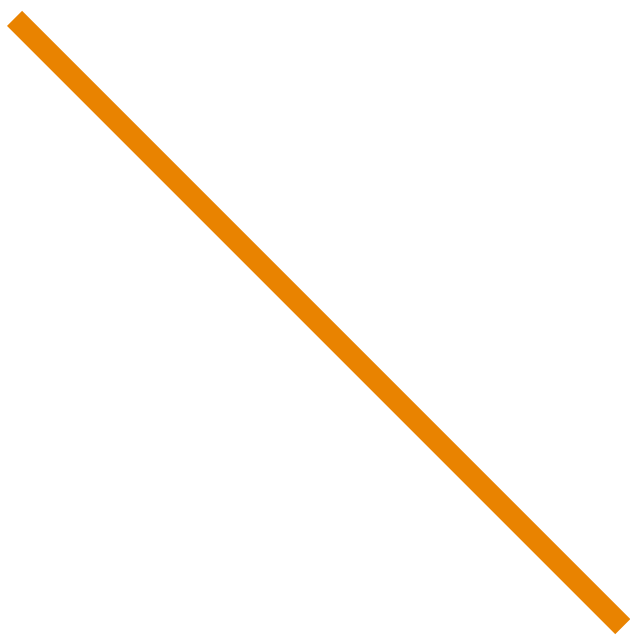


# Hybrid CDN



# Essential Guide

**EG**

# Introduction

Broadcasting video and audio has rapidly developed from the send-and-forget type transmission to the full duplex OTT and VOD models in recent years. The inherent bi-directional capabilities of IP networks have provided viewers with a whole load of new interactive viewing possibilities. Even though the reliability of the internet can be occasionally challenging, discerning viewers do not make any allowances for variances in internet delivery and demand the same quality of service and experience as they get from traditional broadcasting.

The internet was never designed to carry the amount of data a modern streaming service offers. Video and audio are relentless in their thirst for data capacity, especially with mobile device delivery where multiple streams are offered within a package to take into consideration the differing environmental delivery conditions. Capacity requirements soon ramp up as more services are offered by broadcasters, and that's before we start looking at the interactive nature of OTT.

Hyper Text Transfer Protocol is the fundamental streaming mechanism used in OTT delivery as it is the common format for portable device players and internet browsers and is ubiquitous for internet delivery for all types of data exchange. Consequently, the whole internet relies on facilitating the HTTP protocol to make it reliably operate.

Delivering streamed video and audio to viewers is more than just taking the video and audio data and sending it across a network. Multicasting is a technology that is constantly being researched and developed but is yet to find its way into the public internet. It is possible to provide multicasting in private networks and is regularly used within the broadcast infrastructure, however, technical limitations restrict its use and instead we rely on a one-to-many mapping of program stream to user.

Each viewers device requires a direct logical connection to the server providing the program stream. As the numbers of viewers increases then so does the demand on these servers. Having a centralized system is inefficient and often impractical. Instead, a distributed server mechanism provides the optimum solution. This results in "edge-servers" being deployed as viewer demand increases.

The edge-server forms part of a complex interaction between the broadcaster, internet and viewer. More infrastructure is required such as the transcoders and playout servers, especially as we start to look at the differences between VOD and OTT, and it is this interaction that lays the ground for Content Delivery Network (CDN).

The internet's network, as provided by a collection of ISPs and intermediaries, delivers the backbone of the network. They do provide infrastructure to help distribute IP datagrams, especially for HTTP systems, but they tend to leave the tuning for streaming to others. CDN is one method of tuning the internet to deliver streamed programming and generally relies on adding storage, packaging and edge servers to the network to facilitate better VOD and live OTT delivery to viewers.

Both private and public CDN's are available with multiple advantages and disadvantages for both. Public CDN's are a generalized solution but private CDN's deliver specific services to broadcasters including increased granularity of monitoring and higher tolerances for data delivery.

The whole internet delivery mechanisms for VOD and OTT has ballooned enormously over the past few years and it can often be difficult to keep up with the technological advances and why we use them. Understanding the "why" is often the starting point to understanding how complex systems work, who uses them and when.



Tony Orme.

This Essential Guide introduces the concepts of CDN's and explains why we need them. It then goes on to discuss both public and private CDN, and how a hybrid model approach adds value for both broadcasters and viewers.

Tony Orme  
Editor, The Broadcast Bridge

# Hybrid CDN



By Tony Orme, Editor at The Broadcast Bridge

Content Delivery Networks (CDN) are gaining popularity as broadcasters move to the OTT method of distribution. But what are CDNs? Who operates them? And how does the hybrid model benefit us? In this Essential Guide, we uncover the challenge hybrid CDNs solve and the practical applications of making them work.

Three fundamental concepts change as broadcasters adopt the OTT method of distribution; we no longer “own” and have complete control of the distribution medium, the network is a one-to-many mesh configuration, and the data path is bi-directional.

Broadcasters transmitting television programs can be sure that when a signal leaves their transmitter, it will reach the viewer. Unless somebody has erected a skyscraper between the viewer and the transmitter, the television pictures and sound will be reliably received.

But to make OTT systems work reliably and efficiently, we must be much more aware of the deeper underlying network capabilities of all the systems between the broadcaster and the viewer.

**TCP and Latency**

By design, IP is a non-guaranteed delivery mechanism. That is, when we send an IP datagram from a server into the network, we can only say, with some certainty of probability, that the IP datagram will be received by the viewers device. To provide a level of guarantee and be sure the viewer sees the broadcast, we must add TCP (Transmission Control Protocol). However, TCP adds latency, but this is an inevitable consequence of using flow control methods.

OTT has some further challenges as broadcasters no longer send just one format for each service. Three viewing formats dominate the OTT market; Android, Apple and Microsoft. These are designed to optimize the viewing experience for mobile devices where the WiFi network conditions may change rapidly. This isn't just due to location, but congestion can be caused by more mobile users entering a location reducing the overall availability of data.

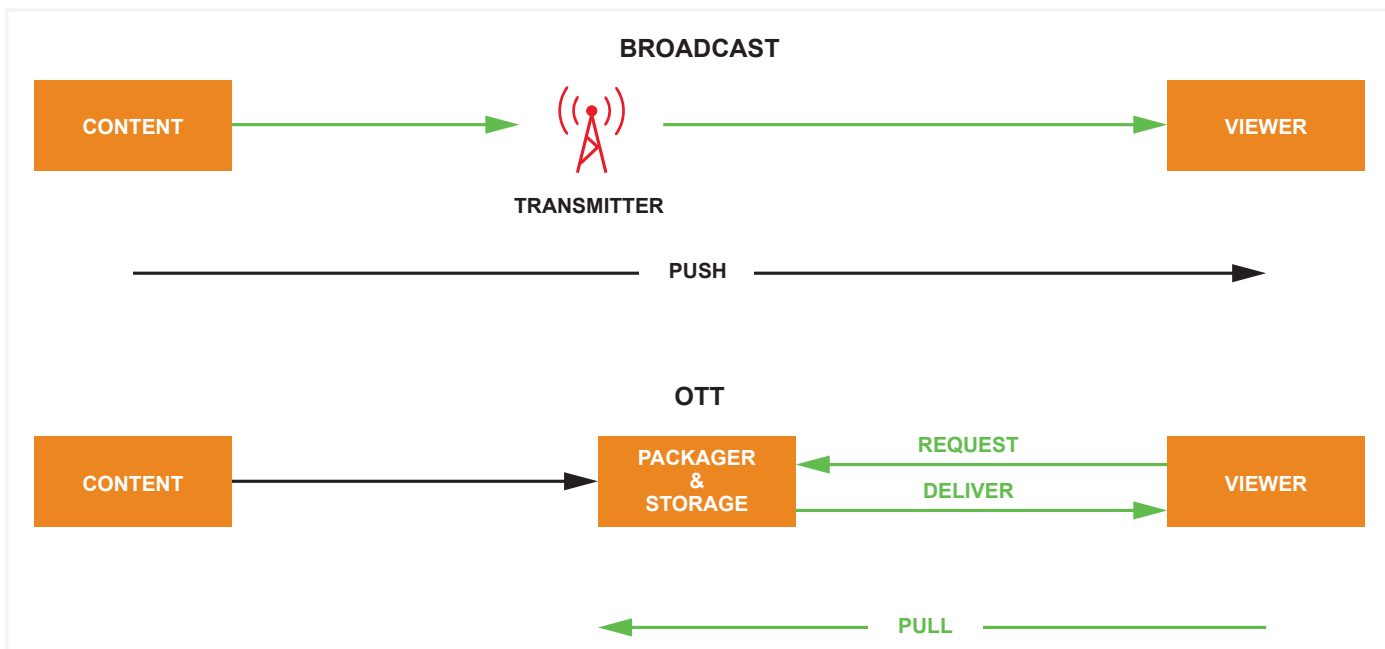
ABR (Adaptive Bit Rate) distribution overcomes congestion to a larger extent. To achieve this, a service is encoded with multiple streams with varying data rates, and for an SD or HD service this may consist of 4-6 streams. As well as providing differing data rates to meet the changing conditions of the network, this bouquet of streams delivers different screen sizes and even frame rates.

**Increasing Streams**

Vendor specific formats further increase the number of streams as a complete bouquet is needed for each of the three main mobile device viewing vendors. Manifest files are also needed and must be reliably streamed to tell the viewing device which type of streams are available and where they can be found.

This all soon mounts up and a service that started as just one stream can soon increase to 18 streams with the associated manifest and description files. From a broadcasting perspective, this can be incredibly daunting, the challenge of providing a one-to-many distribution system over the internet is difficult enough, but when we then increase the number of streams to 18 per service, then life becomes incredibly challenging.

The broadcaster also has to divide the streams into smaller data packets to improve distribution throughout the network. Referred to as chunking, or segment sizing, the stream is separated into smaller packets as the design works on the IP-retry principle. DASH requires three, four second segments (adding 12 second latency) and HLS requires three, ten second segments (adding 30 second latency) to achieve lock and synchronization by the receiver. This results in a trade as the CPU prefers to process larger packets, but to achieve ultra low latency, the segment size has to be dramatically reduced to allow video players to lock to the stream.



OTT delivery fundamentally differs from traditional RF broadcast as the viewers device requests video and audio in the form of data and the broadcast server responds by sending the requested information. Segmentation and packaging of the stream help achieve this as the mobile device requests the next in sequence packet segment. This keeps the viewers device memory buffer full to achieve smooth video playback and distortion free audio. The Packager and Storage processes are split between the Origin (where the ingest, recording, storage, packaging and encrypting takes place) and the Cache servers (either Intermediate or Edge Cache) so that VOD can be cached and the content can be more efficiently distributed as multiple files and streamed from the edge as opposed to streaming through the network, and live programs are held in the Edge Cache in fast storage or memory so that each bit-rate of the live stream can be provided should multiple request be initiated from viewers.

In terms of operation, topology, and technology, OTT distribution is worlds apart from traditional RF broadcast and one of the major challenges to be overcome is dealing with who owns the network.

Telco's have been providing managed networks to television stations for nearly as long as there have been television stations. Analogue video, audio and SDI distribution services have all been available to us. But these bespoke services attracted a hefty price tag. The beauty of IP distribution, specifically with OTT over the internet, is that the distribution costs to broadcasters are orders of magnitude lower. But the price we pay for less is a reduction in control.

### Internet Limitations

When broadcasters started experimenting with OTT they soon realized that the internet could not withstand the amount of traffic being streamed across it. Although the public internet may appear to be "free", at some point, somebody somewhere must provide the infrastructure. To a larger extent this fell to the major Internet Service Providers (ISPs).

To understand public CDN it's worth considering the ISP business model. Essentially, ISPs make their money by providing a bidirectional data pipe to our homes. They also provide bolt on services such as television and film channels, but the bulk of their revenue comes from providing the data service to homes. There is really no incentive for them to boost data connection between two POPs (Point of Presence), for the public CDNs to then flood it again with even more streaming traffic.

Furthermore, a CDN builds on a managed network, whether private or public. The CDN fundamentally consists of the origin server, storage and edge-server. These components, combined with the network, make up the CDN. It's entirely possible for a CDN provider to work with a network provider, but not necessarily own the network.

A public CDN shares resource and doesn't guarantee bandwidth or latency. It may improve distribution compared to normal internet systems, but the components that make up the CDN are shared amongst several users, in the case of television, this would be several broadcasters.

### Dedicated Premiums

When broadcasters leased SDI circuits from Telco's, they paid a high premium for guaranteed bandwidth and latency. Telco's were able to deliver on this as they costed exclusivity into the service. However, the business model for the public CDN providers does not guarantee exclusivity. In fact, they actively promote sharing of the data circuit across many clients. Taking advantage of the distribution of data from statistical analysis, public CDN providers are able to provide an average data rate and latency.

But the devil is in the detail and averages often mask underlying peaks. For example, the latency may average at 10ms, but it might peak at 100ms, or even more, and this could have disastrous consequences for the reliability of the service. In the new connected-world, broadcasters are fighting growing competition from other service providers. They know all too well how easy it is for a viewer to switch to another channel should their service fail.

Due to the complexity of the internet it's often difficult for the broadcasters to determine where a problem has occurred and who is responsible for it. This is not necessarily a fault of the public CDN or ISP providers, but instead, is just a consequence of how the system works. One of the benefits of OTT distribution is that broadcasters are taking advantage of a system that can be shared. But with sharing comes compromise, and that is exactly what broadcasters are doing when they are using a public CDN.

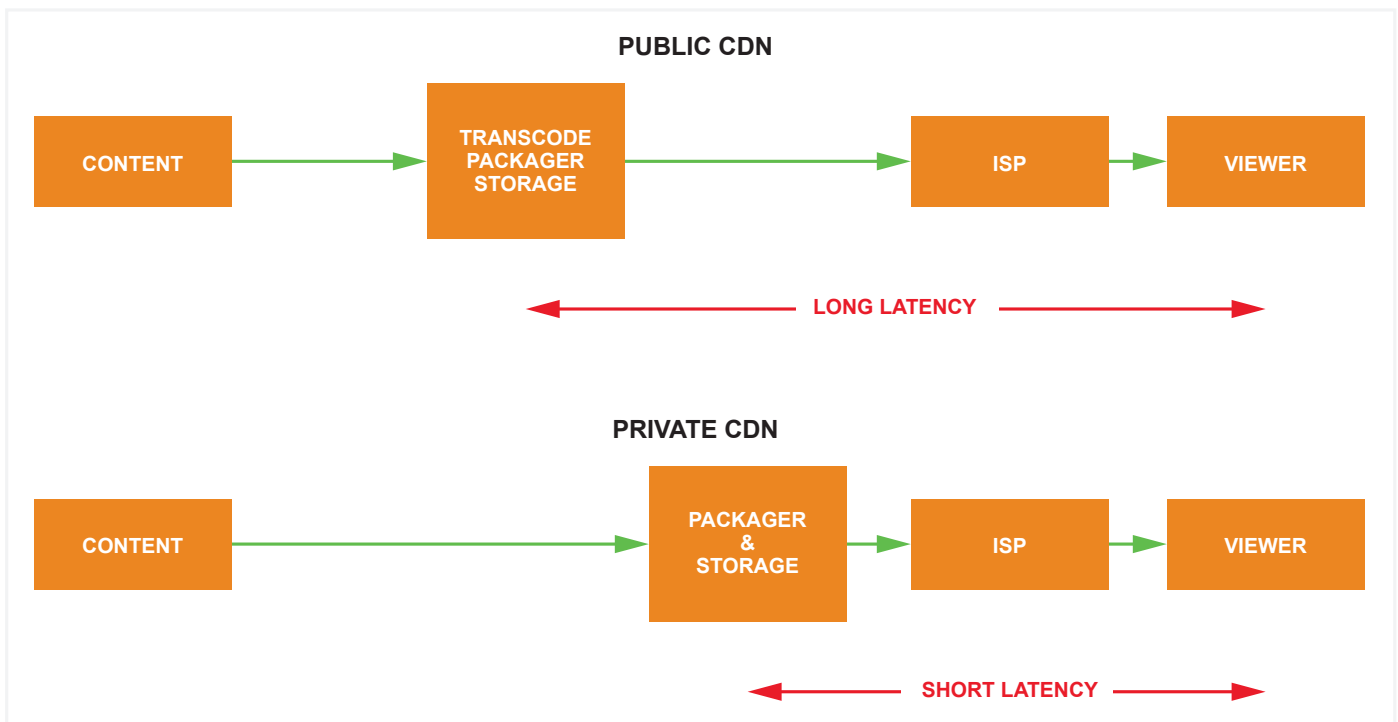
Although viewers have come to expect the flexibility of watching their favorite programs on the device of their choice, they also expect the same quality of service traditional RF broadcast mediums have always provided. From the viewers perspective, they don't really care how the signal reaches their mobile device, they just want to watch what they want, when they want, and how they want. Viewers expectations for live events are even higher as nobody wants to miss the winning goal in a premier game.

### CDN Control

Another alternative is to use a complete private CDN service as it gives a broadcaster full control over the distribution. They can place monitoring probes where they like and be able to monitor the system to incredibly granular levels of detail. Generally speaking, a private CDN consists of installing cache servers and monitoring systems, it does not include the fiber network or home delivery. The costs of this vary depending on the use-case and private CDN can be cheaper if sufficient viewers are watching. As a very rough rule-of-thumb, 50,000 viewers is the tipping point at which private CDN becomes cost effective. However, the number of viewers may change depending on the programs being streamed. For example, if a broadcaster routinely streams 500Gbps per day, it can be much more cost effective to use Private CDN. But if there are spikes up to 2Tbps one per month, it might be best to use the Public CDN for the extra 1.5Tbps. This isn't a linear relationship and many variables influence the point where public, private, or public and private become more efficient.

The third option that exists is the hybrid CDN model.

Public CDN's do have many advantages; they are cost effective and have deep reach into many of the ISPs needed to broadcast to wide audiences. However, they are shared, and this causes compromises for the viewers quality of service and experience. But this compromise doesn't necessarily affect the whole network, there will be many parts of the distribution that are working well enough for the public CDN to operate adequately and meet the expectations of the viewer.



As well as providing managed services, private CDNs provide the opportunity to move the packager and storage closer to the ISP and viewer. This takes advantage of the shorter latency between the viewers mobile device and the playout servers resulting in a much-improved quality of experience for the viewer. In this diagram, the transcode, packager and storage is closer to the customer in the private CDN than it is with the public CDN.

It's often difficult to monitor a public CDN. This is as much of a logistical challenge for the public CDN provider as anything else. Public CDN providers may want to move their equipment, upgrade it or just service it. Within the organization, there will be tight operational controls on when these tasks can take place through change control notices. These require agreement from all affected parties within the organization and are notoriously difficult to get signed off. It would be even more difficult if they had to require sign-off from many other broadcasters if their monitoring equipment was also installed.

Private CDN's do allow monitoring to be installed so the broadcaster can keep close control of how the network is performing. Again, broadcasters using private CDN's have better access to the managed network so are able to determine any anomalies quickly, but to provide private CDN for a complete OTT distribution network would result in a high level of complexity and cost that few broadcasters would be able to deal with.

There is less chance of the private CDN being shared between broadcasters. It's possible there could be some sharing but the private CDN service provider will establish brick-wall rate control between clients so that minimum viable latencies, bandwidths and jitter can be guaranteed as part of a larger service agreement.

### Hybrid CDN

However, if we keep the use of private CDN for routinely delivered content or areas we know congestion is likely to occur, then they can be used as part of a public-private partnership strategy. CDN capacity planning is critical here and with the correct monitoring, broadcasters will be able to determine where best to focus their resources.

Furthermore, greater efficiencies can be gained through the placement of storage, encoding and packaging servers. In our example discussed earlier, the packaging process would require approximately 18 streams of video and audio leaving the broadcaster for each service. Even the time delayed +1hour services would need to be streamed from the broadcaster. As these are effectively the same stream delayed by one hour, this is an incredibly inefficient method of operation.

Moving the storage and packaging as close to the edge as possible is a much more efficient and effective method of operation. In terms of processing power, packaging is relatively cheap. However, transcoding is resource hungry and tends to be centralized. By "the edge", we mean moving as close to the consumer as we can. This may well mean installing equipment directly in the ISPs POPs and in doing so the required data capacity needed in the core internet is also reduced and greatly optimized.

Instead of sending three versions, each containing six ABR streams (one each for HLS, DASH and Android), one version with six-bit rates needs to be sent over a private CDN and the packaging can be provided at the edge. At the ISP, the private CDN will have sufficient infrastructure installed to provide the services needed to package the streams for ABR distribution to the ISPs clients, who are in effect, the broadcaster's viewers.

## Efficient Internet Delivery

This method reduces the pressure on the core internet resulting in better delivery of the program stream to the ISP.

Monitoring is much more effective as the private CDN provider will have monitoring solutions installed as part of their service thus giving much greater transparency to the broadcaster. This will not only help with managing key parameters such as jitter and latency but will also assist in predicting viewer demand, so when it is at its lowest, the broadcaster can take advantage of the caching aspect of the private CDN and distribute their on-demand programs to the storage servers or consider using the computing processing power of the edge servers.

Placing distributed storage within the private CDN as close as possible to the consumer will reduce congestion on the core internet as users will effectively grab their program streams from the edge storage and associated servers and not have to request program segments back from the broadcaster. It's inevitable then that the latency between the viewer and the ISP is significantly lower than that between the viewer and broadcasters, thus further reducing delay and the potential for the dreaded "wait, loading" symbol on the viewers device.

## Predictable Delivery

Live programming is greatly improved as the distribution from the broadcaster to the ISP and then viewer becomes more predictable and will result in a much better quality of experience. If the core internet infrastructure isn't having to contend with the spontaneous, bursty and unpredictable behavior of the network caused by the on-demand services, then the service is bound to be better.

The edge storage and servers used for on-demand can also be used for the live programming and distribution. Again, only a reduced number of streams needs to be sent to the edge devices and they can provide all the packaging, processing and mobile device interaction needed to make the program delivery reliable and comparable to OTA broadcasting.

The combination of private and public CDN solves many of the challenges broadcasters face when providing premium OTT services. Combined, they provide the best of all worlds. That is the client reach of public CDN with the control and quality of service a private infrastructure delivers.

But providing private CDN's isn't just about delivering faster connectivity to POPs within the internet. It's also about distributing the core infrastructure technology to take it as close as possible to the viewer. OTT fundamentally differs from traditional OTA as there is significant interaction between the playout server and the mobile device. Instead of pushing the program stream, as happens in OTA, to the viewer, the mobile device requests packets of streams from the playout device as it needs them.

## Improving Viewer Quality of Experience

ABR technologies are greatly affected by any latency in the delivery network as they fundamentally rely on TCP to allow them to work reliably. As the latency increases, as would happen over a highly utilized core internet link, the reliability of the program greatly deteriorates leading to the "waiting buffer" issues and even picture break up. Moving the interaction between the mobile device and playout server to the edge greatly reduces latency resulting in a much-improved user experience, and therefore viewer retention.

Hybrid CDN also scales to the cloud. Many broadcasters are taking advantage of cloud-playout using virtualized storage and servers. Often, the public cloud provider has data circuits directly into the main ISP POPs throughout the world. This further encourages the use of private CDN as the edge servers can be placed close to the viewer in the ISP.

Hybrid CDN should be a key consideration for any broadcaster looking to provide reliable OTT for their viewers. The combination of deep reach and control is a winning formula. Working together, it's even possible to have the public CDN acting as a backup to the private CDN. Quite often in broadcasting we find that a partnership of methods provides the best overall solution, and Hybrid CDN is one of those.

# The Sponsors Perspective

## Reaching OTT At Scale - Pitfalls And Opportunities

By Ryan Nicometo, SVP of Product, Vecima

OTT delivery continues to expand to meet the relentless growing consumer demand. This trend shows no chance of abating and technologists are continually looking to innovation to scale infrastructures accordingly. In this sponsors perspective, Ryan Nicometo, SVP of Product for Vecima, presents the OTT Scale – Maturity Quadrant to help decision makers clearly see the potential technical challenges and opportunities to meet the demands of OTT delivery.



### OTT On The Rise

OTT video consumption continues to grow, and quickly. In January 2020, Conviva reported global streaming volumes in 2019 were 58% higher than 2018. In February 2020 the Streaming Video Alliance stated over 80% of consumers are streaming more video than they did three years ago. And over the last 6 months, major media companies have made big announcements about new OTT service launches that are significantly more ambitious compared to previous efforts.

OTT is growing fast, but despite the many Exabytes of content already being delivered annually, we're only at a relatively early stage of adoption. As more and more people shift to OTT consumption, what we currently measure as hours of online video viewing per person per year will become hours per person per day, just as traditional TV viewing has been.

Supported by





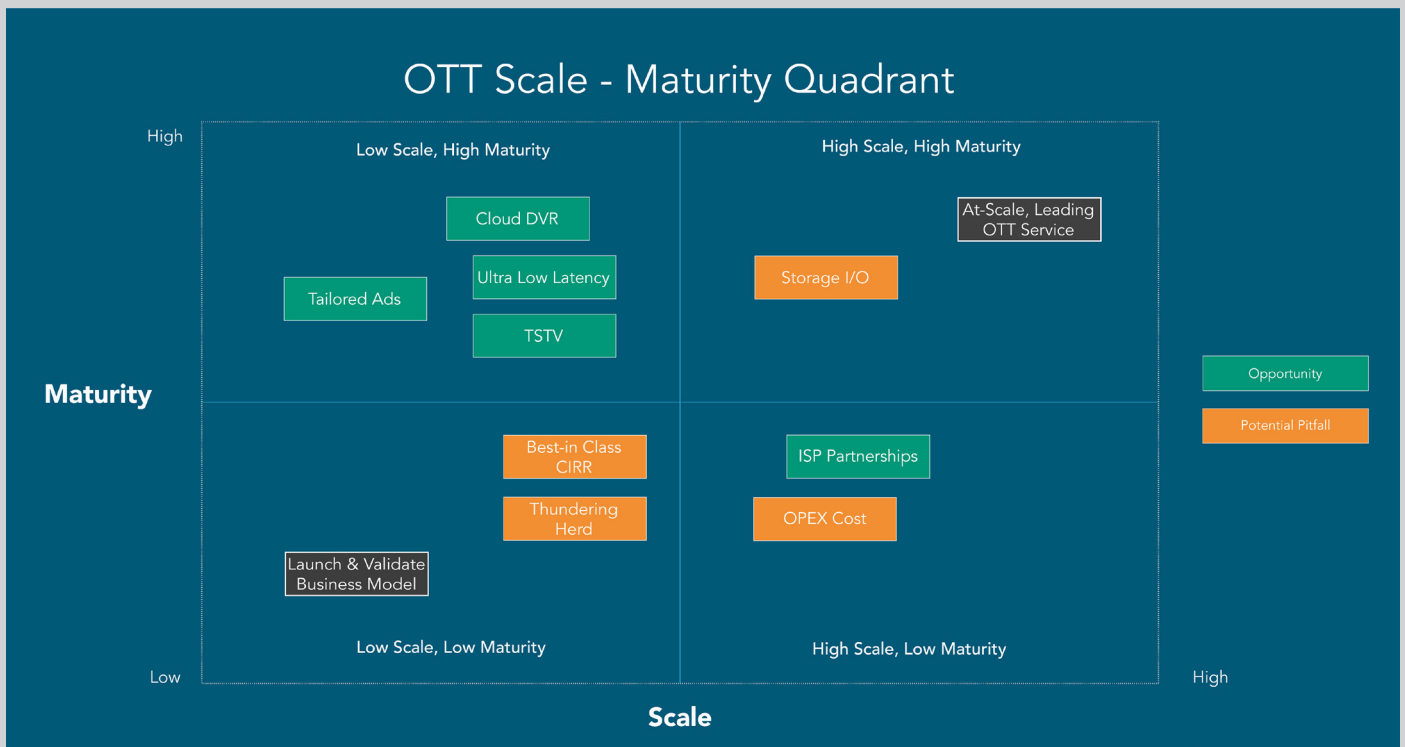


Figure 1. OTT Scale – Maturity Quadrant.

The world’s largest OTT Operators like Netflix, Hulu, YouTube TV, and Sling TV have reached “scale” already. But what does it mean to achieve this milestone? Once you get there is your service static and unchanging, or is there still work to be done? The answer depends on how you got there.

Vecima has created an **OTT Scale - Maturity Quadrant** to describe the relationship between Scale and Service Maturity. Using the model, decision-makers can clearly see the potential technical challenges and opportunities along the path towards an industry leading service.

**Ready For Launch**

In the early stages of growth, platform stability issues need to be resolved, from back office systems required for authentication, billing, etc. to the Content Delivery Networks (CDNs) themselves to handle the high number of requests. We’ve all seen the headlines about CDN platforms performing badly under the load of a large audience.

**Thundering Herd** is an issue created when a large number of simultaneous requests for the same content, which is common in live events like sports and news, savvy operators will want to avoid – fail on this and your audience cannot grow. **Connection Induced Rebuffering Ratio (CIRR)**, is a metric of stream delivery quality – fail on this and your audience will either leave or complain.

Supported by



The end-game is to reach the **At-Scale, Leading OTT Service** position. In this environment, the quality of content and the ability to meet consumer needs is paramount. Netflix holds this position on a global basis but is being hotly pursued by the rest of the field and has many local in-country competitors.

**Growing Up**

Once you’ve successfully launched an OTT service, the next step is either growing the number of subscribers or maturing your offering by adding additional services.

Some key service offerings include:

- **Time Shifted TV (TSTV)** – adding high-performance recording and storage capabilities to allow linear/live pause, rewind, restart and look-back functionality. This service is typically offered for live content only and is not available on many OTT services today, however, is commonly found on competitive platforms for cable and telco operators. TSTV requires an investment in high-performance storage infrastructure.
- **Ultra-Low Latency** – reaching levels of < 5-8 seconds of visual latency from Encoder to Client is the goal to bring OTT services in line with cable, IPTV and satellite delivery. This requires deep integration in the Encoder-Storage-Origin-Cache-Player chain to handle sub-second fragment sizes. Working within the established HLS and DASH standards is important to maintain a scalable, open platform.

- **Targeted Ads** – personalizing the video content with targeted advertising enhances the user experience while also driving incremental revenue opportunities for the OTT provider. The video edge provides the integration point (“glue”) between the consumer knowledge held in the back office and the ad campaign information held in the ad decision system. In addition to caching linear or VoD content, the video edge leverages the existing edge-based manifest re-writing process to provide a dynamic and scalable solution for audiences of any size.
- **Cloud DVR** – for parity with set-top-box services, cloud DVR allows OTT customers to record and keep linear/live content. Normally a service provider offers an amount of storage capacity to their customers, and manages content recordings and total usage optimizations in a central or regional storage cluster. For many OTT Operators this may be a nice-to-have feature while for others it will differentiate them from their competition. To deploy Cloud DVR requires scale-out, high-performance storage that can stream content from fast storage while cost-effectively archiving content which will be played at some point in the future.

All these features require investment in the Content Delivery Platform, either in new capacity or in new integrations. At Vecima, we anticipate that OTT operators moving towards fully mature, business impacting service will make these investments in infrastructure for the maximum financial return. The OTT pioneers have led the way so far and proven these types of offerings are not only compelling to users, but ultimately profitable for their organizations.

### Crawl, Skip Walk, Run

How a platform will scale is often an afterthought to many system designs. It is usually more important to just get something delivered quickly. Quite naturally, it is difficult to see scaling issues until they appear. But by starting with a platform that is actually designed to operate at significant scale, it is possible to avoid a big growing pain down the road.

In the lower-right quadrant of figure one, a relatively immature OTT service can reach high scale by having the right content and a cost structure attractive to users. Imagine a scenario where the creator of the highest value content in the world prices their offering low compared to competition. Or imagine a household -name broadcaster offers exclusive programming, plus linear channels and VOD content on its OTT platform. You would expect a high take rate. If that content creator doesn't have their own infrastructure, they'll need to use cloud based computing and public CDNs to distribute their content.

It is under this set of conditions that **OPEX Costs** can quickly become a challenge. Leveraging infrastructure from various cloud and public CDN providers comes at usually a high cost. There are several examples of these cost models being poorly understood, which led to business cases not working out the way they were planned to.

As a way to reduce OPEX costs, the Edge Caching component of the Content Delivery Platform is the key. Deploying Edge Caches inside ISP networks is already common amongst the largest OTT Operators. This strategy helps relieve backbone network traffic, and therefore OPEX costs from public CDN providers. It can enable a change from OPEX to CAPEX if desired. And adjusting the standard pay-per-GB model for CDN services to a more predictable pay-per-Gbps model is attractive to large OTT Operators.

Importantly for delivering OTT at scale, the Edge Cache can enable closer **ISP Partnerships**. ISPs share the consumer with the OTT Operator, so there is mutual benefit in serving them well. By deploying caches closer to the consumer in collaboration with ISPs, there is also a nice side benefit of latency reduction for live services.

### High Scale, High Maturity

Time Shifted TV, Cloud DVR, and even high consumption rates of On Demand all can drive massive amounts of **Storage Input / Output (I/O)** workload on an OTT provider's platform. This amount of I/O can break public clouds and privately deployed storage clusters alike.

The storage platform has to handle high-volume recording, storing, and streaming, across a wide range of content and client devices. Starting with a platform that is designed and proven to scale makes sense in order to avoid a major content migration just as the OTT service reaches high scale.

### The Vecima Advantage

At Vecima we think that OTT Operators will, over time, need to take full control over the full **Content Delivery Platform**. While we don't expect OTT operators will look like traditional cable and IPTV operators that build their own networks, we expect OTT operators to take a high level of control over the full content delivery platform to optimize the experience of their audience. We expect them to build end-to-end content delivery platforms – a network of streaming, storage and monitoring servers – that optimise the performance, cost and business efficiency of serving content to their audiences.

Regardless of which quadrant you find yourself in, Vecima has products and services that can help you avoid the challenges we've highlighted here, and others.

Supported by

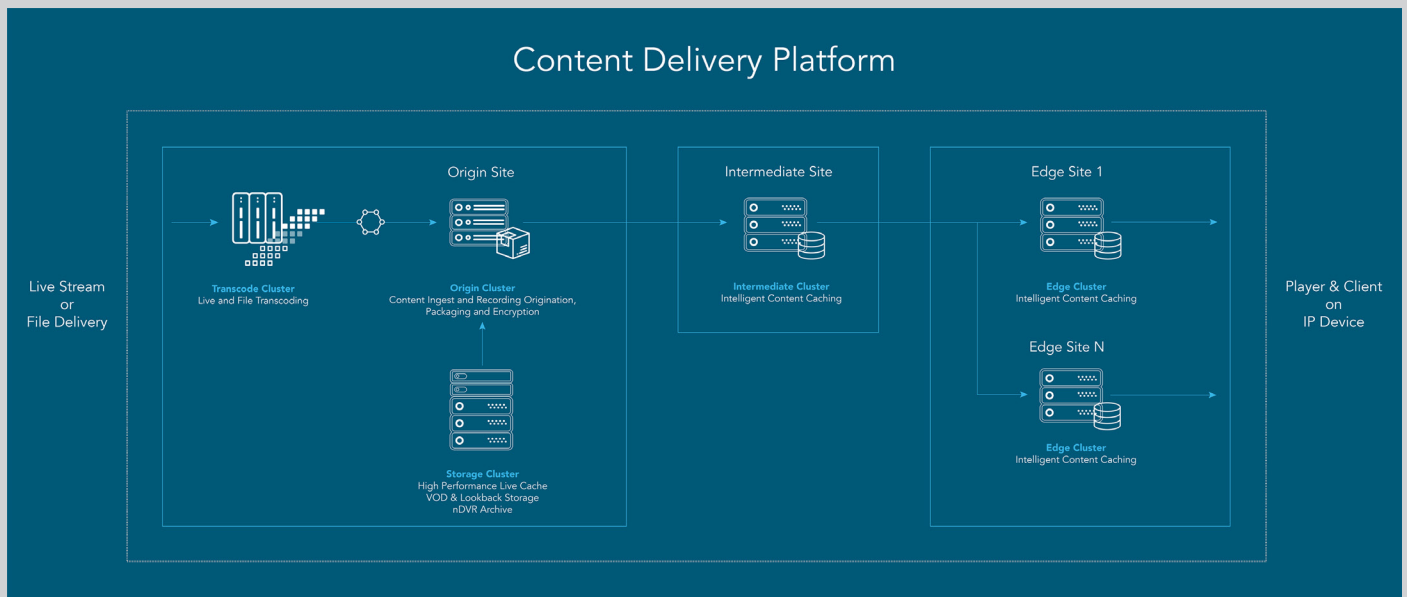


Figure 2. The Vecima Content Delivery Platform.

### Build or Buy?

When the decision has been made to insource portions of your OTT offering, “build or buy” is the next decision. As an experienced CDN product development company, we obviously believe there are more benefits to buying than building. By buying the right platform that has been proven to scale, OTT Operators can avoid the inevitable pitfalls of developing a platform for themselves. Partnering with the right technology partner should yield faster and more cost-effective results. If you’re considering building your own OTT platform, you really should talk with us before embarking on this journey.



Ryan Nicometo, SVP of Product, Vecima.

### The End Game

If content is King and content distribution is Queen, then the Content Delivery Platform can be a major differentiator for an OTT service. Its performance, features and cost will heavily influence your customer satisfaction, brand reputation, and profitability.

It is easy to see content delivery platforms as a commodity, and arguably some of the underlying technology is. But how that technology is deployed, configured, developed, and how it scales and performs under pressure, are not commoditized at all. We are moving into a space where the industry hasn’t been before – OTT at scale. Choosing the right technology, founded on a strategy of achieving scale with least risk, is going to be one of the most important influencers of success.

Vecima stands at the ready to assist your organization in planning, growing, maturing, or scaling your OTT service.

Supported by



For hundreds more high quality original articles and  
Essential Guides like this please visit:

[thebroadcastbridge.com](http://thebroadcastbridge.com)



04/2020

Supported by

