Science and Technology Facilities Council

XRootD: Achieving 100Gb/s Data Transfers

Tom Byrne, Jyothish Thomas, James Walder
SCD, RAL-STFC

CHEP 2024, Kraków, Poland
19–25 October 2024

Science and
Technology
Facilities Council

Trying (and failing) at

XRootD: Achieving 100Gb/s Data Transfers

Tom Byrne, Jyothish Thomas, James Walder
SCD, RAL-STFC

CHEP 2024, Kraków, Poland
19–25 October 2024

Motivation

- LHC Experiments and WLCG Sites preparing for HL-LHC conditions
  - Data Challenges (e.g. DC24) essential for testing capabilities and scales
- New communities of large-scale experiments, e.g. SKA
  - Nominal data rates of 100Gb/s from each (of two) SKA Telescopes at in full operations.
  - Up to 1EB of data annually
    - (See SKAO Top-Level Roadmap)
- Need to have:
  - Increasing file sizes (10–50 GB) ?
    - With moving each file faster
  - Increasing throughput
    - Moving all data faster
- Deal with contention of the network for different VOs ? (Not discussed here …)
- While the future may be Object Storage, support still required for POSIX-like access …

HL-LHC expected CERN – T1 data rates

T1 Rates	HL-LHC Minimal	HL-LHC flexible
T1 (Total)	4.8 Tb/s	9.6 Tb/s
e.g RAL	610 Gb/s	1.22 Tb/s

DOI: 5532452

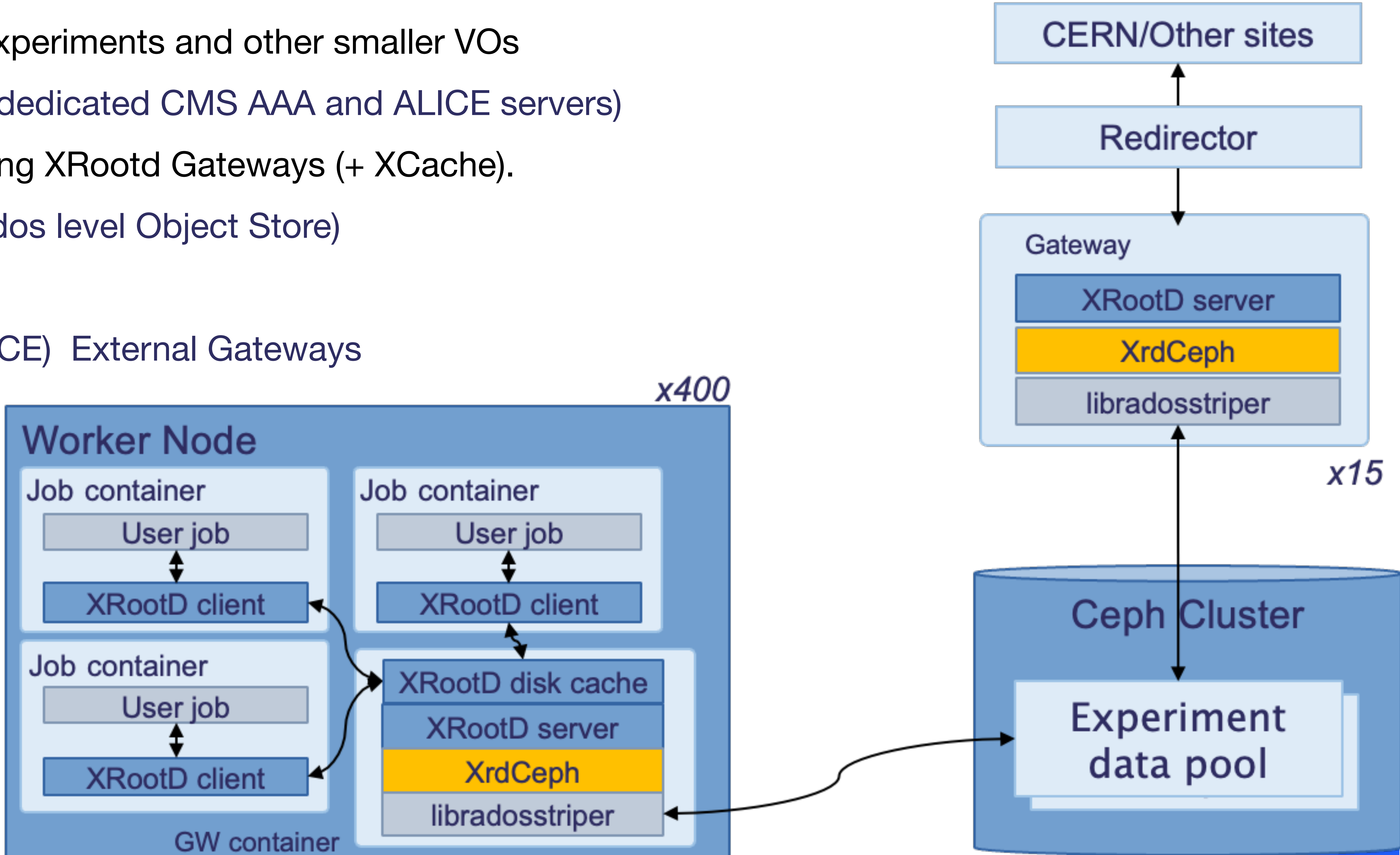
Milestone	Year	Primary Activity	Estimated Data Rate (Low)	Estimated Data Rate (Mid)
AA4	>2030	Full SKA design	216 PB/year, 55 Gbps	400 PB/year, 100 Gbps

Shari Breen @ SKAO

Current (WLCG) Production Deployment @ RAL

- UK WLCG Tier-1 Facility: (Rutherford Appleton Laboratory)
  - Serves all main WLCG Experiments and other smaller VOs
- XRootD External Servers (+ dedicated CMS AAA and ALICE servers)
  - Worker Nodes also running XRootd Gateways (+ XCache).
- XrdCeph plugin to Ceph (Rados level Object Store)
  - > 70 PB (usable) storage
- ~ 15 (excluding AAA and ALICE) External Gateways
  - 25Gb/s NICs

See talk: J.Thomas (next) for more details

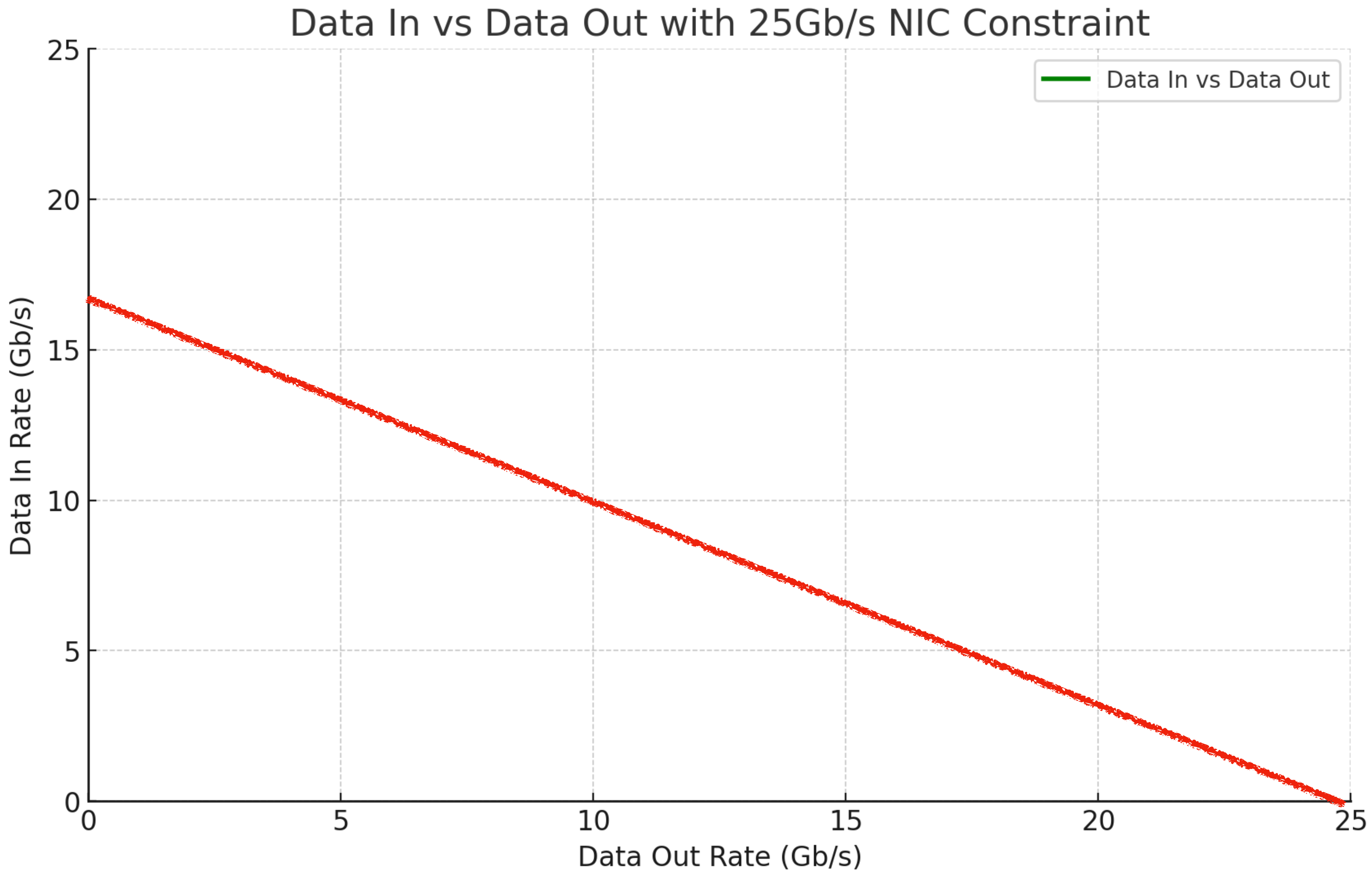


Current (WLCG) Production Deployment @ RAL (2)

- Gateway DTN (Data Transfer Nodes) act as interface between External clients and Ceph OSDs.
  - i.e. Every byte in => bytes out (and vice versa)
  - Also, Checksum is performed on the server after transfer to OSDs
    - Additional bytes read back in to gateways on Write.
- Assuming similar Reads and Writes; 25Gb/s NICs; => ~ 1GB/s file transfer in and out.
  - require ~ 75 hosts for HL-LHC (naively) (with saturated NICs)

- So O(15–20) hosts for HL-LHC, if can saturate 100Gb/s … ??
- Use On-the-fly checksumming to avoid additional read
  - (Or, OSD-level checksums)

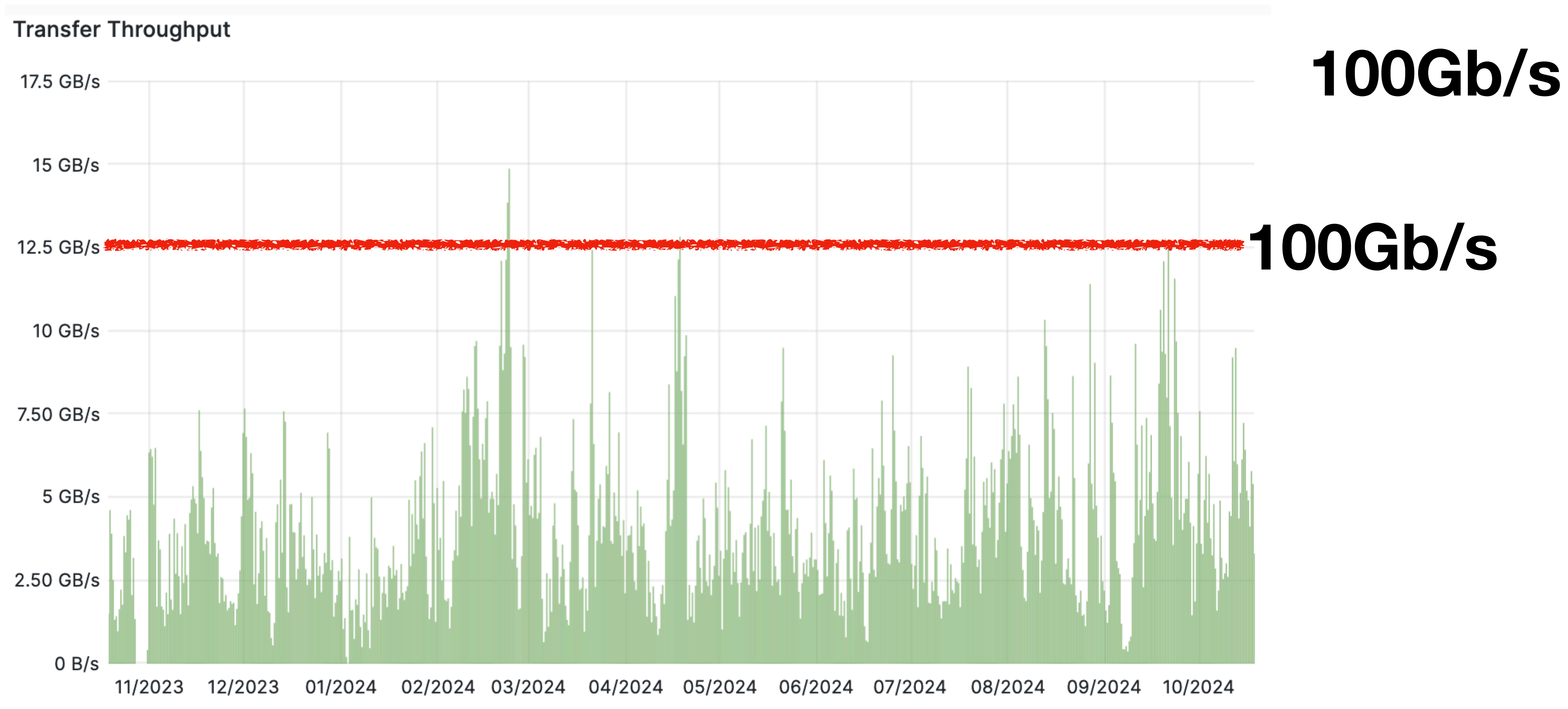
- Significant effort in the future of Networks,
  - Unlikely to be the primary constraint.
- Can software (e.g. XRootD) support such throughput, and, can the storage keep up …



Current Setup: RAL (3)

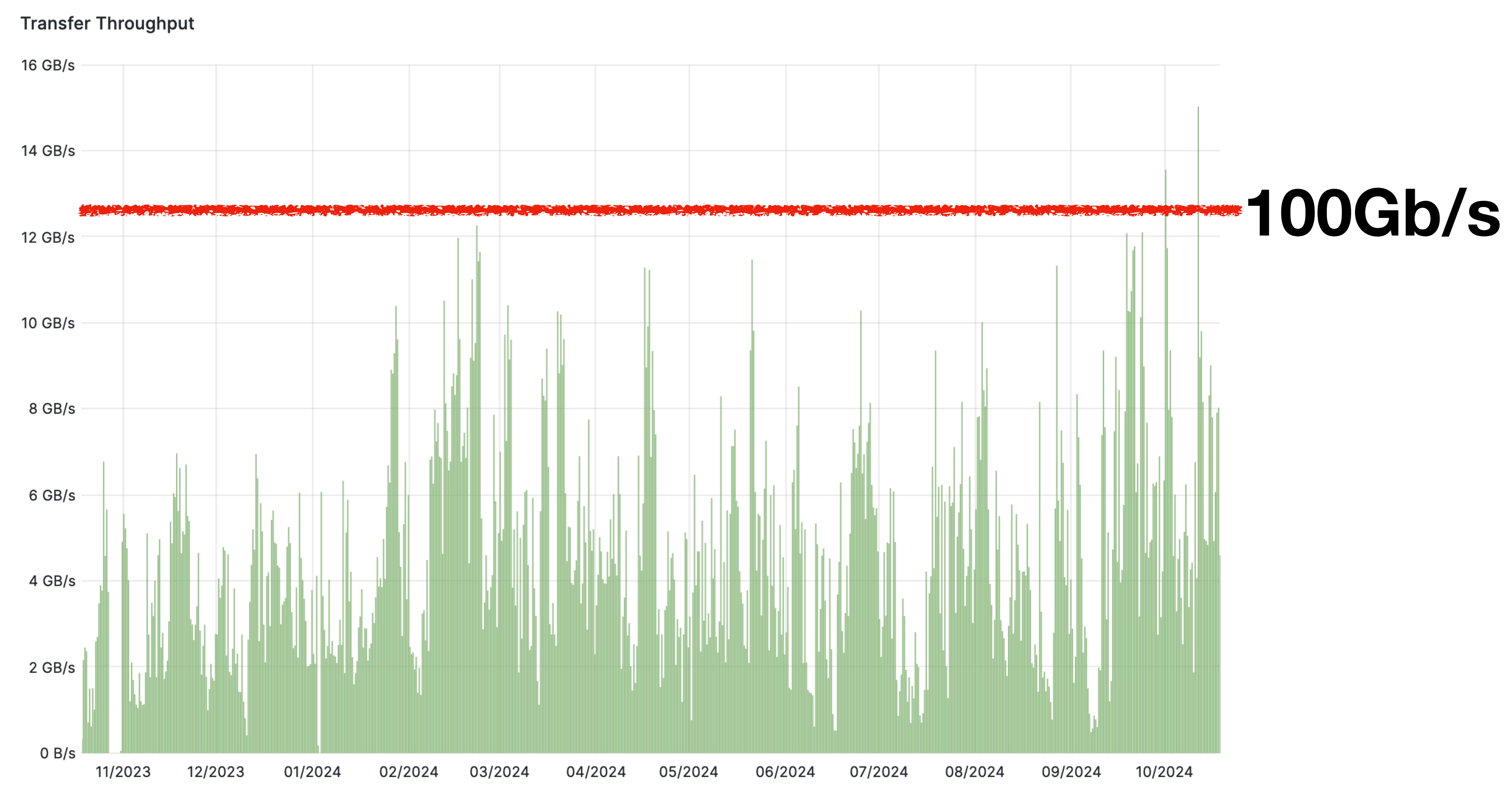
- RAL currently able to saturate the ~available 100Gb/s of OPN connectivity, when required (includes also traffic to WNs from CMS-AAA).

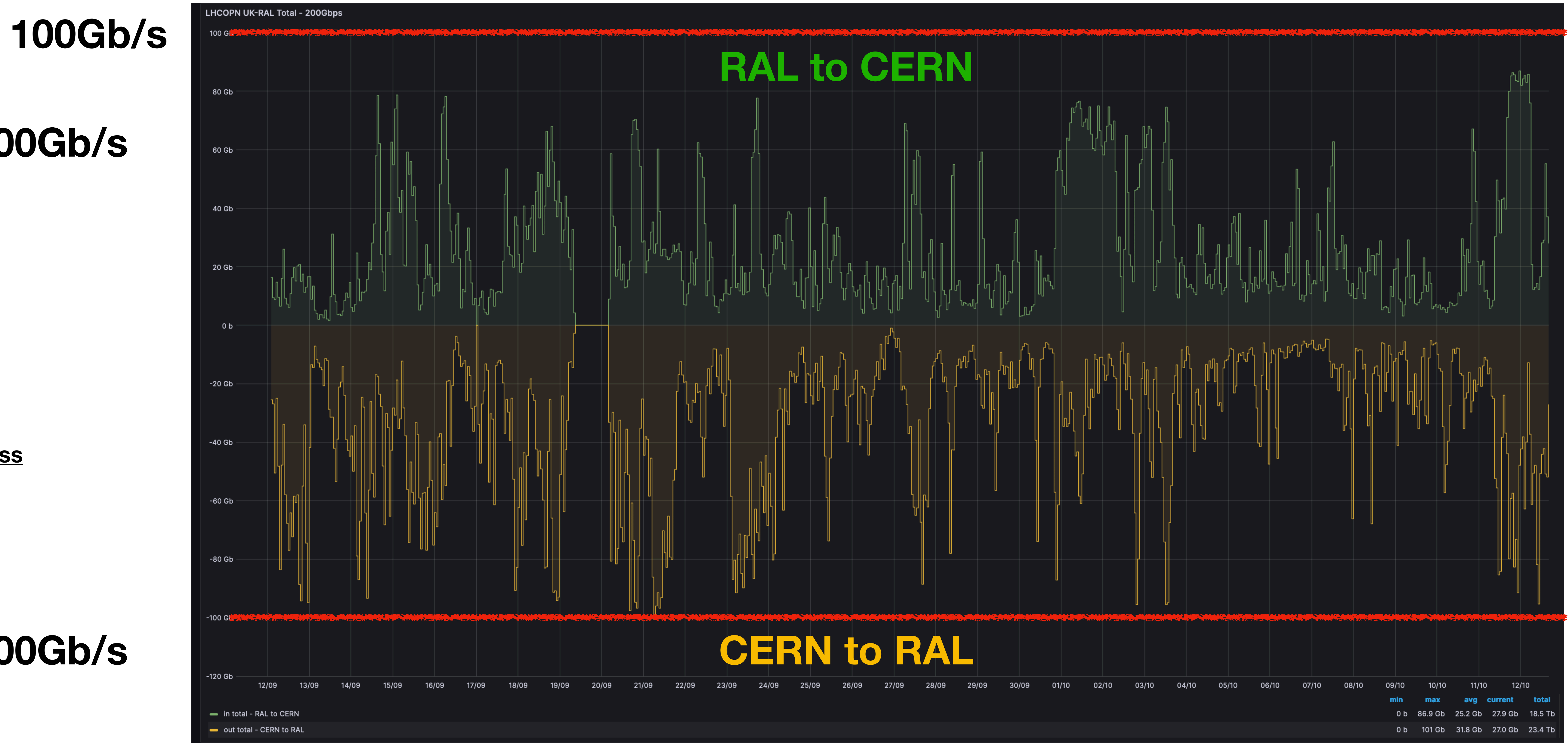
Writes into RAL [Disk] (FTS)



FTS: Ingress

Reads from RAL [Disk] (FTS)





LHCOPN RAL

Achieving 100Gb/s Connectivity

- To achieve 100Gb/s connectivity, requires:
  - ❑ Well tuned hosts (e.g. apply Esnet tuning (100G tuning))
  - ❑ Uncongested Network / negligible packet loss
  - ❑ Jumbo frames (not in the tests here, but for long-haul transfers considered vital)
  - ❑ Maximum throughput via optimisation of:
    - Data rate per flow
    - Max number of flows that can be handled efficiently
    - Prioritisation of flows (e.g. SENSE).
  - ❑ Backend storage capabilities …

Test Setup and Configuration

- Physical hosts with similar spec to existing hardware, with addition of:
- Mellanox ConnectX-6 100Gb/s dual-port NIC
  - Bonded NIC, with dedicated VLANs for internal and external traffic in DMZ
  - MTU 9000 (remote storage tests ran at - typically - 1500)

XRootD server
PowerEdge R650xs
2x Intel Xeon Gold 6326 2.9G. (32Cores / 64 Threads)
128 GB RAM
Centos 7

- Ceph cluster:
  - Development cluster (for testing outside of prod workloads)
    - (Experimental cluster delayed commissioning)
  - 4 nodes
    - mixture of 4 and 8TB HDDs
    - mixture of 12 and 16 drives per host
  - 5% flash per OSD
  - 25Gb/s networking
  - CephFS – kernel mounted

Deneb dev cluster theoretical limitations

- 54 HDDs: ~10,000 IOPS
  - With 8+3 EC => ~1000 client write ops per second.
- 4 nodes; 25Gb/s => 100 Gbps
  - Network overhead for EC shard redistribution is 1*(10/11), so theoretical throughput is ~50% of available bandwidth
  - So ~50Gbps or 6.25GBps

TCP Testing

- Basic iperf3 testing for baseline network connectivity.
  - Iperf3 (3.15beta) used, with multithreaded streams
- 'Nearby' (2.6ms) JANET iperf3 server for remote tests

- By 16 parallel flows have saturated to ~ 100 Gb/s on egress and ingress rates

Streams	Egress [Gb/s]	Ingress [Gb/s
1	9.3	10.6
2	18.5	18.7
4	36.9	36.9
8	61.8	80.4
12	76.4	96.8
16	97.1	97.1
24	97.8	97.7
32	98.3	98.3

net.core.rmem_max = 2GiB

net.core.wmem_max = 2GiB

net.ipv4.tcp_rmem = 4096 128k 1GiB

net.ipv4.tcp_wmem = 4096 128k 1GiB

net.core.default_qdisc = fq

XrdBlackhole

- Motivation:
  - Tools such as iperf3 perform tests of network / host-level tuning
  - Needed mechanism decouple storage backend from XRootD layer
  - Possible options: nullfs, writing to /dev/null
  - In addition, needed a way to generate a set of test data for reading

- XrdBlackhole :
  - Simple XrdOss plugin to XRootD
  - Data written in is dropped in the OFS layer
    - Performs simple metrics calculations
  - Minimal FS mocking
  - Ability to read data from 'crafted' path name (e.g. file_*100GB*_001)
- Still a WiP.

XrdBlackhole: Local Tests

- Run tests from the XRootD host:
  - 10GiB File in tmpfs memory ~ 8Gb/s for a single file
  - **Writing** files in parallel, saturating around ~ 32 Gb/s of write speed.

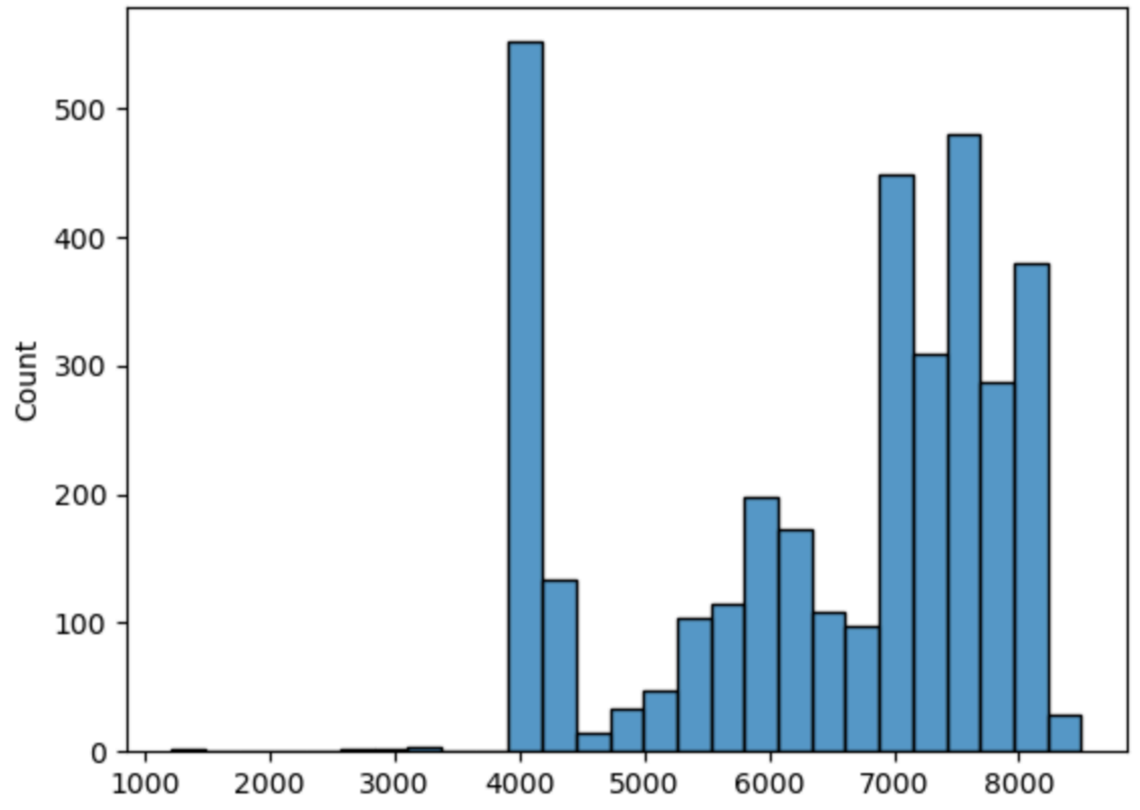
Parallel Files	1	2	4	8	16
Total throughput [Gb/s]	8	13	24	30	32

- For **Reading** (to /dev/null) via curl client (macaroon authZ)

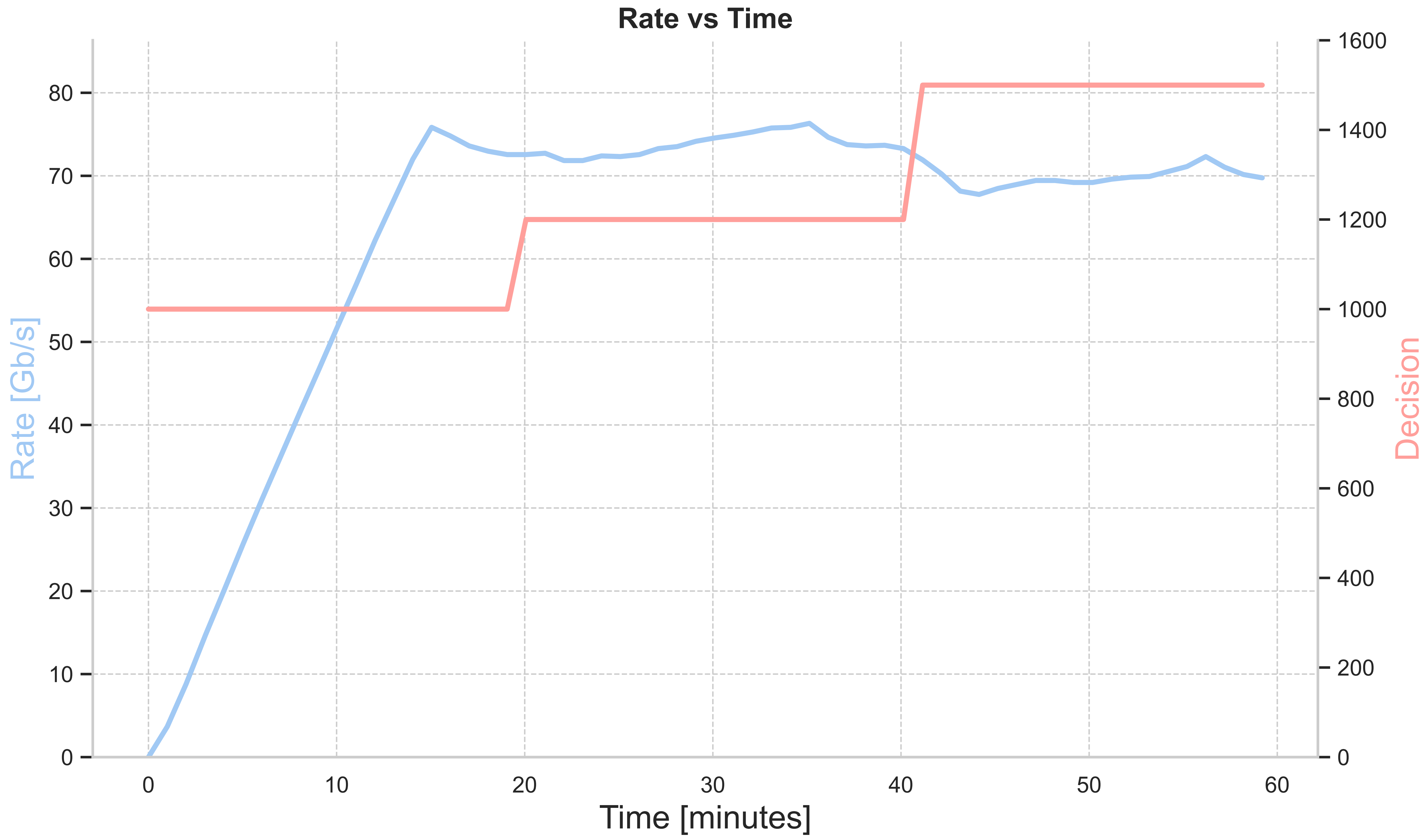
Parallel Files	1	2	4	8	16	32
Total throughput [Gb/s]	14.5	27	39	91	150	184

Writing from Remote Hosts

- Choose a source site of ‘infinite’ capacity (aka CERN)
  - Write into the “blackhole”
  - No additional file checking (e.g no checksum verification)
- FTS to schedule transfers
  - Modify site limits and Link limits to control number of simultaneous transfers.

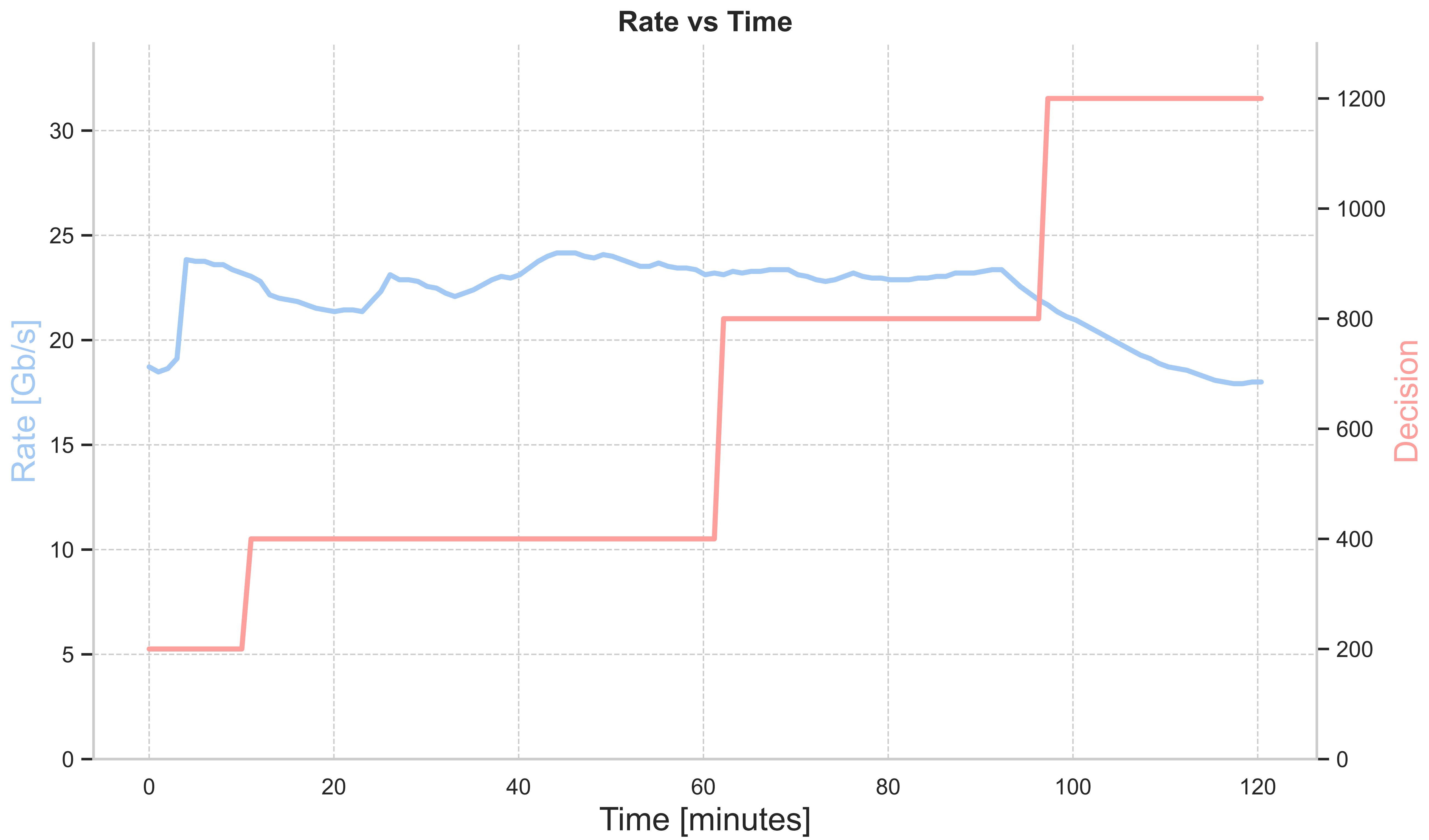


- Dataset: ~ 3.5k files, mean size 6.5GB
- Max throughput ~ 76 Gb/s @ 1200 concurrent transfers
  - Average 50Mb/s / per file
- Increases to 80Mb/s /per file for 300 concurrent transfers



CephFS Testing

- Managed to achieve writes into a single xrootd client (no storage connected) up to ~75 Gb/s
  - Connected storage is likely to be the real bottleneck …
- Redo tests:
  - Connected to development CephFS Cluster
    - (unfortunately new experimental cluster not available in time).
- CephFS allows for different stripe counts on directory / files
- Expected theoretical limit of 50Gb/s for this cluster
- Maximum achieved rate ~ 24 Gb/s
  - Testing from 200 – 1200 concurrent transfers



File Layout Tests

- Control the mapping of files to their RADOS objects (using xattrs)
- **stripe_unit**
  - determines how much data is written to each Ceph object across the OSDs
- **stripe_count**
  - how many units of data to write to consecutive OSDs
- **object_size**
  - how much data can be packed into each RADOS object

- Modified via setfattr (get with getfattr)

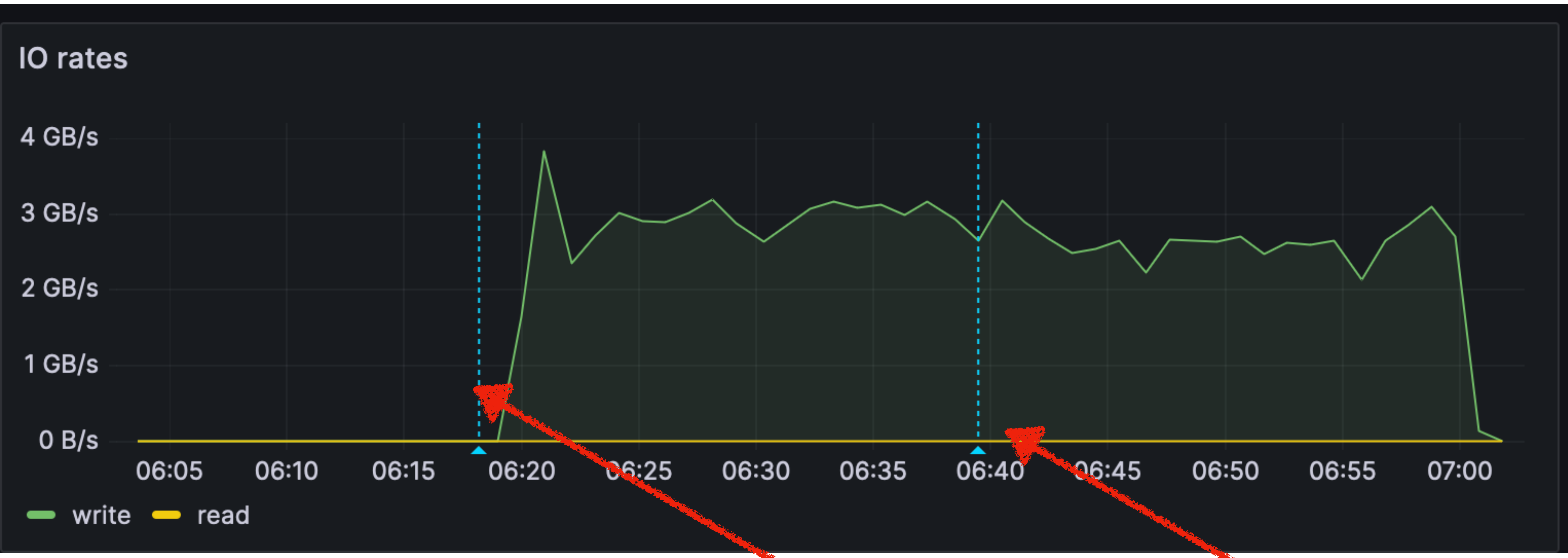
```
$ touch file
$ getfattr -n ceph.file.layout file
# file: file
ceph.file.layout="stripe_unit=4194304 stripe_count=1 object_size=4194304 pool=cephfs_data"
```

https://docs.ceph.com/en/reef/cephfs/file-layouts/

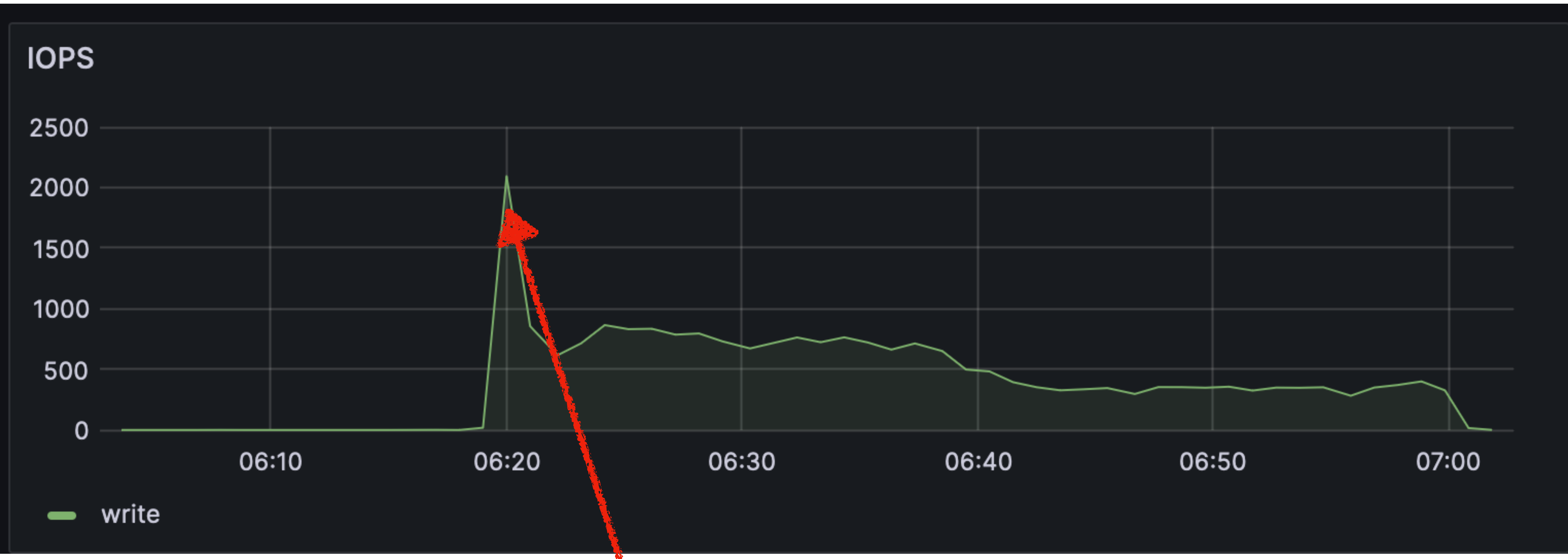
- Concentrate on cluster IOPS and IO rates

File layout tests (2)

Ceph monitoring IO Rate



Ceph monitoring IOPs



FTS Limit = 400 FTS Limit = 200 Initial FTS overshoot

- Some variation with FTS concurrency
- Max of 2.5k IOPs (5k theoretical max)
- Small overall change of write rate.

- Tests needed to compare read rates
  - And with random access patterns

Stripe unit	Stripe size	Stripe count	IORate [GiB/s]	IOPs
8	16	2	2.8	480
4	16	1	2.8	500
1	16	1	2.4	600
2	16	1	2.9	1500
1	16	2	3	2500

Looking Forwards: Cluster for UKSRC (SKA)

- UK is expected to hold ~ 200 PB of SKA data on disk for at full operations
  - Some mixture of POSIX and (presumptuously) Object Storage (S3?)
- Primary Usage:
  - Data ingest from the two telescope sites (~ 30 – 40 Gb/s)
    - Archive to Tape storage
  - Data replication to other SRCNet Nodes (~ 30 – 40 Gb/s)
  - Internal Data movement within UK (O(100) Gb/s (bursty))
  - User and Production workflows (TBC)
- Extrapolating from this study (dev cluster),
  - and based on 4PB (usable) Ceph cluster (being deployed)
  - Assume O(1) Tb/s (for headroom and internal Ops and Workflows)
  - By 24 PB (usable) should be approaching needed IOPS (in theory?)
- Delivering required output rate will be a future study

New Ceph cluster (4PB usable) for UKSRC

- 312 HDDs: ~60,000 IOPS
  - With 8+3 EC this corresponds to ~6000 client write ops per second.
- 13 nodes with 25+25Gb/s connectivity - 650 Gbps
  - Network overhead for EC shard redistribution is 1*(10/11), so theoretical throughput is ~50% of available bandwidth
  - Reduce factor 2 due to this study
  - So ~ 150 Gb/s

Summary

- 100Gb/s of throughput not established:

- Development CephFS cluster theoretical limit at 50Gb/s
    - Managed ~24Gb/s in practice
    - With new cluster O(300) OSDs need to repeat tests
- Decoupling the (destination) storage:
    - Achieved 75 Gb/s writes (to /dev/null type plugin) from disk-backed source site
    - Possible only with multiple 100s of concurrent transfers however
        - Testing with XrdBlackhole <-> XrdBlackhole would be interesting

- Future work:
    - Repeat and extend tests with new 4PB cluster (in commissioning)
    - Verify S3 vs POSIX vs 's3 via POSIX-fuse mount' performance speeds
    - For CephFS, tuning of MDS mapping to directories, etc may be critical.