RAVENNA
AES67 & ST 2110 built-in

White Paper

Author: Andreas Hildebrand,
RAVENNA Technology Evangelist

"Your flight will be delayed by 5 ms" Understanding Latency in AoIP Systems

What is latency and where does it come from? Why does latency vary? What minimum and maximum latencies can be expected? This White Paper will provide answers to all these questions and explain buzz words and abbreviations such as packet time, PDV, software jitter and more...

ABSTRACT

The scene is set by explaining the fundamental differences between circuit-switched and packet-switched networks — in other words, the principal difference between traditional analogue & digital and network-based transport mechanisms. While circuit-switched networks basically are only affected by the speed of light, and thus latency is a direct result of distance, packet-switched networks add several more factors to the calculation. Even more complex, most of these additional factors are highly dynamic, depend on other factors or can be configured to match particular applications or use cases. In this article, we look at important factors like network technology and topology, stream and packet configuration, device implementation and many more aspects. It also describes the highly dynamic effect of other network traffic with respect to latency.

While all of these factors appear to imply that IP-based media transport technologies are not applicable to use cases with low latency requirements, readers will learn how to counter these variable factors and what latencies are actually achievable, depending on device capabilities and configuration.

Note: This White Paper is based on a presentation held by the author at 32. TMT 2023 in Dusseldorf.

Content

1. SETTING THE SCENE

1.1 WHAT IS LATENCY?

"Latency can be defined as a time delay **between** the cause and the **effect** of a physical change in the system being observed. It is a **consequence** of the **limited velocity** at which any physical interaction can propagate [...] which is always less than the **speed of light**."[1]

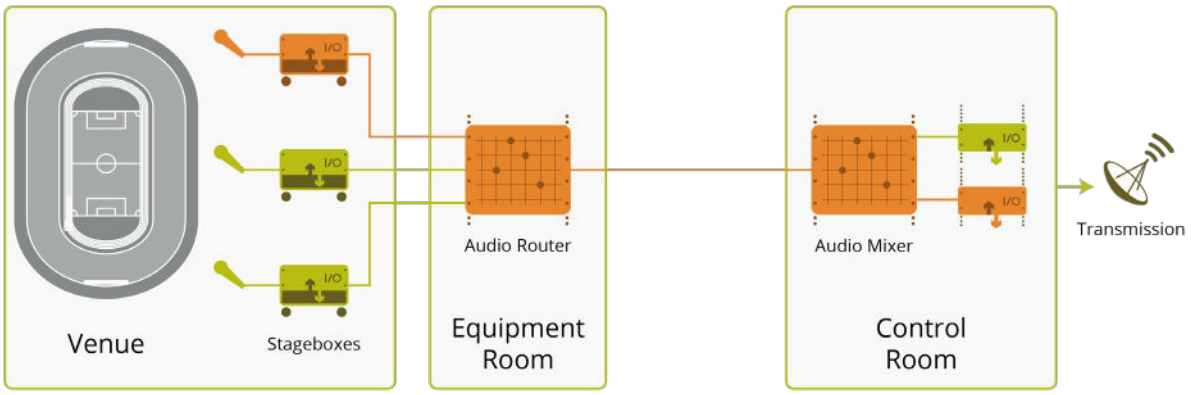
In media distribution systems, latency is commonly defined as the delay between the origination or capturing of an audio-visual event and its reproduction at the desired end point (i.e., a speaker, a screen or a processing device). In short, the time it takes for a data packet to go from one place to another. In general, we seek to keep the latency at a minimum, unless a particular application requires specific delays (i.e., time-aligned playback between independent speakers or lip-sync between audio and video etc.). For the remainder of this article, we discuss latency in audio distribution systems, but the majority of aspects apply equally to video distribution systems.

1.2 "TRADITIONAL" AUDIO SIGNAL DISTRIBUTION

Traditional audio signal distribution is based on a wired approach where designated cables are run between participating devices. In the analogue domain, purpose-built copper cables with varying diameters, lengths and connectors are used. In analogue installations, signal quality unrecoverably decreases with distance, limiting the span on individual connections depending on physical and electrical characteristics. While signal degradation over distance is also inherently part of the digital domain, signals can be recreated without loss as long as the digital information can be unambiguously recovered.[2] The digital domain also includes fiber as a transport medium, adding the capability to span larger distances between devices or to increase the signal channel count on a single "connection"[3]. Typically, digital transport options define the transport medium (the physical layer) and the signal format (how signal information is modulated/encoded); examples for digital transport formats are AES3, AES10 (MADI) or SDI (includes video, audio and ancillary data in a multiplexed format).

Both transport domains are based on so-called **"circuit-switching"** which means that for transport of information between any two devices a dedicated path is established (usually by running a dedicated connection, i.e. a fiber or a copper cable). Since signals may cross processing or routing devices (matrices or cross-point switchers), multiple connections in a row may be used to transport a signal from its origin to the designated destination. A classic example of analogue circuit switching is a POTS[4] switchboard:

They represent a fixed connection between sender and receiver, even if the signals pass through several intermediate devices. An example of a typical digital broadcast live production system is illustrated below:



fixed connection between microphone and mixer output

fixed / deterministic latency

= circuit switched routing

Example of Digital Signal Distribution in a Live Broadcast System

[1] https://en.wikipedia.org/wiki/Latency_(engineering)
[2] For example, recovery methods in the digital domain include FEC, redundant transfer and other methods.
[3] Employing time-division multiplexing (TDM) or frequency-division multiplexing (FDM)
[4] POTS = Plain Old Telephone Service

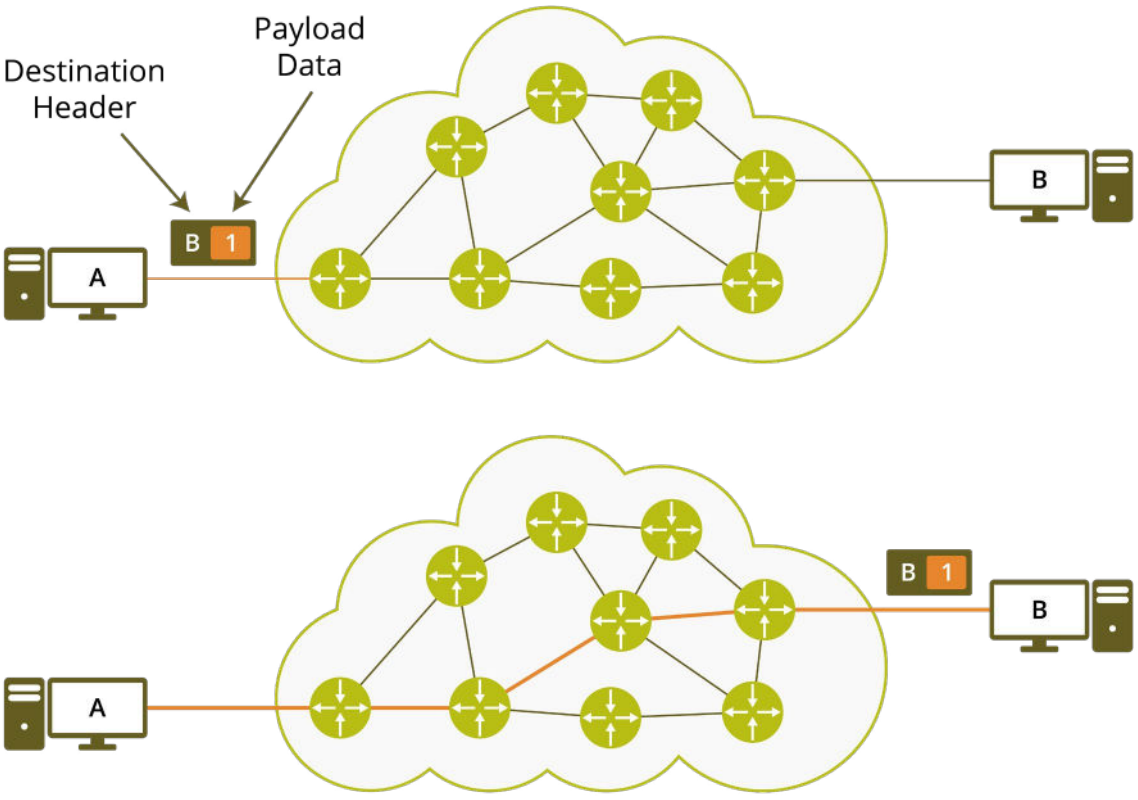
The benefit of such an approach is the (usually) fixed and determinable latency between the input device and the designated destination. Neglecting the individual processing delays of intermediate devices — which of course significantly contribute to the overall latency but are irrelevant in the context of differences between traditional and network-based signal distribution as they apply likewise to both domains[5] — the latency depends solely on the overall distance a signal must travel, which in turn depends on the speed of light (for both fiber and copper media):

Distance	Time
1 m	3.3 ns
1 km	**3.3 µs**
6 km	**20 µs (1 sample @ 48 kHz)**
300 km	1 ms (mandatory AES67 packet time)
4000 km (NYC – LA)	13.3 ms

Approximate Light Signal Travel Times

1.3 NETWORKED-BASED AUDIO DISTRIBUTION

A network constitutes a general-purpose distribution system in which the transport infrastructure is separated from the data content to be distributed. The data to be transported is partitioned and wrapped into packets which are forwarded using various protocols. This type of data transport is called **"packet switching"**. The most important difference to circuit-switched systems is that a physical link does not establish a dedicated connection between a sender and receiver, nor does it imply a particular data format. In other words: plugging a network cable into a jack does not establish a connection with a dedicated receiver, nor does the physical type of cable imply a particular data format which can be transported on this link. Consequently, the most important information required to successfully transport packets across a network are their origin and destination addresses:

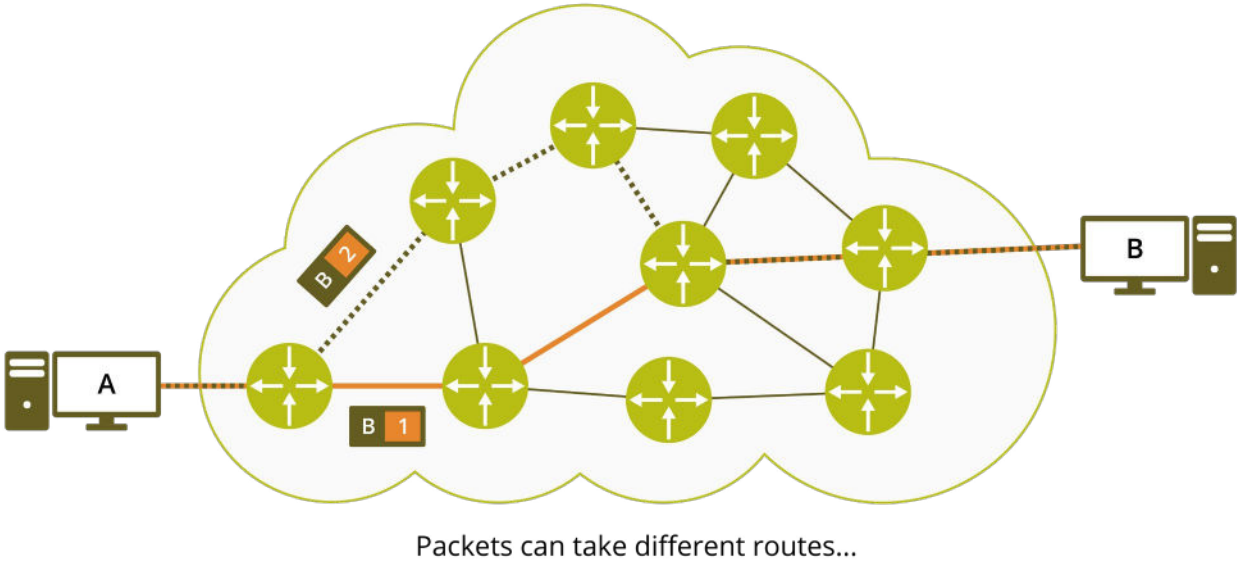


Principle of IP Packet Switching

[5] "Processing" in the digital domain includes necessary signal conversion (i.e., A/D or D/A conversion) as well as signal encoding into the desired format (i.e., data compression).

Further information is typically used to establish particular handling and forwarding behavior by the network infrastructure devices (i.e. information on size of packet, particular routing information, forwarding priority etc.). Since the network is agnostic to the type of data transported in a packet, the receiver requires additional information on the content (the type of the encapsulated data) in order to properly process it. A real-life analogy is the parcel transport service provided by public shipping carriers (i.e., DHL, UPS etc.):

Just as in this real-life analogy, depending on the network structure, individual packets may even take different routes between the same sender and receiver:



Different Routes in a Meshed Network

In summary, networked-based media transport is specific to these characteristics:

- Content-agnostic general purpose transport infrastructure
- Physical connection separated from logical connection
- Flexible bandwidth utilization — signal bandwidth is not constrained to a specific media link technology anymore, multiple signal flows can use the same link
- Bandwidth capabilities increasing with underlying network technology
- Highly adaptable to new media standards (new formats only require new payload format definitions)
- Requires dedicated interfaces at the end devices only (i.e., no special-purpose routers anymore)
- But: no deterministic transport latency anymore... because:
  - There is no fixed connection — each packet may take different routes to reach its destination
  - Each connection can be used by multiple flows, loads on a particular link may vary dynamically
- Thus: transport latency typically varies

The following section further details those characteristics and describes what methods are used to reduce or constrain the negative effects of varying latency to achieve a stable and deterministic signal transport on IP networks.

2. LATENCY IN NETWORKED MEDIA SYSTEMS

In networked media systems, latency is a result of various static and dynamic factors:

- Underlying network technology (mostly the network speed) and topology
- Network jitter or packet delay variation (PDV)
- Configuration of a particular flow (the specific packet setup)
- Device implementation
- Any additional delays to achieve (playout) alignment amongst related streams

2.1 NETWORK TECHNOLOGY

In most cases Ethernet will be used as the underlying technology. Ethernet technology is available in different speeds, with Gigabit Ethernet (GbE) being the predominant technology today. GbE provides a raw data throughput of 1 Gbit/s. The Ethernet protocol itself uses a small amount of this bandwidth for Ethernet frame identification and separation (preamble + inter-frame gap), so that a maximum bandwidth of 974 Mbit/s is available for data being transported on a GbE link[6].

Still in use for less demanding throughput are devices with 100 Mbit/s "Fast Ethernet" (FE) interfaces. On larger networks or in (end) devices requiring larger throughput (i.e. backbone switches or video devices) 10 GbE or even faster Ethernet technologies up to 400 GbE are also common.

The minimum forwarding time for an Ethernet frame with the maximum size allowed in a modern switch is as listed in the table below:

Network technology	Network speed	Frame transmission time (MTU)
FE	100 Mbit/s	258,58 µs
GbE	**1 Gbit/s**	**25,86 µs (~ 1 sample time @ 48 kHz)**
10G	10 Gbit/s	2,6 µs
40G	40 Gbit/s	10,4 µs
100G	100 Gbit/s	260 ns

Minimum Forwarding time per Ethernet switch

As a rule of thumb, the minimum forwarding delay for an Ethernet frame on a GbE network is about 1 sample time (@ 48 kHz) per "hop"[7].

Those numbers apply to switches with "store-and-forward" technology and hardware switching which is the predominant technology of modern switches. Other switching technologies[8] may increase or decrease performance, in particular if switching involves software-based algorithms which may be used in routers or firewall devices with advanced processing or filtering functions (i.e., deep packet inspection).

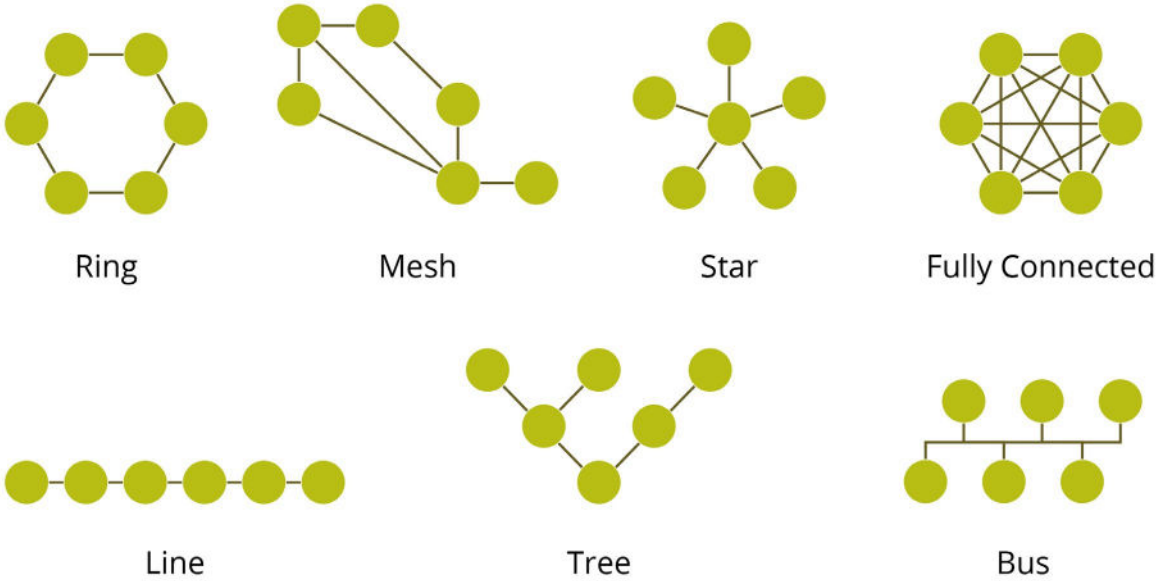
2.2 NETWORK TOPOLOGY

As previously mentioned, the number of hops a particular stream of packets must perform contributes directly to the minimum delay. Consequently, the network topology may have significant impact: shortest delays can usually be expected in star-shaped topologies as the number of hops between any two devices is minimized.

[6] Note that this number does not include Ethernet frame, IP, TCP or UDP etc. overhead and is valid for Ethernet frames with a maximum payload size of 1500 bytes (MTU). Shorter packets would decrease the bandwidth accordingly.

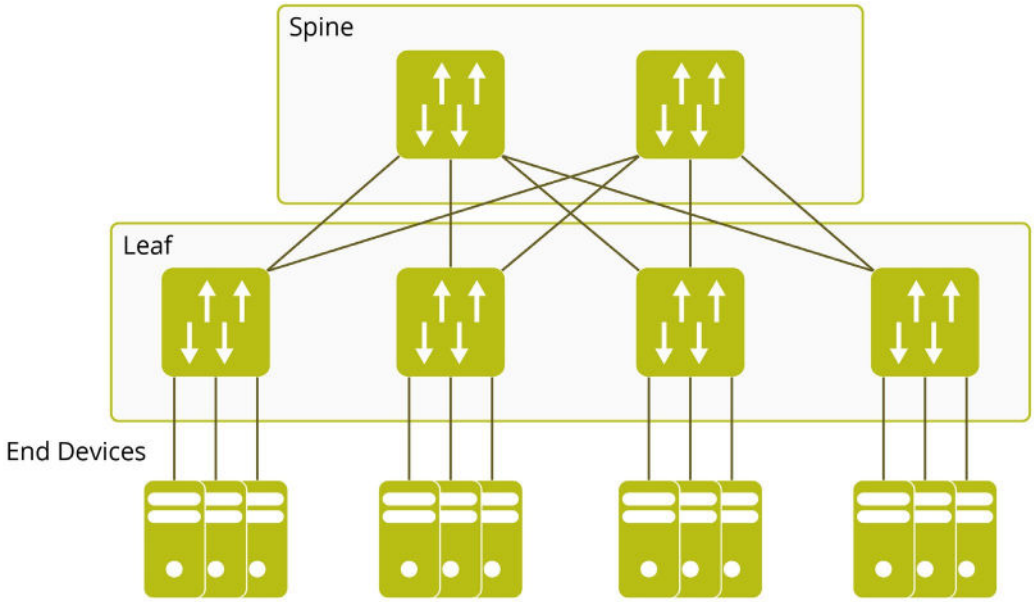
[7] A "hop" is a link between any 2 Ethernet interfaces; i.e., a connection between a sender and a receiver with one switch in between is counted as 2 hops.

[8] "Store-and-forward" means that an Ethernet frame is fully ingested and validated before it is forwarded to the queue of the designated egress port; switches designed for fastest throughput may use "cut-through" technologies where a frame is already forwarded to the designated egress port queue once the Ethernet header containing the destination address has been received. This may reduce the residential time significantly, at the trade-off of potentially forwarding invalid packets.



Basic Network Topologies

However, star-shaped topologies are only practical in small networks; more common are tree-shaped and "spine-leaf" topologies. While a classic tree structure can suit smaller to medium sized networks, a spine-leaf topology can accommodate a large number of devices and traffic flows, provides good redundancy and is thus typically found in data centers or larger corporate environments. The maximum number of hops is typically limited to 2 times the number of topology layers.



Spine-Leaf Architecture

Other topologies like ring and line (also called "daisy-chained") are often found in mobile or install applications where devices offer input and output ports to connect one device to the next. While this is certainly a low-cost option for building networks as deployment of switches is minimized or not even necessary, these topologies definitely increase the latency when traffic needs to flow between the most distant points as each device in the chain adds another hop.

Another factor which may contribute to latency is, of course, the overall distance a packet has to travel as the propagation delay over distance is limited to the speed of light. While this factor has less of an impact in local area networks, it certainly plays a role in wide area application. As a rule of thumb, you have to add a delay of about 1 sample time (@ 48 kHz) per every 6 km (see table in section 1). An interesting number in this table is the delay of 1 ms per 300 km, which equals the default packet time in AES67 (see discussion of stream & packet configuration below).

2.3 NETWORK JITTER ("PDV – PACKET DELAY VARIATION")

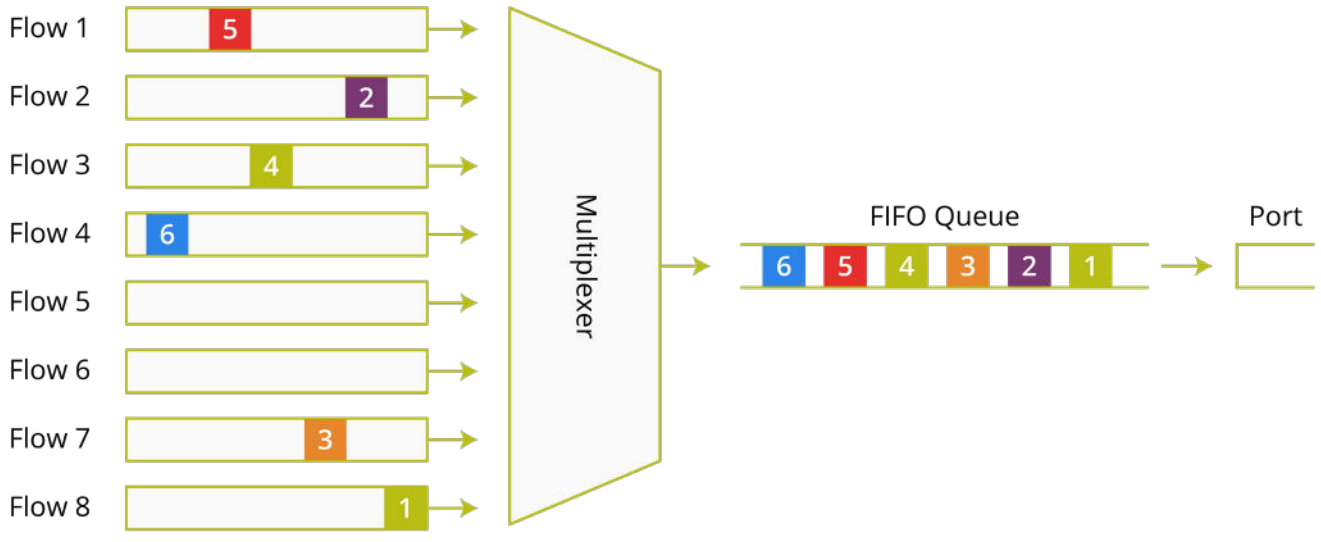
Since the factors discussed above can mostly be considered as static and deterministic, they contribute to the minimum transport delay in any given network. However, network-based transport adds another important variable: network jitter or packet delay variation (PDV). Packet delay variation depends on a number of factors:

- Individual switch performance
- Packet routing
- Dynamically changing traffic situation on individual links

The individual switch performance mostly depends on the switch technology, overall (backplane) switching capacity, available memory for address tables and packet queueing and more. While any of these factors may add variable delays to packet forwarding, they are mostly capacity constraints which — if exceeded — may result in packet dropping. The impact on any of these individual switch characteristics is not something which can easily be predicted or accounted for and is thus not further discussed here.

Depending on the network topology, individual packet routing through the network may be a variable factor. In networks where several possible routes between any two devices exist (i.e., meshed networks, but also spine-leaf architectures), routing changes may occur in the middle of a stream which may result in transport delay variations due to changes in the number of hops a particular stream has to perform. Also, with routing changes the link speed may be different on certain segments, also contributing to a variable delay. However, if the network technology and the potential routing options are known (which is usually the case in managed networks), the worst-case transport latency can be determined and accounted for as a potential factor for PDV.

Another important contributor to PDV is dynamically changing traffic conditions, in particular on aggregating links (i.e., from leaf to spine). When traffic arriving at different ingress ports or from different streams needs to be forwarded to the same egress port, the packets will be lined up in a FIFO[9] queue on that particular egress port as they arrive at the switch. Depending on the current traffic situation, this may result in a significant delay for individual packets as they have to wait for all preceding packets in the queue to be forwarded:

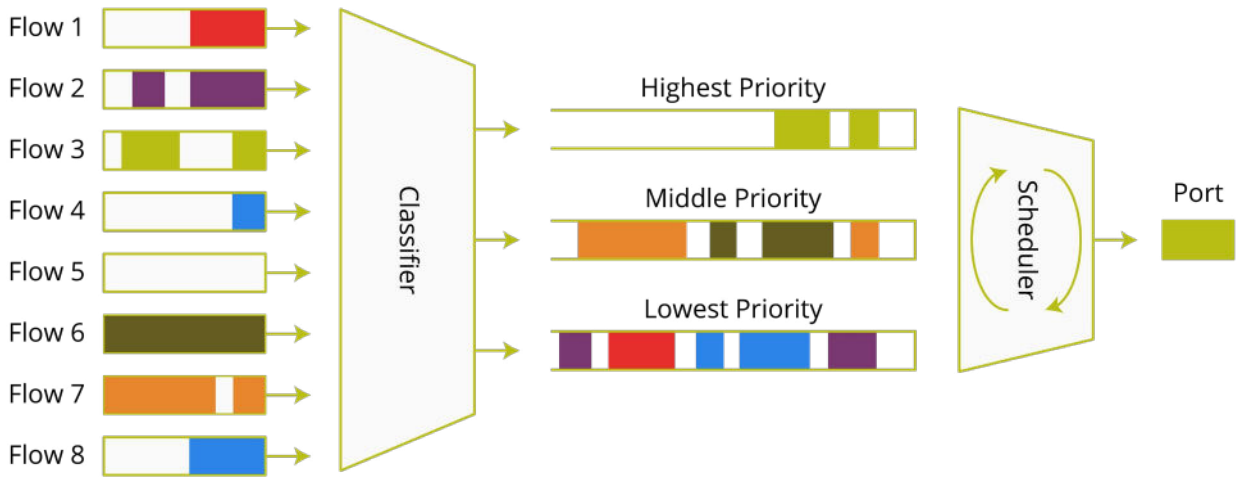


Packet Queueing

[9] FIFO = first in, first out

2.3.1 LIMITING PDV WITH QOS

This situation can be countered by employing "Differentiated Services" (DiffServ), a Quality of Service (QoS) method available on the IP layer. This mechanism is basically available on all managed Ethernet switches these days but needs to be explicitly enabled and configured. Although switches usually process and forward packets based on their Ethernet header information, with DiffServ enabled they can "snoop" into the IP header and evaluate the Differentiated Services Code Point (DSCP) field. Depending on the DSCP value the switch can now store a particular packet into the configured priority queue for the designated egress port. When it's time for the next packet to be forwarded, the egress scheduler will fetch the next packet from the highest available priority queue and forward it to the egress port. The diagram below illustrates this behavior:

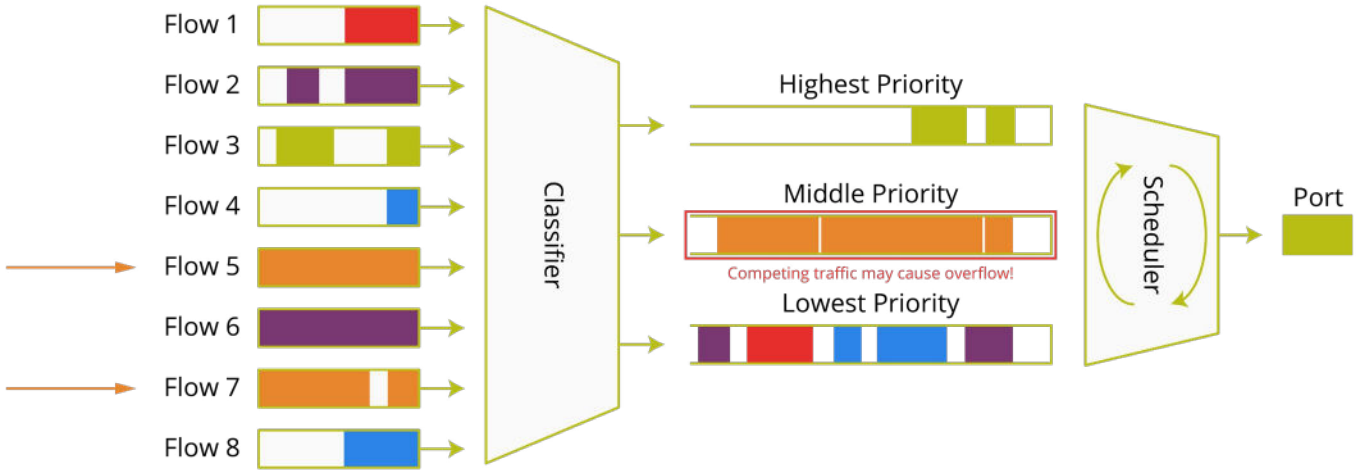


Packet Forwarding with DiffServ QoS

Packets from flow 3 have the highest priority (DSCP) marking and are stored to queue 3 upon arrival, packets from flow 6 and 7 have the next priority marking and get lined up in queue 2 as they arrive, while packets from all other flows with no priority marking are stored into queue 1[10]. The egress scheduler looks for packets in the highest queue first, before forwarding packets from queue 2 or 1. DiffServ works in a similar way to the boarding process at airport gates:

With this QoS method, the residential time (packet delay) and the packet jitter (packet delay variation) due to dynamic traffic aggregation can be minimized for prioritized traffic (the higher the prioritization, the lesser the delay variation). However, DiffServ depends on proper configuration of switches and end nodes, is prone to misconfiguration or rogue acting end nodes and is not a guarantee against packet loss, which may still occur if the available bandwidth on the egress port does not match the accumulated bandwidth of all incoming flows designated for this port. Dropping will start with lowest priority packets first, but may well affect higher priority traffic, depending on the actual dynamic traffic situation:

[10] Typically, switches offer 4 or 8 priority queues per egress port; here, 3 queues have been used to simplify the illustration.



Queue overflow with DiffServ QoS

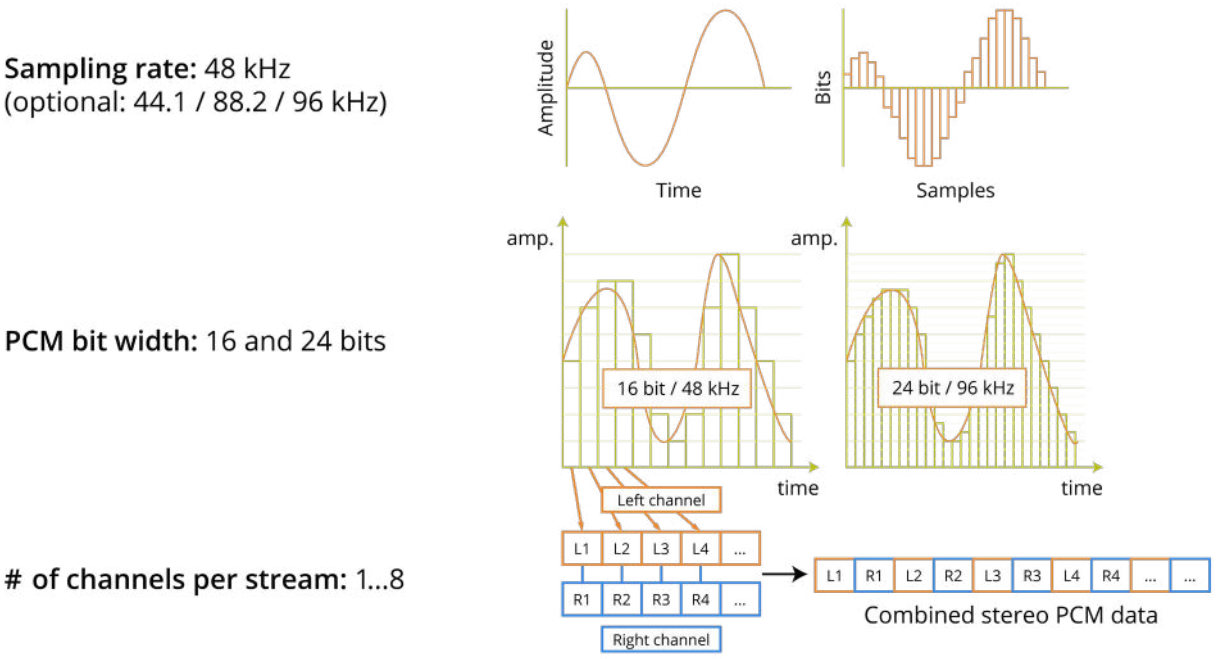
Network administrators should always monitor the traffic pattern on critical links and increase available bandwidth where needed (or take other operational measures to reduce critical bandwidth situations).

2.4 DIGITIZATION & STREAM / PACKET CONFIGURATION

Further factors contributing to network latency are the digitization and the packetization of a particular audio signal.

2.4.1 DIGITIZATION

In a traditional digital audio system, all bits are sampled and transported in a serialized form. Digitization depends on bit depth (the number of bits used to digitize a particular analog signal value, i.e., 16 or 24 bits) and the sample rate (the number of samples taken from an analog signal per second, typically 48 kHz or 96 kHz). Depending on the chosen transport format, multiple channels can be distributed on a single line (i.e., 2 channels on AES3, up to 64 channels on MADI). Once all data for a particular audio frame (samples x no. of channels) is digitized, the individual bits are serialized and forwarded in an isochronous manner:

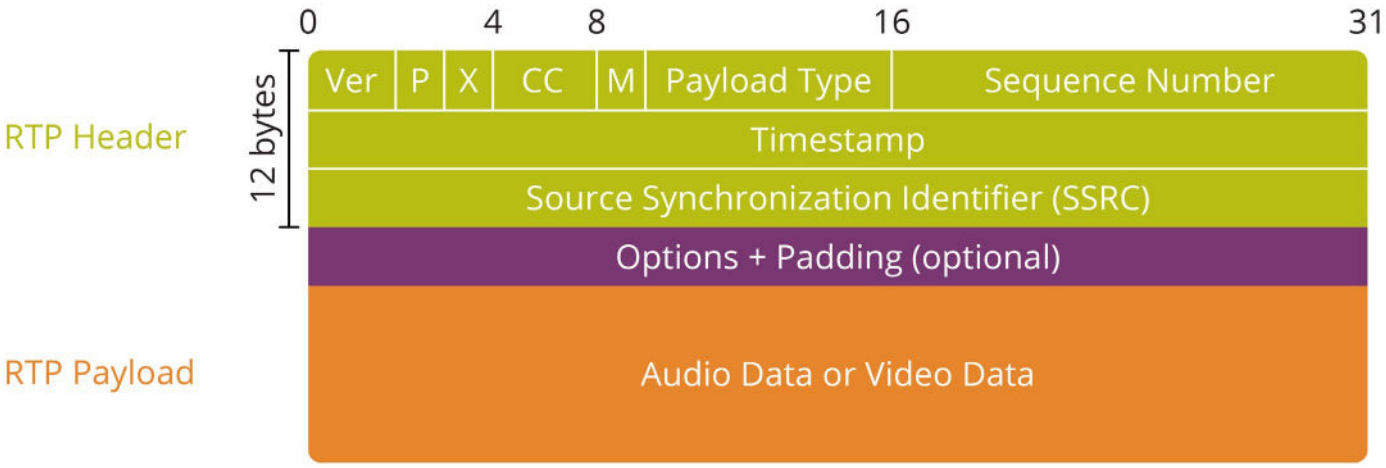


Principle of Digital Sampling

The necessary data encoding (or protocol framing) requires very little overhead, the data rate for a 48 kHz, 24-bit stereo signal transported with AES3 is ~ 3 Mbit/s.

2.4.2 PACKETIZATION

In a networked audio system, the digitized audio data is transported in data packets. The protocol being used in AoIP systems is called RTP (Real-time Transport Protocol). A single RTP packet consists of an RTP header and a payload section:



RTP Packet Format

The RTP header has a fixed length of 12 bytes and the payload section can hold up to 1460 bytes of audio data. Since RTP packets are transported over IP/UDP, the total overhead per packet is 40 bytes[11]. Since it is not very bandwidth-efficient to transport every single audio sample in an individual RTP packet, multiple audio samples from a number of related audio channels are collected in a single packet. The more samples (or audio frames) are collectively transported in a single packet, the more efficiently the available bandwidth can be used as the total packet overhead is always constant; the following table depicts a few possible configuration options and the resulting bandwidth efficiency:

# Samples	Packet Time @ 48 kHz sample rate	Packet Rate (packet/s)	Channels per Flow	Bandwidth Efficiency	Bwidth (Mbit/s)	
					Stream	Total (*8)
1	20.8 µs	48,000	1	5%	32.64	261.12
1	20.8 µs	48,000	8	30%	**40.70**	
6	125 µs	8,000	1	24%	6.40	51.20
6	125 µs	8,000	8	71%	**14.46**	
48	1 ms	1,000	1	71%	1.81	14.46
48	1 ms	1,000	8	95%	**8.87**	

Packet Configuration Options vs. Bandwidth (Efficiency)[12]

In order to save bandwidth, it might seem that an obvious strategy would be to choose a packetization format which uses as much of the maximum available payload space per packet. For example, for a stereo signal digitized with 48 kHz at 16 bits, 365 audio frames would fit into one RTP packet. However, the drawback is the "packet time": in order to fill up the available space, the packetization process has to wait until the required number of audio frames have been digitized before the packet can be forwarded to the network. Consequently, larger packet times implicitly increase the overall latency. In the example above, the packet time (and with it, the absolute minimum latency) is 7.6 ms which is already too long for many professional applications.

[11] Ethernet protocol adds another 18 bytes of overhead (22 bytes for Ethernet with VLAN tagging).

[12] Formula: bandwidth efficiency = payload size / total Ethernet frame length; with total Ethernet frame length= payload size + 40 bytes UDP/IP/RTP header + 18 bytes Ethernet frame overhead.

A good balance between bandwidth efficiency and achievable latency is a packet time of 1 ms (the default configuration for AES67 streams), which — at 48 kHz and 24-bit sampling — results in an overall data rate of ~ 3 Mbit/s and 83 % bandwidth efficiency for a stereo stream; other examples are shown in the table above.

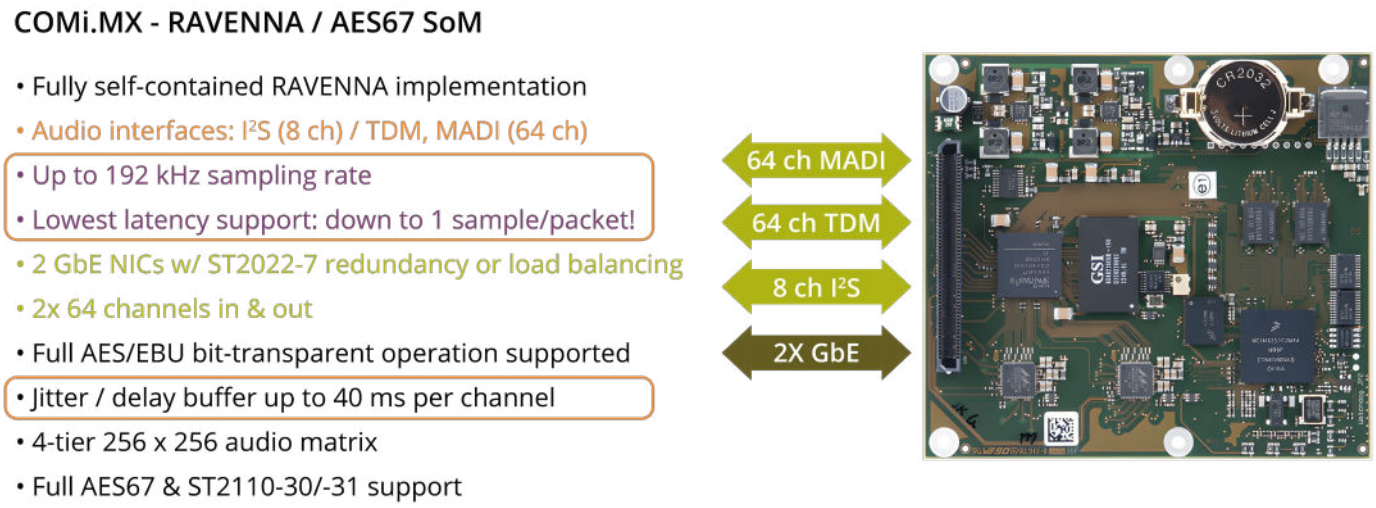
While higher sampling rates can help lower the latency — i.e. at 96 kHz any given number of samples is digitized at half the time compared to 48 kHz — in practice, AoIP systems usually offer the packet time as a given configuration parameter. At 96 kHz sampling rate, 1 ms packet time equates to 96 samples being transported in one RTP packet, resulting in the same packetization latency as at 48 kHz.

2.5 SENDER / RECEIVER IMPLEMENTATION

Another factor contributing to overall latency is the type of sender/receiver implementation. Implementations and their respective influence on latency can be differentiated by three typical platform models:

- Hardware implementations, typically based on FPGA or other dedicated processing circuitry.
- Embedded platforms, typically featuring low-level or Linux-based coding on specialized processors, optionally with certain dedicated hardware support, i.e., ARM-based platforms with dedicated functional and/or I/O support like system clocking with PTP timestamping, AD/DA conversion etc.
- Generic software platforms like Windows or Linux PCs with no dedicated hardware support (i.e. no PTP support, no dedicated clocking or audio I/O)

Hardware-based implementations can typically process packets at line speed and forward the digitized audio data without further processing delay to/from their audio interfaces, resulting in lowest achievable latencies and high channel capabilities.



Example of an FPGA-based Hardware Implementation

On the contrary, pure software-based implementations rely on operating system support for packet handling, the need to reassemble PTP timing and media clocking through pure software algorithms; and they also have to use the interface routines provided with the respective audio interfaces.

The audio data has to be passed back and forth between various library functions and drivers and often needs to be transferred several times between kernel and user space routines. Software Interrupts and/or callbacks often have a non-deterministic run-time behavior while other applications running on the same systems may interfere with required system resources (i.e. available processor time, system memory, network bandwidth etc.). As a result, the minimum achievable latency is significantly higher due to the less deterministic system behavior. In practice, a Virtual Sound Card implementation on a Windows PC may typically operate at latencies in the 10 ms range and may even increase if the VSC is running on a virtual machine.

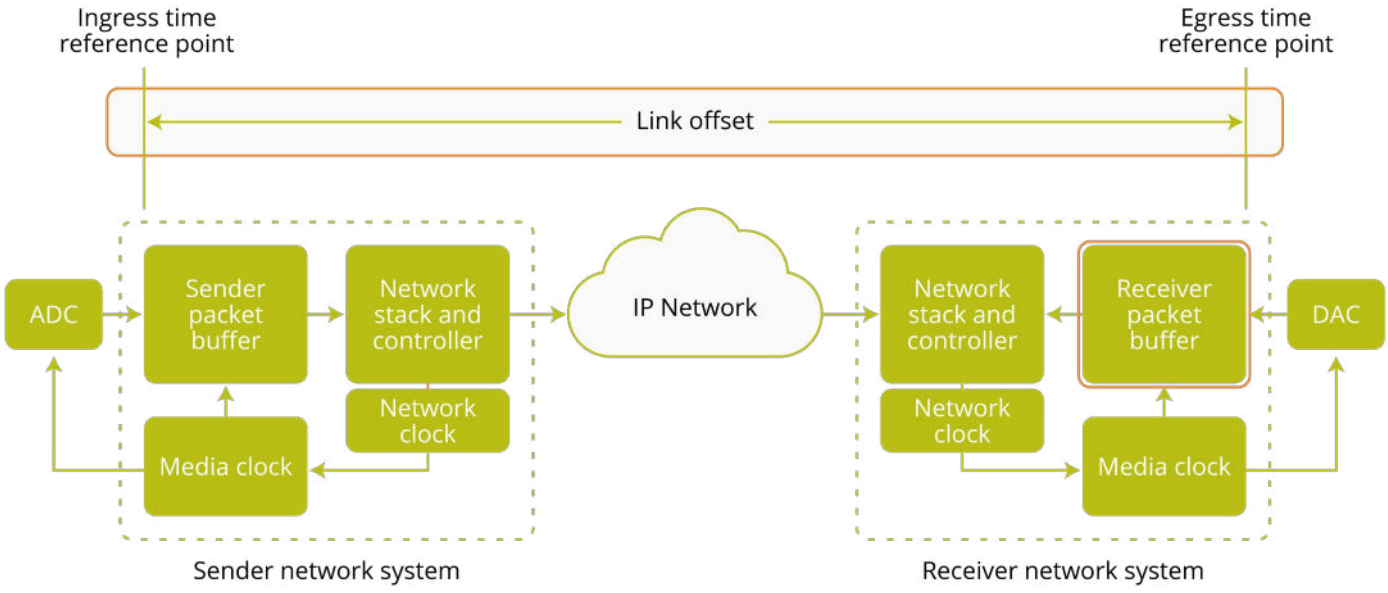
Virtual Sound Card Examples

Implementations on embedded systems — although typically less performant than a generic PC — can take advantage of specific functional support and may achieve processing latencies in the lower single-digit milliseconds range.

2.6 LINK OFFSET

All factors discussed above contribute to the network latency experienced in AoIP systems. While some factors just play a minor role in local or campus-wide networks, they become of significance in wide-area application (i.e. network technology and distance). A major factor contributing to latency is the packet time configuration as this determines the absolute minimum latency which cannot be undercut. Depending on the size of the network and the dynamic traffic situation (i.e., number of concurrently transported streams, in particular on core links), significant further packet delay variation may build up.

In order to achieve an isochronous and undisturbed playout, the worst-case latency must be accounted for with an ample playout delay configuration at the receiver. This parameter is called “link offset”:



The Link Offset in AoIP Systems

If the link offset parameter is too small, certain packets may arrive too late due to a temporarily higher PDV and consequently, some samples may miss their designated playout time, resulting in audio interruption at the playout stage. On the contrary, a larger link offset ensures safer playout, but implicitly results in higher overall latency.

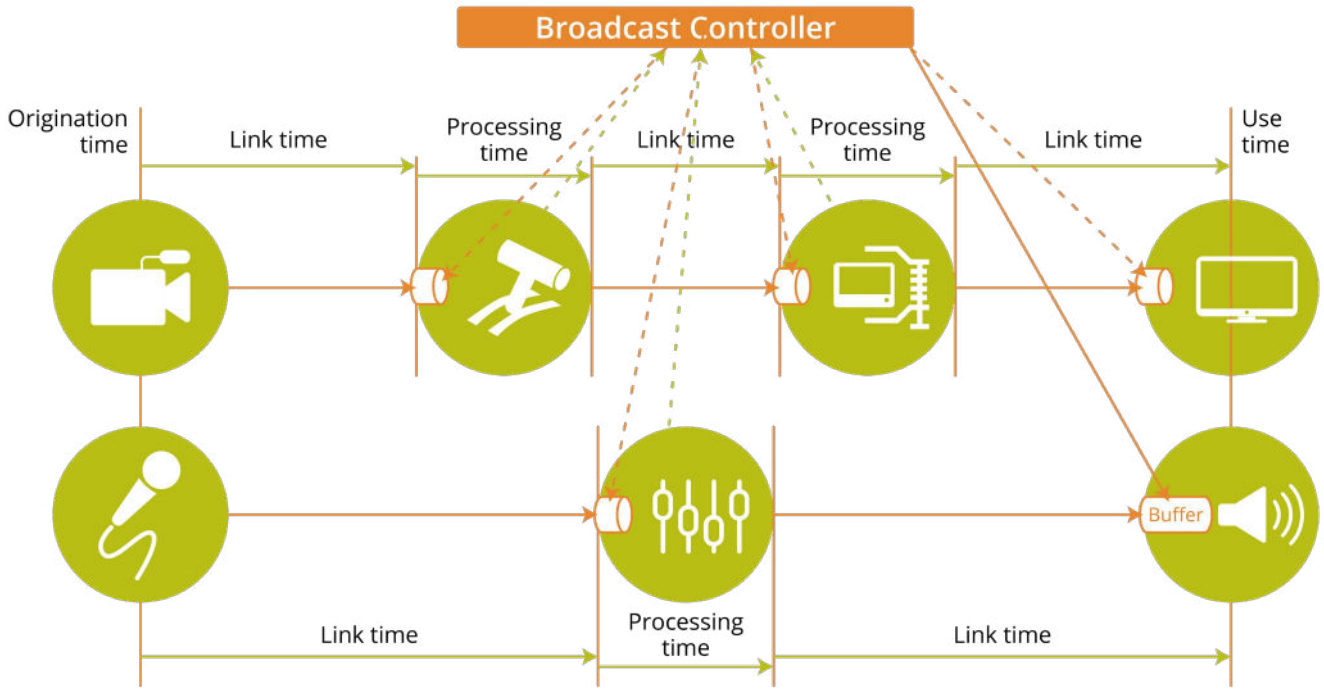
In any case, the receiver needs to provide ample storage capacity to cope with any PDV: incoming packets which have accumulated very little transport delay need to be buffered until the designated link offset time has been reached. While this is less of a problem in local networks with mostly hardware-based or embedded system implementations where link offsets can typically set to single-digit milliseconds, it may become an issue when VSCs and/or VMs are part of the system and receiver implementations do not offer ample buffer capacity to cope with the PDV[13]. Wide-area applications may require even larger buffer capacity to cope with further transport delays and increased PDV; therefore, some implementations offer 40 ms buffer capacity or beyond per channel.

2.7 STREAM ALIGNMENT

In order to align the playout of a particular stream (or a number of related streams) at various independent receiving devices, all devices would have to be configured with the same link offset value. While some devices offer link offset configuration on a per stream basis, others offer link offset to be configured as a device-wide or even system-wide parameter only; yet other implementations have a predefined value that is not configurable at all.

The existence of ample receiver buffer capacity in combination with configurable link offset values enables system-wide playout alignment, even in applications with mixed media (audio and video).

In the example below, a video signal captured by a camera and the related audio signal, captured by a microphone, travel as independent streams through various processing stages with different internal processing times. At the end, lip-synced playout is desired. This can be achieved by configuring ample link offset values for the individual streams being transported between the various stages. A management system can interrogate for the internal processing delays and can calculate and configure the required link offsets through both chains so that final playout of audio and video are perfectly aligned:



Production workflow timing

[13] AES67 describes an interoperability gap in this respect, as receivers are only required to provide 3 ms minimum buffer capacity (with 20 ms recommended), but senders are allowed to send out packets with a variation of up to 17 ms (to accommodate VSCs as AES67-compliant senders). While most AES67-compliant receiver implementations provide the recommended buffer capacity, some receiver implementations are strictly limited to the minimum capacity of 3 ms.

3. SUMMARY

Traditional digital audio transport is based on the principles of circuit switching — audio is transported in a serialized form on dedicated point-to-point connections with specific cables, using defined protocols and formats (i.e., AES3, AES10). Latency is basically defined by cable length (not taking any conversion or processing delays into account).

Network-based audio distribution is based on packet switching — audio data is transported in chunks (packets) on a general-purpose network infrastructure using generic transport protocols and formats (i.e. RTP/UDP/IP). The transport infrastructure is neither purpose-built nor being used exclusively. The transport protocols are based on various transport layers covering packet routing and content-agnostic payload transport. Latency depends on a number of additional factors of variable impact:

- Underlying network technology
- Network topology
- Network jitter (PDV)
- Stream/packet configuration (packet time)
- Sender/receiver implementation
- Stream alignment

QoS can lower the impact of PDV, while ample stream & packet configuration enables control of the minimum possible latency. An appropriate link offset covering all (variable) latency effects needs to be configured per stream to ensure uninterrupted playout or processing at the receiver stage.

Takeaway: While sub-millisecond latency is achievable with a capable network infrastructure and proper configuration stream configuration, a typical latency with an AES67 default configuration can be expected in the 2 - 3 ms range. Ample receiver buffer capacity enhances the scope towards wide-area applications or to accommodate software-based senders.

For more information check out the RAVENNA website (www.ravenna-network.com).

Andreas.Hildebrand@lawo.com
www.ravenna-network.com