Assuring quality of service in the Columbus Ground Segment Network

Stefan Maly [1], Gábor Szücs[1]
Telespazio Deutschland GmbH, Talhofstraße 28a, 82205 Gilching, Germany

and

Dr. Osvaldo Peinado[2]
DLR – Münchner Straße 20 - 82234 Weßling, Germany

The ISS Columbus Ground Segment Network is a complex communication network which connects sites located in the USA (NASA), Russia (RSA), France (ATV-CC), Germany (COL-CC) and several user centers all across Europe. For the WAN communication between the different control centers and facilities MPLS technology is being used while the LAN setup employs Ethernet technology.

One of the challenges within the IGS network is the simultaneous transport of different datastreams like TM/TC, voice, video, science data, etc., with separate and sometimes competing needs for quality of service parameters. While travelling from end to end, the datastreams cross different realms of the network. In addition limitations have been imposed by the wide area network provider. Some of those, well known in theory, have proven to yield surprising effects in reality.

While most of the limitations were known at the design phase of the overall structure, a few have revealed themselves later during the test and implementation phases and had an impact on operations.

The network in its present design is being used for more than two years.

This article will present the building blocks and design parameters that shaped the setup of the network as it is being used today. Unused alternatives will be shortly discussed and the reasons for the choices that lead to the current setup will be given.

A short outlook of the future development of the network will be presented together with a discussion of the limitations and consequences that cost driven technology changes imply.

[1] Network Engineers IGS, Telespazio Deutschland GmbH, Talhofstraße 28a, 82205 Gilching, non AIAA Members.
[2] Ground Operations Manager, Communications and Ground Stations, DLR – Münchner Straße 20 - 82234 Weßling, Germany, non AIAA Member.

Acronym List

AC	=	ApplicationClass
AF	=	Assured Forwarding
AH	=	Authentication Header
BRI	=	Basic Rate Interface ISDN circuit with two data channels and one signaling channel
CDN	=	Content Delivery Network
CE	=	Customer Edge
CS	=	Class Selector
DSCP	=	Differentiated Services Code Point
EF	=	Expedited Forwarding
ESP	=	Encrypted Security Payload
GPC	=	GeneralPurposeClass
GRE	=	Generic Routing Encapsulation
HSRP	=	Hot Standby Redundancy Protocol
IGS	=	Interconnected/Interconnection Ground Subnet
IKE	=	Internet Key Exchange
IP	=	Internet Protocol
IPsec	=	Internet Protocol Security
ISDN	=	Integrated Services Digital Network
MAC	=	Media Access Control
MPLS	=	Multi-Protocol Label Switching
MSS	=	Maximum Segment Size
OID	=	Object IDentifier
OSPF	=	Open Shortest Path First
PABX	=	Private Automated Branch Exchange
PE	=	Provider Edge
PRI	=	Primary Rate Interface of ISDN (2 Mbit/s in Europe / 1.5 Mbit/s in the US)
QOS	=	Quality Of Service
RFC	=	Request For Comments (an internet protocols suite standard)
RTC	=	RealTimeClass
SLA	=	Service Level Agreement
SNMP	=	Simple Network Management Protocol
SPI	=	Security Parameter Index
TCP	=	Transmission Control Protocol
TDM	=	Time Division Multiplex
TM/TC	=	Telemetry/Telecommand
UDP	=	User Datagram Protocol
VC	=	VoiceClass
VoIP	=	Voice over Internet Protocol
VRF	=	Virtual Routing and Forwarding instance
WAN	=	Wide Area Network

I. Introduction

THE Interconnected Ground Segment is a MPLS based communication network which connects the control centers located in the USA (NASA), Russia (RSA), France (ATV-CC), Germany (Col-CC) and several user centers all across Europe. The part of the network maintained by Columbus personnel is called IGS Relays and Nodes while the part maintained by the wide area network provider is called IGS WAN.

The network structure matches the main dataflows. It is hub and spoke with the Columbus Control Center (Col-CC) at Oberpfaffenhofen, Germany as the hub. Another possible network structure would have been a full mesh, but it was decided against that solution because data sparsely flows directly between the spoke endpoints. From an administrative perspective hub and spoke was easier to maintain.

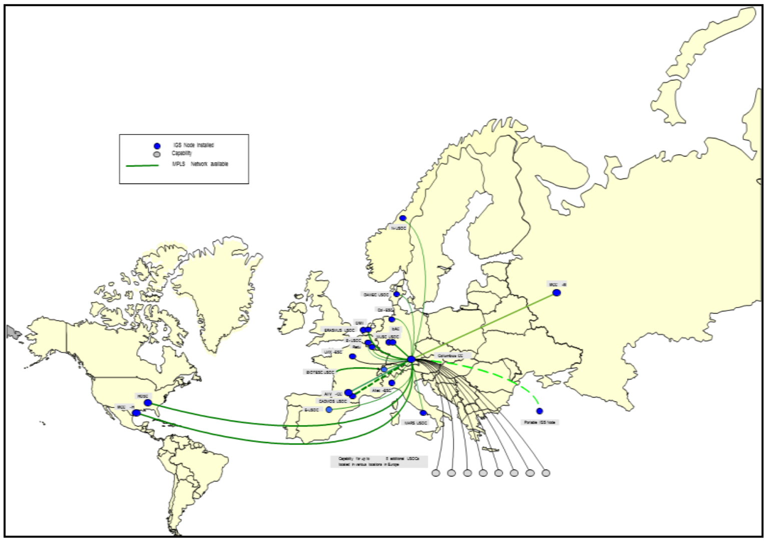


Figure 1 - Overview of the Columbus Ground Segment Network

The data being transferred across the IGS network is high level divided into three logical subgroups: Operations, operations-support and highrate[3]. The names hint at the logical type of data. Within each of these subgroups concurring datastreams are being transferred. Inside the LAN part those have been separated into different networks and even split up onto different hardware. In the backbone the subgroups are separated via virtual routing and forwarding instances (VRFs).

As the bandwidth locally available is much higher than the sum of all datastreams being processed, the separation of the logical subgroups via different hardware has proven to be enough to ensure the proper delivery of the data. At the point where the data enters the realm of the wide area networks the situation is different. Due to the cost structure of the wide area network providers, bandwidth is limited throughout the entire MPLS part of the network. Traffic coloring via DSCPs is used in order to provide the necessary handling information to the WAN provider. When entering the CE equipment all possible DSCP values are accepted. However the combinations decrease rapidly as the provider reserves certain values for internal purposes. Therefore, before the packets are being passed on from the CE to the PE equipment, some get their DSCP values remapped. This is laid out in more detail in the chapters "II The life of a packet in the IGS network" and "III Multicast, Encryption and DSCP values".

Between CE and PE equipment the packets still are not MPLS tagged. That happens when they enter the backbone area. Further restrictions apply. Within the backbone, the provider hosts many customers in parallel and has restricted the granularity of possible traffic classes to just four different ones: VoiceClass (VC), ApplicationClass (AC), RealTimeClass (RTC) and GeneralPurposeClass (GPC). This is not to be mistaken with the lack of separation. Through proper configuration of the edge and backbone devices it is still made sure that data from one customer does not end up at another customer; however all data of all customers flows in parallel through the MPLS backbone and is being transferred across the same lines which effectively means that while being separated per customer, they still compete for the available service levels.

[3] While this may sound tiny in comparison to local area network figures, within the worldwide IGS network, data at a speed of 27-32 Mbit/s is considered highrate.

II. The life of a packet in the IGS network

When entering the CE equipment, the packets are being examined for their DSCP values. The three different subgroups described above use different combinations of those. Common to all three subgroups is the use of CS2 and CS6 for management data like routing protocols and for key exchange. Within the subgroup operations, EF is being used for voice data and AF21 for privileged data. Regular data is being flagged default. Within the subgroup operations-support AF41 is being used for video streams, everything else – besides the above mentioned CS2 and CS6 datastreams is flagged default. Highrate utilizes AF41 for its datastreams, nothing else besides CS2 and CS6 is used in addition. This is a special case as default exists but is not being used. Reason for that is the fact that the provider does not allow for 100% of the ingress traffic to be privileged. The limitations imposed by the service provider are discussed in further detail in the chapter "IV Limitations imposed by the service provider".

As CS2 and CS6 are reserved by the wide area network provider, those values just exist in between the IGS Relays and Nodes and the CE equipment. Once they enter the CE equipment these values are being remapped. CS6 becomes AF22 and CS2 becomes AF23. While travelling from the CE equipment to the PE equipment, those packets are thus not being handled much differently from other privileged operations traffic.

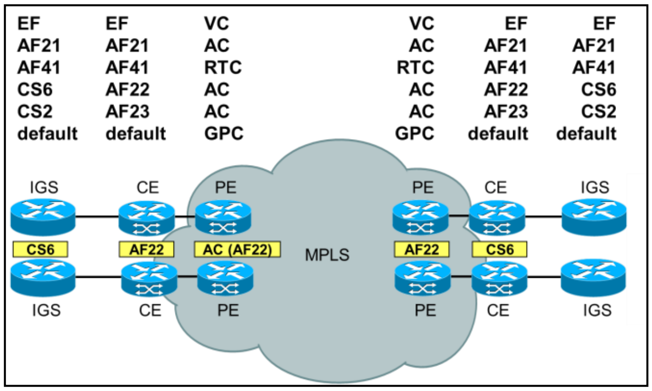


Figure 2 - Remapping DSCP values and Traffic Classes

At the PE router the MPLS network starts. The original packet becomes payload of the MPLS packet. Thus the original IP header with its DSCP value is encapsulated and preserved. Granularity of traffic distinction decreases further as just four backbone classes of traffic remain.

All DSCP AF2x packets are being mapped into the backbone ApplicationClass (AC) while all DSCP AF4x packets are being mapped into the backbone RealTimeClass (RTC). Voice packets with a DSCP value of EF become backbone VoiceClass (VC), everything else becomes backbone GeneralPurposeClass (GPC). While it is still made sure through configuration that none of those packets end up on the CE devices of a different customer, within the backbone those packets are being transferred across the same lines as the packets of other customers of the wide area network provider.

At the other end of the MPLS network the original IP packet is being extracted from the data part of the MPLS packet. The DSCP information still exists and is being used on the way to the CE equipment. As on the way between the CE equipment and the IGS equipment the DSCP values CS2 and CS6 become available again, packets with a DSCP of AF22 are being mapped back to CS6 and those with AF23 to CS2. At this point the original DSCP information becomes available again and the packet leaves the wide area network just as it has entered it on the other side.

III. Multicast, Encryption and DSCP values

Multicast is not supported by the wide area network provider[4]. As the provider does not support multicast, the according packets, as for example OSPF messages, need to be converted into several parallel unicasts before being sent on to the CE equipment. This is being achieved in the Relays and Nodes part of the network via GRE tunnels which have been set up between all locations.

Additionally there exists a requirement to encrypt all traffic which leaves the Relays and Nodes part of the network. Therefore packets are being encrypted as they leave the IGS routers towards the CE routers. There existed different choices for the key distribution scheme. At the time the decision was made however the automated key distribution was not considered stable enough to implement, instead a manual key distribution with shared secrets

[4] Actually, at the time the network was migrated from ATM into its present setup, none of the wide area network providers was willing to support multicast. Therefore utilizing multicast in the WAN never was an option.

was employed. This would not have scaled well with a full mesh but with the chosen hub and spoke structure, key management is no problem. The encryption scheme being used transfers the entire original IP packet into the data portion of the encrypted packet and generates a new IP header for the encrypted packet with just the IP addresses of the two IGS router interfaces at both ends. The rest of the packet becomes unreadable without the proper key.

During both steps, while encapsulating with GRE and when encrypting the entire packet, the outer IP header is newly composed. In the first step the source and destination IP addresses of the GRE tunnel become visible while the original IP address information resides within the data portion of the GRE packet. In the second step the GRE packet becomes the data portion of an encrypted packet which shows in its IP header just the outer interface addresses of the IGS routers towards their respective CE equipment as IP source and destination.

Due to the double encapsulation of the original IP packet the traffic coloring information would become lost were there not the possibility to configure the devices in a way that they duplicate the original DSCP value while encapsulating. This way it is made sure that the original DSCP value shows up both in the header of the GRE and of the encrypted packet.

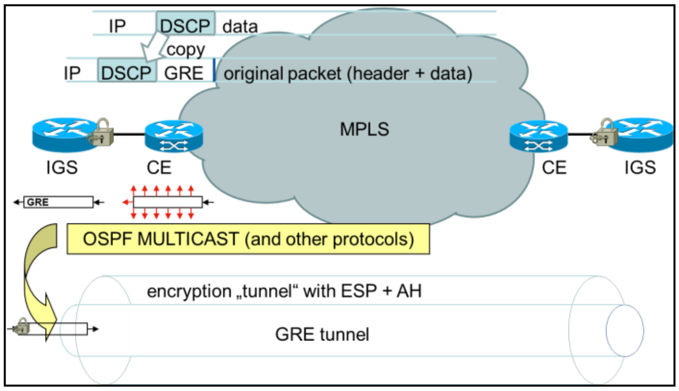

IP
DSCP
data
copy
IP
DSCP
GRE
original packet (header + data)
MPLS
IGS
CE
CE
IGS
GRE
OSPF MULTICAST (and other protocols)
encryption „tunnel“ with ESP + AH
GRE tunnel

Figure 3 - Multi Encapsulation and Traffic Coloring

The provider acts accordingly on the DSCP values seen in the encrypted packet, not even knowing what type of data is hidden within. In an overflow condition where more traffic is being poured into a VRF than allowed by the SLA, the excess is being dropped. What exactly gets dropped is governed by the DSCP values. As long as the limits for privileged traffic are not violated, only lower prioritized packets are being dropped. Of course if the privileged traffic is overbooked, even those packets have to be dropped. At this point it has to be kept in mind that due to the double encapsulation and the encryption no “intelligent” dropping mechanism at the provider equipment is possible. Not even information like the stream identifier of the original IP packet is available.

It should also be noted that the double encapsulation comes at a price. The maximum segment size (MSS) inside a GRE tunnel being transferred within a crypto tunnel has been reduced to just 1300 bytes. This allows besides the additional space needed for the double GRE and IP/AH/ESP encapsulation for a couple of additional IP and TCP options. One out of those TCP options is especially important: To facilitate the transition from 1500 to 1300 bytes, a mechanism in the transmission control protocol (TCP) has been used. TCP option number four (TCP:4 – MSS – see RFC1122) allows for setting the MSS upon initialization of a session. The IGS routers add or adapt the TCP MSS option during the three way handshake accordingly with a value of 1300 bytes. Thus for TCP, the end devices do not even have to be set up in a special way. This comes out of RFC793 Section 3.1, which says that every TCP/IP stack must implement TCP option four. For the user datagram protocol (UDP) a similar mechanism does not exist. Applications utilizing UDP have to be configured explicitly in a way that their datagrams won’t exceed 1300 bytes. If this is not done, the applications still work, but fragmentation occurs and the efficiency of the transfer network decreases.

IV. Limitations imposed by the service provider

Some of the limitations were already mentioned. The service provider limits the number of different DSCP values within the area between CE and PE equipment and allows just four different traffic classes to exist within the MPLS backbone. Additionally the max bandwidth for the traffic classes is being limited as it is paid for on a monthly basis. Bandwidth is being regulated in a multi-step approach. First of all there is a limitation of the overall bandwidth on the local loop[5]. Second there is a limitation of the overall bandwidth within each VRF. Last but not least there is a limitation of each defined traffic class within each VRF.

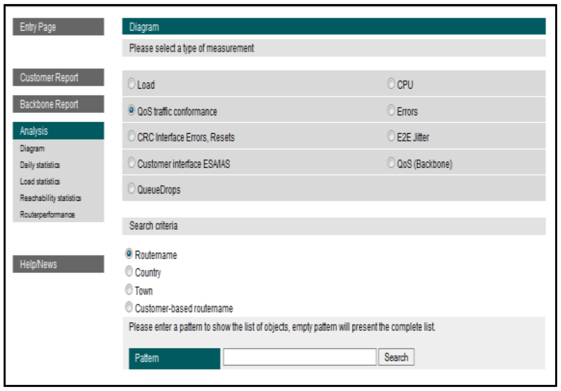
[5] The term “local loop” in this context designates the network between CE and PE equipment.

Also, the provider has imposed limitations as of how many per cent of the overall bandwidth of each VRF may be used for privileged data. The idea behind this is that if there is no low priority data to be pushed out of the way, the concept of high priority data becomes meaningless. For sure in the backbone it makes no sense to have one customer who transfers just data flagged as high priority. This would effectively take a big share out of the available bandwidth for all other customers, which could in turn lead to a situation where everyone uses high priority and flagging data as high priority does not ensure a reliable transfer of that data across the backbone for anyone any more.

At certain locations there exist fiber extensions from the CE equipment to a secondary site. This is the case when two sites are located close to each other and purchasing a second local loop is far too expensive in comparison to a solution where a campus network or something comparable can be used. The limitation from the service provider side in those cases is, that within their network equipment there exist no separate bandwidth limitation schemes for the two collocated sites. Instead it is the duty of the IGS Relays and Nodes network team to make sure none of the two locations can overbook. In order to achieve this, cascaded policy maps have been established at the hub and traffic limitations at the remote locations. An automated Excel™ based solution exists which acts as a general tool for bandwidth calculations and at the same time automatically generates the necessary commands for the fiber extension sites.

V. Troubleshooting QOS issues

The backbone part of the network is only visible via a web based portal application. QOS data is being collected and made accessible via selectable submenus. When analyzing QOS problems, a good starting point is a submenu for policy monitoring within the provider network. Selecting a privileged datastream from the available choices allows for the display of a graph with the current bandwidth distribution.



Please select a type of measurement

○ Load
○ CPU
● QoS traffic conformance
○ Errors
○ CRC Interface Errors, Resets
○ E2E Jitter
○ Customer interface ESA/AS
○ QoS (Backbone)
○ QueueDrops

Search criteria

● Routername
○ Country
○ Town
○ Customer-based routername

Please enter a pattern to show the list of objects, empty pattern will present the complete list.

Figure 4 - The service portal

The graph tells whether the traffic has been colored correctly and thus ended up in the right class. Overbooking becomes visible in form of a flat line at the top of the graph, exactly at the height of the configured max value for that type of data. The graph is near real-time. Depending on the type of graph (CE/backbone) the data is collected on a per minute or on a per 5 minute basis. It has to be taken into account that short spikes will not show up in the graph as it is showing the one/five minute average. Another small effect is the fact that the scheduler which collects the values from the corresponding SNMP object identifiers (OIDs) is not 100% accurate in terms of measurement times. It might well collect a value a couple of seconds earlier or later than the specified time, which results in waveform patterns being displayed where a flat line of measurements is expected. The way those diagrams have to be read is, that curves which have the same amount of deviation below the line as they have above the line are considered to be flat and therefore do not point towards a problem. Curves having more deviation in one of the two directions are considered to point towards a problem. To put it most simple: “Symmetrical patterns are good – asymmetrical ones are bad” (see Figure 5 - Monitoring a datastream). While this is something the network engineer needs to get used to, on the other hand the number of parameters monitored is quite impressive.

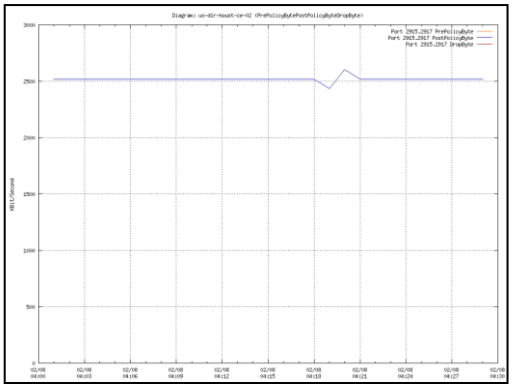


Figure 5 - Monitoring a datastream

Even parallel policy mappings for the different classes of data are available, making it easy to find bottlenecks and bandwidth misconfigurations. In the case of a fiber extension location, a secondary analysis is needed where the cascaded policy maps within the IGS routers are being examined. If the output of the web portal is not detailed enough, there exists the possibility to connect to the command line of the CE routers. In there the policy-maps can be displayed on a per 30 second basis. It is not possible to do the same on the PE equipment or in the case of CE switches.

VI. Lessons learnt in Two Years of Operations

What seems stunning on first glance is the real extent of how much the percentage limitation for privileged data in a VRF impacts monthly cost. For example in order to reliably transfer a 32 Mbit/s highrate datastream, an overall bandwidth of 55 Mbit/s within the highrate VRF has to be purchased. This high overall bandwidth has to be purchased in order to keep the percentage of high priority traffic at bay. But calculating the according percentages does not alone explain the extremely high overhead. In addition there are other factors influencing the transfer of highrate data.

If data is being transferred via TCP, the window size does not start at a high value right away. This generates an overhead of buffered data that on top of the regular data rate has to be transferred in parallel in order to catch up with the initial slow start. Also when packet losses appear – and a certain amount of packets are allowed to be lost, even within the most expensive SLA – the Nagle algorithm[6] will slow things down as well. This can be mitigated by having selective acknowledge and forward acknowledge enabled and by tuning the parameters for initial TCP window size and buffer sizes for data reassembly, however the overhead of additional data which needs to be transferred in parallel to the ongoing data stream still remains.

UDP data does not have the problem of the slow start as mentioned above for TCP with a smaller initial window; nevertheless the packet loss problem still applies. If an application cannot afford to lose a certain amount of data, a watchdog mechanism needs to be included. This will request a retransmission of missing data, thus still adding to the overall bandwidth.

Another interesting effect surfaced concerning voice data. A noticeable increase in jitter appears when in parallel there is a high fluctuation in the amount of other prioritized data being sent. Of course the pure theory states that the data flagged EF is transmitted immediately and without delay, in reality it seems that there exists a queue just like in any other case. It's just that sending from this queue does not require possessing a token.

The lesson learnt here is that configuring a buffer which just fits the regular jitter measurements for the line will eventually fail due to fluctuations within the other datastreams. Within the IGS the jitter buffers have been almost doubled to be on the sure side. Luckily this still allows for staying within the max delay values possible for pleasant voice communications (< 150ms).

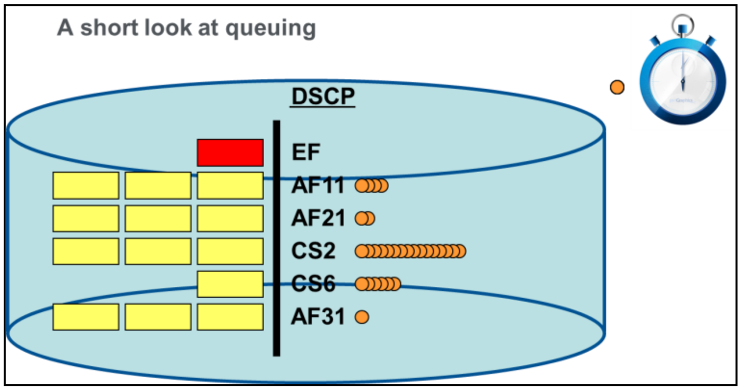


Figure 6 - Queuing

A very special situation occurred when suddenly no data was being transferred across the MPLS network any more while no error was detected by the wide area network provider. This had to do with the fact that the MPLS paths are being negotiated in a different plane than the data transfer plane. This way the routing information was still intact and even control traffic from the network provider was still present while the label distribution protocol did not function properly for a couple of connections any more. All data being sent across the affected MPLS paths under those circumstances gets lost. A detailed analysis of this very special scenario you can also find in a poster presentation that exists on this very topic: “Communication Black Holes in Ground Segment Networks"

[6] RFC 896 „Congestion Control in IP/TCP Internetworks“

VII. Future Network Changes

In order to minimize operating and maintenance cost, the IGS endpoints have to be simplified. While redundancy at the control center locations remains, mapping the subgroups operations, operations-support and highrate onto separate hardware will be ended. Instead all dataflows will be concentrated on the same hardware. Besides more powerful hardware, which will be installed as systems become deprecated and reach their end of service time, implementation of dot1q priority mapping[7] within the local networks might become necessary in order to prioritize certain traffic. Whether or not dot1q needs to be used in the future is subject to tests carried out in a test environment at Col-CC. The number of VRFs in the provider network will not be affected by this as it will still be necessary to separate the dataflows of operations, operations-support and highrate.

Also processing power which resides within the locations at the spokes of the hub and spoke structure will be centralized at the hub. This way superfluous datastreams which result from transferring unprocessed data across the wide area network can be eliminated. In addition for many tasks having a remote desktop connection to the processing engines is fully sufficient. The extract of this processing can then be downloaded offline as a low priority datastream.

The usage of a content delivery network for transferring video information to the remote locations is being taken into consideration. It is clear that a reliable transfer across the internet is not possible unless a CDN is being employed. Offloading the video bandwidth from the MPLS can significantly reduce needed bandwidth and as a result monthly recurring cost. However a CDN also costs money. A detailed analysis will show whether money can be saved while the flexibility in video distribution increases.

[7] This kind of mapping utilizes a layer2 mechanism for prioritizing data. The corresponding field is introduced directly behind the source MAC address in the Ethernet frame. While this ensures that also switches which are unaware of this function still are able to forward the packets, the mechanism has to be supported for the entire chain of switches and routers in order to function properly.