GridPP
UK Computing for Particle Physics

Unified Storage Systems for Distributed Tier-2 Centres

Greig A. Cowan, Graeme A. Stewart, Andrew Elwell

University of Edinburgh & University of Glasgow

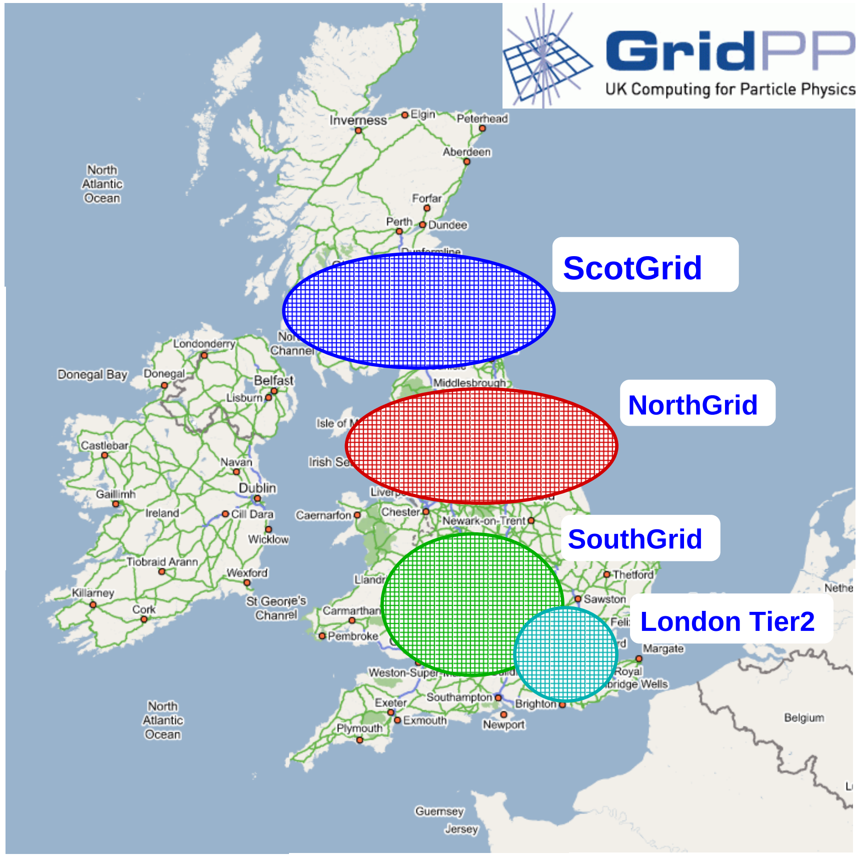
THE UNIVERSITY OF EDINBURGH

UNIVERSITY *of* GLASGOW

Overview

1. Using Tier-2 resources efficiently.
   - Accessing data across the WAN
2. Testing setup
3. Results
   (a) File open times, read times and rates
   (b) Server response
4. Future work
5. Summary

Distributed Tier-2s



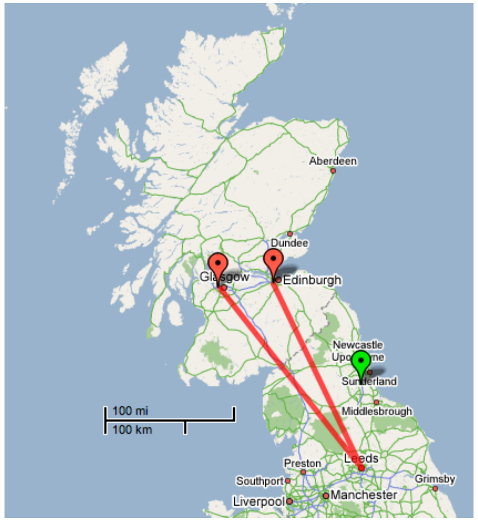
- GridPP organised into four regional Tier-2s.
  - Helps with deployment and operations.
  - Cross-site support.
- Can we do better on a technical level?
  - Can we pool resources to use them more efficiently?
    * **Storage at one site could be regarded as being "close" to CPU of another.**

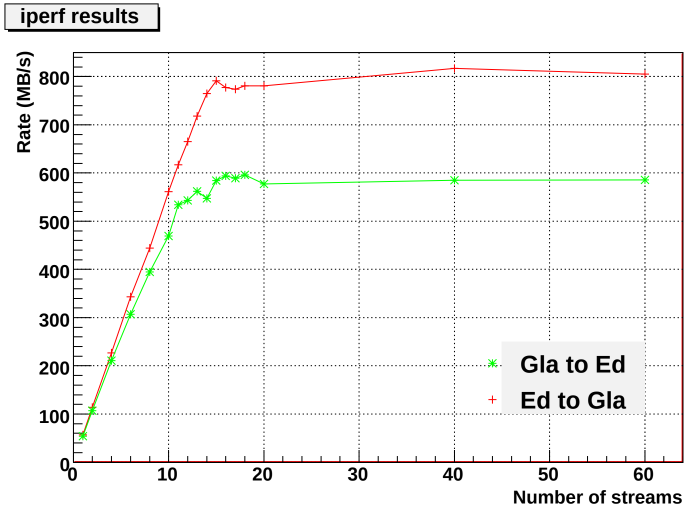
Access to data

- Currently, jobs are sent to the compute element which is local to the data.
  - Users running a selection algorithm over a dataset.
- Often more efficient for the jobs to process the data directly on the SE.
  - Use POSIX-like protocols (rather than copying entire file to the WN).
    - `rfio` for CASTOR and DPM (with gsi)
    - `(gsi)dcap` for dCache
    - ROOT provides `TGFALFile` to allow access to these SEs on the grid.
- Problem: if batch farm where data is located is full, then jobs cannot run.
  - Other sites in Tier-2 may have spare capacity.
    - **Inefficiency in system.**

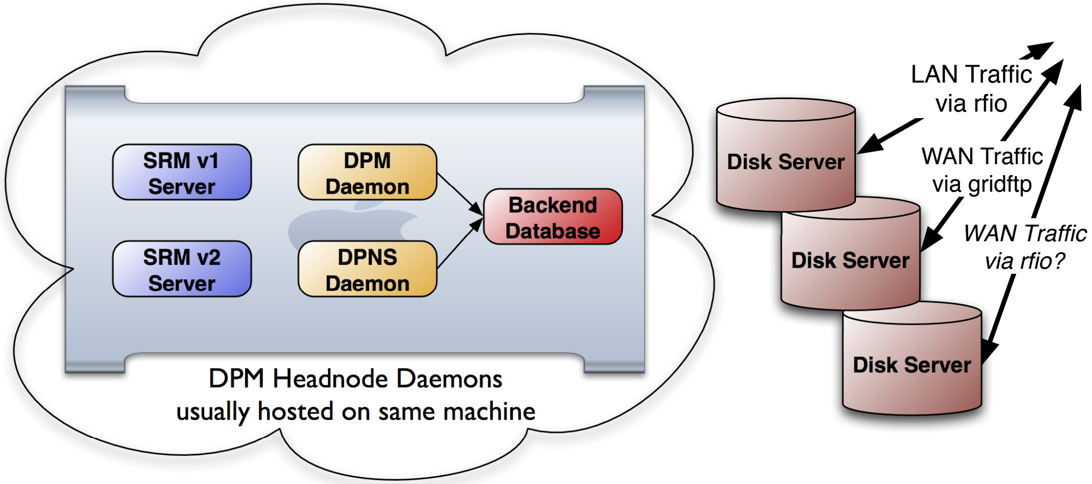
Access to data across the WAN

- Can we use the POSIX protocols to access storage across the **wide area network**?
  - Will this be transparent to users?, i.e.,
    - ∗ **Can they access data at the same rate?**
    - ∗ **Does the efficiency of their jobs remain the same?**
- Production JANET-UK network between UKI-SCOTGRID-GLASGOW and ScotGrid-Edinburgh. RTT ~ 12s.





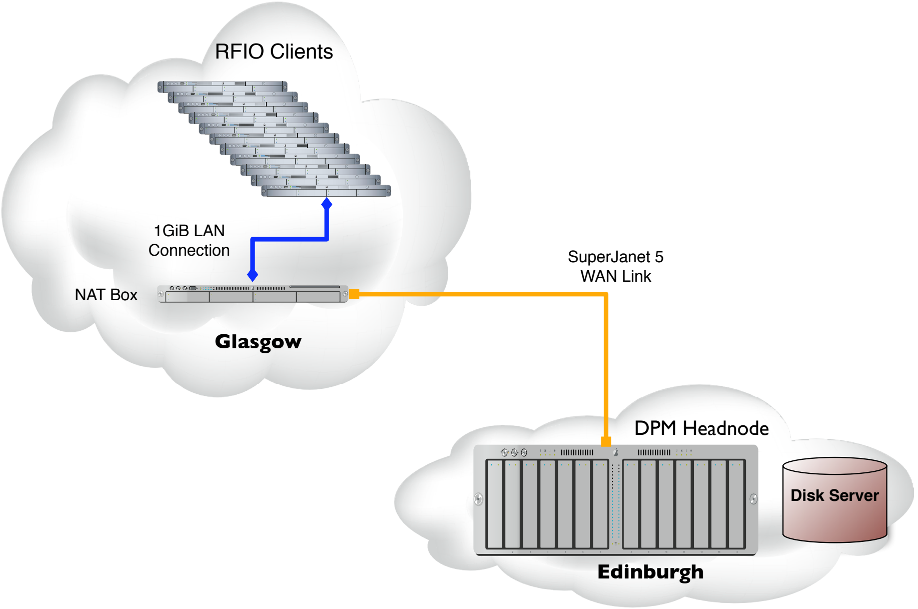
Tier-2 storage



- DPM developed by EGEE as a lightweight solution for disk storage management at Tier-2 institutes.
- See other talks/posters at CHEP07 for further details.

Testing method

- We wrote our own RFIO client application.
  - Reading data appears to be the main use case.
  - Configurable to meet needs of our study, i.e.,
    - **RFIO mode**
    - **read block size**
    - **reading pattern (sequential, skipping, random)**
    - **Allows us to stress the SE.**
  - "Skipping" means that we read a block of data, then skip ahead M blocks and read again, until EOF.
- Seed client onto N nodes and simultaneously start reading 1GB files from ScotGRID-Edinburgh DPM.



RFIO protocol

- GSI-enabled protocol which allows POSIX file operations, permitting byte-level access to files.
  - clients require a X.509 Grid certificate signed by a trusted CA.
  - can use RFIO over the wide area network.
  - Ports must be opened in site firewall.
- RFIO library allows the client to choose from four modes of operation (see `rfiosetopt()` man page):

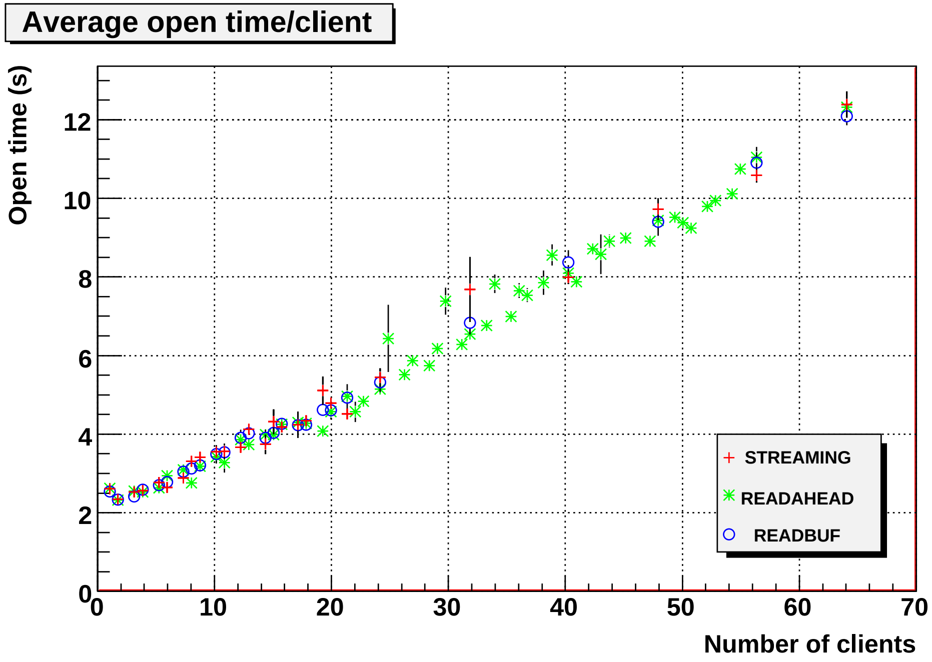
1. `NORMAL`: one call per read.
2. `RFIO_READBUF`: fills internal buffer to service requests.
3. `RFIO_READAHEAD`: uses internal buffer and reads until EOF.
4. `RFIO_STREAM`: separate TCP streams for control and data.

RESULTS

File open times

Sequential reading

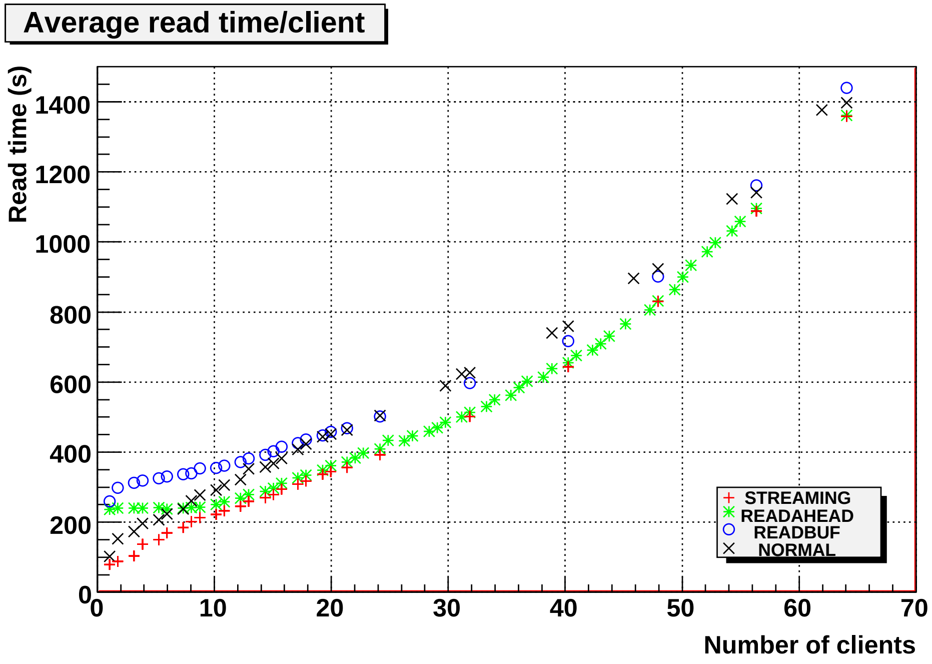
- Linear increase in the open time with client number.
- Large number of clients can increase open times up to > 12s.



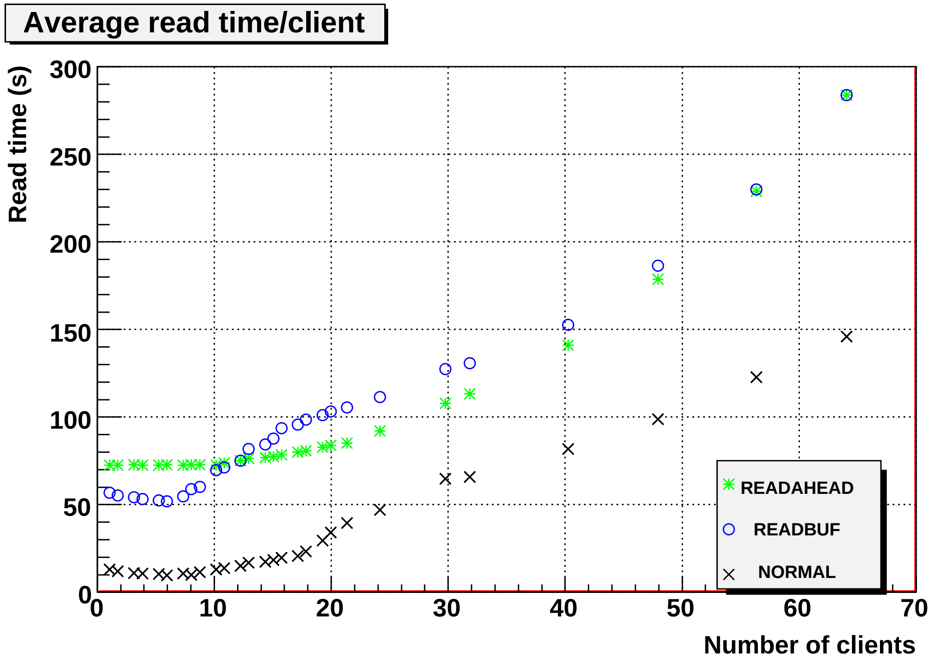
GridPP
UK Computing for Particle Physics

Read times

Sequential reading



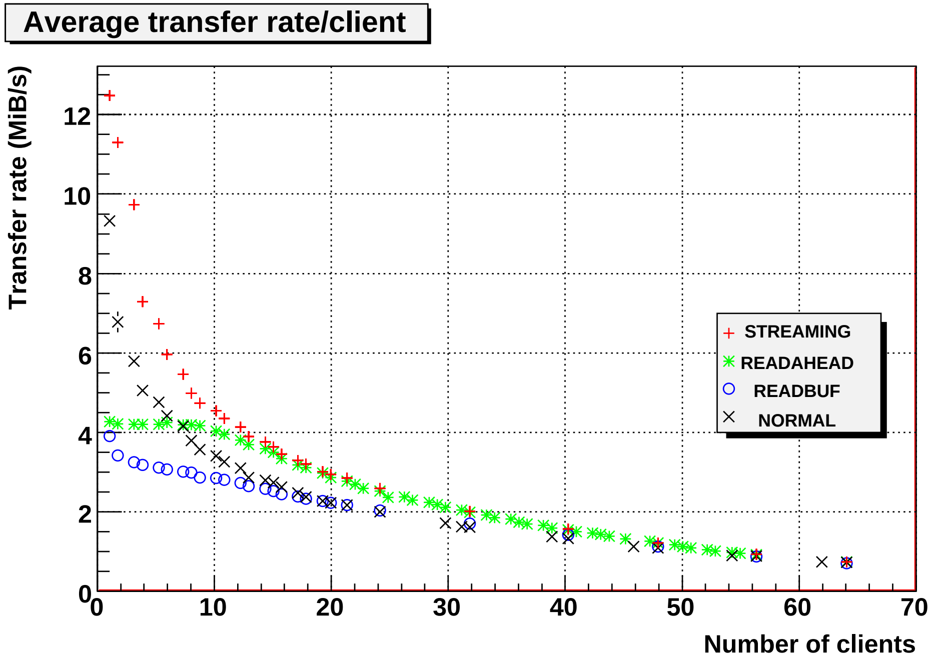
Skipping through 10% of the file



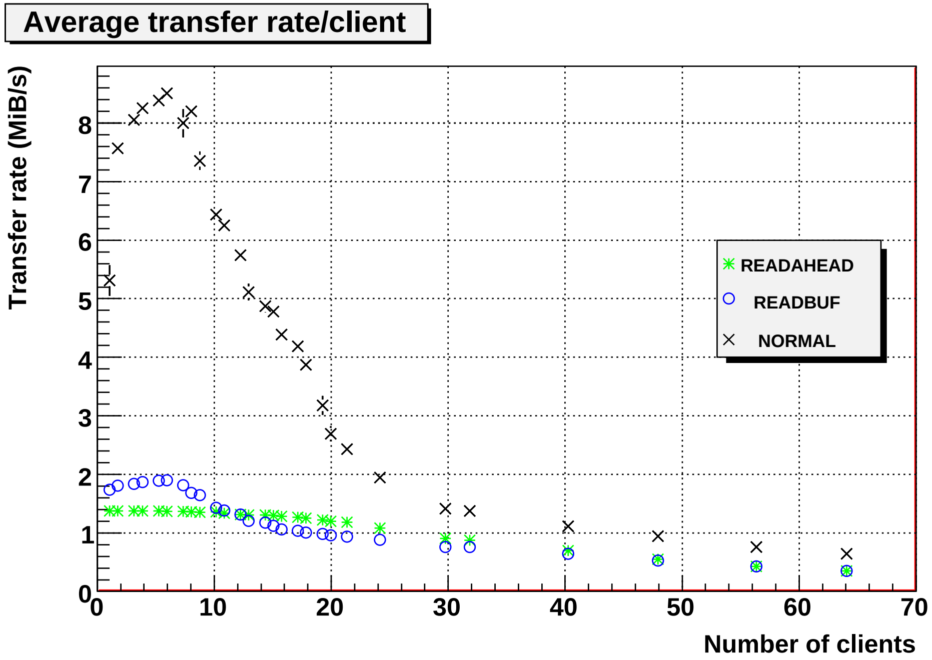
- LHS: `STREAMING` comes out on top for small number of clients. Not much difference for large number.
- RHS: `NORMAL` mode leads to optimal access.

Read rates

Sequential reading

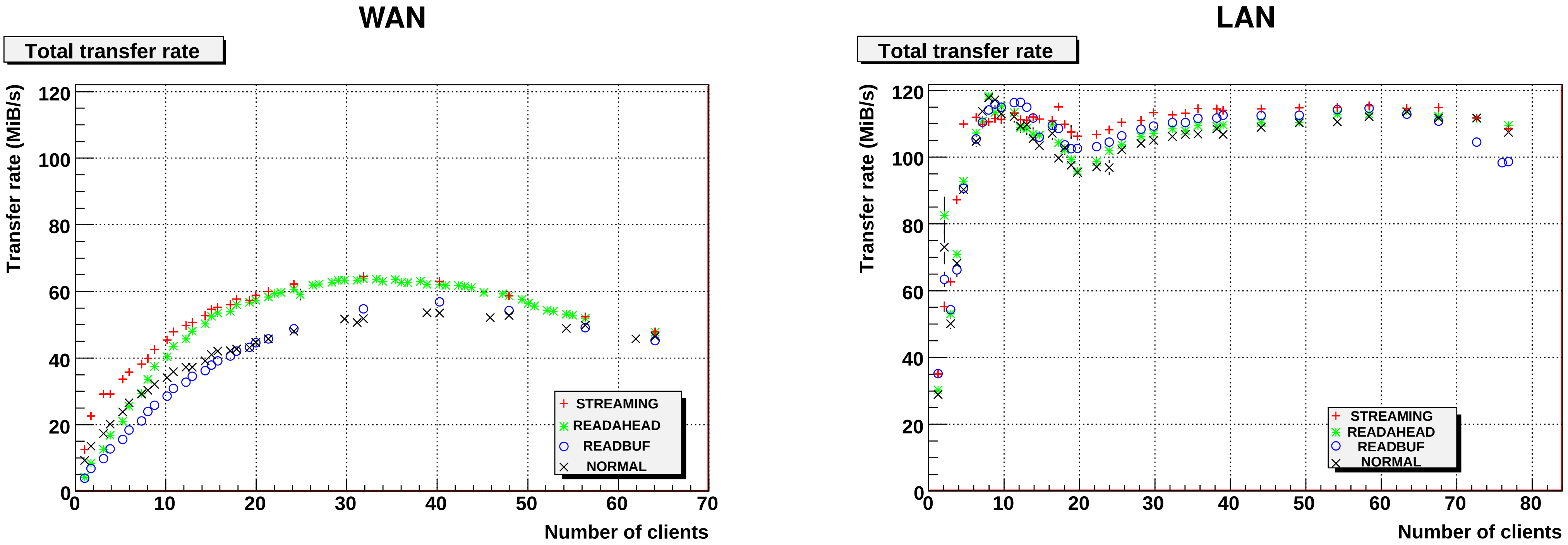


Skipping through 10% of the file



- Large number of clients, rates down to ∼1MiB/s per job (NB single DPM server).
  - **ATLAS software expects rates of O(10)MiB/s per job in 2008.**

Comparison with LAN access



- Peak total rate across WAN ∼65MiB/s.
  - Contention on the network. Max expected - 100MiB/s.

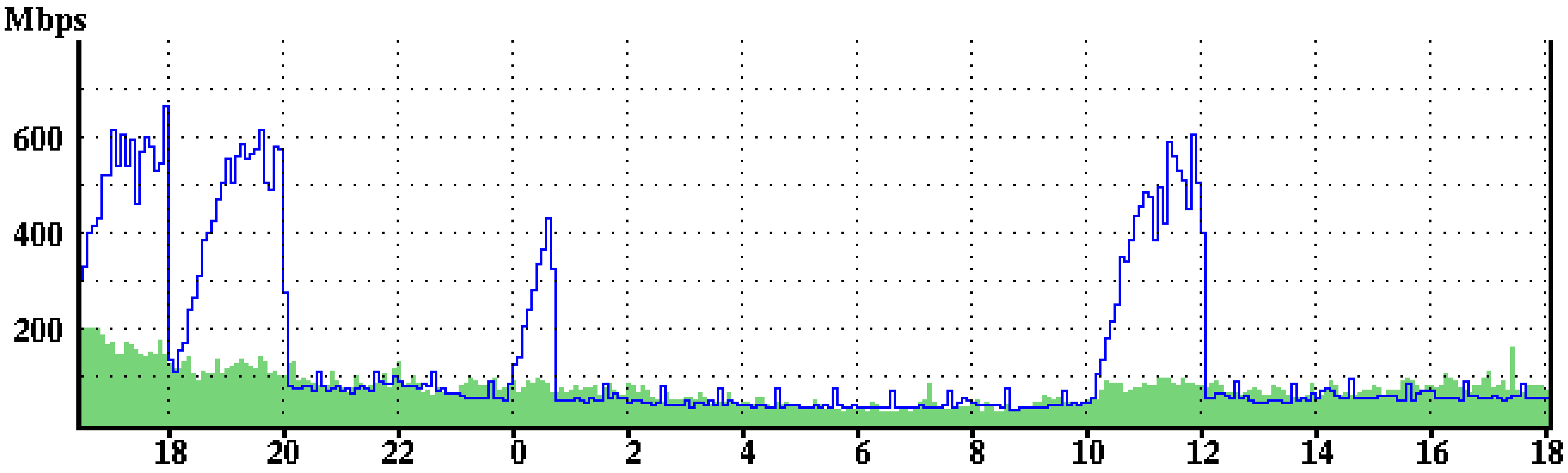
- Peak total rate across LAN ∼110MiB/s.
  - Single server. Dedicated bendwidth.

- Becomes **IO-bound** at a large number of clients, rate begins to decrease.

GridPP

UK Computing for Particle Physics

Impact on network



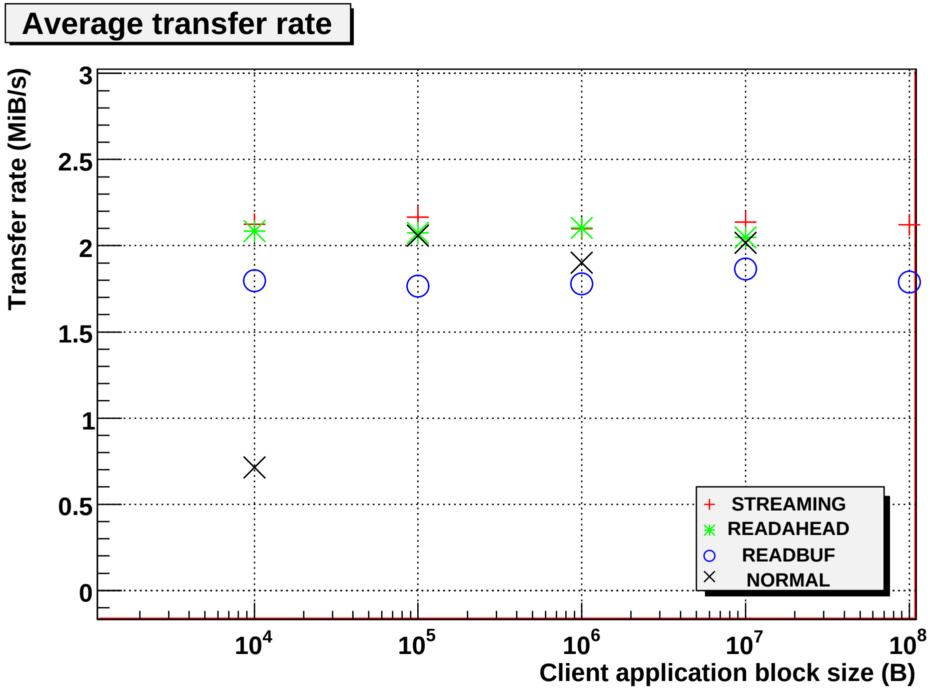
- Background traffic on the production network $<$100Mbps.

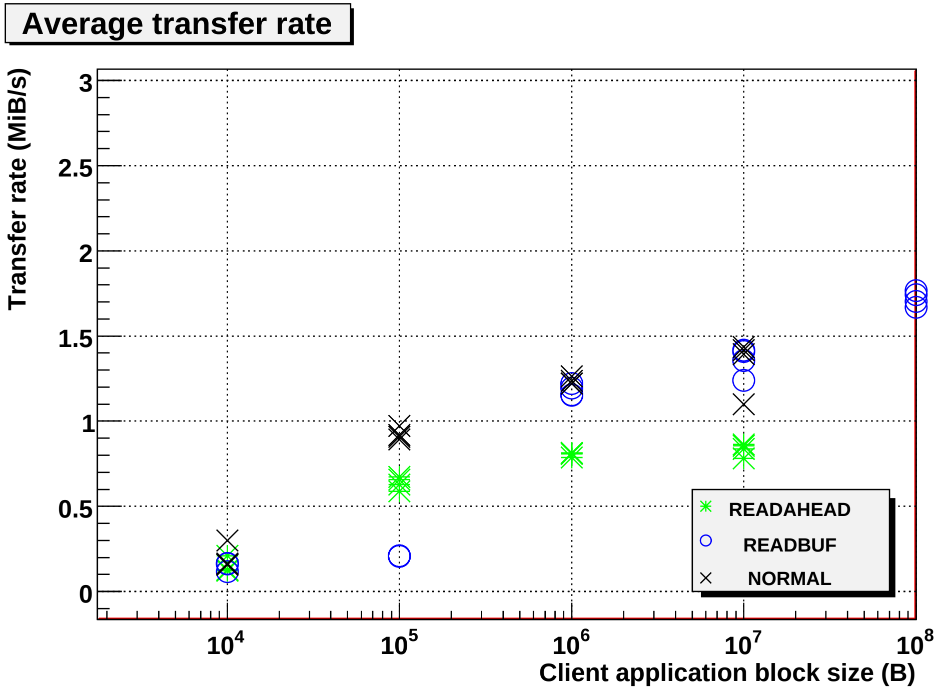
GridPP
UK Computing for Particle Physics

Application block size

Sequential reading

Skipping through 10% of the file





- No change when sequentially reading the file.
- When skipping through the file, higher rates achieved with larger block sizes, particularly for READBUF mode.

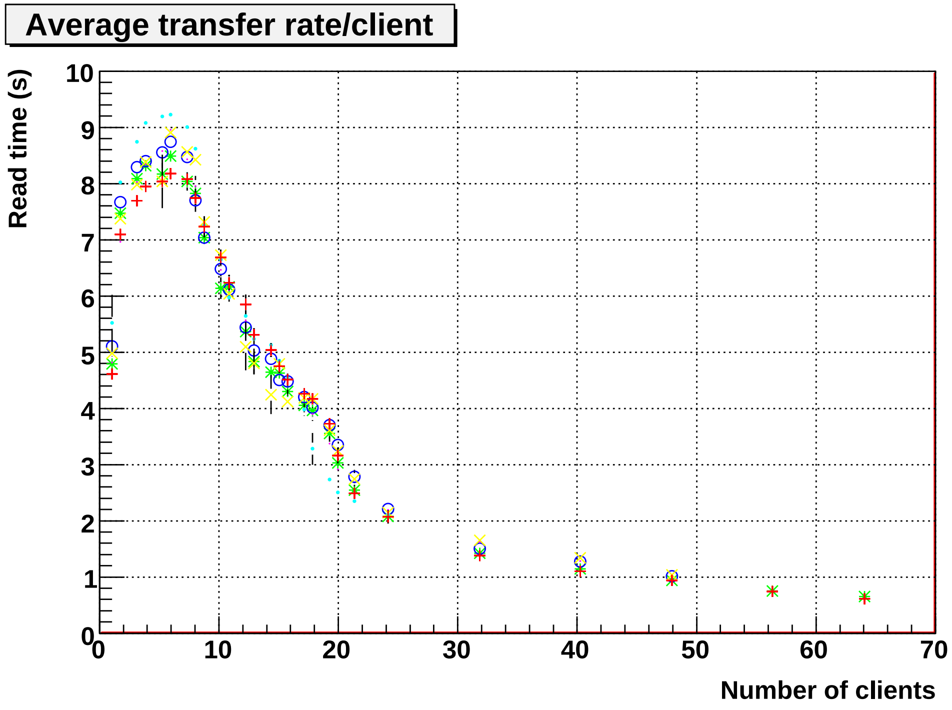
Client TCP parameters

- Since we are moving data across the WAN, TCP kernel parameters **could** have a impact on the data throughput.
- Initial work looked at increasing the maximum TCP window size.
- e.g., in `/etc/sysctl.conf` we varied parameters such as, `net.ipv4.tcp_rmem` and `net.core.rmem_max`.
- Looked at increasing window sizes from 0.5MB up to 16MB.

GridPP
UK Computing for Particle Physics

Variation with client TCP parameters

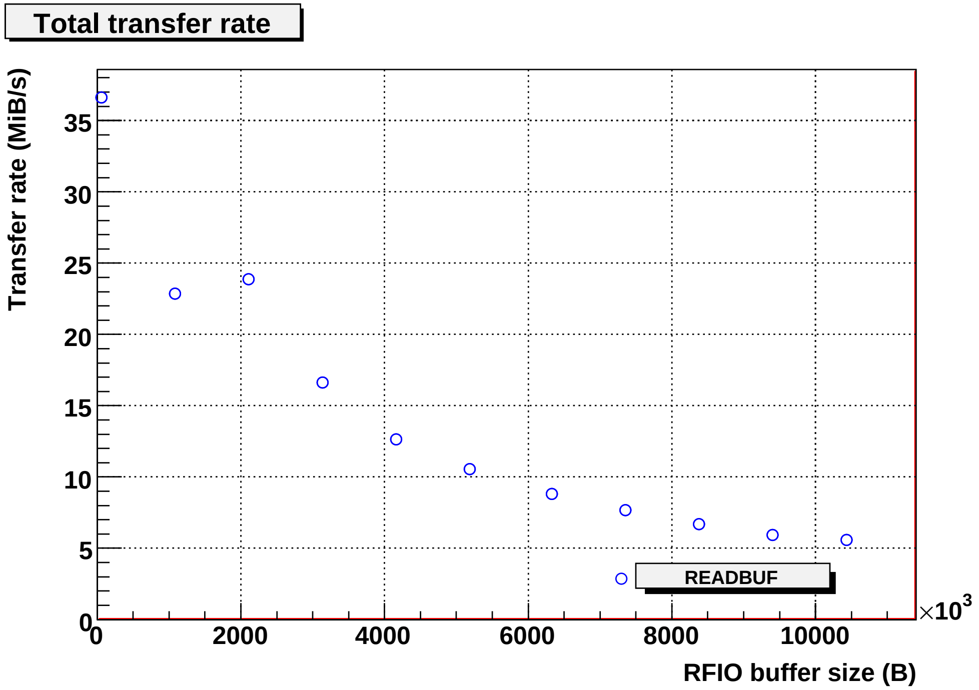
- Different colours correspond to different TCP window sizes.
- Very little difference.
  - Probably expected when such a large number of clients are simultaneously reading data.
  - Slight improvement at small client numbers with a larger window.
- Application optimisations probably required before tuning the networking parameters.



Transfer rate vs. RFIO buffersize

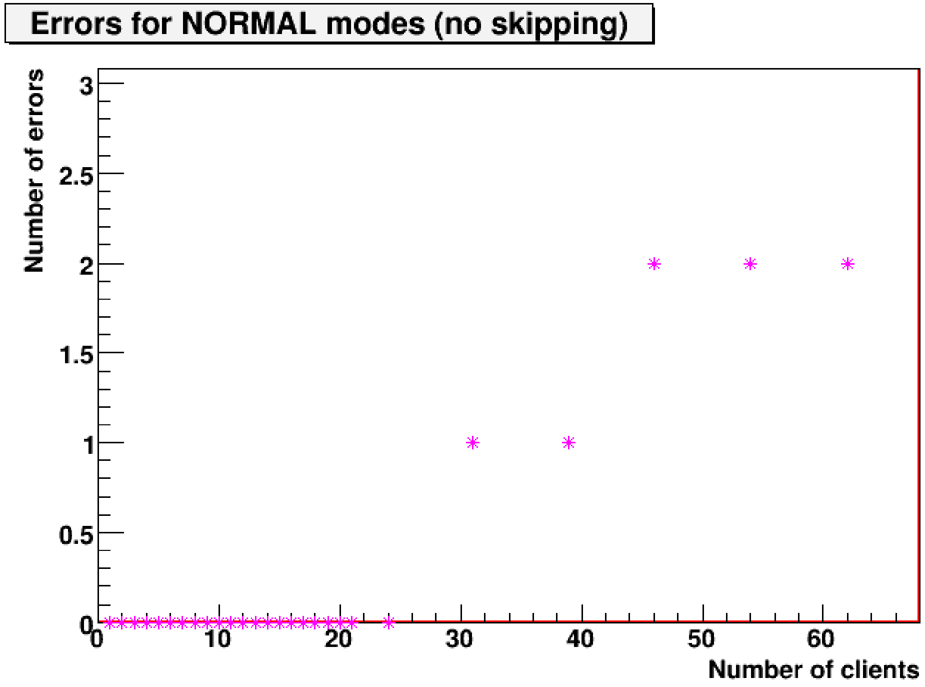
- RFIO READBUF mode uses a fixed size client side buffer for data transfer.
  - Parameter is `RFIO IOBUFSIZE` in `/etc/shift.conf`.
  - Can we see any dependence on the size of the buffer?
  - Plot shows that for a constant block size of 1MB, increasing the RFIO buffer leads to a reduced total transfer rate.

Skipping through 10% of the file

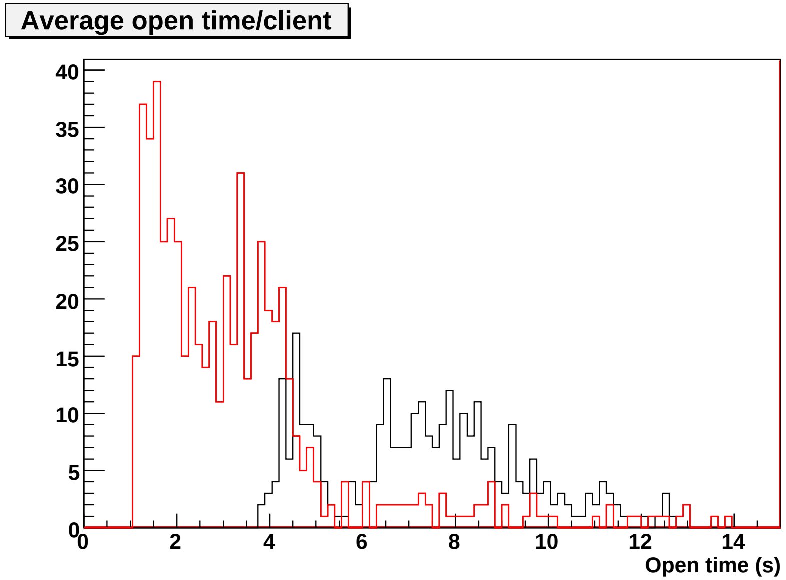


File access errors

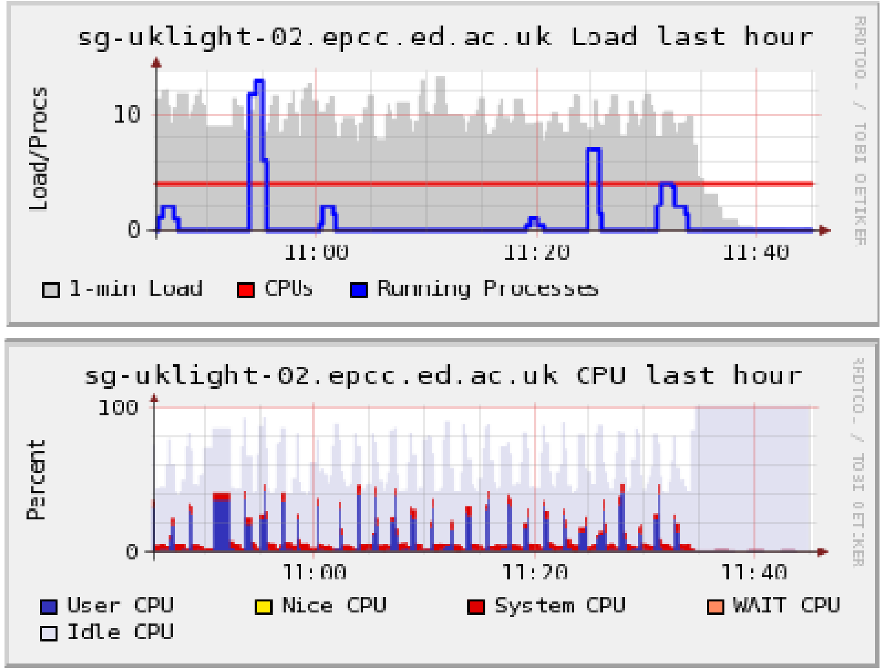
- Server performance degrades slightly when many clients simultaneously attempt to open files.
  - We are intentionally stressing the system.
- Substantial improvement over versions of DPM $<$ 1.6.5, which could not support more than $\sim$40 opens per second.



Server load



Difference in the open time for < 20 clients (red) and ≥ 20 clients (black).



Load on single DPM server when 30 clients are simultaneously reading using NORMAL RFIO mode.

Future Work

- Would like to repeat tests using **lightpath** as this gives:
  - Dedicated bandwidth that will not impact on other users.
  - Smaller RTT, of order 2ms.
- Alternative data access patterns. Extreme cases.
  - i.e., Use 1 client to open 1000 files on the SE and then send them to `sleep(3000)`.
- Run some real analysis jobs.
  - ROOT `TTreeCache` will allow efficient data access across WAN.
    - *See talk 284 at CHEP07*
- Create a single DPM that spans both Glasgow and Edinburgh sites.

Summary

- Using DPM and RFIO, our study has shown that it is possible to access storage across the WAN.
- This opens up possibilities for optimising storage and CPU usage within ditributed Tier-2s.
- Principle could be more widely applied to the Grid.
  - Rather than having many replicas of files spread over the Grid, closely linked sites could access a replica within their geographical region.
    - **ATLAS already have a cloud model for data management...**
- Studied the ideal cases where clients were behaving as expected.
  - How does the system respond in non-optimal cases?
- We saw good utilisation (60%) of the production network, but these rates may not be sufficient when large numbers of clients are running.
  - Need to investigate the potential of dedicated lightpaths.