FAX and Panda WAN access: performance & reliability findings

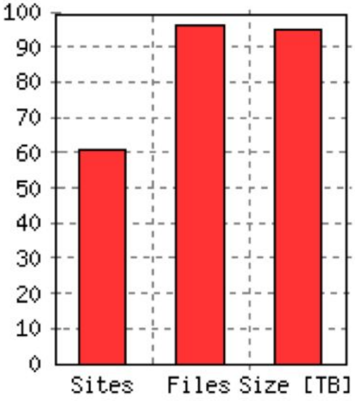
Ilija Vukotic • University of Chicago

Joint US ATLAS Software and Computing / Physics Support Meeting
LBNL, August 22, 2014

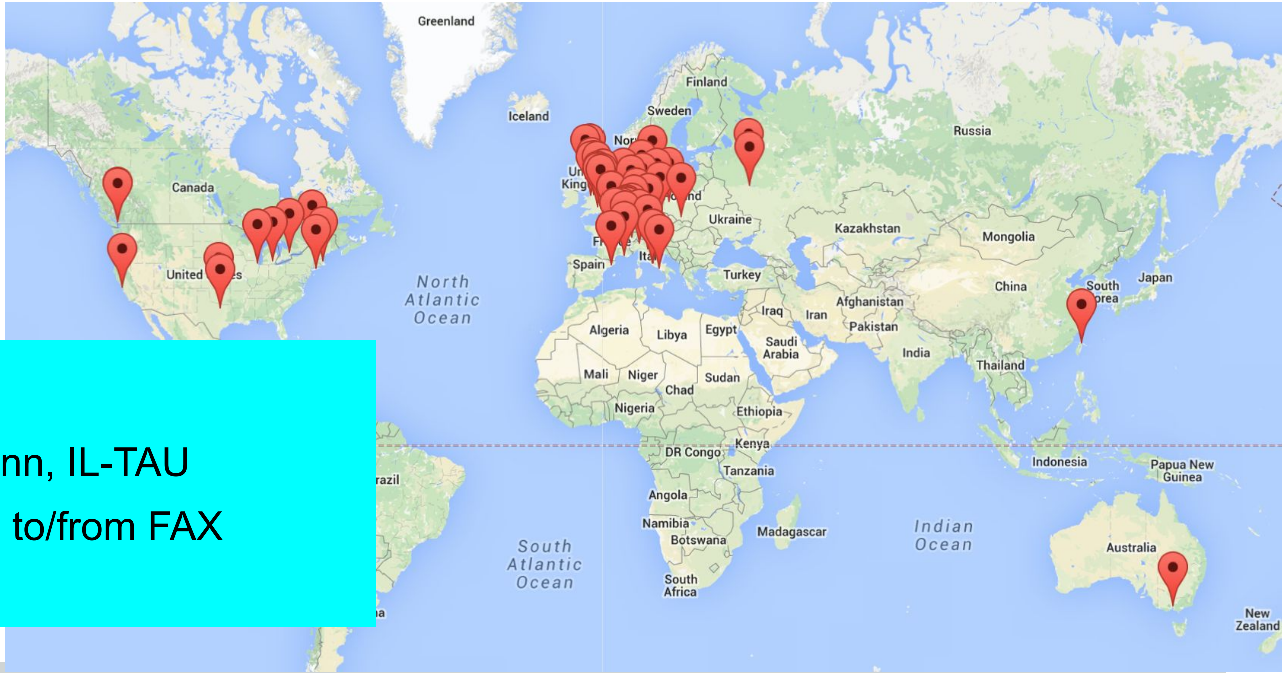
Introduction

- FAX status & coverage
- Review of usage modes by Panda
- Site controls
- Pilot changes
- Failover metrics
- Overflow metrics
  - Direct submission tests
  - Actual performance
- Summary and Conclusions

FAX deployment



Goal reached ! >96% files covered



Just several more sites to go:

Tokyo, Technion, Weizmann, IL-TAU

Some sites seriously limit rates to/from FAX

FAX status

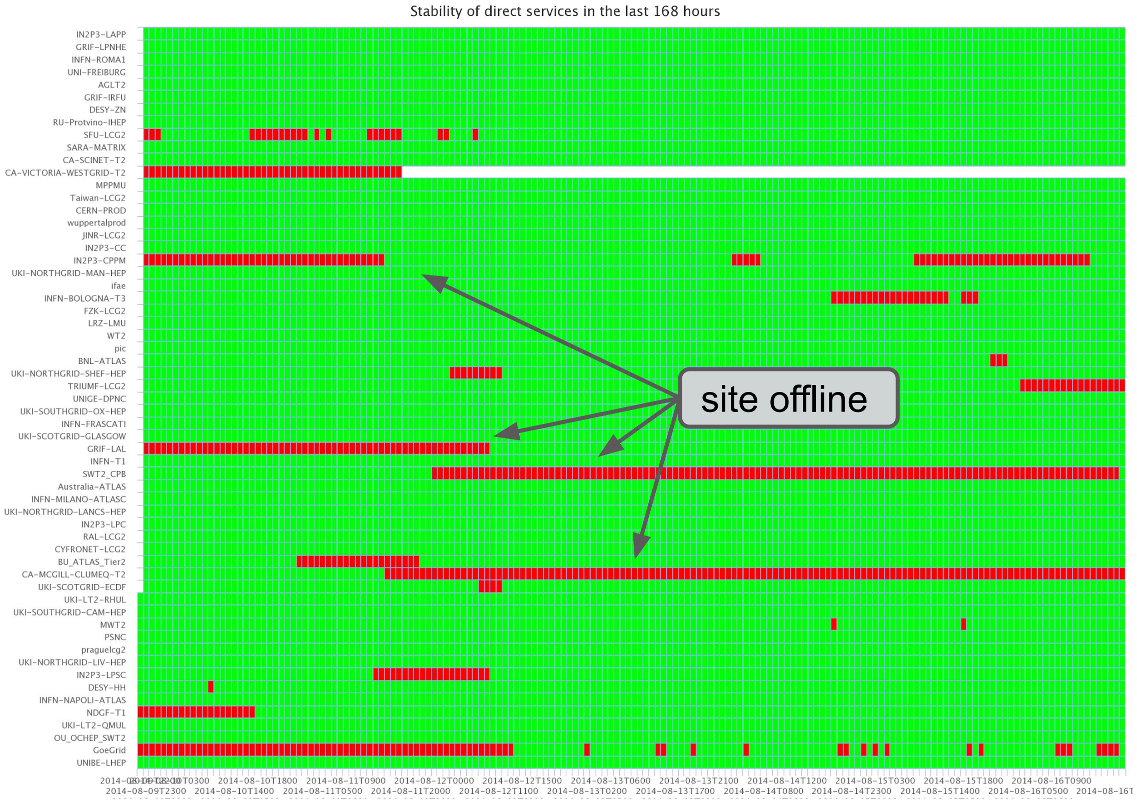
Quite stable running.

Issues usually simple and quickly resolved:

- full log disk
- not updated CRL
- unstable server

Soon we will start move to XRooD 4.0.3

- remote debugging
- higher performance
- sub-file level caching



Panda WAN modes

- Failover

  In the case stage-in fails due to a temporary SE related problem, the pilot will re-attempt the stage-in a second time after a few minutes. If that fails as well, the pilot has the option to attempt stage-in from a remote SE using FAX.

- Overflow

  When deciding where to broker a task, JEDI can estimate that it is better to send it to a site that does not have the input data and let it read from the FAX, that let it sit in the queue of the site that has the data.

- Explicit overflow

  If a user explicitly requires CE that does not have the input data, the task will be brokered to that CE and FAX used to get the data.

Site controls

Failover*

Setup per panda queue using two AGIS fields:

- **allowfax**=True will enable FAX retries.
- **faxredirector** sets the FAX access point to be used. For optimal performance it should be set to the site's closest redirector.

*turned on by default for all queues on March 2014

Overflow

In addition to allowfax and faxredirector queue* should set:

- **wansinklimit**** - limits the bandwidth that jobs overflown to the site can use.
- **wansourcelimit**** - limits the bandwidth that site's FAX endpoint can deliver to jobs overflown elsewhere.

* should be per site and not queue.

** zero value turns off overflow in that direction.

Overflow - technical details

Decision making

based on:

- “cost matrix” - lists expected rate for a single file transfer between an analysis queue and a FAX endpoint.
- a measure of how “busy” are the queues

Access point

While JEDI sets a variable “source site” for each job, actual FAX access point will be the faxredirector set for the destination queue.

Limit enforcement

two issues:

- lag between cost matrix measurement and job start
- can’t predict how fast jobs will start running

JEDI first sends 10 “scout” jobs. When all of them finish it calculates average per job bandwidth used.

Only sends more jobs if the limit is expected not to be breached.

This is not a very hard limit, so site’s limits are conservative.

Failover

- Running for more than 6 months.
- The last available numbers* show in average ~100 failover jobs/hour of which half finish correctly.
- There is a standing task to re-implement the failover monitor in BigPandaMon.

*Due to contamination of the failover log with the overflow messages, we don’t have current numbers.

Explicit overflow tests

- Used JEDI to submit tasks testing all the combinations* of remote data access in US.
- Each task has 252 ROOT jobs reading 3 x 4.5GB SMWZ D3PDs files.
- Interested in success rate, performance, comparison to local data access
- Important to know: In case an analysis queue doesn’t allow direct data access, pilot will xrdcp file from FAX to WN’s scratch area (BNL,BU).

*Some links not tested:

SWT2 had downtime.

ANALY_SLAC has setup issue.

ANALY_AGLT2 never got jobs delivered to it.

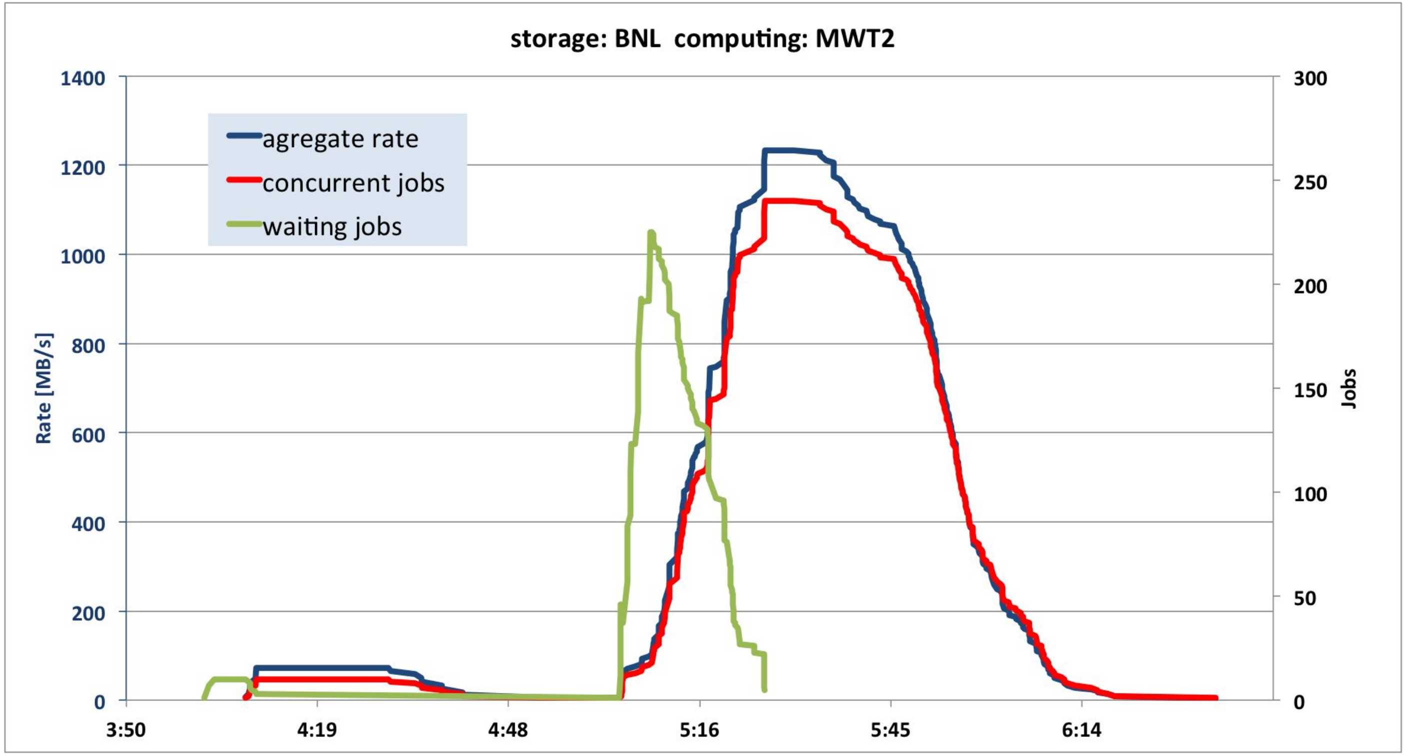
Explicit overflow tests

Failure rate at 1‰ level (4/5040).

	ANALY_BNL_SHORT	ANALY_BU_ATLAS_Tier2_SL6	ANALY_MWT2_SL6	ANALY_OU_OCHEP_SWT2
AGLT2	0	0	0	0
BNL-ATLAS	0	0	0	0
BU_ATLAS_Tier2	0	0	1	0
MWT2	3	0	0	0
OU_OCHEP_SWT2	0	0	0	0

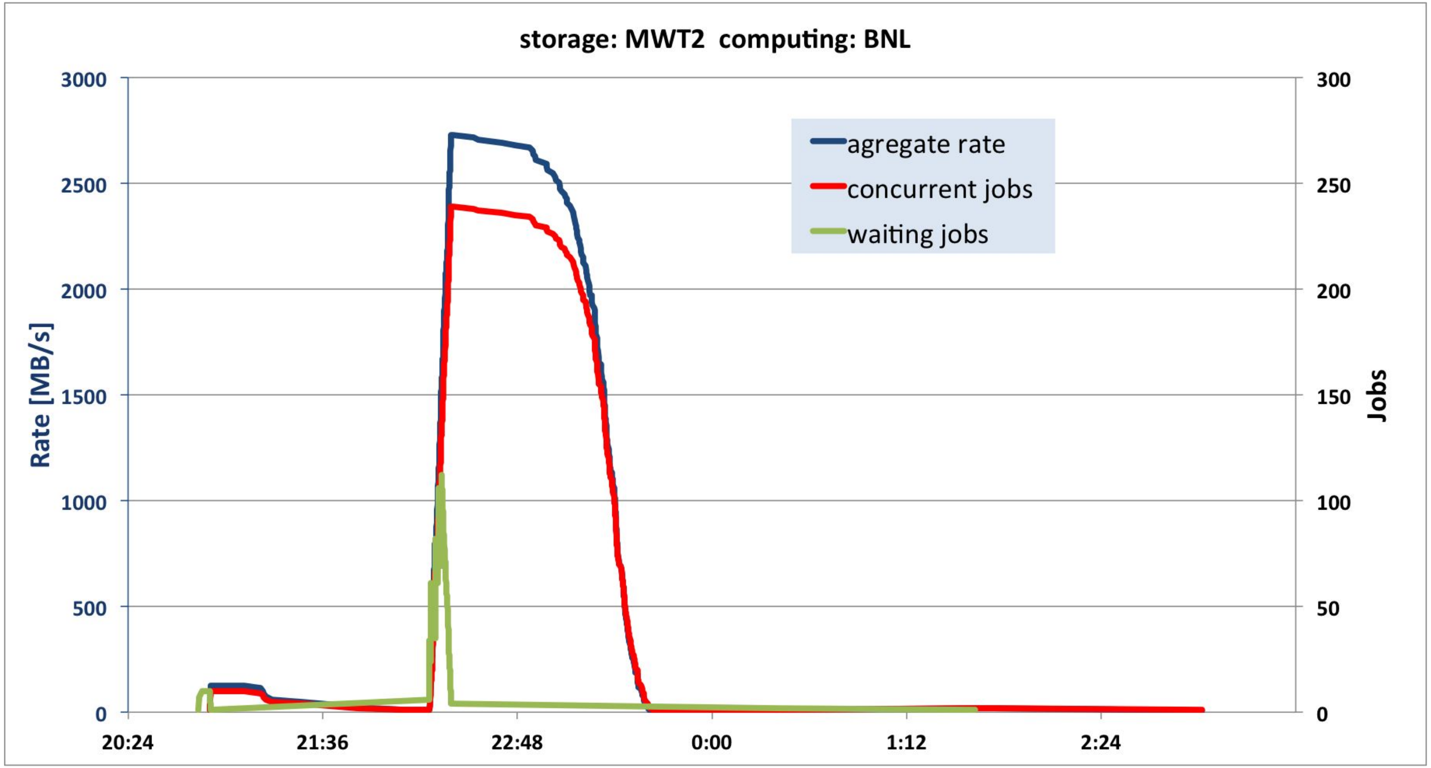
Explicit overflow tests

source/destination	BNL/MWT2
finished/failed	252/0
from start to end	2:32:21
max running jobs	240
average job duration	37:40.5
max waiting jobs	225
average waiting time	10:13.1
max rate	1233.68 MB/s
average rate	5.22 MB/s
average CPU time	1060.82
average WALL time	2123.29
average CPU eff.	0.5



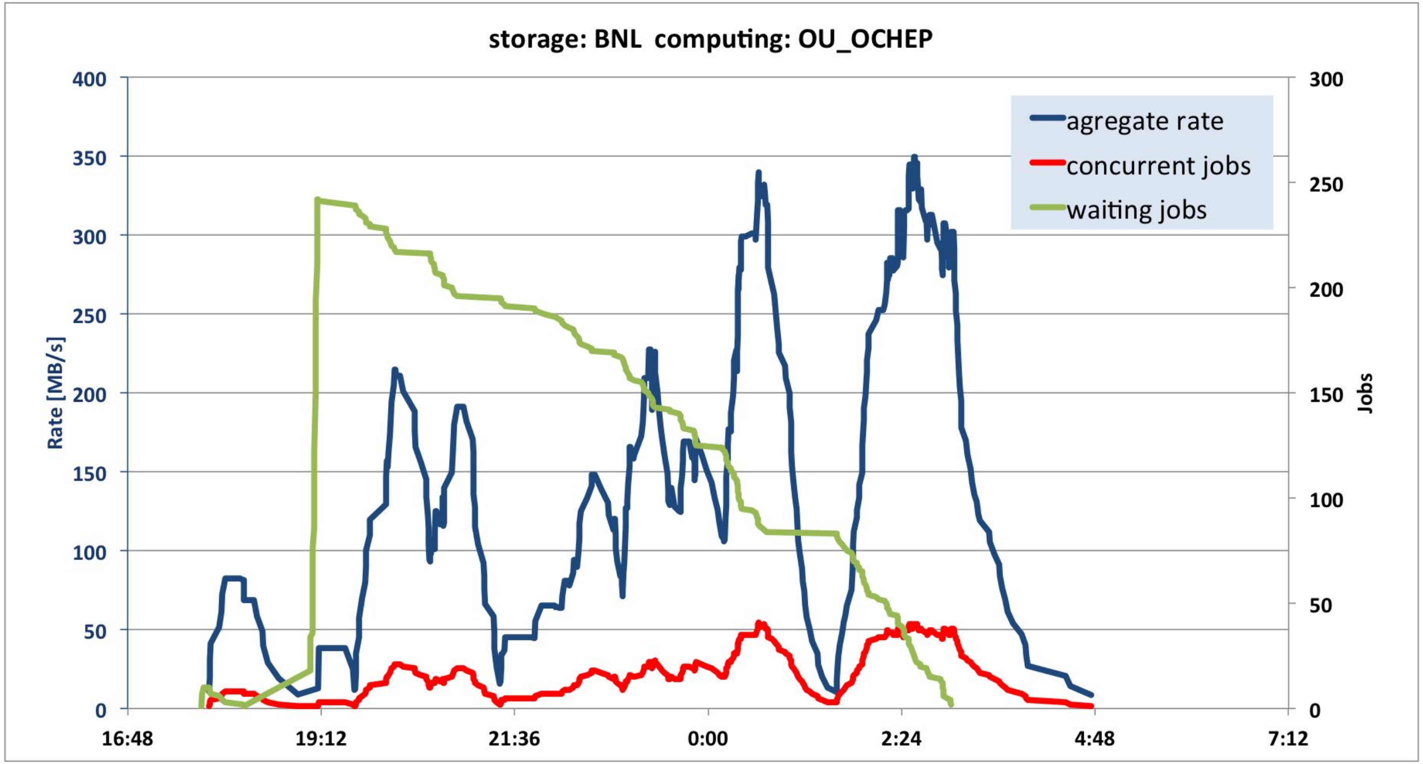
Explicit overflow tests

source/destination	MWT2/BNL
finished/failed	252/3
from start to end	6:11:25
max running jobs	239
average job duration	59:04.9
max waiting jobs	112
average waiting time	03:32.7
max rate	2728.98 MB/s
average rate	11.44 MB/s
average CPU time	936.62
average WALL time	963.4
average CPU eff.	0.97



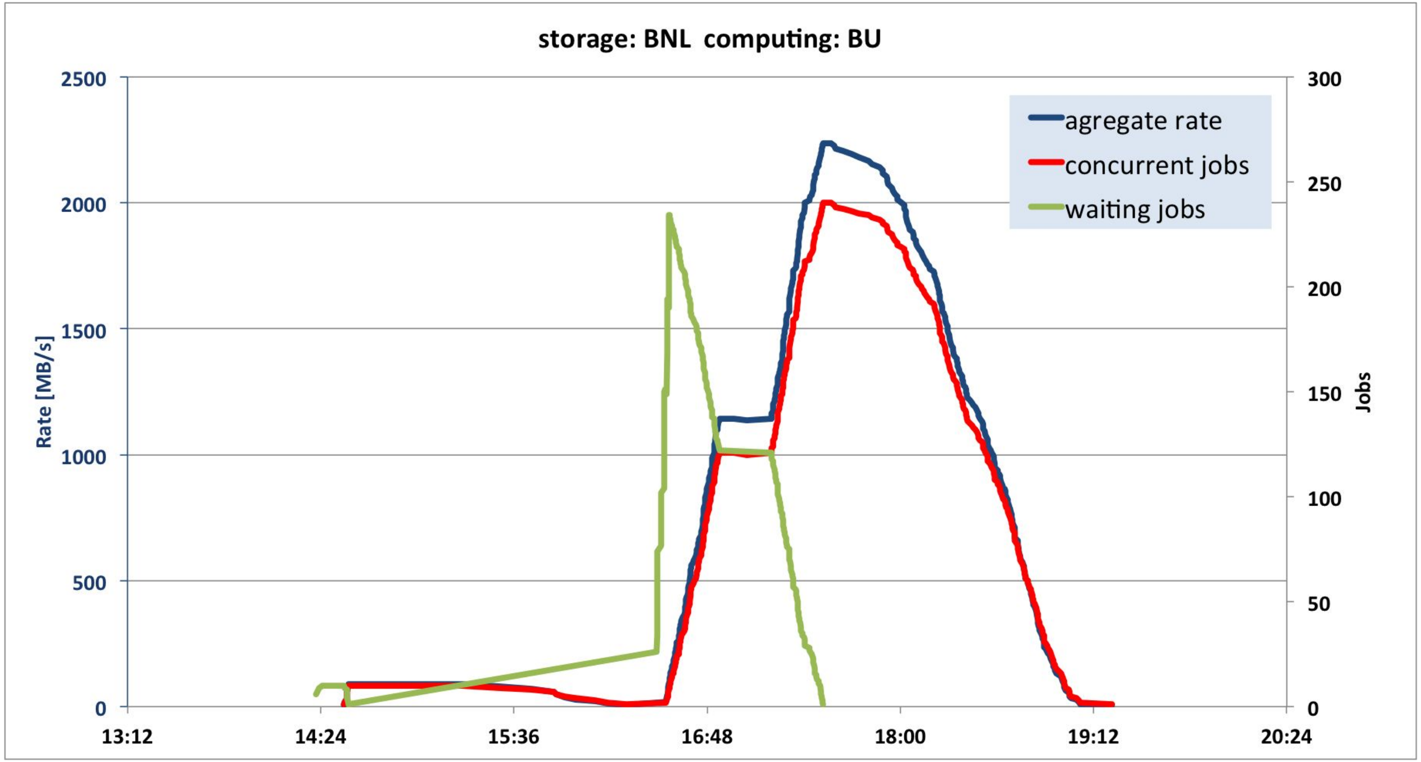
Explicit overflow tests

source/destination	BNL/OU
finished/failed	252/0
from start to end	11:02:11
max running jobs	41
average job duration	36:40.9
max waiting jobs	242
average waiting time	30:03.8
max rate	349.58 MB/s
average rate	9.03 MB/s
average CPU time	1115.98
average WALL time	1266.52
average CPU eff.	0.88



Explicit overflow tests

source/destination	BNL/BU
finished/failed	252/0
from start to end	4:56:34
max running jobs	240
average job duration	26:55.0
max waiting jobs	234
average waiting time	29:28.2
max rate	2235.7 MB/s
average rate	9.31 MB/s
average CPU time	1091.38
average WALL time	1218.26
average CPU eff.	0.9



Overflow startup

On August 12th, JEDI replaced PanDA.

Regular users jobs may be overflown to FAX.

Our goal: have 5-10% of all the jobs use WAN access, by the time of new data taking. Have at least 50% of the CPU efficiency of regular jobs.

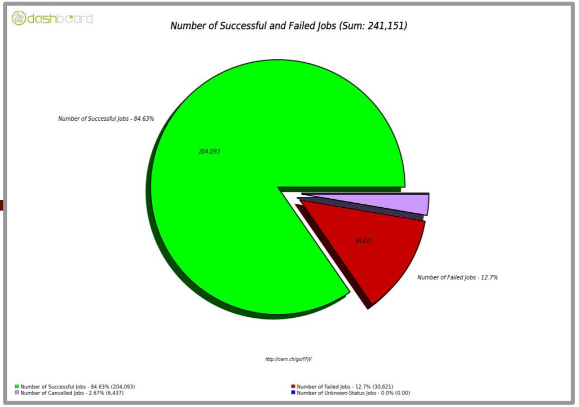
Decided to start slow:

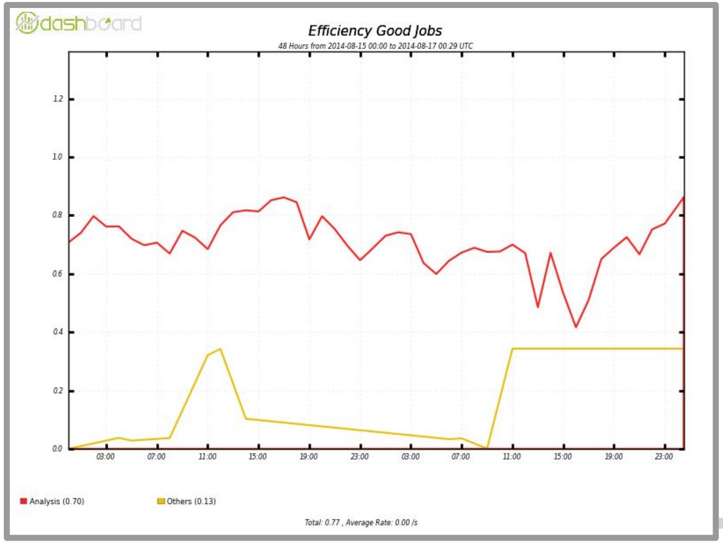
- Only US
- Only Analy queues
- Very high cut on expected per file transfer rate (25MB/s)

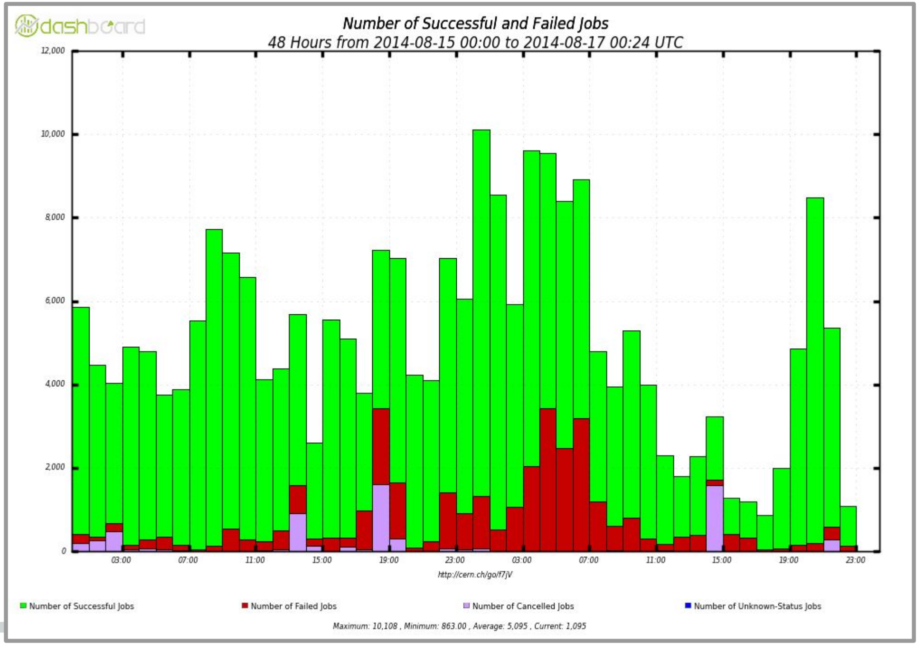
Monitoring through ADC http://dashb-atlas-job.cern.ch

Regular jobs

	regular	overflow goals
jobs per hour	5000	250
job efficiency	85%	>80%
cpu efficiency	70%	>35%







Pilot/JEDI changes

Found and still finding kinks in the system

Bug fixes:

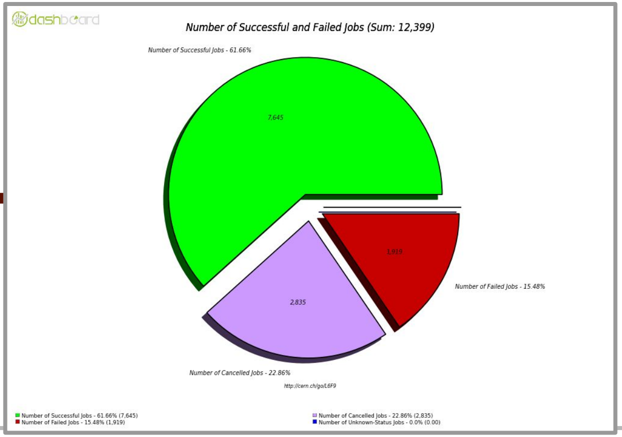
- failover log messages were sent for overflow jobs too. Spoiled monitoring.
- was taking the first replica returned, sometimes this was a TAPE replica where files are not in RUCIO format.
- TRIUMF-LCG2 has the FAX endpoint inaccessible to their WNs. This prevents payload file from being accessed. Will be solved by TRIUMF.
- obsolete cost matrix data copied from SSB to AGIS. Fixed.
- JEDI checks of the destination scratch size.

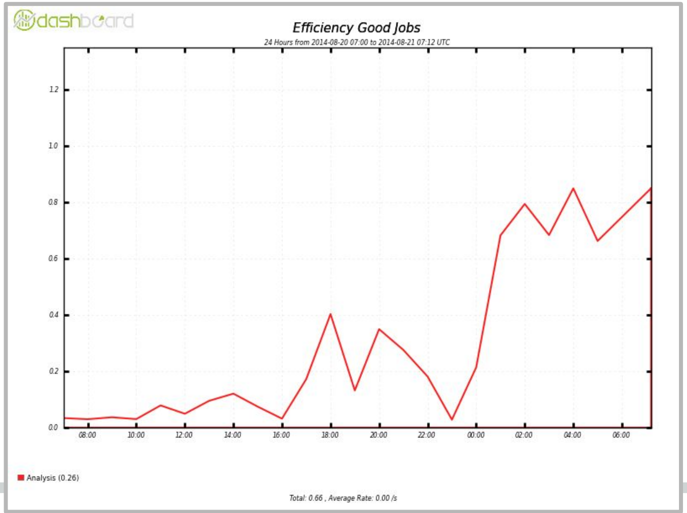
Tunings:

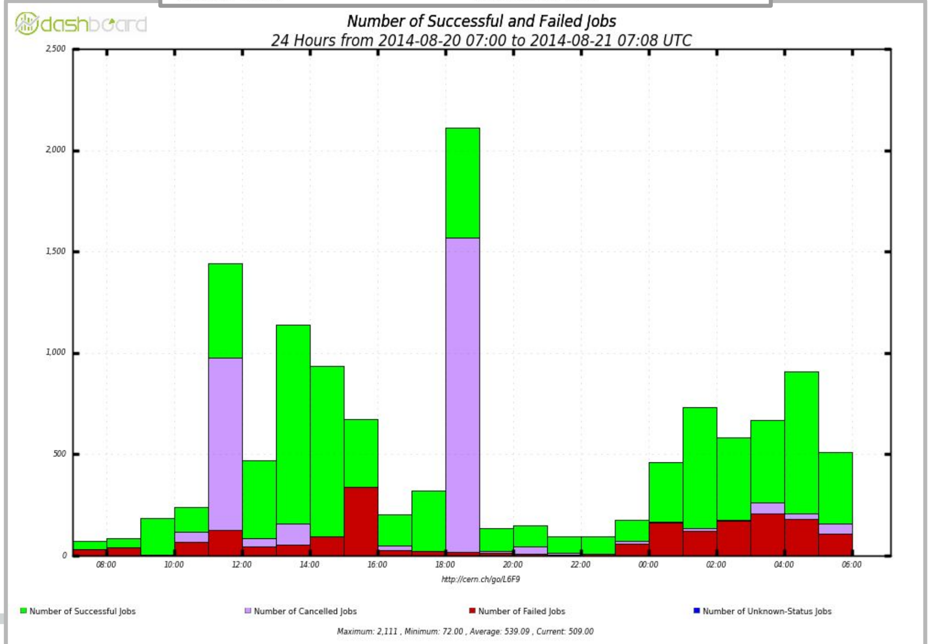
- new cut off on cost rate (was 50 MB/s now 25 MB/s)
- US requirement removed

Overflow jobs

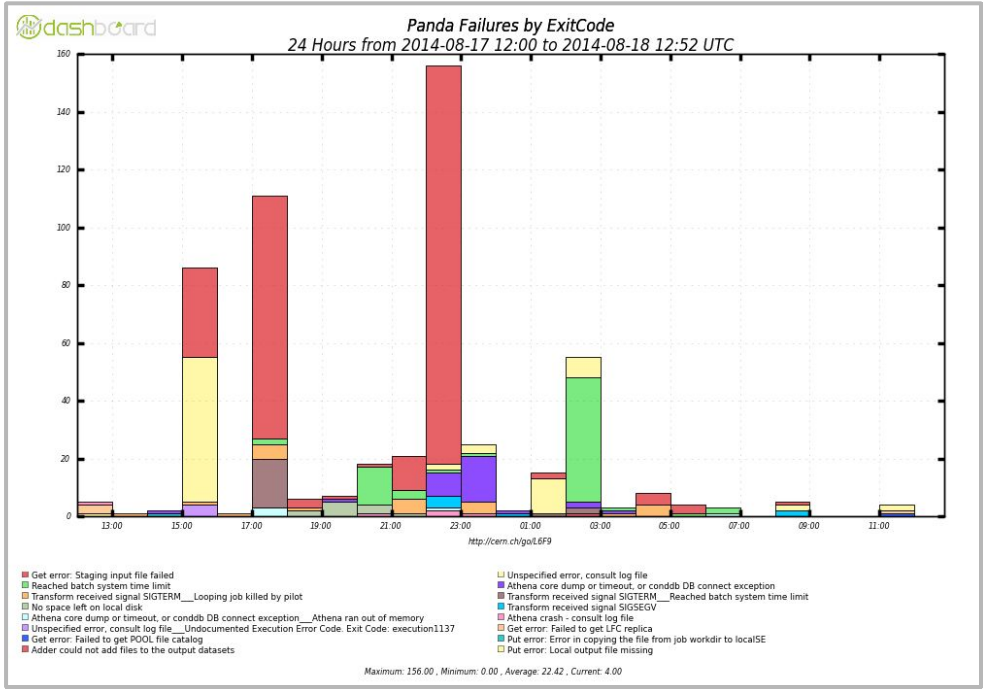
	overflow goals	measured
jobs per hour	250	500 (to one site)
job efficiency	>80%	83%
cpu efficiency	>35%	26%







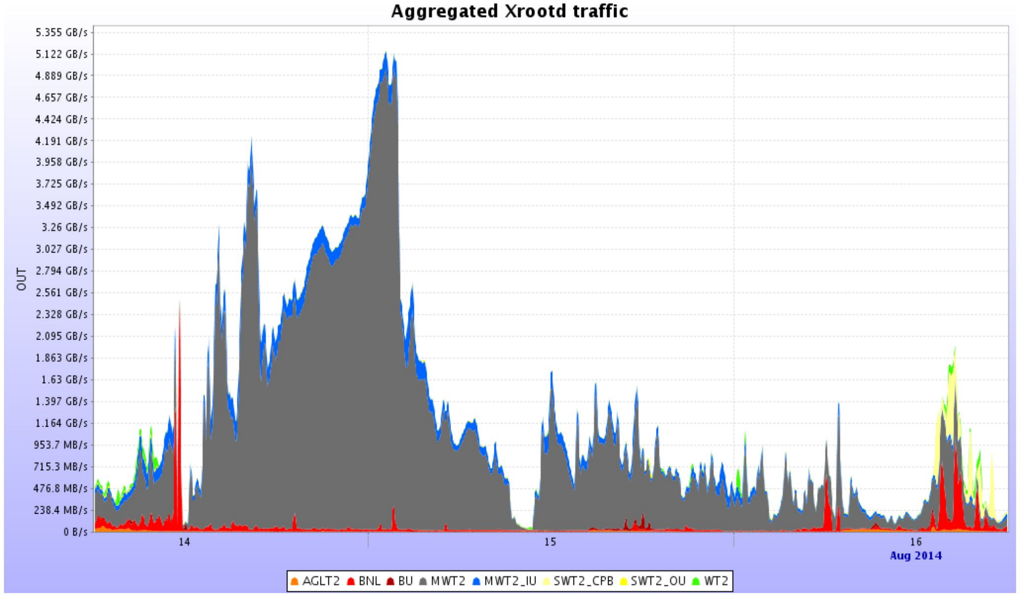
Overflow jobs

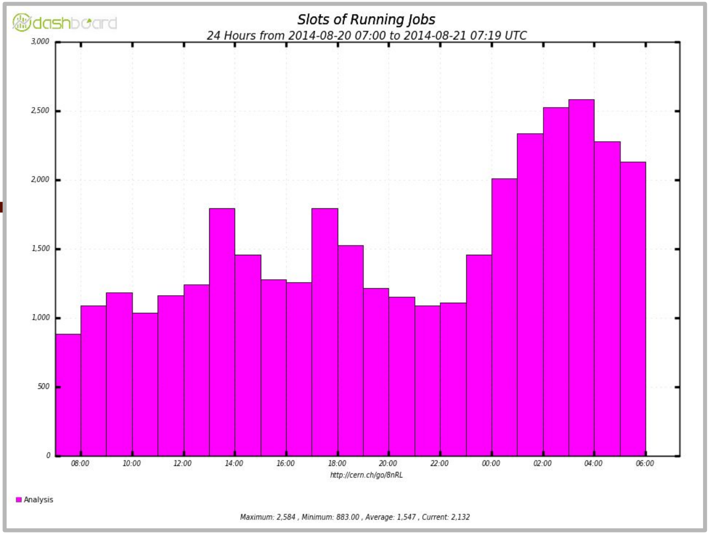


- Stage in errors mostly disappeared.
- Transform error rate does not seem higher than in regular jobs.
- more statistics needed.

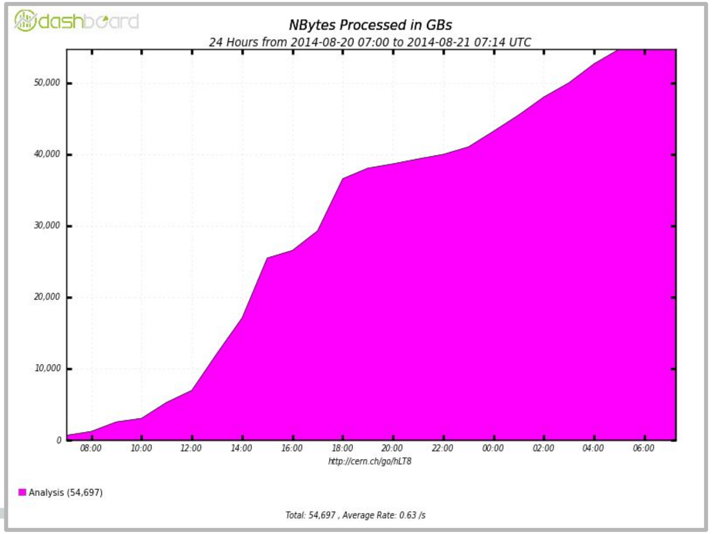
Overflow jobs

Not really stressing the system.


Aggregated Xrootd traffic


Slots of Running Jobs
24 Hours from 2014-08-20 07:00 to 2014-08-21 07:19 UTC

Maximum: 2,584 , Minimum: 883.00 , Average: 1,547 , Current: 2,132


NBytes Processed in GBs
24 Hours from 2014-08-20 07:00 to 2014-08-21 07:14 UTC

Total: 54,697 , Average Rate: 0.63 /s

Conclusions

Explicit overflow:

- Low failure rate.
- Satisfactory performance.
- Still some configuration fixes to be understood and fixed.

Overflow:

- There is a need for it - enough jobs overflow even with very stringent limit on per file rate.
- Needs debugging and tuning.
- At the scale of 500-1000 jobs per hour (US only) not a serious load on the infrastructure.
- Job’s CPU efficiency will benefit a lot from switch to xAODs.