# UK site experience with direct I/O and Copy to Scratch

Alastair Dewhurst, Gareth Roy

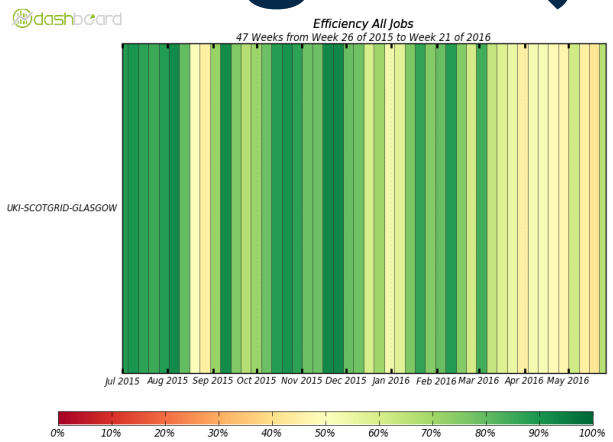
# Outline

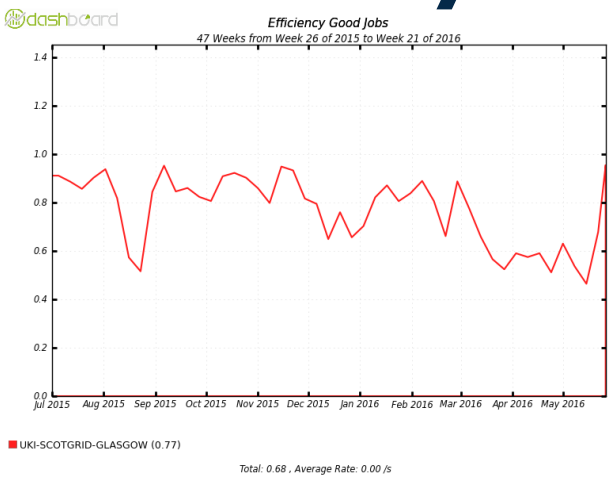


- Recent issues at Glasgow with copy to scratch
- Direct I/O at RAL with CMS
- Hardware purchasing
- Requests to ATLAS

I have tried to keep the arguments general, there will always be a few exceptions.

# Glasgow job efficiency



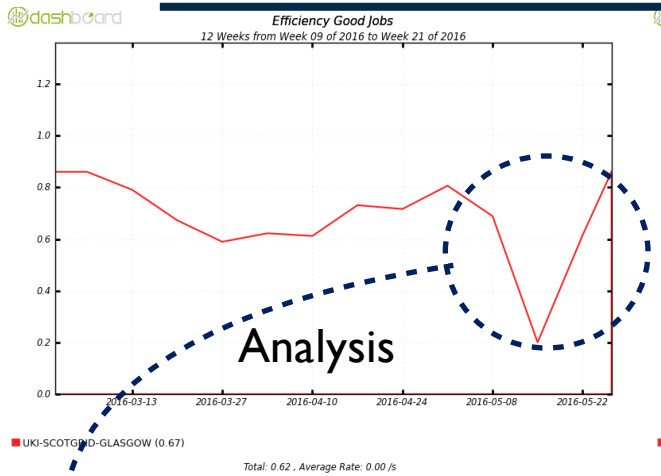



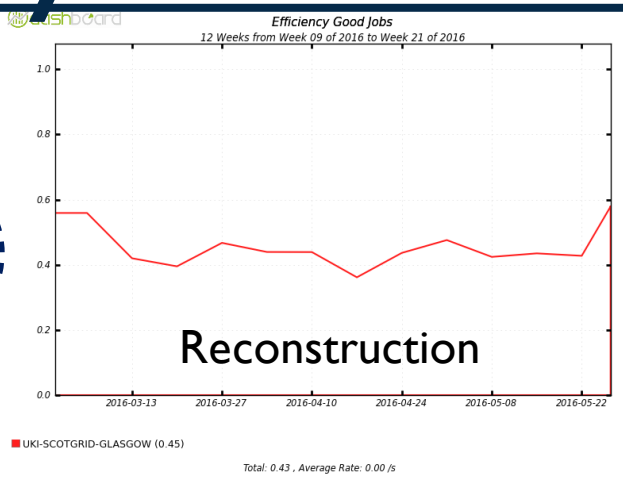

| Date Range | % Eff |
|---|---|
| Last 52 Weeks | 77 |
| 1/7/2015 - 1/1/2016 | 83 |
| 1/1/2016 - 26/5/2016 | 70 |
| Last 12 Weeks | 61 |

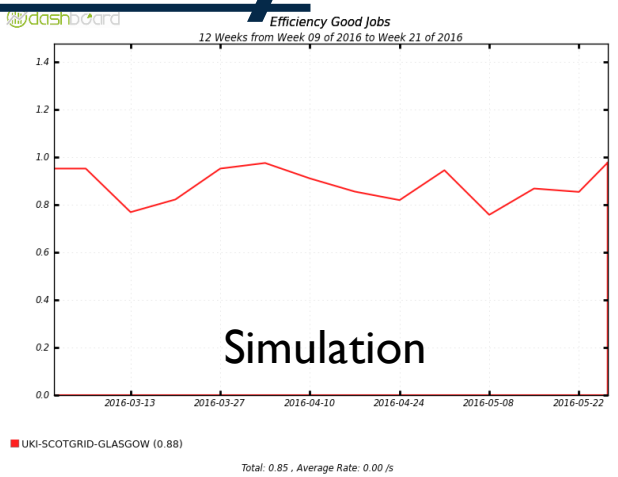
| Queue | Access Type |
|---|---|
| ANALY_GLASGOW_SL6 | Copy to Scratch Using xrdcp |
| UKI-SCOTGRID-GLASGOW_SL6 | Copy To Scratch |
| UKI-SCOTGRID-GLASGOW_MCORE | Copy To Scratch |

- Uki-ScotGrid-Glasgow, 4800 cores, 3.2PB disk.
- 80% ATLAS share of resources, primary customer.
- Noted a steadily decreasing efficiency since the beginning of the year, poor efficiencies showing on local accounting.

# Efficiency March + May



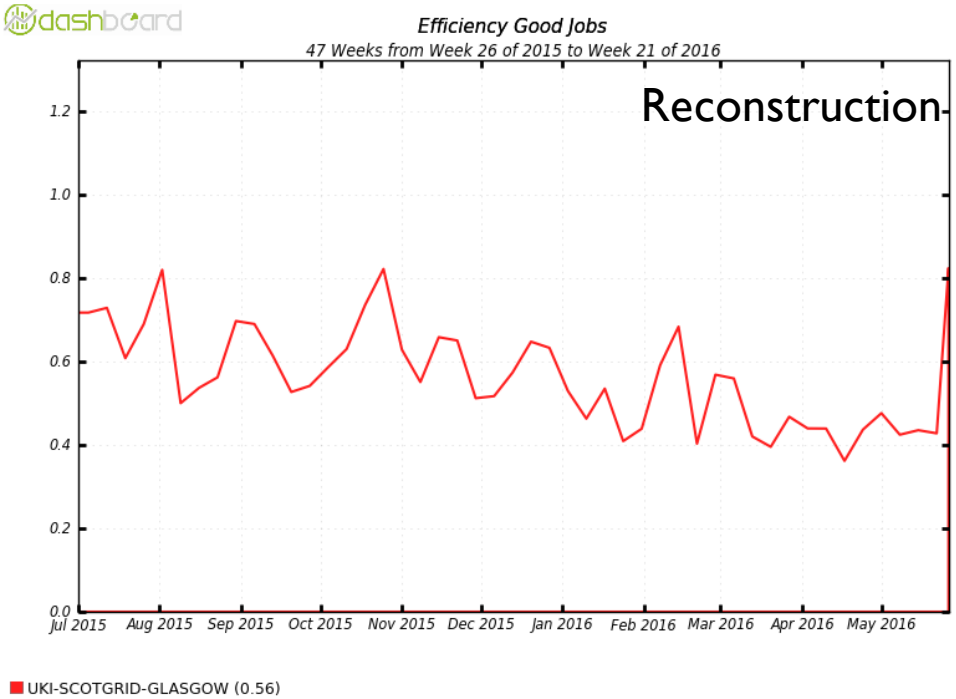

Average Efficiency 67%

Average Efficiency 45%

Average Efficiency 88%

- General inefficiency of MC Reconstruction (~45%) at Glasgow using "copy-to-scratch", larger payloads adds additional load (see next Slide).
- 2 week period where analysis efficiency dropped <20% due to payloads using **lcg-cp** to stage ~60GB per job.
- High I/O loads lead to CVMFS caches becoming broken and approx 5% of WN unmounted filesystems at periods of peak load.

# MC Reconstruction




- To begin with efficiency was reasonably high but has been declining since late 2015.
- Increase in memory usage and storage requirements for scratch disk seem to indicate resource starvation on disk.
- Indicative jobs using 45GB of scratch and in some case use up to to 9.4GB of SWAP (see below).
- On a 32 WN 4x MC Reconstruction can cause large number of IOPS and 100% utilisation of disk.

*On Thu, Apr 21, 2016 at 6:33 PM, Andrej Filipčič <andrej.filipcic@ijs.si> wrote:*

*But, your site is big, and maybe you run more pile jobs per node than others UK T2s. One possibility is also that the nodes are tight with memory, pile jobs use a lot of memory, so not much is left for VFS cache. Or even worse, if they are swapping.*

Hi Andrej, Gareth,

I checked one of the last jobs MC reco finished at UKI-SCOTGRID-GLASGOW_MCORE:

http://bigpanda.cern.ch/job?pandaid=2830687640

and the job was swapping (maxswap=9.4 GB). If I look at the average of the maxswap for jobs at the site is 8.3 GB/job. This could explain the problems with the disk if several jobs of this type land in the same physical node.

What I see is that the memory limit (rss) of the queue is 16 GB/job but these jobs have maxrss=13.7 GB, maxpss=21.7 GB and maxvmem=35.5 GB. So this should be the explanation.

Best regards, Andreu

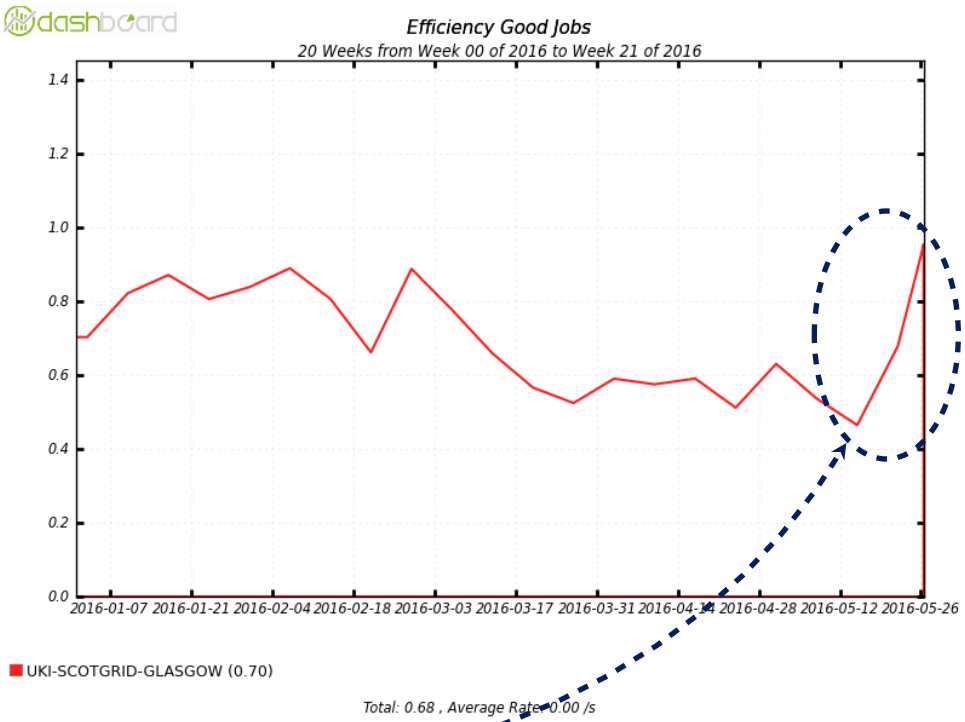
Best regards, Andreu

# Resource starvation

- Many WNs of various generations across the Glasgow estate have begun to exhibit large amounts of %iowait, sometimes as high as 40%
- Further investigation showed many WN disks maxed out on total number of IOPS. It was originally thought this was associated with large amounts of SWAP but appears to be exacerbated by large data set staging.
- 100% utilisation of I/O was leading to poor efficiency and problems with CVMFS cache (accessing and emptying) an WN nodes crashing.
- A Perfect Storm of I/O requirements:
  - Increase SWAP requests due to larger than normal memory requirements
  - Full CVMFS caches due to new releases
  - Larger data sets from both PROD and ANALY queues
  - Not enough IOPS per spinning media to meet all needs.

# Initial solution

- Reduce size of Analysis Sandbox to 20GB
- Reduce total Analysis share and cap at 500 running jobs.
- This improved the <20% efficiency problem.
- Still need an overall solution to the poor MC Reconstruction efficiency.
  - Likely need to limit number of Reconstruction jobs per individual node.
  - As we can't tell the type of payload before hand, looking to see if HTCondor has a resource usage solution.
- Overall performance has begun to improve (but may be payload related).



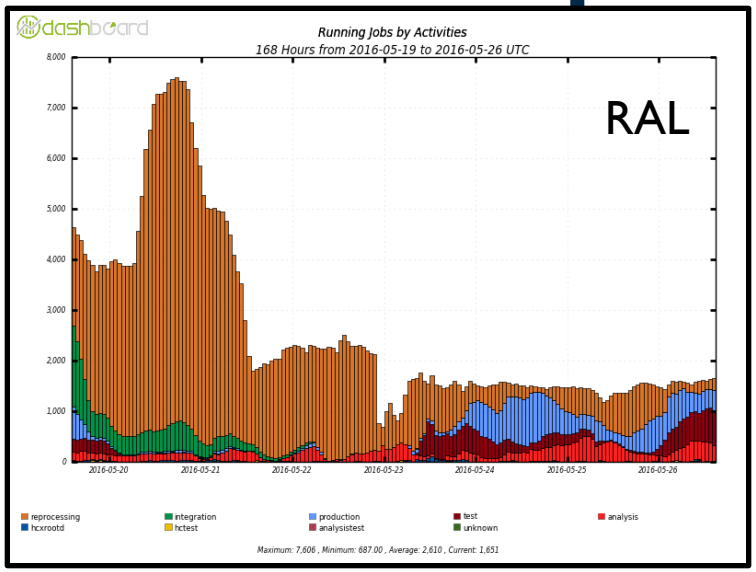
# CMS at RAL

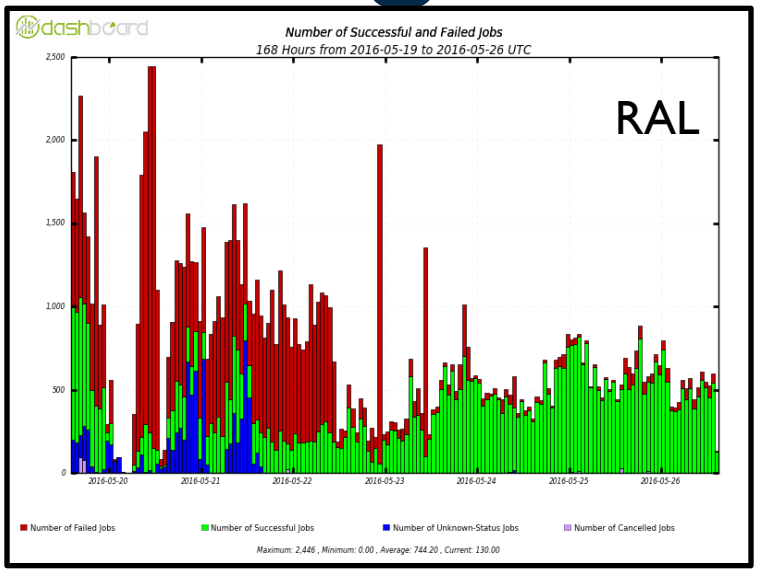


- CMS use XrootD direct I/O for all their jobs (at RAL).
- RAL provide CMS with 2310 TB of storage in their disk pool.
  - Across 23 disk servers.
- RAL frequently observe problems with job efficiency due to direct I/O stressing storage.
  - Problem observed at other Tier 1s as well.

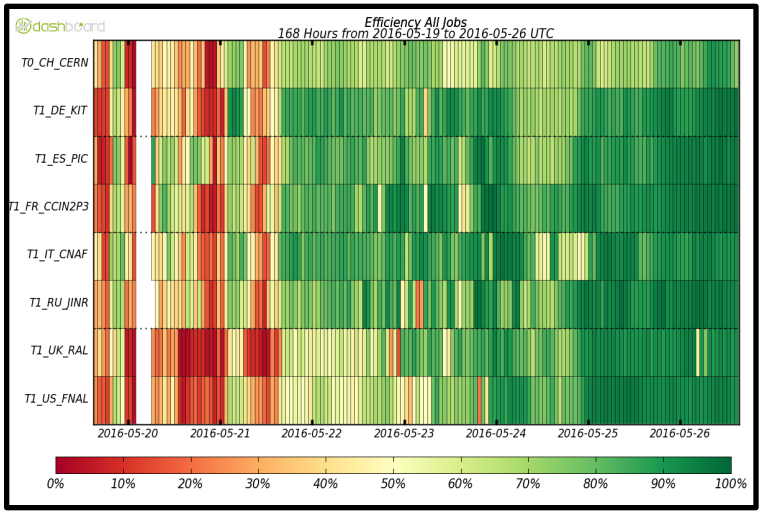
# CMS Re-processing





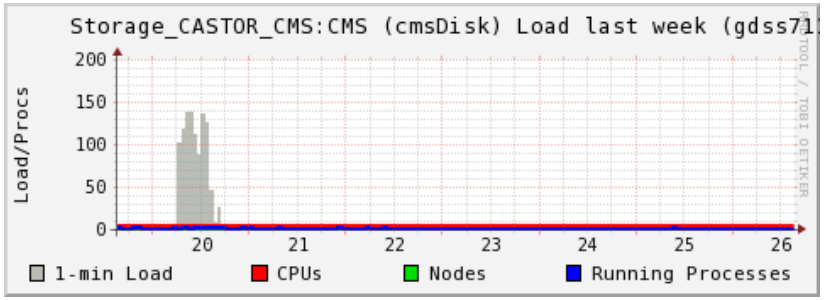

- Large number of data intensive re-processing jobs.
- Significant job failure rate
- Very low job efficiency at all T1 sites.

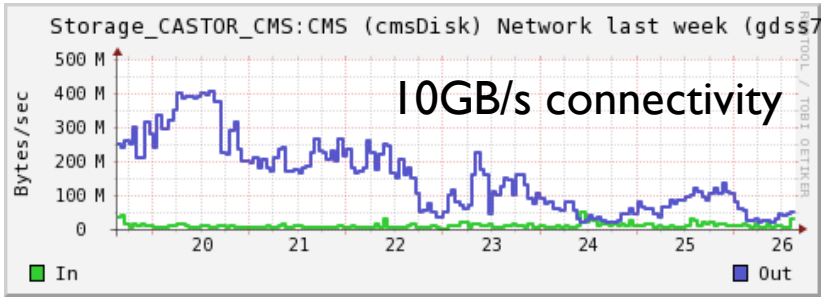


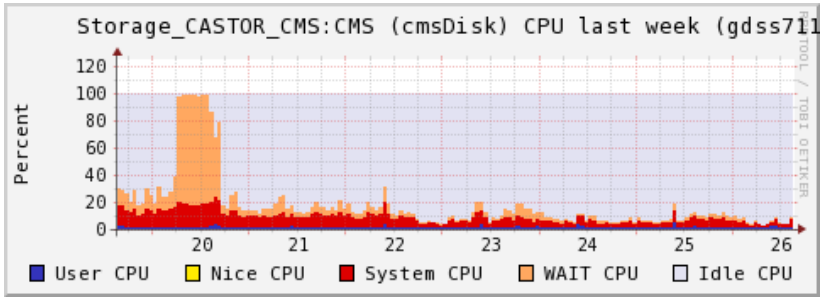
# Impact on disk servers

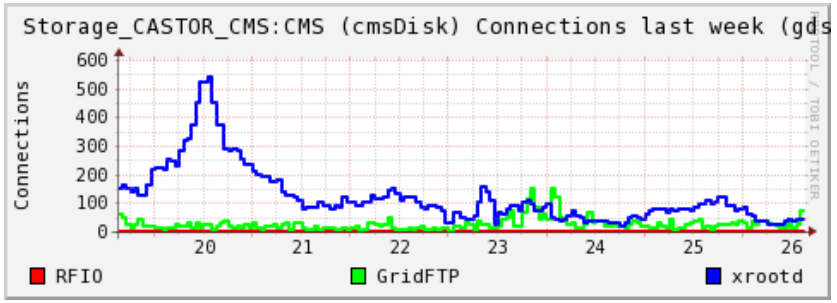


- When the number of connections reaches a critical point (200 – 300) Load and Wait CPU jump significantly.
- Total throughput caps out far below actual limit.



# CMS 'Optimization'



- Time out limit on XrootD connections to storage has to be removed
  - For ATLAS it is 2 hours.
- CMS transfers don't go through scheduler
  - We lose ability to protect storage.
- ATLAS Analysis jobs use Direct I/O at RAL.
  - Works well and even when we see spikes of connections we don't generally see load problems as they aren't accessing much data.

# Comparison



- Copy to Scratch
  - More straight forward for Storage to deal with.
  - Easier for pilot to identify problems.
  - WN can struggle with data intensive jobs
- Direct I/O
  - Better for jobs that require only a very small fraction of data in a file.
  - Takes manpower to optimize.
  - Storage can struggle with data intensive jobs

# CPU procurement



- CPU procurement normally focuses on maximizing HEPSpec06 with little attention paid to everything else.
- New CPU at CERN now comes with SSD storage.
- RAL considering SSD for next year WN
  - Probably will wait until the year after.
- SSD:
  - Significantly better disk performance.
  - Possibly less scratch space per job slot.
- For the next few years we will be working with CPU resources with significantly difference performance when performing data intensive work.

# Requests



- Make setting the access method easier in AGIS and clearer in the logs
- Do not override site settings
- For the longer term future have a method to flag I/O intensive jobs to sites.
  - Use direct I/O.
  - Use copy to scratch to an SSD.
  - Just run fewer.
- Note: There is a meeting on Friday at 9:30am CERN time to discuss this further