Viable distribution of Multi-channel Audio-over-IP for Live and Interactive “Voice Talent” based Gaming using High-quality, Low-latency Audio Codec technology

a technical white paper

by APTX

www.aptx.com

© APT Licensing Ltd., 2009

ABSTRACT

The delivery of multi-channel audio – from mono to surround sound – in real-time over public IP networks for the purpose of interactive crowd-participant gaming presents a significant design engineering challenge to games developers, console manufacturers, ISPs and CDNs. Leveraging expertise gained in professional broadcasting and recording studio post-production, APT has developed a robust and scalable audio codec technology that meshes with popular gaming systems to realize low-latency distribution of high-quality audio for immersive, instantaneous audio experiences in massively multi-player online games involving interactive audience responses.

INTRODUCTION

This paper aims to inform, educate and stimulate interest in the possibilities and challenges in the area of audio-over-IP for multi-player interactive gaming. It draws on the expertise of many audio engineers in the fields of broadcast and telecom to bring together the reasons and rationale behind the problems and issues. It also aims to bring to the table information that may well be the solution to some of the problems and prepare the industry for a full interactive audio experience across IP networks.

THE STATE OF AUDIO-OVER-IP FOR GAMING

Bringing an immersive, interactive audio experience to a vast number of discreet users geographically dispersed across the world creates a unique set of challenges to be overcome. As technology has improved and bandwidth has increased, so the quality of games and user expectation has increased. The present and next generation of gamers see the potential of the IP networks that surround everyone to be used for interactive experiences. Whilst demonstrations of these technologies can be seen and heard across smaller networks, the challenges for larger networks remain.

The geographical distribution of the participants and audience places challenges in the path of interaction. Whilst IP networks provide almost unlimited connectivity to the most inaccessible places on the planet, and the most technically congested at bandwidths that can support higher and higher quality content, it remains a challenge to use these networks for interactive transfer of audio. Systems such as CobraNet [1] and EtherSound [2] distribute audio over local area networks very efficiently with minimal latency.

 © APT Licensing Ltd., 2009

System	Latency (ms)
CobraNet	5.33 @ (F_s = 48kHz)
EtherSound	0.0012

Table 1 Streaming system latency

The CobraNet system delay is related to the 256 sample packet used for transport, whilst the EtherSound system has a sample delay of only 125µs making it highly efficient in terms of latency.

To try and summarise the problems:

- **Delay/Consistency**: This comes from the characteristics of a packet network. The flexibility that packet based networks bring also beings challenges for consistency of delivery. The staggered and inconsistent arrival times, Jitter, of the packets, requires that end points buffer enough audio to mask the longest arrival time.

- **Delay/Throughput**: Transferring a packet over a large geographic area, through congested regions produces challenges for network providers. The ability to have a fair and reasonable approach to all users, opposes the need in real-time systems for guaranteed bandwidth availability.

- **Delay/Audio compression**: Today there are some great technologies for compression audio to bit rates that can be realistically transferred across even the most restrictive bandwidth circuits. The difficulty faced by these algorithms that the analysis of audio to reduce it to the bandwidths required often creates delays that prohibit its use in real time environments.

- **Delay/Packet Optimisation**: Each router/switch in a network will have a queuing rule applied to the traffic input and output from its ports. These often relate to the size and frequency of the packets. Practically a small number of large packets may traverse the route or switch more efficiently than a large number of small packets. Small packets take less time to fill with audio than large packets. So a conundrum exists, in that one optimisation creates a problem for another element of the network.

- **Quality of Service**: Many publically available services operate on a best effort to provide the best quality for service to all its subscribers in as fair a way as possible.

THE NETWORK

IP networks are based on many varied technologies as they traverse the globe. They can be synchronous, asynchronous, variable bandwidth, guaranteed bandwidth, fibre, copper, wireless, simplex and duplex. They

© APT Licensing Ltd., 2009

invariable use multiplexers, routers, switches and modems to connect end points. Each of these technologies has inherent characteristics around delay, throughput, packet-size and management control.

THE END-POINTS

Most consumers have network attached equipment capable of receiving audio and video streams. Laptops, game consoles, set top boxes even personal media devices. The capabilities of these devices also vary in terms of their ability to process the stream in real time and provide a truly good user experience. The plethora of equipment available also raises the issue of interoperability and re-usability. How many pieces of equipment under the TV is enough?

The technology used to stream audio integrates with the TCP/IP stack that resides within most user equipment. In most cases this is used for file transfer under TCP control under a guaranteed delivery mechanism. This mechanism is too slow for real time operation due to the re-send mechanism present in the protocol. The UDP protocol, or send and forget is much more efficient in terms of delay, with the resultant downside that it is not guaranteed that the audio or media will actually be delivered.

THE AUDIO

MP3 has captured the consumer audio compression ideal in that it has transcended from a being a technology to a product in its own right. It captured a market segment in relation to the timing of its introduction and the availability of other technology, hard disk drives, solid-state memories, available at the time. The next generation of DVD and gaming technology uses high quality audio up to 96kHz sampling at 24bit. For a stereo stream this results in a stream 4.608MBps. Even a reduced quality offering at 48kHz/16bit creates a stream of 2.304MBps. Technology to reduce this to a useable transmission bandwidth is freely available and of a significantly higher quality than MP3. Some of these technologies are detailed below in terms of bit rate, delay and quality.

Algorithm	Delay (ms)	Bitrates (kbps)
apt-X Live	1.8	> 32
AAC-LD	>20	>16
Linear	0	1150
MP3	200-300	16-128

Table 2 Algorithm comparison

© APT Licensing Ltd., 2009

Audio for real time applications has been researched widely to see if the networks can be used for applications such as online jamming, wireless audio interaction, and the effect of delay on user experience and ability to react and manage the delay.

CURRENT ISSUES

SYNCHRONIZATION: The streaming of data in one direction over an asynchronous network implies that the receiver monitors the rate of the incoming stream and adjusts its play-out rate to match that of the incoming stream. The same principle applies to a game requiring duplex connectivity. The receiver would be locked to the initial Rx stream and therefore lock the rate, of the Tx stream, to that of the Rx stream. In this way the system self regulates and provides an inherent synchronous connection across all parties. Ad-hoc networking between individuals requires a common clock of the ability to provide a central point of reference for everyone to use as the electronic metronome for the performance or interaction.

BANDWIDTH: Streaming into a network at low bandwidth can be handled relatively efficiently using port forwarding or multicasting to minimise the bandwidth required at all points in the network. However the returned audio from each end point needs to be funnelled to a central point. Take a simple example of an online game requiring 30 participants. Audio is streamed at 64kbps from each user, but that user should therefore have to receive 29 times that bandwidth from the other participants, almost 1.92Mbps in real time. The diagram below visualises this scenario. Routers that have multiple downstream connections require greater capacity in a unicast model.

Streaming to a large audience can be a challenge: the bandwidth required to unicast to a million users – even at low bandwidths – requires an enormous amount of dedicated bandwidth. For example, a simple audio feed using a high quality audio compression technology at 256kbps, to 1 million users would require 256Gbps + approx 10-30% overhead for the IP network. Most publically available networks around the world are unicast based and provide a one to one link for media streaming. The next generation of technology widely available today is multicast. This provides the ability to send one stream to many different end-points with minimal bandwidth usage. Each end-point requests the stream be forwarded from the multicast address of the media they want to use. The stream is then forwarded to that endpoint. A plethora of technologies are now available to optimise the amount of bandwidth used in this type of system.

© APT Licensing Ltd., 2009

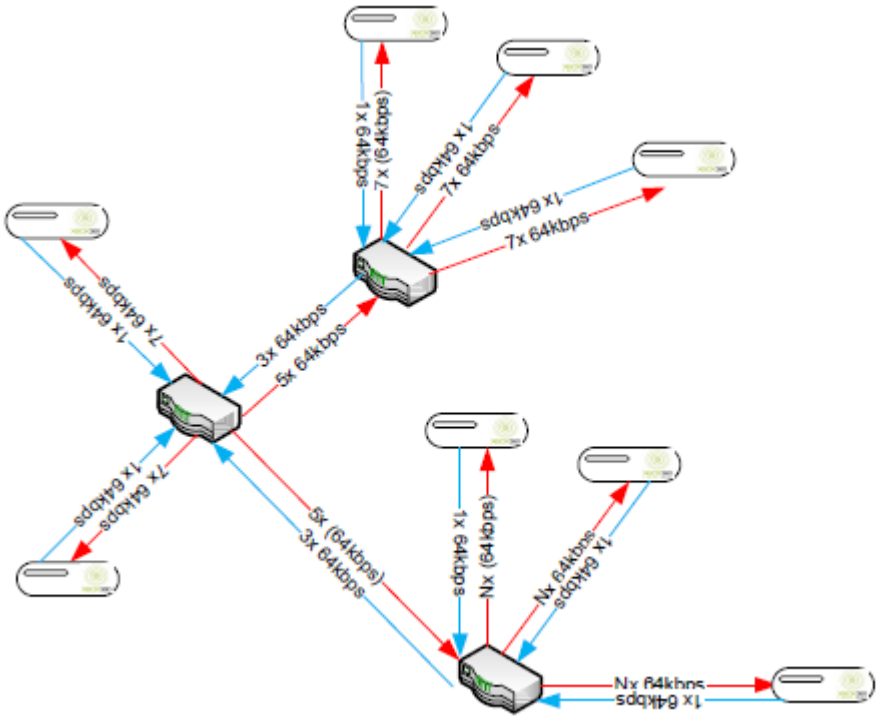

Figure 1 Multi-player game

AUDIO COMPRESSION: Streaming in one direction is not delay sensitive, streaming in both directions is delay sensitive causing problems of interaction and compensation among users. Expedient trans-coding of audio for mixing, ad insertion requires the ability to tandem code, or encode and decode the same material many times in order to post process as the audio traverses the network.

DELAY: In the paper "The Effects of Latency on Live Sound Monitoring" [3] authors Jon Boley and Michael Lester indicate that the acceptable level of latency for monitoring is between 1.4ms and 43.0ms depending on instrument and audio content. Translating this into a real-time gaming situation puts extreme pressure onto the infrastructure and console to provide a true real time experience that does not become tiring. The internet spans the world and therefore in relation to online applications, participants could therefore be geographically dispersed across continents.

© APT Licensing Ltd., 2009

Origin	Destination	Delay (ms)	Packet Loss (percent %)
LA	Sydney	197	5.89
LA	Hong Kong	214	4.42
LA	Frankfurt	173	0.2
LA	London	176	5.93
LA	Boston	110	2.85
LA	Dallas	47	0.66
LA	San Francisco	26	0.78
LA	Rio de Janiero	229	4.45

Table 3 Delay examples

Adding to this delay is the jitter experienced by the receiver due to the difference in time between the expected arrival time of the packet and the actual arrival time of the packet. Receivers need to compensate for this by creating a buffer equal to or greater than the largest jitter time experience by the receiver. The jitter may well vary from several ms to several hundred milliseconds.ombining the transport delay with the jitter gives the user an idea of the inherent delay in the system before audio or processing is added. For example, the maximum delay of the packet shown in figure 2 from Los Angeles, California to Cambridge, near Boston, Massachusetts is 110ms, yet the average delay is on 80ms. The jitter or compensation required is therefore greater than 30ms.

© APT Licensing Ltd., 2009

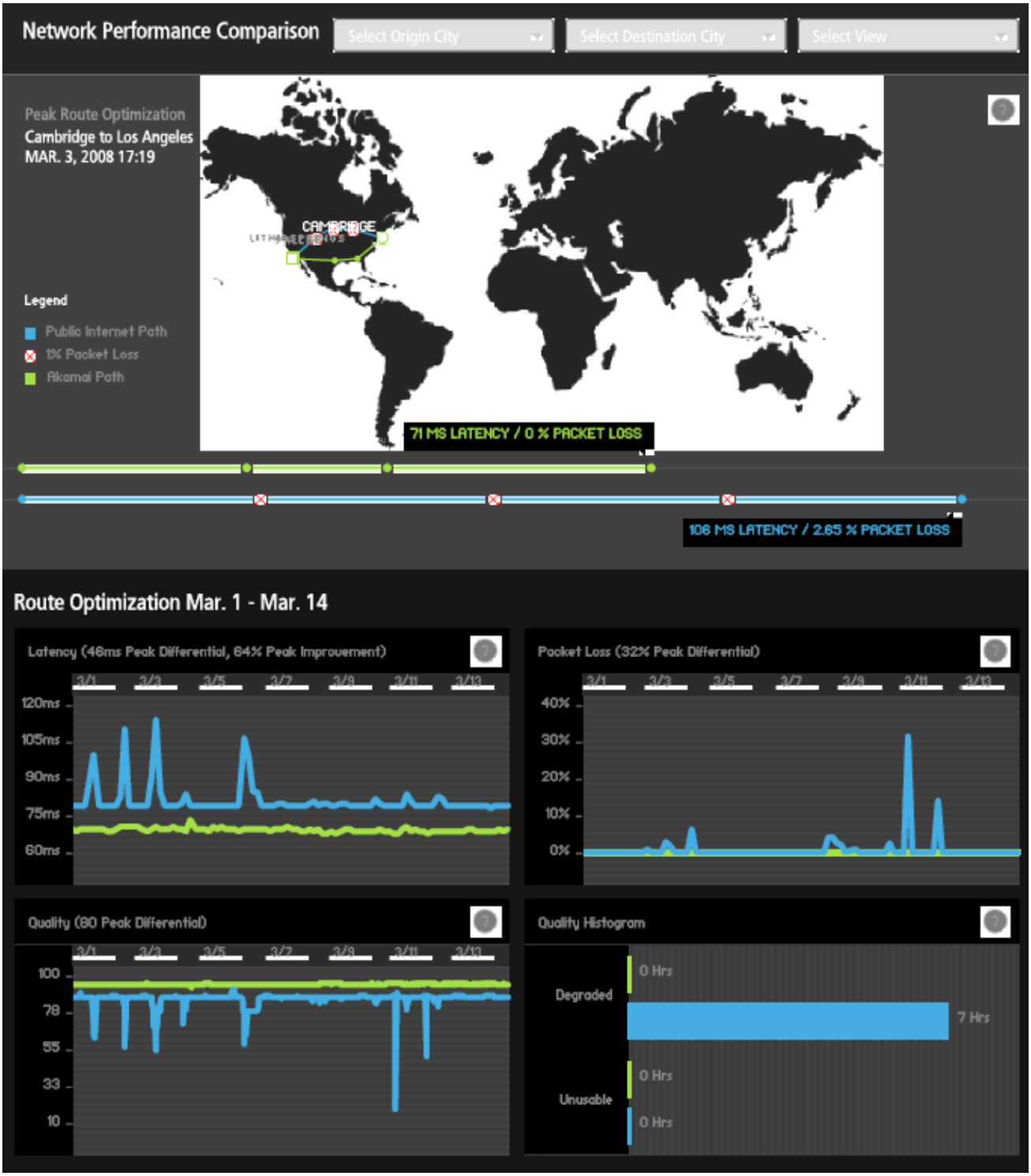

Figure 2 Network transfer example

As can also be seen from the screen-capture, the managed network (in green) – while not decreasing the transport delay by a great amount – has reduced the jitter to a few ms, therefore improving the end-to-end delay.

© APT Licensing Ltd., 2009

THE APPLICATIONS

Streaming media to millions of users can be done today: the BBC iPlayer, SHOUTcast, and various other technologies are available for this purpose. This unidirectional unicast stream however, while practical and available, has physical limits in relation to scalability. Adding another million users requires more and more bandwidth from the network. Without the additional bandwidth, congestion will take over and the quality of the end user experience will deteriorate. Figure 3 and figure 4 show how the bandwidth is disseminated through unicast/multicast networks respectively.

Figure 3 Unicast audio distribution

© APT Licensing Ltd., 2009

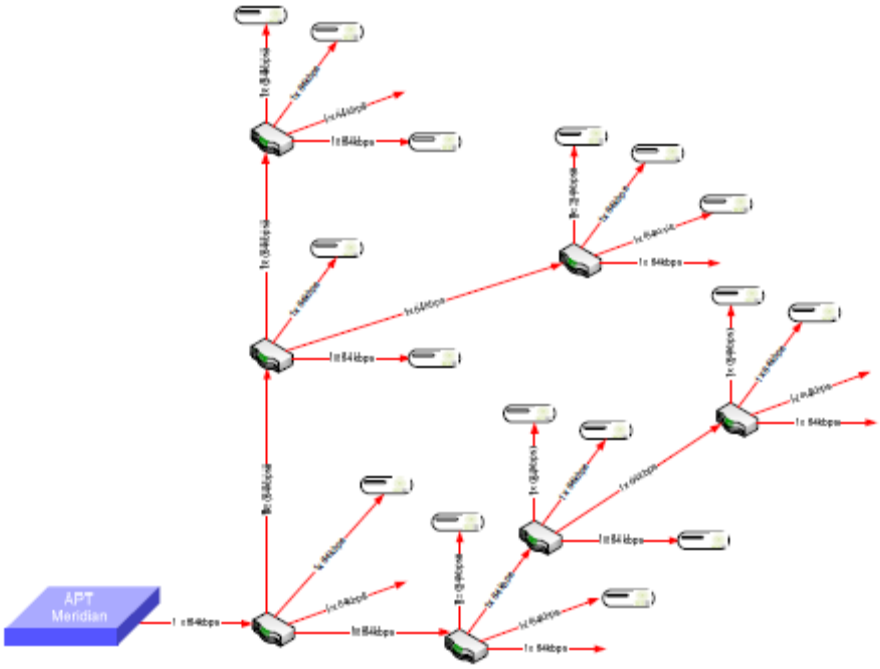

Figure 4 Multicast audio distribution

Historically, streaming audio over telecommunications networks has been used to provide a mechanism to locate the talent in one location and send or distribute the audio to another geographical location/studio. The bodily relocation of "voice talent" – the voice-over provided by specialist actors – is expensive and problematic; moving audio data is relatively simple by comparison. The ability to harness voice talent to an online game – in the most simple and cost-effective way possible – brings the ability to have immersive audio experiences through the internet. Transmitting a concert or an interview through a network to an almost unlimited audience, and be able to create the ambience and feedback through sound, and have direct interaction between all participants in real time must surely be the next generation in entertainment.

The on-line interactive game moving to sharing audio and video in real-time has major technical limitations with current network topologies. The delay through the network is the starting point for the problem, compounded by either lack of bandwidth or coding delay of the compression technology used. Therein lay the technical challenges facing the future of the truly immersive online experience. The solution is to have a low-delay network capable of delivering the content to the end-points, and to have a system that is not only scalable in terms of bandwidth, but also in terms of functionality.

The second issue is the ability to encode and decode media, audio and potentially video as fast as possible without severe loss of quality. Audio compression technology usually falls into two distinct groups, high

 © APT Licensing Ltd., 2009

quality/high bandwidth, large delays or low quality, low bandwidth, low delay. Over the last few years technologies have emerged that fill the missing niche in the market, that of low delay, low bandwidth and high quality.

apt-X Live is an ADPCM based algorithm capable of up to 8:1 compression at greater than 12kHz audio bandwidth at a dynamic range of greater that 85dB. The encoding delay is only 1.8ms and the decoder delay is also 1.8ms. The algorithm is not frame based but contains a built in synchronisation mechanism using less than 3% of the allocated bandwidth. The algorithm was developed with live performance in mind and targeted the lowest delay possible with the highest quality audio possible. The algorithm is also designed to have a linear phase response, so it can encode multiple channels of audio and reproduce almost perfect phase coherency across all encoded channels.

The other key issue for media streaming and interactive audio is that of synchronisation. The source and destination will ultimately not be exactly frequency locked and therefore the potential for audio glitching due to buffer over-runs and under-runs is a possibility. In the broadcast market technologies such as GPS have been used to synchronise the source and destination. This is somewhat impractical for consumer products and installations.

CONCLUSIONS

In conclusion, for the ultimate immersive audio experience, the delay between users needs to be reduced to less than 120ms so 60ms one way for voice gaming and for interactive music less than 10ms. Given the technology available today this is still a step change away from where it could be in the next ten years.

REFERENCES

1) CobraNet: a combination of software, hardware and network protocols designed to deliver uncompressed, multi-channel, low-latency digital audio over a standard Ethernet network; developed in the 1990s, CobraNet is widely regarded as the first commercially successful implementation of audio-over-Ethernet. *http://en.wikipedia.org/wiki/Cobranet*

2) EtherSound: developed and licensed by Digigram, EtherSound is one of several Audio-over-Ethernet technologies currently used in audio engineering and broadcast engineering applications. *http://en.wikipedia.org/wiki/EtherSound*

© APT Licensing Ltd., 2009

3) "The Effects of Latency on Live Sound Monitoring" Michael Lester and Jon Boley, Shure Incorporated (Niles, IL): AES New York 2007. *A subjective listening test was conducted to determine how objectionable various amounts of latency are for performers in live monitoring scenarios. Several popular instruments were used and the results of tests with wedge monitors are compared to those with in-ear monitors. It is shown that the audibility of latency is dependent on both the type of instrument and monitoring environment. This experiment shows that the acceptable amount of latency can range from 42ms to possibly less than 1.4ms under certain conditions. The differences in latency perception for each instrument are discussed. It is also shown that more latency is generally acceptable for wedge monitoring setups than for in-ear monitors.*

4) Akamai Technologies: on-line latency and jitter measurement examples from this Content Delivery Network (CDN) provider. *http://www.akamai.com*

MORE INFORMATION

More information about **apt-X Lossless** and other **apt-X®** series audio codecs, including comprehensive technical data and specimen commercial licensing documentation, is available from the appropriate contacts given or send a request e-mail to *licensing@aptx.com*

www.aptx.com

APTX – *the apt-X® licensing company*

APT Licensing Limited

Whiterock Business Park

729 Springfield Road

Belfast, BT12 7FP

Northern Ireland, UK

 © APT Licensing Ltd., 2009