History photos: A. Shevel reports on CSD seminar about new Internet facilities at PNPI (Jan 1995)

Distributed computing in HEP [Grid prospects]

Andrey Y. Shevel

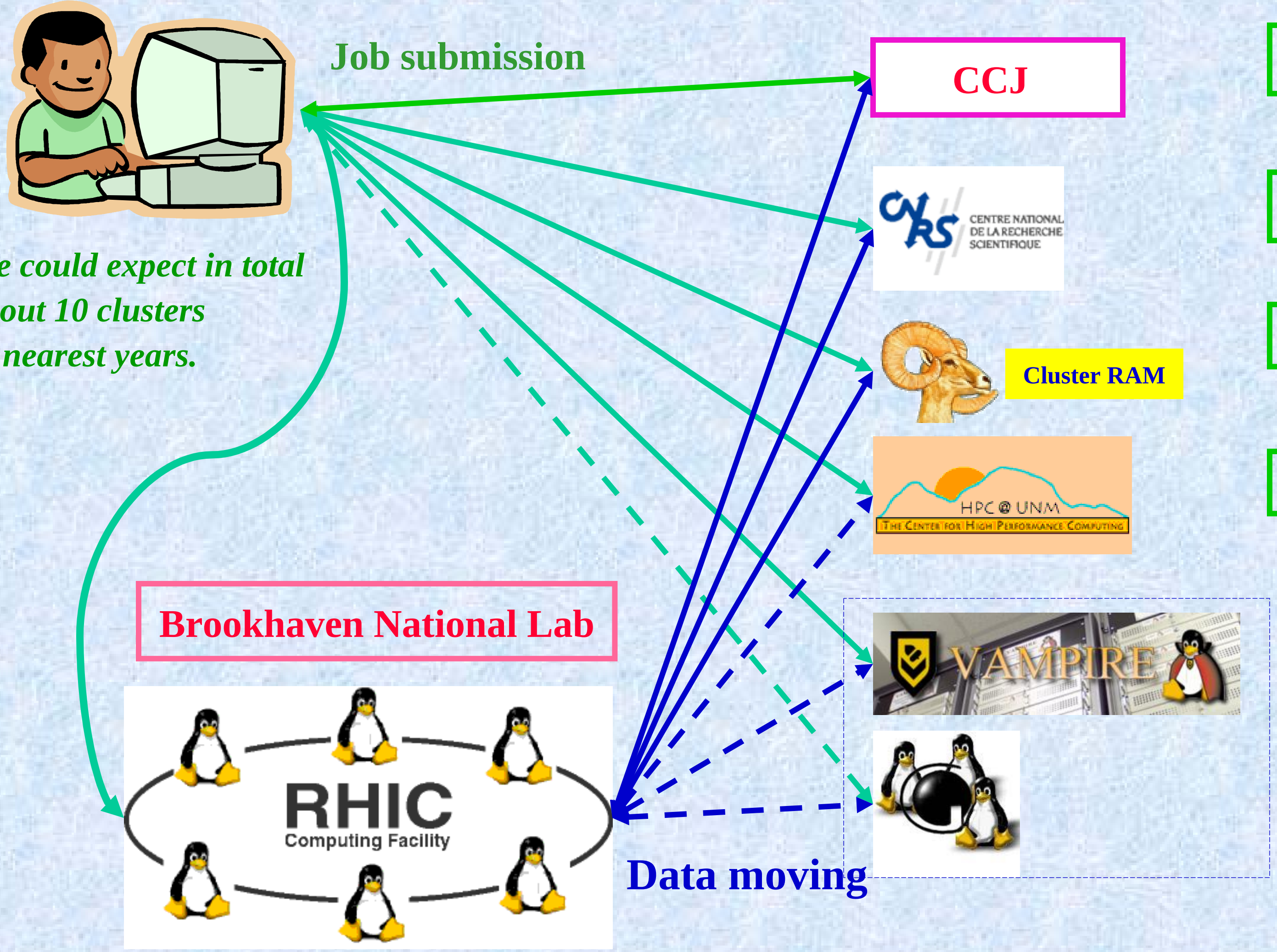
PHENIX Job Submission/Monitoring in transition to the Grid Infrastructure

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

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



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

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*/phenix-grid.html)

Other PHENIX conditions

- 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).

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)*
+ PHENIX job monitoring tool was developed on the base of BOSS *(http://www.bo.infn.it/cms/computing/BOSS/).*

"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.

[388] A Lightweight Monitoring and Accounting System for LHCb DC04 Production

+ Complete implementing technology for run-time installation for remote clusters.

[476] CHOS, a method for concurrently supporting multiple operating system.

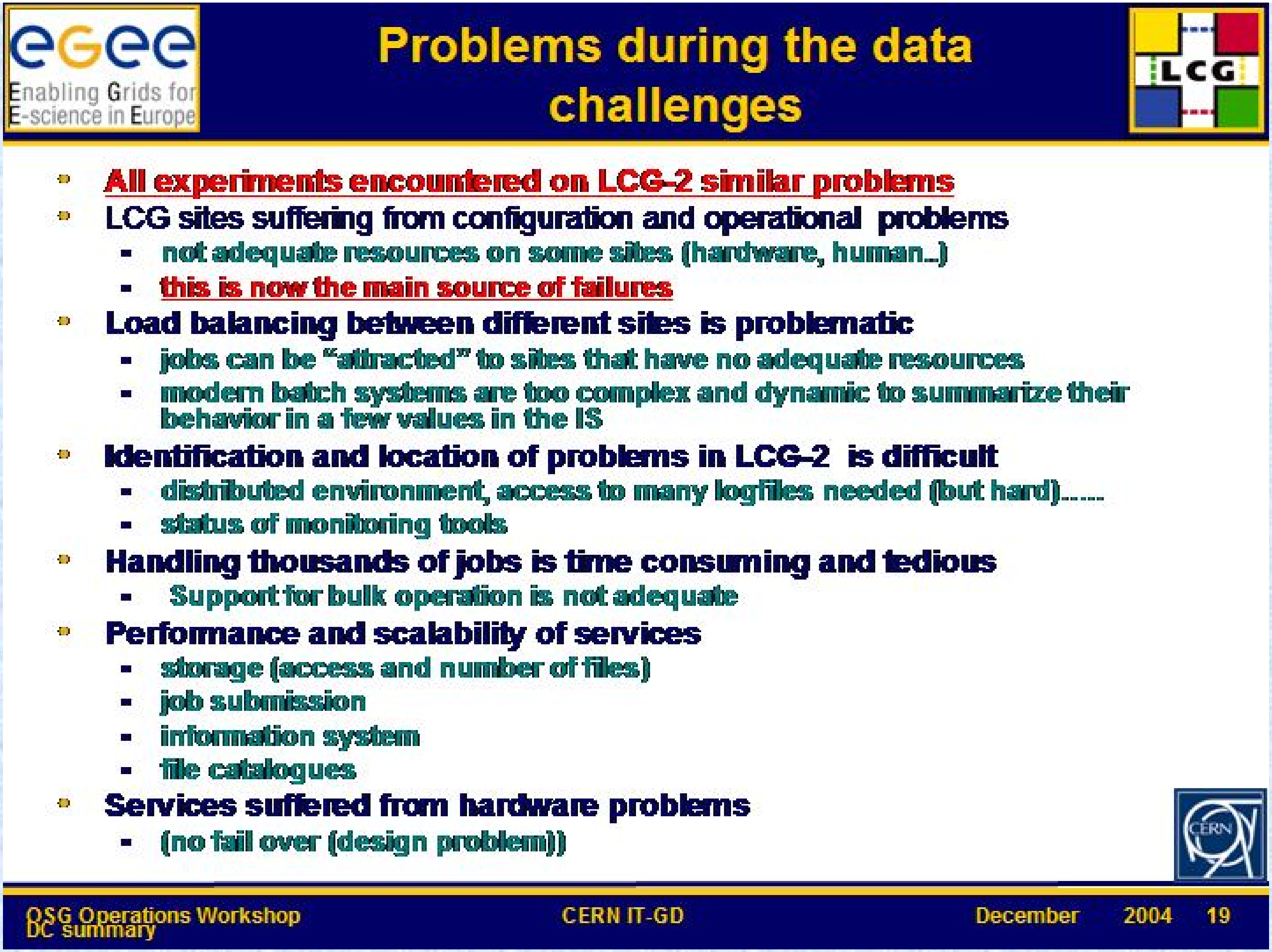
+ More checking tools to be sure that most things in multi cluster environment are running well – i.e. automate the answer for the question "is account A on cluster N being PHENIX qualified environment?". To check it every hour or so.

[455] Application of the SAMGrid Test Harness for Performance Evaluation and Tuning of a Distributed Cluster Implementation of Data Handling Services

+ Portal to integrate all PHENIX Grid tools in one user window.

[443] The AliEn Web Portal

[182] Grid Enabled Analysis for CMS: prototype, status and results



Problems during the data challenges

- All experiments encountered on LCG-2 similar problems
- LCG sites suffering from configuration and operational problems
  - not adequate resources on some sites (hardware, human..)
  - this is now the main source of failures
- Load balancing between different sites is problematic
  - jobs can be "attracted" to sites that have no adequate resources
  - modern batch systems are too complex and dynamic to summarize their behavior in a few values in the IS
- Identification and location of problems in LCG-2 is difficult
  - distributed environment, access to many logfiles needed (but hard)......
  - status of monitoring tools
- Handling thousands of jobs is time consuming and tedious
  - Support for bulk operation is not adequate
- Performance and scalability of services
  - storage (access and number of files)
  - job submission
  - information system
  - file catalogues
- Services suffered from hardware problems
  - (no fail over (design problem))

OSG Operations Workshop
DC summary

CERN IT-GD

December 2004 19

My Summary on CHEP-2004

+ The multi cluster environment is PHENIX 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.

Many flavors of grid systems (no 100% compatibility)

+ Grid2003
+ SAM
+ EGEE
+ NORDUGRID

….

SAM looks most working but …

SAM development was started at 1987…

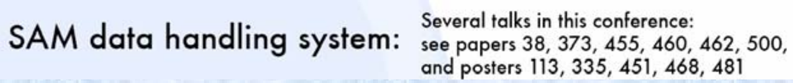
What was mentioned often …

+ Data handling issues:

D-cache;

xrootd;

SRM ([334] Production mode Data-Replication framework in STAR using the HRM Grid)



SAM data handling system: Several talks in this conference: see papers 38, 373, 455, 460, 462, 500, and posters 113, 335, 451, 468, 481

+ Security issues.

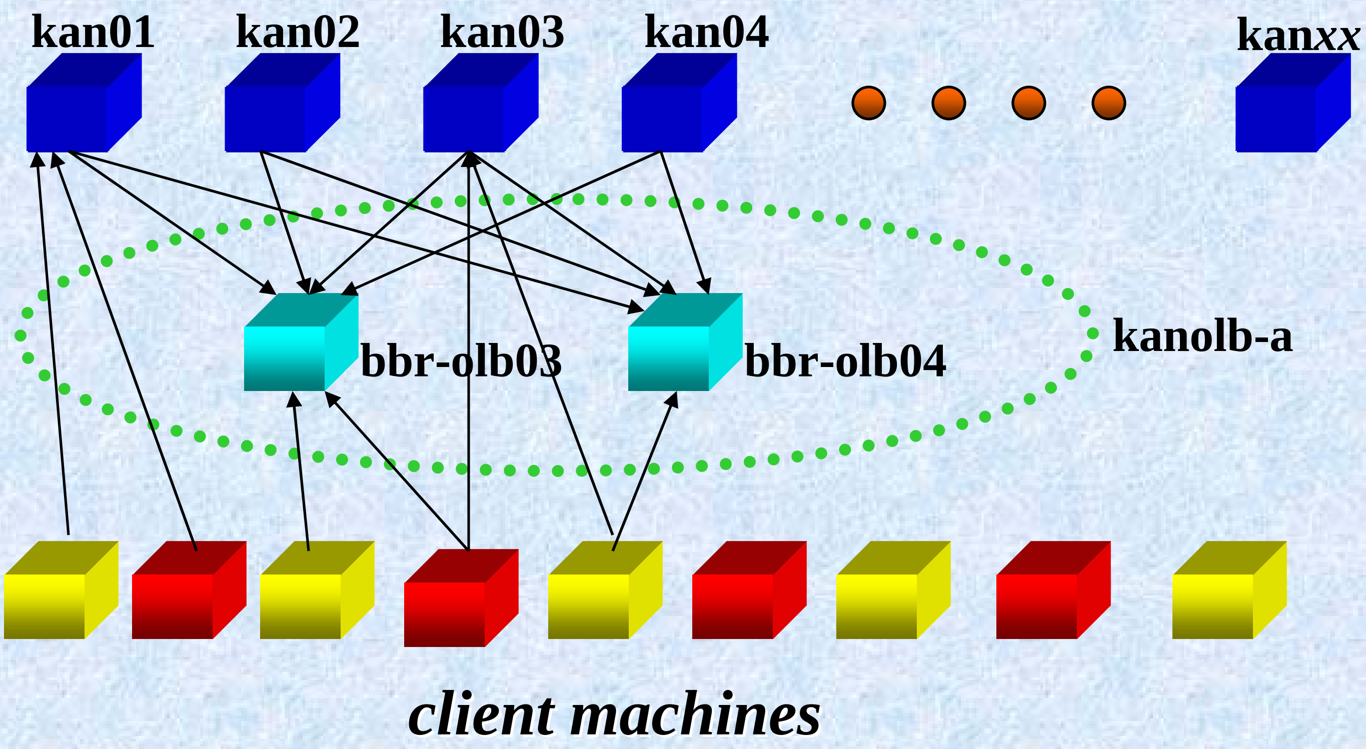
+ Grid Administration/Operation/Support centers.

+ Deployment issues.

Development hit – xrootd

(Example: SLAC Configuration)

http://xrootd.slac.stanford.edu/presentations/XRootd_CHEP04.ppt



Grid prospects

+ Many small problems are transformed into one big problem (Grid :-).
+ Advantages (point of balance of interests)
    - for funding authorities;
    - for institutes;
    - for collaborations;
    - for end users (physicists).

Estimates

Let us introduce the variables:

T - total time for the computing task with using only local cluster;

τ - reduced time for the computing with using additional cluster;

t_l - average time for processing of one portion of the data on the local cluster;

t_r - average time which is required to process the one portion of the data on remote cluster;

t_o - the average overhead time which is required to perform any additional operations (for example time for the data transfer) per one portion of the data on remote cluster;

D - total number of the data units which have to be processed;

S - speed: the number of the data portions at one unit of the time;

α - accelerating (speeding up) of the computing (in times) due to use additional cluster;

- for only local cluster:

$$S = \frac{1}{t_l}$$

and total time for the computing is

$$T = \frac{D}{S} = D * t_l$$

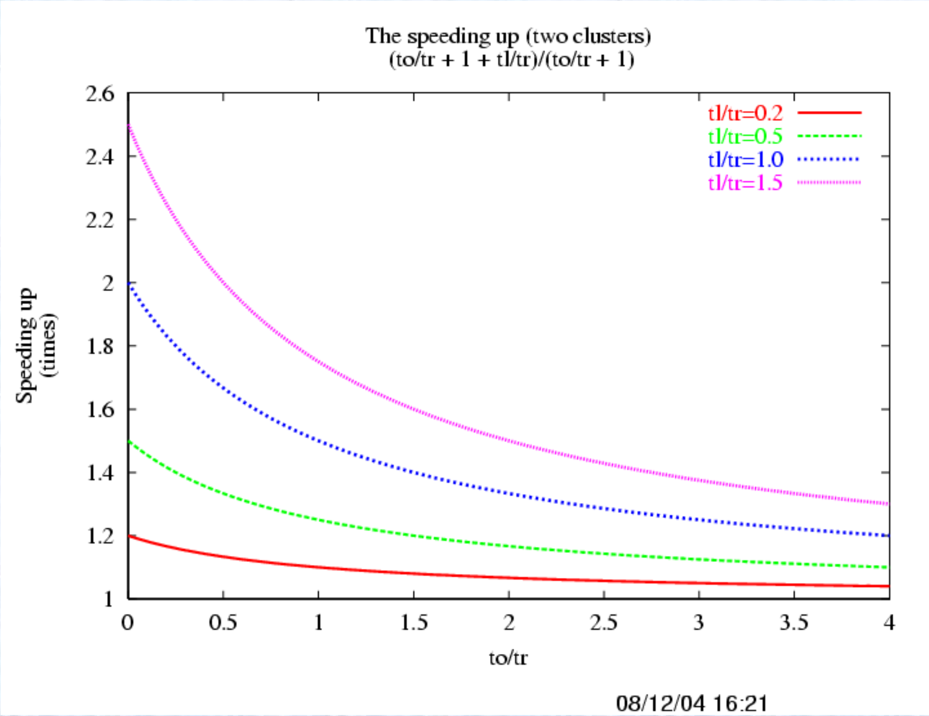
- for two clusters (local + remote)

$$S = \frac{1}{t_l} + \frac{1}{(t_o + t_r)}$$

and total time for the computing is

$$\tau = \frac{D}{(\frac{1}{t_l} + \frac{1}{(t_o+t_r)})}$$

$$\alpha = \frac{\frac{t_o}{t_r} + 1 + \frac{t_l}{t_r}}{\frac{t_o}{t_r} + 1}$$



Grid computing advantage (simulation versus analysis)

- The simulation on Grid structure implies high volume data transfer (i.e. overheads);
- On other hand the data analysis assumes limited data transfer (once for relatively long period, may be once per ½ year).

Conclusion
PNPI role in Grid

- Anybody who plans to participate in accelerator physics simulation/analysis has to learn the basics of Grid computing organization and collaboration rules where you plan to participate (to get Grid certificate as the first step).
- In order to do so HEPD has to keep up to date own computing cluster facility (about 10 TB of disk space and appropriate computing power) and external data transfer throughput 1-5 MBytes/sec.