

EDITING ON GENERIC STORAGE OVER 10GIGABIT ETHERNET: A FINAL CUT PRO USE CASE

P. Defreyne

Co-authors L. Andries and K. Segers

VRT-Medialab, Belgium

ABSTRACT

At present, broadcast companies are increasingly implementing and optimizing file-based workflows for standard and high definition media production. In this respect, the Flemish public broadcaster in Belgium, VRT, has built the Digital Media Factory (DMF) for news production and generic programs in 2007. The workflows are based on the concept of central storage. This paper describes the use case where multiple Final Cut Pro HD editing clients use the central storage cluster as real-time media storage. This is often referred to as in-place editing, or editing on generic storage, which can be considered a recent trend in media production. The challenges for the storage and network infrastructure will be investigated. The insights and experiences presented are valuable to all technical practitioners aiming to optimize their file-based post production infrastructure.

INTRODUCTION

A typical file-based workflow...

At VRT a central storage system was built to support enterprise wide media workflows. Guaranteed throughput, scalability and high availability are the key requirements. A media asset management (MAM) system was installed on top of this storage to manage the media. To enable basic media workflows on the central storage, a number of services have been defined, as shown in figure 1. The ingest services allow for media to be ingested into the MAM. The new media is ingested from news feeds, tape, P2 cards, SDI feeds and more. Often the media needs to be converted to a different compression and file format, respectively called trans-coding and rewrapping. Also a low resolution version needs to be created.

The low resolution media can be viewed and preselected using the browse clients. Media is preselected to be made available for post production. After post production the finalized media is checked-in and archived in the MAM. This media can then be sent to the playout system at the scheduled time.

This type of workflow can be found in most file-based broadcast companies.

... And corresponding dataflows

Looking at this basic workflow from a dataflow point of view, things get more complex. Every service mentioned in this workflow is connected to the IT network. As can be seen in figure 1 almost every service has its own storage system. It is incorrect to assume that media is ingested over the network at real-time speeds. Instead during ingest media is stored on the local storage system of the ingest service. The media is then sent to the

central storage as a high throughput file transfer. If the media needs to be trans-coded or rewrapped, the media file is transferred to the local storage of the trans-coding or the rewrapping service. Once finished the media is transferred back to the central storage. The converted media is now available on the central storage.

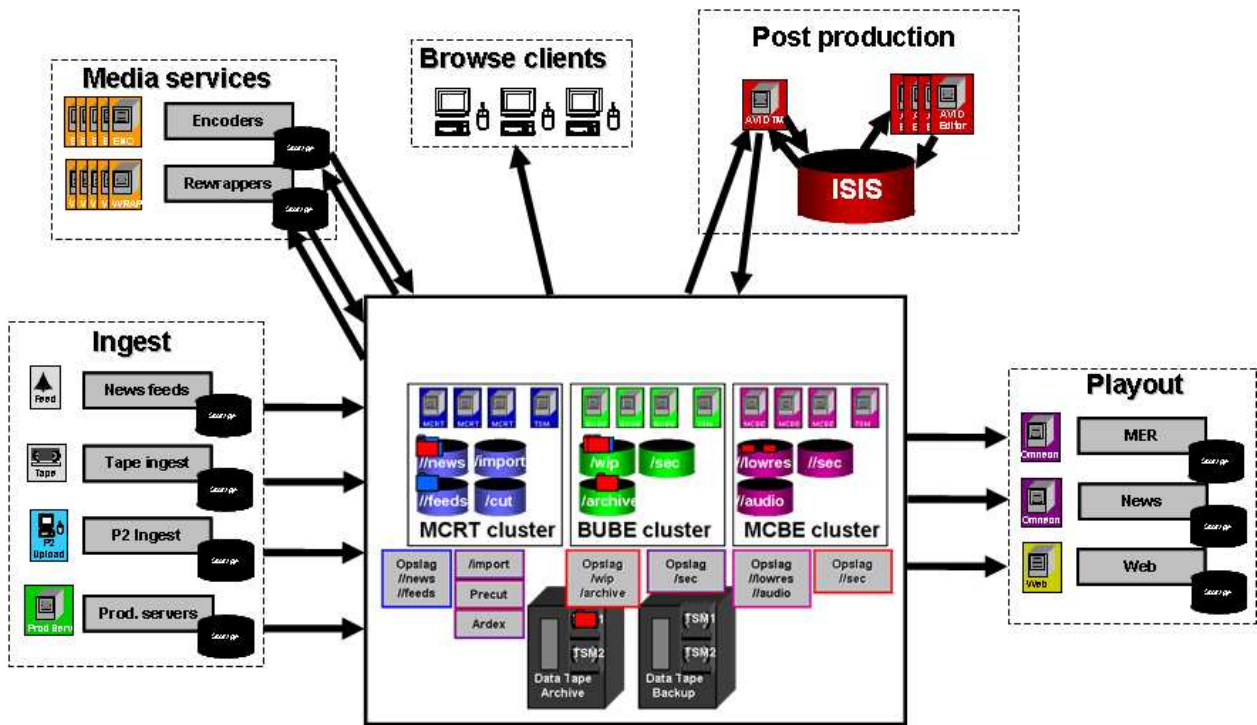


Figure 1 – Central Storage and Media Services at VRT

However the media is not yet available for post production. Using the low resolution browse clients, media is preselected for post production. This preselected media needs to be transferred to the production storage system. However at VRT all media is stored as an MXF OP1A file format. Since the post production environment only accepts MXF OPAtom file formats, the media needs to be transferred to the rewrapping service before the transfer to the post production storage system. Then the media is available for post production. After post production the finalized media is sent back to the rewrapping service to generate a new OP1A file. Finally the media is transferred to the central storage. The media can then be transferred to the local storage of the playout system.

CENTRAL STORAGE DATAFLOW OPTIMIZATION

The workflows at VRT are based on the concept of central storage. The workflows show a tight integration between the services and the central storage. However from a dataflow point of view there is no tight integration. Every step from the workflow requires media to be transferred back and forth between the central storage and the services. The dataflow point of view shows that a basic workflow results in a complex dataflow with many transfers over the network and multiple copies and versions of the same media. These types of dataflows are very inefficient for three reasons. First, most dedicated local storage systems of media services are proprietary systems. This means that these systems can only be used for the particular service. Usually these systems are very expensive. They are based on proprietary architectures and therefore require additional support contracts. Second, the large amount of file transfers puts a heavy load on the network. These very large file transfers generate a high throughput. It has been shown in earlier research that interference between these transfers can lead to failed transfers. Finally it should be noted that each transfer adds time to the workflow.

Since the central storage and the services have no close integrations from a dataflow point of view, complex dataflows are required to execute the basic workflows. Efficiency can be gained by providing a better integration between the central storage and the services. The goal is to run the services directly on top of the central storage. This means that services do not longer require their own proprietary storage since they have direct to the media on the central storage system.

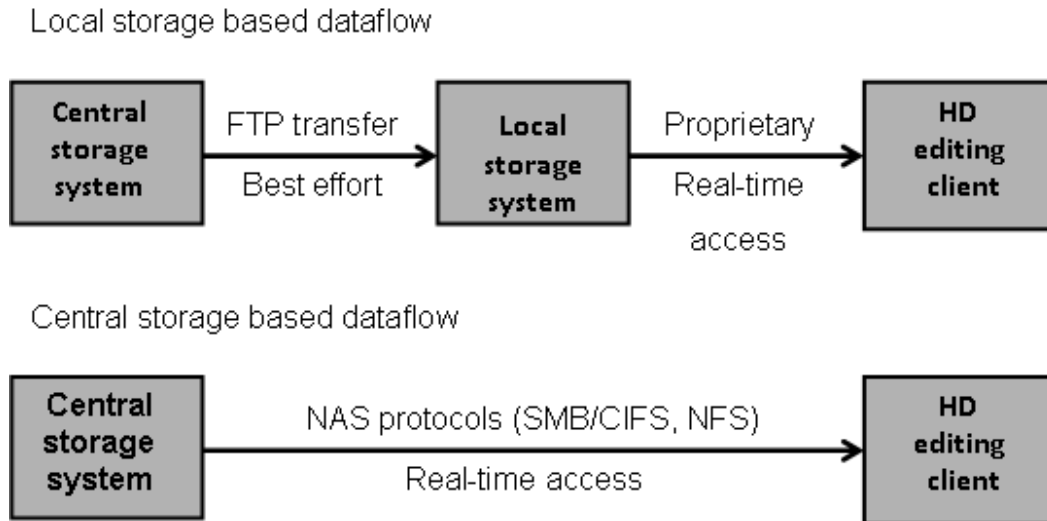


Figure 2 – Integration between Central Storage and HD editing clients

In order to get a clear understanding of the technical requirements, a challenging use case has been chosen. This use case is Final Cut Pro HD editing in a post production environment based on central storage. Both the storage and network domain have been investigated.

From a technical point of view this requires a radical new approach for the central storage and the network architecture. This can be seen in figure 2. On local storage based dataflows, the central storage and network infrastructure offered a best effort performance with soft timing constraints. FTP-based file transfers do not require a guaranteed throughput and failed transfers can always be restarted.

With central storage based dataflows, the HD editing clients require media access with hard real-time constraints from the central storage. Both the storage system and the network need to deliver a guaranteed throughput to each HD editing client and meet the real-time constraints, even under high loads.

THE STORAGE SYSTEM

The design of the storage system needs to deliver real-time performance for multiple parallel clients. Real-time performance means that guaranteed throughput is required for each client. To be able to claim guaranteed throughput the architecture of the storage system needs to be understood completely. This knowledge is used to build a mathematical model of the system. This way performance specifications can be calculated and guaranteed by design.

VRT-Medialab has developed its own central storage system. The architecture is based on generic IT components, with scalability, guaranteed throughput and high availability in mind. The architecture of the storage system is shown in figure 3. Performance is not achieved by scale-out but by optimizing each layer of technology. These layers are the disks, the disk network, the storage servers, the cluster network and the client accessible servers. The client accessible servers have no direct attached storage but have high bandwidth access to the file system. The GPFS file system was chosen for its parallel

throughput, scalability and redundancy. The architecture allows independent scalability of storage capacity, SAN throughput and client accessible nodes. This storage architecture is the result of years of research and development by VRT Medialab. It has recently been developed further and commercialized by a spin-off company.

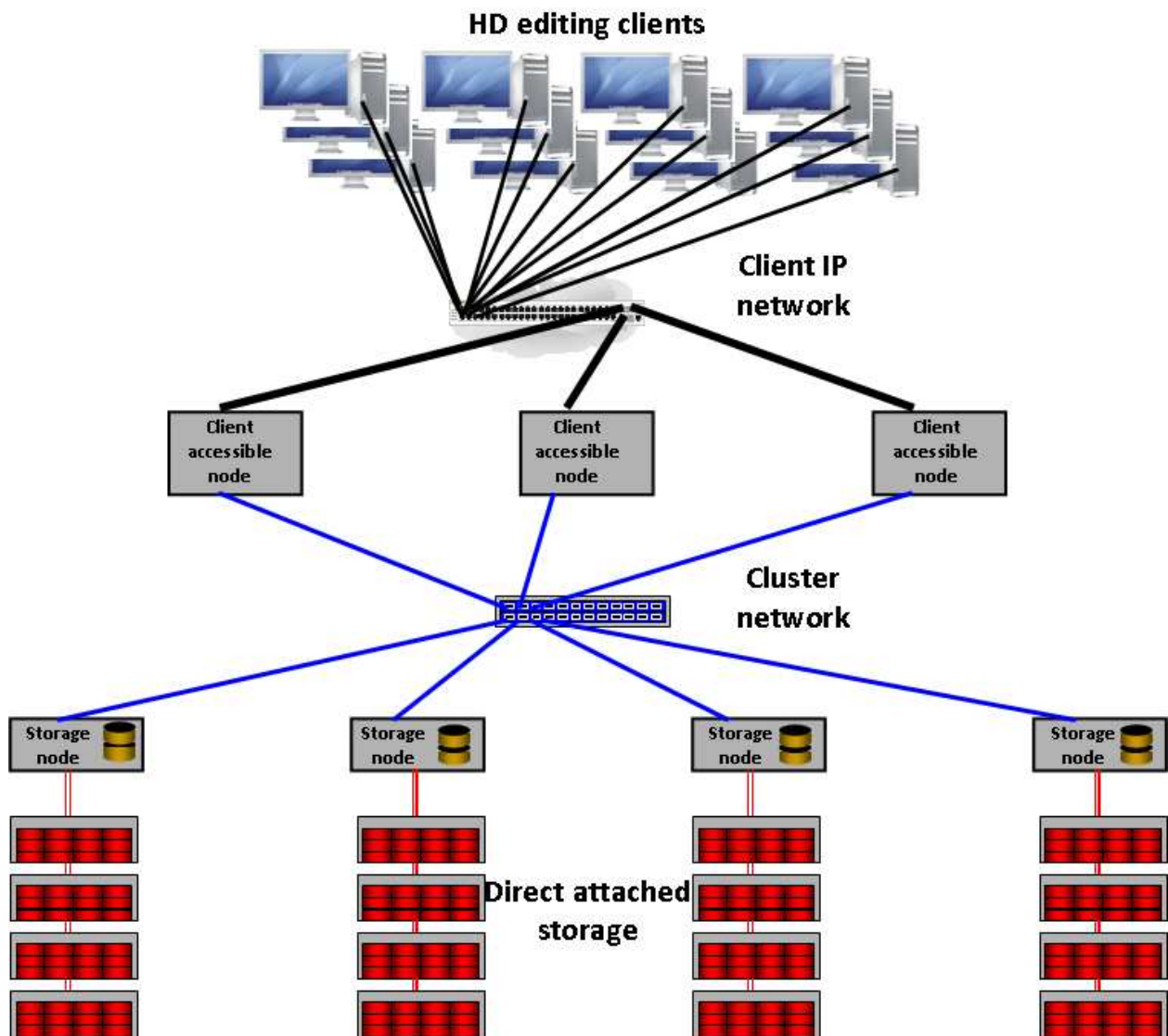


Figure 3 – Architecture of the storage system

THE NETWORK

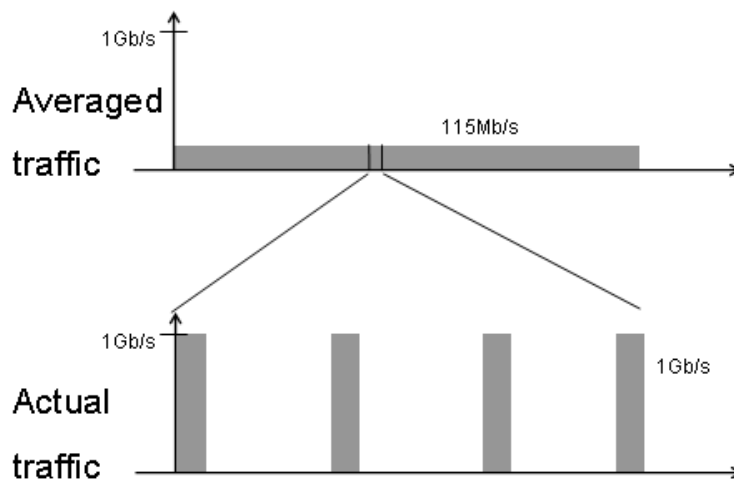


Figure 4 – Media traffic from an IT and a technical point of view

To use the storage cluster as real-time storage for the HD editing clients, the client accessible servers export the file system over the client network using networked attached storage (NAS) protocols, such as SMB/CIFS and NFS, as shown in figure 4. These NAS protocols will replace the protocols used by commercial post production storage systems, such as Avid Unity ISIS or Apple XSAN.

Using the storage system described above, parallel access to the storage system with guaranteed throughput can be achieved. This means getting the data from disk to the client accessible server. It will be shown in this section that getting the data from the server to the client over the client IP network poses new challenges.

First it will be shown how media traffic is described from an IT point of view using the Final Cut Pro use case. Then the corresponding technical point of view is explained. Next two basic examples will illustrate why the IT point of view is no longer valid in media environments. Finally the conclusions are formulated as general guidelines for media network design.

Media traffic from an IT point of view

IT network traffic consists of short messages, such as emails, web pages, database applications, etc... Network load is measured in terms of average throughput. Usually the throughput is averaged over a period of one second or more. The average throughput is then used to dimension the required bandwidth of the network. The network is often overdimensioned to avoid most interference between the traffic. The remaining interference is short-lived and has little or no visual impact on the performance.

In the use case Final Cut Pro clients are playing DVCPRO HD streams. One DVCPRO HD stream has an average of 115Mb/s, as shown in the upper part of figure 4. This average bandwidth of this stream fits easily over a network link with an available bandwidth of 1Gb/s (= 1000Mb/s) or 10Gb/s (= 10000Mb/s).

This is how IT and media solution providers look at media over IT network technology.

Media traffic from a technical point of view

In order to understand the network load from a technical point of view, throughput should be measured on a much smaller time scale. In the lower part of figure 4 it can be seen that the average throughput of 115Mb/s has little meaning on this smaller time scale. The traffic is no longer a continuous stream of data. Instead at regular intervals large bursts of data are sent at wire speed, which can be 1Gb/s or 10Gb/s.

Example 1: 10Gb to 1Gb transition

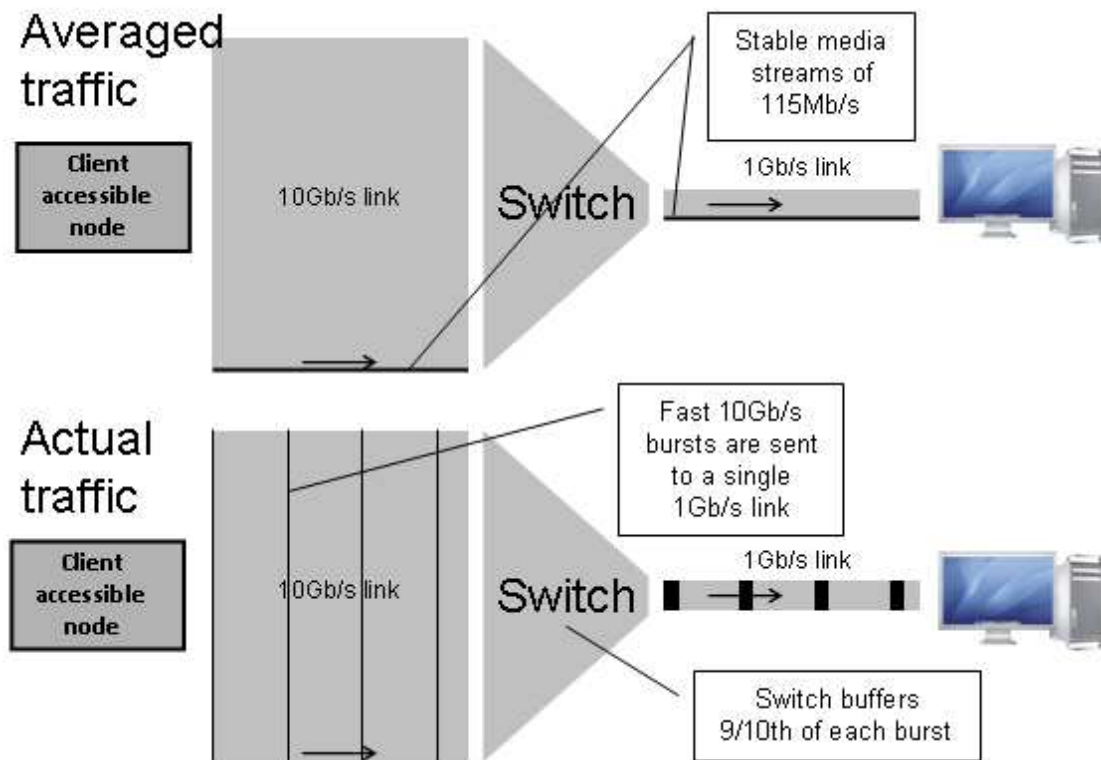


Figure 5 – Example 1: 10Gb to 1Gb transition

Most HD editing clients don't require more bandwidth than a single 1Gb/s (1000Mb/s) network link to the client accessible server. However the number of 1Gb/s ports on a server is limited. With the arrival of 10 Gigabit Ethernet (10000Mb/s) the client accessible servers can support many HD editing clients over one or more 10Gb/s links. A basic network design using 10Gb/s links is shown in figure 5. All HD editing clients connect over 1Gb/s to a local network switch. The server is connected to the switch over a 10Gb/s link. Suppose one client is playing one DVCPPro HD stream. From an IT point of view there is plenty of bandwidth available on each link, therefore the application should work. The technical view shows a different view. Each burst is sent to the switch with a speed of 10Gb/s. From the switch to the client this burst is sent at a ten times slower speed, 1Gb/s. As a consequence 9/10th of the burst must be buffered by the switch. The bandwidth mismatch is called a 10:1 oversubscription. If the buffer capacity of the switch is smaller than 9/10th of the burst size, packets will be dropped. Packet drop means that the server will need to retransmit the lost packets. These retransmissions will consume bandwidth and add latency. As a consequence the real-time requirements can not always be met. The editing application can start dropping video frames and might eventually stop working. Most switches can enable a pause mechanism on its ports, which pauses the link before the switch drop packets. Using this mechanism the 10Gb link could be paused during bursts. However the link will also be paused for every other client connection over this link, causing additional latency and interference. The properties of the traffic, the server and switch must be understood completely to determine if pause mechanisms work for a particular setup.

Example 2: 2Gb to 1Gb transition

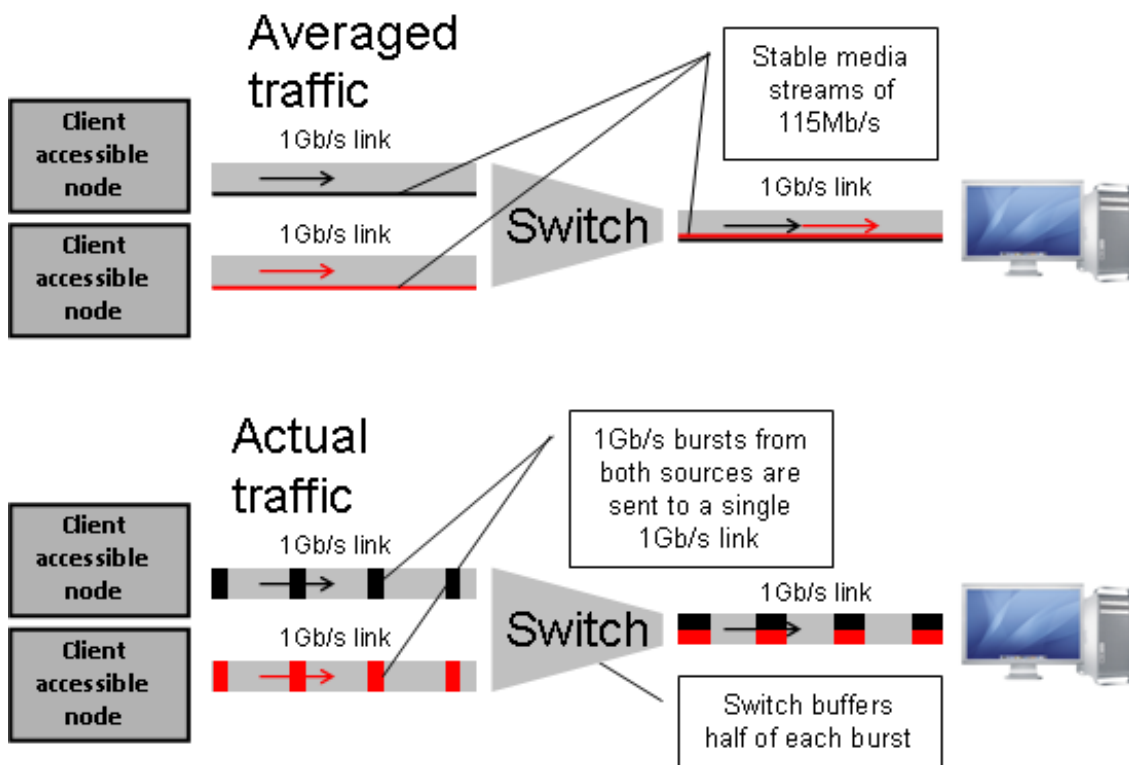


Figure 6 – example 2: 2Gb to 1Gb transition

This example describes the test setup as shown in figure 6. In this example there are two servers, each exporting the central storage. Both servers are connected to the switch over a 1Gb/s link. In the setup one client plays one stream from two different servers. Again from an IT point of view the network design provides plenty of bandwidth. From a technical point of view there is a 2:1 oversubscription, since the bursts from the servers are sent with two times a bandwidth of 1Gb/s. The client only has 1Gb/s of bandwidth, which means that the switch needs to buffer half of each burst. If the switch buffers are too small, packet drop will occur. Again the application runs the risk of not working properly.

Conclusions for media networks

Both examples show that despite ample bandwidth the application risks not working properly. By only determining the average throughput it is impossible to detect the cause of the problem. Looking at the traffic from a technical point of view, it was shown that oversubscription on small time scales could cause the switch to drop packets. These are the basic guidelines to build a working media network infrastructure from a technical point of view:

- Measure the size and speed of each burst. For QuickTime files the tool Dumpster allows to check the chunk size, which is equal the burst size. For MXF-based files the network traffic should be traced and analyzed for each application.
- Choose a switch with large buffer capacities. Configure the switch to allocate the maximum amount of buffers for each port.
- Dimension the buffer requirements using the worst case scenario.
- Determine the amount of oversubscription considering the amount of streams, clients, switches, ...

- Packet drop should be avoided. TCP retransmission mechanisms are too slow to match the real-time constraints. Also during retransmission the traffic is not deterministic and throughput can no longer be guaranteed as a consequence.
- Be extremely careful when enabling a pause mechanism on a link to avoid packet drop.
- Do not assume that a design based on IT best practices, will result in a working solution. Be aware of oversubscription at small time scales.

CONCLUSION

At VRT the architecture of the Digital Media Factory includes a central storage system and services with local storage. This results in a mismatch between central storage based workflows and the corresponding dataflows. While the workflows suggest a tight coupling between the central storage and the media services, there is no tight coupling from a dataflow point of view. This results in many file transfers and expensive dedicated local storage at each service.

To optimize the dataflows, it was proposed that the central storage could be used as media storage. Services do not longer require their own local storage since they have direct access to the media on the central storage system. A storage system with guaranteed throughput is required. The network will become an even more critical component of the design. IT best practices are not valid in a media context. The media traffic needs to be measured at a much smaller time scale. With this information the buffering requirements of the switches can be calculated for each network architecture.

ACKNOWLEDGEMENTS

The author would like to thank Luc Andries for years of research on implementing IT storage and network solutions in a media environment and thus providing the context for this paper. The author would also like to thank Jeroen Van Aken, Koen Segers, Luk Overmeire and Stijn De Smet for their contributions to this research.