# phase 4

# media streaming with IBM cloud video streaming

# Live Video Delivery System Built for Scalability

## Introduction

With online video traffic continuing to surge to ever-greater heights, many online content distributors are finding themselves taxed trying to keep up. This becomes especially critical when dealing with live video assets, where the added value proposition to the viewer is largely tied to viewing that content in the moment. As a result, disruptions due to congestion become mission critical. Another piece of the puzzle is the fragmentation in possible viewing options, both across devices or even due to the strength of the viewer's connection. This presents two major challenges in reaching huge audiences, both in the ability of the system to scale effectively and also being able to reach viewers regardless of their personal setup.

Realizing the critical nature of live streaming and the need to reach a large, diverse audience with this content, IBM Watson Media has invested in an intelligent, dynamic and scalable solution. This paper covers how the IBM Video Streaming service is specifically tuned for live video through a combination of reliability, scalability, and QoS (quality of service). It goes in depth on how IBM Watson Media manages both the delivery side and also how content is live transcoded to greatly expand the content's potential reach.

## Dynamically Scalable, Global Architecture for Live

Unlike traditional VOD (video on-demand) serving architectures, IBM Watson Media's dynamically scalable, intelligent architecture is built to handle sudden traffic surges. This caters to the unexpected nature of live broadcasting, which can have unpredictably large audiences. For example, an enterprise live streaming a product launch can find themselves initially with a modest audience only to have viewership sky rocket when the actual product is about to be revealed.

Consequently, IBM Watson Media has built out its own smart, dynamic network, including a CDN (content delivery network) with POPs (points of presence) in the United States, Asia, and Europe. This network is specifically architected for scaling live and linear broadcasts where a massive number of viewership is consuming the same content simultaneously. To address this, the Video Streaming architecture was designed specifically with the key challenges of live video transmission in mind. These specially tuned software and infrastructure solutions include:

- A proprietary media server designed for serving a large number of high quality streams per server
- Proprietary TCP/IP packet-level optimization for mitigating network problems, such as packet loss, congestion, jittery transmission, etc.
- An intelligent, dynamic network of multiple public CDNs that allows IBM Watson Media to dynamically scale across both its own proprietary CDN, as well as leveraging the resources of the world's largest CDNs through per-user QOS monitoring and provider switching

The combination of IBM's own CDN with external, global CDNs allows IBM Watson Media's network to scale virtually to support millions of simultaneous connections. This allows a high quality of service and also cost-efficiency across the globe, as enterprises don't need to provision separate delivery methods to increase their global reach or establish backup solutions.

IBM Watson Media's deep experience with the major global and regional transit providers and CDNs allows its network to achieve three unique feats:

- dynamically and economically scale to meet sudden demand for streaming data
- optimize viewer sessions at the geographical level
- provide good quality even if transit providers or CDN edges are failing through detecting quality problems and switching between stream sources or even CDN providers

## The IBM Watson Media Server

The IBM Watson Media Server is a proprietary server application developed for the purposes of large scale live stream delivery with complex server-side business logic and access control. The IBM Watson Media Server maintains a bidirectional connection with each connected client. Content is typically ingested via RTMP while delivery supports HTTP streaming, HTTP live streaming (HLS) protocols, and also legacy RTMP. The media server provides IBM with scale and flexibility, empowering enterprises to stream content without having to worry if their video content might go viral and attract huge audiences.

The bidirectional connection allows not just for multimedia streams but also statistical, telemetry and access control data to travel back and forth between server and clients. With full control over the server-side (application) layer, it is possible to control individual clients, get statistical or telemetry data and aggregate them in real time.

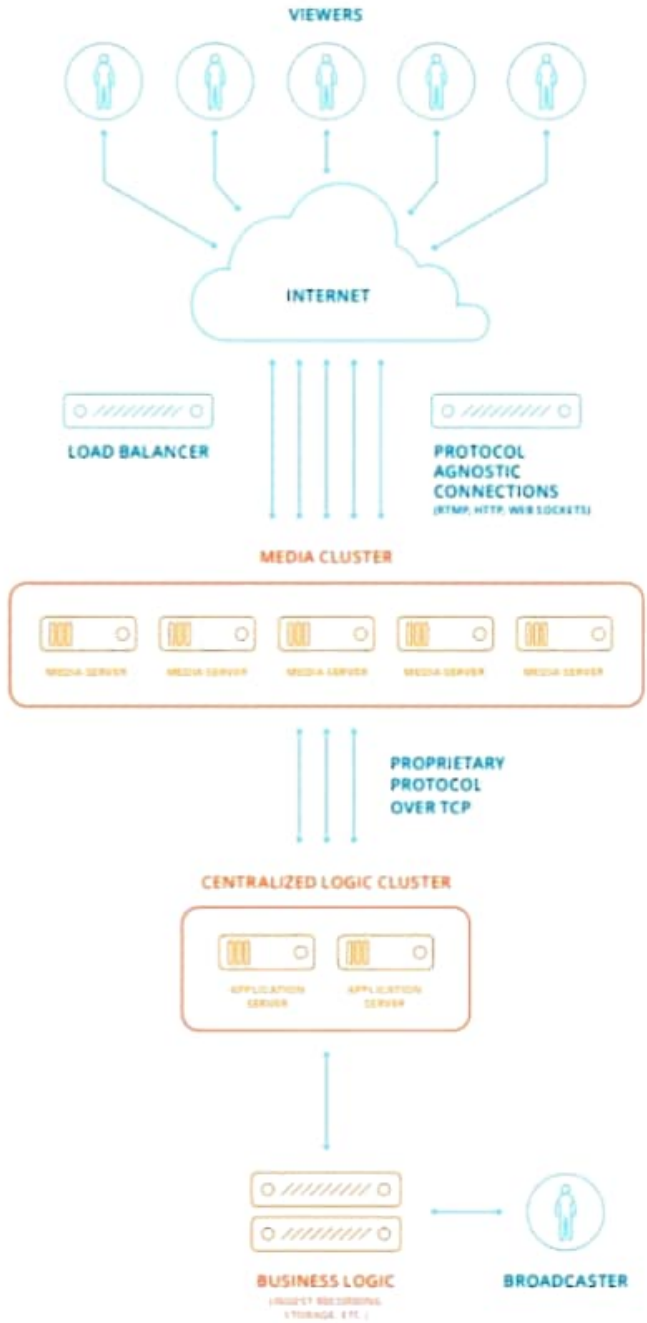



Figure: Bidirectional Connection

## Supported Protocols and Codec

Having total end-to-end control over the serving, transcoding and chunking technology allows IBM Watson Media to be flexible in terms of delivery methods and codec formats. That said, the industry is moving toward HTML5 delivery on desktops and other playback methods. So even though IBM Watson Media can provide support for several protocols through a Flash version of the player, such as VP6, the recommended codec format for sending an RTMP stream is H.264 for the video codec and AAC as the audio codec.

In total, the following protocols are supported, but keep in mind some need to be delivered over IBM Watson Media's Flash player as opposed to its HTML5 player:

### RTMP

Commonly used TCP (Transmission Control Protocol) based real-time protocol, supports H.264, H.263 and VP6 for video along with AAC, MP3 and Nellymoser for audio delivery.

### HTTP STREAMING

IBM Watson Media's adaptive HTTP chunk based protocol can support HTML5 or Flash playback. For HTML5 desktop delivery, through HTML5 MSE, the service uses mp4 chunks and supports H.264 for video and AAC for audio delivery. For Flash on desktop, the service uses FLV container chunks and supports H.264, H.263 and VP6 for video, AAC, MP3 and Nellymoser for audio delivery.

### HTTP LIVE STREAMING (HLS)

Apple's live streaming standard used on mobile and CE devices. Supports H.264 and AAC.

## State of the Art Live Transcoding and Chunking

Device capabilities such as performance, preferred multimedia formats, or even the supported transport protocols are currently very fragmented – as a result one format or protocol that works well on one device may not work (or perform poorly) on another device. IBM Watson Media's proprietary live transcoding solution enables an optimal viewing experience across different devices. From one single ingested stream, ideally high definition, it can produce multiple audio and video streams optimized for the characteristics of the connected devices.

The transcoding layer accepts various input formats and outputs H.264 with AAC sound. As previously mentioned, though, it's recommended to use H.264 as the video codec and AAC as the audio codec so that content can reach HTML5 players on desktops. As for the output, it's produced in multiple resolutions, frame rates and bit rates, which will be playable on a wide spectrum of multimedia devices.

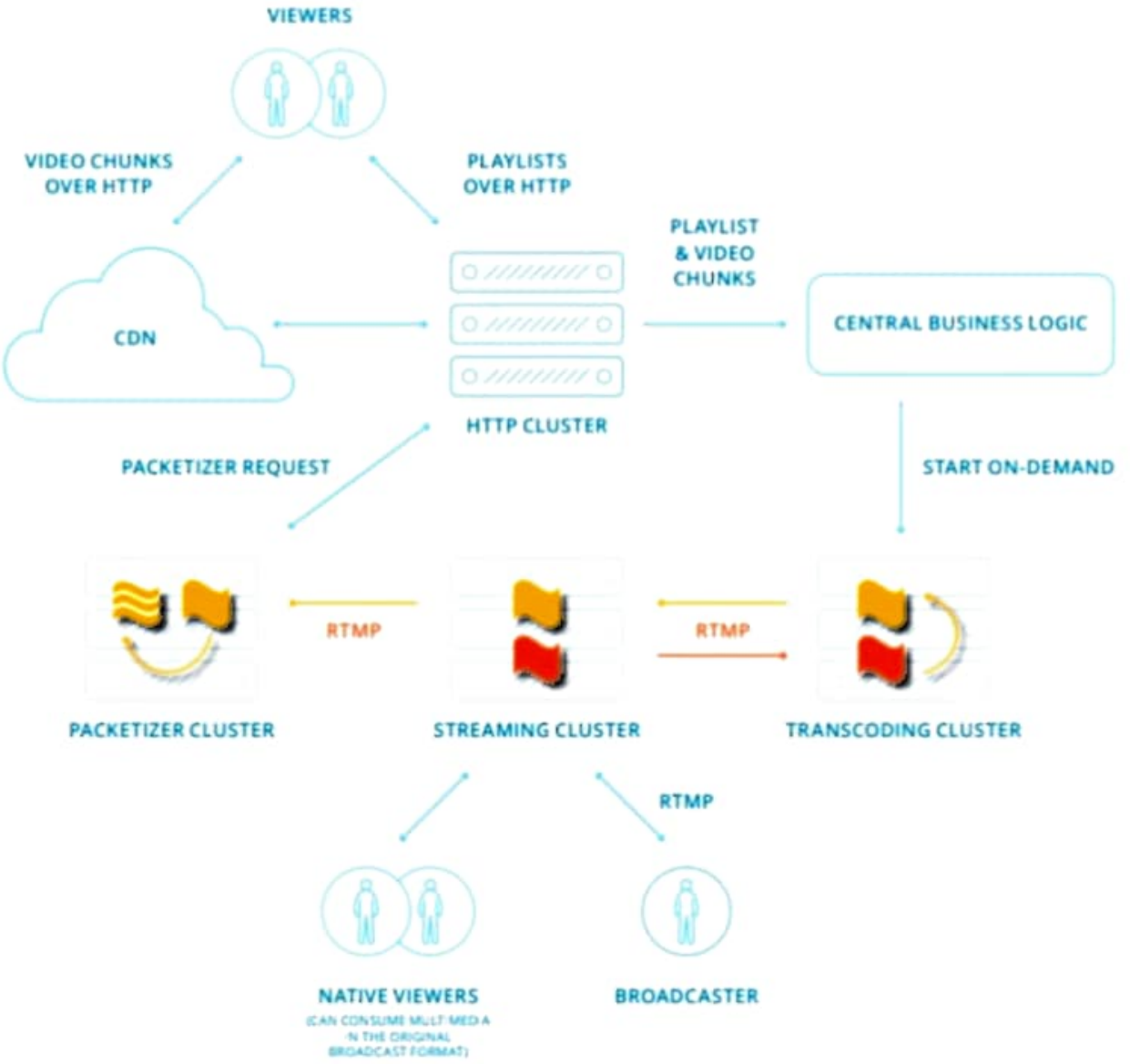


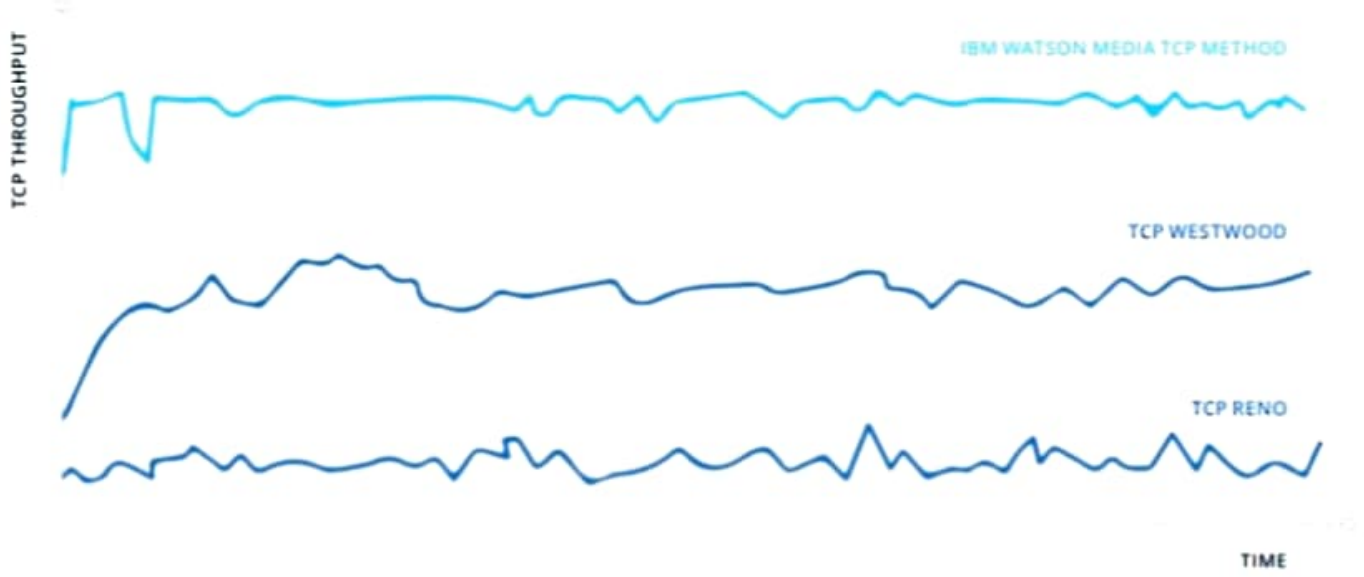

Figure: Transcoding and Chunking Process

## Effect of Packet Loss on TCP throughput

Multi-format transcoding, being very processing-expensive, is intelligently controlled server side through starting/stopping individual processes in order to deliver the required formats but nothing more. This saves capacity in the transcoding cluster.

The transcoding algorithm is also proprietary; it leverages GPU support and is capable of transcoding multiple video frames in a single GPU cycle. This is achieved by transcoding one large video frame which consists of multiple video frames laid out on a single video pane.





## Network QOS Advantage: IBM Watson Media TCP Method

IBM Watson Media uses a novel congestion control algorithm targeting both emerging wireless networks such as LTE, WiMax, Wi-Fi, HSPA as well as high speed long delay (high BDP) networks.

TCP (Transmission Control Protocol) is a reliable transport layer protocol that is widely used on the Internet. It is the underlying protocol for HTTP, RTMP and many other applications. Generally speaking, the congestion control algorithm is an integral module of TCP that directly determines the performance of the protocol. Standard congestion control algorithms such as TCP-Reno and TCP-NewReno performed well for several decades but are found to perform poorly over wireless and high Bandwidth Delay Product (BDP) links.

To improve TCP performance over wireless and high BDP networks, many TCP variants have been proposed, including TCP Westwood and TCP Veno for wireless applications along with Compound TCP, TCP CUBIC and FAST TCP for high BDP networks. Although these algorithms have achieved performance increase in their respective target applications, designing a TCP congestion control algorithm that performs equally well in both wireless and high BDP networks is still a challenge. On the other hand, with the deployment of wireless networks, such as LTE, WiMAX, as well as high bandwidth, real time applications such as multimedia over TCP/HTTP, it is required for the TCP congestion algorithm to handle both wireless connections with random radio related losses as well as congestion-introduced issues typical for wired high BDP networks.

IBM Watson Media's TCP congestion control algorithm achieves major improvement in this highly important area. The quality increase in stream viewing and broadcasting gained by improving this technology puts IBM Watson Media ahead of alternatives in terms of streaming QOS and delivery costs.

**Traditional TCP congestion control algorithms are found to perform poorly over wireless networks:**

- **Traditional TCP: Packet loss = Congestion**
- **Wireless networks: Packet loss ≠ Congestion**

The performance of IBM Watson Media's TCP method on a Wireless network with the following attributes: 5Mbps available bandwidth, 1% packet loss, 100ms round trip time.

## Quality Optimization of Delivery

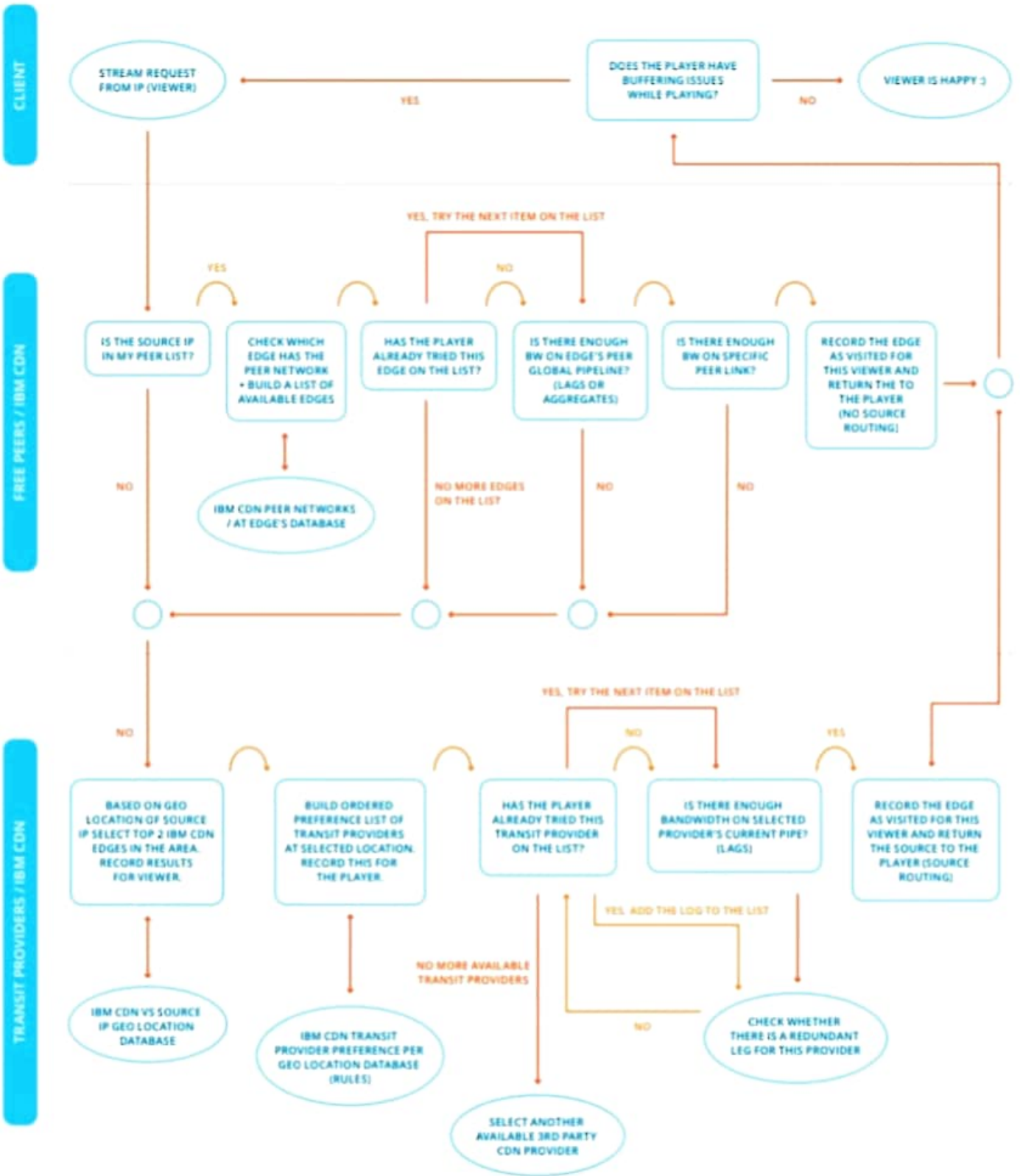
IBM Watson Media utilizes several delivery sources, some of which are internal while others are global 3rd party CDNs or regional 3rd party CDNs.

The choice of delivery source is optimized on a per-user basis, prioritizing for optimal quality. To execute this, IBM Watson Media collects real-time usage information about the peering lines and transit lines as well as usage on the supported CDNs and also geo based quality information from connected viewers.

The simplified decision making flow is as follows:

- Select an edge cluster based on the geolocation of the visitor
- Based on the user's IP address and AS code check if IBM Watson Media can serve the content from any of the peering lines supported in the given edge cluster. Use a peering line with free capacity if available.
- In the absence of peering capacity check if there are any usable transit lines with free capacity on the given edge cluster. There is an algorithm managing usage on transit lines and optimizing them for balanced 95th percentile traffic.
- When falling back to a third party CDN, CDN choices are prioritized partly by historical performance in the given region.

The following diagram displays the steps of the described logic:

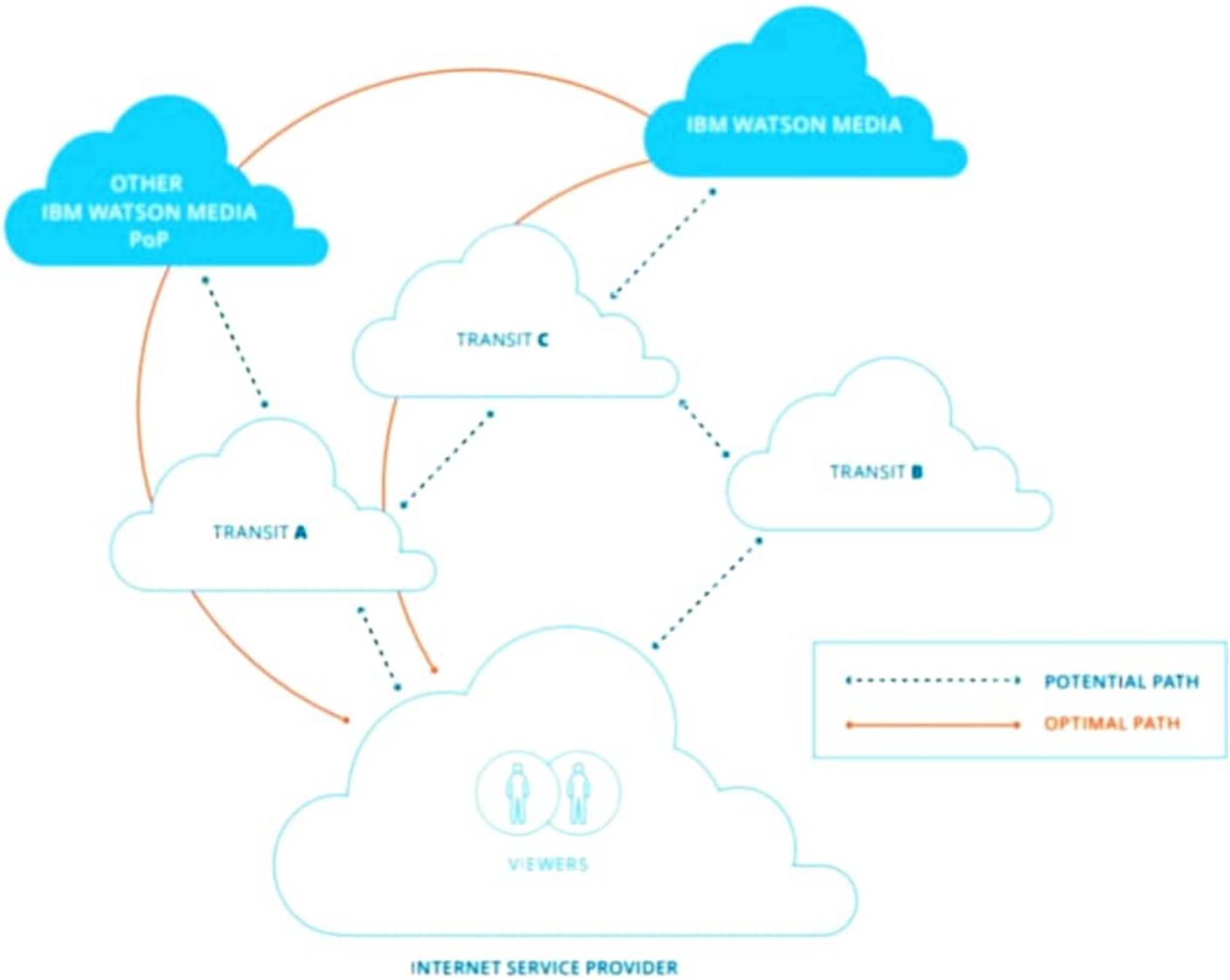



IBM CDN = IBM Watson Media Content Delivery Network

## QOS Optimization in a Multi-Provider Environment

While cost optimization plays a part when directing viewers to a specific delivery option, IBM Watson Media does not compromise on stream quality. In fact, users watching content will get served in a manner that delivers optimal available quality.

The following diagram shows a simple ISP topology and the Video Streaming Platform. IBM Watson Media is able to detect congestion or other network problems which affects the quality of the stream delivery and switch between stream sources (re-route the stream) to provide optimal experience to viewers.



## This intelligent optimization is made possible by a number of techniques:

- Static server side geo rules: transit lines and CDN providers have variable level of quality of service in different regions of the world. While some of them may work well in the US they may perform less well in Asia or Europe and vice versa. IBM Watson Media has a global, up to date picture of their performance and uses this information to make optimal decisions on a per user basis.

- Real time client side quality module: while the playback client will start with the predictably optimal cost / quality choice of delivery method it is possible that the preselected source may have quality issues during playback. To mitigate this the client continuously monitors quality related performance metrics, and in case of QOS issues seamlessly switches over to the next available stream source.

- Real time server side quality module (beta): the next evolutionary phase of the client side switching module, the server side quality module keeps a tally of QOS reports coming in from playback clients per provider and AS number uses this information to predictively keep new clients from using a problematic provider-AS pair in case there repeating issues detected on that pair.

## Ultimate Scale with Optimal Quality

The Video Streaming platform was built with the goal of being able to not just scale effectively for massive events, but also do so in a manner that presents improved quality for viewers. This includes removing geographic barriers, for a truly global solution through being able to tap into a diverse pool of CDNs. It also encompasses live cloud transcoding of video, to offer an optimal experience regardless of the viewer's connection speed while also simultaneously delivering video to any device.

This platform and technology is available for Video Streaming, which has delivered content for over one million concurrent viewers on a single broadcast. It's a highly scalable solution that allows enterprises to deliver content, from product launches to media events, at tremendous viewership scale.