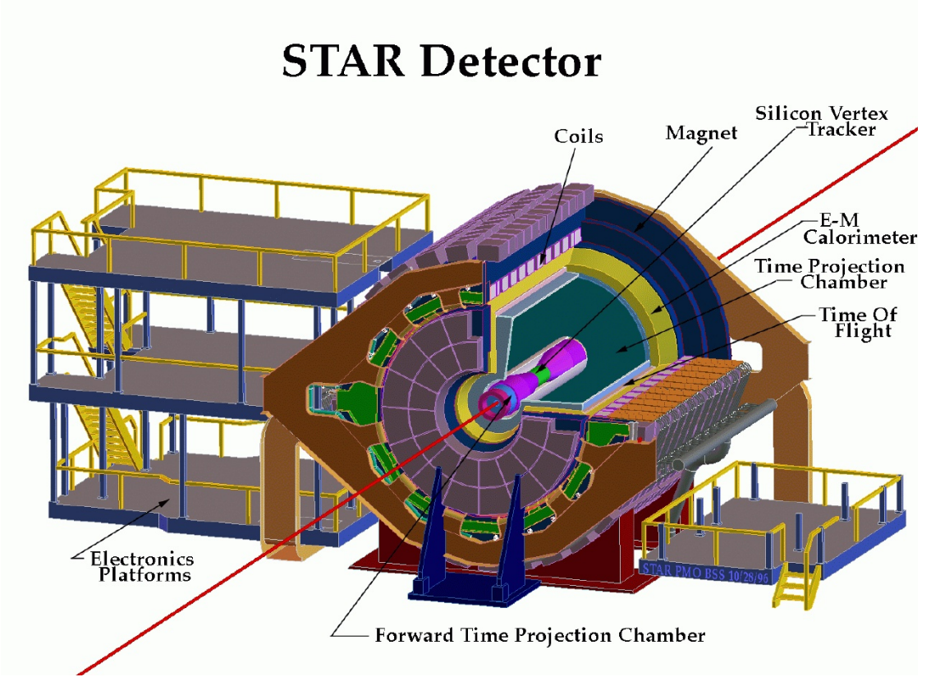
Setting up Tier2 site at Golias/Prague farm

Petr Chaloupka, Pavel Jakl, Jan Kapitán, Jerôme Laurét[2], Michal Zerola
NPI ASCR Prague, [2] Brookhaven National Laboratory

Introduction

The volume of scientific data acquired by High Energy Nuclear Physics (HENP) experiments at RHIC has been steadily increasing over past years reaching scale of hundreds of Tera-bytes per year and is expected to reach Peta-byte scales in the near future. Demands for data storage and subsequent processing in reasonable time are enormous and growing out of scope of a any single computing center.

Prague heavy-ion group participating in the STAR experiment has been a strong advocate of local computing as the most efficient way of data processing and physics analyses. In order to take full advantage of available local resources a Tier2 computing center has been set up at a regional Computing Center for Particle Physics (Golias farm). It is the biggest site in the Czech Republic fully dedicated for particle physics experiments.


STAR Detector

STAR experiment

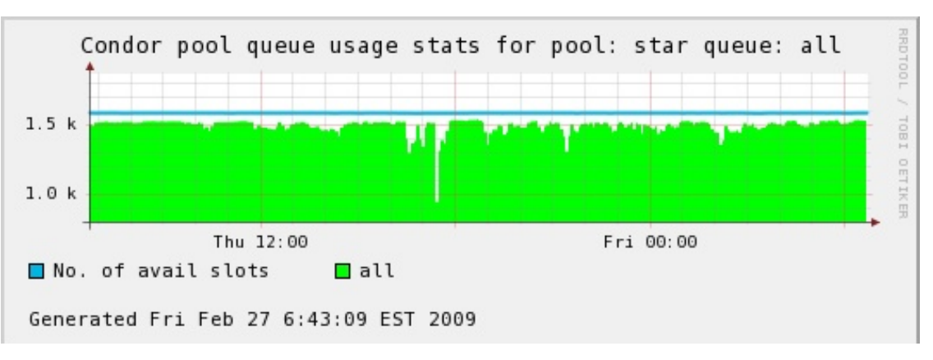
- International collaboration of ~600 scientists from 55 instututins in 12 countries
- Produces 1 PB of data each year and grows
- Already 15 PB of stored data
- Over 17 Millions of files
- Running ~1500 jobs concurently

Present status of STAR computing at RCF

The RHIC Computing Facility (RCF) at Brookhaven National Laboratory is the Tier0 computing facility for the STAR experiment.

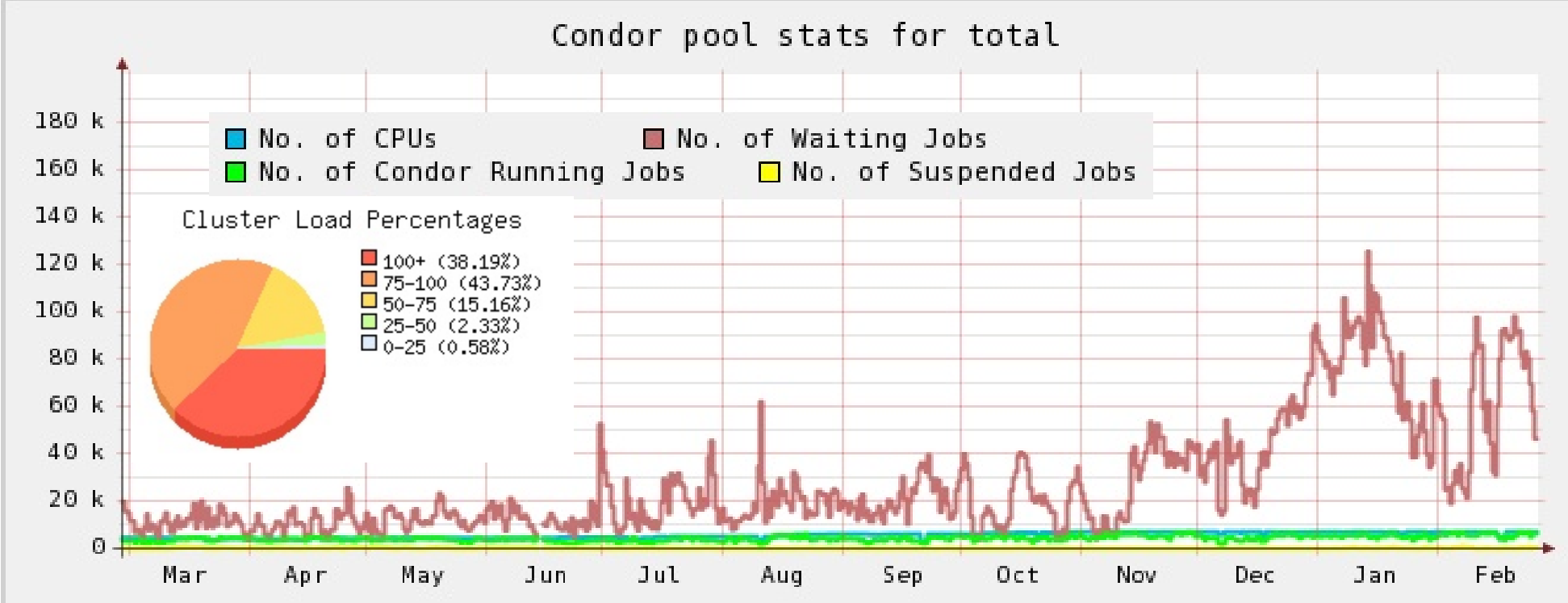
RCF resources for STAR:

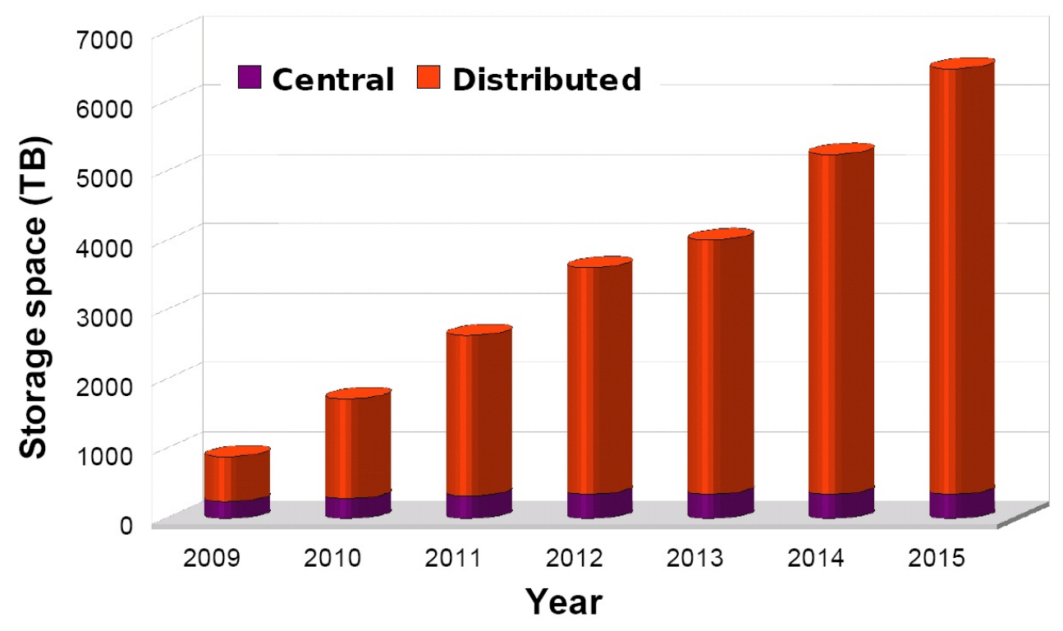
- Linux computing farm - 345 computing nodes/1500 CPU for reconstruction and physics analyses
- HPPS - high capacity tape storage
- centralized and distributed disk storage - 500TB


Condor pool queue usage stats for pool: star queue: all

Generated Fri Feb 27 6:43:09 EST 2009

Steadily growing amount of taken experimental data increases demand for data storage and processing power and decreases resources available to individual users for their physics analyses. This together with need to work remotely over often lagging network renders work of far away collaborators to be cumbersome and ineffective.


Condor pool stats for total



At many places computing and storage resource are available at smaller local computing farms that can be utilized by Tier2 centers for data processing and physics analyses of the local physics groups.

Paradigm of local computing: "bringing data to users"

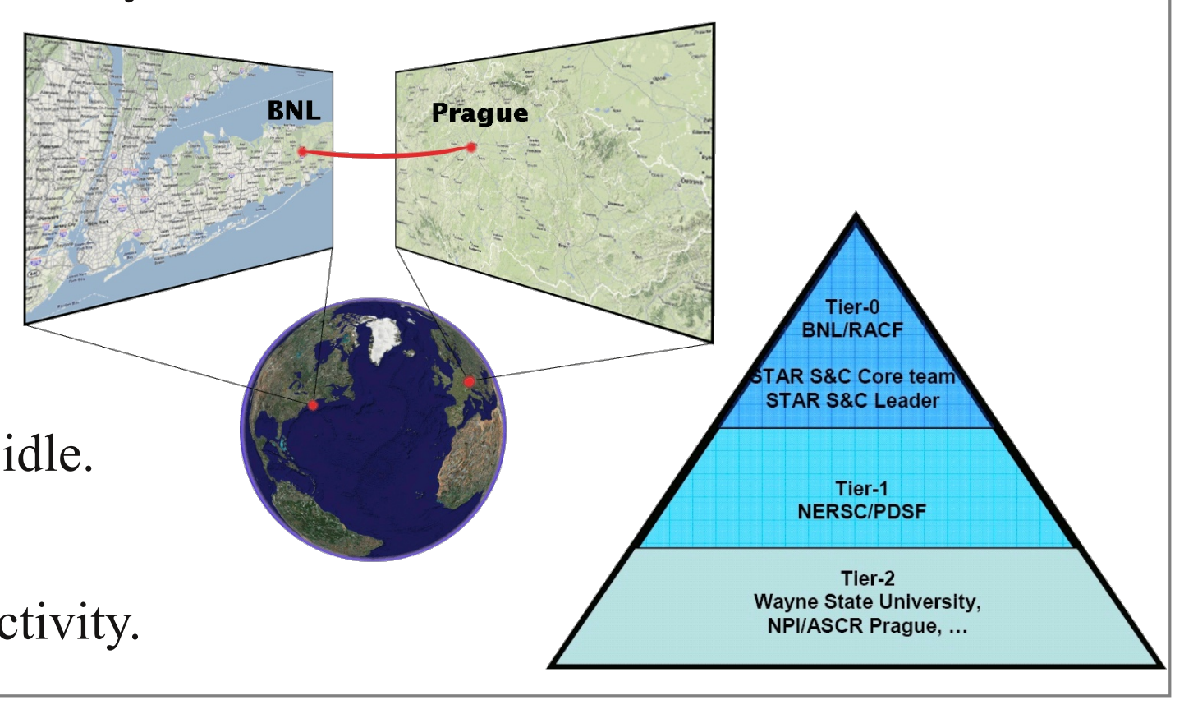
Leveraging of local resources for running physics analyses and simulations on localy stored data.

Utilization of local resources:

- Storage and computing elements
- Use of already existing infrastructure (network, cooling..).
- Local manpower

Gains from local computing:

- Share CPU power - use extra CPU when in need, borrow to others when idle.
- Sharing of maintenance costs with other experiments.
- Fast interactive work, reduce development cycles, increase science productivity.



Golias farm at Prague

- **National computing center** for processing data from various HEP experiments. Located in the Institute of Physics in Prague. Officially started in 2004, basic infrastructure already in 2003.
- Certified as a Tier2 center for LHC Computing Grid Collaboration with several Grid projects: LCG, EGEEII (Enabling Grids for E-science Europe) and WLCG (Worldwide LHC Computing Grid).
- Provides computing services for ALICE, ATLAS, D0, AUGER, H1, STAR, Solid State Physics.

- **CPU:**

  end 2008: Golias in total 450CPU and 50TB storage.

  For Heavy Ion Physics group (Alice and STAR) 220 CPU: about 2% of the total ALICE CPU power.

  expected expansion during 2009: 1550 CPU and 216 TB.
- Excellent **network connectivity** provided by the CESNET syndicate:

  1 Gb/s to Europe network GEANT2.

  10 Gb/s CESNET central site: FNAL, BNL, GRIDKA FZK Karlsruhe, Taipei,Inst.of Phys., ChU, CTU,NPI Rez/Prague.
- **Middleware/grid tools**:

  gLite: CE, SE: classical + DPM, BDII, vo-boxes (ALICE, ATLAS), LFC (ALICE), PBSPro, AliEn.
- **Batch system PBSPro**: configured to keep all CPUs job-busy. So, when your project is at the moment the only one having jobs in the batch queue, you get all the free CPUs regardless of your official share.

STAR computing at Golias - solution components

- Local CPU: sharing of idle CPU power between STAR and ALICE
- Local data storage - Disk Pool Manager (DPM)

  18 TB of cost-effective storage space
- GRID aware data transfer tools - RFIO protocol
- STAR software framework:

  locally compiled STAR libraries

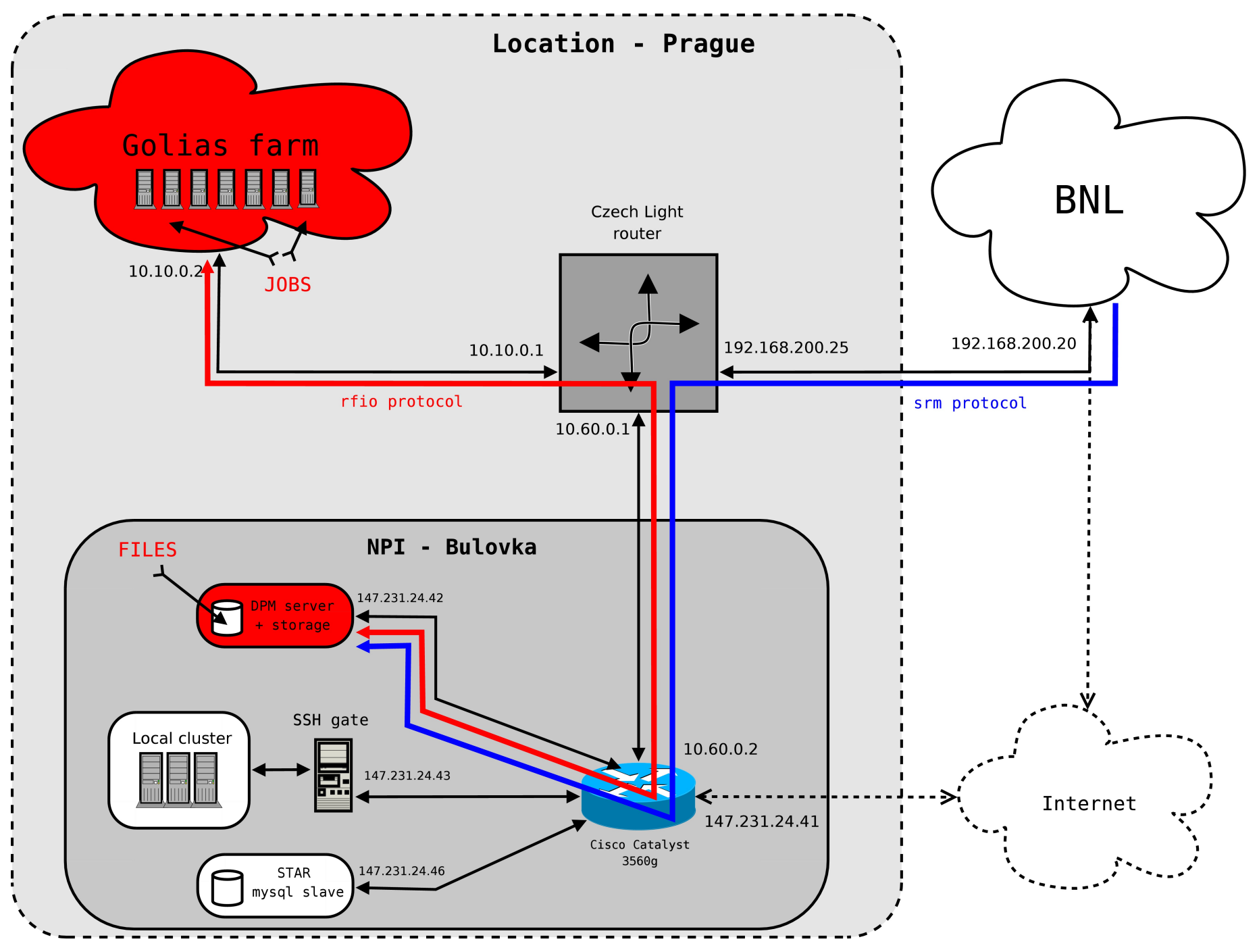
  root4star - experiment specific version of ROOT supporting RFIO protocol

  STAR Unified Meta Scheduler (SUMS)
- Local MySql database mirror of STAR database

  allow for efficient running od detector simulations
- High speed/low latency network connection:

  between DPM and Golias

  to Tier0 center and CERN



Local storage solution: DPM

Key component in our local computing strategy: Use of low cost solutions for high-capacity storage: servers packed with high-capacity disk arrays (no fancy expensive systems like e.g. HP StorageWorks, NAS).

Hardware:

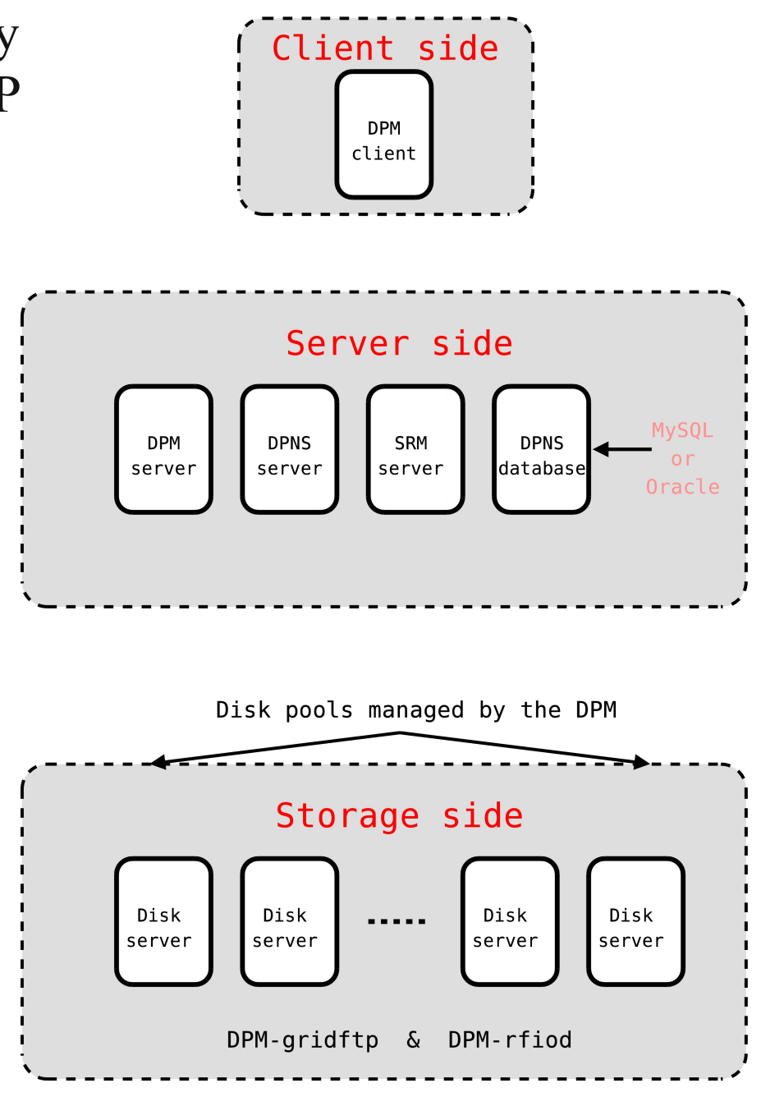
LSI MegaRAID with 48 SATA 500GB disk, hot-swappable
cost ~20kCZK/1TB (700USD/1TB)

Downsides:

needs possibly more maintenance then expensive solutions.
data pool controlled by database - longer access time
no kernel drivers - not mountable like NFS

Upsides:

easy adding of new data pools
not mounted to each work node - light on network traffic
supports RFIO protocol - directly accessible from ROOT

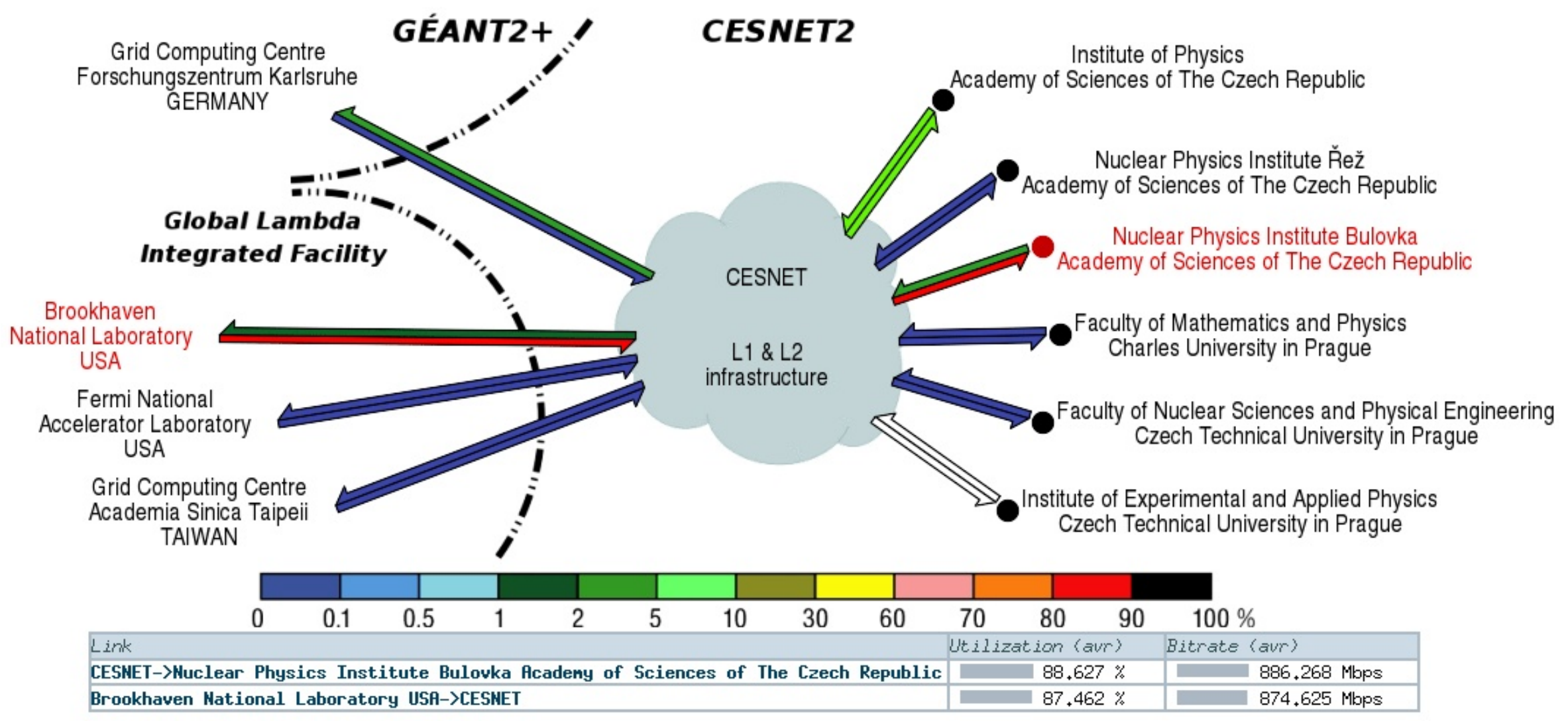


Direct 1Gb/s connection to GOLIAS – seen as a component of the farm from outside

Network connection

Experience with remote access, editing, job submission and retrieval of results from Tier0 center over standard Internet connection with high latency led to considering the option of local computing.

- 1Gb/s connection between DPM storage and Golias farm
- since 2006 connected to CzechLight network
- dedicated link between BNL and Golias allows for full 1Gb/s data transfer between Tier0(BNL) and Golias using SRM tools


GÉANT2+
CESNET2
Grid Computing Centre
Forschungszentrum Karlsruhe
GERMANY
Global Lambda
Integrated Facility
Brookhaven
National Laboratory
USA
Fermi National
Accelerator Laboratory
USA
Grid Computing Centre
Academia Sinica Taipeii
TAIWAN
CESNET
L1 & L2
infrastructure
Institute of Physics
Academy of Sciences of The Czech Republic
Nuclear Physics Institute Řež
Academy of Sciences of The Czech Republic
Nuclear Physics Institute Bulovka
Academy of Sciences of The Czech Republic
Faculty of Mathematics and Physics
Charles University in Prague
Faculty of Nuclear Sciences and Physical Engineering
Czech Technical University in Prague
Institute of Experimental and Applied Physics
Czech Technical University in Prague
0
0.1
0.5
1
2
5
10
30
60
70
80
90
100 %

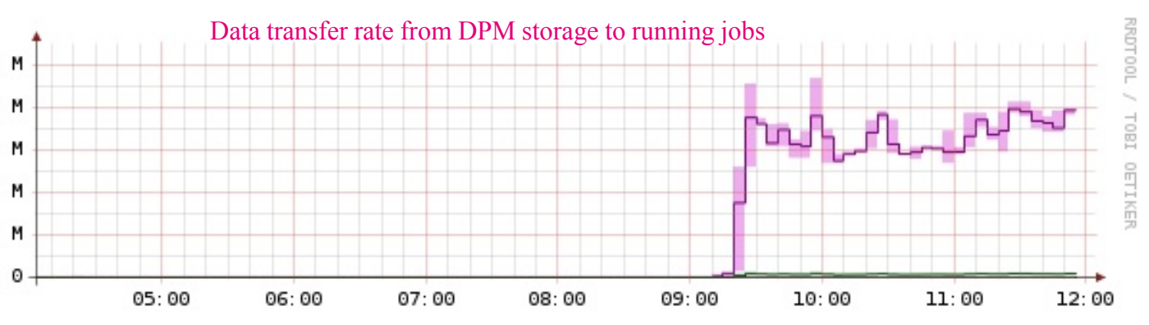
Link	Utilization (avr)	Bitrate (avr)
CESNET->Nuclear Physics Institute Bulovka Academy of Sciences of The Czech Republic	88.627 %	886.268 Mbps
Brookhaven National Laboratory USA->CESNET	87.462 %	874.625 Mbps

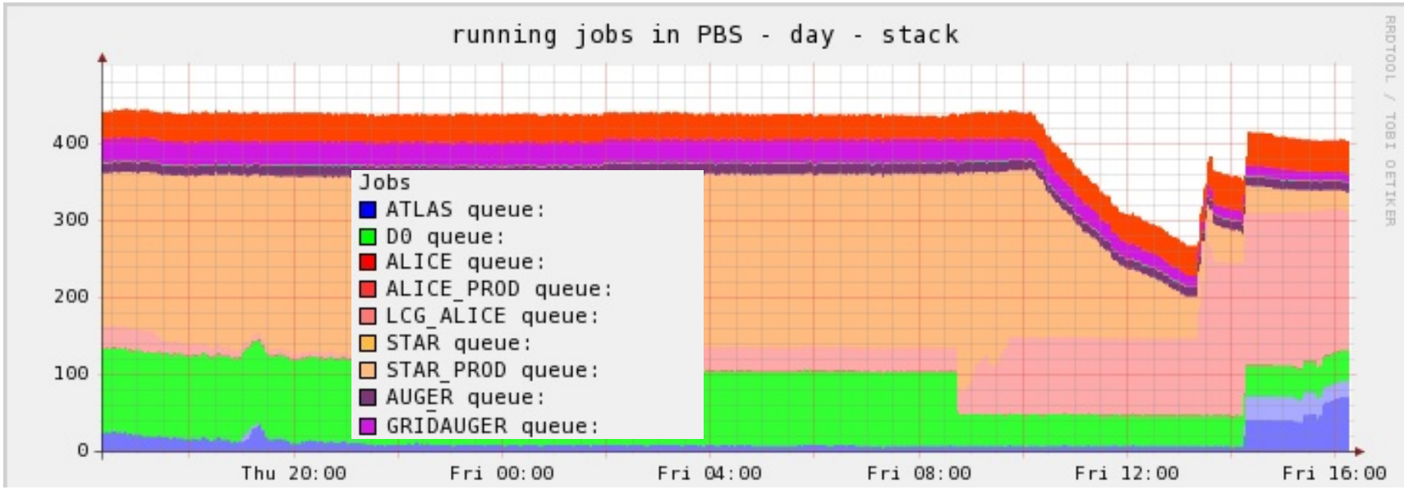
Performance

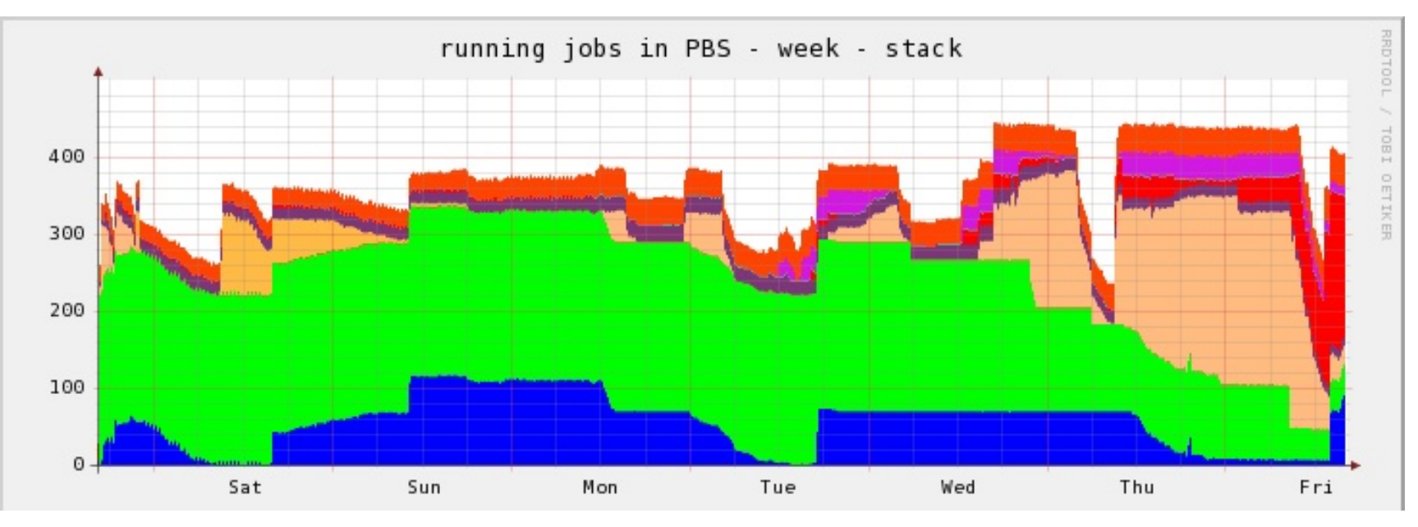
- Installation of complete STAR software framework Golias including job scheduler allows for very easy migration of jobs between Tier0 and Tier2.
- Setup of farm allows for using of idle computing power. For the STAR experiment this means maximum of about ~200 running jobs when needed.
- Always ready minimal computing power, corresponding to actually bought CPU, even when farm is under full load.
- The setup was tested by the computing needs of the Prague physics group of the STAR experiment in a conditions typical for small to mid-size physics working group. While such a group does not need all time high processing power, it needs intermittent on-demand access to computing resources. The most common task performed at the STAR/Golias site are physics analyses on locally stored, usually preprocessed data sets on the order of ~ 10-1000GB.
- The system is capable of handling ~200 running jobs with data transfer of up to 1Gb/s from the main storage to the running jobs.
- Perceived future bottlenecks:

  running more then 200 jobs requires appropriate increase in data transfer rate. The database driven DPM - throughtput may not be sufficient.

  manpower: Increase in the low cost storage capacity in future may require more service.


Data transfer rate from DPM storage to running jobs


running jobs in PBS - day - stack


running jobs in PBS - week - stack

Conclusions and perspectives

- The setup of the local STAR Tier2 at Golias/Prague demonstrates feasibility and usefulness of the local computing approach.
- Sharing of resources help to utilize locally available resources maximum and proves especially useful for running analyses of local physics group that needs intermittently large computing power.
- The low cost storage solutions using Disk Pool Manager - DPM has shown to be light-weigh, yet reliable enough, solution to handle data transfers up to 1Gb/s from Tier0 and to running jobs.
- In future it is perceived that GRID computing will be a viable option for STAR computing at Golias.

> With minimal effort and concentrating on moving data alone, we can go a long way to support a very active and vibrant Physics program as well as sustaining Computer science

The authors wish to thank the team of Golias farm for their support of the STAR Tier2 project. We also thank the BNL computing team for continuous support and development of the off-site computing. We would like to extend our gratitude to involved people from ESNet, CZNet.

Petr Chaloupka, Michal Zerola, Jan Kapitán, Pavel Jakl - NPI ASCR, Prague

CHEP09, Prague