IP for Audio Broadcast Networks

Jon McClintock
Commercial Director

jmcclintock@aptx.com
www.aptx.com

Transporting broadcast quality audio over IP is the new “must have” technology for radio and TV networks. However, IP as a transport mechanism has a number of inherent characteristics that could potentially pose problems for codec manufacturers and broadcasters alike.

It is a fact that while massive IP bandwidth is now available it comes with some preconditions. These include:

1) The packets do not always arrive in the order in which they where sent.
2) The transport protocols introduce a natural delay to the link.
3) Dropped packets are always a possibility.
4) Bandwidth is not guaranteed.
5) Packet Delays through the network are variable.

Manufacturers have already put in place methods and protocols to cope with the plethora of issues surrounding audio over IP including forward error correction, concealment and security mechanisms.

Experience has also shown that the use of MPEG derivatives in contribution and distribution networks has resulted in almost unusable coding delays and serious degradation of audio quality follow several psycho-acoustic passes. The layering together of the inherent delays associated with MPEG and IP will create an impossible scenario for any broadcaster wishing to put live content on-air.

This paper aims to explore the audio coding technologies available and determine if an answer exists that can benefit the broadcast community by delivering a solution that maximizes the benefits of IP, retains acoustic integrity and ensures that the coding delay is reduced to a workable value, i.e. under 20 milliseconds.

IP Network Considerations

Latency

All networks have transport latency due to the natural laws of physics. Transporting an electronic signal through whatever medium takes a finite amount of time that cannot be removed. In a switched network there is both the standard transmission delay and also the packetizing delay to contend with. By definition a packet must be assembled and consists of a header plus payload. The size of that payload can be varied but ultimately it consists of an audio sample. Take for example, a system that uses a four to one compression algorithm and has a packet size of 128 bytes. That’s 512 audio samples, equal to 666 usec in Mono and 333 usec in stereo. Then take the time it takes to propagate through the UDP stack after being assembled into an RTP (Real-time Transport Protocol) packet. In real, live-unit tests this is approximately 20-30 ms. It is important to realize that this natural latency increases as the sample frequency decreases. With this inherent latency in the protocol stack the additional delay of the audio coding is critical in real time audio applications.

Lost packets

Lost packets are a fundamental feature of switched networks and something that all audio codecs must learn to live with or suffer the consequences. Losing an audio packet quite literally translates into losing audio, which is a major problem in the broadcast environment. Several solutions exist, ranging from ignoring the problem to masking it through to retransmission. If one discounts ignoring the problem as a solution, what remains are the masking and retransmission scenarios.

Packet size

As with all packet-based systems there is a trade off between the size of the packet, the processing overhead and network traffic/congestion. Smaller packets increase overhead with the additional byte required to populate the packet header. Smaller packets are transported through the network at the same speed but an entire packet can be received more quickly simply due to its size. The difficulty with smaller packets is the increased likelihood that a packet will be lost or received out of sequence. Larger packets have a problem once they exceed the Ethernet limit of around 1500 bytes. The packet is then fragmented and transmitted in two or more parts, before being re-assembled at the opposite end. This complicates the receiver with no real benefit.

UDP versus TCP

Transfer control protocol (TCP) is a command and response architecture that guarantees delivery through an exchange of information between sender and receiver. This interaction requires time and has proven unworkable for streaming applications requiring low latency or real time operation. User datagram protocol (UDP) is simple send/receive architecture with little in the way of payload protection or guaranteed delivery. It does, however, lend itself to streaming applications since there is less processing delay on the protocol.

Jitter

These are packets that are received either side of the predicted arrival time. This is a feature of packet switched networks given that any packet can take any route from source to destination. Thus packet jitter will affect the way in which the receiver must handle the data sent. Packet arrival jitter can be significant and can amount to several seconds depending on the network. The buffer depth is therefore critical in allowing the codec to provide enough time for the packet to be received and decoded before its playout time. Reducing the buffer size reduces the jitter time mask available and increases the potential of being forced to drop packets that have arrived beyond their playout time.

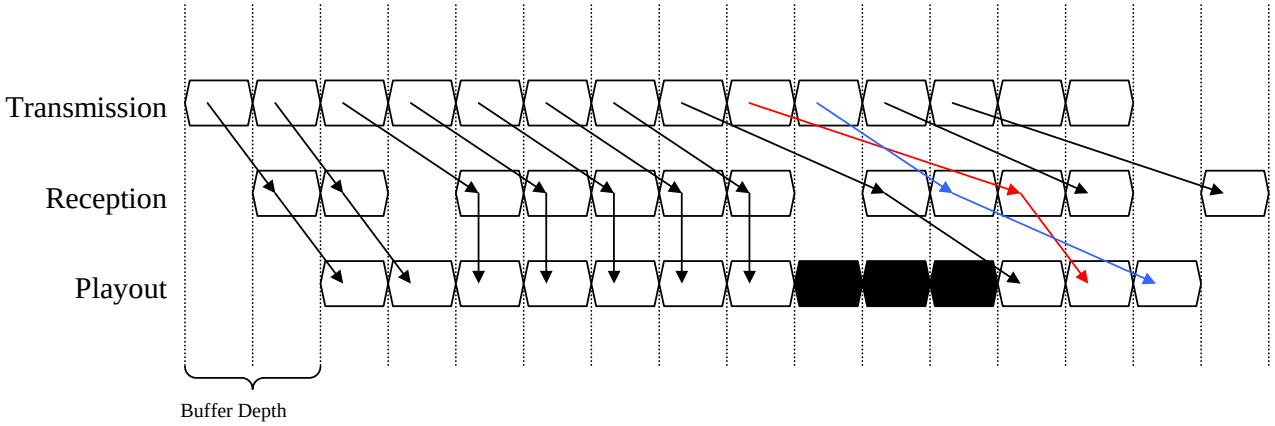

Figure 1 Network Jitter Effects

The above diagram shows the effect of network jitter on the reception of audio and its subsequent playout through an audio system. The buffer depth is usually set in milliseconds; in this case it is set to a two-packet buffer. For the purpose of this example this allows for up to a two-packet jitter delay in the system. Provided the network jitter is low the system is unaffected and plays out the packets received in sequence. However, should the jitter increase, there is the distinct possibility that the packets will arrive after the determined playout time. In this example the packets have to be dropped, which results in the audio being corrupted.

Correction/Concealment/

Overview

No forward error correction is currently used in the UDP packet, which makes it susceptible to bit errors. It does however speed up the processing since this additional correction stage is bypassed. Forward error correction also has implications for end-to-end systems. For example, to what extent should the FEC correct the errors i.e. all or just partially? Alternatively, does a failure indicate that a retransmission should occur?

Concealment

Assuming retransmission is not practical then the decoder must implement some type of concealment to prevent or conceal audio loss.

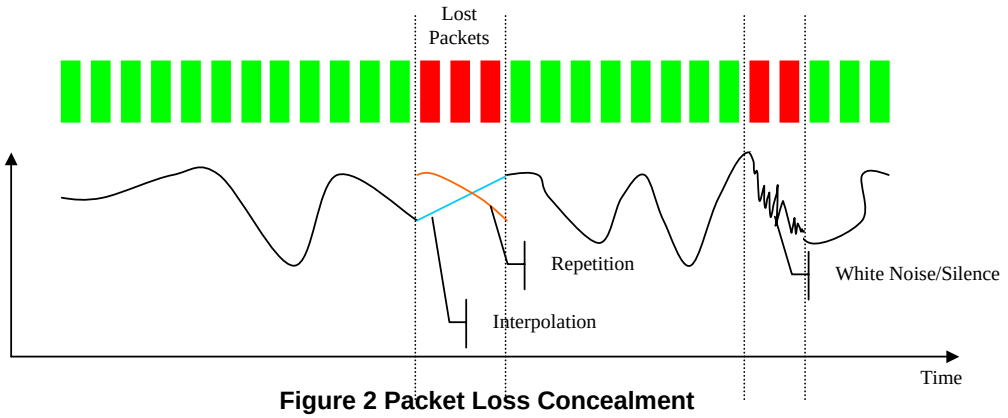

Figure 2 Packet Loss Concealment

Various methods can be used to conceal errors in the final reproduction of the audio due to packet loss. They range from simple repetition of the last good packet received, to silence/noise injection, or interpolation and retransmission. All have an impact on the reproduced audio.

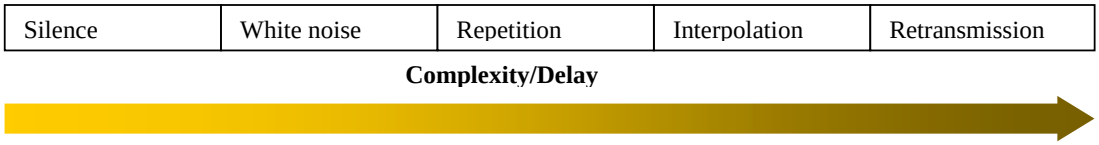

Figure 3 Perceived audio quality

In listening tests the injection of silence produced unacceptable breaks in the audio that led to a level of incoherence. The use of white noise improved the intelligibility of the reproduced audio but was again noticeable. The use of repetition of the last known good frame produced favorable results. The use of interpolation/pattern matching/waveform substitution to conceal the loss of packets is possible but the benefits versus complexity are governed by a law of diminishing returns. The results of these techniques are all governed by subjective improvements in audio quality and are also subject to the amount of audio lost that is being concealed or repaired.

Correction

The use of Forward Error Correction to ensure packet recovery can be effective in audio streaming applications but it has implications for real time applications due to the processing and data overheads associated with FEC algorithms.

The determination of how much FEC to use has to be related to the losses experienced with the medium being used, since adding more overhead to a heavily congested medium may exacerbate the situation. Also, the scheme used must be tailored to the type of loss being experienced. For example, is the FEC designed to correct burst packet loss, or a percentage of lost packets if the packets, which are lost, are non-consecutive?

Technique	Overhead	Complexity	Scope
FEC per n-1 packets	Low	Low	Uniform Loss No Burst Recovery
FEC packet per packet	Low	High	Full recovery possible
FEC packet per n packets	Med	High	Burst loss recovery possible depending on scheme Increased delay

Table 1 FEC techniques comparison

To maintain compatibility and interoperability between codecs the FEC information should be sent via a separate port so that the audio codec does not become confused if it cannot handle the FEC scheme.

Clock skew

Assuming data is received from a central master unit, then each slave receiving the stream must adjust its playout to match that of the master, otherwise the buffer will either overrun or under run. While this can be done quite easily, problems can occur in multicast and multiple unicast. In a simple network setup there is only one master and many slaves. IP makes it possible to send and receive audio feeds to and from anywhere in the network. The difficulty is that each sender and receiver has a different master clock; if they are sending at their basic rate they will eventually over-run or under-run the buffer. This is further complicated if multiple streams are being received from different sources. It implies that each stream must be monitored and the play-out adjusted to track the master clock.

More basic systems ignore this and simply have a strategy that allows for over-run and under-run to occur before restarting the system.

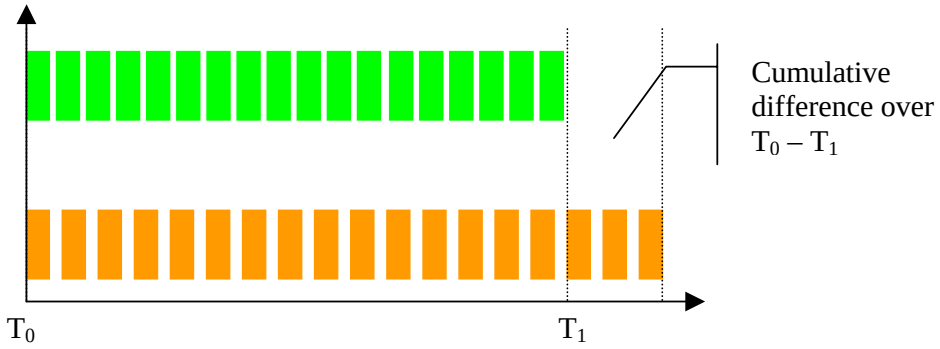

Figure 4 Clock Skew

Retransmission

This has implications for the real time operation of the unit given the processing delay and use of Transmit and Receive buffers. A retransmission must insert the retransmitted packet into the buffer at the appropriate point.

RTCP feedback can be used to signal to the sender that packets should be retransmitted. There are several benefits and problems associated with the retransmission of lost packets.

Advantages:

1) Complete correction of fault in the packet stream.
2) Very effective for small groups and one to one streaming.
3) Non-real time applications will benefit from this method.

Disadvantages:

1) Transmission overhead increases as packet loss increases. Packet loss is generally due to a high degree of network traffic. Retransmission will only exacerbate the problem in this scenario.
2) Retransmission incurs at least three passes through the network - the initial send, the retransmission request and the reception of the new packet. In networks where the latency is high this prevents retransmission being used for live streaming applications.

In networks using multicast or multiple unicast the retransmission of packets from multiple sources will contribute a substantial overhead to the network traffic and may also swamp a single sender unit if all slaves request retransmission of a single packet.

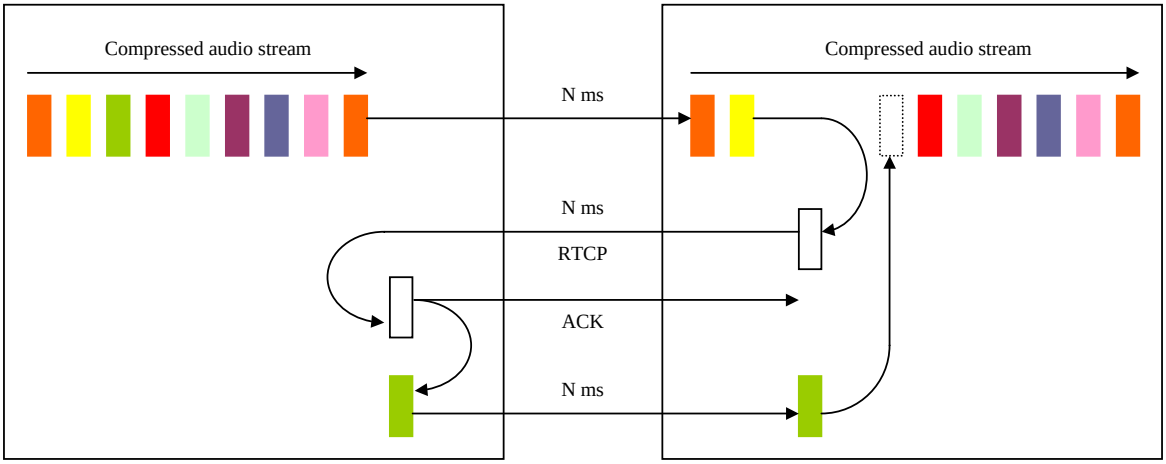

Figure 5 Retransmission Strategies

If retransmission strategies are used in real time applications to ensure no audio dropout, then they will have to incorporate enough buffering to compensate for all possible packet loss scenarios.

Audio Algorithms

Having thoroughly investigated the intricacies of IP as a method of moving program content from Point A to Points B, C, D and through to Point X, the next step is to look at the best method of layering in audio on top of the transport stream. In essence there are two options – PCM / linear or using compression to reduce bit rates. Within compressed there are two sub-options - perceptual or ADPCM.

PCM or linear audio is well defined in terms of the audio - what you get in should be what you get out, assuming there are no problems relating to analog-to-digital conversions, signal-to-noise ratios or quantization issues. The compelling reason not to choose linear is directly related to the data bandwidths required.

A stereo signal sampled at 44.1 kHz, with a word depth of 16 bit, will require a data rate of 1.411Mbit/s (plus 10 – 15% overhead and additional for FEC and synchronization algorithms). This data rate bandwidth will cause stress on the IT network passing the data. If the broadcaster adds in additional channels (5.1 or more Stereo signals), deepens the word depth to 24 bit and increases the sampling frequency to 96kHz (or even 192kHz for the small furry animals that happen to be listening), it soon becomes apparent that what was a benign solution has now turned into a network nightmare.

Compression
Making the decision to use compression opens up an interesting argument. Two options are available:

1. The perceptual based algorithms using psycho-acoustic based principles that can generally be described as “Lossy”. Some examples are MPEG Layer II, MPEG Layer III (MP3) and AAC (including the myriad of derivatives). These algorithms are heavily processor hungry and remove content that is perceived to be irrelevant. As such, they result in content that vaguely resembles the original (especially after several passes) and has a long latency i.e. 50+ milliseconds.

2. The other option is to use the relatively non-destructive Enhanced apt-X algorithm, which is based on ADPCM principles. This algorithm offers a low delay of less than 2 milliseconds and has exceptional acoustic properties. These acoustic claims have been confirmed by independent listening tests (The most recent listening test undertaken was with a group (approximately 20) of Chief Engineers from the GWR group (now GCAP after they merged with the Capital Group). This was a double blind listening test with 10 audio samples (different genres (Classical, Pop), a cappella, spoken voice (Male / Female). Enhanced apt-X was tested along with, MPEG and J.41. Enhanced apt-X was shown to be indistinguishable from the original PCM). The Enhanced apt-X algorithm can also offer word depths of 16, 20 & 24 bit, thus significantly improving the dynamic range to greater than 110dB.

Working on the assumption that the IP transport stream will naturally introduce a minimum delay of 20+ milliseconds, the latency of the compression algorithm then becomes an imperative when considering the design of a broadcast network. In essence, using a perceptual coder will render the solution unusable for any level of live event, talkback or off-air monitoring. Whereas using Enhanced apt-X will offer broadcasters a viable alternative IP, at best case, will have a delay of 20 milliseconds. MPEG (of various guises) will have a delay starting at 50 milliseconds. In total this will be a minimum of 70 milliseconds, end-to-end or 140milliseconds for a round trip. Enhanced apt-X has a delay of 1.87 milliseconds; add that to the 20ms of IP, then we have less than 22ms end-to-end or fewer than 44ms for a round trip.

Along with the well-documented features of low latency and audio performance, Enhanced apt-X also has an embedded word pattern to help connection and synchronization. This feature, AutoSync, aids the ability to quickly synchronize i.e. 3 milliseconds on start up or drop out. AutoSync, used in conjunction with the Predictive nature of Enhanced apt-X (allows for the masking of lost (small) packets), can act as an alternative to an FEC algorithm. Thus reducing overheads and additional latency due to the complexity of the process.

On a more subjective issue, using multiple passes of a perceptual codec (for example, consider the final emission for HD Radio or DAB) will result in content heavy with artifacts. Ultimately these will cause “listener fatigue,” swiftly followed by users tuning to another station that sounds better because it uses less destructive coding algorithms.

Summary:

IP as a transport mechanism for broadcasters is here to stay because it allows radio broadcast networks to bundle their audio with data, reduce operational costs and amalgamate IT and Audio into a single department. However, these massive and well-defined advantages come with some safety warnings - networks should be well managed, packets should be prioritized and correct choices should be made with regard to audio compression. Assuming all these boxes are ticked, then broadcasters will enjoy the benefits of the transition away from synchronous networks without running into serious problems.