anevia®

YOUR NEXT GENERATION TV, NOW

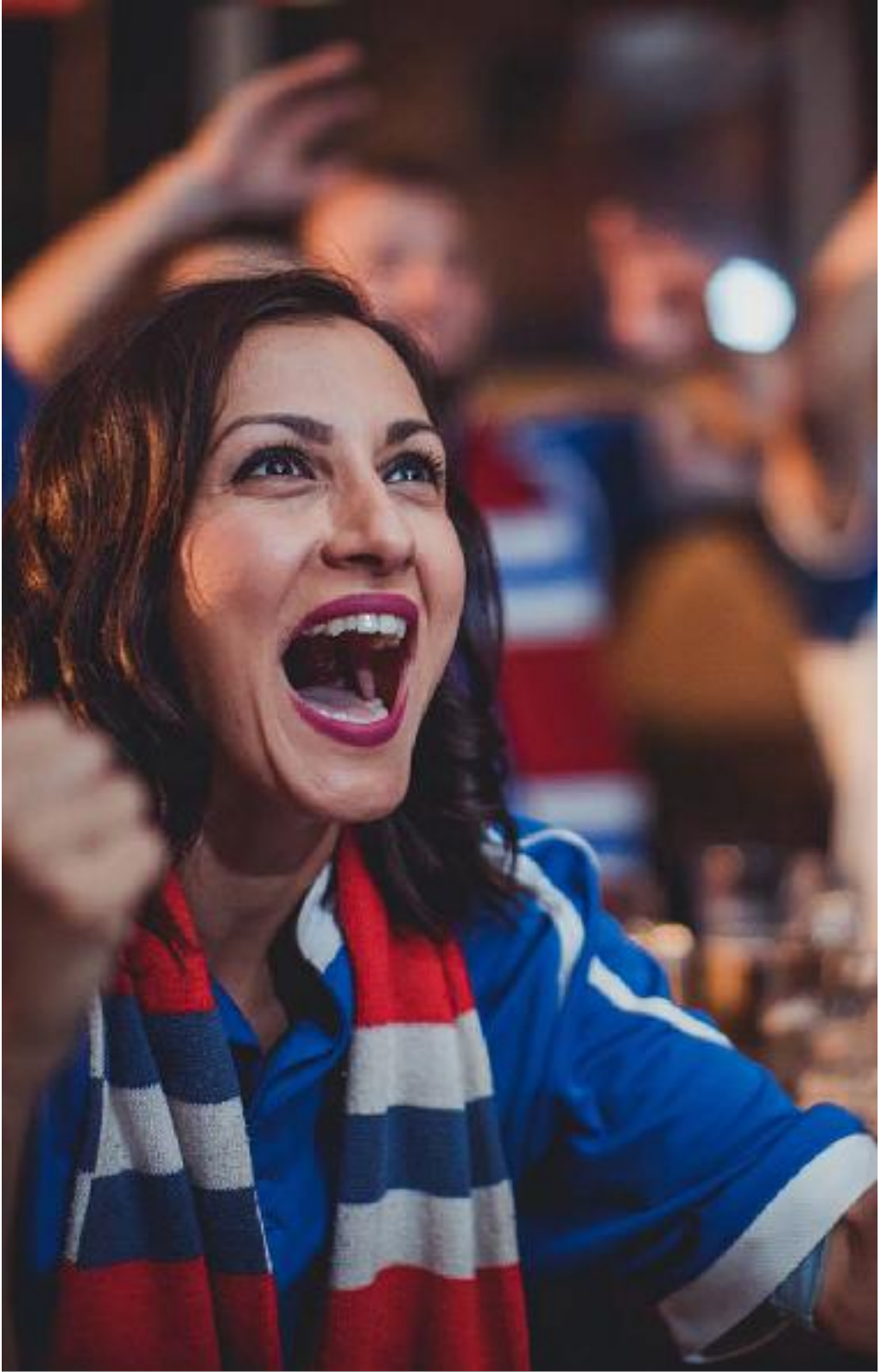
What the 2018 World Cup has taught us about video streaming.

With reduced delay, OTT streaming is set to become the norm.

Part 1 – The business perspective

TABLE OF CONTENTS

anevia®
YOUR NEXT GENERATION TV, NOW



1. INTRODUCTION

This business-focused eBook is the first of a two-part series.
Part 2 will address the technology aspects of stream delay reduction.

Over the past eight years, more and more people have watched FIFA world cup football matches over the Internet, reaching new highs every year. But this year, it's not just the growth that was spectacular, but absolute numbers: online streaming of FIFA 2018 smashed all records with 70 million viewers in a country like India, where football isn't a national sport.

The disruptive novelty is that for most millennials, streaming is their A-plan, and broadcast is the backup. By 2022, when Qatar hosts the FIFA event, this could potentially go way beyond millennials to reach the general population. As a mainstream solution Over-the-Top will be used to 'just watch' live content.

Live streaming is just starting to enter the mainstream though, and to convince viewers it needs to deliver the same kind of experience that people are used to getting through broadcast TV. The main difference that we still see between OTT streaming and broadcast is stream delay.

This issue is becoming critical for many service providers and in this eBook, we look at why the entire value chain is adopting Low Latency strategies to reduce delay.

Definition and a brief history of how we got here

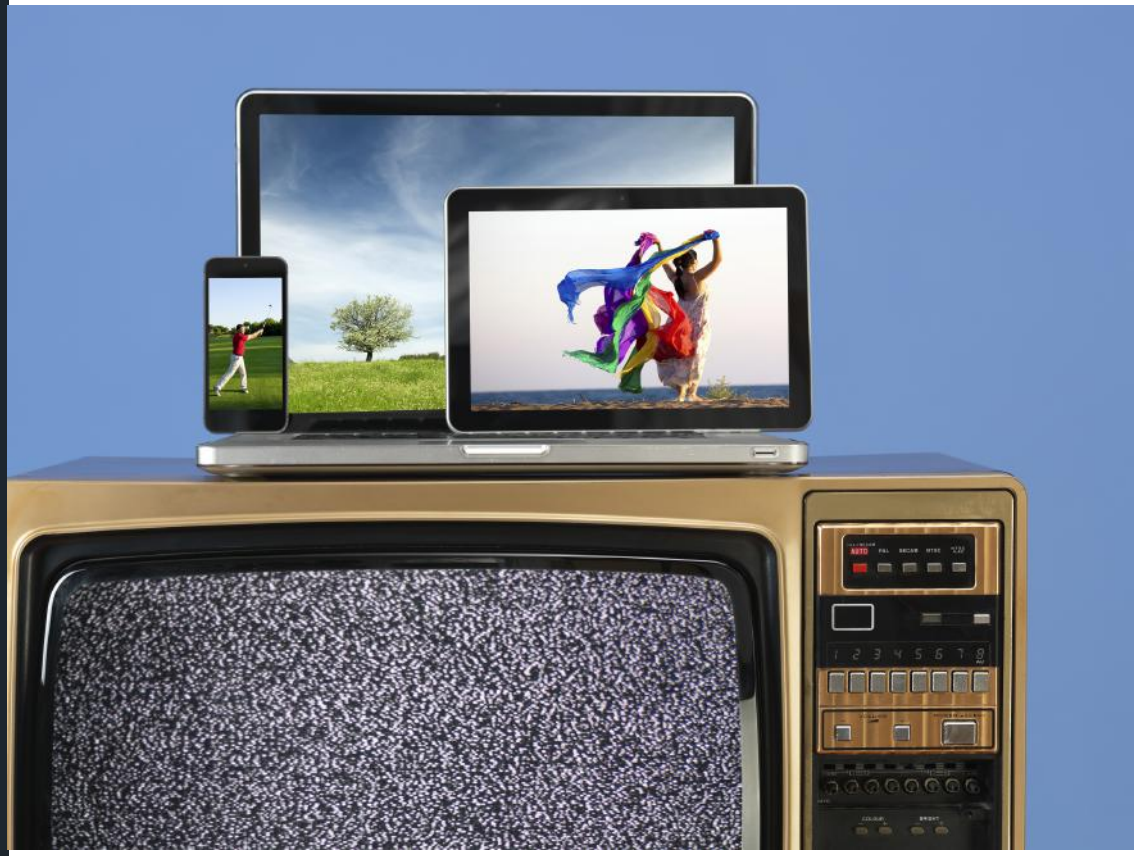
Streaming is as old (and as young) as the Internet, with TCP protocols such as RTMP underpinning early successes of technologies such as Flash.

These early approaches had a fundamental scalability issue, as they required server infrastructure to keep track of what each client was doing, making it difficult to reach millions of subscribers.

OTT was only able to take off when the industry broke free of the scalability constraint. Adaptive Bitrate (ABR) streaming was the technical breakthrough that kicked off the OTT revolution a decade ago.

Since then, its essential characteristics, from a business perspective, haven't changed. It offers the best possible stream quality whatever the network conditions and the device capability.
And today's OTT streaming architectures naturally fit into a Cloud setup, with all its inherent promise of scalability.

The 2018 World Cup has seen the largest OTT stream growth so far on a global scale, and as with previous World Cups, the issue of stream delay or latency is again a key industry topic.

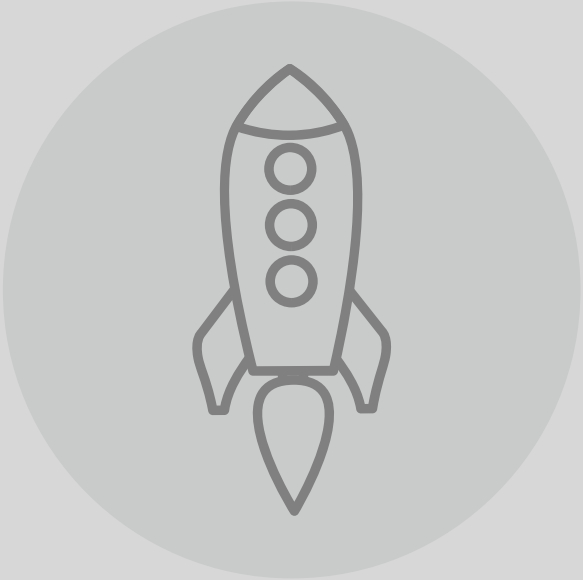


CLARIFYING LATENCY VS. DELAY

The word latency is now often used to mean video delay, sometimes also called lag. Let's try to explain.

Latency is "the delay before an action completes following an instruction to start that action". At a low level, the action can be a transfer of data, and from a user-experience perspective, it can, for example, be the time it takes for a video to start after pressing the play button.

In live TV, the time between real-time action and the video displaying on the screen in a subscriber's home should be called **delay**, but most industry stakeholders refer to it as latency. We strive to use the appropriate term.



Beyond the scope of this eBook, the responsiveness of a user interface is sometimes also referred to as latency.

Other related concepts that get mixed up with delay and latency are **channel-change time** and **buffering**.

Streaming challenges and limitations

Reliability has been the primary target of OTT streaming detractors because it can still be a potentially crippling issue for two main reasons. Immaturity of technology and web infrastructure have led to many scare stories, such as the 2015 basketball matchup between Wisconsin and Kentucky which was to be the most-watched Final Four game in 22 years – but turned out to be a bit of a fiasco online. Other scare stories involve un-anticipated iOS updates happening at the same time as a major live event over the same network.

This issue is receding fast, but even if it never entirely disappears, it could become irrelevant – meaning that OTT streaming can still become predominant.

Indeed, today's OTT streaming is unreliable only in comparison to broadcast TV. Broadcast TV is just a reference point for people up to and including generation X. Millennials and younger generations grew up on YouTube - you only need to watch the reaction of a family with parents over 40 and on the rare occasions when members all watch streaming video together.

If the service is briefly interrupted or the quality goes very low, the parents typically complain and maybe try to fix it, while the kids move to something else on their mobile while waiting for the issue to repair itself.

It is probably the adoption of streaming by older segments of the population that will drive quality requirements ever higher.

2. The strategic importance of OTT

The industry is rife with stories about how network operators such as Telcos and ISPs are forced to carry ever-increasing amounts of traffic for video streaming-based services such as Netflix. The business risk is that these network operators derive no revenue opportunities from subscriptions, transactions or advertising; but if they attempt to block or slow down such services, they could lose broadband subscribers. (This issue also evokes the net neutrality debate, which is painful for operators and beyond the scope of this eBook.)

Business opportunities do exist though, through negotiations between OTT providers and traditional pay-TV operators. Pure OTT players such as Netflix and Google's YouTube can locate storage caches inside an operator's network, which enables the operator to reduce peering and improve the end-user experience. Such services can resemble a halfway-house between OTT and IPTV.

For operators of pay-TV services, the main threat of not getting onto the OTT bandwagon is dropping out of the content loop – they risk being dis-intermediated by content owners who go directly to their subscribers using their own OTT services. Numerous examples already exist, including the trail-blazing BBC iPlayer in the UK or HULU in the US. There is also an increased threat of piracy through content redistribution over the Internet. While all forms of pirate operation threaten pay-TV, more recent pirate OTT services present a higher risk to TV operators' own OTT efforts.

Success stories like Netflix illustrate that simplicity and convenience are essential, but that OTT streaming must be reliable for continued success. OTT can and should be a fantastic opportunity for TV operators. Indeed, "TV everywhere" used to be on TV operators' long-term roadmap and ABR technology has made it feasible to deliver TV streams to many more devices over unmanaged networks with varying levels of performance and congestion. The terms "TV Everywhere" and multiscreen may have fallen away, but the concept is alive and kicking in the TV apps that broadcasters already have or are deploying – and OTT and ABR enable it.

The explosion of OTT devices driving streaming

As more and more OTT-enabled devices hit the market, more and more people are streaming live video.

According to a recent Comscore webinar, as of April 2018, 80% of Wi-Fi households in the United States own at least one OTT device. The report shows that these figures are significant because 79% of households with at least one OTT device stream content.

In a similar Conviva study, the company found that more viewers are watching TV over streaming OTT devices – such as Roku, FireTV, Apple TV, gaming consoles and Chromecast – than over mobile platforms, though that segment continues to grow too.

So, are OTT devices, in the home and mobile, exploding in popularity because people want to live stream? Or are people live streaming more, because they can?

Whatever the case, streaming live video over unmanaged networks is a reality that is fully embraced by consumers, and advertisers are not far behind.

The explosion of OTT offerings driving streaming

OTT started just as free IP telephony from a Telco perspective, or free YouTube-type service from a content operator perspective, and seemed significant neither as a threat nor as an opportunity.

Things have changed dramatically in just the last five years, and there is now a wide range of offerings from new entrants, Telcos and content providers.

Netflix, the market leader, is still experiencing strong growth – even beating its own growth forecasts. And Hulu's growth is just as astonishing, estimated at 39% year-on-year.

But beyond these pure players, so many hundreds of broadcasters, cable-TV operators and telecom operators are offering OTT services targeting specific niches. Video is now core to any OTT offering.

Operators coming from traditional broadcast TV, often serving older demographics, can still differentiate their offer by delivering a high-Quality-of-Experience OTT over unmanaged networks.

streaming video alliance

> "The Streaming Video Alliance Executive Director Jason Thibeault told us that "OTT pure-play like Hulu is still driving the live streaming industry forward. Incumbent operators are taking cues from them and offering similar packages.
> So live TV is slowly taking over as the driver from SVoD."

3. The impact of latency on user experience

There is no absolute target value for stream delay, even within a particular use case. Users accept trade-offs and their evaluation of what's 'good enough' changes over time. In the past, technologies have come to reduce unwanted delays even beyond what users needed - the case of channel-change time is a telling example of such an overshoot.

Just over half a century ago, when markets started seeing more than one TV channel, users would have to manually tune into different stations, usually by turning a knob on their TV. Then buttons came, followed by the remote control. Channel change was then almost instantaneous. With the first digital TV transmissions before the turn of the century, channel-change time became dependent on the ability to decode a new channel (i.e. wait for the first full frame).

There was little resistance as we moved from almost instantaneous channel change to around 2 seconds because that extra delay came with the arrival of so many more channels that viewers accepted the compromise.

In the early days of Internet-based TV, starting a new stream often took up to 8 seconds. Here, the trade-off didn't work anymore, and IPTV systems rapidly improved to the previous benchmark of around 2 seconds.

Some vendors, such as Microsoft, came up with dedicated fast channel-change solutions reducing zapping time to below half a second. The considerable marketing effort they put into selling these solutions created a buzz, and some operators worked on reducing channel change time to below 2 or even 1 second.

Fifteen years on, it seems that the 40-year-old 2-second benchmark is still here since channel change that takes more than 2 seconds is considered too long, but once you've got to 2 seconds or less, users don't care that much.

Coming back to stream delay, the moral of this story is that, if today's OTT video delay is often unacceptably long, there is a sweet spot, probably somewhere between 2 and 6 seconds. Getting much lower, in the milliseconds' range, would only be useful in a few extreme cases, such as flying a drone.

Beyond video streams, within the UI as a whole, latency in milliseconds remains a critical factor for responsive user experiences.

Note there are some rare cases where increased delay can be desirable for sensitive operations, as in reality TV, to be able to censor content in the event of an unexpected incident during a live transmission.

4. Video use cases and delay

While researching for this eBook, we assumed that reducing stream delay would be at the top of all operators' agendas. While this isn't yet always the case, we did find a varied mix of use cases around delay.

As you can see below, we have identified cases where ultra-low-latency is already a make-or-break feature, so reduced stream delay is not a case of a solution looking for a problem to solve. In the aftermath of the 2018 World Cup, consumers are becoming more and more aware of what stream delay is and what it could mean to their viewing experience, and industry bodies such as the Streaming Video Alliance (SVA) are taking a proactive view on reducing streaming delay.

But many operators have not yet seized the opportunities that come with reduced delay and latency, simply because they still have to recoup the investment in their first-generation streaming platforms.

Arian Koster, KPN Business owner for TV, told me that **"Stream delay isn't yet a make-or-break issue as we have the market-leading solution that others measured against, but we are looking into it in anticipation."**

"The motivation to reduce latency is to reduce churn due to poor user experience,"
Jason Thibeault, SVA Executive Director

Live sports: often important to reduce stream delay

Live sports are the most frequently cited area in which delay reduction could be critical.

The spoiler effect around football matches, where you hear neighbors cheering for a goal you haven't yet seen, intuitively speaks to us all. In most markets though, major sports sell the broadcast rights in exclusive deals. So when a delayed OTT stream has to compete with broadcast, it is usually through multiple services of the same operator. Use cases where subscribers watch the same event in close proximity, but over different distribution platforms, are still quite rare.

> Motor Sports Network France CEO Olivier Dufour told me that **"In almost all markets, sports rights are exclusive, so there are rarely any competitive issues around streaming of a particular sports event. Of course, there is abundant competition from other events, from watching a movie or simply from going out with the kids and having an ice-cream in the park!"**

> According to Alchimie's Philippe Rambourg, who previously designed OTT streaming platforms at Canal+, **"Despite potential spoilers when you hear 'goal' screamed by your neighbour and you're still 30 seconds away from the action, OTT is more often a peace-maker in the home. Sports fans can watch OTT in another room while other household members watch their content on the main screen."**

> Rambourg went on to say that in a multiscreen environment **"When subscribers use the big screen and OTT devices for different content, stream delay is not a problem. Issues occur at scale only when broadcast TV or IPTV are used alongside OTT streaming, like with a Multicam app."**

The well-known exception is for major events involving a national team (e.g. the World Cup). In such cases, a national channel carries the match, and that channel is available on several competing broadcast and OTT streaming platforms.

Although this use case only appears on a large scale every few years, it is critical for many operators. For the 2018 World Cup, the BBC warned users that there would be at least a 20-second delay on streaming compared to broadcast TV.

In the French market, there are over 20 Free-To-Air channels that can be received on all of the approximately 20 million set-top boxes available in households.

With four competing tier-1 Telcos as well as a few smaller players, viewers could use stream delay as one of the decision-making criteria in choosing a service provider.

There is even potential for a monetisable use case in which a TV operator could use managed constant stream delay to notify viewers of a major action, a few seconds before it happens on-screen.

In such a use case, an alert could be sent in real-time to the client software so that a beep or flash occurs say 3 seconds before a goal.

This notification would be particularly relevant for younger viewers who often multitask while watching an event.

> "It's one of the flaws in an emerging technology," Neil Hall BBC Sports Head of Product, in a Guardian article.

BBC

RACING POST

> Racing Post's John Matthews told me that **"People under 30 don't watch a football match all the way through. When an action is notified to them, their attention is drawn. Live is still super important, but not many eyeballs stay on it except during key moments."**

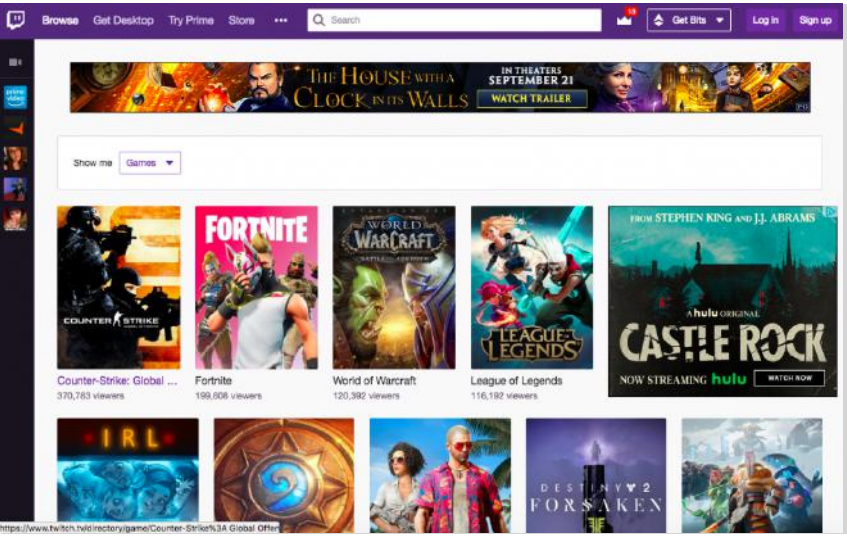
The BBC is working on new techniques to reduce delay and latency in video streaming, and demonstrated some of these at the 2018 IBC trade show. Different approaches include streaming MPEG CMAF "chunked segments"; shortening the media buffer used by players; and optimising their own media processes.

Regular TV channels: would like to reduce delay

For live TV in general and news in particular, having a stream that often gets delayed up to a minute can be an issue in the age of instant social media. However, as it takes a few seconds or more to write a post or a tweet and then at least 5 to 10 seconds for it to appear, this is not yet a major issue for most TV stations - as long as the video delay remains below 30 seconds. For now, even for breaking news, reducing delay below this 30s barrier is highly desirable, but not yet a must-have.

Re-linearised assets: not yet worth spending money on reducing delay

Two opposing trends affect channels made up of linearised assets such as music clips, TV series, kids' content or movies. On the one hand, they are becoming less relevant with the rise of on-demand platforms, but on the other, AI techniques are at last allowing personalisation to work. As we see interactivity enter into the fray more, where you can, for example, skip a clip you don't like, lower latency becomes a selling point. For all other aspects though, this sort of channel is impervious to stream delay; only start-up time can have a negative impact on user experience.



Live online game streaming: the counterintuitive exception

With millions of gamers, Twitch is the platform offering the most significant number of different live streams in the world. Stream delay on Twitch is often just below 10s, and in some cases, such as when streamers and viewers want to chat, this is barely good enough. We should also note that Twitch recently released a beta test that brings stream delay down to 2s for players with a good ISP connection. However, when players are streaming live games, such as League of Legends (LoL), that stream delay is so short that players could gain a strategic advantage by viewing the game from an opponent's perspective. In cases like these, operators can artificially increase the delay to 1 to 2 minutes.

5. Emerging trends in streaming

Now let's take a look at the innovations and changes already happening around us.

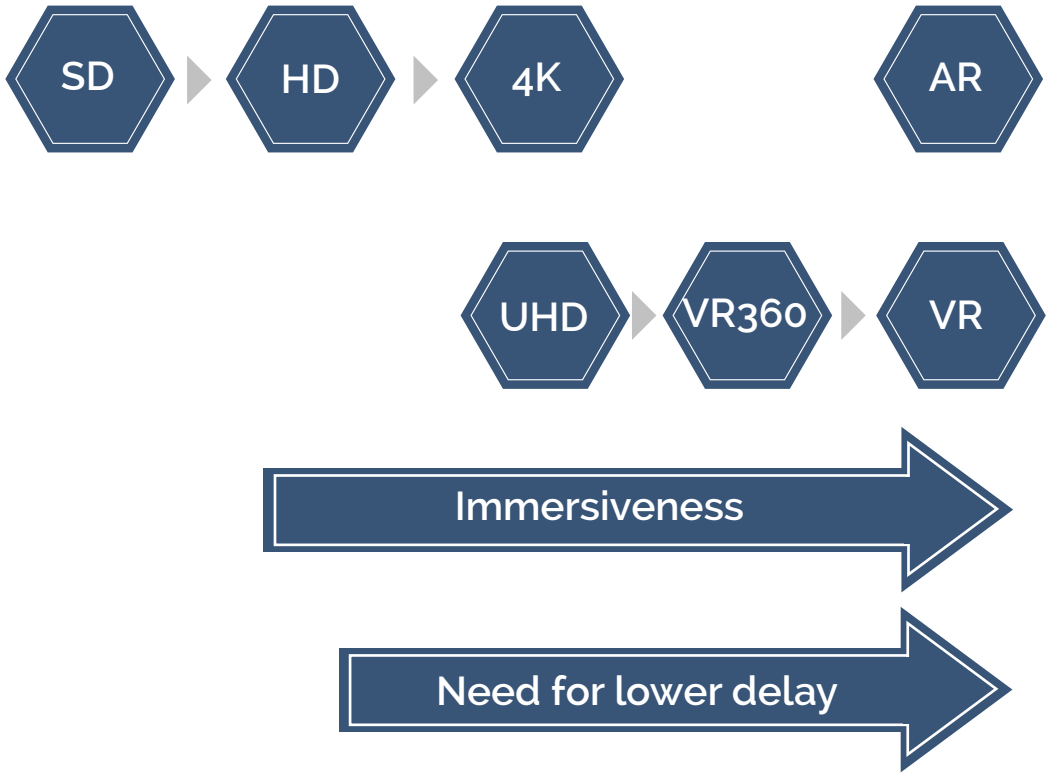
A refocused strategy

A current trend for strategic thinking within many Telcos is that they have lost the ongoing battle for advanced services to the likes of Google, Amazon, Apple and Facebook. They see their long-term salvation in refocusing on their core infrastructure strengths, and in this case, that means improving reliability and reducing latency or delay.

They have been disintermediated out of many high-end services that they were banking on such as VoD, and they are fighting tooth and nail to not also be disintermediated out of their core network services.

They are counting on low delay giving them a competitive advantage, so irrespective of the use cases described in this eBook, we expect some network operators to work on reducing streaming delay for its own sake.

Taking a holistic view of the long-term evolution of entertainment, we can see a need for lower latency throughout the entire user experience.



AI's impact on the streaming market

AI is enabling more powerful content navigation paradigms through more accurate personalisation, which generally leads to simpler UIs that now often include voice. The fuel of AI is data analytics, which enables significant improvements in user experience such as voice control.

More central to this eBook is the enhancement we are starting to see in pure video quality through better optimization of streaming parameters. Having a more sophisticated toolset, thanks to AI, operators have greater control over the trade-offs and, where required, can prioritise delay over other aspects such as bandwidth or the number of ABR profiles.

Techniques such as Content- or Context-Aware Encoding (CAE) are progressing in leaps and bounds, thanks in part to automated content analysis where we see AI techniques brought to video processing.

Ultra-Low-Latency killer apps: Gambling, next-gen interactive game shows and multi-cam

In addition to the live sports use case, that calls for reduced delay, we have identified three killer apps with viable business models requiring sub-2s stream delay:

BETTING

Ultra-Low-Latency will become critical for betting use cases, but for now, many betting environments do not yet include in-event betting when using live video streams. Ultra-Low Latency (ULL) streaming is used in the specific use case of video poker, where the game host is particularly alluring, adding a reason for people to continue to place bets. In such a use case, each second gained is turned directly into profits for the gambling company.

> “The Racing Post’s John Matthews told us that **“Today, OTT streams are still decoupled from the core betting infrastructure in most cases, but when we use them, good enough is usually 5 to 6 seconds.”**”

INTERACTIVE LIVE TV

Next-generation interactive live TV programs will require Ultra-Low Latency. It’s not a very well-kept secret that in the early days of SMS voting over a decade ago, TV shows had to guestimate the results as telecom platforms couldn’t cope. Nowadays TV programs such as the Eurovision song contest can accurately count (and monetise) millions of votes via SMS. If we use apps to interact with TV shows, for example, to interactively participate in a live debate with a multi-part discussion, stream delay must be reduced to a maximum of 10s, if not under 2s. If the interaction uses video in both directions (as with a live Skype phone-in), even a second of extra delay becomes an issue.

MULTICAM

Operators have been running Multicam trials for decades. In the ‘90s, when all feeds were broadcast, users could zap between a multitude of different feeds covering a Formula 1 race. These services haven’t been widely adopted as average users tend to play with the system for a few minutes to create their unique feed, before settling down to watch the race on a traditional channel edited by a professional. In the age of OTT streaming, there have been attempts to revive such services by broadcasting the primary stream and then streaming numerous other feeds OTT. Two major European operators have artificially delayed the broadcast stream by 1 minute to ensure all streams could be synchronised.

New video formats such as UHD and VR360

VR360 offers an immersive video experience presented in 360˚ requiring a head-mounted display for premium experiences, but it can also be viewed on a TV, using the remote control to navigate.

VR360 user experience must be responsive, and experts agree that latency between head movement and the UI updating must be below 20 milliseconds to avoid motion sickness.

The user with a headset never sees or uses 80% of the pixels sent across the network with a brute force approach that sends the whole 360˚ of video to the end device. The bandwidth vs. resolution trade-off cannot be found to deliver a good-enough UX at a reasonable cost, especially when we consider higher resolutions, even at today's sub-par premium of 1K x 1K pixels per eye. Indeed, we know that a single human eye is probably capable of discerning at least 8K x 8K, which would mean that a premium VR360 experience could one day require above 1 Gbps.

That is one reason why solving VR360 challenges can need at least the bandwidth of 4K or 8K resolution video. Viewport-adaptive streaming schemes are a way to enable VR360 without colossal bandwidth requirements.

Tiled VR streaming is a method where the overall latency perceived by the end-user is kept below 20 ms with the help of a CDN.

As elsewhere in this eBook, latency is the most critical factor here.

6. Takeaways

As we've seen throughout this first eBook, many factors need to be taken into account when considering the importance of delay and latency within live OTT streaming.

So far, we have seen some cases where it is already critical to reduce stream delay, and we believe these are likely to grow as:

- Live events such as sports drive awareness of stream delay,
- Costs decline,
- Live streams further integrate with social media,
- Truly interactive TV increases,
- "OTT streaming" becomes the de-facto standard for watching live content,
- New latency-dependent use cases such as tiled live VR360 become a reality.

Part 2 in the series will address the technical challenges and opportunities such as the case for edge-processing.

Ultra-Low Latency solutions are already in the market, and it seems inevitable that in the very near future the norm will be for live streaming – at least of premium events – to deliver a video experience that is at least as good as broadcast. As new interactive features are added, Ultra-Low Latency will play a significant role in making streaming the preferred video experience.

Check out "What the 2018 World Cup has taught us about video streaming", Part 2, to learn about Anevia's solution for Low and Ultra-Low Latency live OTT:

- Includes Anevia's Low Latency live encoder for a full Low Latency video processing chain,
- Anevia's Low Latency packaging and CDN solutions,
- Takes advantage of CMAF-LLC packaging and HTTP chunked delivery,
- Can reduce latencies to as low as 1 second, from camera capture to tablet display.

7. About the Authors

Benjamin Schwarz

Ben is recognized as an international thought leader in the converging media space. After ten years in IT, he was CTO of an Internet start-up. Ben spent the next ten years at Orange on strategy, innovation and launching media services. Since 2008 he has run his own company helping international clients with innovation, strategy, business modelling, technology and communications. Ben has written many White Papers that intertwine a business and technology narrative. He is an active member of the Ultra HD Forum and is now involved in blockchain-related projects. He lives in Paris, and his social media name is nebul2.

Jérôme Blanc

Jerome Blanc is EVP Compression Products at Anevia, where he oversees the development of the Genova video encoders and live OTT packagers. He was formerly COO at Keepixo, an Allegro DVT spin-off recognized for its video technology expertise, which was acquired by Anevia in 2018. With a strong technical background and in-depth knowledge of video compression, he co-developed Genova's first H.264 video transcoder engine. Jerome holds an Engineering Degree and a PhD in computer vision and image processing.

ABOUT ANEVIA

Anevia is a leading OTT and IPTV software provider of innovative multiscreen solutions for the delivery of live TV, streaming video, time-shifted TV and video on demand services. The company offers a comprehensive portfolio of video compression, multiscreen IPTV head-ends, Cloud DVR and CDN solutions to enable viewers to enjoy a next-generation TV experience – anywhere, anytime and on any screen - including 4K UHD content. The solutions are widely adopted by globally-renowned telecom and pay-TV operators, TV broadcasters and video service providers in hospitality, healthcare and corporate businesses.

Founded in 2003, Anevia has a track record of being first to market with advanced video technologies. The company is a member and active contributor to several TV, media and hospitality industry associations.

Headquartered in France, with regional offices in the USA, Dubai and Singapore, Anevia is listed on the Paris Euronext Growth market.

For more information, please visit

www.anevia.com