TIMELY AND SECURE:
REAL-TIME PERFORMANCE CHALLENGES OF CONTENT SECURITY

Reza Rassool, Chief Engineer,
Widevine® Technologies Inc.

Abstract

Encryption, authentication, and key distribution are the mainstays of digital rights management (DRM) and conditional access (CA) systems in modern entertainment networks. In the traditional DVB CA security model[1], entitlement control messages (ECM) and entitlement management messages (EMM) are inserted into an encrypted MPEG stream. These messages are received in a timely manner by a subscriber device, to enable it to access the stream data. In more modern delivery networks, watermarking, fingerprinting, and digital copy protection are additional processes that have been inserted into the pipeline to secure the business of on-line entertainment. All these security processes introduce measurable temporal distortions in bandwidth, latency, and jitter to the smooth flowing of entertainment content to subscribers. While basic real-time requirements stem from linear broadcast applications, file-based delivery imposes a new set of constraints that challenge engineers to deliver content in a secure and timely manner. File-based distribution calls for security processing that scales, persists and is faster than real-time. This paper quantifies the potential temporal distortions in the DVB CA security model, detailing the perceptible effects on channel change time, temporal jitter and latency.

INTRODUCTION

Television and movie content is, by its very nature, a temporal experience. Frames and samples are presented to us in rapid succession to give the illusion that we are observing objects in motion and listening to sound. The audio-visual illusion relies on precise timing of the delivery of each sample of content. It is this most basic illusion that must be maintained to ensure that the delivered content is received in the condition intended, and achieves its potential value in maintaining the attention of its audience. Traditional over the air broadcast networks delivered a consistent stream that did not suffer from the temporal distortions of modern packet-based networks.

User Perception of Timing

Psychologists would identify audio video timing as a *hygiene* factor. Herzberg suggests a model for human motivation[2] wherein certain essential factors are considered pre-requisites. Only once these hygiene factors are satisfied can we be motivated by other factors. Herzberg's original work concerned the motivation of employees. Since then, Herzberg's work has been applied to consumer motivation. Consumers are typically motivated to visit restaurants based on the menu rather than the quality of service. A certain level of service is a pre-requisite; in the same way viewers expect to enjoy a movie where each frame and sample is delivered on time.

It turns out that hygiene factors cannot motivate a consumer. But, their deficit can certainly demotivate, in the same way that poor service would negatively affect the enjoyment of the meal - no matter how good the menu. For the audience of this paper it is especially important to understand the hygiene factors of the digital television business. One of these is audio video timing. Surprisingly few studies have explored the area of user perception of temporal distortion in audio and video.

Distortion by Frame rate changes

The illusion of motion can be maintained at quite a low frame rate. It is surprising that utility is found in video conferencing systems operating at less than 10 fps. The content frame rate should not be confused with the display frame rate. Even though the old silent movies had 16 frames of content per second, the projectors would display each frame three times resulting in a display frame rate of 48 fps. The display frame rate is critical and is related to, but not identical to, a physiological concept called the flicker fusion threshold or flicker fusion rate. Light that is pulsating below this rate is perceived by humans as flickering; light that is pulsating above this rate is perceived as being continuous. The exact rate varies depending upon the person, their level of fatigue, the brightness of the light source, and the area of the retina that is excited. Few people perceive flicker above 75 hertz for CRT monitors. A flicker free display is a hygiene factor. The content rate and the display rate must be controlled independently. The display rate must be tightly locked to a steady clock, while the content must be delivered at a rate that was intended to maintain the illusion of motion. It is well known that content played at the wrong content frame rate adversely affects the viewer experience. Even the most dramatic Lillian Gish movies seem comical when played at 24 fps. But at the original 16 fps, the content delivers the intended impact.

Distortion by Audio frequency changes[3]

In many respects, audio timing needs to be more stringent than video timing. In the range 1kHz to 8kHz, the human ear can detect a pitch shift that results from a change of frequency as little as 0.2%. Wow and flutter is particularly audible on music with oboe or piano solo. While wow is perceived clearly as pitch variation, flutter can alter the sound of the music differently, making it sound 'cracked'. There is an interesting reason for this. A 1 kHz tone with a small amount of flutter (around 0.1%) can sound fine in an echo-free environment, but in a reverberant room constant fluctuations will often be clearly heard. [4] These are the result of the current tone 'beating' with its echo. What is heard is quite pronounced amplitude variation, to which the ear is very sensitive.[5]

Distortion by Jitter and Latency

While over-the-air broadcasts deliver isochronous streams in real-time, modern networks burst packets of data that need to be buffered in memory for variable lengths of time and need to be processed to differing extents depending upon the type of data in the packet.

Jitter is caused when a processing element in the pipeline operates in bursts. The result is that the time each data packet spends in the element, from input to output, is not constant. Even though the long term flow rate through the processing element may be constant, the instantaneous rate fluctuates. Jitter causes a problem for subsequent downstream processing elements. Either their buffers overrun due to receiving a burst of several packets, or their buffers run empty due to gaps between the bursts of data packets. Jitter is resolved by larger buffers or by ensuring that each processing element operates in a timely manner. Timestamping each packet on arrival and holding the processed packet, until a fixed period after the timestamp, before outputting it, reduces jitter to the resolution of the timestamp but introduces a fixed latency.

Buffer overrun in an element often results in packet loss, while buffer under-run requires the processing element to deploy an under-run strategy. An MPEG decoder, for instance, would repeat the previous frame at the output if the input buffer under-runs. The viewer sees a freeze frame.

Latency is a fact of life in transmission systems. Latency must be constant so that end-to-end propagation delay can be used to schedule live or real-time events. Provided that all elementary streams undergo the same latency then the content is delivered as intended. Viewers with both cable and satellite systems may notice the different end-to-end delays of each service when they switch from one to another. Buffering, introduced to smooth out jitter, adds to the end-to-end delay.

Distortion by AV Synchronization drift

Reeves and Voelker[6] reported on a Stanford University study. When audio precedes video by 5 video fields, viewers evaluate people on television more negatively (e.g. less interesting, more unpleasant, less influential, more agitated, less successful). This difference is not large, but it is statistically significant. Viewers can accurately distinguish between a television segment that is in perfect synch, and one that is 5 fields out of synch. Viewers cannot accurately tell the same segments are 2.5 fields out of synch but their evaluation of content is negatively affected.

In 2003, an ATSC Implementation Subcommittee (IS)[7] studied the issue of AV synchronization. They said that the overall audio-video synchronization error is the algebraic sum of the individual synchronization errors encountered in the chain. While a given synchronization error may cause either a positive or negative differential shift in audio video timing, the video signal is typically subjected to greater delay than the audio signal, and the tendency is therefore toward video lagging behind audio.

IS finds that under all operational situations, at the inputs to the DTV encoding devices, the sound program should be tightly synchronized to the video program. The sound program should never lead the video program by more than 15 milliseconds, and should never lag the video program by more than 45 milliseconds. In MPEG-2 the end-to-end delay from an encoder's signal input to a decoder's signal output is regarded as constant.

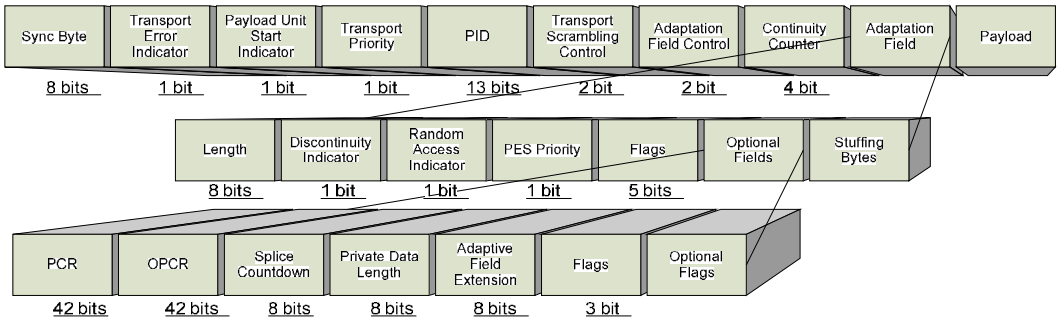


Figure 1 PCR in adaptation field of MTS header

This end-to-end delay is the sum of the delays from encoding, encoder buffering, multiplexing, transmission, de-multiplexing, decoder buffering, decoding, and presentation. Presentation time stamps are required in the MPEG bit stream at intervals not exceeding 700 milliseconds. The MPEG System Target Decoder (STD) model allows a maximum decoder buffer delay of one second. Audio and video presentation units that represent sound and pictures that are to be presented simultaneously may be separated in time within the MPEG transport stream (MTS) by as much as one second. In order to produce synchronized

output, IS finds that the receiver must recover the encoder's System Time Clock (STC) and use the Presentation Time Stamps (PTS) to present the audio-video content to the viewer with a tolerance of +/-15 milliseconds of the time indicated by PTS.

MPEG TIMING MODEL

MPEG supports timing metadata that may be inserted at encoding of the elementary streams and at packetizing of the MTS. These timestamps are read by the decoder to ensure the real-time performance of the stream. An MPEG-2 encoder includes System Time Clock (STC) as a reference time.

The system adds an STC value to the coded AV data as a time stamp for each unit of presented information, and then multiplexes the resultant data. Next, the multiplexing system inserts a reference clock so that the receiver may regenerate the STC on decoding. The receiver places each unit of coded data in a buffer to generate a delay, then decodes and presents the data unit when its time stamp matches the STC. This process corrects the temporal offset between the video and audio streams caused by multiplexing.

Timestamps, tables and their constraints

The MPEG-2 System Standard defines two types of timestamp that are added during encoding: Presentation Time Stamp (PTS), indicating time of presentation, and Decoding Time Stamp (DTS), indicating decoding start time. The multiplexed MPEG transport stream (MTS) includes a Program Clock Reference (PCR), a timestamp marked periodically in the adaptation field of the MTS header.

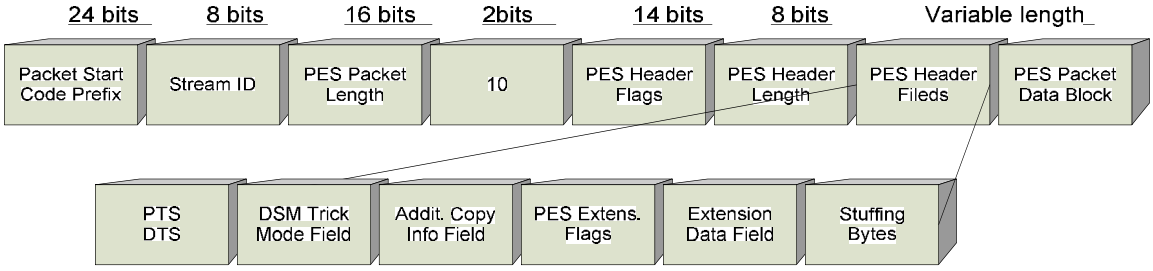


Figure 2 PTS and DTS in PES header

The PCR allows the receiver to regenerate a system time clock to match the timing of the encoding process. The PTS and DTS timestamps are sent in the PES headers. The Program Map Table (PMT) associated PIDs with program(s). The Program Association Table (PAT) associated program number with PMT. The Conditional Access Table (CAT) associated PIDs with private streams.

ATSC and DVB tighten the MPEG-2 constraints on timestamps and tables.

Clock Recovery Schemes[8]

An ideal MPEG decoder would implement a numerically-locked loop (NLL) to regenerate the 27MHz system clock from incoming PCR values. 27MHz was chosen because that is the frequency used to drive the video display electronics. Since the DVB specification requires that PCR values are inserted in the stream, at most, 40ms apart this requires the NLL to operate between 25Hz and the MTS packet frequency, 2500Hz (for a 3.75Mb/s stream).

188 byte packets are received, demuxed, and placed in the buffers according the PIDs. Each frame of MPEG data in the PID buffer contains its own timestamp. Once decoded each frame is stored in the display buffer tagged with its own PTS.

As shown in figure 3, the NLL[9] contains a 27MHz VCXO (voltage controlled crystal oscillator), a variable-frequency oscillator based on a crystal which has a relatively small frequency range. The VCXO drives a forty-eight bit counter. The state of the counter is compared with the contents of the PCR and the difference is used to modify the VCXO frequency. In practice, the transport stream packets will suffer from transmission jitter, and this will create phase noise in the loop. This is removed by the loop filter so that a large number of phase errors are averaged over time before affecting the VCXO. The 48bit counter is divided by 300 to produce a 33bit counter. The 'decode' module retrieves MPEG data from the PID buffer when the 33bit counter matches the DTS of the frame. The 'display' module similarly retrieves a decoded frame from the display buffer when its PTS value matches the 33bit counter.

A heavily damped loop will reject jitter well, but will take a long time to lock. Lock-up time can be reduced when switching to a new program if the counter is jammed with the first PCR value in the new program.

	description	MPEG-2	ATSC	DVB
PTS	90 kHz clock 33bit counter	Interval <0.7s Jitter	Interval <0.7s Jitter <15ms	Interval <0.7s Jitter <15ms
DTS	90 kHz clock 33bit counter	Interval <0.7s	Interval <0.7s Jitter <15ms	Interval <0.7s Jitter <15ms
PCR	27 MHz clock 48bit counter	Interval <0.1s Jitter <4ms	Interval <0.1s Jitter <4ms	Interval <40ms Jitter <0.5ms
PAT	Lists PMT PID	Interval not specified	Interval <0.1s	Interval <0.5s
PMT	Lists prog. PIDs	Interval not specified	Interval <0.4s	Interval <0.5s

Table 1 Timestamps, Tables and their constraints[10]

In legacy receivers, the NLL module could not be implemented in software by the main CPU so it was either implemented in hardware or was radically simplified. Both choices have led to issues in the performance of legacy receivers. One simplification was to replace the NLL with a 48bit counter driven from the video electronics 27MHz clock. This results in a clock that does not dynamically adjust to reproduce the original timing of the encoder.[11]

MPEG TS OVER UDP

A typical IP packet carrying MPEG-2 video-streaming data consists of seven MTS packets, each containing 184 bytes of payload and 4 bytes of header. This results in 1316 bytes, plus the packet overhead – 8 bytes for the UDP header, 20 bytes for the IP header, 14 bytes for the Ethernet header. (Fig.4)

What temporal distortions result from packet loss?

UDP is an unreliable transmission mechanism. Packets can be lost. Loss of IP packets may occur for multiple reasons — bandwidth limitations, network congestion, failed links, and transmission errors. Packet loss usually results in bursty behavior, commonly related to periods of network congestion. Depending on the type of transport protocol

used for the video streaming, a packet loss will have a different impact on the quality of the perceived video. When UDP is used, the lost packets will directly affect the image, as the information cannot be recovered and the image will simply be corrupt or unavailable. When using TCP, a packet loss will generate a retransmission, which can produce a buffer underflow and, consequently, a possible frozen image.

The loss of one UDP packet results in the loss of 7 MTS packets. In a 3.75Mb/s CableLabs stream, one UDP packet represents about 2.8ms of content. Assuming also a 192kb/s audio stream, there will be 18 MTS packets containing video, for each one containing audio. This means that a single UDP packet has a high chance of containing no audio.

The loss of a silent UDP packet results in the video stream jumping forward in time by 2.8ms while the audio stream is undisturbed. Now, a well behaved MPEG-2 decoder should present each video or audio "frame" at the scheduled time – when its PTS/DTS values match the recovered system clock.

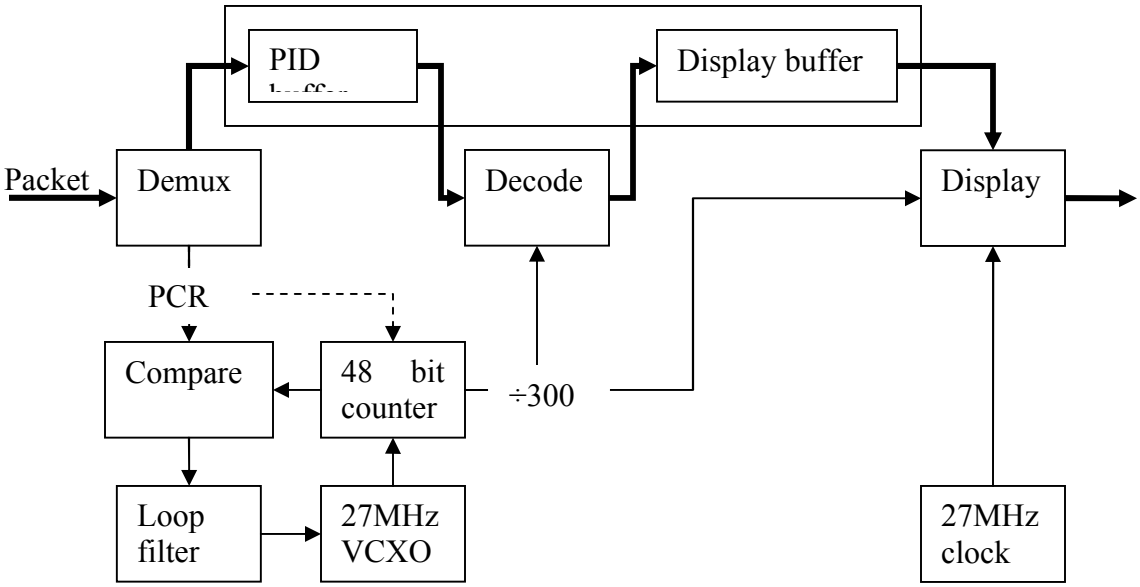


Figure 3 Clock regeneration with NLL

Legacy receivers that omitted the NLL suffer loss of synchronization. The PCR value is looked at only when a program switch occurs and thereafter the system clock runs locked to an internal reference such as the CPU clock or video display clock. This means that packet loss results, inexorably, in loss of AV synchronization. After 15 lost UDP packets, the drift is noticeable. In a network with just 0.01% loss, the sync drift would be noticeable after one hour of continuously viewing the same channel. A simple user controlled remedy is to reset the counter. This is achieved by switching to another channel and then switching back. Try it at home!

Variable MPEG processing delays

The MPEG-2 specification states that video or audio elementary-stream access units that do not contain B pictures are to be transferred immediately from the main buffers to the decoders at the time denoted by its PTS. The STD then decodes and outputs the data in the main buffers when the STC matches the PTS.

However, a video elementary stream that includes B-pictures requires that I and P pictures be decoded before decoding the B-pictures, and it is for this reason that the decoding time and presentation time of I or P pictures differ. In particular, the specification states that I or P picture data are to be transferred immediately from the main buffer to the decoder at the time denoted by DTS. The system decoder then decodes and outputs the I-picture or P-picture in the main buffer when the STC matches the DTS. Thereafter pictures are held in a re-order buffer until its PTS matches the STC.[12]

This means that packet loss can cause drastically different perceived distortions whether the lost packet contains I, B or P data. Higher compression results in greater temporal distortion. A lost UDP packet from a 700kb/s H.264 stream represents 15ms!

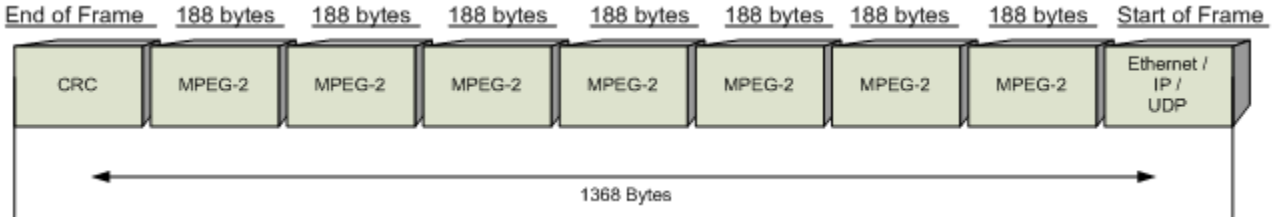


Figure 4 UDP packet contains seven MPEG transport stream packets

DVB-CA SECURITY MODEL

The DVB-CA security model comprises a combination of scrambling and encryption to prevent unauthorized reception. Encryption is the process of protecting the secret keys that are transmitted with a scrambled signal to enable the descrambler to work.

ECM

The scrambler key, called the control word (CW) must, of course, be sent to the receiver in encrypted form within an entitlement control message (ECM). The CW is valid for a particular crypto-period (CP) which is typically 10 seconds long. ECMs must be received and the CW extracted and decrypted in advance of MTS packets in the associated crypto-period. If the ECM is not available for the associated crypto-period in time, then the content cannot be decrypted and the subscriber will suffer service loss.

The ECMs are repeated every 0.1s to ensure that the stream is still decryptable even under severe packet loss. The ECM stream takes up about 1% of the stream bandwidth. The ECMs are transmitted in a separate PID that is multiplexed in with the original stream. The original stream is already time-stamped. The injection of ECMs causes jittering in the PCR values of the original transport stream. An important feature of DVB-CA multiplexer is to perform PCR correction to compensate for this jitter. ECMs of moderns CA systems now carry more that just control words. Watermark metadata, extended copy control information, and other metadata would cause the ECM to grow beyond a single MTS packet.

In the absence of PCR correction the packets could arrive in an untimely manner – outside the 0.5ms jitter spec of the DVB standard.

EMM

The CA subsystem in the receiver will decrypt the control word only when authorized to do so; that authority is sent to the receiver in the form of an entitlement management

message. This layered approach is fundamental to all proprietary CA systems in use today. In a traditional DVB network the EMMs are transmitted in-band, in another PID multiplexed with the content. In IPTV networks the EMMs can be sent out-of-band via a reliable communication channel.

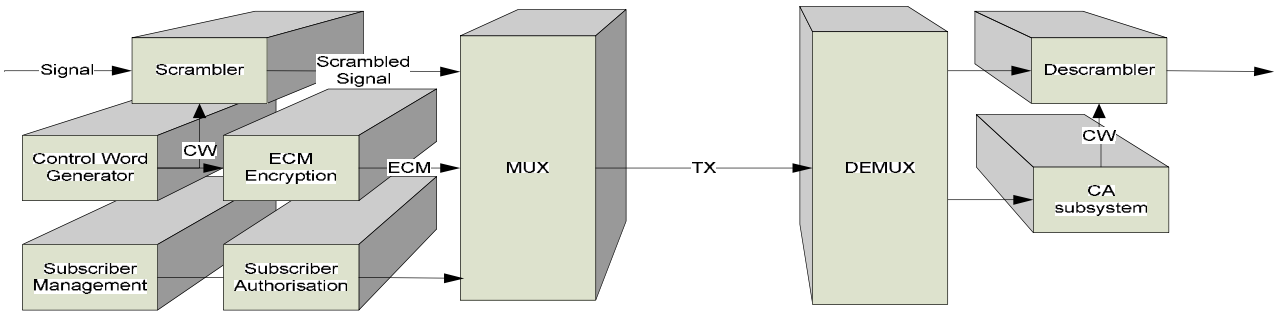


Figure 5 DVB-CA model

SECURITY TIME CHALLENGES

Security affects the timing of streams, by introducing jitter and consuming transmission bandwidth, through the insertion of ECM packets.

Client-side CPU load

The client CPU is burdened with an increasing load to support more sophisticated security systems including processes that insert watermarks into the content, monitor content to generate fingerprints, and monitor the receiving device to ensure that no theft is occurring. This occurs as the client device labors under a six-fold load increase in the transition from standard to high definition. It means that timing is becoming ever more critical in modern client devices.

Timely arrival of keys

In the case of linear content the EMMs are typically sent to the client at the time of subscription to the channel or bouquet of channels. EMMs are revoked and re-issued to rotate entitlement keys -typically on a monthly basis. Unidirectional networks such as satellite need to handle EMMs with special care. The carousel transmission of the EMMs has to strike a balance between reliability and security.

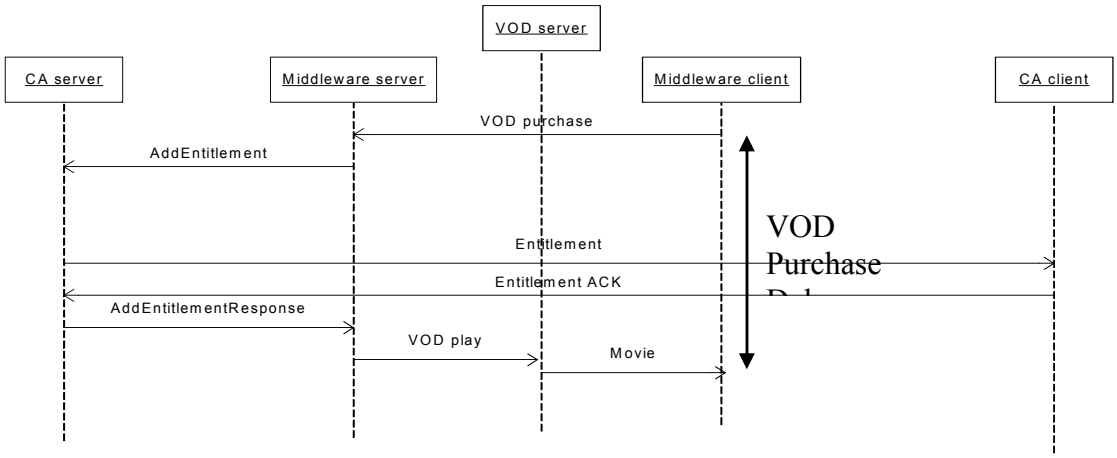
The service operator must ensure that all the subscribers receive the EMMs for the services to which they are entitled, while also being careful not to expose an EMM for too long to hackers.

Blocking EMMs revocation, an obvious hack to thwart key rotation, would allow a subscriber to access a service long after the subscription has expired.

In bidirectional IPTV networks, with reliable TCP/IP communication, the EMMs can be issued on a just-in-time schedule. EMM acknowledgements can also become part of the business logic of the service.

In the case of an impulse VOD purchase, the EMM cannot be pre-staged on the subscriber's set top box. To reduce 'VOD Purchase Delay,' the latency between the client's purchase request and the start of the movie, the timing of the security communications needs attention.

Figure 6 VOD sequence diagram showing unmitigated VOD Purchase Delay



Ordinarily the delay would be several seconds. One method to mitigate the delay is to leave a leader of the movie in the clear. The leader duration is just longer than the maximum VOD purchase delay. An even more secure solution is to encrypt the leader with a key that is only issued to those clients that have subscribed to the VOD service. This approach means that the Middleware server can issue the 'VOD play' command to the VOD server as soon as it receives the VOD purchase message. The subsequent CA communication will then occur in parallel with the playing of the leader. The EMM for the movie will arrive at the STB in time to decrypt the remaining duration of the movie.

Timed entitlement

Normally EMMs are issued and revoked by the CA server. In this case the time is maintained by the server. In the advent of secure processors, secure memory, and secure clocks, the CA client can be safely implemented to operate with a higher degree of autonomy. In these more secure devices the client can receive EMMs with richer rights expressions. A simple timed entitlement includes the start and end time. The client will only use the entitlement after the start time and will purge it after the end time. This would allow a customer to download a movie onto a mobile device and watch it on a plane or boat, disconnected from the CA server.

Secure time

Manipulation of the client clock is a well-known hack to retain expired service. Secure clocks are tamper proof and are protected against unauthorized changes. The clock can be set through a secure protocol each time the client connects with the CA server. Time should be maintained independently of local time-zone, using UTC/GMT, to avoid errors caused by the CA client and CA server operating in different time zones. The CA server should itself obtain time from a trusted NTP source.

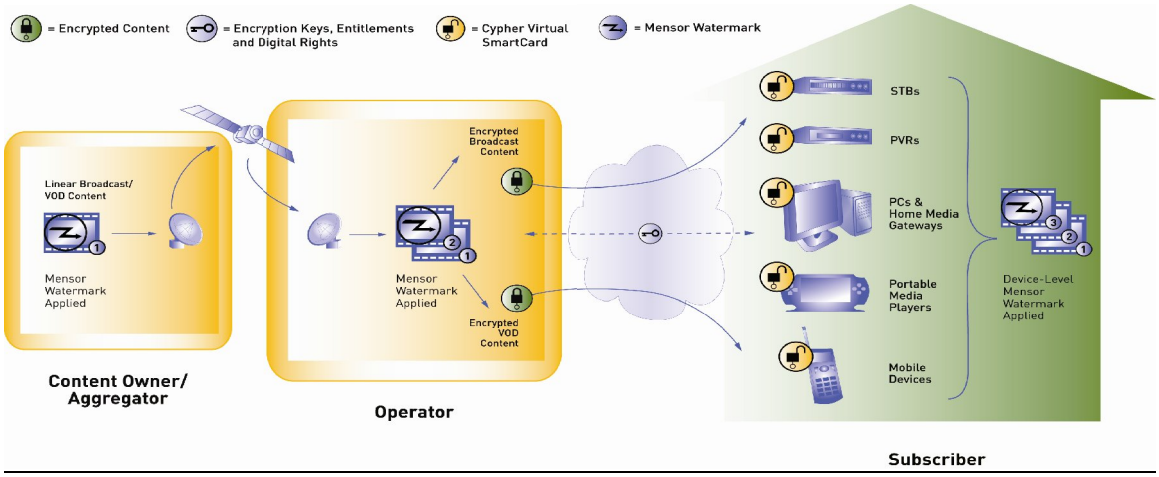


Figure 7 Hybrid CDN

File-based transmission

Modern content delivery networks (CDN) transmit linear and file-based content. These hybrid networks aggregate content and distribute files to remote service operators that each serve separate populations of subscribers. In a network, as in figure 7, the security processing has the traditional real-time requirements at the point of displaying the content on the client device. However the rest of the network treats the content as files. Files are transmitted from the aggregator to the operators as fast as the satellite transponder allows. In this environment files may be secured by different conditional access systems in different legs of the pipeline. Each file may need to be encrypted at the aggregator, decrypted at the operator and then re-encrypted with the operators CA of choice. Then the file is served to the subscriber where it is decrypted, decoded, and displayed in real-time. At the operator, however, bulk crypto processing should happen as fast as possible as files are pitched from the aggregator at 20 times faster than real-time. As described earlier in the paper, legacy CA systems that have implemented their stream parsing and crypto-processing in hardware have built their systems around the real-time clocking requirements. These real-time scramblers and descramblers cannot be easily retooled for the task of bulk encryption or decryption required to secure a file-based CDN.

TVN Entertainment[13], a VOD service operator, delivers 5000 hours of file-based content per month to affiliate operators around the country.

This network is both encrypted and watermarked. In 2007, benchmarking tests TVN Entertainment showed that an off-the shelf single rack unit server, running Widevine Cypher® DRM software, can encrypt or decrypt one gigabyte of MPEG file in 1.5 minutes while a traditional real-time DVB scrambler takes 35 minutes.

CONCLUSION

At IBC07, SMPTE and EBU[14] jointly declared:

The current methods of timing and synchronization for television, audio and other moving picture signals rely on standards that have been in place for more than 30 years. While these standards have proven to be robust solutions that have served the industry well, they are predicated on technologies that are becoming increasingly inappropriate for the

digital age with, for example, networked content sharing or higher frame rate HDTV image formats; they now impose unacceptable limitations for the future.

The time constraints on MPEG streams are onerous. The addition of security processing provides an extra timing challenge for control logic of servers and clients. Legacy CA systems suffer from a lack of flexibility in dealing with the timing of a faster than real-time CDN. The oversimplification of the timing logic or its implementation in hardware precludes traditional CA crypto-processors from reaching the performance of modern software bulk encryptor / decryptors. The evolving security landscape is challenged with emerging requirements for watermarking, fingerprinting and copy protection. A software solution is best placed to rise to these challenges - both timely and securely.

References

[1] ETSI (1997), TS 101 197-1 Digital Video Broadcasting (DVB) DVB SimulCrypt Part 1: Head-end architecture and synchronization.

[2] Frederick Herzberg, 'The Motivation to Work' (1959), Work and the Nature of Man (1966), The Managerial Choice (1982); and Herzberg on Motivation (1983).

[3] E. Alexandra Athos et al (2007), Dichotomy and perceptual distortions in absolute pitch ability, PNAS, September 11, 2007, vol. 104, no. 37,14799

[4] Audition, by Pierre Buser and Michel Imbert, English translation by R. H. Kay, MIT Press, Cambridge MA, 1992

[5] CD Audio Demonstrations, by A. J. M. Houtsma, T. D. Rossing, W. M. Wagenaars, Philips 1126-061.

[6] Reeves and Voelker (1993), Effects of Audio-Video Asynchrony on Viewer's Memory, Evaluation of Content and Detection Ability, Stanford University.

[7] ATSC (2003), ATSC Implementation Subcommittee Finding: Relative Timing of Sound and Vision for Broadcast Operations, Doc. IS-191, 26 June 2003.

[8] Tryfonas and Varma, Timestamping Schemes for MPEG-2 Systems Layer and Their Effect on Receiver Clock Recovery, IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 1, NO. 3, SEPTEMBER 1999, Page 251

[9] Watkinson (2001), The MPEG Handbook, Page 333, Focal Press

[10] Isnardi (1999), MPEG-2 Systems, Sarnoff Corporation, August 25, 1999

[11] SS. Bindra (2006), Studio Systems, July – August 2006

[12] Yoshimura (2002), Technologies and Services on Digital Broadcasting (5), Broadcast Technology no.11, Summer 2002

[13] Dom Stasi (2007), Broadband Business, CED Magazine, Sept 2007

[14] EBU, SMPTE announce joint task force on time, synchronization, Broadcast Engineering, Sep 15, 2007