DØ Reprocesssing with SAM-Grid

Joel Snow on behalf of the DØ reprocessing team
D. Wicke, M. Diesburg, G. Garzoglio, C. Ay, A. Barnovski, I. Bertram,
Y. Coadou, G. Davies, L. Duflot, D. Evans, D. Gillberg, E. Gregores,
M. Hildreth, V. Hynek, R. Illingworth, T. Kurca, D. Lamb, P. Lebrun, S. Lietti,
Z. Liu, P. Love, P. McGuigan, P. Mercadante, J. Meyer, P. Mhashilkar,
T. Nunnemann, D. O'Neil, S. Salih, J. Steele, T. Stewart, F. Villeneuve-Seguier,
J. Yu

Outline

- Task and Implementation
- Certification
- Status
- Summary

Data Reprocessing

Improved detector understanding and new algorithms require rereconstruction

Computing Task

	2005 (p17)	2003/4 (p14)
Luminosity	$470\,\mathrm{pb}^{-1}$	$100\,\mathrm{pb}^{-1}$
Events	1G	300M
Rawdata 250kB/Event	250TB	75TB
DSTs 150kB/Event	150TB	45TB
TMBs 70(20)kB/Event	70TB	6TB
Time $50s$/Event	20,000months	6000months
(on 1GHz Pentium III)	3400CPUs for 6mths	2000CPUs for 3mths
Remote processing	100%	30%

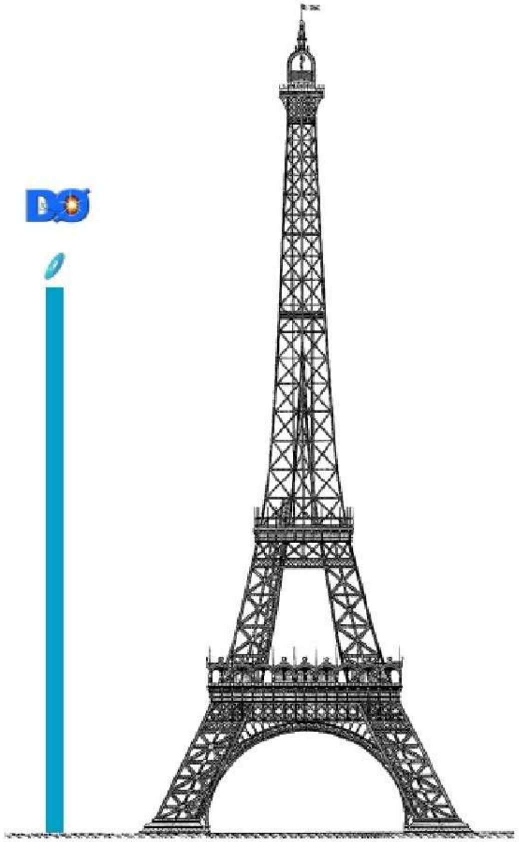
Central Farm (1000CPUs) used to capacity with data taking.

Data Reprocessing

Improved detector understanding and new algorithms require rereconstruction

Computing Task

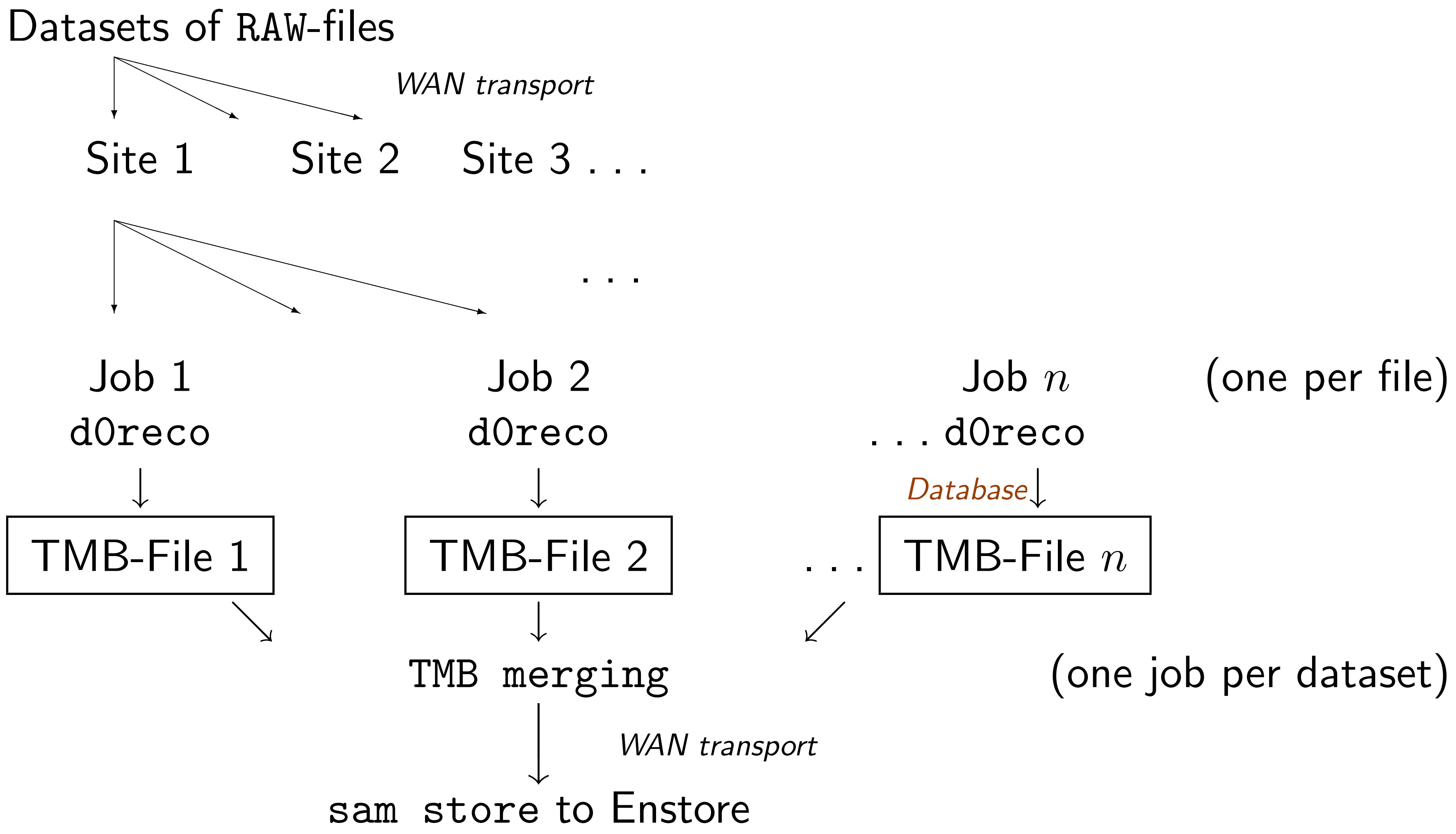
	2005 (p17)
Luminosity	$470\,\text{pb}^{-1}$
Events	1G
Rawdata 250kB/Event	250TB
DSTs 150kB/Event	150TB
TMBs 70(20)kB/Event	70TB
Time $50s$/Event	20,000months
(on 1GHz Pentium III)	3400CPUs for 6mths
Remote processing	100%



A stack of CDs as high as the Eiffeltower

Application flow

Overview



Datasets of RAW-files

WAN transport

Site 1 Site 2 Site 3 . . .

. . .

Job 1 Job 2 Job n (one per file)

d0reco d0reco . . . d0reco

Database

TMB-File 1 TMB-File 2 . . . TMB-File n

TMB merging (one job per dataset)

WAN transport

sam store to Enstore

Implementation

SAMGrid was chosen to implement this task on distributed systems.

- provides common environment for `d0reco` at all sites.
- allows common operation scripts (`d0repro`).

Production Step

- Each dataset processed through `d0reco` in one grid job.
- The grid jobs spawns one batch job per input file
- Resulting intermediate files are stored to SAM durable location (disk)

Scalability was improved by a factor of 500 to 1000(!)

Merging Step

- Merge TMBs after all RAW-files of a run, $\mathcal{O}(100)$, are successfully processed.
- But there are crashed and failures.

⇒ Merge only those that succeed; recover independently.
Book-keeping is essential to avoid merging one TMB into two merged-TMBs.

At any stage SAM will know what happened to a file

Sequential data Access via Meta-data

SAM is a data handling system organized as a set of servers working together to store and retrieve files and associated metadata, including a complete record of the processing which has used the files. SAM is designed for the following tasks:

- Track locations and comprehensive metadata for each file in the system.
- Provide storage utilities to add a file to a permanent storage location.
- Cache files on local disk for the duration of the requesting job or longer.
- Deliver files on request to systems that are SAM enabled.
- Utilize file location and system information for performance optimizations.
- Track processing information down to the level of per-file delivery and consumption status.

Book-keeping in SAM

SAM knows

a) from which RAW-file(s) a given TMB was created
b) with which version of which program it was created
c) which RAW-files were consumed by a given (set of) project(s).

⇒ SAM know about successes.

- By checking a) duplication of data in merging can be avoided.
- By asking "all RAW minus those for which TMB exists" those that failed can be found [uses a) and b)].
- By checking c) those that failed can be found, also.

SAM is sufficient to avoid data duplication and to create recovery jobs

Jobs and Information Management

Job submission and distribution was handled via JIM.

- JIM guarantees a uniform global interface to the system.
- Software releases are distributed via SAM (no pre-installation).
- All site peculiarities are parametrised in JIM.
- Provides a common software environment for `d0reco`.
- All sites run the same scripts.
- Provides site independent methods of job submission.

JIM was already successfully applied for MC production.

Book-keeping in JIM

JIM provides a local XML database at each site

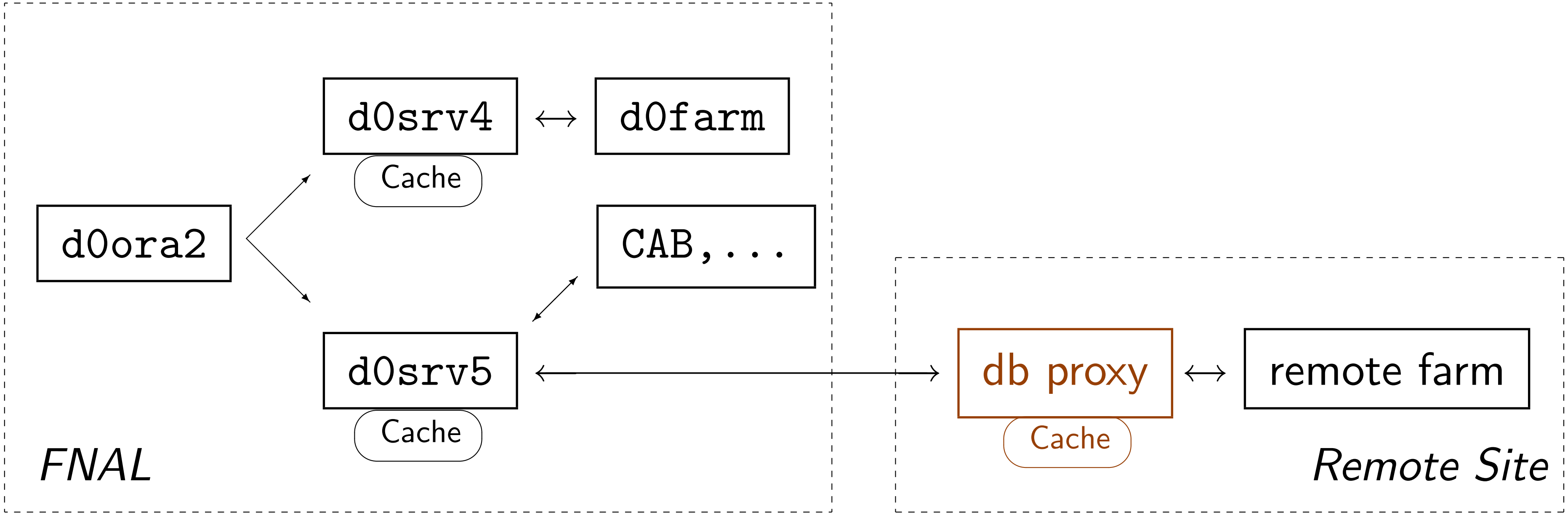
- Contains information about:
  - the definition and status of a grid job
  - which batch jobs created from a grid job
  - status of each batch job
  - which files were created by each batch job
  - detailed error conditions in case of failure
- This information quickly allows identification of errors

JIMs XML-DB was used to facilitate *error recovery*

Access to Calibration Database

Direct database access from Europe *much* to slow

Now: Database proxies

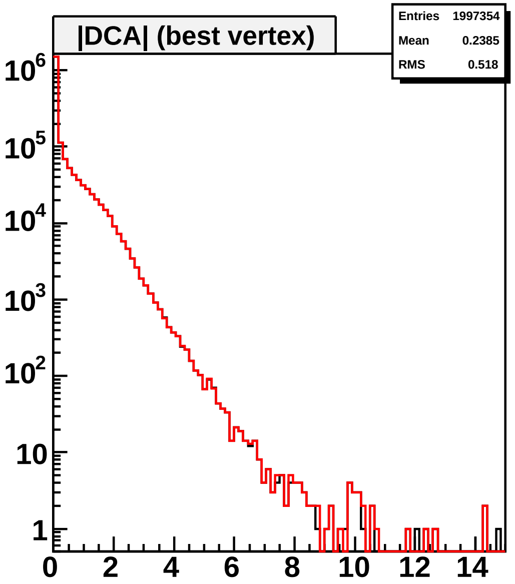


- Proxies were installed and tested at most sites.
- Proven to fix the problem.

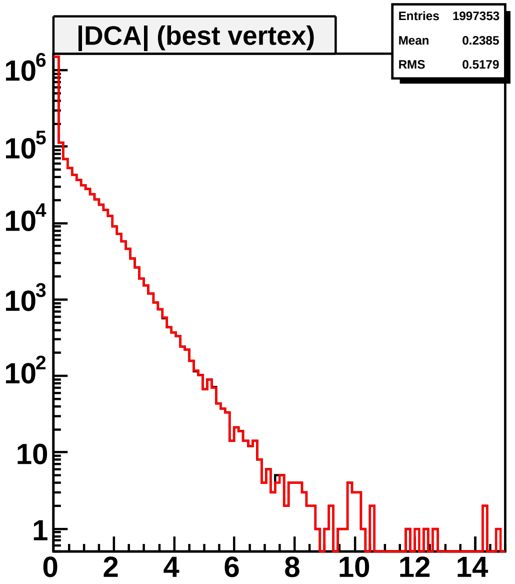
Certification of Sites and Code

- Compared SAMGrid production to conventional production on `d0farm`.
- Compared SAMGrid production at each site to `d0farm` production.
- Compared merged to unmerged TMBs at each site.

Lead to significant improvements in `recocert`



D0Farm JIM vs Lyon JIM



D0Farm Std vs JIM

Operation Scripts

The application flow started with a `dataset` to be processed.

Needed scripts

- which determine the full or partial success of processing.
- that submit the corresponding (partial) merging jobs.
- which determine the full or partial success of merging.
- that create and on request submit the recovery jobs for both steps in case of (partial) failure.

Was implemented using

- SAM for obtaining the information about files and
- JIM to submit jobs.

These scripts were common to all sites

Error Handling and Recovery

Beside unrecoverable crashes of `d0reco` there will be *random* crashes.

- Network outages
- File delivery failures
- Batch system crashes/hangups
- Worker-node crashes
- Filesystem corruption

To recover we need exact knowledge of what failed and what succeeded.

Book-keeping

1. of succeeded jobs/files
   needed to assure completion without duplicated events.

2. of failed jobs/files
   needed to trace problems in order fix bugs and to assure efficiency.

D0Repro (Basic commands)

- Support for certification
- Submission (and recovery) is done by

```
sub_production.py <dataset> <d0release>
sub_merge.py <dataset> <d0release>
```

- Determination of production and merge status (poor man's request system)

```
check_production.py <dataset> <d0release>
check_merge.py <dataset> <d0release>
```

- Manually modify status of jobs

```
set_status.py [production|merge] [approved|held|finished] <dataset> ...
```

Typical workflow:

1) `sub_production.py ...` (investigate/retry in case of failures)
2) `sub_merge.py ...` (after production is finished; retry if failed)
3) `set_status.py ... finished ...` (in case of unrecoverable failures)

D0Repro (Autopilot functionalities)

- Investigate status of all active requests `check_all.py`
- Clean completed/finished datasets `clean_completed.py`
- Display status of all active requests and suggests `auto_pilot.py`
  - recover production if less than 5% failed
  - submit merge if unmerged files exist and last job was production
  - optionally approved additional production jobs (one per automatic merge submission)
- Run commands suggested by autopilot `source Autopilot.sh`

This chain could be run in a loop (with 1 or 2 hours delay).

Autopilot was built on the experience of reprocessing.
Significantly reduced work-load of operations

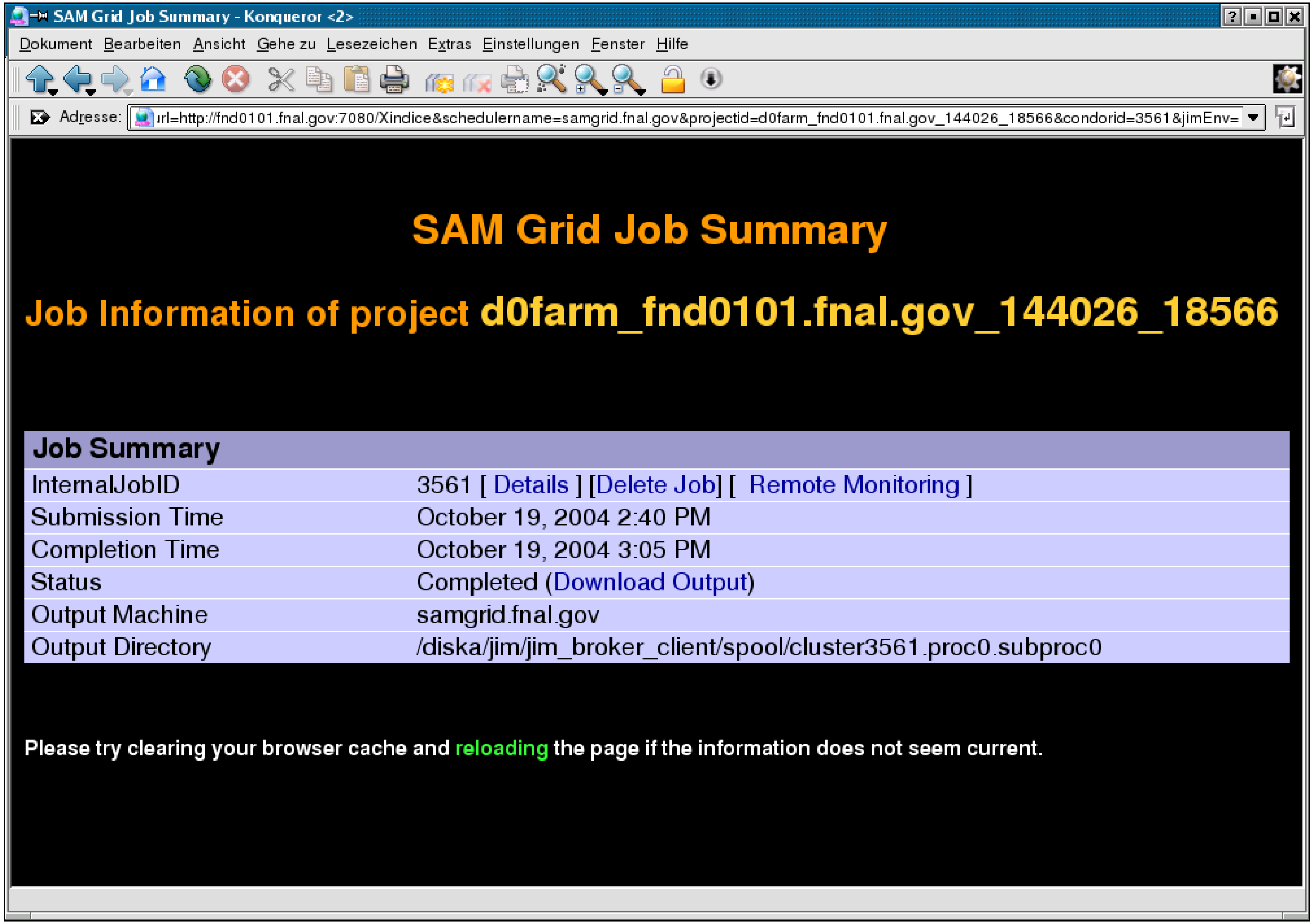
More that 90% of the operational work is to chase and fix failures.
Reliable book-keeping (taken from SAM) is prerequisite to implement these tools.

SAM-Grid Developments

Reprocessing stimulated the development of improved scalability and reliability for SAM-Grid. Developments included:

- Implementation of data queues
- Database access using proxies
- Implementation of application-aware grid services
  - that is, to configure different applications with different policies
  - for example, for the use of the storage or data queues

Some Screen-shots


SAM Grid Job Summary - Konqueror <2>
Dokument Bearbeiten Ansicht Gehe zu Lesezeichen Extras Einstellungen Fenster Hilfe
Adresse:

url=http://fnd0101.fnal.gov:7080/Xindice&schedulername=samgrid.fnal.gov&projectid=d0farm_fnd0101.fnal.gov_144026_18566&condorid=3561&jimEnv=

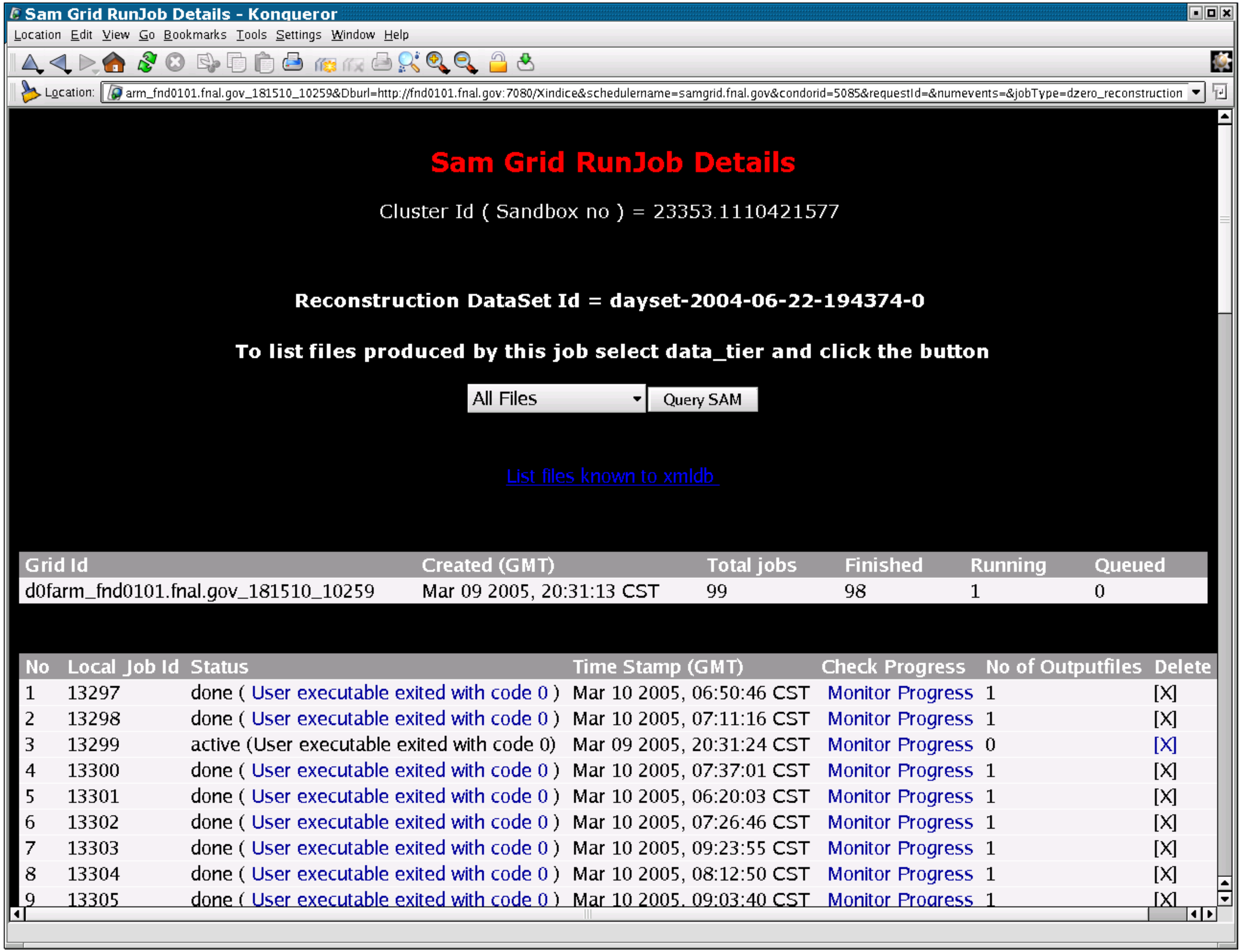
SAM Grid Job Summary

Job Information of project d0farm_fnd0101.fnal.gov_144026_18566

Job Summary	
InternalJobID	3561 [Details] [Delete Job] [Remote Monitoring]
Submission Time	October 19, 2004 2:40 PM
Completion Time	October 19, 2004 3:05 PM
Status	Completed (Download Output)
Output Machine	samgrid.fnal.gov
Output Directory	/diska/jim/jim_broker_client/spool/cluster3561.proc0.subproc0

Please try clearing your browser cache and reloading the page if the information does not seem current.

Some Screen-shots (2)


Sam Grid RunJob Details - Konqueror
Location Edit View Go Bookmarks Tools Settings Window Help
Location:

arm_fnd0101.fnal.gov_181510_10259&Dburl=http://fnd0101.fnal.gov:7080/Xindice&schedulername=samgrid.fnal.gov&condorid=5085&requestId=&numevents=&jobType=dzero_reconstruction

Sam Grid RunJob Details

Cluster Id (Sandbox no) = 23353.1110421577

Reconstruction DataSet Id = dayset-2004-06-22-194374-0

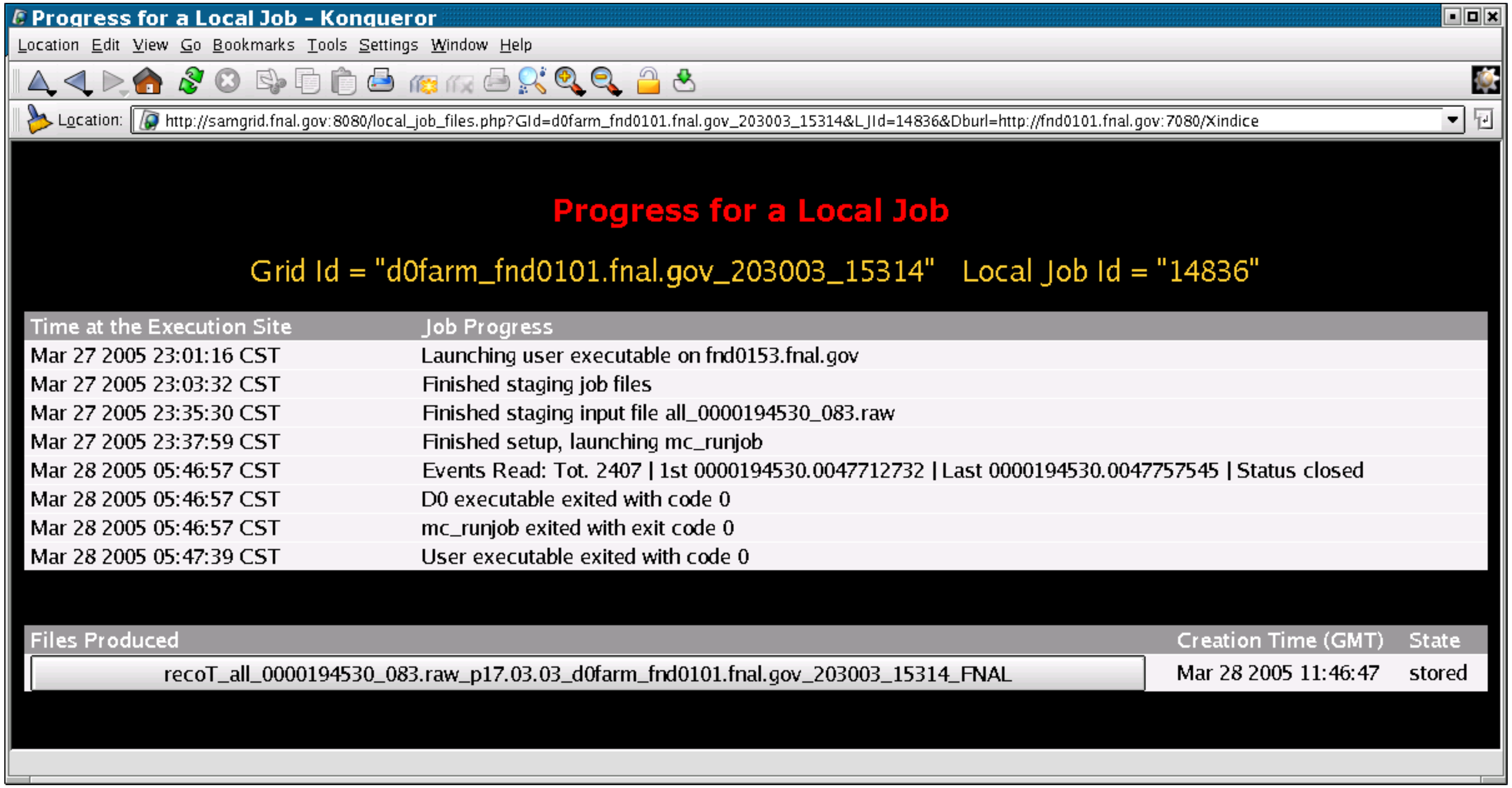
To list files produced by this job select data_tier and click the button

List files known to xmldb

Grid Id	Created (GMT)	Total jobs	Finished	Running	Queued
d0farm_fnd0101.fnal.gov_181510_10259	Mar 09 2005, 20:31:13 CST	99	98	1	0

No	Local Job Id	Status	Time Stamp (GMT)	Check Progress	No of Outputfiles	Delete
1	13297	done (User executable exited with code 0)	Mar 10 2005, 06:50:46 CST	Monitor Progress	1	[X]
2	13298	done (User executable exited with code 0)	Mar 10 2005, 07:11:16 CST	Monitor Progress	1	[X]
3	13299	active (User executable exited with code 0)	Mar 09 2005, 20:31:24 CST	Monitor Progress	0	[X]
4	13300	done (User executable exited with code 0)	Mar 10 2005, 07:37:01 CST	Monitor Progress	1	[X]
5	13301	done (User executable exited with code 0)	Mar 10 2005, 06:20:03 CST	Monitor Progress	1	[X]
6	13302	done (User executable exited with code 0)	Mar 10 2005, 07:26:46 CST	Monitor Progress	1	[X]
7	13303	done (User executable exited with code 0)	Mar 10 2005, 09:23:55 CST	Monitor Progress	1	[X]
8	13304	done (User executable exited with code 0)	Mar 10 2005, 08:12:50 CST	Monitor Progress	1	[X]
9	13305	done (User executable exited with code 0)	Mar 10 2005, 09:03:40 CST	Monitor Progress	1	[X]

Some Screen-shots (3)


Progress for a Local Job - Konqueror
Location Edit View Go Bookmarks Tools Settings Window Help
Location:

http://samgrid.fnal.gov:8080/local_job_files.php?GId=d0farm_fnd0101.fnal.gov_203003_15314&LJId=14836&Dburl=http://fnd0101.fnal.gov:7080/Xindice

Progress for a Local Job

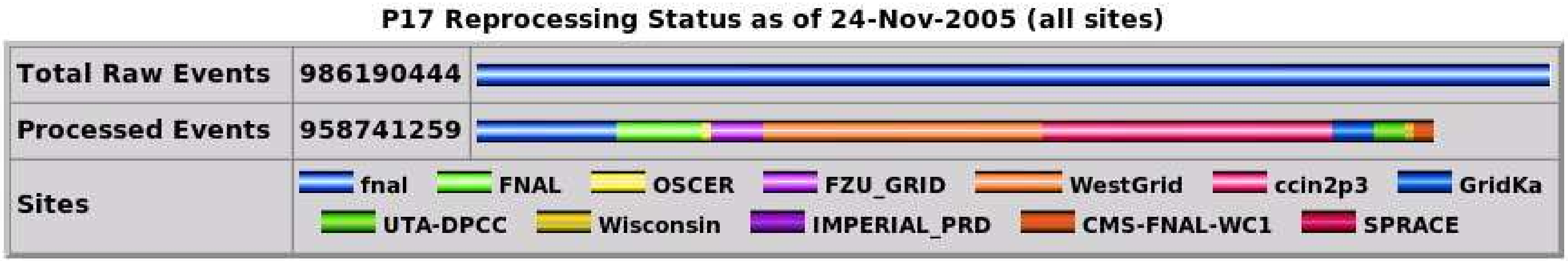
Grid Id = "d0farm_fnd0101.fnal.gov_203003_15314"   Local Job Id = "14836"

Time at the Execution Site	Job Progress
Mar 27 2005 23:01:16 CST	Launching user executable on fnd0153.fnal.gov
Mar 27 2005 23:03:32 CST	Finished staging job files
Mar 27 2005 23:35:30 CST	Finished staging input file all_0000194530_083.raw
Mar 27 2005 23:37:59 CST	Finished setup, launching mc_runjob
Mar 28 2005 05:46:57 CST	Events Read: Tot. 2407 | 1st 0000194530.0047712732 | Last 0000194530.0047757545 | Status closed
Mar 28 2005 05:46:57 CST	D0 executable exited with code 0
Mar 28 2005 05:46:57 CST	mc_runjob exited with exit code 0
Mar 28 2005 05:47:39 CST	User executable exited with code 0

Files Produced	Creation Time (GMT)	State
recoT_all_0000194530_083.raw_p17.03.03_d0farm_fnd0101.fnal.gov_203003_15314_FNAL	Mar 28 2005 11:46:47	stored

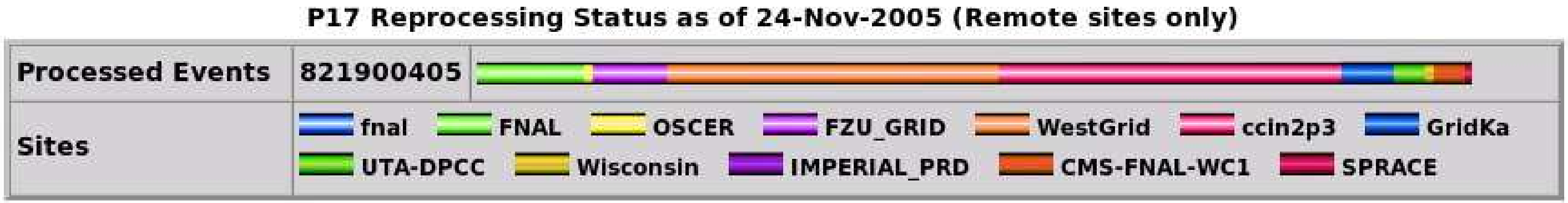
Status

Reprocessing effort started on 25-March-2005 in Lyon and Westgrid.



P17 Reprocessing Status as of 24-Nov-2005 (all sites)

Total Raw Events	986190444	
Processed Events	958741259	
Sites	🟦 fnal 🟩 FNAL 🟨 OSCER 🟪 FZU_GRID 🟧 WestGrid 🟥 ccin2p3 🟦 GridKa 🟩 UTA-DPCC 🟨 Wisconsin 🟪 IMPERIAL_PRD 🟧 CMS-FNAL-WC1 🟥 SPRACE	

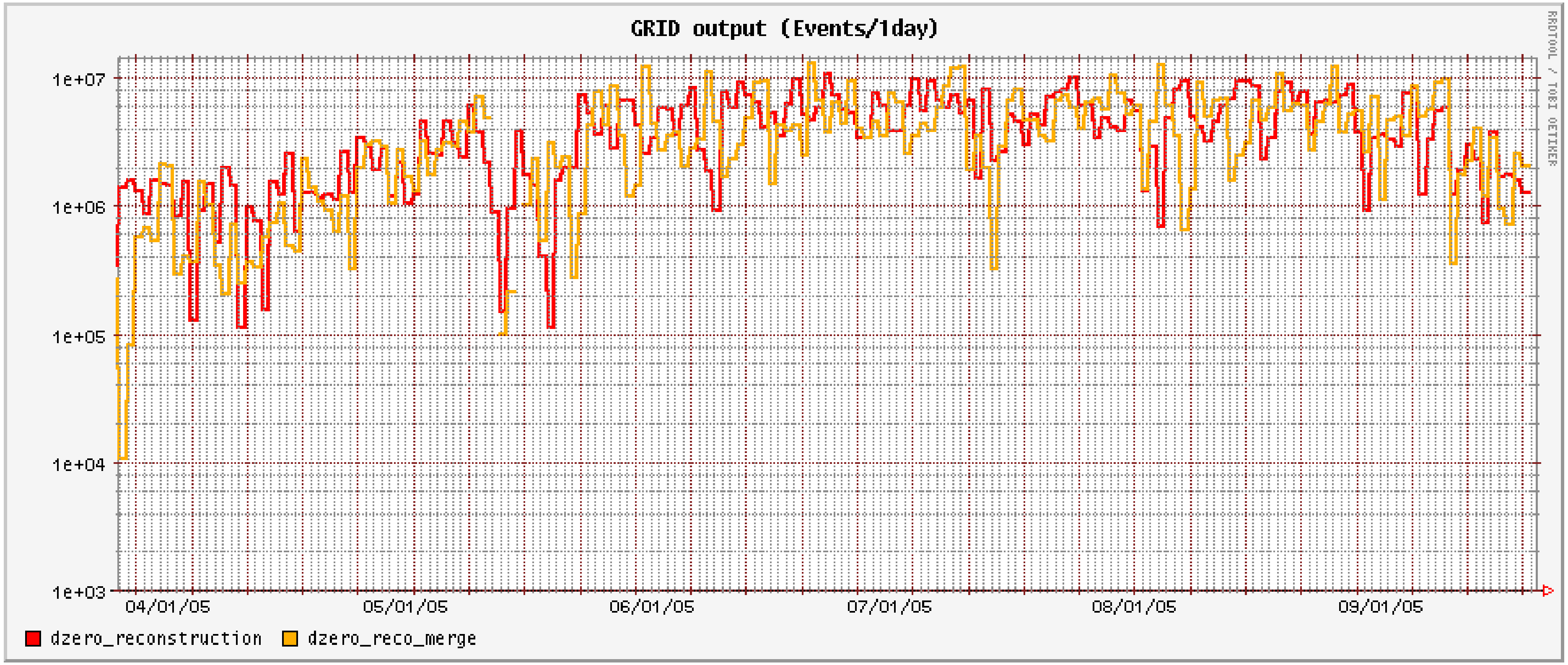


P17 Reprocessing Status as of 24-Nov-2005 (Remote sites only)

Processed Events	821900405	
Sites	🟦 fnal 🟩 FNAL 🟨 OSCER 🟪 FZU_GRID 🟧 WestGrid 🟥 ccin2p3 🟦 GridKa 🟩 UTA-DPCC 🟨 Wisconsin 🟪 IMPERIAL_PRD 🟧 CMS-FNAL-WC1 🟥 SPRACE	

As of 24th Nov. all remote sites finished reprocessing.
958.7M of 986.7M events are completed, i.e. 97.2% done.

Production Speed

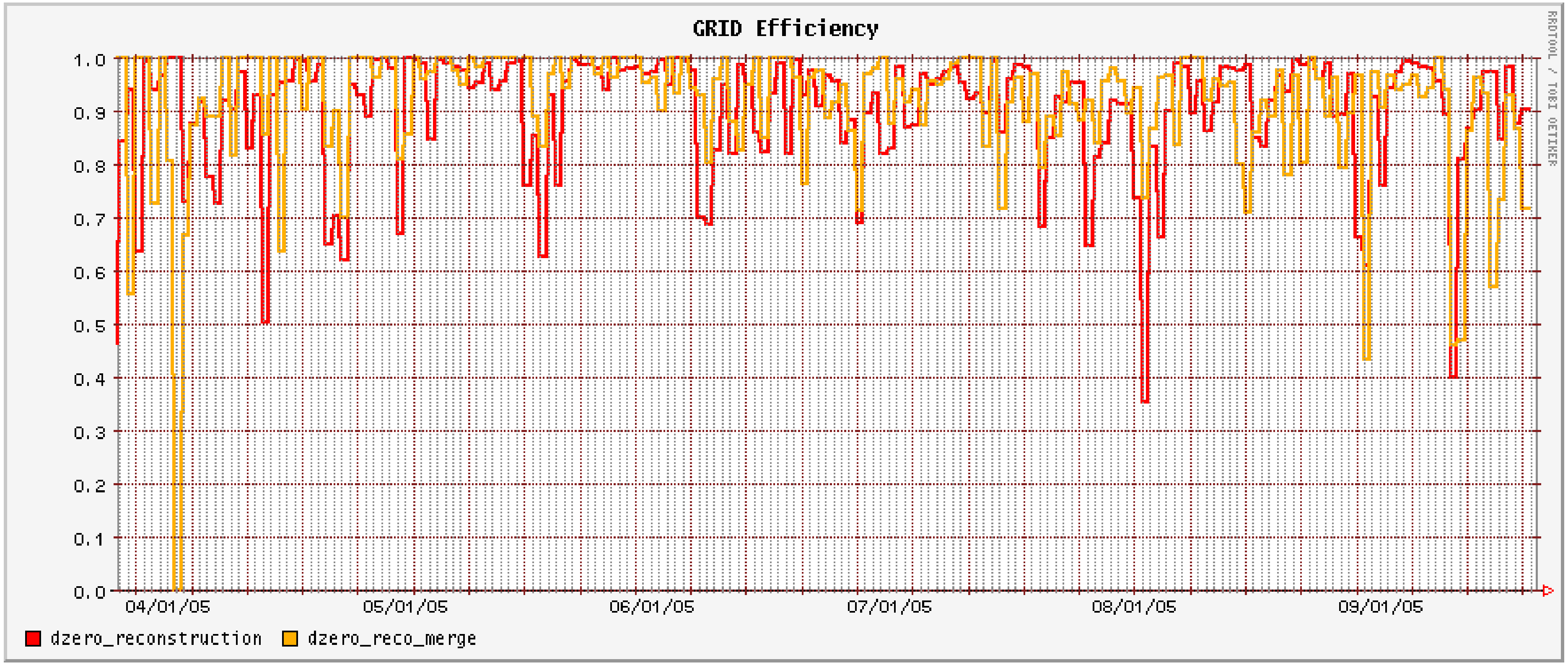


Production speed in Events/day* Needed 6M/day for 1G in 6 months.

Reducing after less than 6 months with sites having completed their assignments.

*Based on XML (by construction pessimistic)

Efficiency



Efficiency is number of batch jobs that produced a file over number of started jobs.

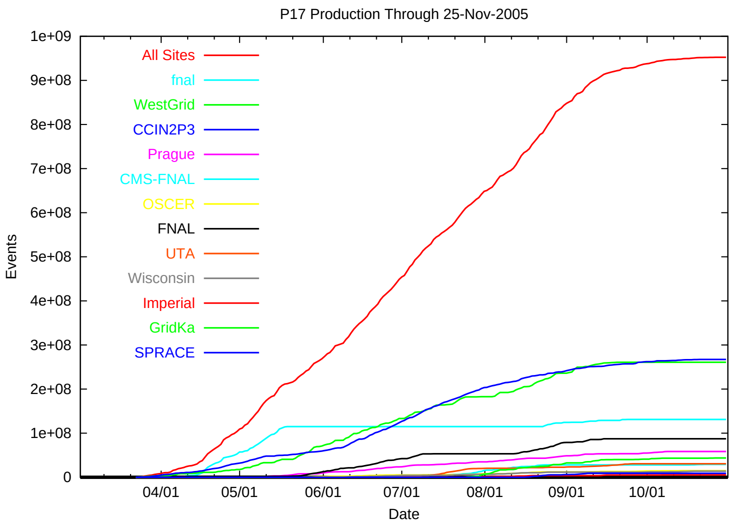
Average job failure rate 10%(!)*

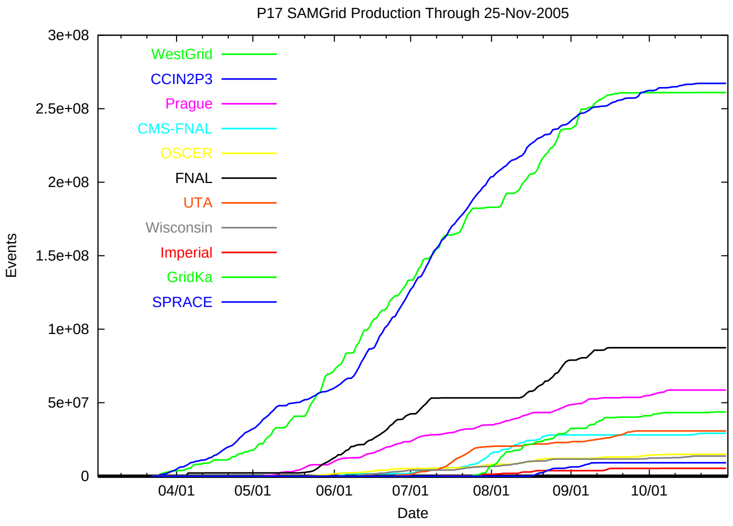
Dominated by failures of services: (Broken SAM, partial broken nodes, ...)

Rate of unrecoverable failures 3.0%**

*Based on XML (by construction pessimistic). ** based on SAM

Integrated number of events (from SAM)





- Deployment of improved infrastructure visible as kink ($\sim$ 25th Apr)
- Started at ~ 2.5MEvts/day.
- Reached up to ~ 10MEvts/day.
- Speed significantly reduced after mid Sep. i.e. after 5.5 months.
- Resources started working on MC

Bulk production completed on schedule.

Planned vs. actual contribution

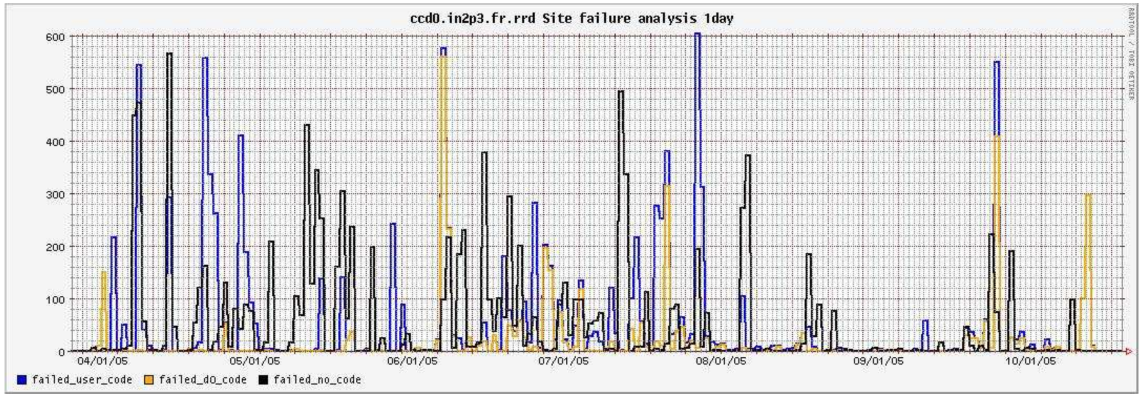
Site	Planned Contribution		Actual Contribution	
DØFarm (Fermilab),	0CPUs	Std:	13.75%	25% on-site
		SamGrid:	8.75%	
CMS Farm (Fermilab)	300CPUs		2.75%	
CCIN2P3 (Lyon)	400CPUs		27.0%	
Westgrid (Vancouver),	600CPUs		26.25%	
FZU (Prague)	200CPUs		5.75%	
GridKa (Karlsruhe)	500CPUs		4.25%	
UTA (Arlington)	230CPUs		3.0%	
Oscer (Oklahoma)	(140CPUs)		1.5%	
Wisconsin	30CPUs		1.25%	
Sprace (Sao Paolo)	(140CPUs)		0.75%	
UK-RAC (UK)	500CPUs		0.5%	70% off-site
External	∼3040CPUs	(1GHz PIII equiv.)		76% SamGrid

Discrepancy isn't a sign of bad work at the sites (in contrary)

This is a warning on how rough our estimates are.

Failure Analysis

Failure patterns very different at various sites

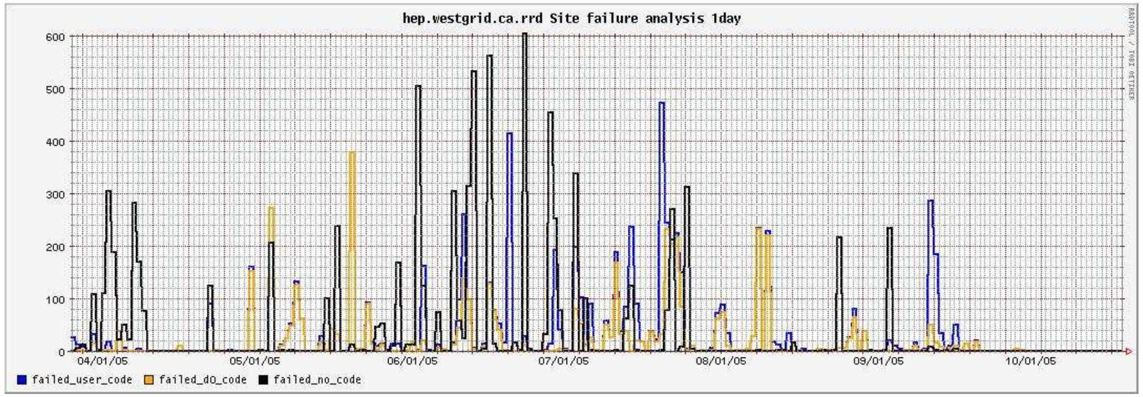


Lyon

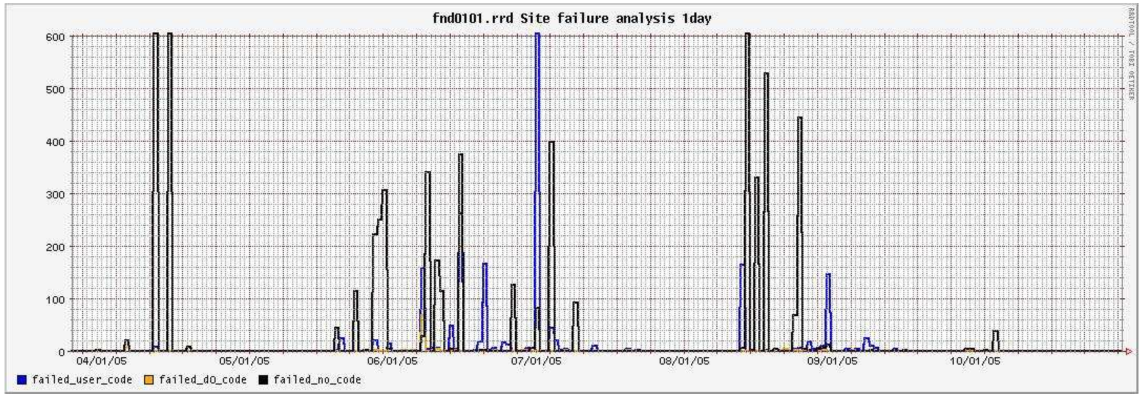
— failed d0reco

— failed mc_runjob

— no exit code at all



Westgrid



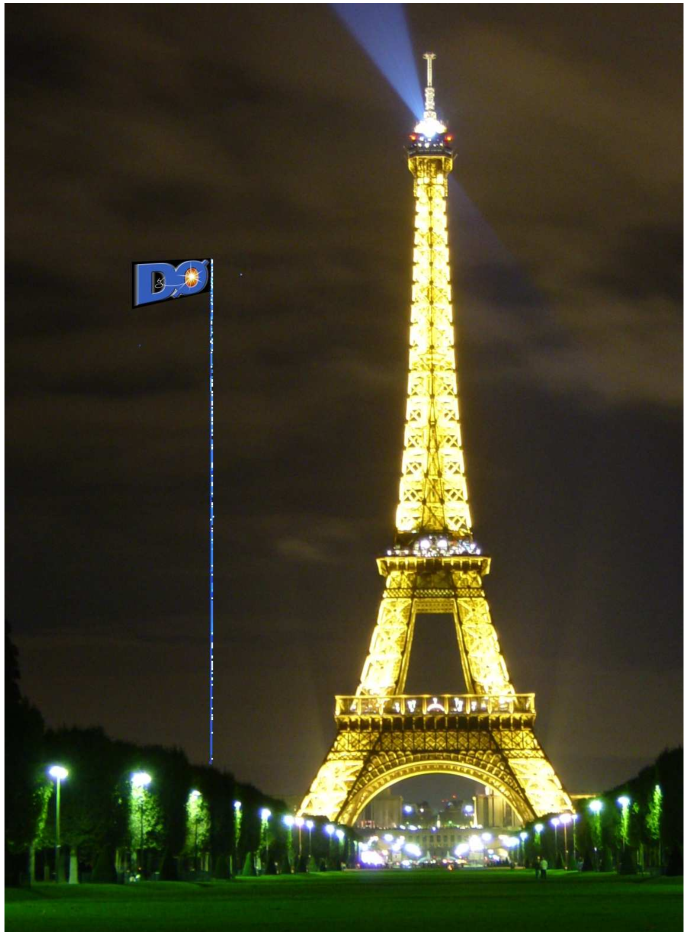
DØFarm

Summary

- p17 data reprocessing effort was 3 to $5\times$ bigger than the 2003/4 effort.
  - 250TB; 1600CPU years. Largest distributed HEP effort.
  - Fully gridified, common tools, 11 sites.
  - Bulk production done on schedule.
  - Recovery of 3% losses ongoing.

- Dataset available for physics doubled
  - $470\,\mathrm{pb}^{-1}$ of $1\,\mathrm{fb}^{-1}$ reprocessed
  - all data available w/ up-to-date reco.

- Grid is starting to return some investment
  - person power intense setup
  - common submission tools
  - sites installed for reprocessing can be used for MC
  - plan to switch initial processing to grid



Acknowledgements

This task required the assistance of many beyond those listed, both at Fermilab and at the remote sites, and we thank them for helping make this project the success that it was.