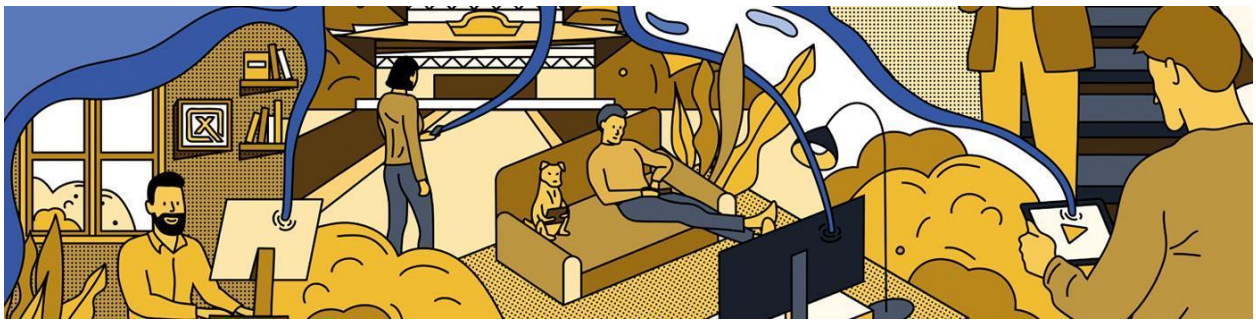


How audience aware encoding?

Marc Baillavoine, September, 2019

More Connected Viewers than BroadcastViewers for the 2024 Olympics?

Video streaming will become the de-facto standard for watching live events in the upcoming years. The target date varies, but there is a consensus that the 2024 summer Olympics will have more “connected” users than broadcast viewers.

In that perspective, the network (CDN) is key. There is likely to be more than 1 billion concurrent viewers for the 100m final, but for now the streaming ecosystem is far from sustaining such a traffic. This bandwidth problem also comes with a cost issue: today, it is more expensive to stream an event than to broadcast it. In this blog, we are going to see why “breaking silos” between the distribution network and the network headend is key to improve the user experience whilst lowering the overall cost for the content owner, and how Audience Aware Encoding(™) can be used to dynamically optimize the CDN and the headend together.

Streaming Architecture and Procurement Silos

Most of the streaming systems have been designed with the broadcast principles in mind: the video headend on one side, and the distribution network (satellite, terrestrial, …) on the other. This made perfect sense in the past because the distribution link was a fixed and static object. Being a “push” model (the content is pushed towards the users), the price of that link was fixed (not dependant on the number of users) and its reliability was not a function of the number of viewers

either. Procurement was separated for these entities, leading to a siloed cost optimization: headend on one side, and satellite capacity on the other.

Most of the content providers haven't changed their approach yet for streaming and will evaluate separately the headend (where the content is prepared) cost and the CDN (where the content is delivered) cost. Moreover, they will also evaluate separately the cost of the hardware and the cost of the software license. But in a world where:

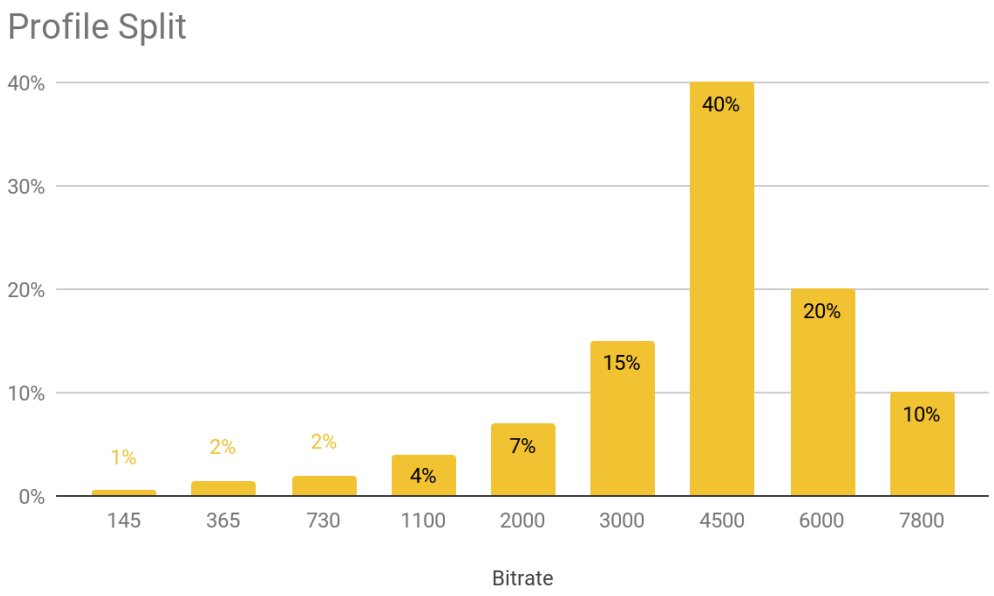
- CPU resources can be provisioned and paid "à la carte",
- CPU resources are collocated with network resources,
- The network condition will vary over time,

It may be beneficial to envision a global TCO rather than using silos inherited from the broadcast world.

Total Streaming Cost

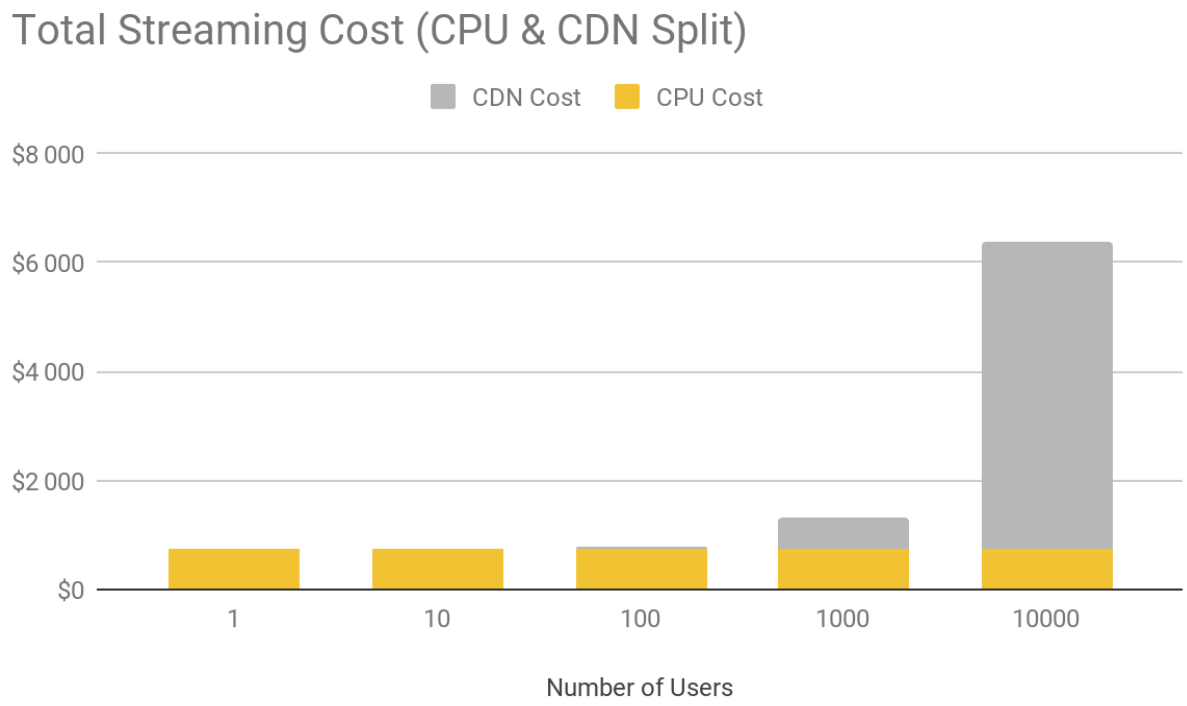
Although several vendor-specific refinements exist, the CDN pricing model is a function of the number of users. The actual price is extremely variable. Some charge only for the egress, some charge for filling the cache or flushing the cache, some will price differently hit and miss … In addition, huge discounts are common for high volumes, making the price list a very theoretical object. For the sake of this post, let's make the following assumptions (leading to ~$0.6 per user per month)

- A pure "egress based" model, with a street price of $0.005 per GB
- A rendition ladder coming from the HLS Authoring Specification (9 profiles from 416x234@145kbps to 1920x1080@7800kbps)
- A split of profiles where most of the users can download ~4.5 mbps
- An average viewing time of 2 hours per day, per user.



Unlike the CDN, the CPU cost does not depend on the number of viewers: it is a function of the number of channels and renditions to process. For determining the price of the CPU, we used a virtual machine in Google Cloud made up of 32 threads/virtual CPUs, and we used ffmpeg with libx264 to build a vendor-neutral comparison. As a matter of fact, this $746 (monthly) CPU is the perfect fit to sustain real time encoding for the defined profiles. There are two immediate conclusions when adding the CPU and the CDN cost:

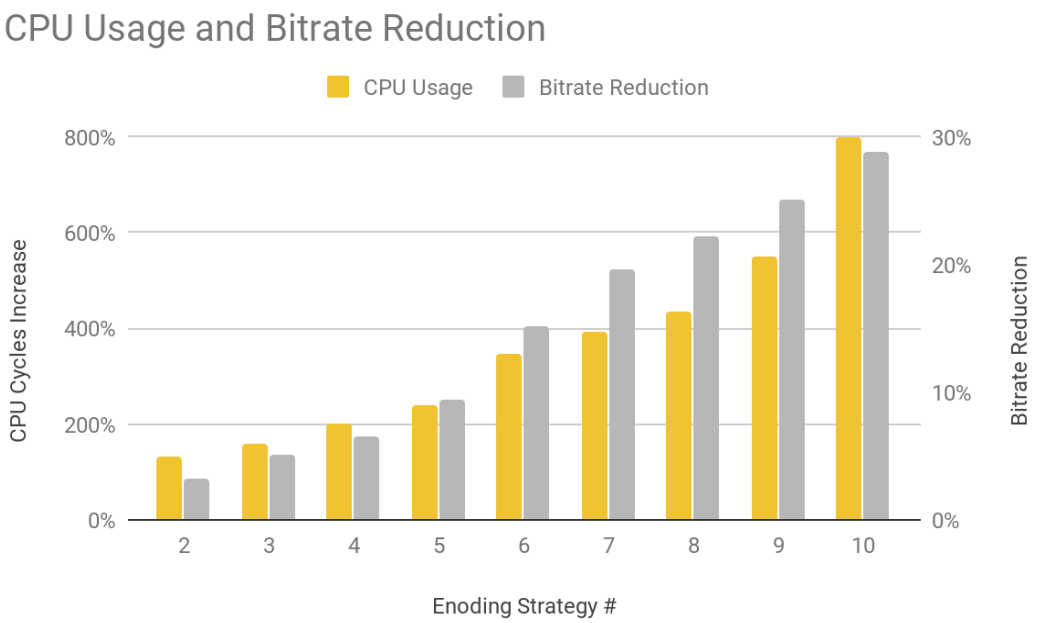
- Below 100 users, the CDN price is negligible: reducing the price means reducing the CPU cost.
- Above 10000 users, the CPU price becomes negligible: reducing the price means reducing the bitrate!



Cloud-Native Processing Strategies

The efficiency of a (decently engineered!) video encoder increases with the number of allocated CPU cycles. In other words, for a given video quality, you can use less bits to encode the content if you have more available CPU cycles, because the encoder will make “smarter” encoding decisions.

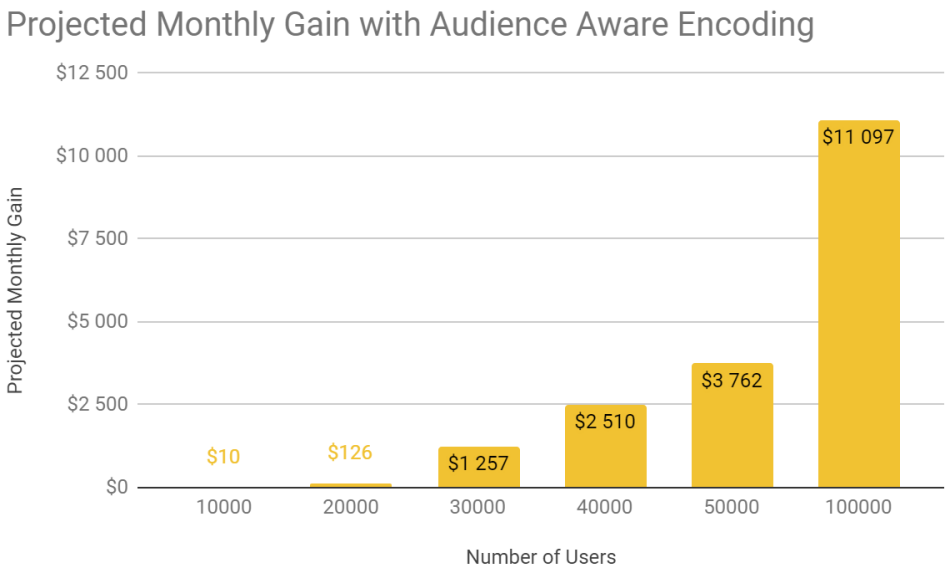
When using a cloud native live solution, there is virtually no limit to the number of CPU cycles that you can allocate to your encoder. This is a paradigm shift, as vendors have always tried to optimize a video encoder efficiency against a given hardware and developers have always been unconsciously limited by the limits of that given hardware. Unlike others, the solution Quortex developed can scale in seconds to use up to 10 times more CPU cycles. We have a full ladder of encoding strategies that all end up with the exact same visual quality, but that have different tradeoffs between bitrate reduction and CPU cost. For instance, on the above graph, the encoding strategy “6” will save approximately 14% of bits against the reference strategy, but will cost approximately 4 times more in CPU cycles.



Audience Aware Encoding (™)

Saving bits will help save on CDN costs, while lowering CPU demand will of course save on CPU cost. If you feed the Quortex solution with the audience information (gathered from the CDN API, for instance), it will permanently take this data into account to adapt its processing strategy and lower the overall cost, without any compromise on the Quality of Experience. We make use of the latest technologies in cloud functionalities to scale in a few seconds and make the best of the cloud resources.

In more details, a callback (for instance, a Google Function or a AWS Lambda function) will be implemented to gather CDN metrics and push it to the Quortex solution. Based on this information, the processing will decide which strategy to use until the next set of metrics and will seamlessly scale up or down. Less viewers? It will probably scale down to limit the CPU cost. More viewers ? It will probably end up scaling up to reduce the CDN bitrate and the associated cost. This process is done automatically, you don’t have to provision any machines or take care of the infrastructure. This leads to significant gains (as depicted on the below graph for the example we used throughout this blog).



Going futher

We have seen how Audience Aware Encoding(™) can be used to globally optimize the TCO of your streaming service. The examples we used in this post did not even take into account the unique ability that we have to further reduce the CPU cost (please read our blog post on that topic). Indeed, using preemptible VMs/Spot Instances for live encoding will further significantly increase the benefits of Audience Aware Encoding(™).

Furthermore, this system can also be used to relieve the CDN in case of an unexpected peak of audience: by reducing the bitrate by 30% is an efficient tool to make sure your content reaches all your subscribers.