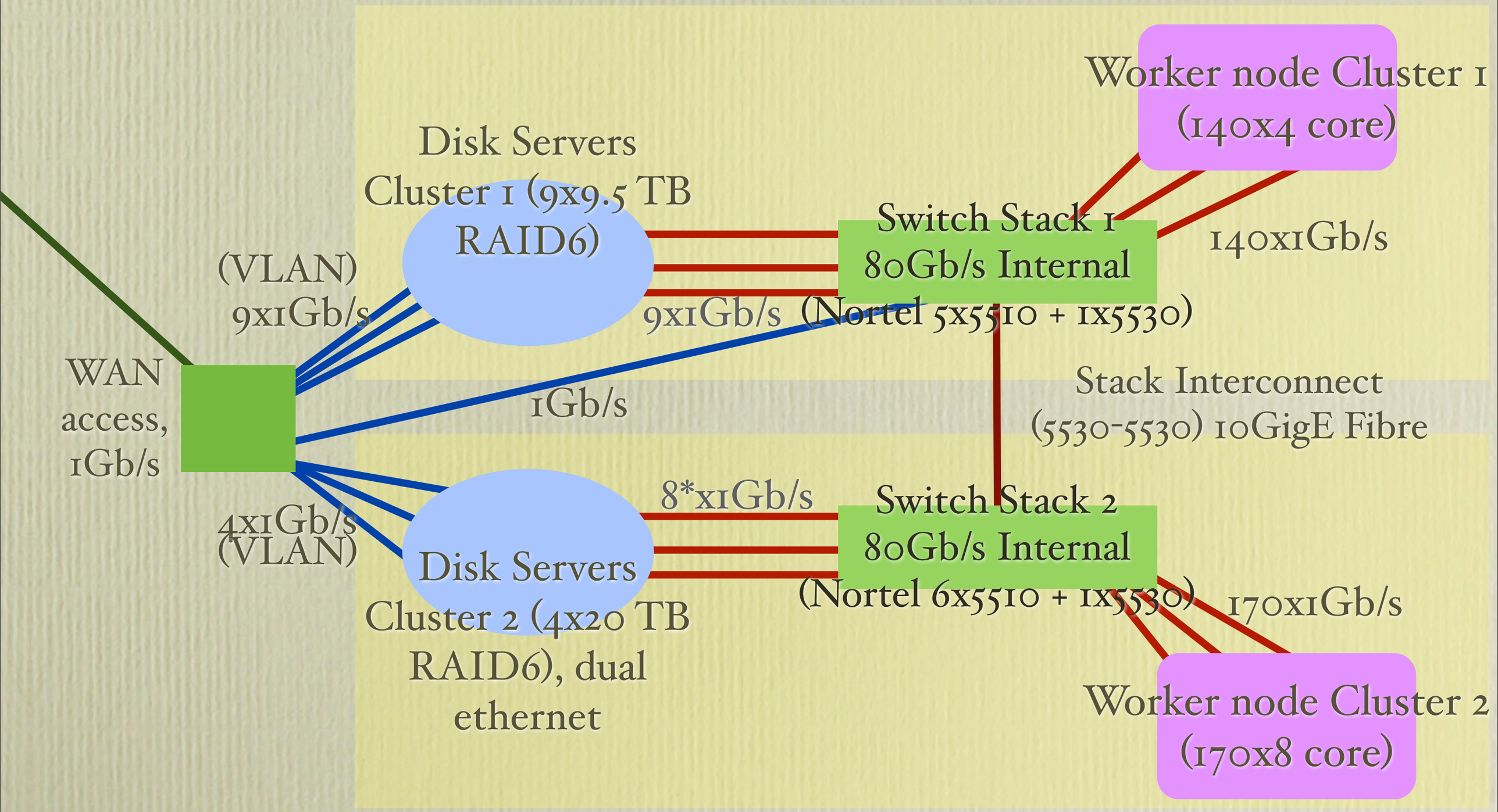
Glasgow Storage Report

(2009: A STEP Odyssey)

Itinerary

- Site network configuration
- STEP'09 for ATLAS @ Glasgow
  - Production, Data distribution (the Good)
  - Analysis
    - PanDA (the Tolerable)
    - WMS (the Ugly)

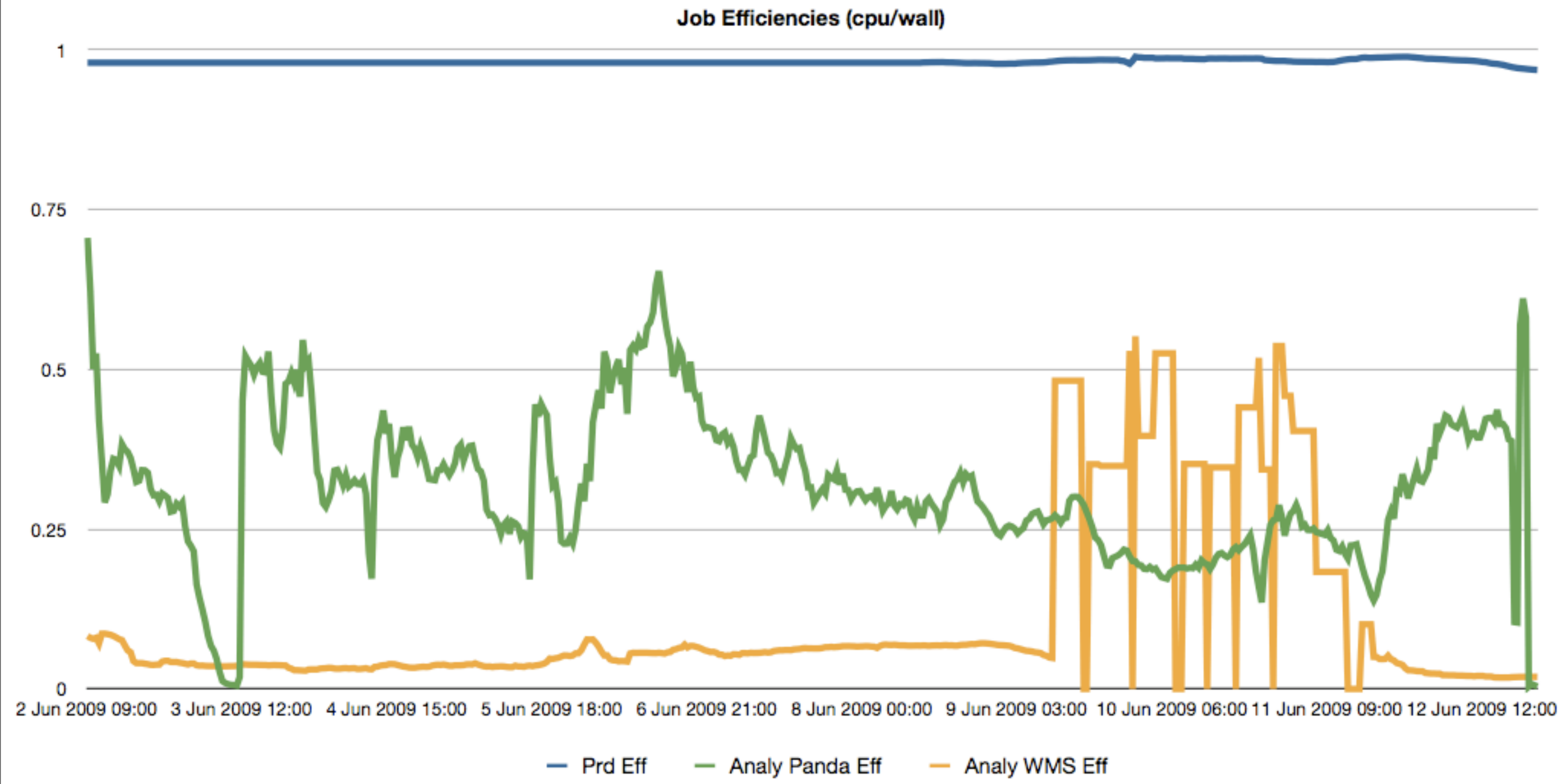
Site Network Infrastructure



Storage Config

- New pool nodes are:
  - 20 x 1Tb disks (RAID 6, Areca), xfs
- DPM head node is 8 core Intel Xeon, 16 Gb (a repurposed WN), DPM 1.6.11
- DPM MySQL instance is a 2 core Opteron (the original DPM head node)
- Daily MySQL dumps, backed up to local disk, and then to remote disk.

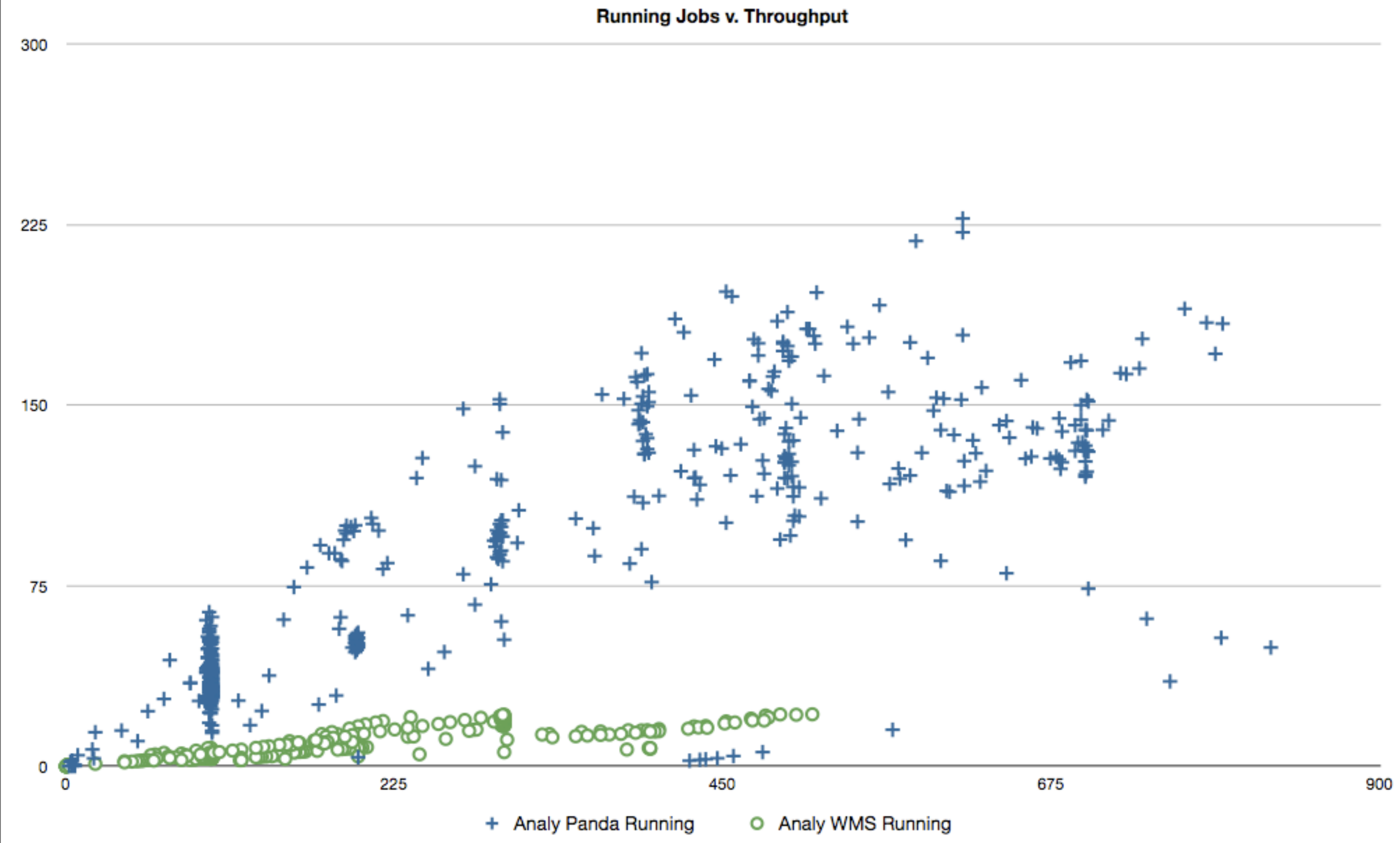
ATLAS Overview



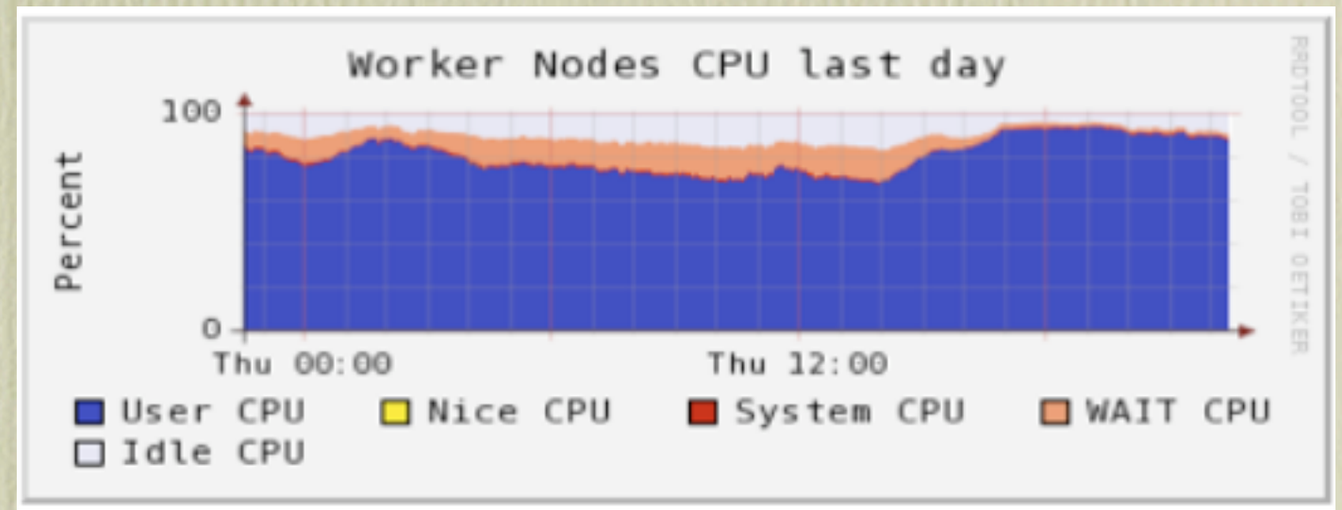
ATLAS Production and Data Distribution

- Production:
  - Very good, as always.
    - 98% efficient @ 1000 jobs (plus otherwise full cluster)
- Data distribution:
  - Passed the metric for 40% AOD+DPD share
  - But only just - “excess” capacity not enough to guarantee a catch-up in short timescale.

ATLAS PanDA Analysis



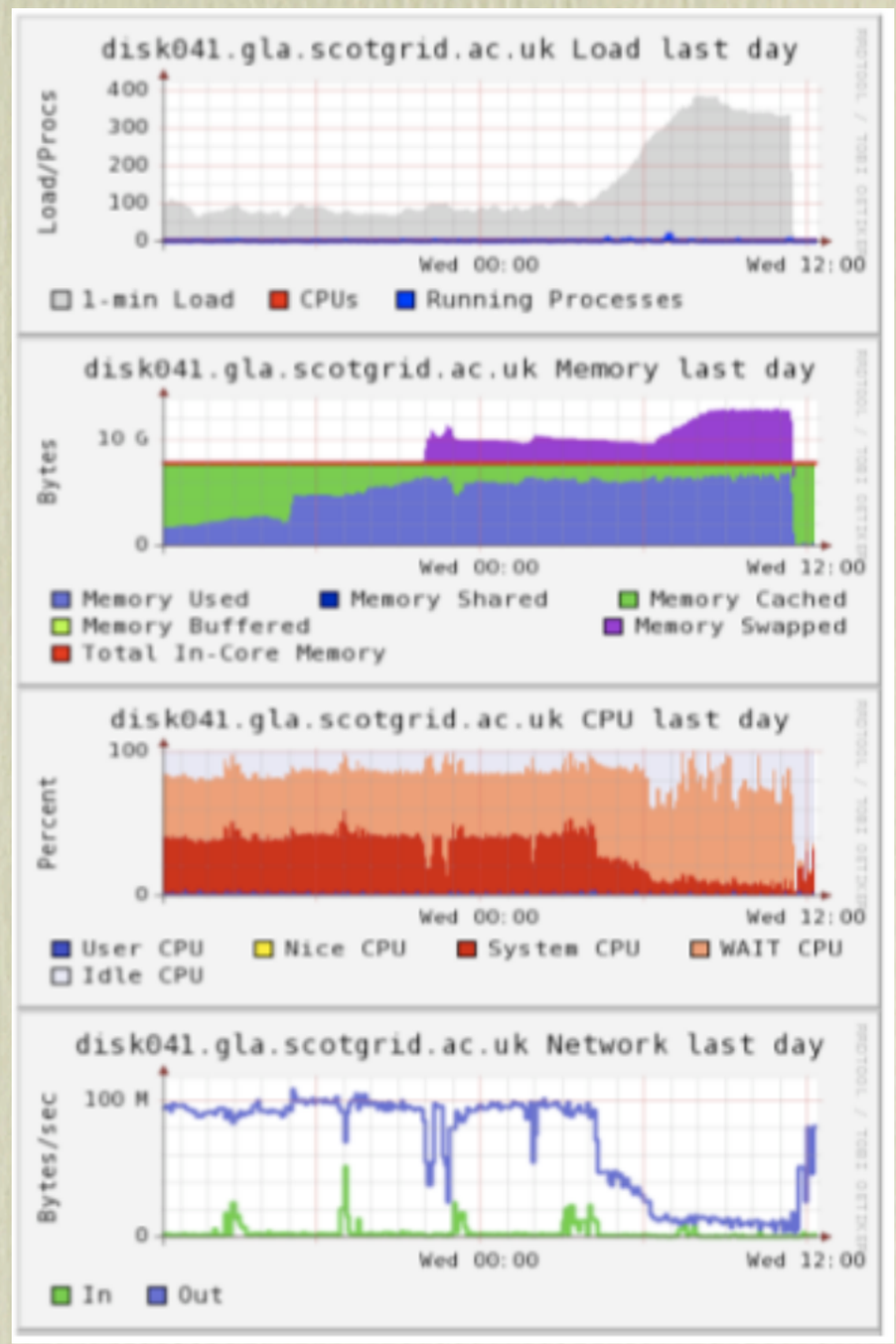
ATLAS PanDA Analysis, 2



- Elbow in throughput/efficiency is due to io contention on Worker nodes themselves.
- We expect this to be a problem on many sites with large numbers of nodes vs single large harddisks.

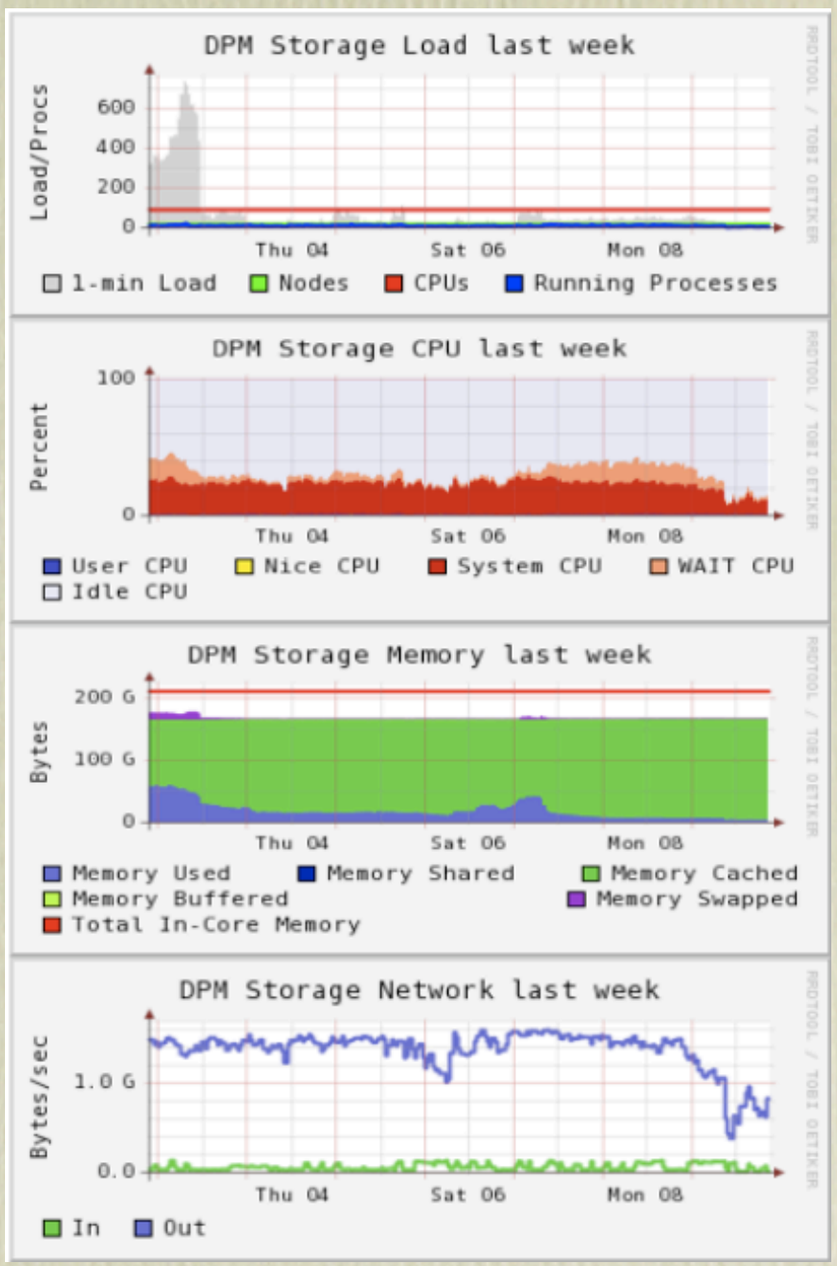
ATLAS WMS Analysis, DQ2-LOCAL

- Is horribly storage intensive (rfio open()s)
- Started off dire for us, due to large RFIO buffer size (which led to disk servers actually exhausting their memory)



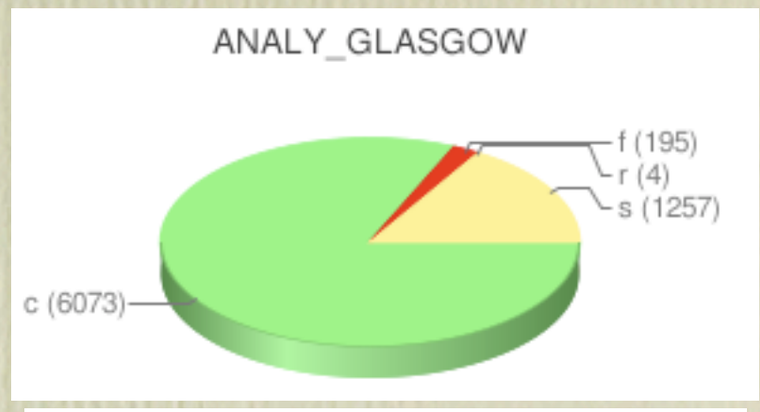
DQ2-LOCAL continued

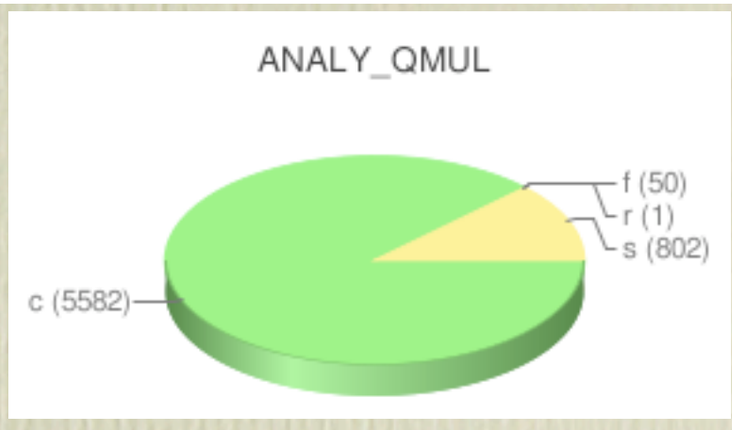
- So, solved that problem, but io demands on disk still very high.
- Probable that the “correct” solution for DPM is Very Small rfio buffer - 4 to 32k.
- (But Lustre may be the “true” solution…)



Post-STEP09 HammerCloud

- Since STEP09, two Hammerclouds have happened vs Glasgow (and QMUL), doing PanDA analysis.
- QMUL "almost beat" us both times (which I am sure Chris is very happy about).
- DPM sites seem to have shown strange rfcp failures under high load, which look worse at Glasgow due to the large number of jobs we ran.





Conclusions?

- What is good for Production is good for PanDA in general.
- But! What was good for (WMS) Analysis is not necessarily good for (WMS) Analysis now.
- Now small buffer-size is optimal.
- Despite high load, sufficient numbers of disk servers *can* be equivalent to small numbers of faster, wider-piped disk servers. (Even with DPM's file distribution logic.)