Building media networks with IT solutions:
Media Data Centre
— the path to performance & efficiency in the media workflow

Luc Andries
VRT / CandIT-media

IP-based architectures are now fully accepted as the standard solution for file-based media production. However, media traffic is intrinsically different from IT traffic, causing classically-designed IP networks to behave unexpectedly differently under this new traffic load. Parameters which have traditionally been used to specify IT traffic, such as average bandwidth, are no longer valid or even relevant to predicting the behaviour of an IP switch when used in this media environment.

File-based media production

The advent and maturing of Internet technology over the last few decades has totally changed the landscape of the IT industry. The absolute success and popularity of (mostly Ethernet-based) IP networks has promoted this technology as the prime architectural choice in most IT environments. Central mainframe computers have in most cases been replaced by distributed client-server architectures, connected by very powerful IP networks.

Steadily, this technology is being introduced in other industries as well. Although adoption and, above all, acceptance of these new technologies was at first occurring rather slowly in the media world, IP-based architectures are now fully accepted as the standard solution for file-based media production and have drastically changed the way that broadcasters operate and function internally. For a long while, broadcasters were using a sequential videotape-based workflow model. However, during the latter part of this decade, they have finally started to embrace Internet technology in their production back-office, leading to a collaborative workflow model *(see Fig. 1)*.

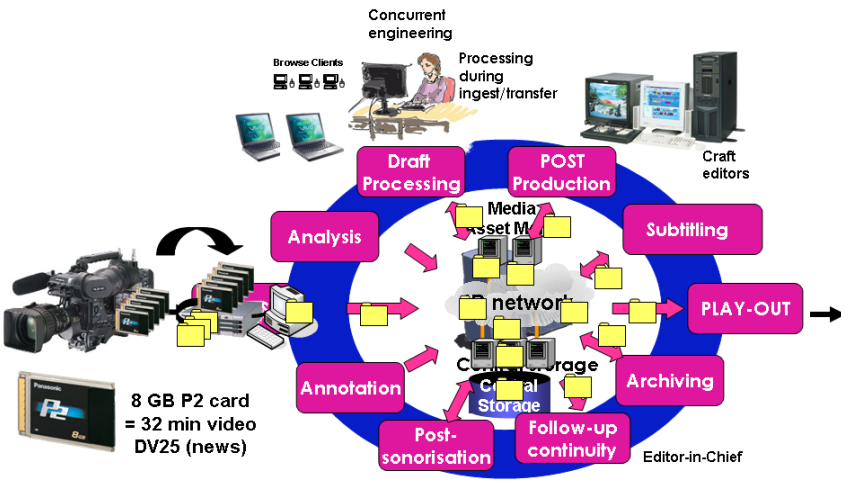

Figure 1
File-based media production = Collaboration Model

Applying an ICT-based infrastructure in video/media production, and IP networks as the means of data transport, introduces a number of possible benefits and is facilitating the fundamental shift from

traditional tape-based video manipulation to a file-based production paradigm. This leap in technology enables video to be treated, processed, stored and transported as ordinary files, independent of the video format. Amongst others, the most profound technology changes are:

- IP network-based access and transport of the media;
- central disk-based media storage;
- server-based (non-linear) video editing or processing;
- software-based media management and media production systems.

Together with the appearance of some standards like MXF and AAF, which provide a generic file container for the media essence, these changes have led to the file-based paradigm of ***media essence***.

Typically, camera crews now enter the facilities with their video stored as ordinary files on memory cards instead of on video tape. The memory cards are put into ingest stations and the files are transferred as fast as possible, preferably faster than real-time, into a central disk-based storage system. Once stored in the central system, everybody can access the material simultaneously. This should lead, in principle, to a much more efficient workflow. Production lead times should become shorter and deadlines should become closer to the moment of broadcasting.

A typical example of such a file-based infrastructure is being deployed at VRT (the Flemish public broadcaster in Belgium), as schematically depicted in *Fig. 2*. On the left-hand side, the video sources are depicted. These include uploads from old tape-based archives, real-time feeds, tape-based inputs (video players), file-based camera inputs and production servers in the studios. At the top, are the non-linear, file-based, high-resolution editing suites. On the right-hand side, several playout channels are listed, including classical linear broadcasting and internet web farms.

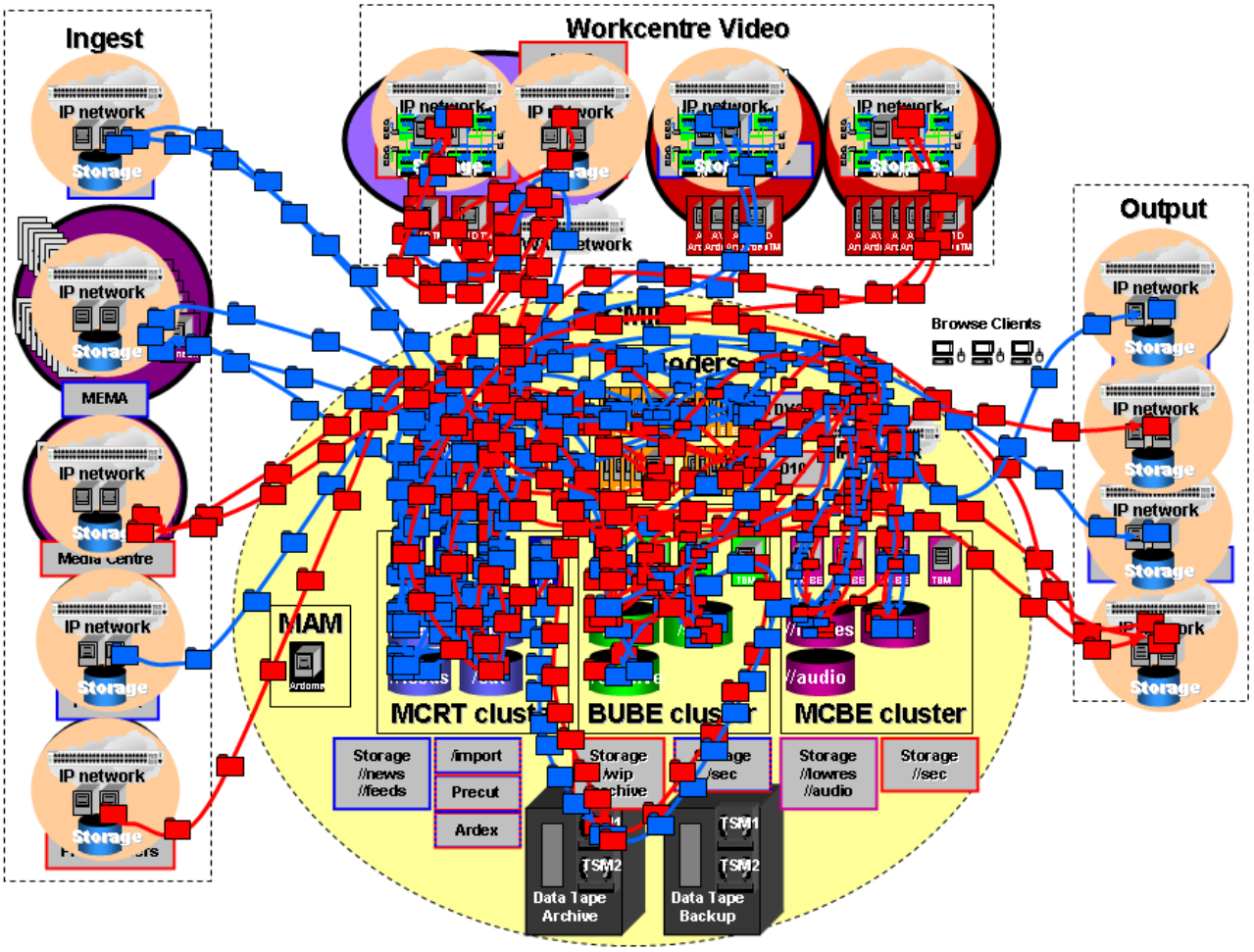

Figure 2
Data flows in the Digital Media Factory at VRT

All these peripheral systems are connected via a central IP network to the central infrastructure, which provides a massive storage warehouse for different flavours of the media, e.g. high-resolution video, low-resolution proxy video and audio. Additional central services such as transcoding and backup/archiving are also included. All media essence is managed by a central **Media Asset Management** system (the MAM system).

At peak times, when many of the different workflows are executed at the same time, a study at VRT estimated that over 500 simultaneous file transfers were being launched on the central infrastructure. If one were to map all the workflows into the data flows and draw the actual file transfers on the infrastructure picture, this would result in *Fig. 2*:

1) Evidently, this puts a very high and largely unexpected load on the central IP network.
2) Transfers share the available bandwidth on some of the links and server interfaces on the network. The network becomes oversubscribed, with mutual interference between different transfers as a consequence.
3) On top of this, IT traffic is intrinsically different from media traffic. Media files are much longer. Traffic is more bursty. Consequently, a classically-designed IP network reacts differently to this new kind of traffic, with unexpected delays and even broken transfers as a result. Packet loss is far less acceptable when dealing with media files. *"A slow e-mail is still an e-mail, but a slow video is no longer a video."*

Behaviour of IP networks in media – the quantum view

IT traffic – such as SAP (business management software) traffic, Microsoft Office documents or e-mails – typically consists of short messages or small files. The IP network is generally used for only relatively small time periods. Transfer speed is not critical. Packet-loss and the resulting retransmission are acceptable.

However, media traffic deals with very large files, generally a few gigabytes in size, that are transmitted at speeds faster then real-time, i.e. in relation to the video compression format used. Hence, media traffic will typically use the link for a large time period and will almost constantly try to use 100% of the available bandwidth of the network infrastructure. The longer this period, the more bursty the traffic becomes. If different transfers share the same link, bandwidth competition between individual concurrent transfers will occur. This will generate packet loss. The resulting retransmissions will decrease the overall efficiency of the transfers drastically. If sustained, this can lead to complete transfer interruption.

IT versus media traffic – cars versus trains

We can make the distinction more clear by using the following analogy. Consider two IT clients, e.g. running Word and Excel, each sending IT traffic at a speed of 400 Mbit/s to a common file server.

Abbreviations

AAF	Advanced Authoring Format
CPU	Central Processing Unit
DCB	Data Centre Bridging
DNxHD	(Avid) Digital Nonlinear extensible High Definition (codec)
EDL	Edit Decision List
FC	Fibre Channel
GPFS	(IBM) General Parallel File System
IB	InfiniBand – a switched fabric communications link
ICT	Information and Communication Technologies
IP	Internet Protocol
IT	Information Technology
LAN	Local Area Network
MAM	Media Asset Management
MXF	Material eXchange Format
NAN	Network-Attached cluster Node
OPAT	(MXF) Operational Pattern Atom (OP-Atom)
PFC	Priority Flow Control
PPP	Per Priority Pause
SAN	Storage Area Network
TCP/IP	Transmission Control Protocol / Internet Protocol
WARP	Workhorse Application Raw Power

The traffic could consist of a small file or even simple commands in response to some keystrokes made by the user. This will result in 800 Mbit/s being received by the server from the two clients. This type of traffic could be related to the way that cars drive on a two-lane highway *(see the left-hand side of Fig. 3)*.

After passing the cross point, i.e. the switch, the two lanes have to merge into one lane. Most of the time, this will happen quite efficiently without too much traffic delay, since streams of cars will nicely "zip" together. In the IT environment, the server will indeed receive traffic at 800 Mbit/s, without much delay. Bandwidth or throughput will nicely add up, linearly in most cases.

However, circumstances are different when dealing with the transfers of large media files *(see the right-hand side of Fig. 3)*. Now, the traffic consists of sustained, very large, bursts of packets arriving back-to-back. Two clients simultaneously sending large files at 400 Mbit/s to the same server will no longer manage to get all the traffic through to the server, without interference. One could, using a similar analogy, describe the traffic as trains running on two tracks, where the switch acts like a junction. If both trains approach the junction at the same time, they will crash into each other, leading to a catastrophe. And traffic will be stopped. Consequently, the media server at the receiving end will not attain an aggregated throughput of 800 Mbit/s, but much less. Throughput can no longer be added linearly.

Ideally, each train should have its own track, or otherwise traffic lights should be installed to manage the possible congestion. Translated to the IP network world, the architecture of the media IP network should ideally provide for separate links for each traffic flow. This is however only technically and practically feasible for very small setups. The alternative would be a traffic management system that takes into account the "long trains" of media traffic.

Clearly, an IP network shows a different behaviour when transferring large media files. Many broadcasters and providers of file-based media solutions are reporting the same problem. The IP network just doesn't behave "as expected". Throughput decreases and becomes unpredictable – transfers are being lost. And above all, these effects are not limited to very large architectures, such as the

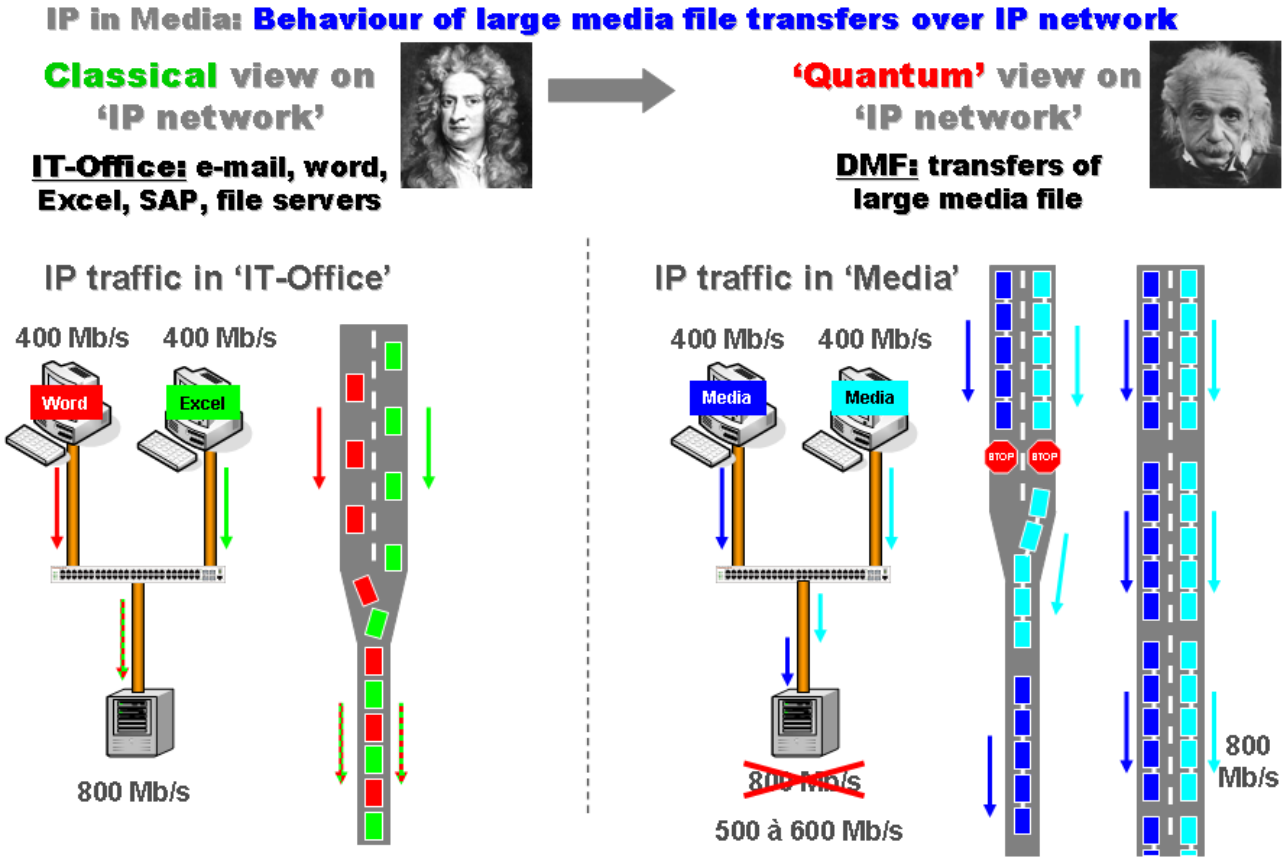

Figure 3
Behaviour of large media file transfers over an IP network

above-described digital media factory at VRT. They also appear in very small installations, dealing with only a few clients and servers.

The quantum view

There is a lot of mystique about what is actually the underlying cause of these problems. And the effect is certainly not visible when applying the classical network-monitoring tools. IP networks, and more specifically IP-over-Ethernet networks, have been deployed by the IT industry for around three decades. It has become ***the*** standard in local area networking in the enterprise world – the LAN. With the ubiquitous adoption of this technology, a large number of network-monitoring practices and subsequent monitoring tools arrived on the scene. These tools are however geared towards monitoring and managing IT traffic on IP networks, and are certainly not optimized or adapted for tracking the specific media traffic described above. They typically deal with measuring throughput by means of averaging over relatively long time intervals.

When IT was being introduced to the media world of broadcasters in the last couple of years, the media engineers embraced the same monitoring tools as the IT industry, to manage their media IP networks. And media engineers have learned to apply the same monitoring practices to detect and resolve network problems as the IT industry has been doing for decades. This could be called the "classical view of an IP network".

In order to deal with "unexpected" behaviour on an IP network in the media environment, the way the network has to be monitored and the tools that have to be used are fundamentally different. Again, an analogy with what happened in the evolution of physics could clarify this point. Since the 17th century, physics has been ruled by the laws and views of Sir Isaac Newton. Nature was described on a macroscopic scale where everything was considered to move or change in a "continuous" way.

However, at the end of the 19th century, the first few experiments appeared on the scientific scene, that could no longer be explained by this "classical" Newtonian view on physics, e.g. black-body radiation, the photo-electric effect, etc. It took the genius mind of Albert Einstein to formulate a radical new view on physics and understand the deeper nature of these unexplainable effects. Einstein stated, in his Nobel prize-winning essay on the photo-electric effect, that nature is quantized on a very small scale and should be described by discrete levels. From that time on, around the start of the 20th century, the view on physics and nature became radically different and "quantum-mechanics" replaced the physics of Newton.

A similar thing has happened in the IP network arena. With the application of IP technology in the media environment, strange effects started to appear that couldn't be explained by the "classical view" held by the IT industry. These effects can only be explained if one starts to look at the network on a completely different timescale, several orders of magnitude smaller than what the classical network-monitoring tools are capable of. At that timescale, concepts such as "average network throughput" become meaningless, since the network starts to behave in a discrete way. A network link is loaded with a packet, or it is idle. There is no such thing as a "bandwidth percentage" anymore. We have to look at the network in a quantized way. This new way of looking at an IP network can be referred to as the "quantum view of an IP network".

Private media cloud solution

Classical IP-centric file-based production infrastructure

The broadcaster's back-office has evolved a lot due to the file-based paradigm. However, this evolution happened in an unstructured, chaotic way. The media solution vendors came with their own very specific answers, without taking into account the complete technology picture.

Most of the architectures are just a "hotchpotch" of products, with each media service having its own local storage, servers and local network, connected to each other in a best-effort mode via the central IP network. As a result, there is a lot of duplication and the complexity has reached a point

where the system becomes unmanageable – a cloud of undocumented, unorganized ad hoc solutions. Moreover, the integration of new products and upgrades to existing ones are getting more and more problematic.

Inter-connecting all these media services in the classical way leads to an extremely IP-network-centric architecture *(see Fig. 4)*. The classical IP switches known and used in the IT environment are no match for that task, given the specific bursty nature of the media traffic.

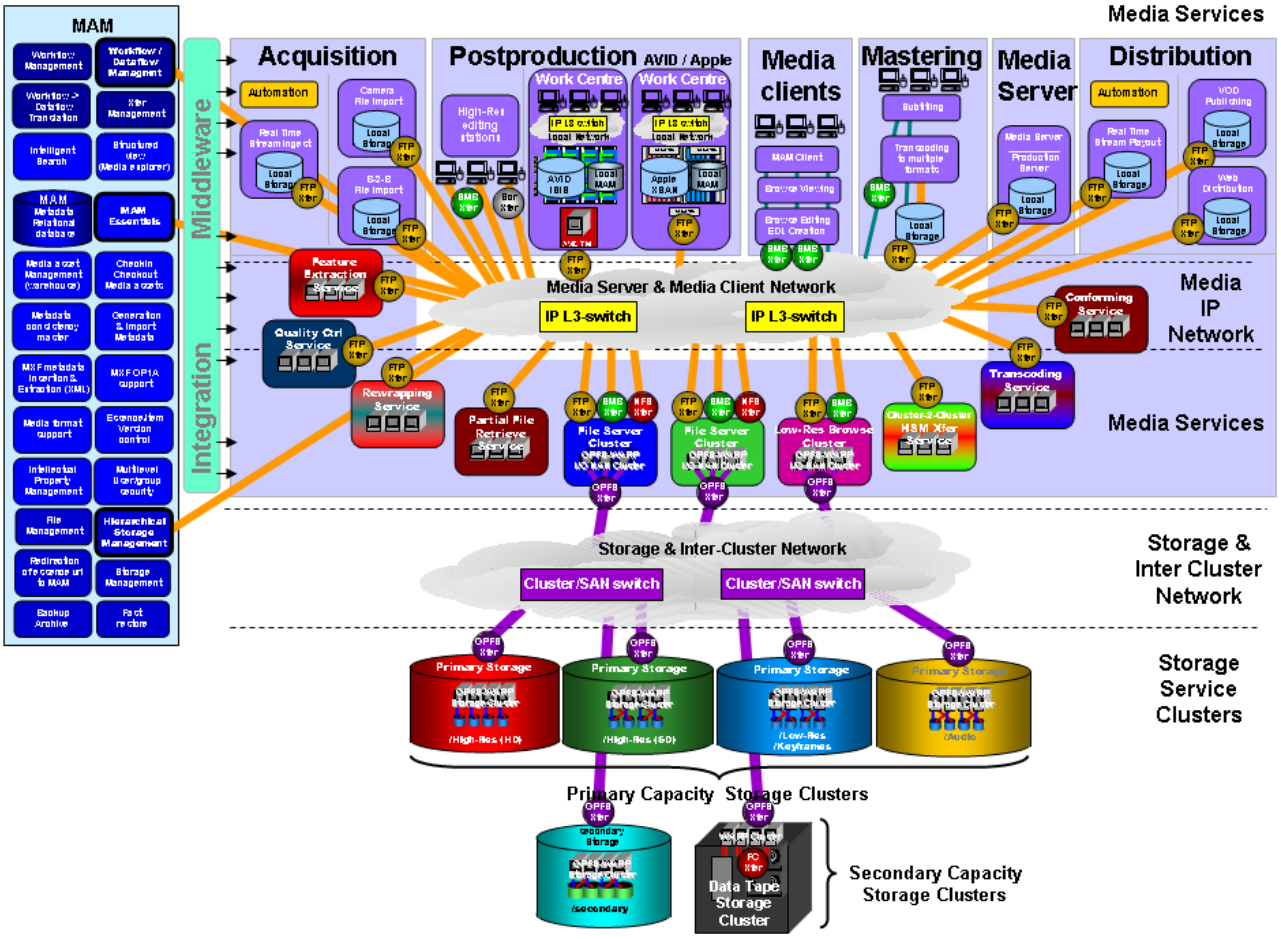

Figure 4
Classical IP-network-centric, file-based, production infrastructure

Today most of the data traffic is launched by the MAM application or the media applications themselves, independent of each other and unaware of the underlying architecture. In the absence of a proper media traffic management system, this leads with certainty to the traffic problems that many broadcasters are already experiencing today.

Because most media services reside in the "client IP-network" in a loosely-coupled way, each with its own local storage and servers, the overall infrastructure consists now of file-based islands. As a consequence, traditional sequential tape-based workflows are often being replaced by almost identical "sequential" file-based workflows. This leads to very inefficient data flows, as files are now being exchanged back-and-forth between these islands in an any-to-any traffic pattern, with many duplicated copies residing in the local storage systems. Because of the bursty nature of the media traffic, packet loss results in unpredictable transfer delays or even transfer loss.

Private media cloud architecture

Many of the essential media services require a processing power platform that is close to a storage service. As described above, media services traditionally use local storage of a proprietary nature, in many cases. Data is then transferred from one local storage system to the next as sequential

steps of the workflow are being executed. This inefficient way of working could be overthrown if all these local server and storage platforms could be unified into a central platform, a **Virtual Media Data Centre**, with the following characteristics:

- guaranteed throughput;
- linear high scalability;
- cost-effective – efficient;
- reliable – redundant – recoverable;
- clustered media service platform capable of supporting multiple OSs (Linux – Windows);
- flexible – maximum efficiency – green by service virtualization.

This would lead to the architecture depicted in *Fig. 5*.

In this scenario, almost all media services could run on the processing nodes of the virtual media data centre cluster, closely connected to the uniform central storage of the cluster, thereby eliminating the need for proprietary local storage.

Now, instead of the central IP network, this scalable clustered storage system could provide the basic platform for the interconnection of these media services, creating a media data centre or private media cloud solution. Since most of the inter-service media traffic would be using the cluster network, the physical data flows would strongly benefit from the lossless characteristics of such a type of network.

Hence, the traffic between these media services will be offloaded from the generic IP network, onto the lossless storage or cluster network. This will, in turn, offload the requirements of the client IP network considerably, making it easier to design a media-aware client IP network, capable of handling the remaining bursty media traffic.

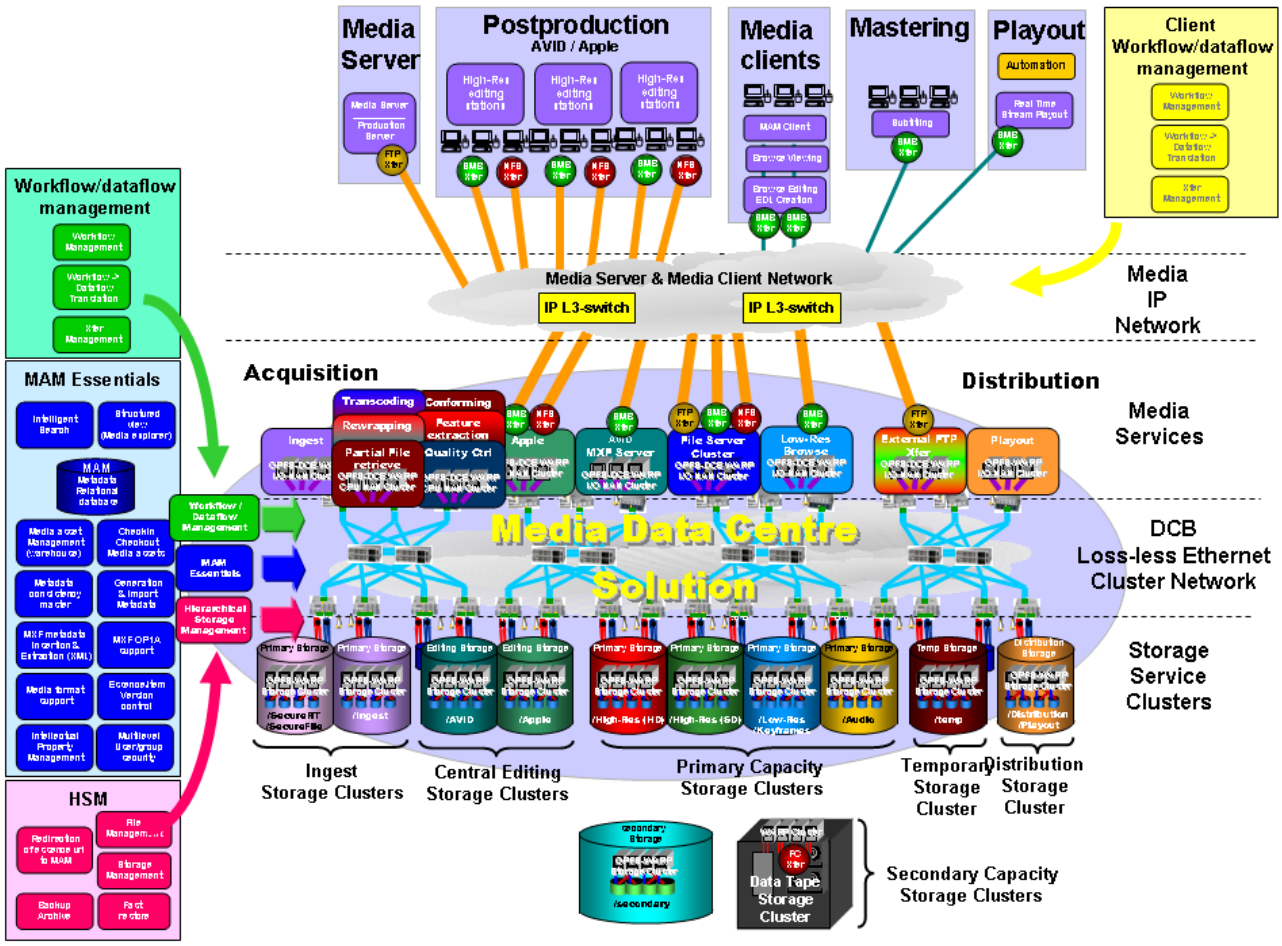

Figure 5
Private media cloud solution

The media data centre – lossless Ethernet

FC-based media storage network

As storage protocols are in general not very well equipped to recover from data loss, they presume the underlying infrastructure to be of very high reliability. Hence, lossless behaviour is one of the most fundamental characteristics of a storage network.

In traditional IT environments, Fibre Channel (FC) is the prevailing storage network technology. It uses a buffer-to-buffer credit mechanism to avoid frame loss. Credits of available buffers are continuously exchanged between ports on the same link. When no buffer credits are available, no packets are transmitted, until the network processes its congestion and buffers become available again. Hence the receiver never needs to drop frames.

However, when there is sustained media storage traffic load, the long bursts interfere with each other in the FC switch buffers, and create a severe efficiency loss, not displayed in a typical IT back-office environment. It has been demonstrated that in more general and extended media storage network topologies, this effect severely impairs the efficiency of the network and therefore limits the scalability of an FC-based storage network in media storage environments.

The ideal media storage network – lossless Ethernet

One can partially overcome these scalability limitations by splitting the storage network into two separate networks – a cluster network and a local storage network.

IBM's General Parallel File System (GPFS) – one of the most powerful media file systems available on the market today – allows for such an architecture *(see Fig. 6)*. A GPFS cluster based on this architecture, called a WARP cluster (Workhorse Application Raw Power cluster), consists of storage cluster nodes and network-attached cluster nodes (NAN nodes). In this model, the storage is directly connected to a storage server, whether locally attached or via a local SAN architecture. NAN nodes are via a cluster network connected to all storage nodes, but are not directly attached to the underlying storage. The NAN node stripes its data requests over all storage nodes, thereby aggregating the available bandwidth of each individual storage node and connected storage subsystems.

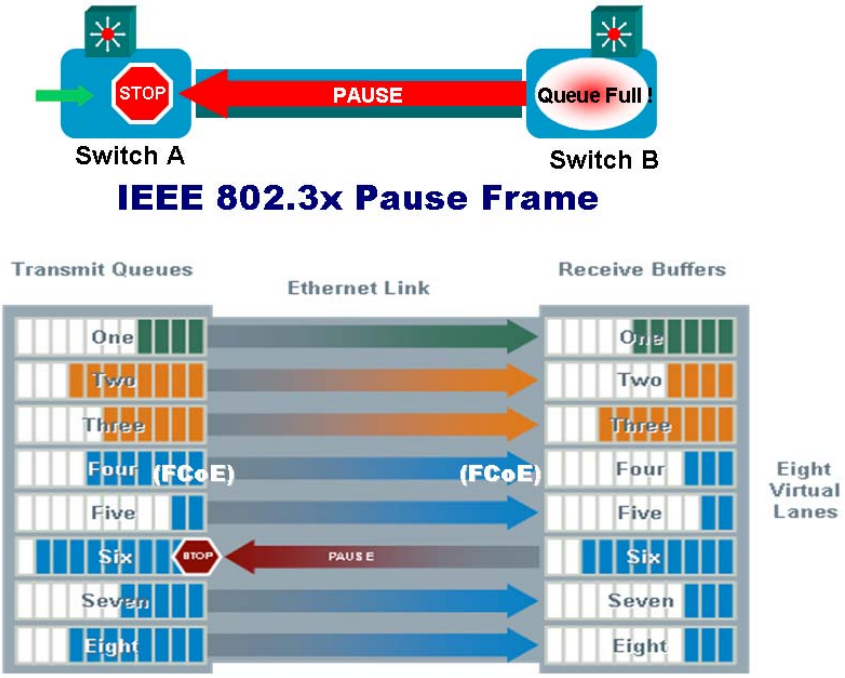

Figure 6
Lossless Ethernet-based WARP cluster

The local storage network connecting the storage to a storage server is much smaller in scale and less complex. It can be designed without oversubscription, thus avoiding the efficiency loss that is described above in the FC-based media storage network. The remaining cluster network is very well defined and has a much simpler topology with a more limited number of devices. Because of the simple well-defined topology and the use of Ethernet as the technology for the TCP/IP based traffic, flows can be very well controlled and the

load balanced over the links. The cluster network is made "lossless" by the application of the IEEE 802.3x PAUSE mechanism Ethernet enhancement *(see top of Fig. 7)*. This provides for a similar link-level flow control as the buffer-to-buffer credit mechanism deployed by FC (or IB).

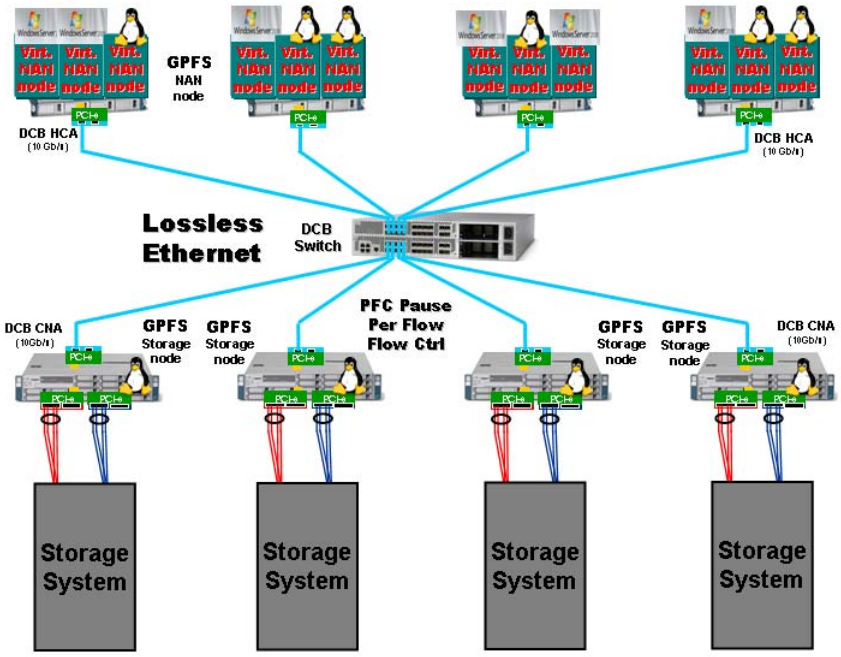

Figure 7
Lossless Ethernet Enhancements

Data Centre Ethernet (or Data Centre Bridging) has another more-advanced flow control mechanism: **Priority Flow Control** (PFC), or **Per Priority Pause** (PPP). IEEE 802.1Q defines a tag which contains a 3-bit priority field. Hence, it can distinguish eight different traffic flows. The PFC flow control mechanism is able to pause traffic labelled with a specific priority or p-value, independent of the other traffic. Each traffic class has its own independent buffers and pause mechanism. Hence, if traffic of one class is filling up its buffer, it can be independently paused, without interfering with the rest of the traffic on the link. The mechanism works the same way as the 802.3x pause but this time selectively-per-traffic-class instead of pausing the whole link at once.

Implementing PFC in the WARP cluster architecture eliminates all traffic interference in the network and thereby creates the near-perfect storage network with absolutely linear scalability *(see Fig. 8)*.

The DCB-based WARP cluster allows for running different operating systems on the physical machines of the processing NAN nodes. Hence, one can design a cluster with Microsoft Windows (Windows Server 2008) on each NAN node, with Linux on each NAN nodes, or with a mixture of

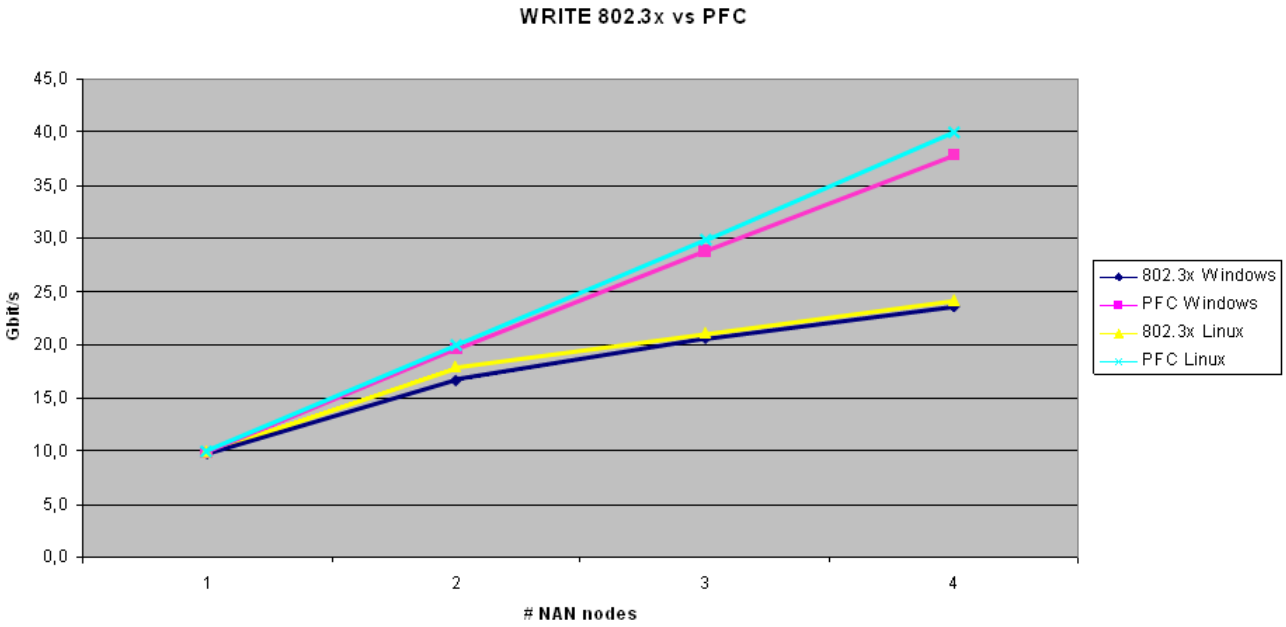

Figure 8
Linear scalability of PFC implementation in the WARP Cluster

Windows and Linux servers as NAN nodes. This allows us to run different media services using different operating systems on the same cluster.

However, we could optimize the utilisation of the resources of these processing nodes by defining multiple virtual machines on the physical NAN nodes, each acting as a cluster node, and as such allowing us to run multiple instances of different operating systems on the same physical machine in the cluster *(see Fig. 6)*. Therefore, one can create a **"Virtual Media Data Centre"** architecture.

Optimized media workflow on a Virtual Media Data Centre

As described above, media services often use local storage of a proprietary nature. Interconnecting all these media services in the classical way leads to an extremely IP-network-centric architecture *(as was shown in Fig. 4)*. While sequential steps of the workflow are being executed, very large media files are exchanged back-and-forth between these islands in an any-to-any traffic pattern. This leads to very inefficient data flows, with many duplicated copies residing in the local storage systems.

Implementing these different media services on the processing nodes of the Virtual Media Data Centre mounted on the clustered central storage would shorten the transport paths and simplify the data flows considerably. Since most of the inter-service media traffic would be using the cluster network, the physical data flows would strongly benefit from the lossless characteristics of such a type of network, leading to a much increased workflow efficiency.

To demonstrate this, a relatively simple workflow will be considered as an example. The same functionality of this particular workflow, implemented on the Virtual Media Data Centre platform, leads to

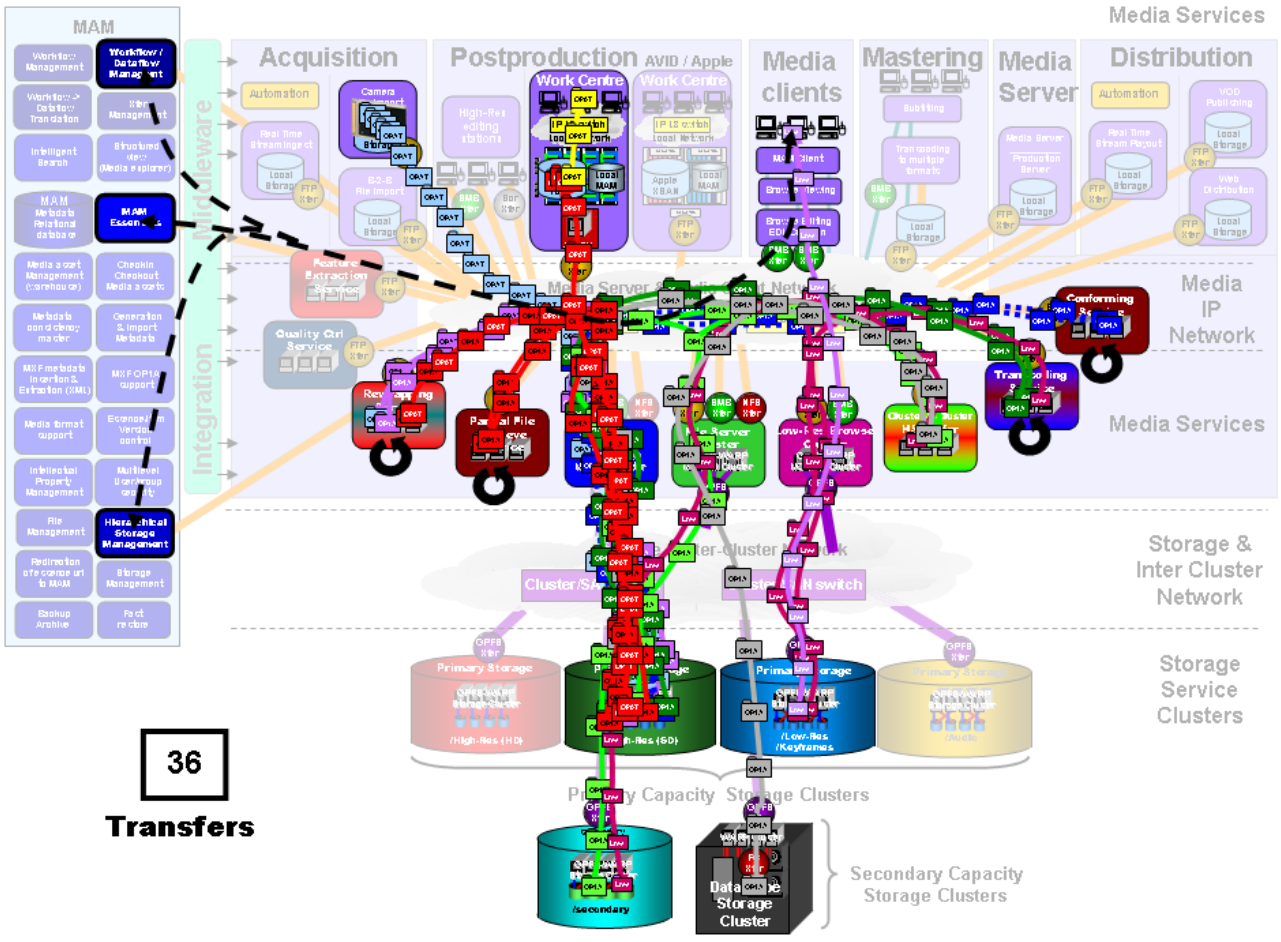

Figure 9
Complex Data Flow Mapped on the Classical IP-Network Centric File Based Production Infrastructure

a much more simplified data flow. The number of file transfers will be decreased by an order of magnitude and the overall execution time will be roughly five times shorter than before.

The workflow example consists of essentially three steps:

1) The material is transferred from a memory card of the file-based camera into the central storage system.

2) A low-resolution proxy is created so that any journalist can view the material and select the relevant clips. The journalist creates an Editing Decision List, EDL, to mark his selection.

3) The system uses this EDL to transport the selected pieces of material to the non-linear file-based editing suite – an AVID Media Composer connected to an AVID ISIS platform in this case. There the journalist, together with the professional editing technician, performs the editing and creates the result again as a media file, or as multiple media files.

Using the media services of the Ardome MAM system, *Fig. 9* shows the actual data transfers required to execute this workflow for each ingested video clip, when mapped on the classical IP-network-centric file-based production architecture of *Fig. 4*. Some of the key transfers are as follows:

- Five transfers, one video file and four audio files, between the memory card of the camera and the temporary folder on the central storage. (Total = 5 transfers)
- Five transfers, one video file and four audio files, to the server responsible for the rewrapping of this media files together into one media file ... and one transfer back to the storage. (6)
- Two transfers from the storage to a conforming server in case the video clip was spread over two different memory cards ... and one transfer back to the storage. (3)
- One transfer to the transcoding engine, to generate the low-resolution proxy version, one transfer of the proxy result to the low-resolution central storage and one transfer of the high-resolution versio back to its final destination on the high-resolution storage, plus two transfers to mirror the file on the disks. (5)
- One transfer to the backup server and another to the data-tape robot for backup reasons. (2)
- One transfer to make a secondary copy of the high-resolution file and another to place the proxy file on a separate storage cluster. (2)
- One proxy transfer for the video selection by the journalist. (1)
- One transfer to the temporary storage location, another transfer to the reverse rewrapping server and five transfers back to the temporary storage location, one video file and four audio files. (7)
- Five final transfers of the video and audio files to the high-resolution non-linear editing work centre. (5)

This leads to a total of 36 file transfers over both the storage and IP network. A total of at least 14 different media services residing on the central IP network were invoked. Although this network is typically Gbit-enabled with even 10 Gbit/s backbone links, packet loss induced by the bursty nature of the media traffic heavily impairs the throughput efficiency to as low as 10% of the theoretically available link bandwidth. This, together with the very large number of consecutive transfers, leads to a very long overall execution time.

The same functionality of this particular workflow, has been implemented on the Virtual Media Data Centre platform *(see Fig. 10)*. The workflow was implemented as follows:

- Files were ingested as OPAT via a NAN node into the GPFS clustered central storage. Immediately after arrival, a hard link was created, linking the high-res media files to the correct directory structure of the AVID project structures. This gave immediate access to the high-res editing clients, without the need for additional moving or copying the files to a different directory.

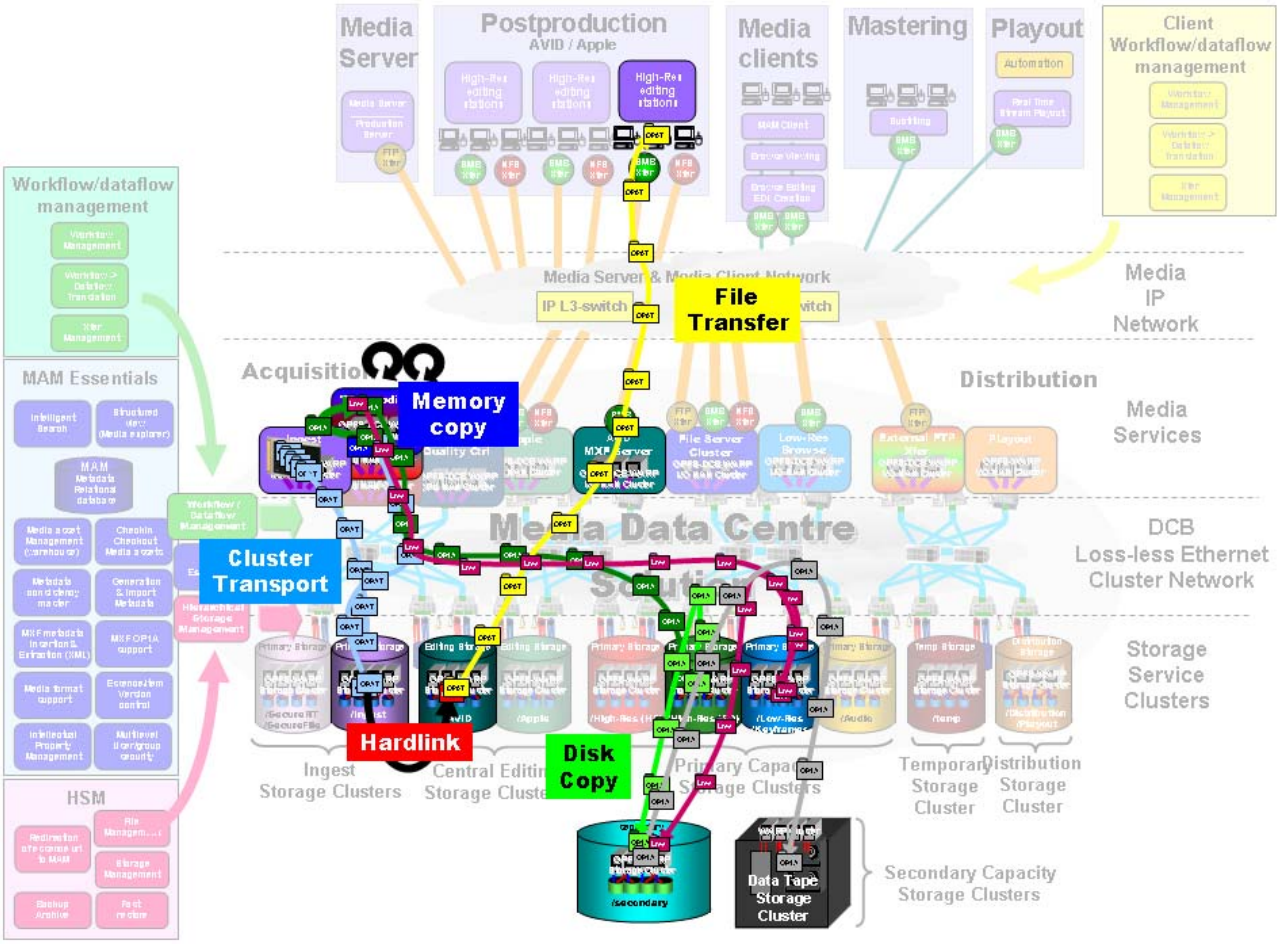

Figure 10
Dataflow mapped on the Virtual Media Data Centre architecture

- Then, the files were read by the rewrapping process running on a virtual Windows NAN node. Rewrapping from OPAT to OP1a was performed at a speed of 5 Gbit/s using only a single core. The lossless 10 Gbit/s cluster network was not the limiting factor in this process.
- Files were being written back to the central cluster directly to the correct final place.
- The files were then passed to a transcoding engine on the Linux virtual machine of the same node. The low-res version was generated by the transcoder and placed into the low-res directory of the central storage. Transcoding was running at 1.2 x real-time speed (using DNxHD 120 Mbit/s HD video).
- Finally, the media item was checked into the MAM system itself.

This workflow implementation clearly demonstrates the following advantages:

- No media left the cluster network during the workflow processes. The IP network was completely offloaded of all traffic.
- All processing steps were performed on the Virtual Media Data Centre making optimal use of the available CPU and memory resources.
- No excess or duplicate copies of the media files were stored.
- No intermediate copies of the files were stored.
- Hard links avoid excess copies of files between directories.
- IP transfers were conducted using the lossless cluster network.
- Being a 10 Gbit/s lossless network, the cluster network wasn't the bottleneck in any of the different workflow steps, contrary to the network in the classical IP-centric workflow of *Fig. 7*.
- High-res material was made available immediately after ingest to the editing clients.

Luc Andries holds a Masters Degree in experimental laser physics. After 15 years of experience in R&D and computer integrated manufacturing in the connector industry, he joined VRT (the Flemish Radio and Television public broadcaster) in 1998, first as a network and storage specialist in the IT department and later as an infrastructure architect in the VRT medialab, the R&D department of VRT. In 2005 he was appointed as top expert in the VRT medialab. His main expertise covers the domain of storage and network technology, such as disk networks, storage area networks (SANs) and IP networks.

As senior architect, Mr Andries presently leads the media infrastructure team at the IBCN research group at the university of Gent, in close cooperation with CandIT-media, a spin-off company of VRT medialab, which specializes in media infrastructure. The research mainly covers the modelling of the capabilities of storage and network technologies. Consequently, it is investigated how that technology can best be applied to the problems posed by the media infrastructure of a broadcaster. The media infrastructure team at IBCN thereby tries to bridge the gap between IT and media.

- No time-consuming transfers of the high-res material to external storage systems were necessary.
- No double wrapping-unwrapping process was required to access the high-res media by the editing clients.
- The number of file transfers was reduced by an order of magnitude.
- The total overall workflow execution time was reduced by a factor of five. The speed-defining factor was the transcoding speed, not limited by the network.

Conclusions

The workflow in media production is very complex and requires the integration of many different media services, needing a continuously-changing capacity to be available to the most critical service resources. Since media traffic is intrinsically different from IT traffic, this causes classically-designed IP networks to behave (unexpectedly) differently under this new traffic load.

An optimal media workflow architecture should provide an integration of both storage and media services into a storage cluster environment, based on a scalable virtual platform: the **Virtual Media Data Centre**. The possibilities created by the addition of the lossless and quality enhancements of Data Centre Bridging, puts the Ethernet-based network at the centre of this infrastructure.

This version: 5 October 2010

Published by the European Broadcasting Union, Geneva, Switzerland

ISSN: 1609-1469

Editeur Responsable: Lieven Vermaele

Editor: Mike Meyer

E-mail: tech@ebu.ch