Science and Technology Facilities Council

Oxford Xcache

(Material from James Walder, Vip Davda)

Overview

- Oxford decommissioned DPM storage (Switch-off end of June 2021)
  - Configured as storageless site prior to this with transfers straight to Worker nodes
  - RAL configured as the SE
  - Production jobs (mostly) stage whole files to WN

- Xcache deployed Mid April - May (2021) (later upgraded)

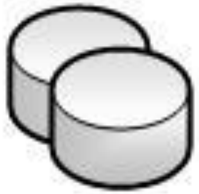
- CRIC / Rucio configuration as Type: “Special”, Token: “XCACHE”
  - Xcache hostname prepended to RAL’s URL path by rucio
- Potential to have a fallback mechanism to other protocols; did not manage to fallback to non-xcache xroot access in case of Xcache failure

- Writes back to RAL go through gridFTP (->WebDav shortly).

Current Hardware / Config

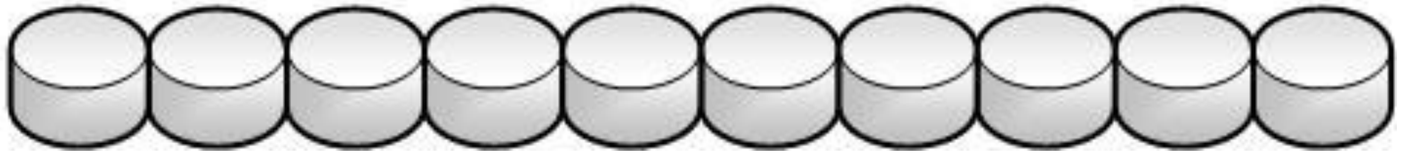
Xcache Meta RAID 1
2.7TB

/xcache/meta/localroot
/xcache/meta/metadata
/xache/meta/var/log
..etc

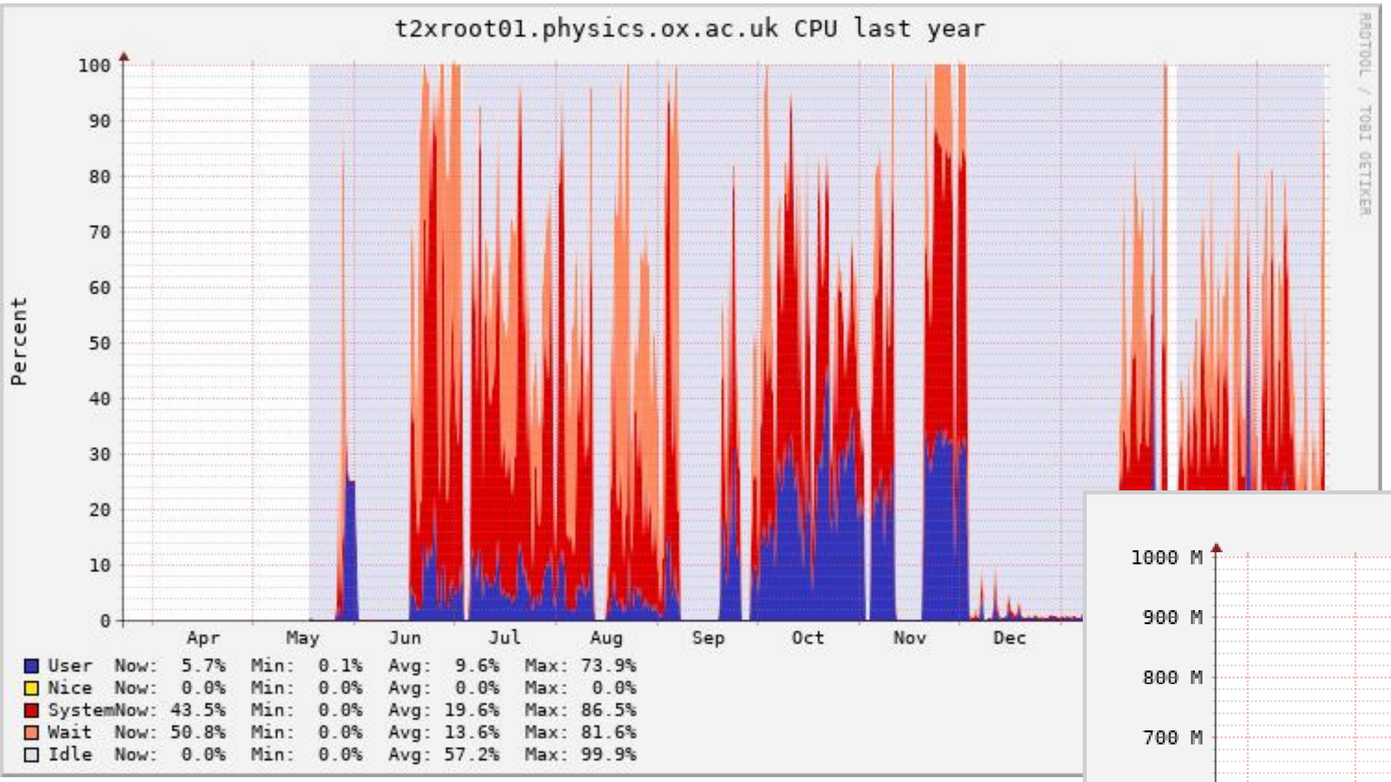


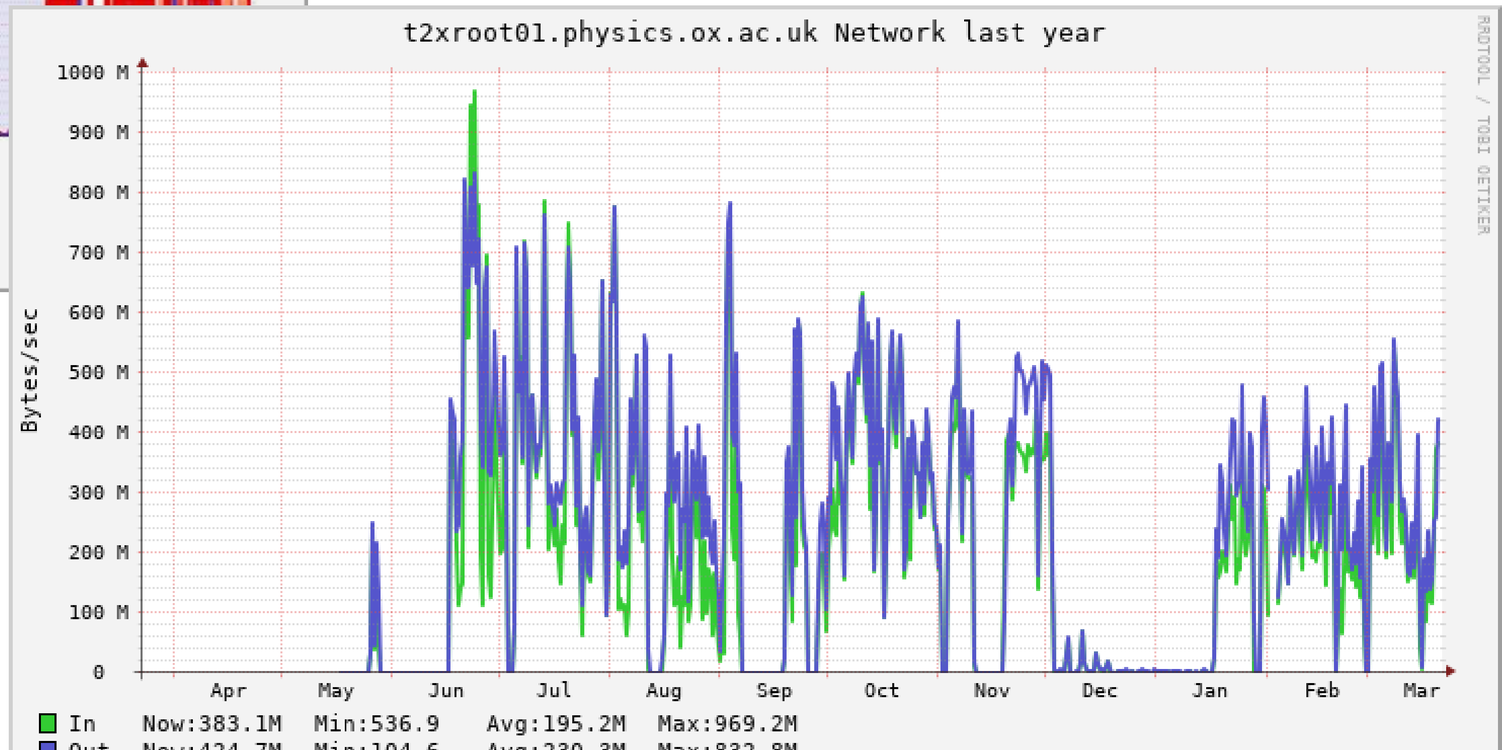
Dell PowerEdge R20xd
CPU: : Intel(R) Xeon(R) CPU E5-2603 v2 @ 1.80GHz
Memory: 48GB

Xrootd Spaces
10 x 4TB



Oxford Local Monitoring





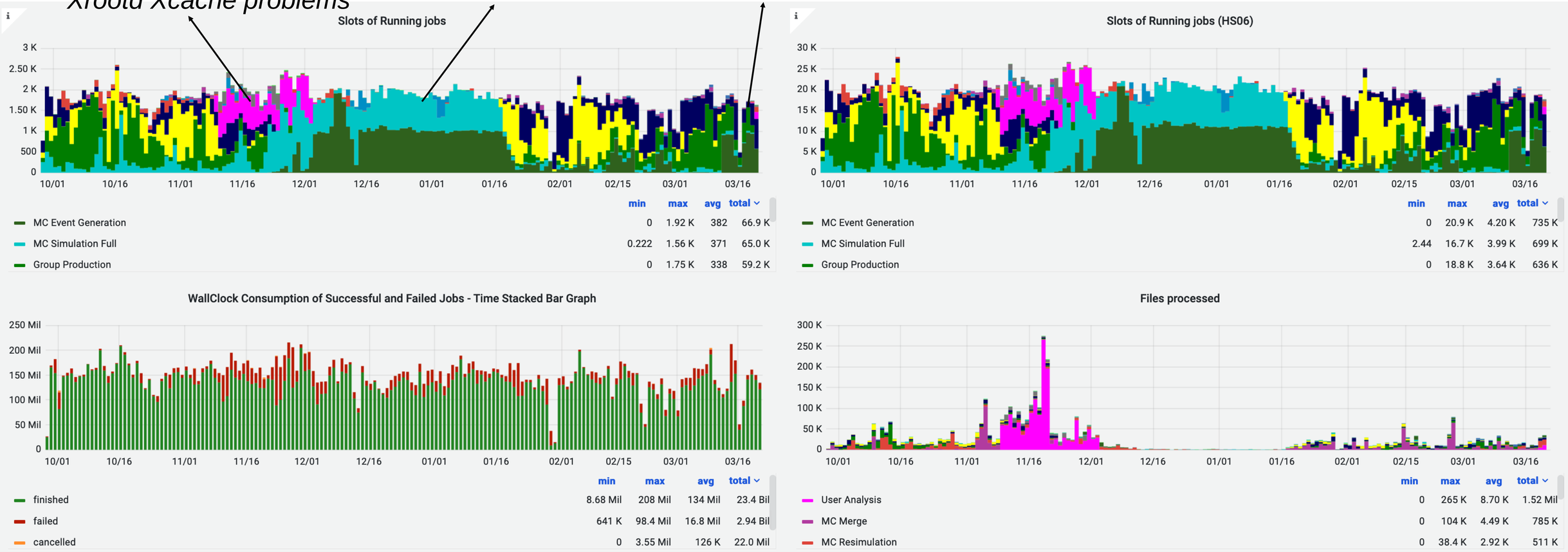
Overview

- Various tests of different ATLAS / Xcache (on/off) configurations attempted.

Analysis (direct IO)
Xrootd Xcache problems

Minimal IO jobs

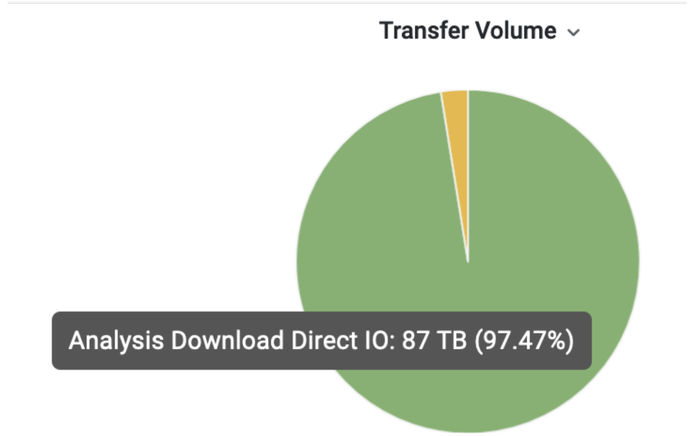
Analysis (direct IO)
(Xrootd 5.4.X on Xcache)

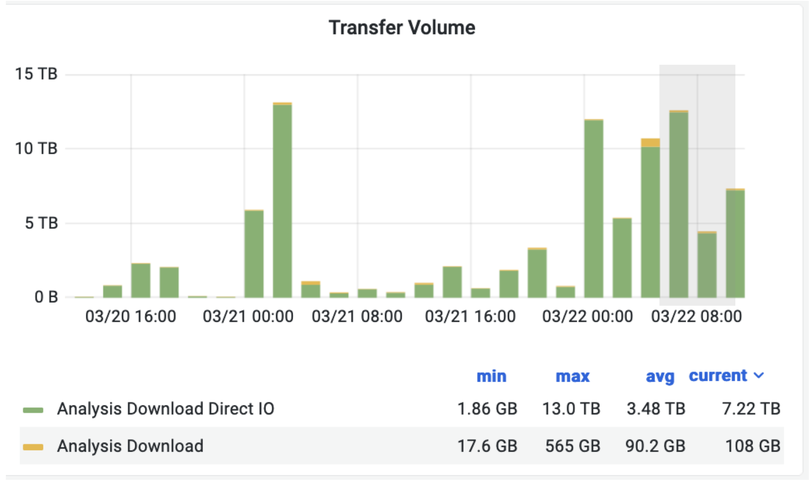


- Initially large problems with direct-io; traced to issues with XrootD, and since fixed in 5.4.1

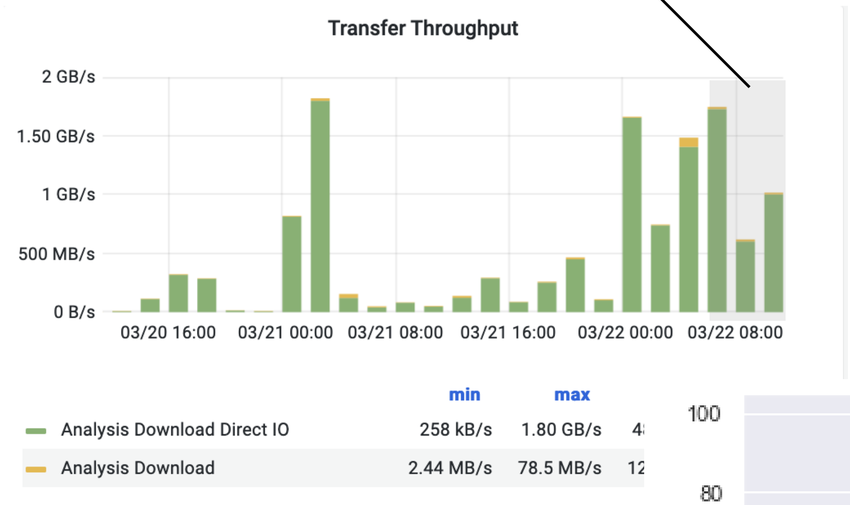
Analysis jobs

- Oxford usually configured to run only Production workflows.
  - Analysis provides useful testbed (providing it doesn't affect users significantly).
- Most data transferred via direct-IO
  - Some still staged; e.g. panda lib (code/binaries) downloaded

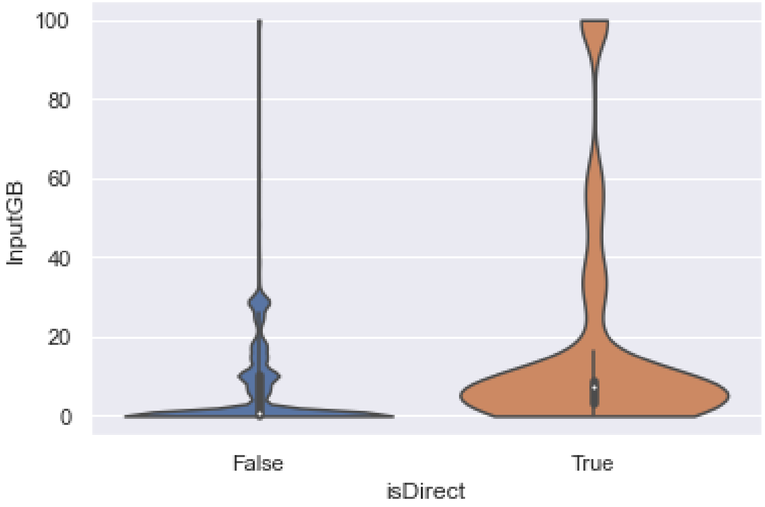




Xcache disabled



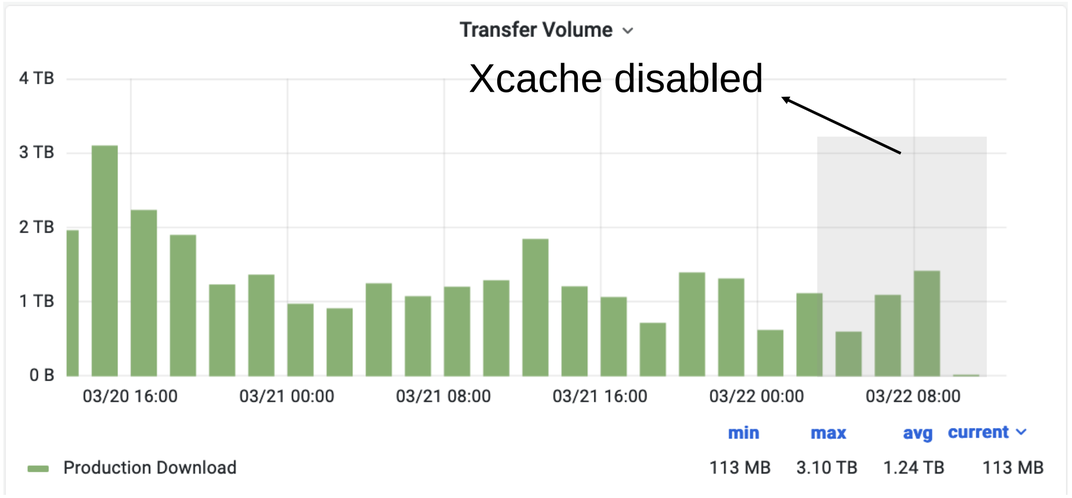
All job types

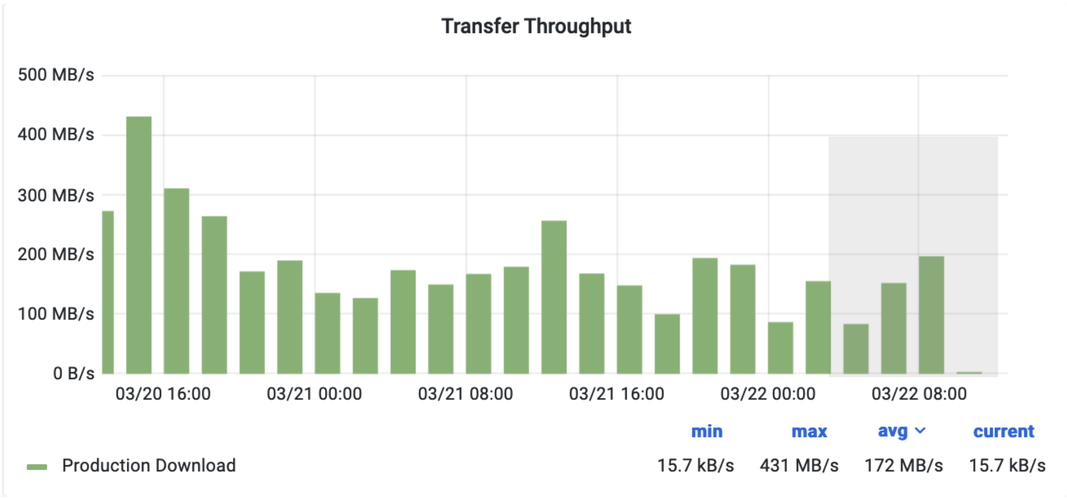


- *Unlikely* that the transfer plots include "only" the data transferred, rather than the nominal size of the file (to be confirmed)
- Direct-io jobs can stream more data per job than staged (and hopefully streaming only the needed data). (i.e not filling up the scratch space)

Production jobs

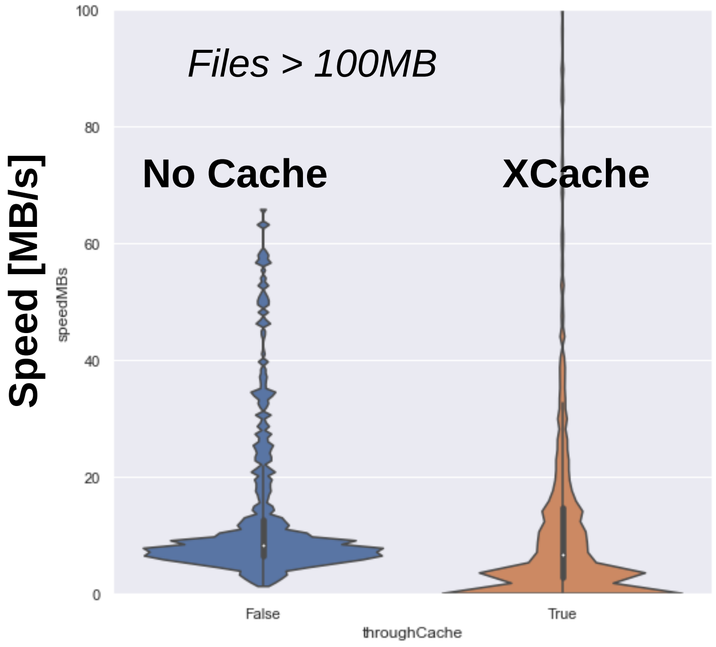
- Production jobs generally stage all data

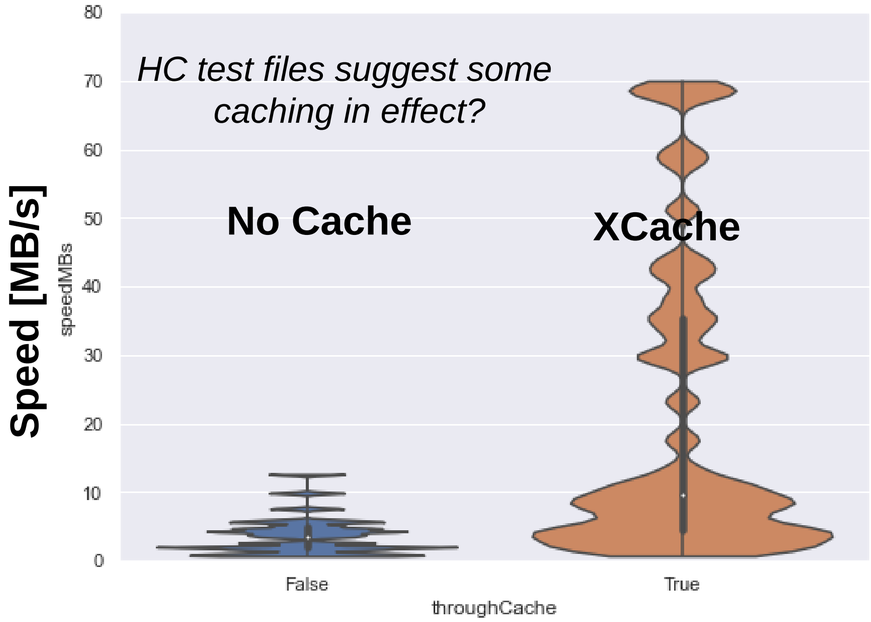




- Mean transfer speeds rather similar between transfers using Xcache or direct;
- Not so simple to remove / isolate external factors

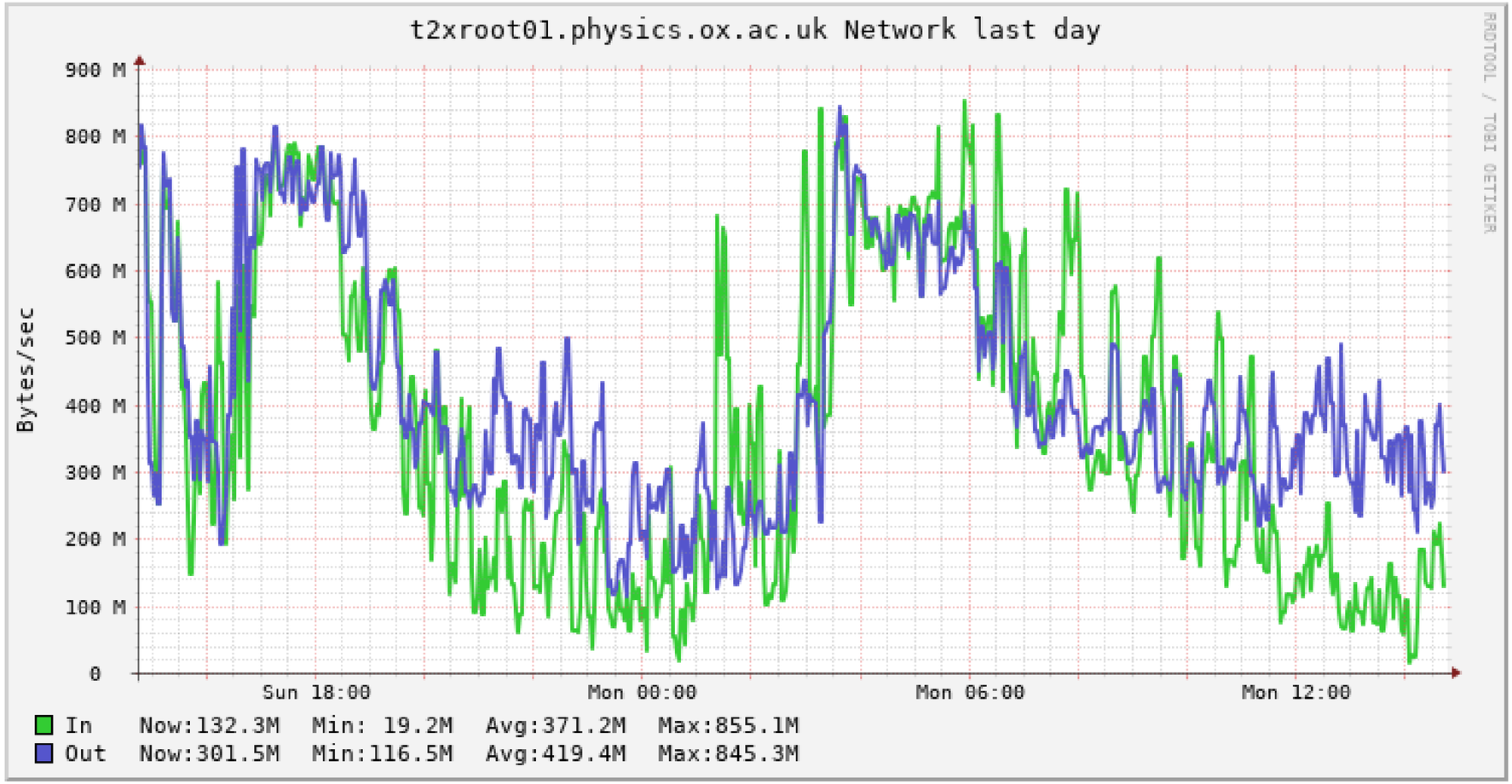
Speed MB/s	count	mean	std	25%	50%	75%	max
No Cache	1061	13.1	12.2	6.5	8.4	12.7	65.9
Xcache	5461	12.9	18.3	2.8	6.7	14.9	174.2





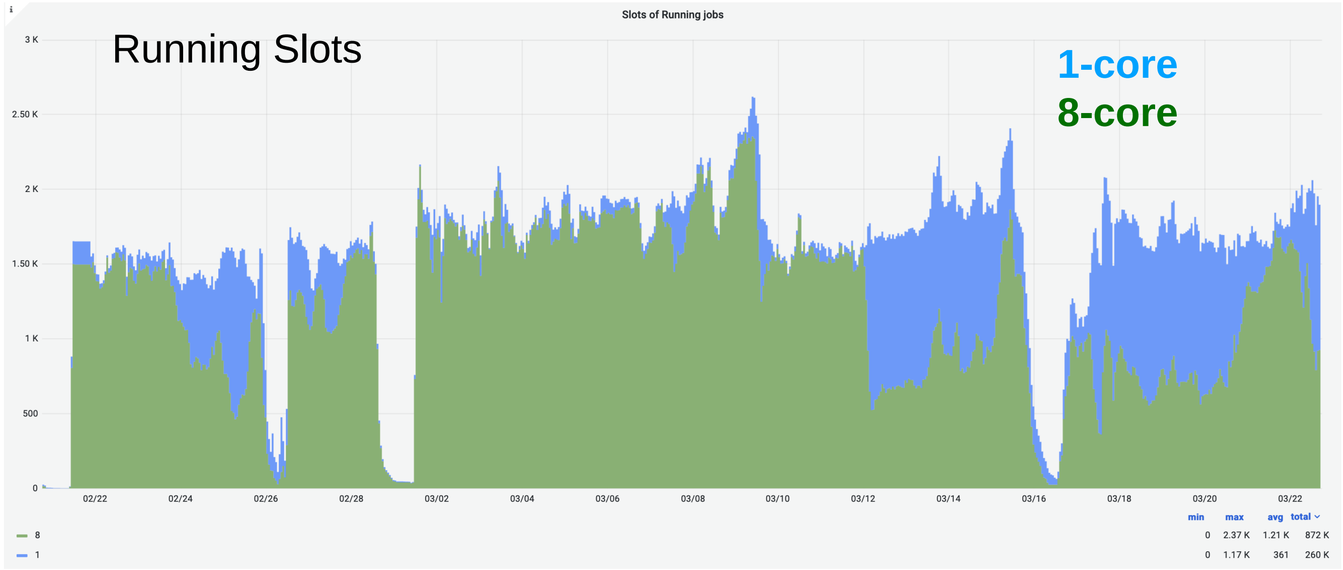
Oxford Network Monitoring

- Evidence of caching?



Site throughput

- Site provides ~ 2k slots, ~ 20kHS06 of compute
- Averages ~ 300MB/s of production throughput
  - => ~ 1.2Gb/s for 10kHS06
- Further breakdown per job-type to be done

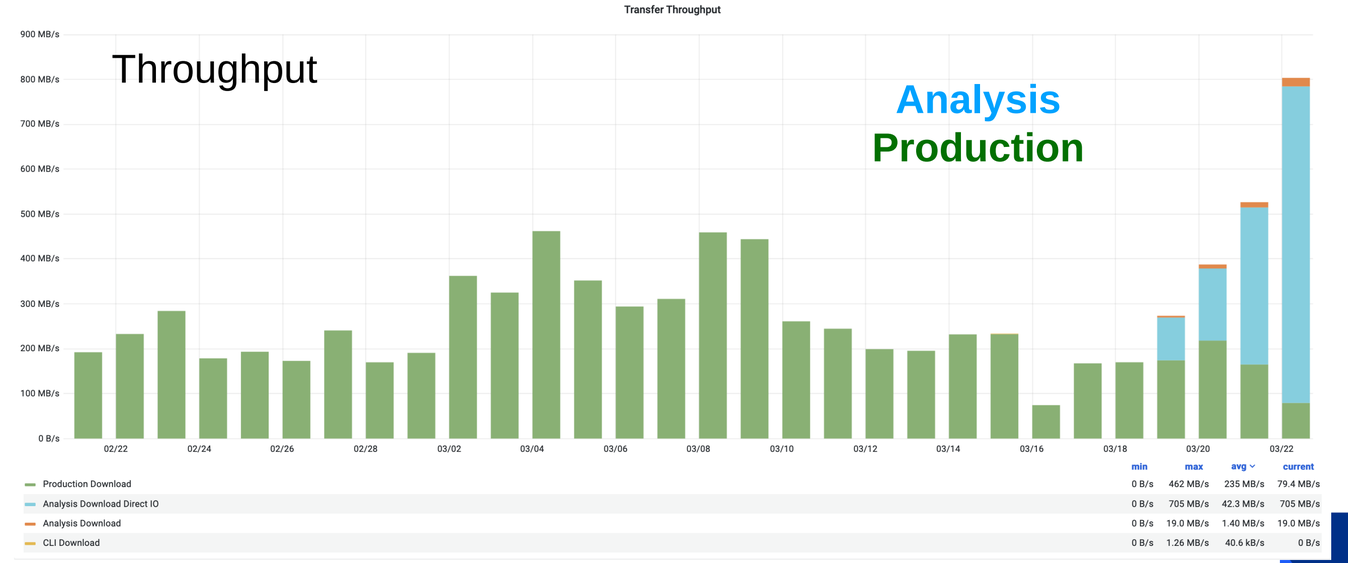


- A previous study (2019) concluded:

- low IO site (a priori diskless= only WAN) : 0.5 Gb/s for 10 kHS06 (1k core)
- high IO site : 5 Gb/s for 10 kHS06 (1k core)
  - Assuming ratio 1 WAN to 5 LAN:
    - WAN : 1 Gb/s for 10k HS06
    - LAN : 5 Gb/s for 10k HS06

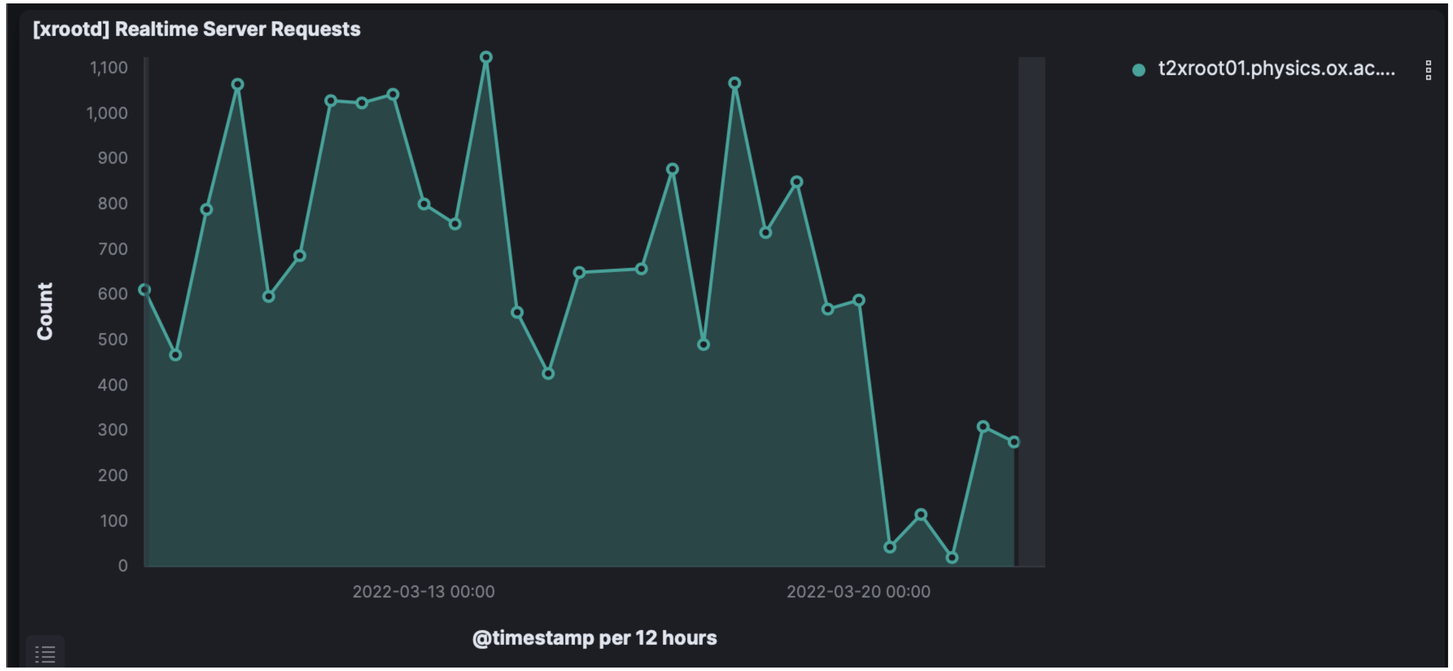
- For the analysis data, *O*(2MB/s) from the plots
  - Also matching up with previous study

- 2 MB/s per job on global average
  - => 16 Gb/s for 10k HS06 or 1k job slots



Edinburgh Monitoring

- Stats collected to Edinburgh's monitoring (https://gridpp.monitoring.edi.scotgrid.ac.uk)



- Suggests a low cache hit rate from (sampled?) data; <1% for files accessed in last 2 days

Oxford’s Xcache Plans

What next?

- Use Xrootd Virtual Placement?

- Use better hardware for Xcache server.
- 25Gb network connection to the local switch stack.
- 20Gb to University core network router.
- 100Gb from University to JANET.

Summary

- Xcache running at Oxford since early last year
  - Adds another point-of-failure in the chain
  - Typical ‘single use’ files (and with a large SE like RAL):
    - Hit rates of the cache are low
- 5.4.1 brings some fixes compared to 5.3:
  - Direct-io analysis jobs able to be run with RAL
- RAL->Oxford good network connectivity; may not be able to test latency well ?
- Tangible benefits may be available to ‘smaller’ sites; but may have less resource to devote to R&D
- More intelligent uses of the cache (e.g. Virtual Placement) to decouple data placement from the job workflow
- Other studies suggest specific ways of configuring disks (e.g. Performance for parallel reads - Use multidisk-mode instead of Raid: https://indico.cern.ch/event/727208/contributions/3444604/attachments/1859894/3056280/XCache-FeaturesEtc-Lyon-2019.pdf