Benefits of grid technology to running experiments

CHEP
September 5th, 2007

Frank Würthwein (UCSD)

Basic Question:

To what extend have running experiments benefited from the LHC investment into grid technology, and deployment thereof?

Many thanks to:

- Hartmut Stadie (ZEUS)
- Christoph Wissing (H1)
- Keith Chadwick (Fermigrid)
- Simone Pagan Griso (CDF)
- Donatella Luchessi (CDF)
- Petar Maksimovic (CDF/CMS)
- Amber Boehnlein (D0)
- Andrew Baranovski (D0)
- Carla Vale (PHENIX)
- Jerome Lauret (STAR)
- Ian Bird (EGEE)

… for their help with this talk!

Outline

- Benefits from Middleware
- Benefits from Infrastructure
- Summary & Conclusion

No attempt is made to be complete. Instead, cover topics by example only.

See also chep 2007 contribs:
BaBar: 48, 344, 368, 455, 456
Belle: 464
BES: 58

CDF: 295,297,298,427
D0: 227
Fermigrid: 234
PHENIX: 219
ZEUS: 336

Benefits from Middleware

Examples from:

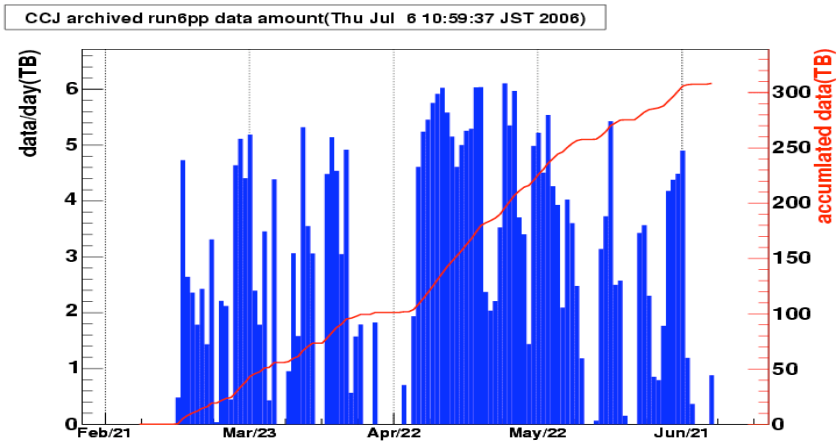
PHENIX & STAR for “traditional” Middleware use.

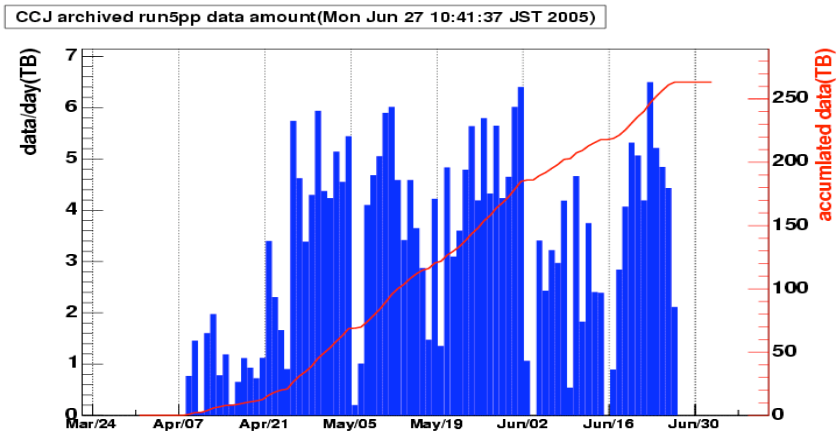
Fermigrid for “unusual” Middleware use.

CDF for “using the SQUID”.

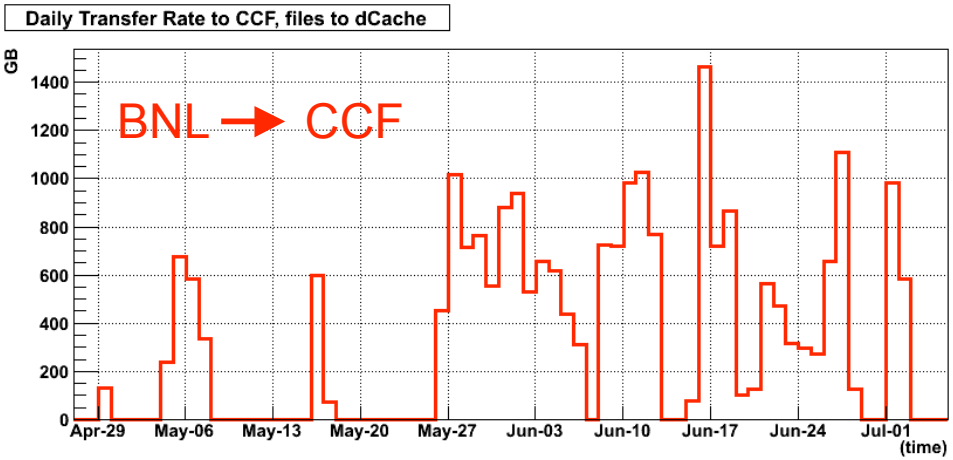
Gridftp & SRM

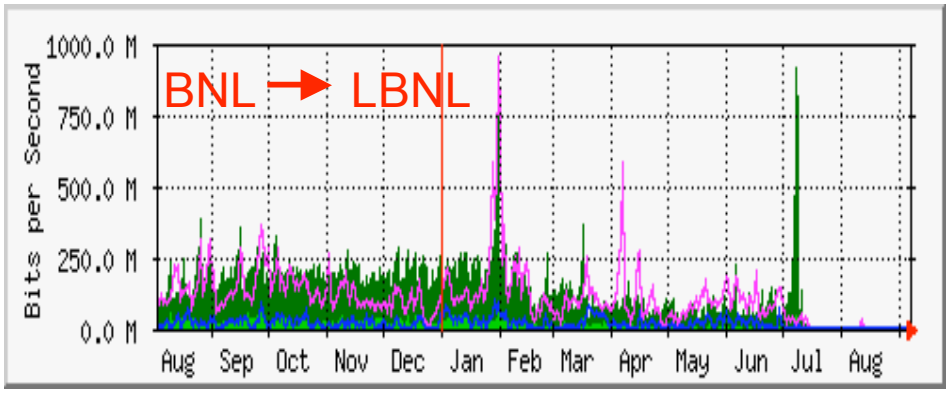
PHENIX "online" transfer
BNL→CCJ @ 2-6TB/day





PHENIX & STAR
data transfers via SRM.



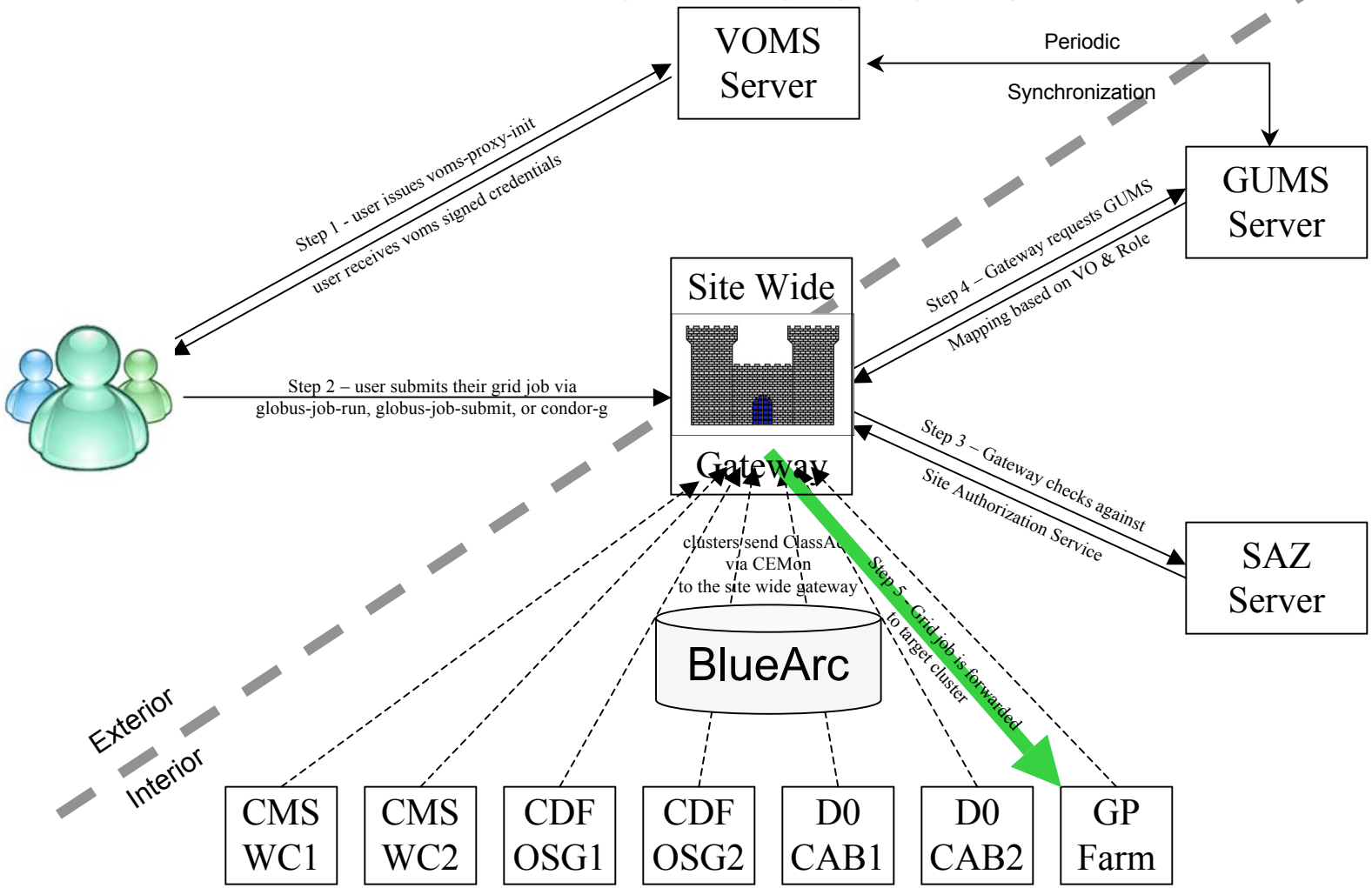


Typically ~0.5TB/day with peaks of ~250Mbps

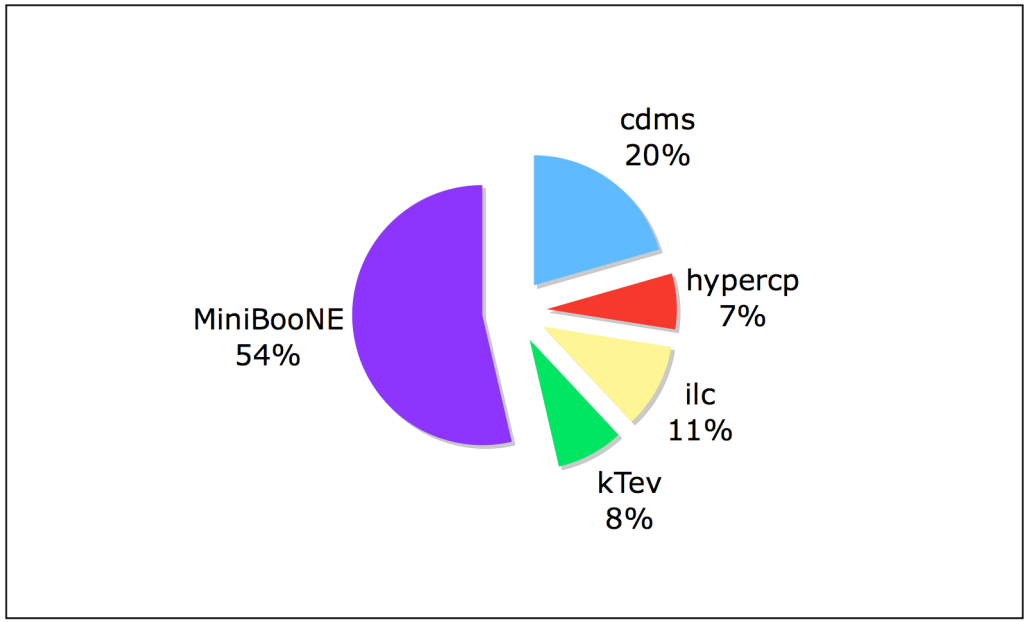
Fermigrid

- In the past, CDF, CMS, D0 owned and operated their own CPU resources at FNAL.
- Since 2005 migration towards grid technology for simpler and more uniform operations.
  - Model of a "campus grid".
  - Expand benefits and knowledge of grid technologies at the lab.
    - Investment in operations aspects, e.g. security, accounting, user management.
    - Investment in "gateway" that aggregates local clusters

FermiGrid - Current Architecture



Fermigrid Utilization

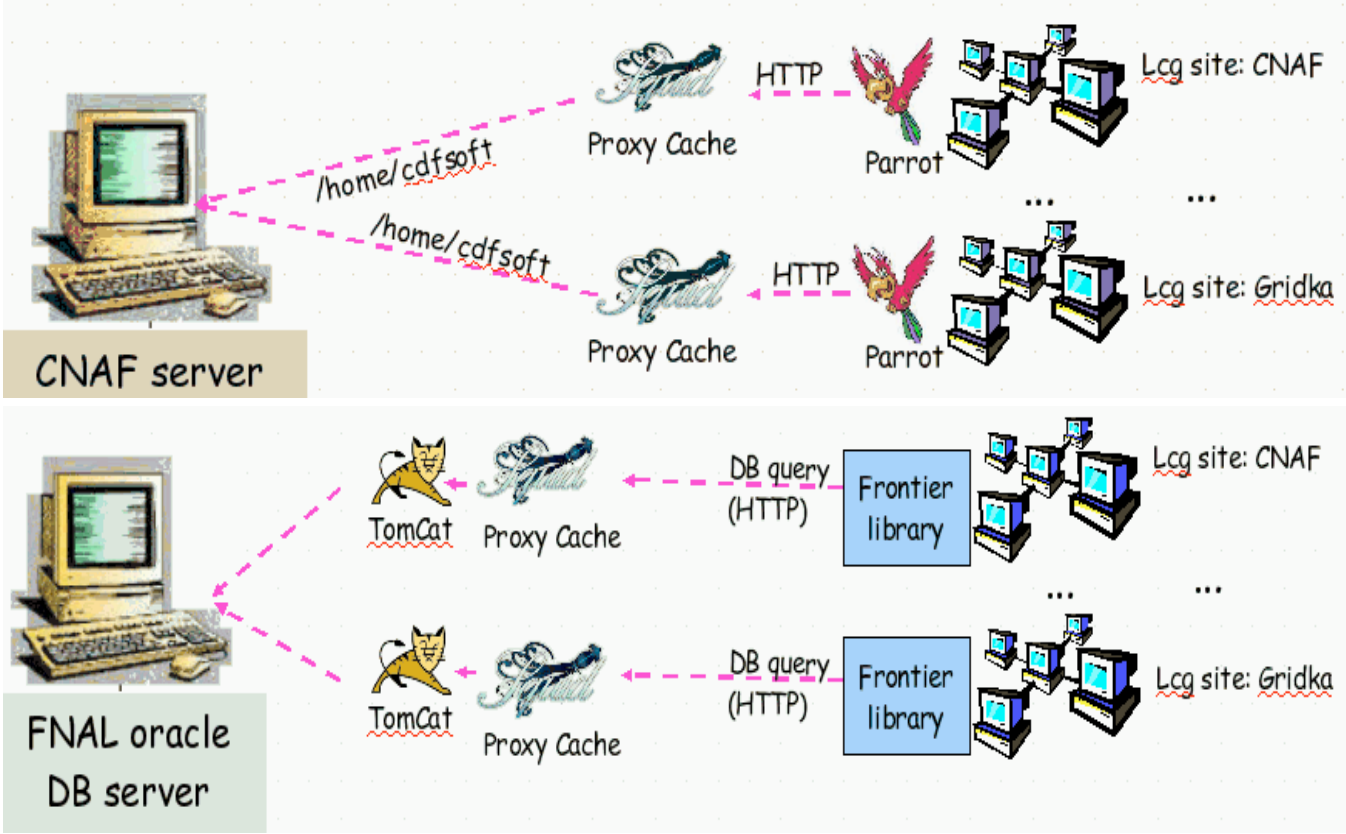


CDF, CMS, and D0 are excluded to make this pie chart.

In addition to HEP & Astro, Fermigrid hosts VOMS for:
Nanohub = Nanotechnology
Gadu = Bioinformatics
I2U2 = Education
OSG = variety of individual PIs

- Up to 12 different VOs per day.
- Up to 230 different users per day with certificates from up to 20 different CAs were seen within the last few months.
  - Note: CDF & D0 still use their resources at FNAL in part via legacy access mechanisms instead of grid.

3 times SQUID @ CDF



Only on EGEE

EGEE & OSG

(same as CMS)

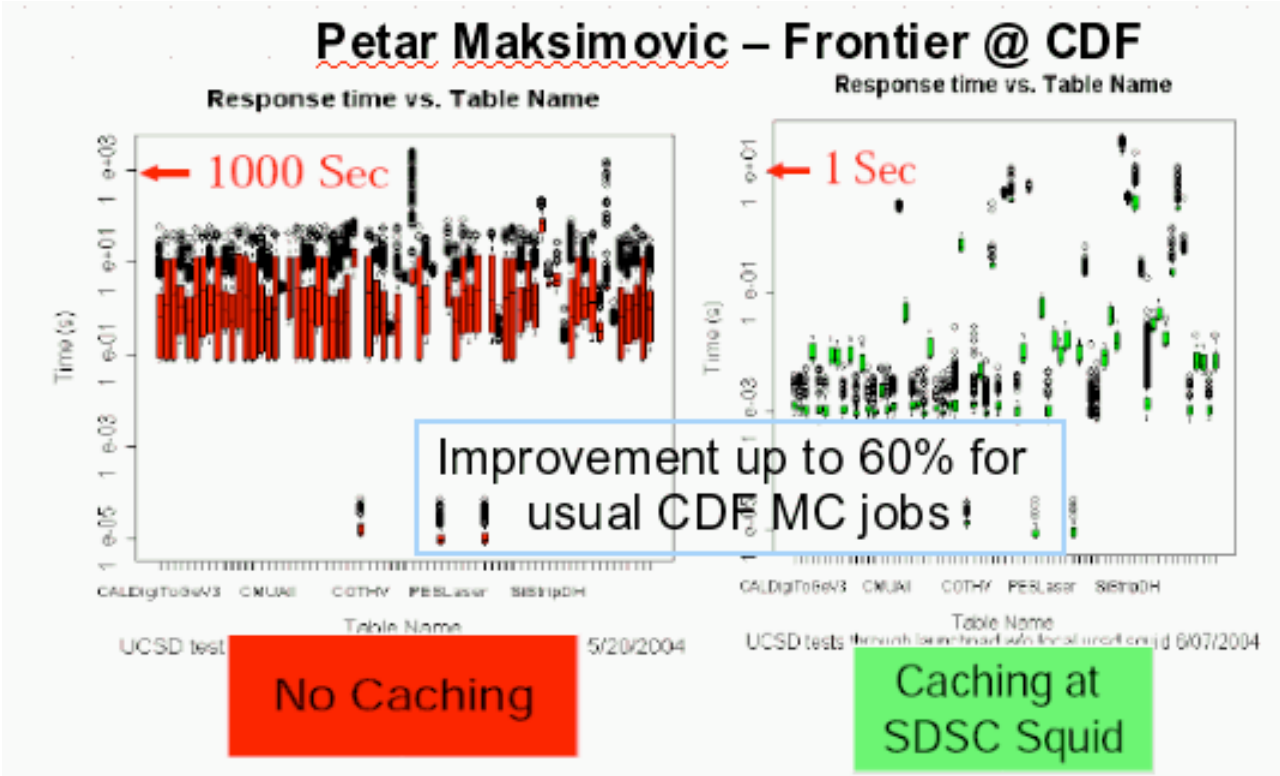
Serving of user application tarballs.

Only on OSG

Performance gains from SQUID

WAN access to remote DB leads to latency dominated apps.

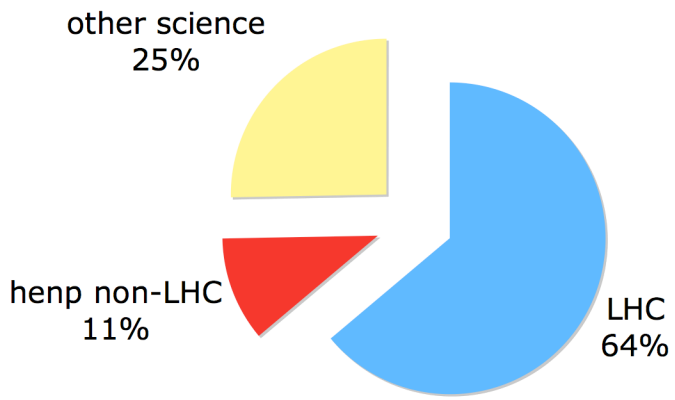
- ~90% cache hit rate for typical CDF apps.
- Up to 60% improvement in execution time at UCSD. Gains probably even larger for more remote sites.



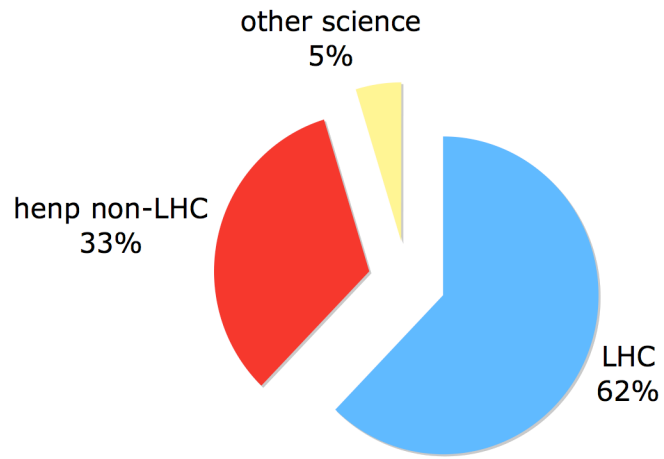
Benefits from Infrastructure

Grid Utilization

EGEE



OSG

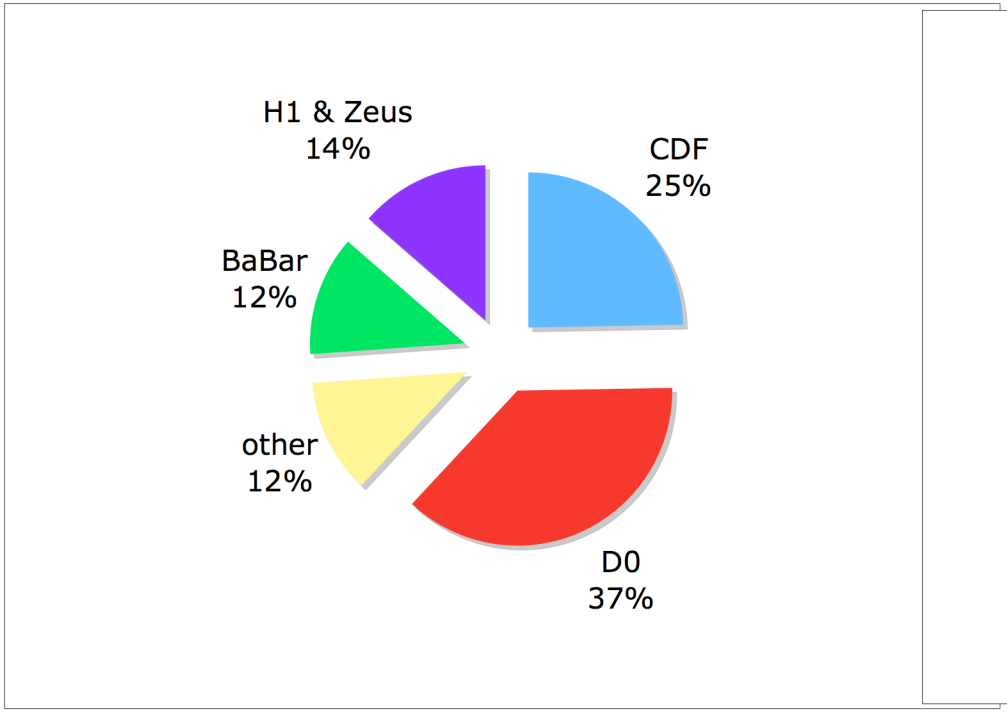


Based on wall clock time.

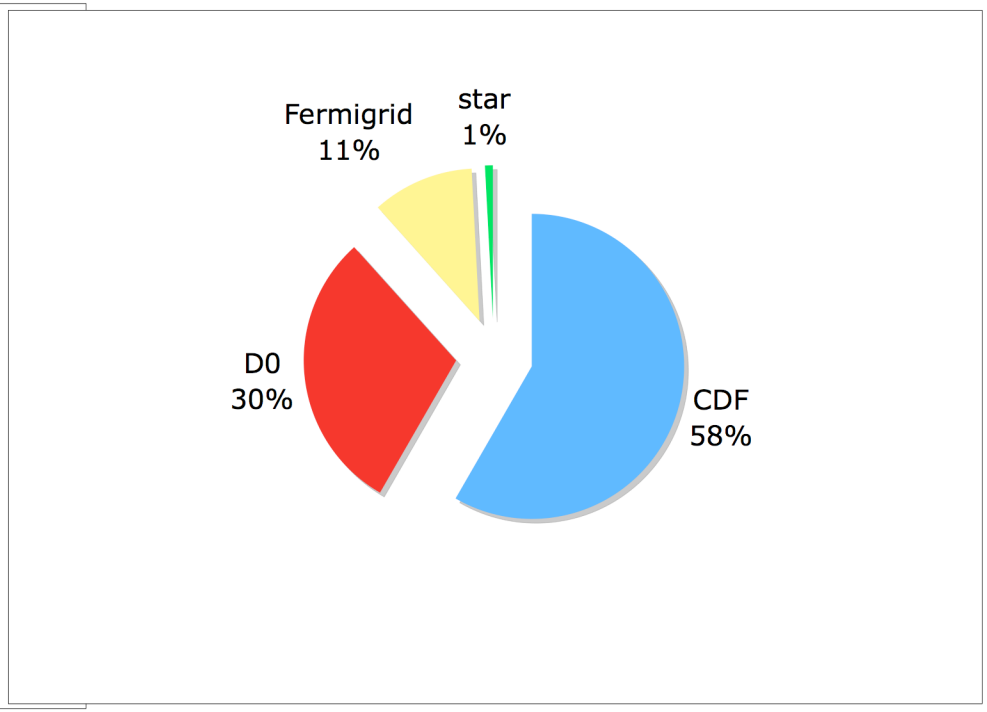
HENP other than LHC is significant fraction of total CPU utilization.

HENP non-LHC

EGEE



OSG



HENP other than LHC is significant fraction of total CPU utilization.

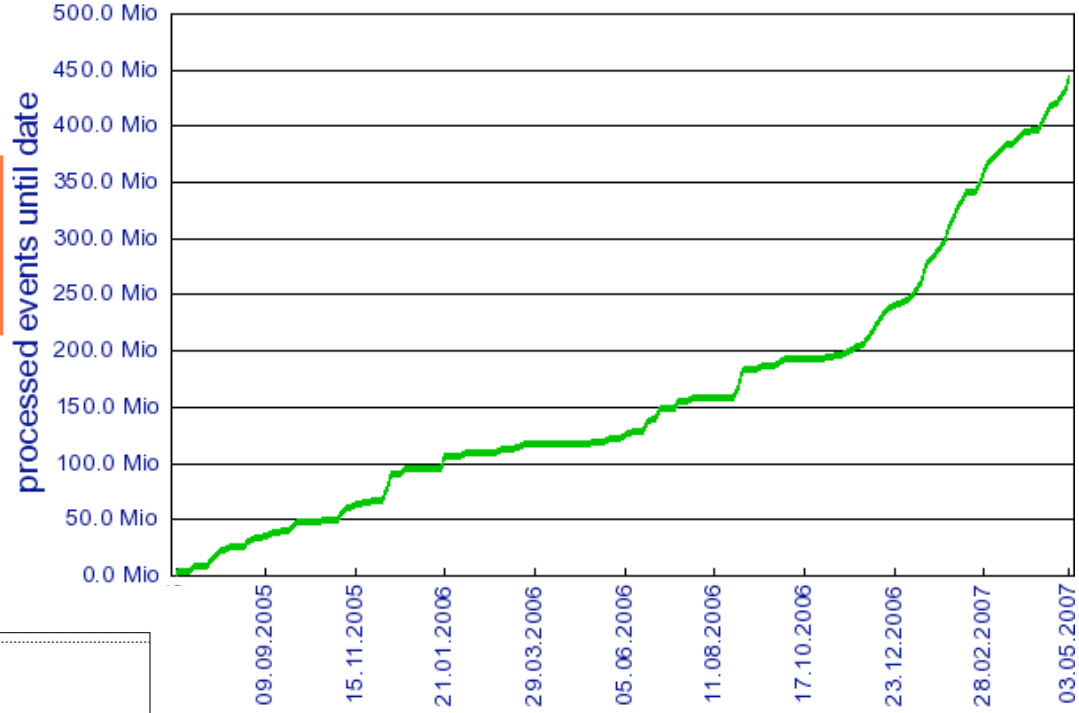
Some Commonalities

- All Experiments tend to generate sizeable fractions of their MC samples on the grid.
- Integrate into legacy operations and interfaces to allow for transparent operations on dedicated and grid resources.
- Focus on resources from collaborating institutions that are made available via grid.
- Implement automated retries to “hide” lack of reliability of grid.

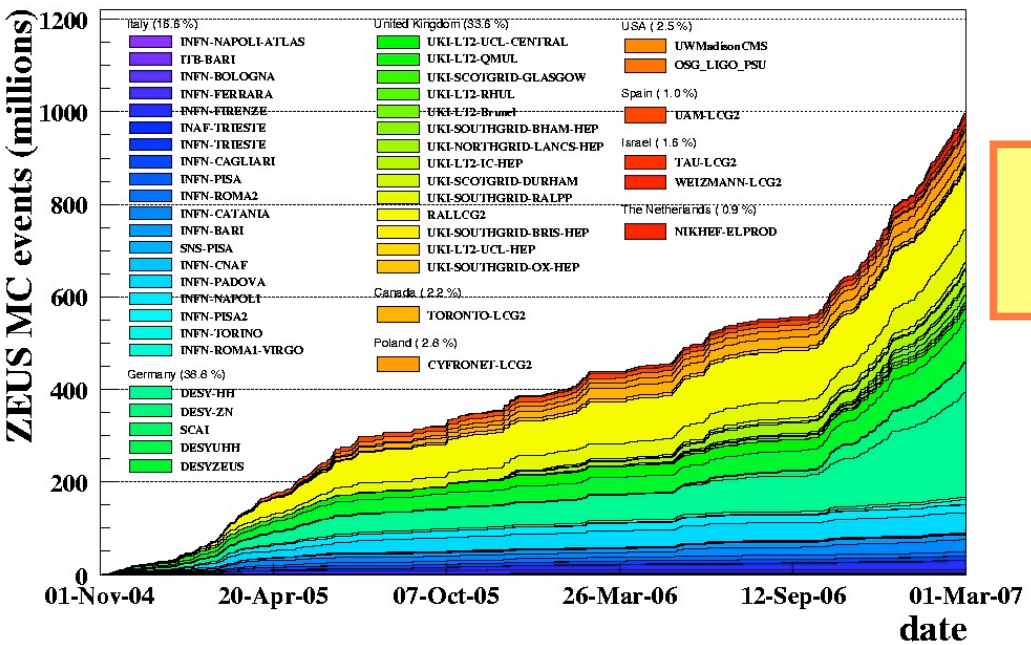
Physics programs crucially depend on MC production on the grid.

H1 & ZEUS MC Production on the Grid.

Almost 0.5 Billion Events by H1
~ 25 TB MC data in last 12 month



Before retry: ~80% success
After retry: 99.9% success!



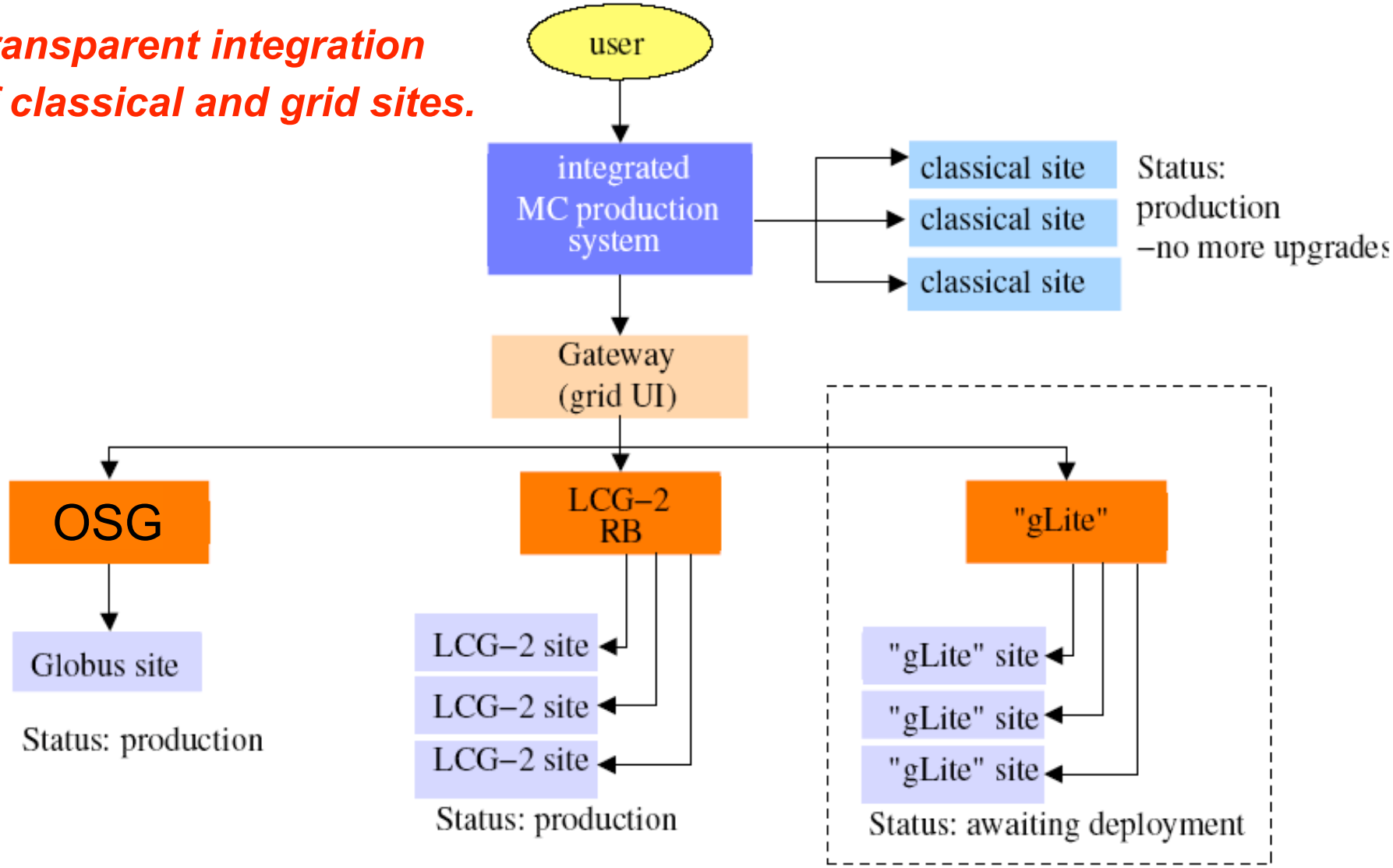
1 Billion Events by ZEUS
~ 80 TB MC data in last 12 month

Shared tools:
ZEUS Grid-Toolkit

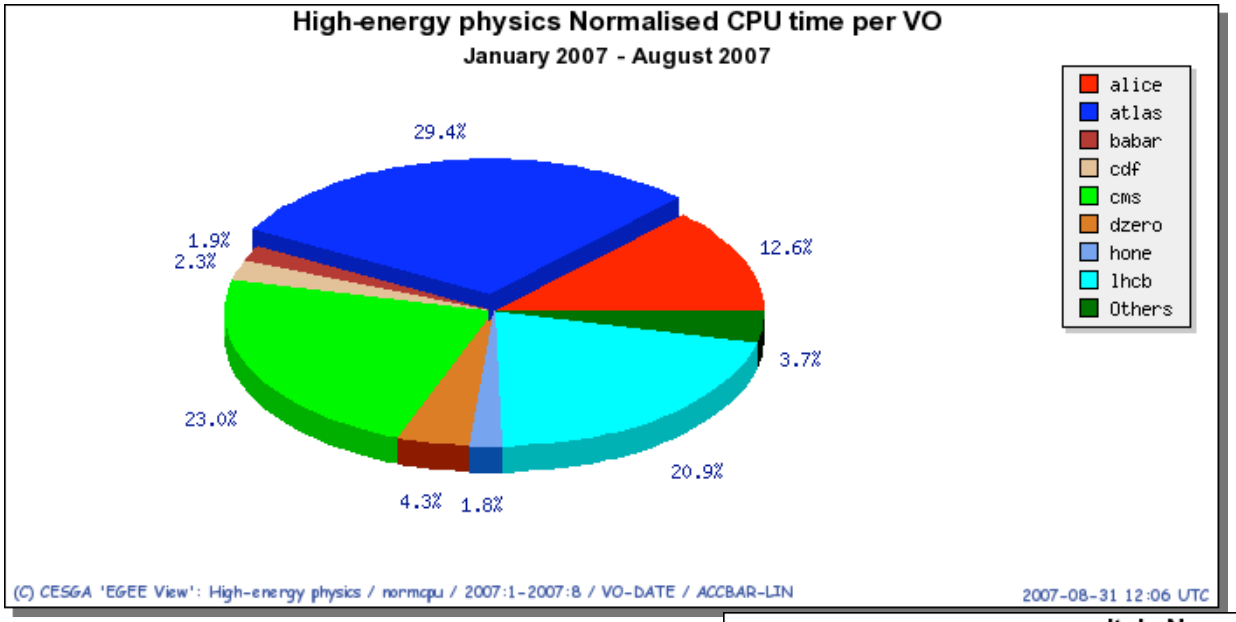
http://www-zeus.desy.de/~wrona/grid/index.php

Bridge to the Grid

Transparent integration of classical and grid sites.

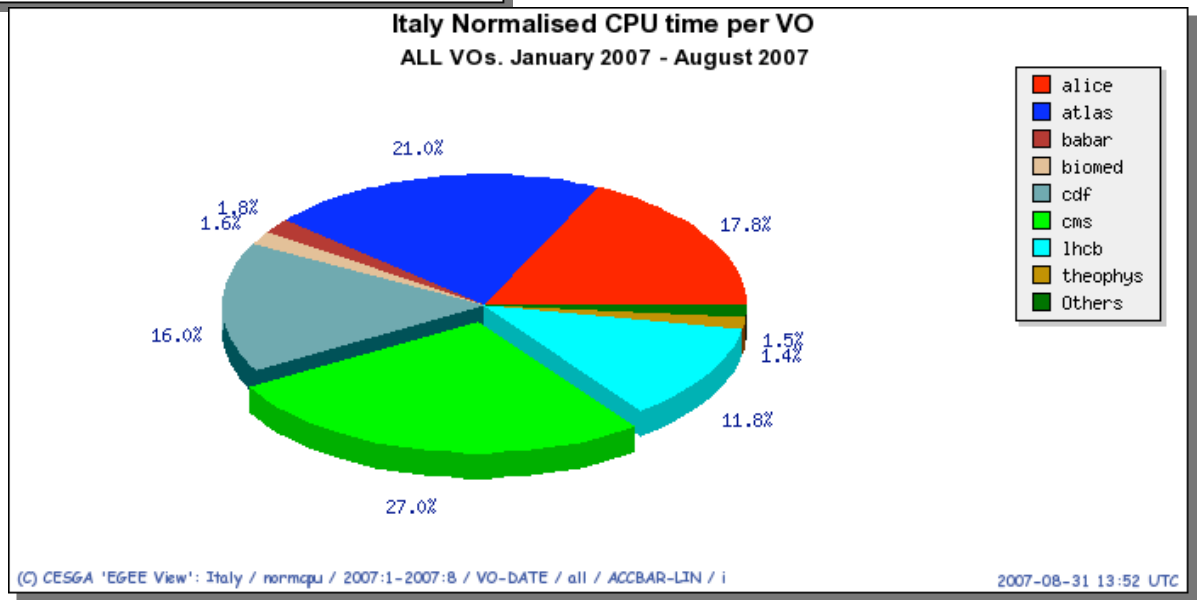


CDF focus on “friendly” sites



2.3% of EGEE total

16% of EGEE in Italy



Simone Pagan Griso

Some Differences

- CDF, STAR, H1, ZEUS use experiment wide MC production teams on the grid.
- CDF also has MC production by individual users for individual analyses on the grid.
- D0 focus on data reprocessing on the grid
- ZEUS, CDF data analysis at host lab via grid interfaces.

Example: D0 re-processing

In Nov ‘06 the D0 experiment asked the OSG to use 1.5-2K CPUs for 2-4 months for re-processing of a dataset (~500M events) for the summer conferences in July ‘07. D0’s own resources were committed to the processing of newly acquired data and analysis of already processed datasets.

By the end of May, re-processing of 445M events was completed. OSG contribution to this effort was

- 286M events
- 286k jobs executed
- 2M CPU hours
- 48TB of input data
- 286M files of final results
- 22TB of output data

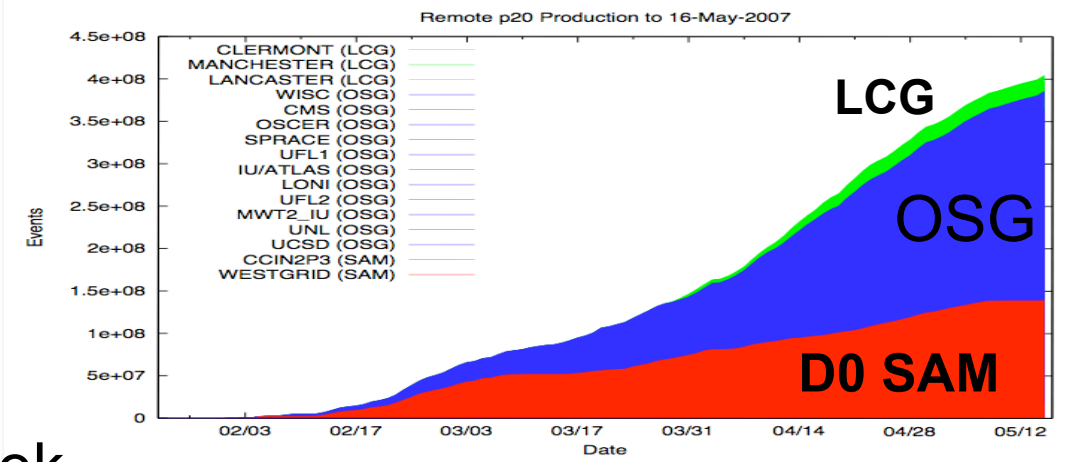
How did the D0 Reprocessing happen?

- The Executive Board estimated there were currently sufficient opportunistically available resources on OSG to meet the request; **They also looked into the local storage and I/O needs.**
- The Council members agreed to contribute resources (processing, data and FTEs) to meet this request.
- D0 had 2-3 months of smooth production using >1,000 CPUs.
- To achieve this
  - D0 testing of the integrated software stack took until February.
  - OSG and D0 staff then worked closely together to reach the needed throughput goals - facing and solving problems
    - sites - hardware, connectivity, software configurations
    - application software - performance, error recovery
    - scheduling of jobs to a changing mix of available resources.

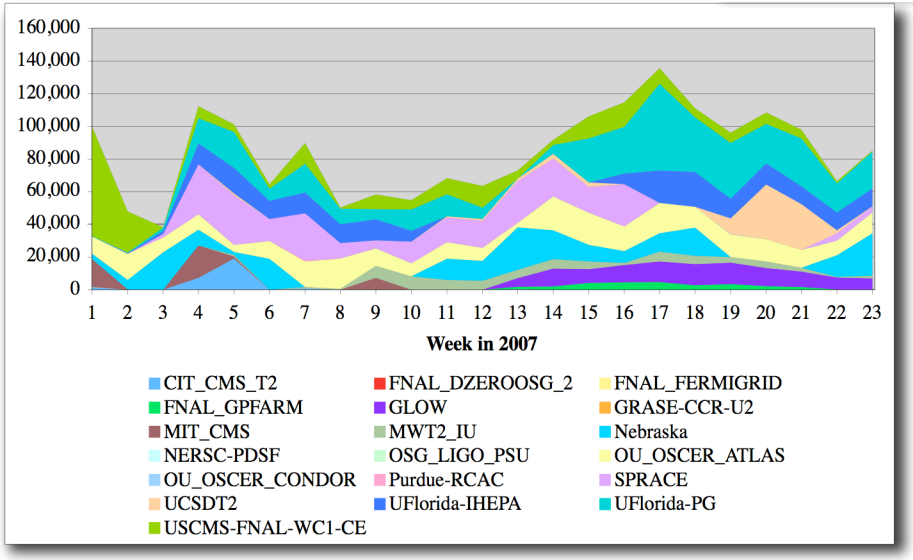
D0 Re-Processing

More than 50% of reprocessing done on OSG …

Total Events



OSG CPUHours / Week



.. most of it on resources that would not have been accessible to D0 without OSG.

E.g. MIT, UCSD, Wisconsin, Purdue, Florida are LHC resources at CDF institutions.

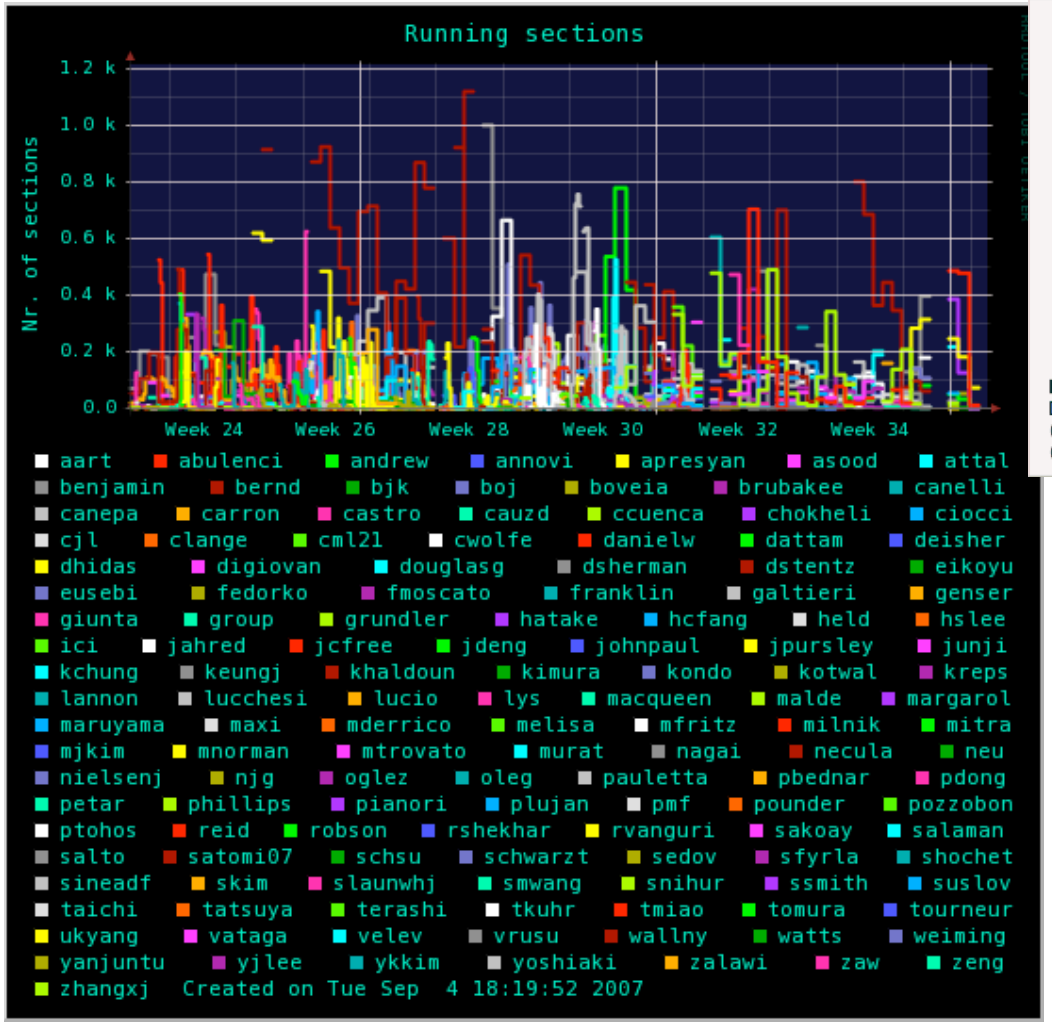
Problems Encountered by D0

- During Start-up phase
  - ~50% of problems related to site set-ups.
    - Work with OSG trouble shooting team and site admins to address those.
  - ~50% due to data delivery failures
    - Tune data handling queue for each site.
    - Maximize use of local SEs (SRMs)
    - “Certify” sites according to transfer performance.
- Throughput Production
  - Over/under subscription of jobs to sites.
  - “unscheduled” downtimes.
- Clean-up phase
  - “long tails” due to insufficient automation of failure recovery.

Data Analysis via Grid@hostlab

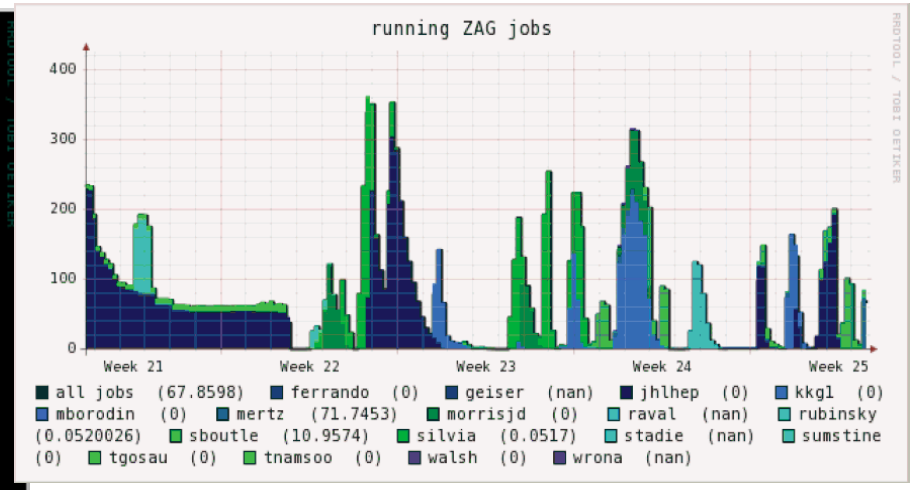
CDF@Fermigrid

(up to 1200 running jobs per user)



ZEUS@DESY

(up to 300 running jobs per user)



Common feature:
Very spiky loads!
A single user gets a lot done in little time!

Conclusions

- Several Running Experiments have fully embraced the grid.
  - MC production on EGEE & OSG.
  - Data Reprocessing on EGEE & OSG.
  - Large scale Data transfers using grid tools.
  - Caching of calibrations, software installations, and application tarballs via squid.
  - Data analysis via grid interfaces at hostlab.
- Integrate grid ops seemlessly into traditional ops.
  - Hide complexity behind “portals” and within “wrappers”.
  - Hide failure rates behind automated retries.