anevia

YOUR NEXT GENERATION TV, NOW

Saving CDN costs: quality-based rate control

New rate control techniques explained for OTT TV

ANEVIA S.A. | 79 rue Benoît Malon, 94250 Gentilly, France | Tél : +33 (0)1 41 98 32 40 | Fax : +33 (0)1 41 98 32 48 | www.anevia.com
S.A. au capital de 192 979,60 € | RCS 448 819 680 CRETEIL | TVA FR15 448 819 680 | SIRET 448 819 680 00043

INTRODUCTION

CDN costs are an important parameter of any OTT TV deployment. This is a concern for everyone in that space:

- Large content providers or service providers aim at the largest possible audience, and CDN **costs grow as audience grows**.
- Smaller actors, for instance in the PEG market (Education, Government, Worship, Ethnic, or other Community channels), may aim at a smaller audience, but then CDN/delivery costs will still represent a **large proportion of their total infrastructure cost**.

Some of these users may not use and pay for a commercial CDN like Akamai, CloudFront and others, but still use a content delivery infrastructure of some sort, even if it's based on their own physical network (for ISPs), or leased lines, or on beefy origin servers just lying on the open Internet. Delivery is always expensive, directly or indirectly.

In short, it shall be no surprise that the larger the number of bits to transport, the higher the cost, and any OTT TV actor should look at hedging against this.

New video encoding technologies, relying on **quality-based Rate Control algorithms**, now allow to reduce those costs, while preserving video quality. We call them **Capped** methods. This paper exposes how.

An executive summary is provided in Figure 1.

EXECUTIVE SUMMARY

What issue are we addressing?	What is the solution?
▪ OTT (HLS, DASH/CMAF) protocols are based on multi-bitrate encoding profiles. ▪ Each profile is defined by its "average/constant bitrate". ▪ It puts constraints on the encoder to produce a given number of bits. ▪ In some scenarios, the encoder will end up **wasting unnecessary bits to match a predefined bitrate**.	▪ The "average/constant bitrate" constraint is loosely defined and can be understood in many ways. ▪ In the encoder, **rate control can be customized** to only enforce the bitrate constraints that are **necessary to OTT TV**. ▪ It will reduce the number of bits if a **target quality** is reached. Savings will depend on scene contents. ▪ In some ways, that is similar to what "Capped VBR" is to "VBR" in the broadcast (non-OTT) world. ▪ It requires expertise on the video encoder and a good understanding of rate control techniques. A good solution **does not degrade video quality**.
Unnecessary delivery/infrastructure costs	**Delivery/infrastructure (CDN) costs savings**

Figure 1: Executive summary.

INTRODUCTION: CDN COSTS: WHO PAYS?

The live OTT TV streams need to reach the customer playback device in some way. That device is connected to a delivery network.

- In a few cases, it's a proprietary and managed network. That's **the case for ISPs offering a live OTT TV service** to their subscribers, for instance. The subscribers have a home gateway that is directly connected to their ISP's network, and they watch live video with OTT TV protocols at home. The ISP's private network is the content delivery network, and obviously, the more bits to transport, the larger the infrastructure and **the higher the cost to the ISP**, even though the ISP doesn't use the services of a commercial CDN. Note that in that case, this is not strictly OTT, but rather OTT protocols over an end-to-end managed network.

- But most of the time, it's "the Internet", an unmanaged network, which means it is made of sections of networks actually managed by different companies. That's what OTT protocols have been designed for: to enable live TV although no strict QoS can be guaranteed end-to-end on the network, and to survive packet loss and network bandwidth variations (whether fixed Internet, or typically mobile connections: wifi and 3G/4G). Then two extreme cases can happen:

    - **Larger providers** need a maximal QoS and a rock-solid, scalable delivery. In that case they use the services of a **commercial CDN**. That CDN company has established its own delivery infrastructure, and interconnections with the other Internet operators, optimally down to the end-customer's last mile. It usually hosts servers, caches, and logic to ensure the most efficient content delivery.
      It guarantees it can scale to whatever bandwidth to be delivered to the end-customers. Commercial CDNs are often **paid per delivered bit** (outbound bandwidth consumed), which is schematically the number of bits that the provider uploaded, times **the number of viewers**. That makes it especially interesting for those large OTT providers to reduce their encoders' output bandwidth as much as can be.

    - **Smaller providers** will be more cautious on costs, and some may just leave their origin servers lying on the open Internet, acting as public web servers. It will come with a lower guarantee of service than commercial CDNs, and to handle outbound traffic correctly and not incur too much traffic loss, those providers will need origin servers powerful enough (storage, caching, connectivity). So even if cheaper than with a commercial CDN, those providers **will still suffer costs related to their amount of traffic**.

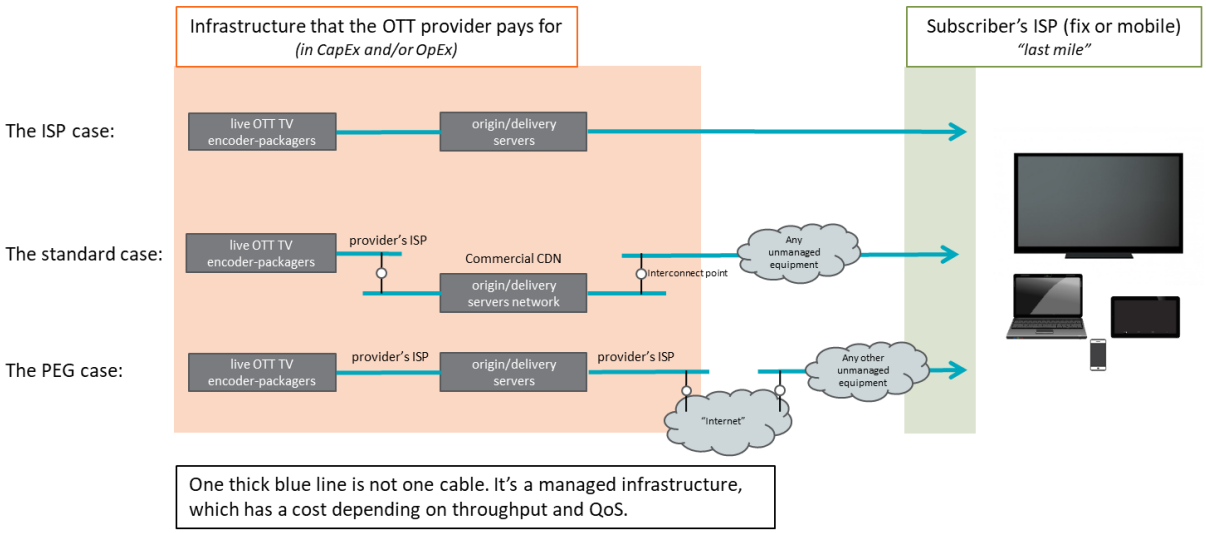

Figure 2: Some delivery schemes. Someone always has to pay.

In conclusion, every OTT TV provider uses a delivery infrastructure and pays for it, at least partially, in a form or another, even providers not using a commercial CDN, but rather their own infrastructure. There is always a (non-capitalized) content delivery network, even when no commercial (capitalized) Content Delivery Network company is involved. **It always has a cost that affects the provider.**

SOME BACKGROUND: WHAT IS RATE CONTROL AND CBR, ANYWAY?

Rate Control is not a trivial matter, and there is a great deal of confusion on what CBR, VBR, Capped VBR… really mean.

CBR

CBR obviously stands for Constant Bit Rate. No one is right or wrong using those words with different meanings, but **what DTH/sat/cable/IPTV operators call a live CBR video stream is defined this way**:

> The video stream can be sent over a transmission link at a fixed speed, and the viewer's set-top box will be able to display it smoothly.
> i.e. the box's internal receiving buffer will never overflow (if it received too many bits), neither underflow (if it's time to display a new video frame, but no data has been received yet).

This is also called "CBR with HRD compliance", or "VBV compliance". Some call it "strict CBR".

It's not as obvious as it seems. It is explained in other words, more graphically, in Figure 3.

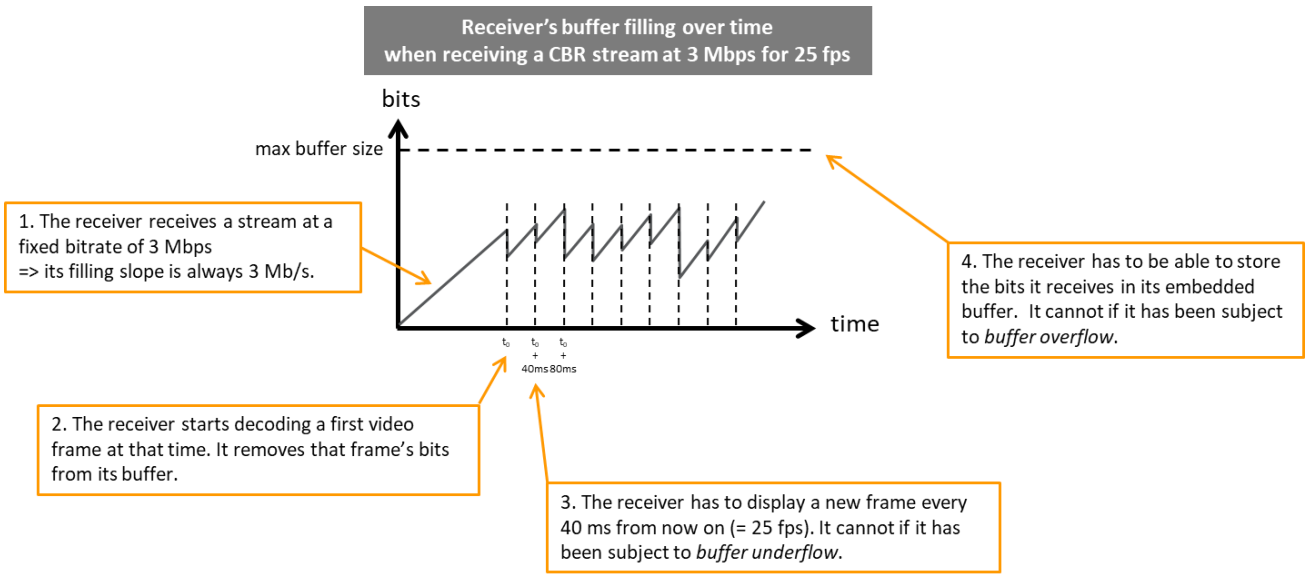

Figure 3: "CBR" as broadcasters mean it.

As can be guessed, a CBR-compliant, or HRD-compliant, video clip can perfectly be created and checked offline. The encoder has to be aware of the speed of the transmission link that clip will be streamed on, decoding rate (fps), and decoder's maximal buffer size. It also has to transmit t_0 to the receiver.

CBR puts such constraints on the video encoder that it cannot freely allocate any number of bits to any image: as it has to enforce “HRD compliance”, the encoder’s CBR rate control algorithm may find itself trapped in a situation where its live video input becomes suddenly more complex to encode, but the encoder cannot spend the necessary number of bits to ensure a good video quality without exceeding the buffer’s limits. This can be alleviated by some look-ahead techniques, at the expense of increased video latency.
In summary: **in “strict” CBR, it’s difficult to ensure a consistent video quality.**

VBR and Capped VBR

VBR is the same, except the filling rate (the slope) can be variable. This is especially interesting if the transmission link itself is of variable speed; or, more frequently, to multiplex several video streams within a fixed CBR transmission link. As it has less constraints than CBR, the video quality can be improved.
Capped VBR is a variation of VBR where the bitrate, and/or the video quality, can be configured with a cap: a limit not to be exceeded.
In normal situations, **VBR enables a better video quality than CBR. Capped VBR uses less bits than unconstrained VBR, and a good Capped VBR also does not degrade VBR’s video quality**, if bitrate allows.

OTT VS DTH: THE CONFUSION OVER RATE CONTROL

What we’ve just seen is applicable to DTH, i.e. traditional broadcast or IPTV.
Our domain, OTT TV, is the encounter of people coming from the traditional broadcast industry (DTH), and people coming from the Internet industry. The Internet industry has long been able to encode video clips with some sort of rate control for a predefined file size and/or quality target.
The software encoders used for that task often expose CBR or VBR settings, but not necessarily with the same meanings as in the DTH world.

OTT and DTH players: a fundamental difference

DTH and OTT decoders are fundamentally different:

- **DTH**: relies on **passive players**: satellite, cable, terrestrial, or IPTV boxes: the decoder receives data passively and cannot do anything else than handle them. That requires that the DTH encoder enforces HRD compliance, for instance **CBR rate control** as exposed above.

- **OTT**: relies on **active players**: tablets, smartphones, PCs (web browsers), HbbTV decoders: the decoder downloads the OTT data (segments, chunks, fragments… according to different terminologies) that it wants, plays them back when it wants, and can even decide to rather download alternative data (e.g. segments at a lower bitrate) if it must. There is no HRD compliance needed, and in the OTT world, "CBR" does not have a clear meaning… **nor necessity**.

What rate control constraints do we really have with OTT?

In the early days of HLS for live OTT TV, Apple recommended that the HLS video be encoded in "*Constant Bit Rate, with a maximum variation of 10%*"[1]. For instance, for a profile at 3 Mbps, what they meant at that time was: "each TS segment of 10 seconds must contain exactly 30 Mb +/- 10%, i.e. between 27 Mb and 33 Mb". In their opinion, that was to facilitate delivery and more easily ensure that playback would not break up. The experiments related in [2] confirm this.

For the exact same reason, some commercial CDNs also recommended that the HLS files are more or less the same size, within the same encoding profile. That was to increase the efficiency of the CDN's internal algorithms for traffic control and caching, and as a result, improve the delivery predictability and QoS consistency.

These constraints are less and less necessary. CDNs now accept to ingest and deliver files with more variability. Even Apple, as a typical example, now recommends this for live OTT TV[3].

1. "*For live/linear content, the average segment bit rate over a long (~1 hour) period of time MUST be less than 110% of the AVERAGE-BANDWIDTH attribute.*"

[1] This is still what they recommend for VOD, but not for Live.
[2] http://streaminglearningcenter.com/articles/bitrate-control-and-qoe-cbr-is-better.html
[3] https://developer.apple.com/library/content/documentation/General/Reference/HLSAuthoringSpec/Requirements.html

2. "*For live/linear content, the measured peak bit rate MUST be less than 125% of the BANDWIDTH attribute.*" The definition of "peak bitrate" is[4]: "*the largest bit rate of any contiguous set of segments whose total duration is between 0.5 and 1.5 times the target duration. The bit rate of a set is calculated by dividing the sum of the segment sizes by the sum of the segment durations.*"

So, what does that really mean?
The first constraint says that, for a profile declared at 3 Mbps, it is now only necessary that the number of bits over a 1-hour sliding window be less than (3 Mbps x 3600 s) + 10% = 11.88 Gb. The second constraint is more difficult to formalize, but as an example, in some circumstances, the number of bits over a 5-second sliding window must also be less than (3 Mbps x 5 s) + 25% = 18.75 Mb.
The first constraint is called by some "VBR 110%". It is a **bandwidth constraint**. The second constraint is a **variability constraint**.

Why is it interesting for OTT TV? A better video quality.

As can be seen from what we've just exposed, the OTT TV constraints, even when they're called "CBR" or "VBR" of some sort, have nothing to do with CBR and VBR in the broadcast world.
Apple's "VBR 110%", as an example, is much more relaxed than traditional CBR or VBR, and thus **enables a better video quality**. To avoid confusion between those different meanings of VBR, some call it "**ABR**" as a generic term, as "the rate control algorithm that is suitable to Adaptive Bit Rate delivery". A fundamental difference is that **ABR rate control for OTT TV has to be chunk-aware**. It is more segment-based than stream-based.
A second important point is that, from these constraints, we see **the encoder is allowed to create smaller segments** if this is sufficient, for instance when a given video quality is reached.

What that means is that it is possible to create new ABR rate control algorithms that save bits, and reduce the OTT segment size if the target quality is reached. In a way, the user configures a "max quality" to not exceed, or a "quality cap". Those algorithms better adapt to the scene complexity, and as an analogy with what Capped VBR is to VBR, we call these rate control algorithms "**Capped ABR**".
The rate control landscape is summarized in Figure 4: the less constrained the encoder, the better the video quality.

[4] https://tools.ietf.org/html/draft-pantos-http-live-streaming-23

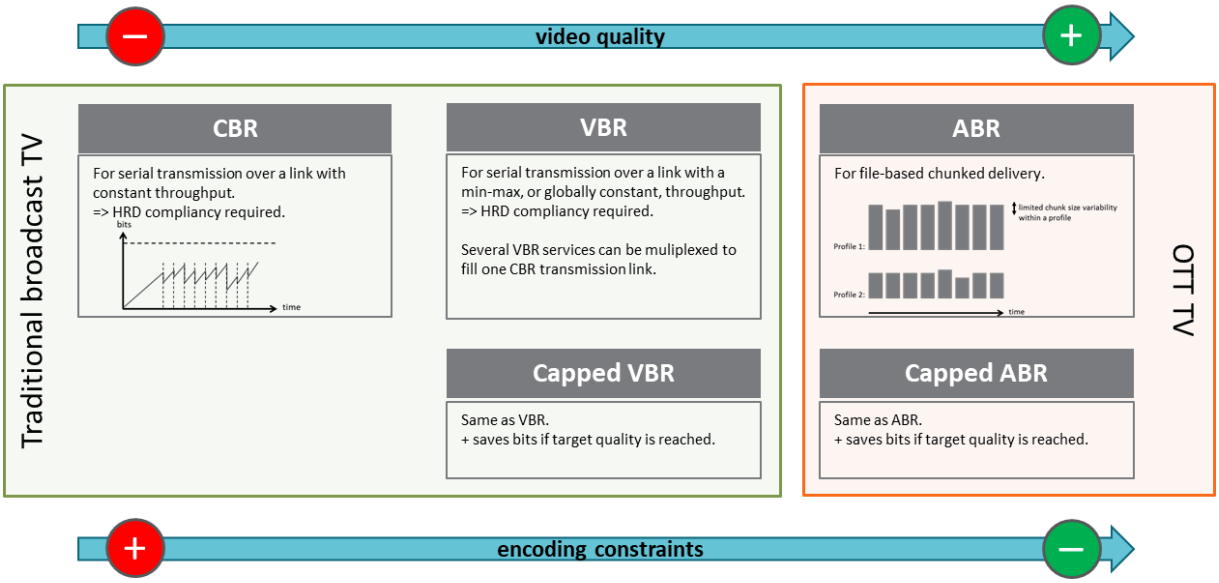

Figure 4: OTT TV relieves constraints on encoding, enabling a better video quality than DTH (broadcast) TV.

"Capped ABR"

For the record,
Table 1 details the case of Capped ABR. In some ways, Capped ABR is to ABR similar to what Capped VBR is to VBR: for instance, if the target video quality is reached, no more bits are delivered. It keeps video quality constant, plus saves bits and CDN/delivery costs when the content is transiently easier to encode. We can call it a "quality-controlled", or "content-aware" method[5].

ABR rate control	Capped ABR rate control

[5] With the same logic, VOD encoding for OTT TV can be optimized with content-aware "per-title" or even "per-scene" optimizations. Netflix uses it. In their case however, they change the bitrate ladder (= the rate control settings), rather than the rate control internal algorithms themselves.

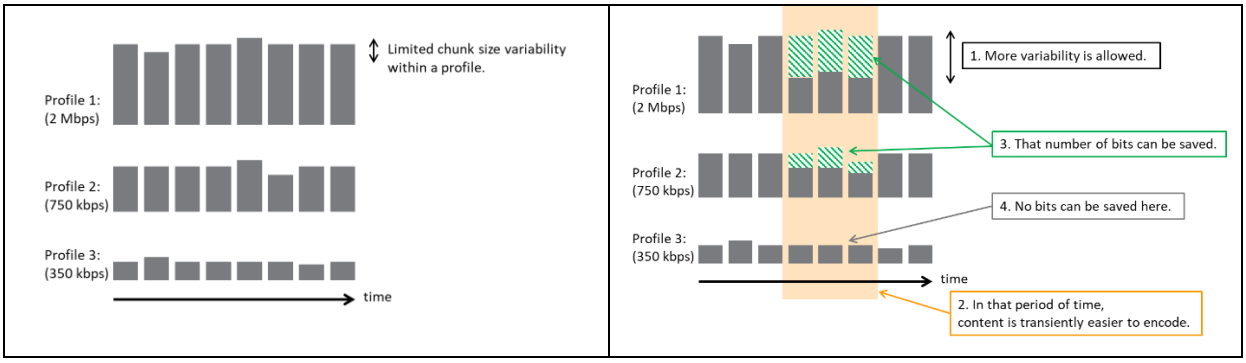

Table 1: ABR vs. Capped ABR.

How to implement Capped ABR

It is clear that Capped ABR is a function that takes place at the video compression stage, and has to be implemented at the core of a video encoder that is specially designed for OTT TV rather than broadcast. Different OTT encoder vendors compete in that space, in at least 3 aspects:

- The **rate control algorithm itself**: as they've always been doing, all vendors spend a lot of time customizing their rate control algorithms, and are very protective about them. Usually based on many years of experience of their codec team, they are an important asset of such vendors.

- The way they define "**video quality**": vendors rely on a **variety of measures**, some being proprietary. Many vendors claim they have designed the closest measure to what the Human Visual System perceives. But more than the measure itself, it's what the encoder and the rest of the rate control algorithm do with it that matters: even a measure like SSIM can be beneficial provided its results are used cleverly to **drive the encoder** to reduce the bitrate while preserving video quality. Some might perform just re-quantization, or mode selection with RDO, or even a complete second encoding pass.

- The workflow: some vendors will need **several encoding passes** to decide where they can save bits on each segment and not degrade video quality too much. Multi-pass in that context is not good or wrong by itself, it depends on with what data it's done (full input picture, or rather simplified input data), and how much it adds to encoding latency.

As seen, not much can be disclosed on the details of how Capped ABR methods are actually implemented. Each OTT vendor will have its own secret measures and algorithms to ensure they can cut the bitrate, on a constant-quality basis.

Efficiency

The system's efficiency cannot be predicted for sure: the encoder will ensure it matches, for instance, Apple's "VBR 110%" constraints, but cannot guarantee how more bits it can save to it. Indeed, content-aware methods… depend on content: if the incoming video stream is always difficult to encode, like a sports channel, and bitrates are already tight, then no savings are to be expected. If it is transiently easier to encode, like mainstream or news channels often are, then savings will occur. **That's exactly what Capped ABR methods enable**.

Also, whether the video quality suffers will depend on each vendor's encoding and rate control algorithms. Some offer a way to **configure the trade-off**, between an "aggressive" setting that allows the encoder to degrade the video quality a bit, and a "conservative" setting that keeps it constant – but reduces bit savings. Another way is to allow the operator to configure a quality cap: an absolute quality score to not exceed, which can be expressed in PSNR or SSIM points, for instance.

In summary, **efficiency can only be measured in real operation** on the operator's own live setup, with its own content and its own tolerance thresholds, rather than predicted in a sure way.

CONCLUSIONS: WHAT'S NEW?

A video encoder is made of many functional parts, which can all be improved. Here we focused on the Rate Control algorithms.

In OTT TV, decoders work differently than in traditional broadcast TV or IPTV, and using traditional rate control algorithms is not justified anymore. Instead, OTT encoder vendors can design custom rate control algorithms, more based on reaching a defined quality (a quality cap) than on enforcing legacy bitrate constraints inherited from the DTH world.

They call those methods with different trade names, and we called them here **Capped ABR** methods.

Such methods save CDN/delivery costs, which would otherwise grow unnecessarily: with some content types, the encoder would waste unnecessary bits to match a predefined profile bitrate, whereas it shall rather match a predefined quality for that profile.

Using rate control algorithms specifically adapted to live OTT TV helps providers reduce their delivery costs and gives them a competitive advantage in this regard.

At Anevia, we strive to provide state-of-the-art encoders and packagers for OTT TV, and we support, among many other features:

- High quality, high density, and low latency encoding modes for H.264 and HEVC;

- User-configurable Capped ABR rate control to save CDN costs;
- All mainstream / or niche / or tricky OTT packaging features as required by professionals: subtitle conversion, many modes of redundancy, DRM, ad signaling, multilingual, parental rating…

Anevia's OTT encoders and packagers provide a solution for any content or service provider wanting to switch to OTT TV with confidence.

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 have been 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.