Exploring performance and scalability using generic IT infrastructure

EBU Media Storage Workshop

21 November 2011

David Butler
BBC Research & Development
david.butler@rd.bbc.co.uk

BBC R&D

Media Storage Workshop 21 - 22 Nov 2011 / Geneva (CH)

Broadcasters are experiencing an enormous pressure to scale up their media storage systems.

Why is it crucial to address the challenges linked to storage systems? What are the specific needs of rich media storage through the whole lifecycle? To which extent is performance measurement in storage and scalability of generic IT infrastructure important?

This workshop will debate those questions and it will set the direction to the new EBU project group on Future Media Storage Systems. Do you want the EBU to address the highest priority challenges you are facing in this field? Then come and share your concerns with us and take part in the solution!

http://tech.ebu.ch/events/media_storage_workshop11

Introduction

- Network Storage
  - What is high performance network storage
  - What should we measure
  - Media Storage Meter open source test tool
  - Difference between cache and disk performance
  - Caching effects
  - Flow control and jumbo frames
  - File fragmentation and scattering
  - Disk type
  - Switch performance
  - Switch connection capability
  - 10 GbE or 1 GbE connected storage
  - Single threaded and multi-threaded access
  - Client behaviour
- Results shown are from real measurements using Media Storage Meter

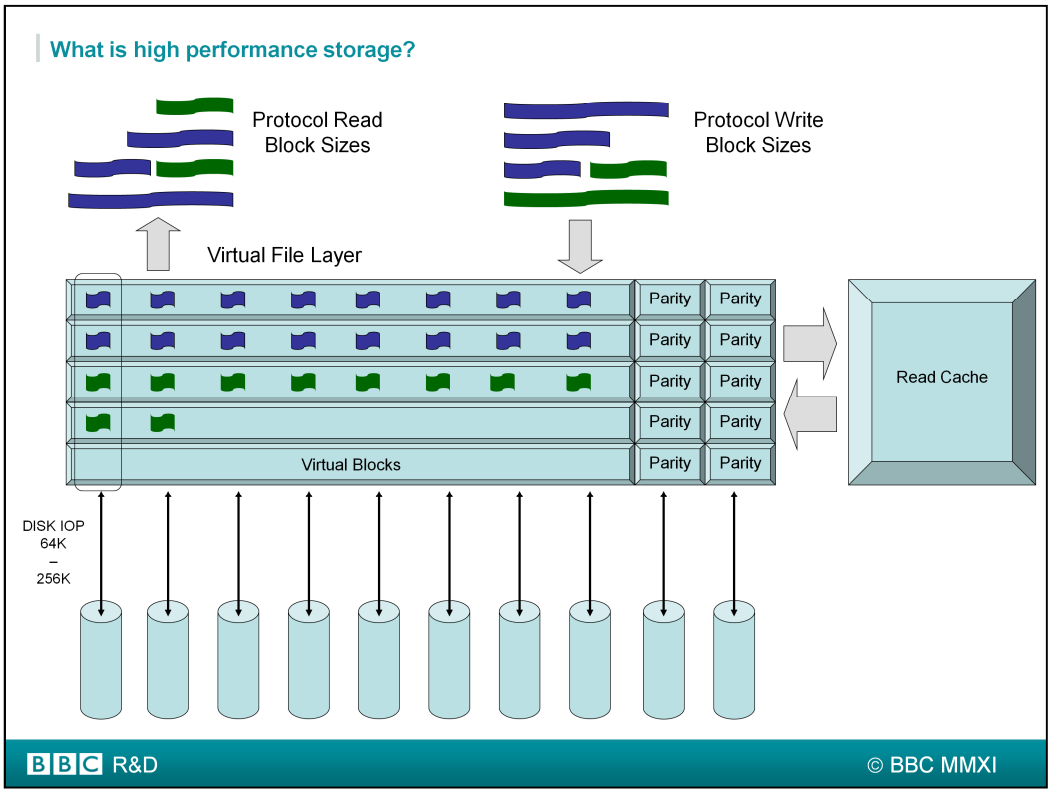
BBC R&D

This presentation provides some results from the testing of high performance network storage, using an open source test tool called Media Storage Meter.

It examines issues that affect performance and how the storage would scale for a production environment.

Both storage and network issues were investigated.

Topics to be discussed are:



Fundamentally storage is limited by the number of disks and disk speed. Spreading files across multiple disks, parallelises disk access, increasing read and write speeds.

Manufacturers of high performance storage use various techniques to improve access speed:

Virtual file access layer
- Virtual 4K block sizes, independent of the disk block size.
- Grouped data blocks and writes to disk in a sequential stream.

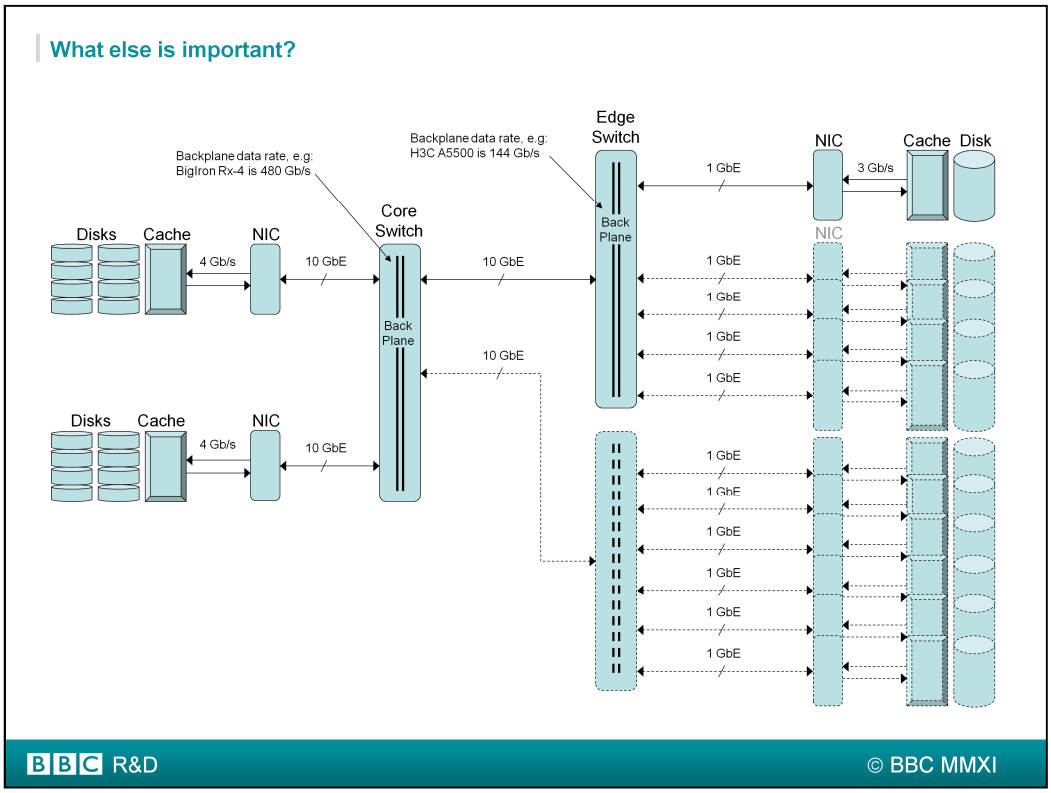
Virtual volumes
- Volumes are allocated from the total aggregate of disks.

Strong Error Correction
- Maintain performance and reliability when disk errors occur

Large intelligent read cache
- Improving access speed by caching data blocks

Intelligent algorithms for disk usage
- Reduce file fragmentation and maintain disk performance.

Also other features for fast back up and recovery. Storage performance is generally optimised for the application. Storage optimised for data base access would perform poorly for media and vice versa.



With increased storage performance, the network and infrastructure become more important.

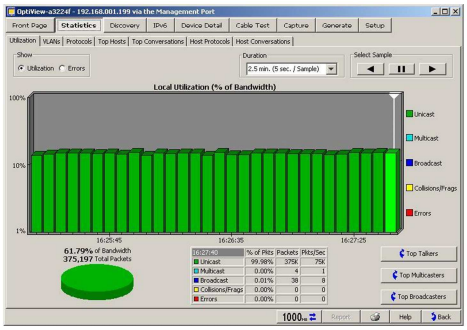
Network storage performance is determined by storage performance and the data path through which everything is moved.

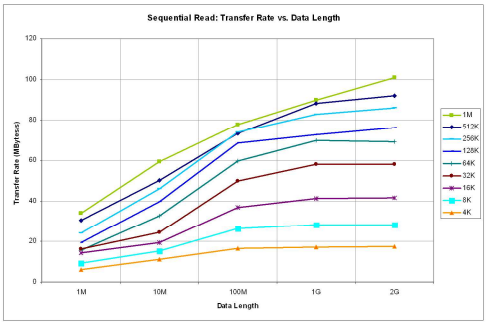
This diagram shows some of the data speeds in a simple storage network, showing typical client bus, network, backplane and storage bus speeds.

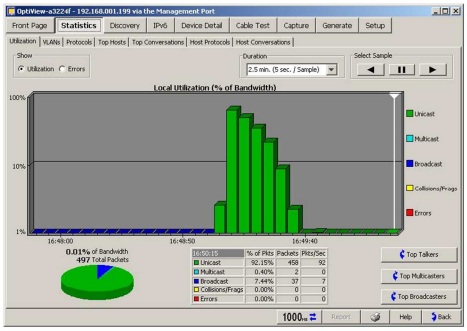
Each point in the data path can be a potential bottle neck, but the bottle neck will move with load and access profile.

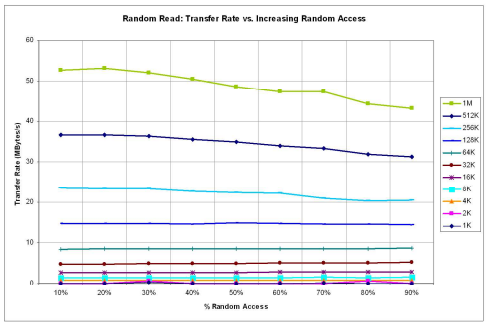
This is examined later in the presentation.

What to measure?









BBC R&D © BBC MMXI

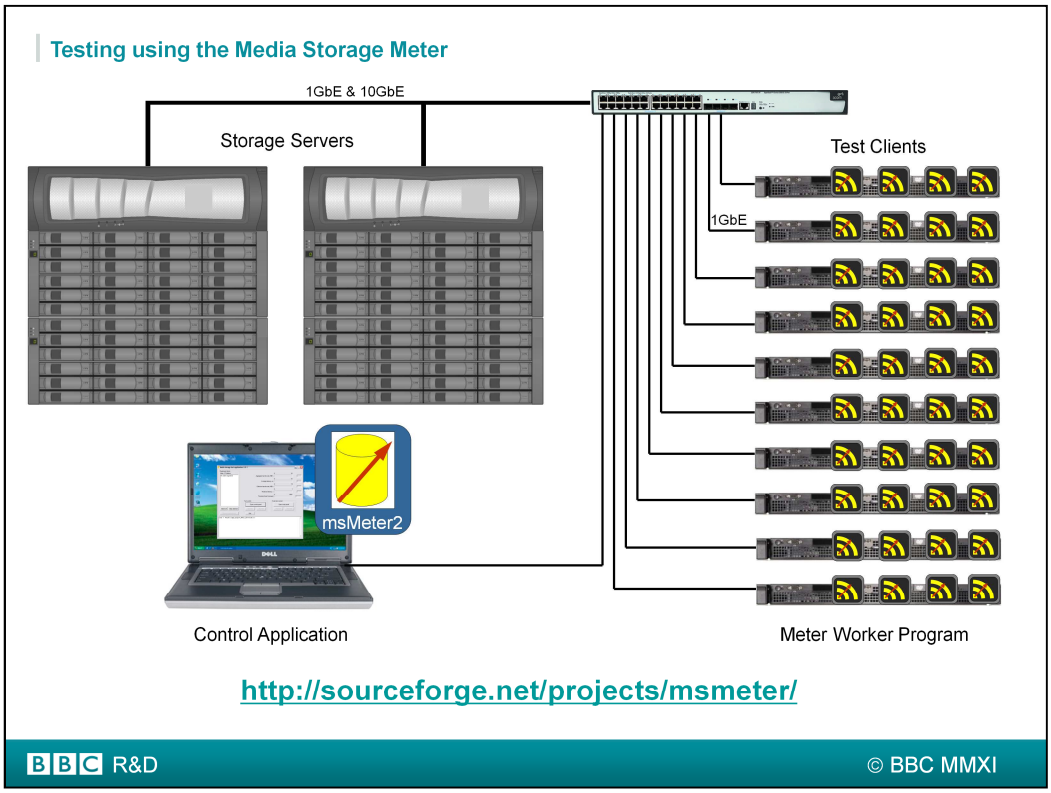
What should we measure to try and bench mark a storage system?

Typically performance is characterised by transfer rate and latency, where data is read and written in blocks:

- Increasing the block size increases the transfer rate, but also increases latency.
- Increasing the number of blocks to transfer, also increases the transfer rate.
- Increasing the mix of random access requests will reduce the transfer rate.

Some file systems also employ read ahead and local caching, which attempts to transfer the entire file into local client memory. This introduces very peaky network behaviour, with very high initial network utilisation. Disabling this results in lower, but more consistent network utilisation.

Results are very dependant on client behaviour and how data is read to and written from the storage.



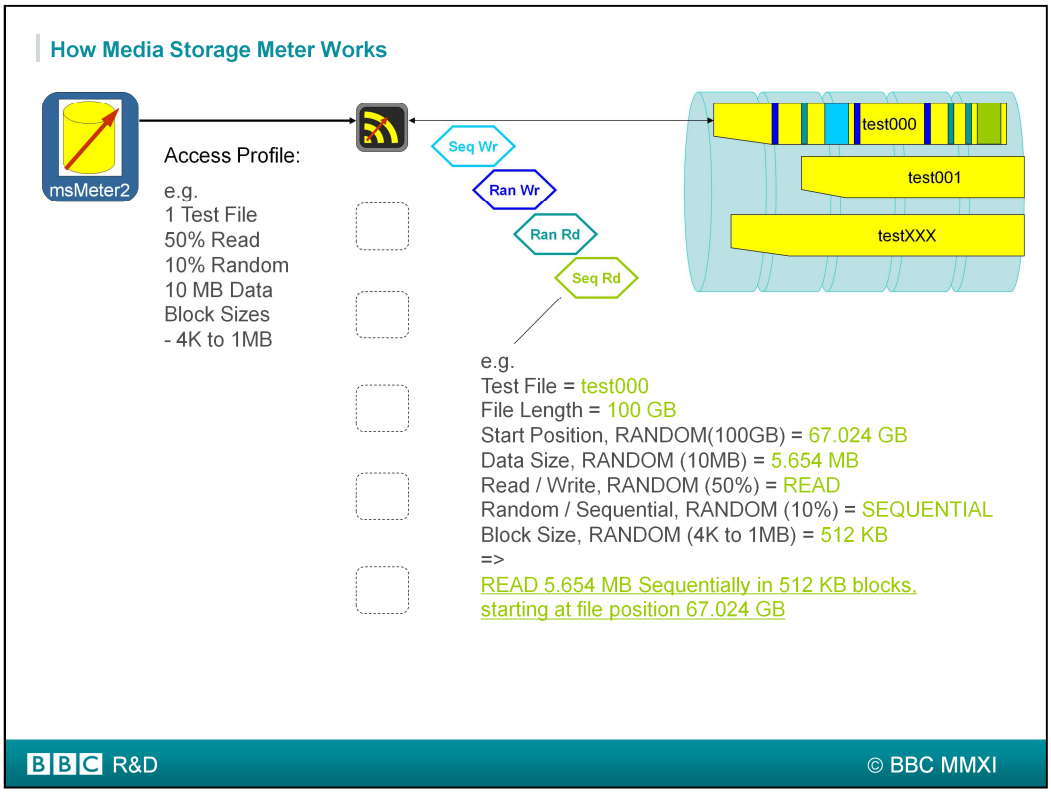
In 2004, the BBC developed and open sourced its open source test tool, called Media Storage Meter. This was updated in 2011 to take into account improvements in operating systems and file systems, such as 64 bit files and large block sizes.

Media storage meter consists of a control application, that runs on Windows, and worker programs that can run on Linux or Windows OS.

The control application controls multiple worker programs running on multiple clients, to emulate the behaviour of production tools.

Specific access profiles are configured for each client, creating read and write access behaviour similar to that seen for video editing or other production processes.

This diagram also shows the test setup used for the results, with the storage and clients connected to a single switch. This made it easier to differentiate between network and storage limitations. Two different switch types were tested.



The control application configures each worker program. For these tests, the access profile is:

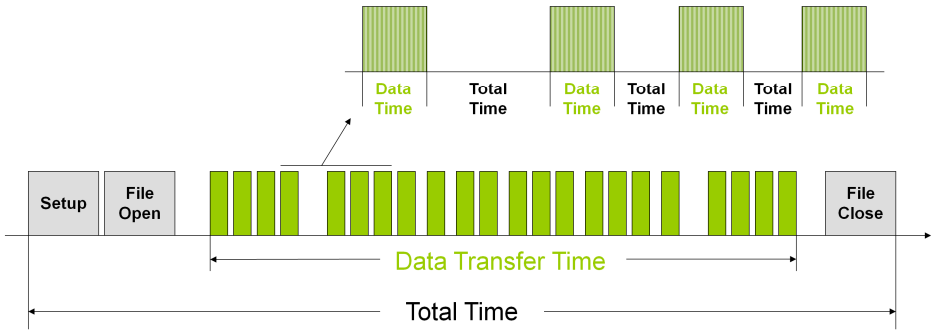
50% read, 50% write.
10% random, 90% sequential.
Up to 10MByte of data for each access.
Block sizes of 4KB to 1MB are used.

This profile emulates what is expected for real production editing. An equal balance of read and write, mostly sequential for the media file, with some random access for metadata.

Each client access the storage based on the profile:
Randomly selects a file start position.
Randomly selects a data size of up to 10MB.
Randomly selects read or write, weighted by 50%.
Randomly selects a block size from 4K to 1MB.

The client continuously selects and performs the read or write operations until paused or stopped.

How Storage Meter Calculates Transfer Rate & Latency



Individual Transfer Rate = $\frac{\text{Number of blocks * Block Size}}{\text{Data Transfer Time}}$

Individual Latency Time = Total Time - Data Transfer Time

BBC R&D © BBC MMXI

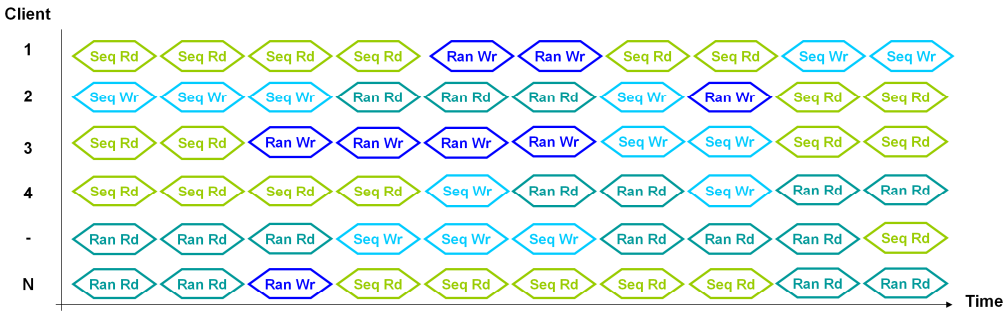
For each transfer operation, msMeter records the overall time and data transfer time.

Data transfer time measures the time taken for block transfers only, with the time between block transfers counted as part of the total time.

The transfer rate is calculated using the data transfer time and the latency is calculated as the difference between the total and transfer times.

This is to differentiate between delays in data block transfer, which form part of the transfer rate and other network storage response latencies.

Scaling with Multiple Media Storage Meter Clients



- **Results do not show instantaneous transfer rate over wire.**
- **Results are averaged over entire measurement by type.**
- Average transfer rate, average latency & maximum latency
  - Sequential Read, Random Read, Sequential Write, Random Write

BBC R&D © BBC MMXI

Unlike other open source tools, msMeter is not reporting the instantaneous transfer rates over the wire.

msMeter reports average transfer rate, average latency and maximum latency for each transfer type; the sequential read, random read, sequential write and random write.

We are more interested in how the storage scales with increasing number of clients, so the results are averaged over the whole measurement.

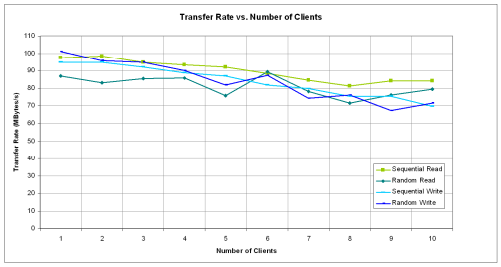
All reads and all writes are not occurring simultaneously, so higher transfer rates per type can achieved than for continuous reads or writes.

Measurements are made while ramping up the number of clients, to show how the transfer rate per client changes.

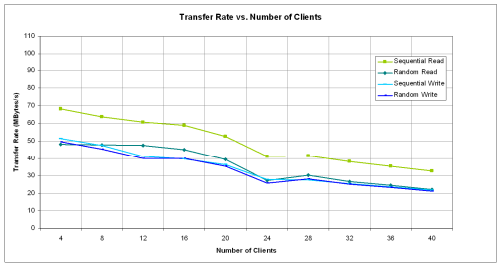
To make the graphs easier to read, only the results for 1MByte block sizes are included in the presentation.

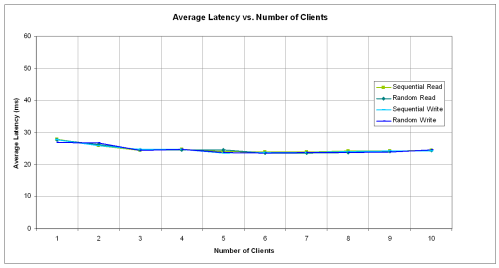
Measuring the cache or disk performance?

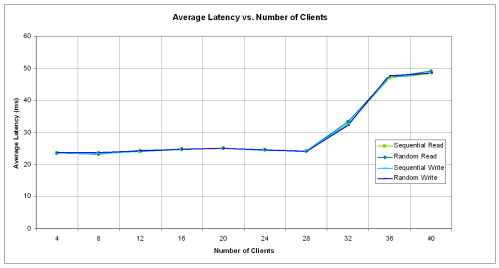
Singe File Access vs. 1 to 10 Users (10GbE)
- 1 x 100GB File, 50% Write, 10% Random, 10MB Data



Multi File Access vs. 4 to 40 Users (10GbE)
- 40 x 100GB File, 50% Write, 10% Random, 10MB Data







BBC R&D © BBC MMXI

In these tests, storage performance largely depends on whether blocks are written and read from the cache or directly to and from the disks.

If clients are accessing a single file, smaller than cache, blocks are mostly read from and written to the cache, giving a high transfer rate.

If clients are accessing multiple files, greatly exceeding the cache, blocks are mostly read from and written to the disks, giving a lower transfer rate.

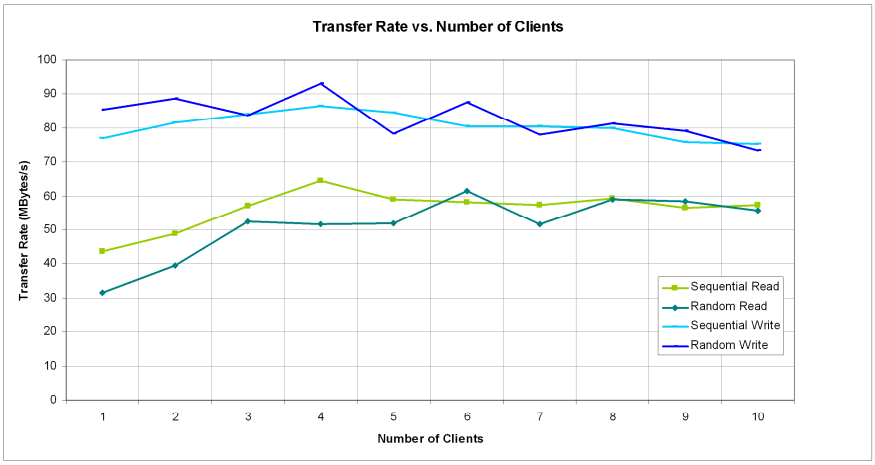
Cache and disk performance both need to be considered when specifying storage, as production staff are likely editing multiple videos simultaneously, often in split or quad screens.

For up to 28 clients, the average latency results are very similar. However for multiple clients and files, with more than 28 clients, the average latency increases with number of clients.

The increasing latency suggests either the maximum performance of the disk IOPs has been reached, or another factor is limiting performance. This is discussed later in the presentation.

What was happening before is important!

Singe File Access vs. 1 to 10 Users (10GbE)
- 1 x 100GB File, 50% Write, 10% Random, 10MB Data



BBC R&D © BBC MMXI

Caching only improves read performance if the required data blocks are already in the cache.

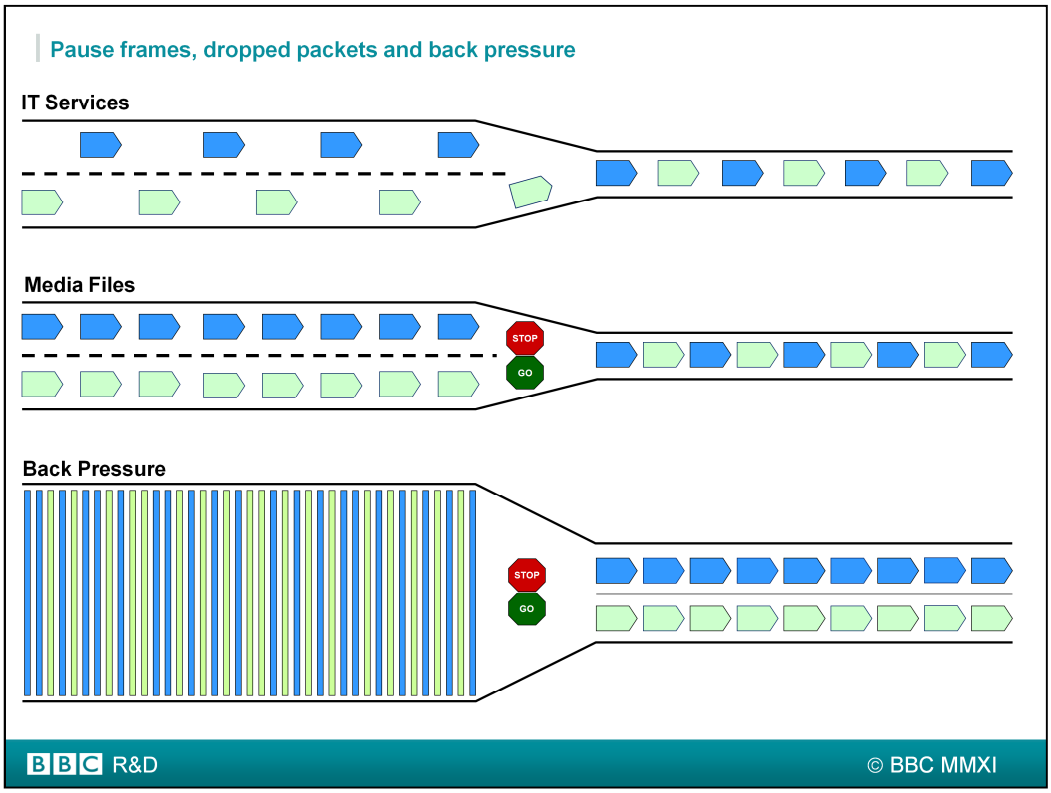
During the test, the numbers of clients increase in 1 hour steps.

Initially, there are no blocks from the test file in the cache, and the read transfer rate is slow. Over time, the storage cache becomes populated with blocks from the test file and the transfer rate increases.

The transfer rate levels of due to opposing performance effects. Caching increases performance, but increasing demand from more and more clients reduces performance.

Write data is cached, for future reads, but this does not effect how fast data is written to disk.

In real production environment, if a production team start editing a new project, the production staff could see a sudden drop in read performance..



For media files there is much more data than for normal IT services.

For IT services, there are fewer packets, so it is relatively easy to multiplex different packets onto a connection.

When accessing media files, there are many more packets, making it harder to multiplex the packets.

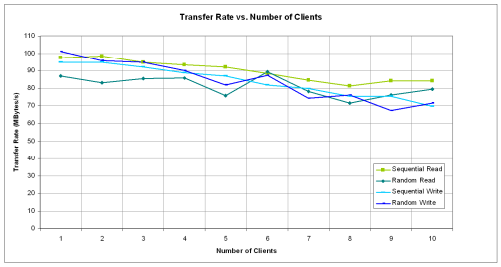
With heavy network utilisation, packets must be dropped or paused. Dropping packets cause further congestion, as packets are re-transmitted.

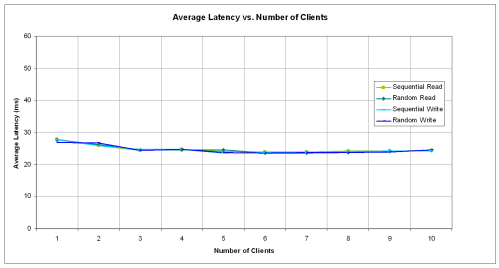
Enabling flow control on a connection allows a switch to send pause commands to the storage, preventing congestion and dropped packets.

Another problem for media files is back pressure, which occurs when data moves from a faster network connection to slower network connections. Data arrives on the high speed connection faster than it can be directed and delivered on the lower speed connections.

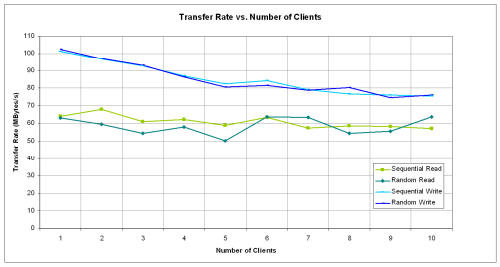
Flow control or no flow control?

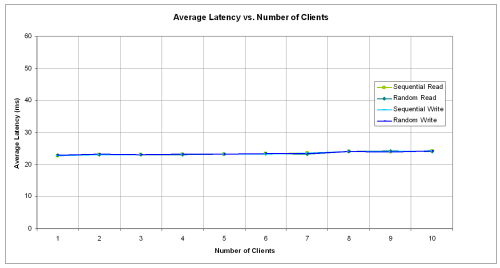
10 GbE Singe File Access with Flow Control
- 1 x 100GB File, 50% Write, 10% Random, 10MB Data





10 GbE Singe File Access no Flow Control
- 1 x 100GB File, 50% Write, 10% Random, 10MB Data





BBC R&D © BBC MMXI

In the tests, data arrives on the 10G connection faster than it can be directed and delivered on multiple 1G connections.

With flow control, transmission is paused to maintain the transfer rate. The read and write speeds are similar, both with a high transfer rate.

Without flow control, packets are dropped to cope with the congestion. This requires re-transmission of data packets, reducing the transfer rate.

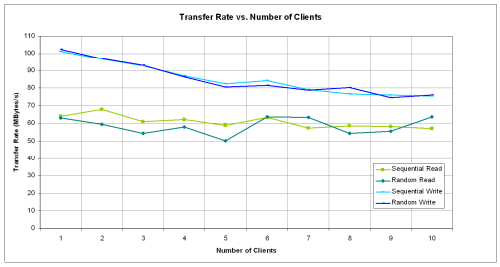
The read transfer rate is considerably lower than with flow control.

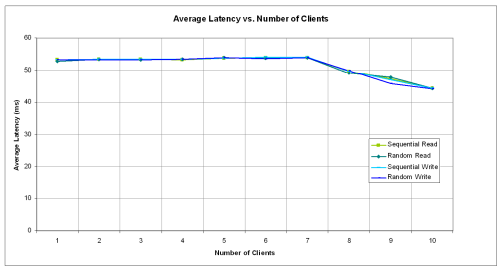
The write transfer rates are very similar, as backpressure only occurs when transitioning from a high speed connection to low speed connections.

The average latency results, with and without flow control, are very similar. The test tool measures the latency as delays that are not part of the actual data transmission. Packet loss, re-transmission and pause frames are all part of the data transmission process and therefore affect transfer rate and are not counted as part of the latency measurements.

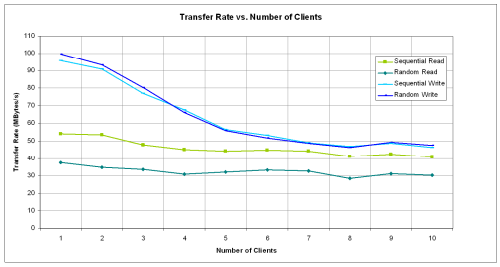
With or without jumbo frames?

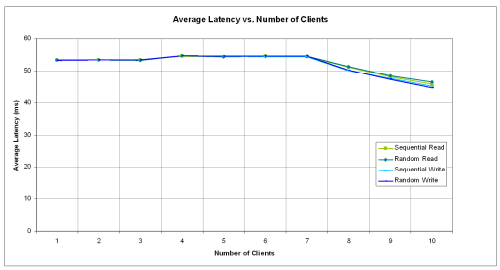
10 GbE Singe File Access no Jumbo Frames
- 1 x 100GB File, 50% Write, 10% Random, 10MB Data





10 GbE Singe File Access with Jumbo Frames
- 1 x 100GB File, 50% Write, 10% Random, 10MB Data





BBC R&D © BBC MMXI

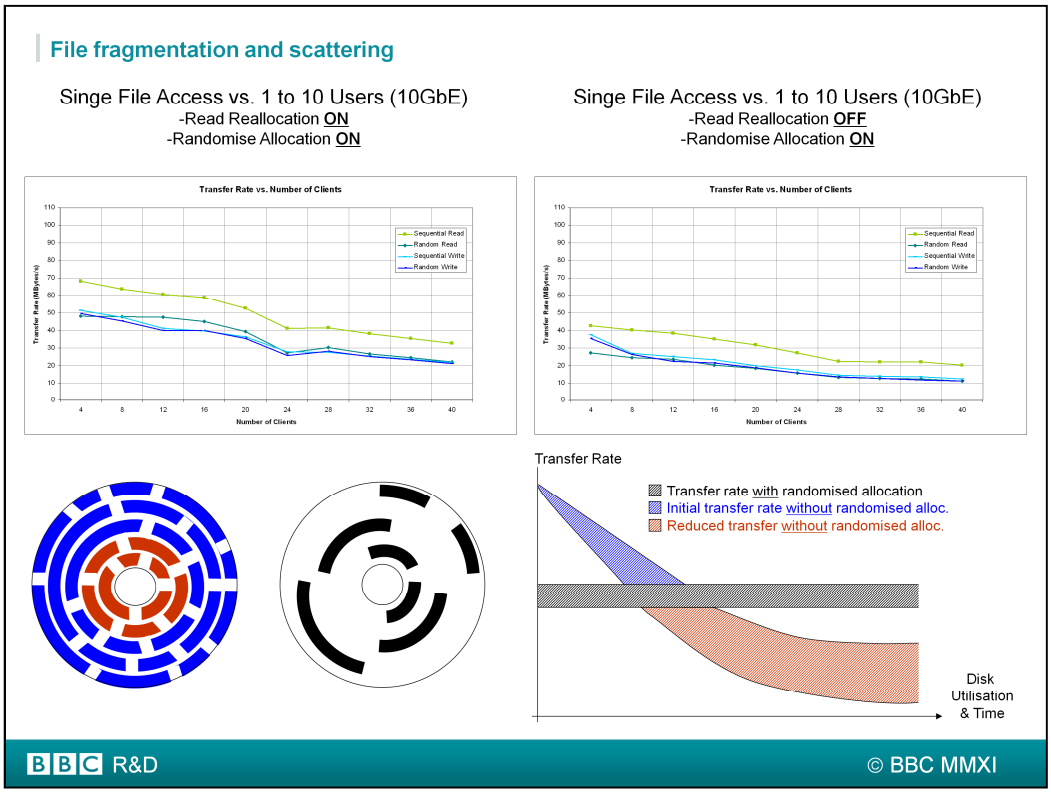
Jumbo frames increase the problem of back pressure, as it is harder to multiplex the larger packets.

The 9000 byte jumbo frames take longer to load and longer to transmit than standard 1500 byte frame. This increases queuing time and delays packets to and from other clients

With jumbo frames there is a larger initial disparity between read rates and write rates due to the back pressure in the switch.

The write transfer rate also decreases more significantly with increasing clients.

In these results the latency graphs are very different from the previous slide, this is covered late in the presentation.



In a production environment, it is important to maintain storage performance over time. Files on the storage can fragment over many reads and writes, reducing performance.

The 2 graphs show the transfer rate performance after several days use. The graph on the left shows results when using a fragmentation reduction algorithm, that intelligently groups blocks from the same file. The graph on the right has no fragmentation reduction. There is a noticeable reduction I performance even after a few days of use.

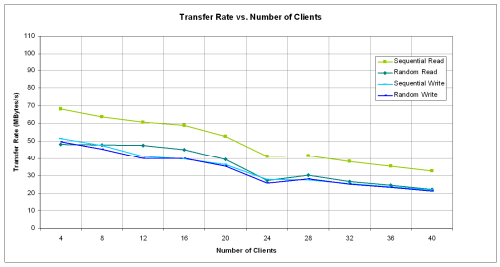
Scattering, also referred to as randomised allocation, is a technique to limit performance reduction with increasing disk utilisation.

Due to the difference in track lengths between the edge and centre of a hard disk, the read write speeds at the edge of a disk are much faster than at the centre. If files are written from the edge inwards, the first files will have very high access speeds. However, access speed will reduce significantly as the disk fills up and tracks close to the centre are used. If files are randomly allocated on inner and outer tracks, the access speed is lower, but the access speed will remain more constant as the disk fills up.

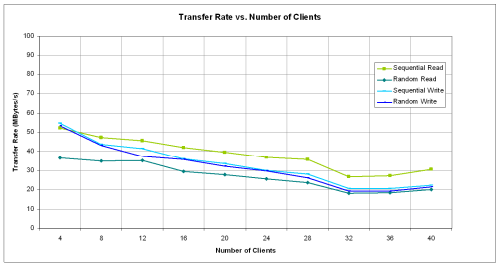
Disk scattering is quite important. With no scattering, If a production system was specified based on file access performance with low disk utilisation, the production tools could be unusable as the disks fill up.

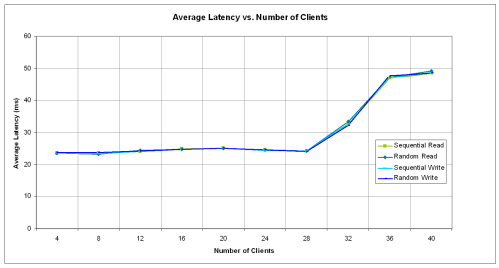
Using SAS or BSAS (SATA with SAS interface)

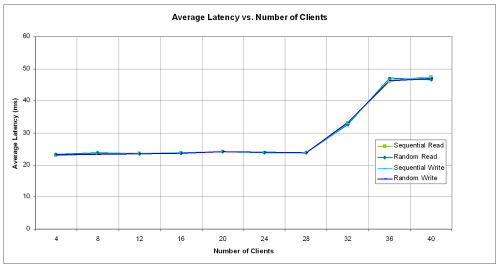
SAS: Multi File Access (10GbE)
-9TB (**23 disks**), 100GB File, 50% Write, 10% Random, 10MB Data
- No flow control, block size 512 KB



BSAS: Multi File Access (10GbE)
-40TB (**68 disks**),100GB File, 50% Write, 10% Random, 10MB Data
-No flow control, block size 512 KB







BBC R&D © BBC MMXI

SAS and BSAS (SATA disk with a SAS interface) offer different performance, reliability and cost.

SAS disk are faster, more reliable, smaller in capacity and more expensive.

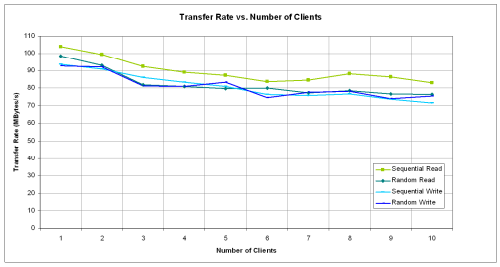
SATA disks are slower, less reliable, higher in capacity and cheaper.

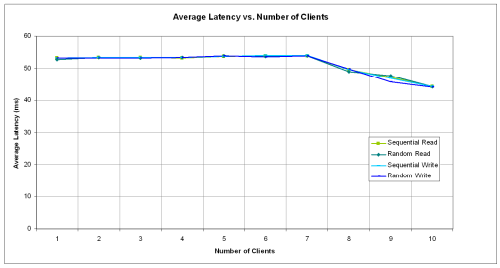
However the results for SAS and BSAS look very similar. The SAS storage consists of 9 TB using 23 disks and the BSAS storage consists of 40 TB using 68 disks, 3 times as many disks.

Disk access IOP speed is determined by both the disk speed and the number of disks. Although individually the SAS disks out perform the BSAS disks, the greater number of BSAS disks provide a similar performance.

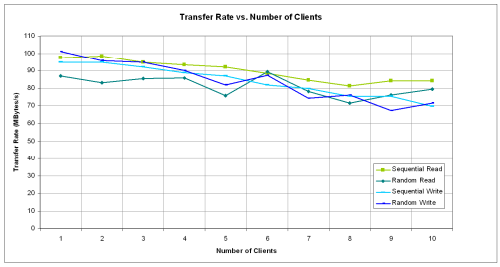
Does switch performance make a difference?

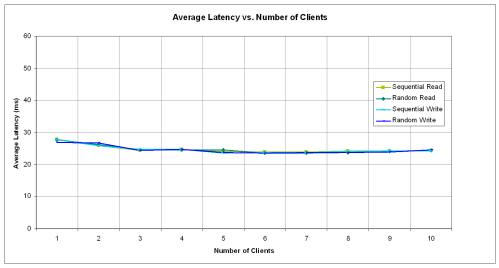
Metro Switch with SAS Single Access (10GbE)
-9TB, 100GB File, 50% Write, 10% Random, 10MB Data
With flow control, block size 1 MB





Edge Switch SAS Single File Access (10GbE)
-9TB,100GB File, 50% Write, 10% Random, 10MB Data
With flow control, block size 1 MB





BBC R&D © BBC MMXI

These results employ different switches to connect the storage. 2 key parameters when specifying a switch are back plane speed (the bus the speed that connects the ports) and buffer size (how much data can be buffered and queued in the switch).

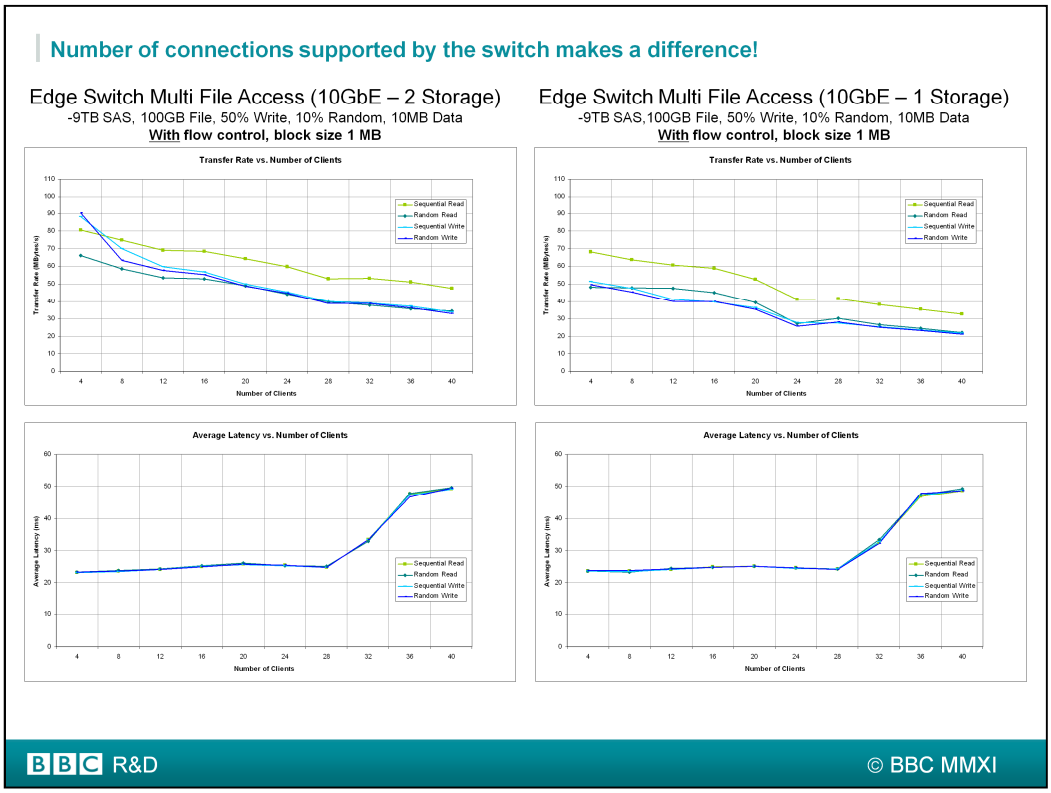
The left results are for a more powerful ‘metro’ switch, which has a backplane speed of 480 Gb/s and 4 x 32 MB tiered buffers.

The results on the right are for a less powerful ‘edge’ switch, with a backplane speeds of 144 Gb/s with a 2MB shared buffer.

The transfer rates for the 2 switches are very similar, indicating that the transfer rates are limited more by storage performance, rather than switch backplane speed.

The average latencies for the switches are very different. This is due to the buffering size and type in the metro switch. The large buffer configuration is normally used to provide QoS for real time services and is actually a disadvantage for low latency file access.

This highlights the issue that a better specified switch does not necessary offer improved storage performance



For the results on the left, 2 network stores support up to 40 clients. For the results on the right, the clients access a single network store.

As expected, the transfer rate is a little higher with 2 storage devices, as the access request are shared over the 2 devices.

However, the average latency values are almost identical. If latency was limited by disk performance, the latency should not increase from 28 clients when using 2 storage devices.

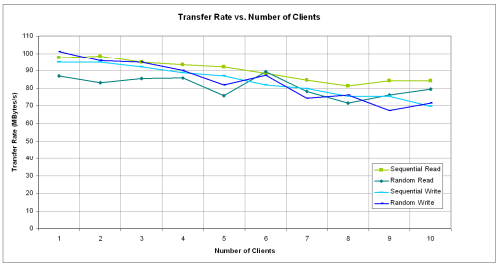
In the tests, 40 clients run on 10 physical clients, with 4 worker programs per client. There are multiple virtual connections, or threads, on each physical connection. Whether using 1 or 2 storage devices, latency increases after 28 connections or threads.

This indicates that performance limitation is in the switch.

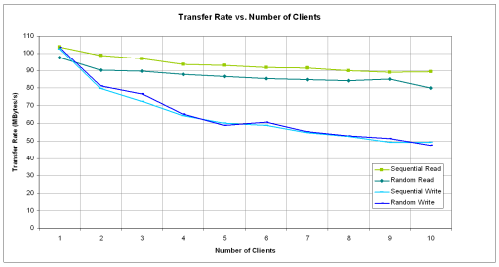
The number of connections supported by a switch is important for storage performance. However, unlike buffer size and back plane speed, this is not a parameter included in the switch specification. In general, a non blocking switch supports enough connection paths for the number of physical ports on the switch.

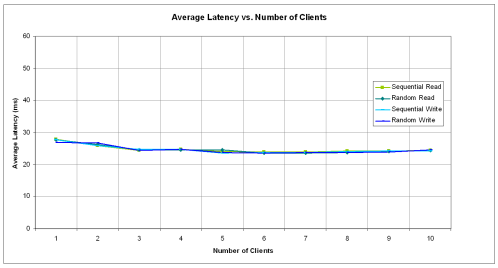
Does 10GbE and 1GbE storage connection speed make a difference?

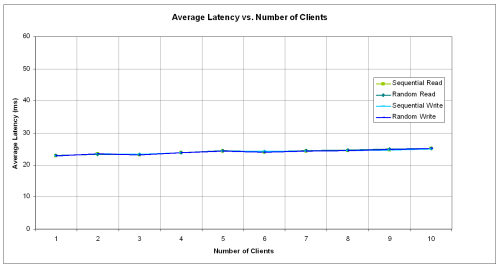
10 GbE SAS Single File Access (10GbE)
-9TB, 100GB File, 50% Write, 10% Random, 10MB Data
With flow control, block size 1 MB



1 GbE SAS Single File Access (1GbE)
-9TB,100GB File, 50% Write, 10% Random, 10MB Data
No flow control, block size 1 MB







BBC R&D © BBC MMXI

Comparing the results for 1 GbE with 10 GbE, using our production profile, the read transfer rates are very similar, but the write transfer rates are significantly reduced.

The average latency results, which are not based on data block transfer, are almost identical.

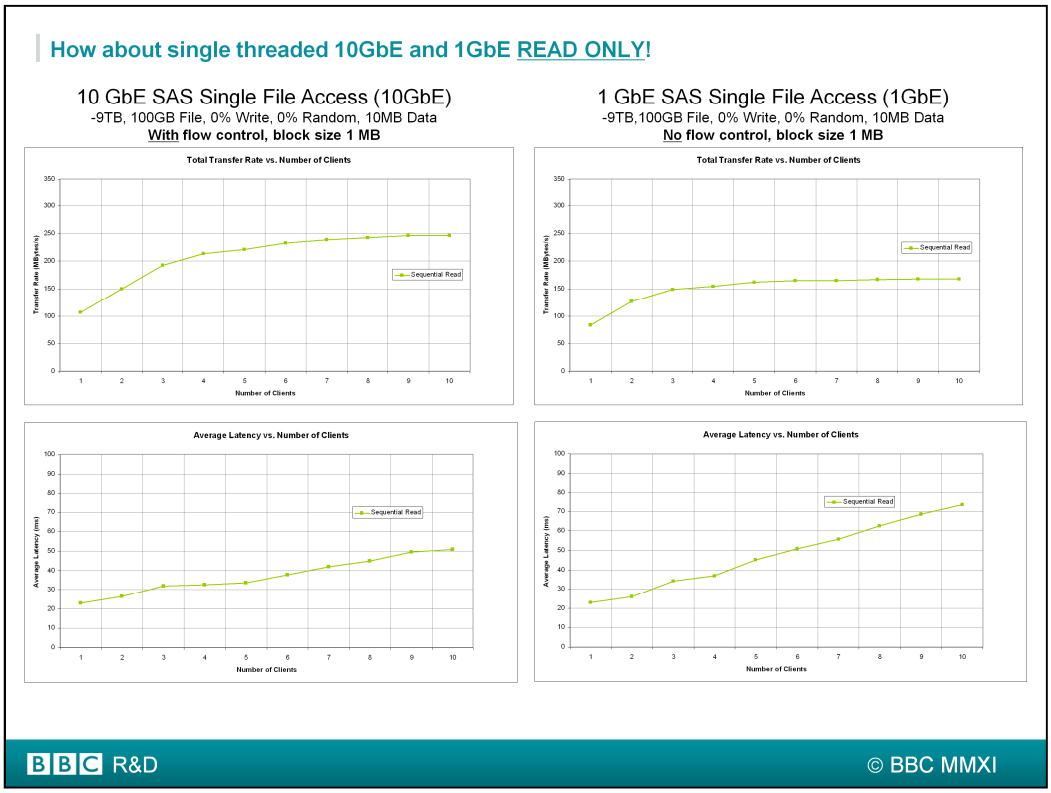
For 10 GbE to 1 GbE data reads, back pressure is an issue.

For multiple 1 GbE to single 1 GbE data writes, congestion is an issue.

In this case, the read results can be a little misleading, as they are very dependant on storage bus speed and the access profile.

The results are averaged over the entire measurement by type. Read and write measurements are not occurring simultaneously.

The results in next 2 slides show things more clearly.



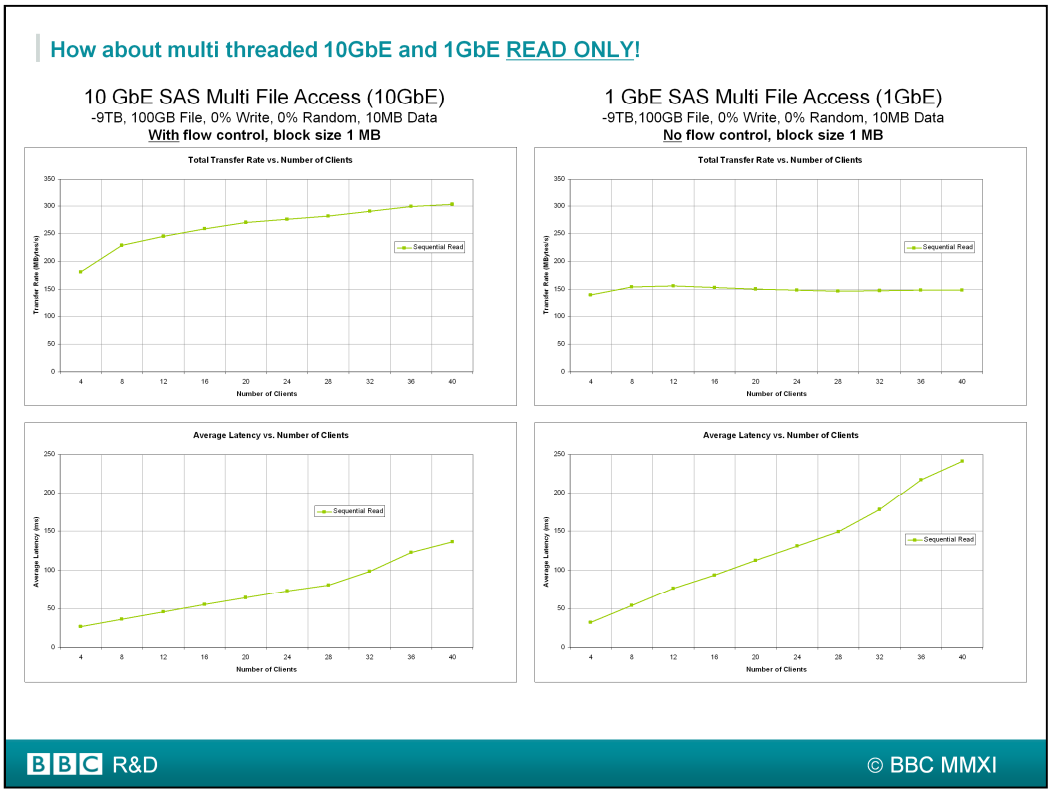
Testing the storage under a heavy load, using just synchronous reads to access a single file, provides a good example of single threaded client access.

The transfer rate for the10GbE storage is greater than for the 1GbE connected storage, but only by roughly 50%.

For 10GbE, the transfer rate should be limited by the speed of the storage internal bus. The theoretical maximum speed on the internal bus is 4Gb/s (500 MB/s).

The results over 10 GbE for this test limits at approximately 2 Gb/s (250 MB/s). Increasing the maximum read length or the multi-threaded access could improve this.

The increased level of congestion for 1 GbE can be seen in both the transfer rate and the latency. The transfer rate limits much sooner than for 10 GbE. The average latency is almost 50% more than for 10 GbE.



Performing a multi-file access test, using a multiple worker programs per client is a good example of multi-threaded client access.

Even though the number of physical clients and connections is the same, multi-threaded access severely increases demand on the storage and network infrastructure.

Looking at the 1 GbE transfer rate, the transfer rate saturates almost immediately at approximately 150 MB/s. The increased congestion is also obvious in the average read latency results, with latencies of up to 250ms.

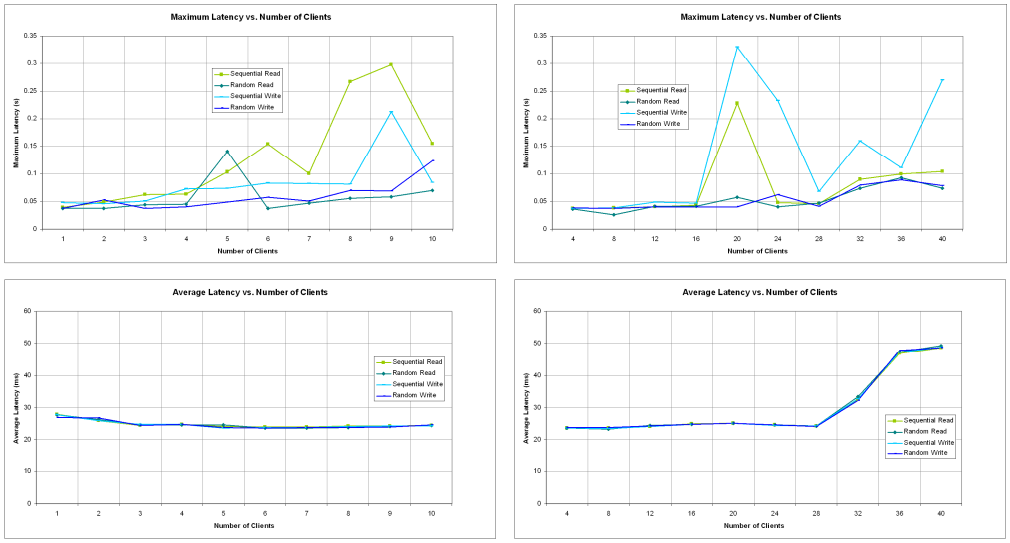
For multi-threaded 10 GbE access, results of up to 2.4 Gb/s (300 MB/s) are achieved. The storage bus speed is 4 Gb/s. For 70 to 80% expected throughput, the maximum expected throughput is 350 Mb/s to 400 Mb/s. The limiting factor in this case is storage bus speed and disk access speed.

The previous slides showed that 1 GbE connected storage scales poorly for write access. These results show that under heavy load, the read results also scale poorly and that multi-threaded access makes a big difference in performance.

Single file & multi file – Peak & Average Latency

Singe File Access vs. 1 to 10 Users (10GbE)
- 1 x 100GB File, 50% Write, 10% Random, 10MB Data

Multi File Access vs. 4 to 40 Users (10GbE)
- 40 x 100GB File, 50% Write, 10% Random, 10MB Data



BBC R&D © BBC MMXI

All the results so far have shown average latency. The maximum latency can vary significantly for both single file and multi file access.

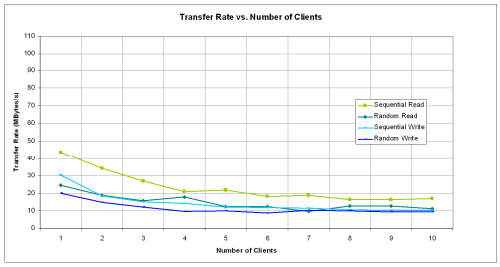
Although the peak latency does tend to increase with increasing number of clients, extreme peaks in latency appear to be random in nature.

These occur when slow disk access coincide with other delays in the system. For example, when consecutive access requests requiring large disk travel coincide with congestion in the storage server, network or client.

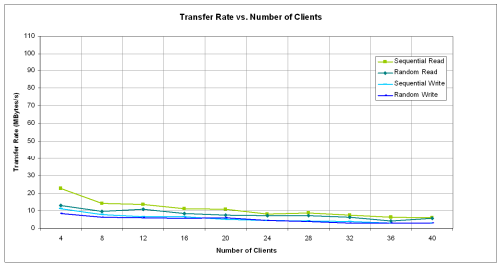
Peak latency is very difficult to predict, interpret and remove, as it is determined by the constantly changing distribution of disk data and statistical effects

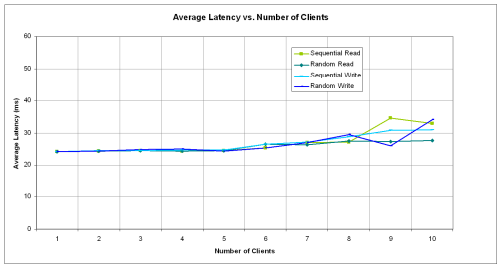
Single file & multi file – OEM Storage

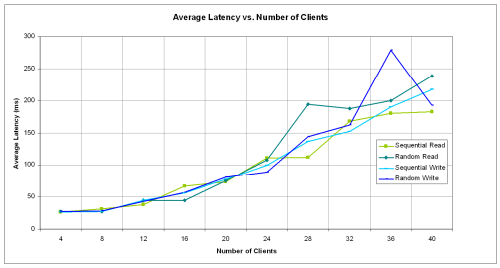
Singe File Access vs. 1 to 10 Users (1GbE)
- 1 x 100GB File, 50% Write, 10% Random, 10MB Data



Multi File Access vs. 4 to 40 Users (1GbE)
- 40 x 40GB File, 50% Write, 10% Random, 10MB Data







BBC R&D © BBC MMXI

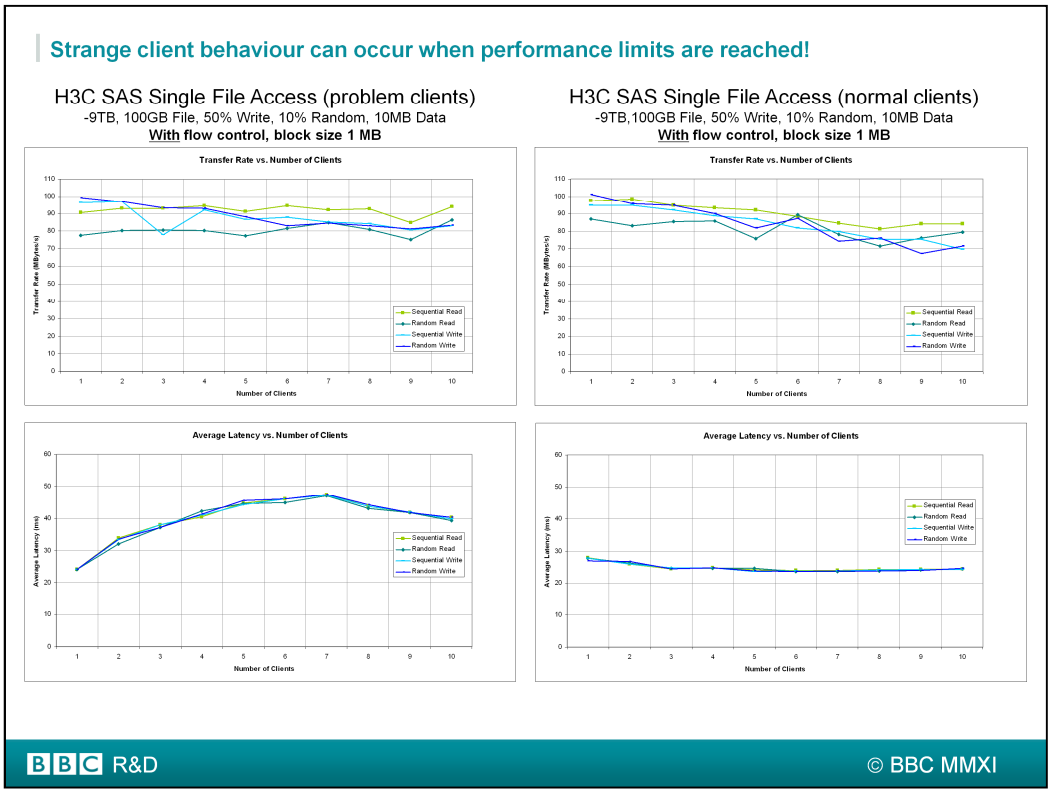
To give a comparison between specialised network storage and off the shelf network storage, the standard single file and multi file tests were performed on OEM storage.

The OEM storage was purchased from an online general IT provider.

For single file access, we see similar average latency results, but the transfer rates are 2 to 5 times lower than for the specialist storage. Also, the OEM storage does not scale well with increasing numbers of clients.

For multi file access, the transfer rates are extremely reduced. Worse still are the average latency results, which increase rapidly with number of clients.

It is unfair to compare the OEM storage with the specialist storage, but it is clear that the OEM storage does not scale well with many users and is not suitable for production use.



Strange issues often occur in real production networks under extreme loads or at the limits of server and client performance. While testing the network storage, 8 out of 10 test clients became fixed in an increased latency mode that could not be removed by hard reset, replacing the network interface card or operating system re-install.

For the test without problem clients, the latency stays very flat with increasing clients. For the test with problem clients, the latency ramps up and levels off. The increasing latency is consistent with the individual behaviour of problem clients, where the latency more than doubles after the first few measurements.

The only difference when viewing captured packets was the size of the TCP window used for problem clients. Either the stored TCP windows size or the TCP window scaling value in the non-volatile memory had been corrupted. The values could not be reset and it required the replacement of the motherboard to correct the problem.

Conclusion

- Application behaviour & file system configuration will have a big affect on performance.
- Consider the cached and non-cached storage performance when specifying storage.
- Production staff could see an initial drop in performance when editing a new project.
- Enable Flow control on a 10G storage connection.
- Jumbo frames should not be enabled.
- Fragmentation reduction or a defragmentation mechanism is required.
- Disk scattering should be employed on all production storage.
- Use SAS or SATA storage where and when appropriate.
- A faster switch does not necessary offer improved storage performance.
- The number of connections that a switch can support is important.
- Use 10 GbE connected storage if possible.
- Peak latency is very difficult to predict, interpret and remove.
- Large difference in performance (and cost) between OEM and specialised storage.
- May see strange behaviour when using clients at the limit of their performance.

End