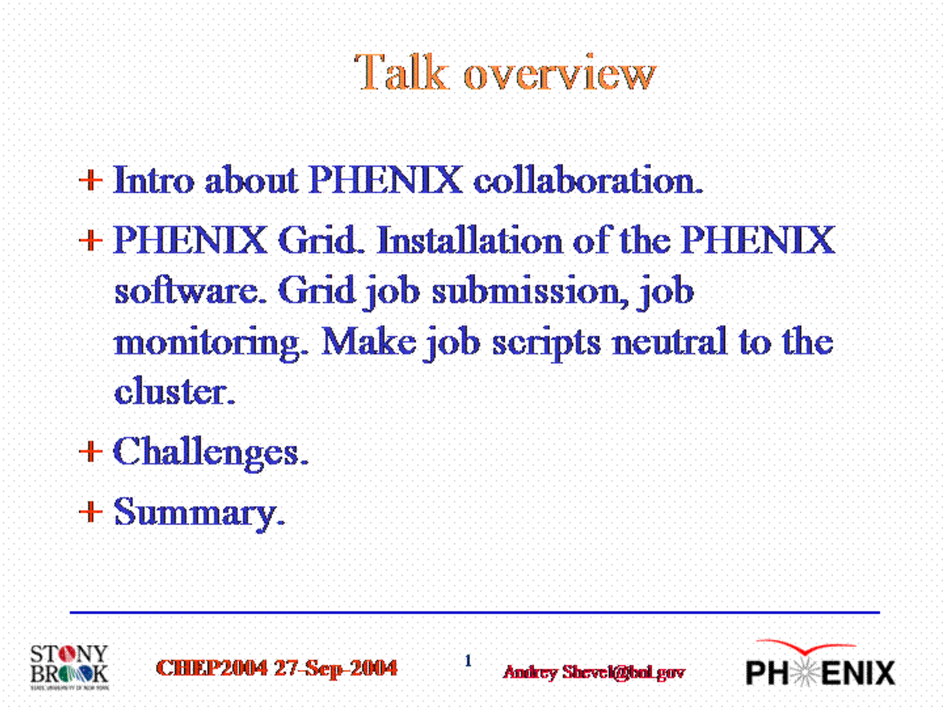
PHENIX Job Submission/Monitoring in transition to the Grid Infrastructure

Andrey Y. Shevel, Barbara Jacak, Roy Lacey, Dave Morrison, Michael Reuter, Irina Sourikova, Timothy Thomas, Alex Withers



Talk overview

- Intro about PHENIX collaboration.
- PHENIX Grid. Installation of the PHENIX software. Grid job submission, job monitoring. Make job scripts neutral to the cluster.
- Challenges.
- Summary.

CHEP2004 27-Sep-2004 1 Andrey Shevel@bnl.gov

Brief info on PHENIX

- Large, widely-spread collaboration (same scale as CDF and D0), more than 450 collaborators, 12 nations, 57 Institutions, 11 U.S. Universities, currently in fourth year of data-taking.
- ~250 TB/yr of raw data.
- ~230 TB/yr of reconstructed output.
- ~370 TB/yr microDST + nanoDST.
- In total about ~850TB+ of new data per year.
- Primary event reconstruction occurs at BNL RCF (RHIC Computing Facility).
- Partial copy of raw data is at CC-J (Computing Center in Japan) and part of DST output is at CC-F (France).

Phenix Grid and Grid related activity

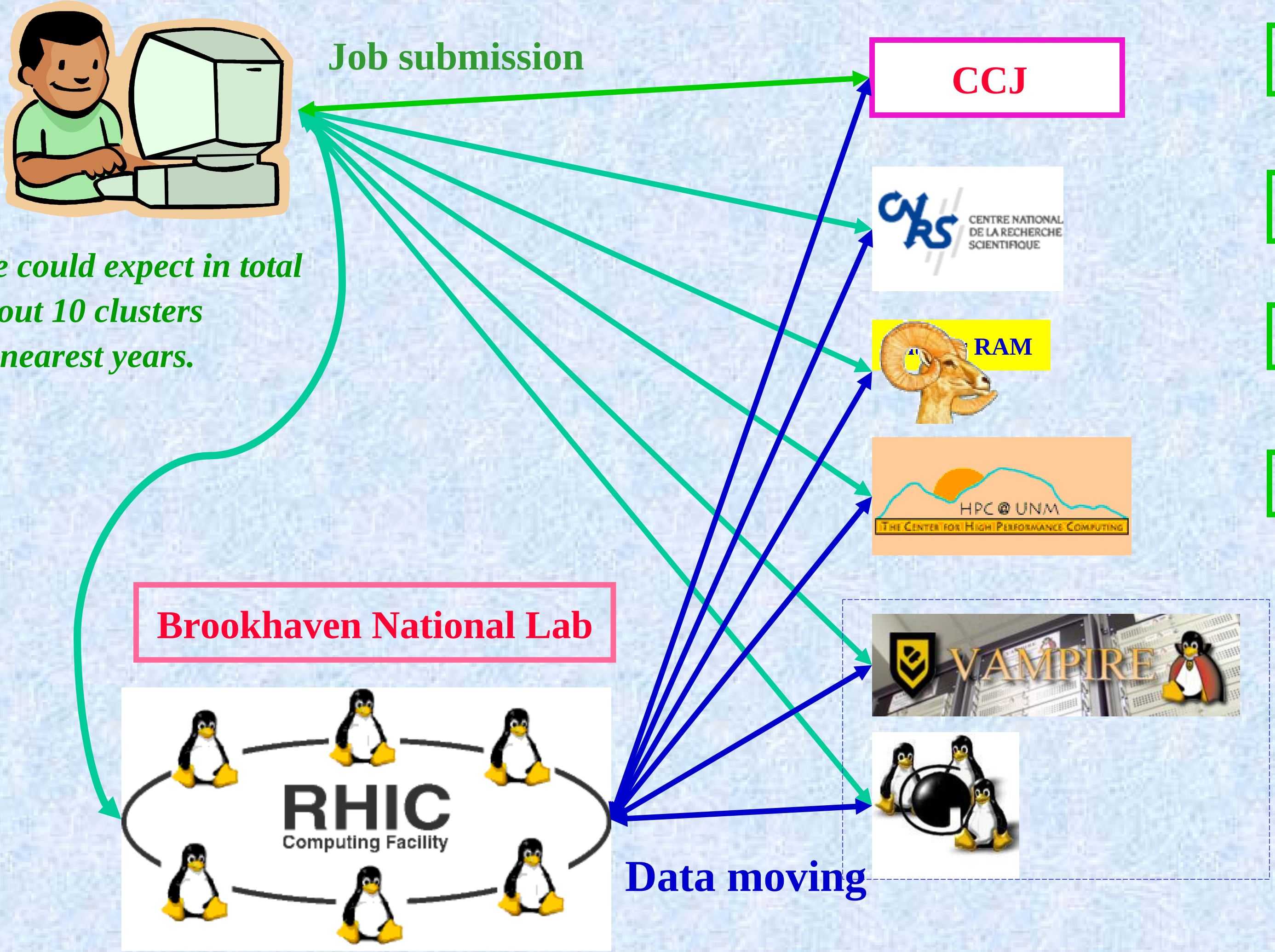
+ In addition to our concentrating now on the job monitoring and submission we are involved into:

  - testing D-cache prototype for file management at RCF;

  - moving most of Phenix databases to Postgres database (separate presentation);

  - participating in Grid3 testbed (*http://www.ivdgl.org/grid3/*).

PHENIX Grid


CNRS
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
RAM
HPC @ UNM
THE CENTER FOR HIGH PERFORMANCE COMPUTING
VAMPIRE
RHIC
Computing Facility

Job submission

CCJ

We could expect in total about 10 clusters in nearest years.

Brookhaven National Lab

Data moving

RIKEN CCJ (Japan)

IN2P3 (France)

SUNY @ Stony Brook

University of New Mexico

Vanderbilt University

PNPI (Russia)

PHENIX multi cluster conditions

+ Computing Clusters have different:

  - computing power;

  - batch job schedulers;

  - details of administrative rules.

+ Computing Clusters have common:

  - OS Linux (there are clusters with different Linux versions);

  - Most of clusters have gateways with Globus toolkit;

  - Grid status board (*http://ram3.chem.sunysb.edu/*)

Remote cluster environment

- + Max number of the computing clusters is about 10.
- + Max number of the submitted at the same time Grid jobs is about 10**4 or less.
- + The amount of the data to be transferred (between BNL and remote cluster) for physics analysis is varied from about 2 TB/quarter to 5 TB/week.
- + We use PHENIX file catalogs:

  - centralized file catalog (*http://replicator.phenix.bnl.gov/~replicator/fileCatalog.html*);

  - cluster file catalogs (for example at SUNYSB is used slightly re-designed version MAGDA *http://ram3.chem.sunysb.edu/magdaf/*).

Exporting the application software to run on remote clusters

+ The porting of PHENIX software in binary form is presumably most common port method in PHENIX Grid:

  - copying over AFS to mirror PHENIX directory structure on remote cluster (by cron job);

  - preparing PACMAN packages for specific class of tasks (e.g. specific simulation).

User job scripts

+ One of the important issues for multi cluster environment is to make sure that all user scripts will run on all available clusters with same results;

+ It could be done at least by two ways:

  - to eliminate from the existing user job scripts all specifics and make them neutral to the cluster environment (to do that we introduced pre assigned environment variables like GRIDVM_JOB_SUBMIT, GRIDVM_BATCH_SYSTEM, etc);

  - change processing scheme – for example - start to use *STAR scheduler [SUMS]* (*http://www.star.bnl.gov/STAR/comp/Grid/scheduler/*) with use same environment variables.

The job submission scenario at remote Grid cluster

+ User needs just *qualified computing cluster*: i.e. with enough available disk space, specific version for compiler and related software pieces.

+ Before job production:

  - To copy/replicate the *major data sets* (physics data) to remote cluster.

  - To copy/replicate the *minor data sets* (scripts, parameters, etc.) to remote cluster.

  - To guarantee that remote cluster with required environment is functioning properly by the set of test tools (scripts).

+ To start the *master job* (script) which will submit many *sub-jobs* with default (for remote cluster) batch system. Also master job might deploy the software components required for *sub-jobs*.

+ During production (job run) and after all jobs were accomplished (after job production):

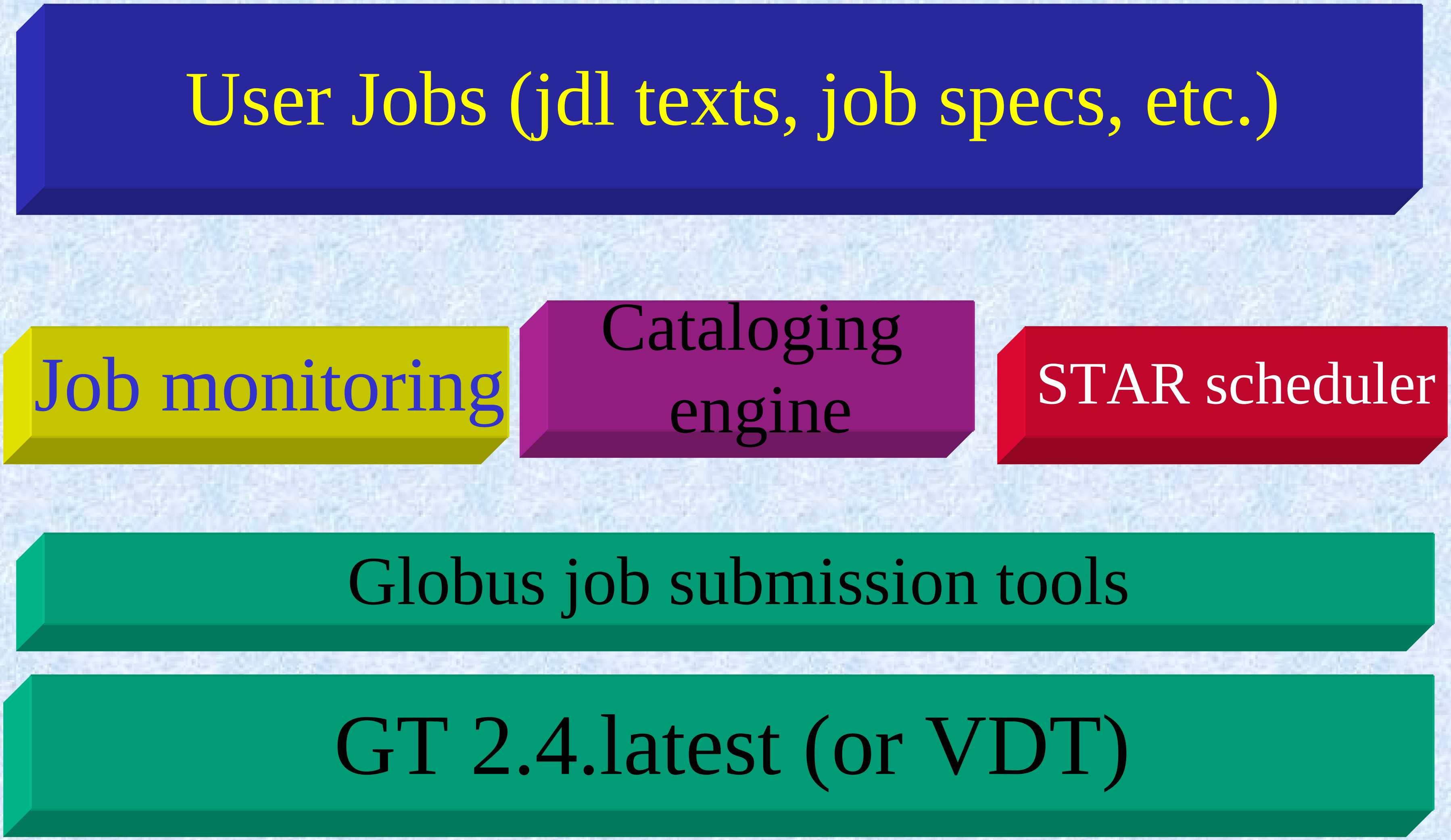
  - To watch the jobs with monitoring system.

  - To copy the result data from remote cluster to target destination (desktop or RCF).

The requirements for job monitoring in multi cluster environment

+ What is job monitoring ?
+ To keep track of the submitted jobs
  - whether the jobs have been accomplished;
  - in which cluster the jobs are performed;
  - where the jobs were performed in the past (one day, one week, one month ago).
+ *Obviously the information about the jobs must be written in the database and kept there. The same database might be used for job control purpose (cancel jobs, resubmit jobs, other job control operations in multi cluster environment)*
+ We developed PHENIX job monitoring tool on the base of BOSS *(http://www.bo.infn.it/cms/computing/BOSS/).*

PHENIX Grid job submission/monitoring



User Jobs (jdl texts, job specs, etc.)

Job monitoring

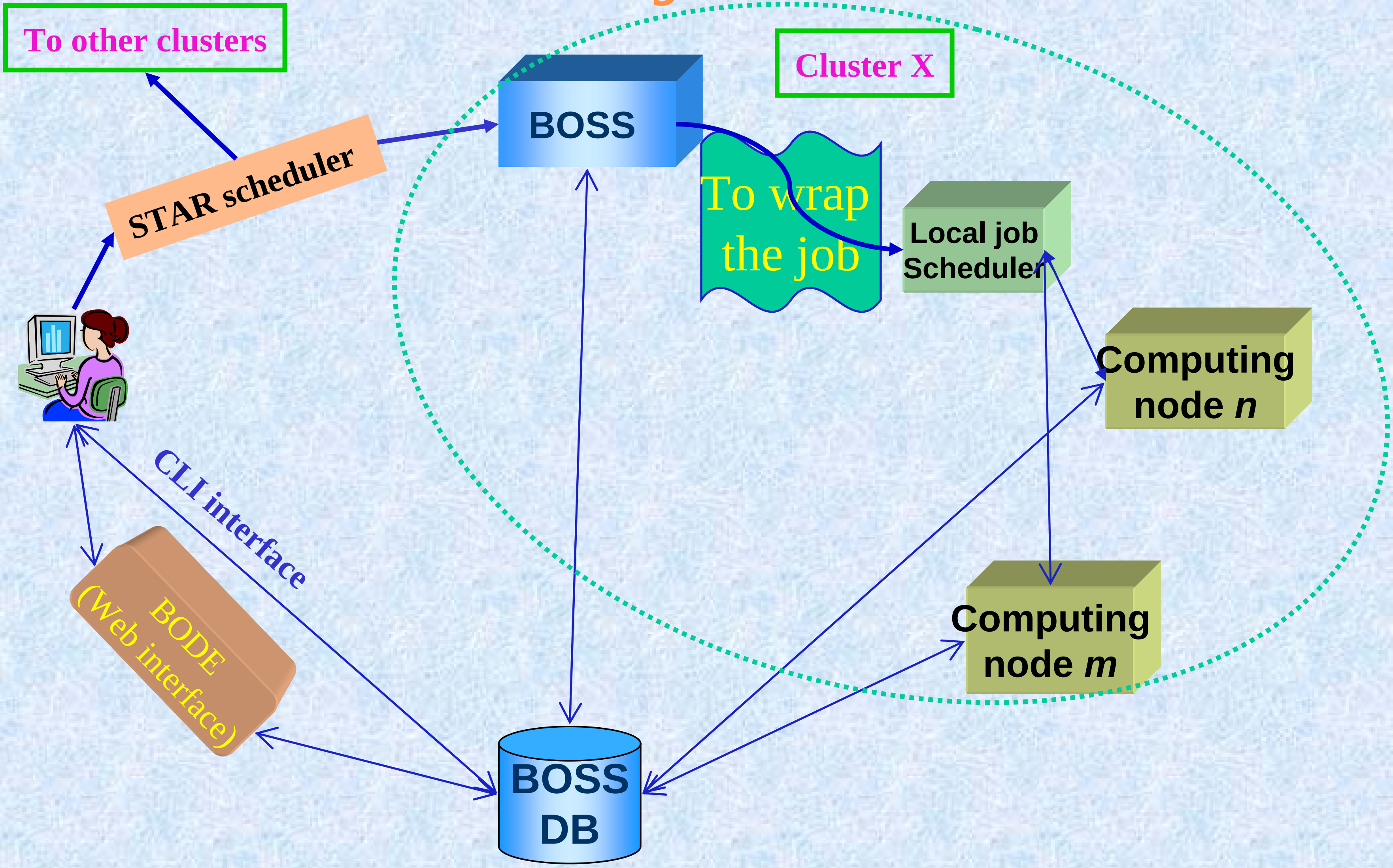
Cataloging engine

STAR scheduler

Globus job submission tools

GT 2.4.latest (or VDT)

Basic job flow

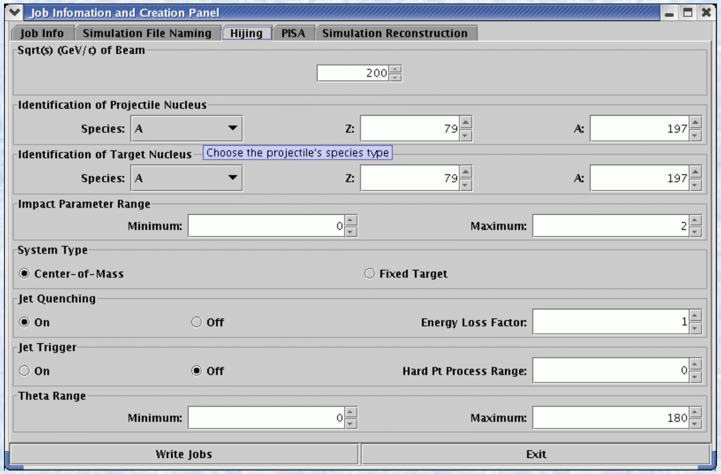


Example for job submission panel

- Need parameter input for event generator, detector response package, reconstruction etc.

- Created GUI to provide easy input specification and hide JDL file creation.

- Use Star-Scheduler for job submission to the Grid and Pacman for job software deployment.



Job Infomation and Creation Panel

Sqrt(s) (GeV/c) of Beam: 200

Identification of Projectile Nucleus
Species: A Z: 79 A: 197

Choose the projectile's species type

Identification of Target Nucleus
Species: A Z: 79 A: 197

Impact Parameter Range
Minimum: 0 Maximum: 2

System Type
● Center-of-Mass ○ Fixed Target

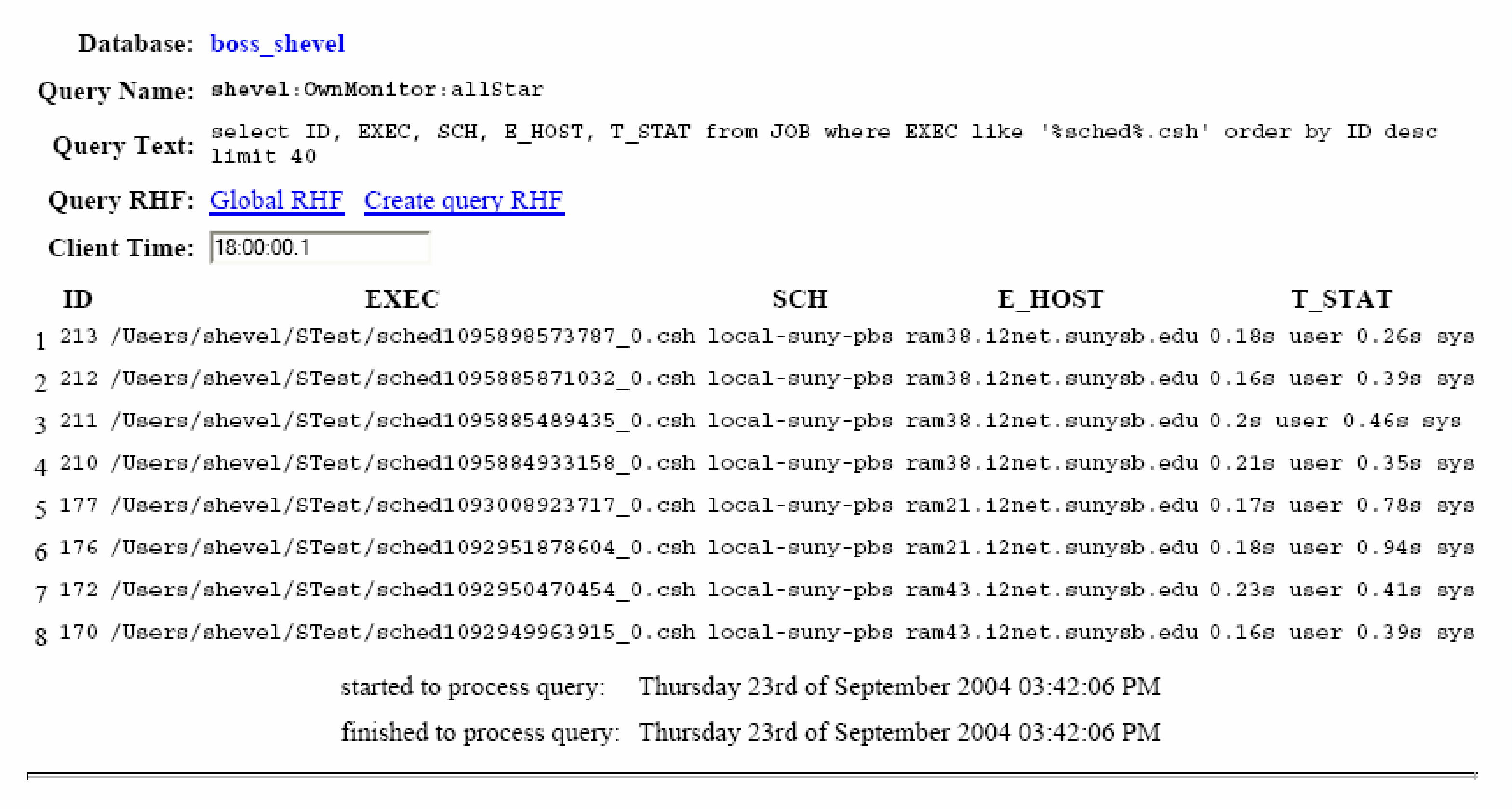
Jet Quenching
● On ○ Off Energy Loss Factor: 1

Jet Trigger
○ On ● Off Hard Pt Process Range: 0

Theta Range
Minimum: 0 Maximum: 180

Continuation of the example

(BOSS/BODE view of jobs submitted by STAR scheduler)



Database: boss_shevel

Query Name: `shevel:OwnMonitor:allStar`

Query Text:

```
select ID, EXEC, SCH, E_HOST, T_STAT from JOB where EXEC like '%sched%.csh' order by ID desc
limit 40
```

Query RHF: Global RHF Create query RHF

Client Time: 18:00:00.1

	ID	EXEC	SCH	E_HOST	T_STAT
1	213	/Users/shevel/STest/sched1095898573787_0.csh	local-suny-pbs	ram38.i2net.sunysb.edu	0.18s user 0.26s sys
2	212	/Users/shevel/STest/sched1095885871032_0.csh	local-suny-pbs	ram38.i2net.sunysb.edu	0.16s user 0.39s sys
3	211	/Users/shevel/STest/sched1095885489435_0.csh	local-suny-pbs	ram38.i2net.sunysb.edu	0.2s user 0.46s sys
4	210	/Users/shevel/STest/sched1095884933158_0.csh	local-suny-pbs	ram38.i2net.sunysb.edu	0.21s user 0.35s sys
5	177	/Users/shevel/STest/sched1093008923717_0.csh	local-suny-pbs	ram21.i2net.sunysb.edu	0.17s user 0.78s sys
6	176	/Users/shevel/STest/sched1092951878604_0.csh	local-suny-pbs	ram21.i2net.sunysb.edu	0.18s user 0.94s sys
7	172	/Users/shevel/STest/sched1092950470454_0.csh	local-suny-pbs	ram43.i2net.sunysb.edu	0.23s user 0.41s sys
8	170	/Users/shevel/STest/sched1092949963915_0.csh	local-suny-pbs	ram43.i2net.sunysb.edu	0.16s user 0.39s sys

started to process query: Thursday 23rd of September 2004 03:42:06 PM

finished to process query: Thursday 23rd of September 2004 03:42:06 PM

Challenges for PHENIX Grid

- + Admin service (where the user can complain if something is going wrong with his Grid jobs on some cluster?).
- + More sophisticated job control in multi cluster environment; job accounting.
- + Complete implementing technology for run-time installation for remote clusters.
- + More checking tools to be sure that most things in multi cluster environment are running well – i.e. automate the answer for the question “is account A on cluster N being PHENIX qualified environment?”. To check it every hour or so.
- + Portal to integrate all PHENIX Grid tools in one user window.

Summary

- The multi cluster environment is our reality and we need more user friendly tools for typical user to reduce the cost of clusters power integration.
- In our condition the best way to do that is to use already developed subsystems as *bricks* to build up the robust PHENIX Grid computing environment. Most effective way to do that is to be AMAP cooperative with other BNL collaborations (STAR as good example).
- Serious attention must be paid to automatic installation of the existing physics software.