Access
  - Sun Grid Internet Access allows enabled applications to open outbound TCP connections to any accessible internet host

- Application Program Interface (APIs)
  - A Java API allowing programatic access to Job submission and data transfer facilities

    This feature is currently in limited access release.

- Network.com plugin for NetBeans IDE

- International Availability
  - The Sun Grid Compute Utility is now accessible by users from 25 countries, including the USA (See Network.com for list of countries)

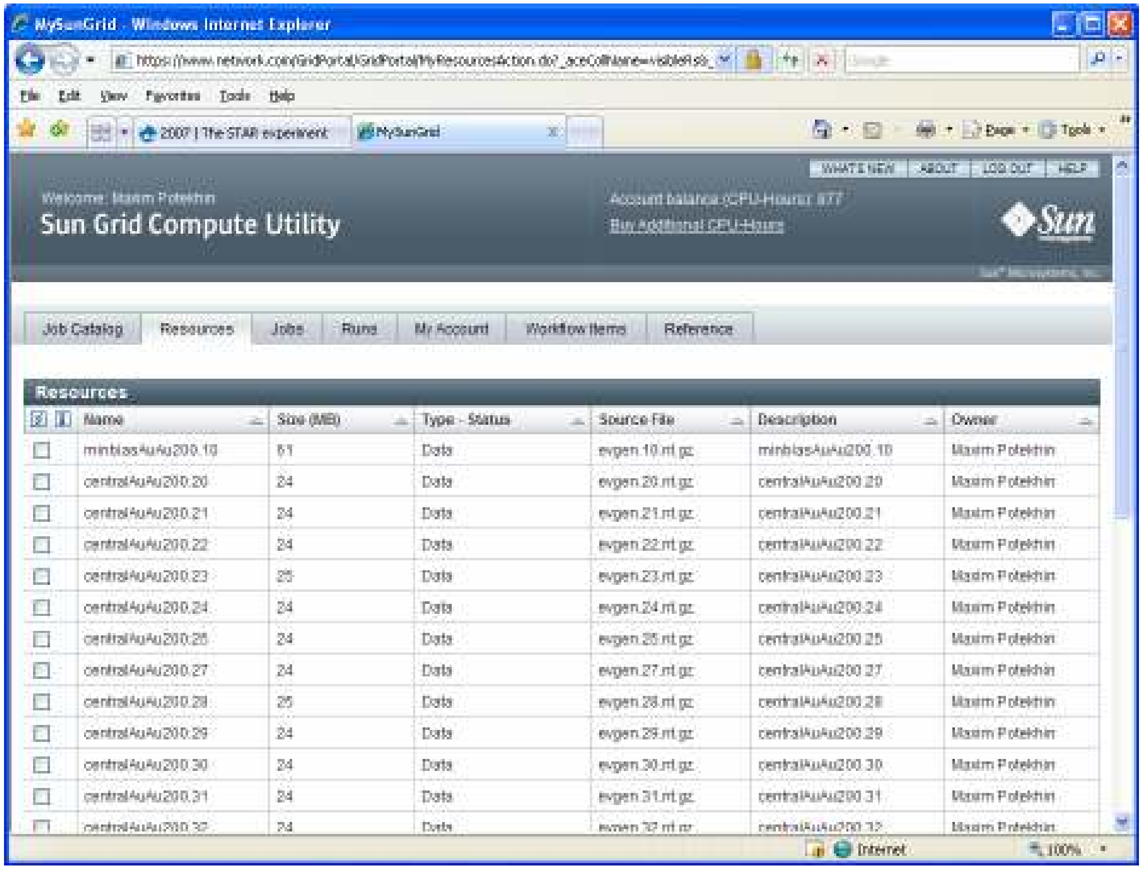
Operational environment of the STAR simulation program – the Sun Grid solution

- What was involved in getting the STAR jobs to run on Sun Grid?
  - a not entirely straightforward port to Solaris 10

- What is the basic scheme of utilizing the Sun Grid?
  - the user uploads compressed archives, which are termed “resources”, to the Sun Grid utility, using the Web interface
  - such resources can contain a combination of any types of data the user decides to put there, including executables, libraries, configuration files etc
  - the user generates a job definition, which contains a list of the resources used, and the script to be executed at job submission time
  - in the job definition, the user can specify values of the input parameters for the script being executed, thus providing a welcome degree of flexibility
  - a job can be submitted and becomes a “run” at execution time
  - resources, jobs and run status are all monitored via the Web interface
  - the data produced by the user job is downloaded back to the user’s host via https, in compressed format

Operational environment of the STAR simulation program – the Sun Grid solution

- the Resources page of the Sun Grid Web interface


MySunGrid - Windows Internet Explorer
https://www.network.com/GridPortal/GridPortal/MyResourcesAction.do?_aceColName=visibleRss_
File Edit View Favorites Tools Help
2007 | The STAR experiment
MySunGrid
Page Tools
WHAT'S NEW ABOUT LOG OUT HELP
Sun
Job Catalog
Resources
Jobs
Runs
My Account
Workflow Items
Reference
Internet
100%

Welcome: Maxim Potekhin

Sun Grid Compute Utility

Account balance (CPU-Hours): 877

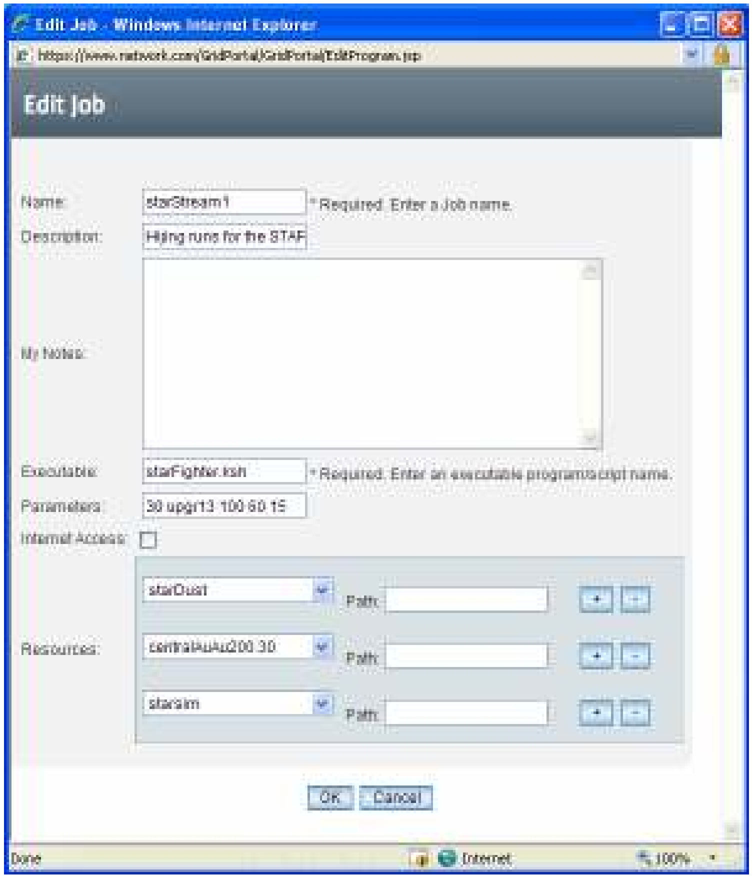
Buy Additional CPU-Hours

Resources

☒	Name	Size (MB)	Type - Status	Source File	Description	Owner
□	minbiasAuAu200.10	61	Data	evgen.10.nt.gz	minbiasAuAu200.10	Maxim Potekhin
□	centralAuAu200.20	24	Data	evgen.20.nt.gz	centralAuAu200.20	Maxim Potekhin
□	centralAuAu200.21	24	Data	evgen.21.nt.gz	centralAuAu200.21	Maxim Potekhin
□	centralAuAu200.22	24	Data	evgen.22.nt.gz	centralAuAu200.22	Maxim Potekhin
□	centralAuAu200.23	25	Data	evgen.23.nt.gz	centralAuAu200.23	Maxim Potekhin
□	centralAuAu200.24	24	Data	evgen.24.nt.gz	centralAuAu200.24	Maxim Potekhin
□	centralAuAu200.25	24	Data	evgen.25.nt.gz	centralAuAu200.25	Maxim Potekhin
□	centralAuAu200.27	24	Data	evgen.27.nt.gz	centralAuAu200.27	Maxim Potekhin
□	centralAuAu200.28	25	Data	evgen.28.nt.gz	centralAuAu200.28	Maxim Potekhin
□	centralAuAu200.29	24	Data	evgen.29.nt.gz	centralAuAu200.29	Maxim Potekhin
□	centralAuAu200.30	24	Data	evgen.30.nt.gz	centralAuAu200.30	Maxim Potekhin
□	centralAuAu200.31	24	Data	evgen.31.nt.gz	centralAuAu200.31	Maxim Potekhin
□	centralAuAu200.32	24	Data	evgen.32.nt.gz	centralAuAu200.32	Maxim Potekhin

Operational environment of the STAR simulation program – the Sun Grid solution

- the Job Edit page of the Sun Grid Web interface



Edit Job - Windows Internet Explorer
https://www.network.com/GridPortal/GridPortal/EditProgram.jsp

Edit Job

Name:	starStream1	* Required. Enter a Job name.
Description:	Hijing runs for the STAR	
My Notes:		
Executable:	starFighter.ksh	* Required. Enter an executable program/script name.
Parameters:	30 upgr13 100 60 15	
Internet Access:	□	
Resources:	starDust	Path:
	centralAuAu200.30	Path:
	starsim	Path:

+
-
OK
Cancel
Done
Internet
100%

Operational environment of the STAR simulation program – the Sun Grid solution

- Has anything been run on the Sun Grid in production mode?
  - YES, a portion of the Hijing simulation for the STAR Tracking Upgrade R&D Project, of the order of ~10^4 heavy-ion events
- What's next for the Sun Grid, from STAR perspective?
  - we will continue running jobs on the Sun Grid, as time permits, while providing feedback to the Sun developers
- Sun Grid: pros and cons
  - ...next slide

Sun Grid solution: pros and cons

- pros:
  - there is no custom software to install on the user’s system, therefore one can jump right into the action
  - there is a hope that in most cases a port to Solaris 10 should be straightforward
  - the user interface is clean and unambiguous
  - having dedicated storage space on the Sun cluster greatly simplifies the management of produced data and saves a lot of upload time, in the context of simulations
  - healthy level of support for end users and developers
- cons:
  - the data transfer to and from Sun Grid is clearly the weak link in the system (based on the feedback, there have been indications from Sun that this will be addressed)
  - as of the time of this writing, there is no programmatic interface (publicly available) to the Sun Computing Utility. This effectively limits the usefulness of this system in the Particle and Nuclear Physics context to running Monte Carlo jobs in modestly sized batches – this was since addressed
  - having promotional (free) CPU-hours is great, however given the nascent nature of the market in computing utilities, the current price point for this resource

Conclusions

- There is a fairly significant (and in fact growing) level of demand from the PWGs, for custom Monte Carlo datasets
  - The “old” mode of operation, whereby load sharing was done by the driver script and the executable, and the jobs were submitted to a local farm, does not scale to meet the new challenges
  - Running simulation via Grid has been the main (and sole) way for simulation in STAR for a while now

- Expanding to a grid-of-grid?
- While a small fraction of our simulation was done on non-OSG resources, running on SunGrid provided
  - An avenue for outreaching to other (commercial) Grid approach and understanding “other” Grid paradigm
  - Resources for running STAR R&D studies
  - Forced to reconsider our simulation workflow and make it portable