SCTE CABLE-TEC EXPO 22
SEPTEMBER 19-22 • PHILADELPHIA, PA

2022 Fall Technical Forum
SCTE • CABLELABS • NCTA

Considerations For the Delivery of Latency-Sensitive, Compute-Intensive Experiences Over a Communication Network

A Technical Paper prepared for SCTE by

Dhananjay Lal
Senior Director, Advanced R&D
Adeia
3025 Orchard Parkway, San Jose CA 95134
(513) 225 4948
Dj.lal@adeia.com

Table of Contents

List of Figures

List of Tables

1. Introduction

For Augmented Reality (AR) and Virtual Reality (VR) devices to become truly immersive experiences, the industry is investing in R&D focused to address how they can achieve a goal of 20 millisecond (ms) motion to photon (MTP) latency [1] by modelling latency in concert with other systems.

In cloud gaming and other cloud content-streaming scenarios, 100ms "button-to-photon" latency is deemed adequate and is achieved through cloud rendering and bit rate encoding specified at the content source. While some latency variations occur in the transmission path, depending on network conditions, this 100ms "button-to-photon" is usually sufficient for these applications. But for AR/VR, prior art has made the case for 20ms MTP latency to avoid spatial disorientation, motion sickness and other adverse experiences for users.

In Cloud gaming and VR/AR, viewers are not passive—their interactivity (e.g., with objects and avatars) requires active modification of the content within the experience itself. In essence, this means that for non-interactive video experiences, content delivery is a "one-way" transfer of bits over the network (after transcoding) to the client (viewer), whereas interactive content delivery is two-way, where the viewer input from the client is used to determine subsequent content, which must then be rendered prior to transcoding and transmission. This imposes tighter latency requirements in AR/VR content rendering. Achieving 20ms MTP is practically impossible with current network architectures unless the cable service provider makes compute resources available deeper in its network.

This paper discusses a practical approach to cable delivery architectures where compute intensity and latency become the primary determinants of the end-user experience and are inherent properties of the content served to the subscriber. It describes a method and system for delivering latency-sensitive, compute-intensive experiences over a network that allows communications service providers (CSPs) to deliver latency-dependent compute to a subscriber, in concert with an application service provider (ASP), by employing seamless interactions between system components.

2. The Need for a Different Kind of Compute: Interactive Content

Current IP-based media delivery, for example, live and VOD (video on demand) TV content, has a latency sensitivity of ~500ms to a few seconds, dependent on various considerations, most importantly network conditions and buffer size at the client. While CSPs that offer programming can serve media experiences from within their own network, i.e., from a server in close proximity to their customers, there is little competitive advantage to be derived for the customer experience. This is evidenced by the rise of OTT (over-the-top) media service providers that provide programming via servers outside the CSP network. Other benefits, however, accrue with customer proximity to the CSPs such as reduced transit costs with shorter backhaul traffic, as well as access to consumer profile data.

In the context of current programming, we may logically separate the CSP that provides the packet route to the viewer from the Cloud, which we may view as a logical entity where content resides.

To explain an IP media delivery architecture, we may abstract out the network from the point of view of the media source in the Cloud and the consumer in the home. The original source file is encoded (compressed) at various bit rates, stored as a multi-bitrate asset (or chunks of the asset, each of which is multi-bitrate) and then delivered to a client based on requested parameters and network conditions, where it is then decoded and displayed.

Encoding, in this context called *transcoding*, remains the cornerstone of 2D media delivery on flat screens, big or small. At the server, transcoding may be performed by a combination of software and hardware, including CPUs, GPUs, FPGAs, etc. Depending on whether the programming is a live-stream broadcast or a VoD asset, the selection of transcoding platform and architecture as well as network protocols can vary significantly.

It is important to note that current media delivery architecture is non-interactive when it comes to content modification. While the quality with which the content is delivered may vary based on network conditions, and ads may be inserted through instantaneous decision-making in local zones, the actual programming content typically does not change based on viewer input.

Interactive content, on the other hand, can be altered by a viewer. In this sense, the viewer is not passive, but rather is immersed in the experience. Examples include games, Augmented Reality/Virtual Reality (AR/VR) experiences and holograms rendered on 3D displays. Interactivity may include content modification, for example, shooting and killing a formidable foe in a FPS (first person shooter) game that prolongs a player's ability to remain in the game (the content), but also content rendering in real-time encoded AR/VR based on the user's viewport, which controls what the user decides to focus on in their real (in case of AR) or make-believe (in case of VR) environment, ultimately changing what is rendered and displayed "on-the-fly". Streaming 360° video VR is a notable exception, where the viewport based on the user's pose is picked from within the pre-encoded video frame sent to the Head-mounted Display (HMD). It is non-interactive because the content is neither rendered nor encoded in real-time.

The key differences between non-interactive media like television and interactive media like gaming/AR/VR may be described as follows. Figure 1and Figure 2 describe the abstracted delivery of media for non-interactive content versus interactive content.

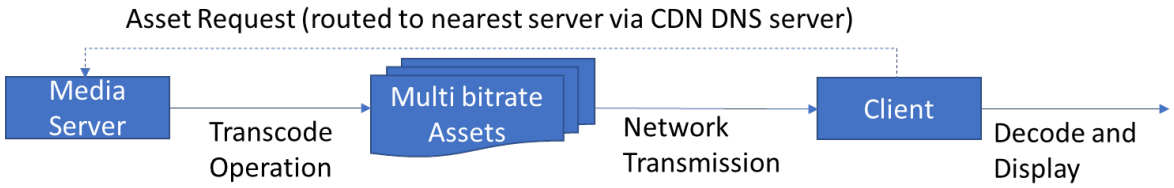


Figure 1- Non-interactive content delivery

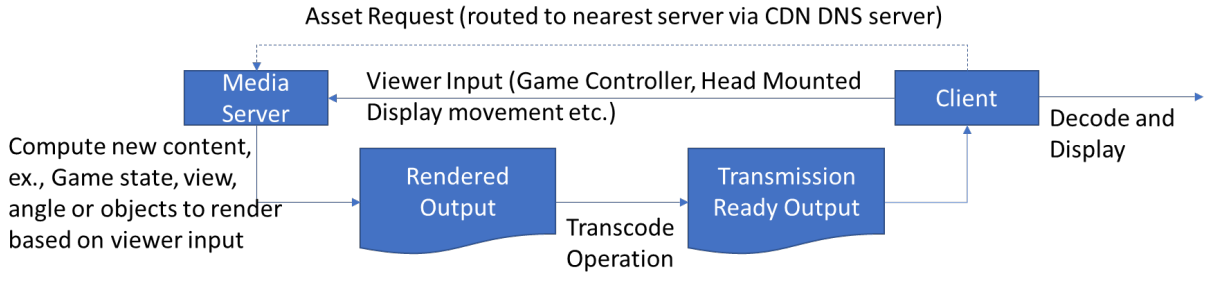


Figure 2 - Interactive content delivery

As mentioned, the key difference between non-interactive and interactive content is that non-interactive content delivery is a "one-way" transfer of bits over a network (after transcoding) to the client (viewer),

whereas interactive content delivery is a two-way street, where the viewer input from the client is used to determine subsequent content, which must be rendered prior to transcoding and transfer over the network.

In the case of non-interactive content, provided there are reasonable network conditions to deliver the bitstream to the client, and the client can buffer for intermittent interruptions in the network or processor use, the viewer gets a good experience. However, for interactive content, the client is typically expecting a higher bitrate (due to higher frames per second and resolution) while at the same time expecting the server to accept and process its current input to determine the subsequent state of the content.

This imposes much tighter latency requirements on interactive content rendering. Using gaming as an example, if the frame displayed at the client is more than three to six frames after the game input was issued, then the user experiences "lag", i.e., feels that the game is non-responsive. Typical game experiences are served at 60 frames per second (fps) which gives the system 50-100ms to present a frame response back to the user from the time that the user issues their input.

There are many public discourses about optimizing latency for Cloud Gaming [2][3]. Figure 3, from Nokia, is a sample representation of the system latency budget for Cloud gaming [4].

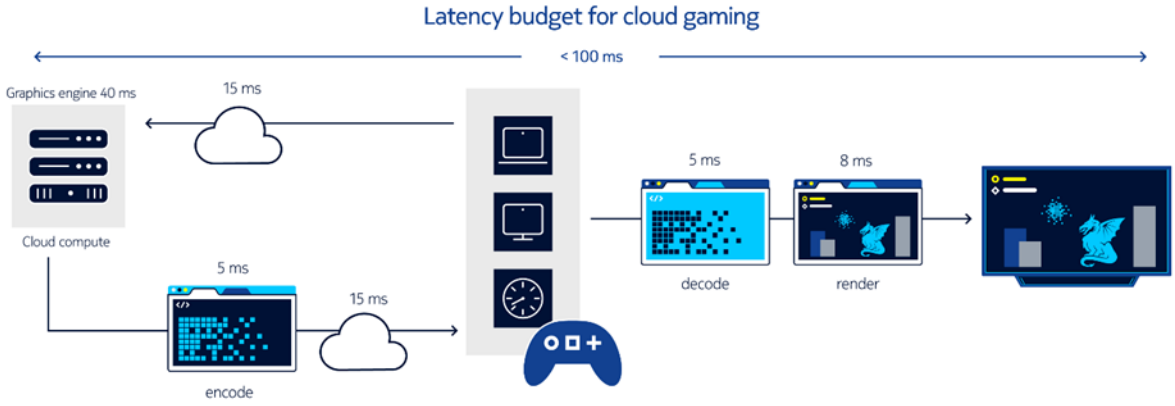


Figure 3- Sample Cloud Gaming System with latency budget

In the above example, the complete response time, i.e., the "button-to-photon" latency, is depicted as 88ms, with the network component of the latency being 15ms one way, or 30ms round-trip-time (RTT).

For a network, meeting 30-40 millisecond RTT latency is a tall order. For example, in a particular home, based on current network conditions, may take 5ms over WiFi and 25ms for DOCSIS, but it could take another 20ms to reach the Cloud due to routing over the backbone network, followed by egress to another carrier network or the Cloud gaming content delivery network (CDN), and ultimately, to the gaming server, i.e., the graphics engine – we assume that the gaming server is not in the same market or Metropolitan Statistical Area (MSA) as the home.

In this example, it becomes challenging to provide such interactive experiences from the Cloud. The unique position of CSPs to affect the quality of experience for services such as Cloud gaming by hosting the service from within their network, coupled with the financial incentives of higher ARPU, is driving some CSPs to bring these services to market.

However, Cloud gaming is only the beginning of immersive media delivery. AR and VR are on the horizon and the rumored launch of Apple Glasses promises to bring these services to mainstream adoption. AR and VR are, however, far less latency-tolerant than Cloud gaming. While public documentation proclaims that 100ms "button-to-photon" (system) latency is acceptable for Cloud gaming, the community of practice has concluded that VR will require 20ms "motion-to-photon" (system) latency. As seen from Figure 3, if a graphics engine takes 40ms to produce the next frame, or if the necessary pipeline steps of encoding and decoding take 10ms, then an exacting standard of 20ms "motion-to-photon" would be nearly impossible even if the network RTT latency is negligible.

In reality, there are several application-level techniques to "relax" the 20ms motion-to-photon latency requirement, such as re-projection, asynchronous time-warp, over-rendering, head movement prediction, etc. This has been validated by our empirical R&D. The type of content, "lean back" vs "lean forward", also determines whether this "motion-to-photon" latency is a strict or relaxed requirement. For example, an experience such as "theBlu" [5], in which the user moves relatively slowly and interacts with elements in the content infrequently, is "lean back", and the user will tolerate greater "motion to photon" latency for an acceptable experience. In contrast, VR "lean forward" content such as the popular "Beat Saber" [6], will tolerate less latency.

Overall, VR and AR will be much less tolerant to latency than Cloud gaming. Currently, the network RTT latency and jitter required to serve a particular AR/VR experience may be proprietary knowledge, dependent on the "motion to photon" latency as well as the technology stack of the Application Service Provider that manages rendering, encoding, decoding, etc. We expect that this will become public knowledge as the industry develops and users demand streaming options for delivery (similar to Cloud gaming today). It can be evaluated by ASPs under test conditions and may also be crowd-sourced through session evaluations of users.

In recent years, the Cable industry has embarked on a journey to explore the delivery of new media beyond video, such as gaming, AR/VR and light fields, over the network. For example, in the 2020 SCTE Cable Tec Expo Keynote demonstration, Charter Communications, together with partners, demonstrated a holographic transmission leveraging the "Power of 10G" [7][8][9][10].

Multiple System Operators (MSOs) have also set up the Immersive Digital Experiences Alliance (IDEA) [11] and are working with technology partners to advocate for standards and ecosystem development of immersive media. As MSOs become converged connectivity providers (wired and mobile), it is important that the Cable industry participate in the ecosystem and lead the definition of standards in Fog Computing and Edge Computing [12][13]. The methodology proposed in this paper crystallizes how MSOs can leverage their network to offer Edge Computing as a service to Application Service Providers (traditionally deployed in the Cloud).

3. The Communication Service Provider Network Hierarchy: Architecture and Considerations

When implemented, the methodology described here allows the CSP to provide latency-dependent compute to a subscriber, in concert with an ASP, for current and future interactive content experiences. We elaborate a method where system components interact seamlessly with each other for compute allocation.

We model compute grade and network latency as the most important factors that affect the end-user's quality of experience (QoE). There may be other factors, such as specifications of an end-user device (ex., head mounted display/HMD); however, due to the nature of interactive immersive experiences (ex., VR

gaming or AR that factors in current context like detection of objects in view, SLAM [simultaneous localization and mapping], spatial anchoring of rendered 3D objects, etc., as opposed to watching a 360-video in VR), the intensity of compute and its latency to the end user become the primary determinants of whether the experience can be served at all. If it can be served, then optimizing delivery, i.e., encoding, streaming, buffering, etc., to match the end user device and pose is a secondary problem that must be solved tactically.

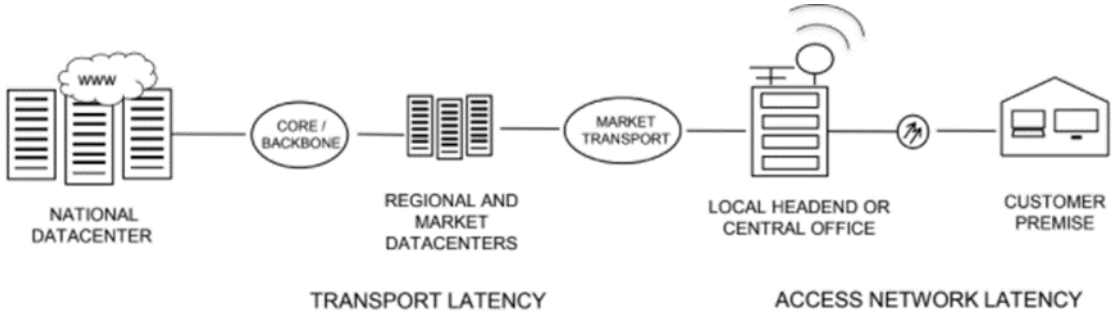


Figure 4- Sample Cable/Telecom Network

Figure 4 depicts the typical network hierarchy of a Cable/Telecom CSP. The access network, i.e., the "last mile" runs between a headend/CMTS (Cable Modem Termination System) to the customer premise for a Cable/broadband network and from a central office/base-station to the customer equipment (ex., mobile phone) for a telecom network. As more latency-sensitive compute is required for immersive media or other applications, it is expected that the central offices and headends may be retrofitted with more compute resources. There is public discourse on CORD (central office rearchitected as a data center) [14] and HERD (headend rearchitected as data center) [15][16] as the logical evolution of these CSP critical facilities.

The local headend/central office, however, is not the lowest level of the network hierarchy. Some of the new cable / fiber deep architectures such as Distributed Access Architecture (DAA) technology [17] [18] create opportunities for putting potential compute "closer" to the consumer along with Hybrid Fiber Coaxial cable (HFC) equipment. When combined with software "virtualization", the downstream migration of DOCSIS functions may also free significant space and power in the headends, for retrofitting general purpose CPU/GPU compute into critical facilities. Similar cell site architecture proposals are also being considered in the 5G technology umbrella. The equipment on the customer premise (CPE, or customer premise equipment) comprises the lowest level in the network hierarchy. It includes cable modems, wireless routers, set top boxes, fixed wireless small cells, and user equipment such as TVs, holographic displays, AR/VR headsets, mobile devices, etc. As embedded compute becomes cheaper and more available, there is an opportunity to provide more capable devices that leverage built-in compute resources like GPUs.

Above the headends and central offices are the market and regional data centers, which may be considered at the same level or at different levels in the network hierarchy, depending on the specific topology of the network. Since these are already data centers, the CSP may augment these with new and more powerful compute like banks of GPUs.

Finally, the CSP network may have a national data center, or the traffic may be exchanged between a carrier network and the CSP network, to terminate in a Cloud location/CDN. This represents the highest level in the network hierarchy. These facilities are typically medium or large colocation data centers, and

providers of various services and applications such as immersive media may upgrade their equipment with adequate compute to serve new experiences.

It is important to note that today's IP media delivery connections are end-to-end, i.e., only have a source and a destination, meaning that, so far, there has been little need to specify a network hierarchy for media delivery. This also derives from the tolerance that current media applications have for network latency.

It is expected that a CSP will have a hierarchy of compute where, starting from the customer premise, each level of the network would typically have greater compute than the lower level, but less compute than the next level. Therefore, at the lowest level in the customer premise, the least amount of compute may be available. The headend/central office will have more compute than the customer premise, but less than the market or regional data centers, and so on. This is because each level of the network is serving more and more customers, requiring a larger serving radius. For example, while a CPE may serve only one home, a Remote-PHY (R-PHY) node [19] may serve 50 homes, a headend may serve 500 homes and a market data center may serve 10,000 homes. This is generally true for connectivity and packet routing today, however, as service providers augment their networks with compute for future applications, this will likely also be true of compute resources.

The other reason for this hierarchy is that it is easier to augment higher levels with compute as they are already data centers, and often have adequate space, power and cooling. Since the demand for compute at any point in time is statistical, adding servers with a bank of GPUs as a consolidated compute asset in a market or regional data center is logistically feasible. By comparison, a CSP may have to arrange for space, power and cooling to retrofit a central office or headend into a data center at the lower level. In a similar vein, a CSP may have to ship upgraded CPE, such as a router or a device to add compute in the customer premise. Therefore, it is relevant that the latency to the compute increases for each higher level in the network hierarchy.

4. Delivering Latency-Sensitive Compute for AR/VR Experiences

In this section, we describe how a CSP network may determine the intensity (or grade) of compute as well as the class of latency needed to serve an immersive AR/VR experience, as well as provide the specified compute to the customer from within its network.

4.1. A Priori Setup, CSP

A. A CSP organizes its compute resources into units and compute grade/intensity. Each unit maps to a self-contained compute system capable of running a class of experiences, such as a Virtual Machine (VM) or a container. Each unit also has a compute grade associated with it, which refers to the quality of its resources, such as CPU cores, RAM, OS, GPU flops and vram. The compute grade may be a simple descriptor (say, a scale of one to ten) or a more complex descriptor that may be more descriptive on individual elements of the compute. Each unit is also associated with a hierarchy level in the network.
B. The CSP sets up a global network compute orchestrator for management of all compute resources, as well as local compute orchestrators in each network compute element from data centers to CPE. A network compute orchestrator negotiates the compute resources on behalf of the subscriber from the network. The negotiation is based on the subscriber's service-level agreement (SLA) with the CSP – higher levels of service authorize higher compute grades for a subscriber, perhaps for longer time periods. Further, a network compute orchestrator may command a local compute orchestrator in a specific data center/headend/central office/CPE to reserve its resources.

C. The network compute orchestrator has real-time visibility into the use of compute resources by receiving responses to queries, success/failure to commands or periodic messages from local compute orchestrators. It organizes this data using an efficient data structure that may be logically equivalent to an Available Compute Resource table wherein each entry represents Compute Resource Label - Compute Grade – Total Units – Available (Free) Units.
D. The CSP maintains a data structure of network latencies required to reach its compute resources from each consumer's home (or a group of consumers' homes), at various levels in the network hierarchy. The CSP may deploy a latency measurement system between several probes at different points in the network. Some latencies may be directly measured and averaged. Other latencies may be deduced by addition or subtraction of aggregated, measured latencies based on knowledge of the topology of the network. Each measurement shall update a previously-averaged value using a weighted approach by allotting a higher weight to the most recently received value. A CPE may also periodically measure its latency to the central office/headend as part of this latency measurement system.

4.2. A Priori Setup, Application/Media Service Provider

A. The ASP maintains a list of compute grade and the (worst case) latency required to serve a particular content/experience to a subscriber.

4.3. Connection Setup

A. Subscriber requests an experience/content to be delivered to the premise on any of their devices. This request is routed to a network compute orchestrator.
B. The network compute orchestrator queries the ASP for the compute grade and the latency required to serve the requested experience.
C. The network compute orchestrator calculates the latencies from the subscriber to each of its compute units in an efficient data structure equivalent to a Network Hierarchy – Latency table.
D. It then finds the lowest level of the network that has a latency equal to or better than the worst-case latency and a compute grade equal to or better than the compute grade for the requested experience using its Network Hierarchy – Latency (built from the point of view of the subscriber) and the Available Compute Resource (global compute resource information) tables, respectively. If the compute resource is not available, then the network compute orchestrator attempts to find the compute unit at the next higher level in the network hierarchy. This continues until either the compute unit is found, or the worst-case latency threshold is exceeded. If the compute unit is not found, the network compute orchestrator informs the subscriber that their requested experience cannot be served at this time.
E. If a compute resource is available, the network compute orchestrator reserves it by issuing a command to the specific local compute orchestrator. It receives a token if the request is successful. It passes this token to the ASP and updates its Available Compute Resource table.
F. If the Application Service Provider receives a token successfully, it proceeds to use the compute unit.

Table 1- Network Orchestrator "Available Compute Resource" Table

Compute Resource Label	Compute Grade	Total Units	Available (free) Units
…..	…..	…..	…..
CPE – Router Subscriber S	3	1	1
CPE – Set Top Box Subscriber S	5	2	1
Headend / CMTS Domain D	9	10	4
Regional Data Center Region R	10	70	33
…..	…..	…..	…..

Table 1 illustrates a network orchestrator Available Compute Resource table. In this example, we show compute grade (intensity) as a simple descriptor on a scale of one to ten. These descriptors can be mapped internally to specific attributes like GPU flops, vram, etc. The table shows entries for one subscriber, S. If the network makes compute available at the CPE level, then there would be entries for each subscriber. A tree data structure may be used for efficient storage and traversal, where the leaf nodes represent CPE at the subscriber level.

Table 2 - Application Service Provider Content Table

Content	Compute Grade / Intensity (descriptor 1-10)	RTT Latency, Customer to Compute (milliseconds)
…..	……	…..
Content-X	8	30
Content-Y	6	10
…..	…..	…..

Table 2 illustrates an ASP content table. In this example, Content-X may be a “lean-back” experience where the user is immersed in a VR session but does not require fast-paced interaction with their environment. On the other hand, Content-Y may be a first-person shooter game in VR, where the targets are moving with high velocity. Thus, the RTT latency requirement for these experiences is significantly different.

Table 3- Network Orchestrator “Network Hierarchy – Latency” Table

Hierarchy	Compute Resource Label	RTT Latency, Customer to Level/Compute (milliseconds)
1	CPE – Router Subscriber S	3
2	CPE – Set Top Box Subscriber S	3
3	Headend / CMTS Domain D	9
4	Regional Data Center Region R	23
5	National “Cloud” Data Center NOC	55

Table 3 illustrates a “Network Hierarchy – Latency” table built by the network orchestrator for serving Subscriber S.

Consider that subscriber S wishes to experience Content-Y, which is both latency and compute sensitive, as observed from Table 2. From Table 3, it is evident that the content must be served from a headend/CMTS or a lower network hierarchy, due to latency constraints. Now, from Table 1, we observe that the CPE in Subscriber S's home does not have the compute intensity to drive this experience. In this case, the compute unit must be allocated at the headend/CMTS. In our example, four units of compute resources with grade nine are available, as seen is Table 1(that is greater than the requirement of six for Content-Y, Table 2). Therefore, one of these units may be allocated to delivering Content-Y in Subscriber S's premise.

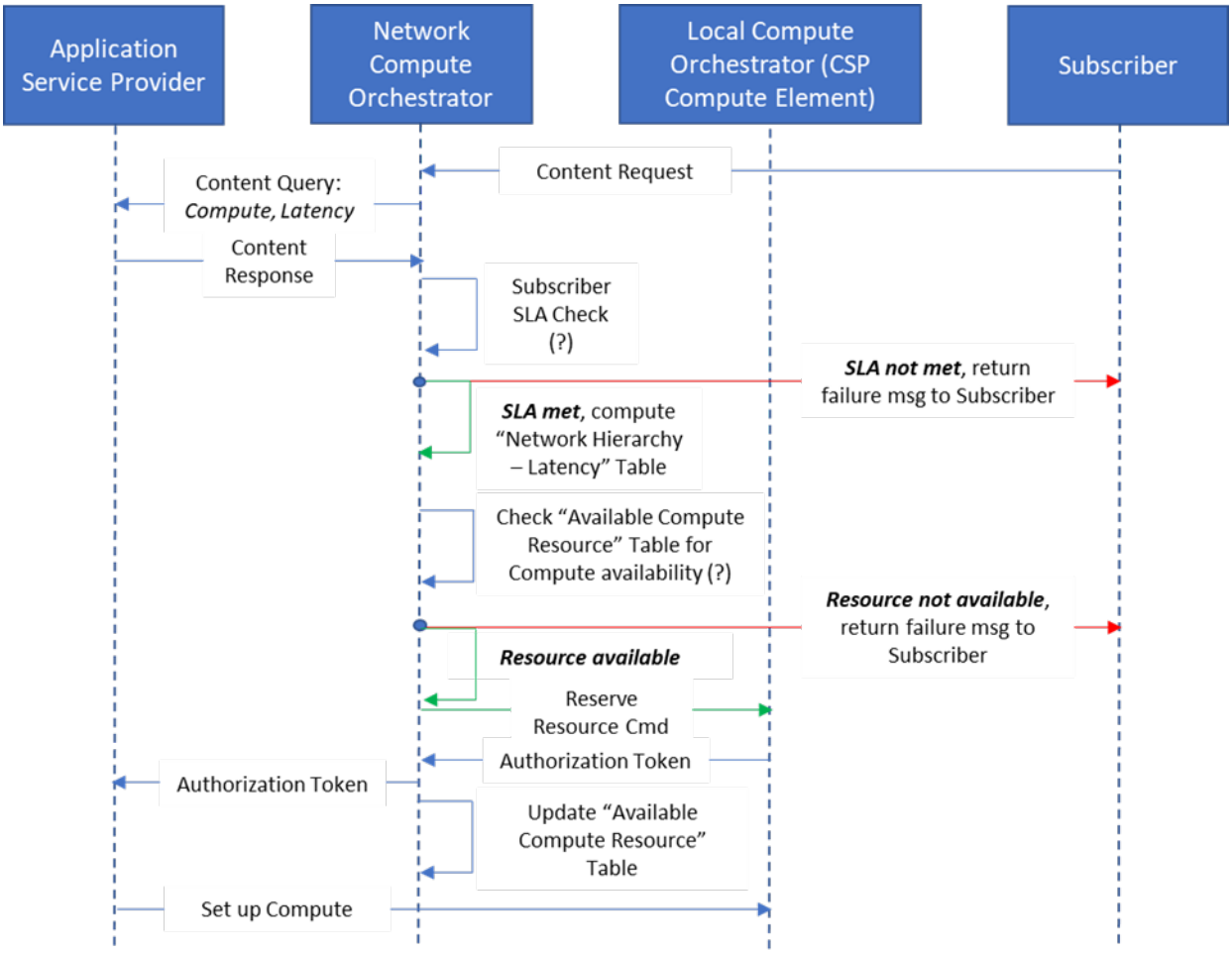


Figure 5 - Sample Session Setup

Figure 5 shows a sample session setup where a subscriber requests an AR/VR experience and the CSP allocates the compute resource for that content from within its network. The subscriber may request the experience either from the ASP that then contacts the CSP for low-latency compute, or directly from the CSP that may forward the request to the ASP. This may depend on the business relationship between the subscriber, the ASP and the CSP. The initial query to the ASP is used to determine the latency and compute intensity requirements. The CSP, after receiving these requirements and checking the subscriber SLA, proceeds to find the compute unit within its network. From the subscriber-centric "Network Hierarchy – Latency" table, the CSP determines the (one or more) levels in the network hierarchy from which the request may be served based on latency. Thereafter, a determination of the critical facility from which the request shall be served is made by the CSP's Network Orchestrator based on latency and

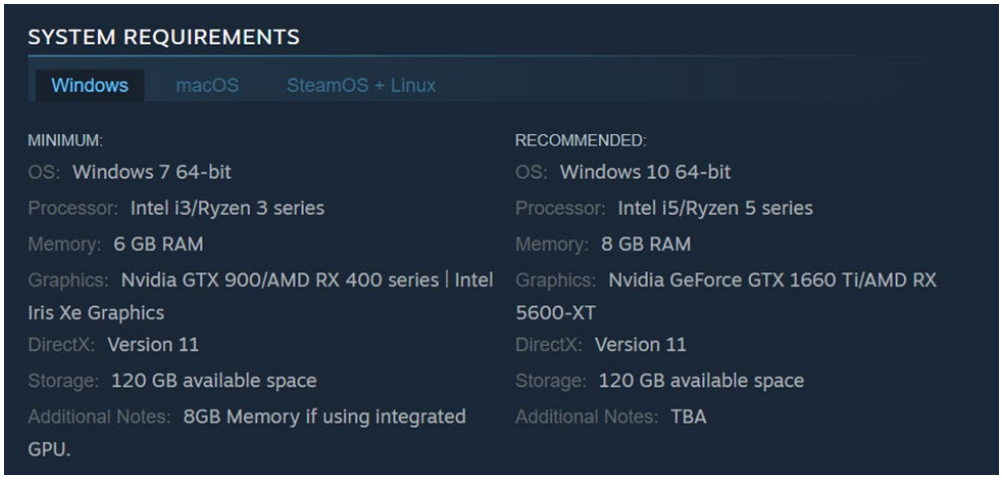
compute intensity. The Network Orchestrator updates its "Available Compute Resource" table accordingly. Once the compute is allocated, the ASP loads the application. It is worth noting that while the subscriber initiates the request, the delivery of content occurs when the ASP and the CSP work together.

5. Discussion of Compute Grade and Latency

In this paper, we used a simple descriptor on a scale of one to ten to denote Compute Grade or Compute Intensity. In reality, gaming/VR storefronts use detailed Minimum and Recommended Compute configurations to specify the grade. Figure 6 illustrates three examples from Steam [20], one of the largest distribution storefronts for gaming and VR:

Grand Theft Auto

Total War: WARHAMMER III



Car Mechanic Simulator VR

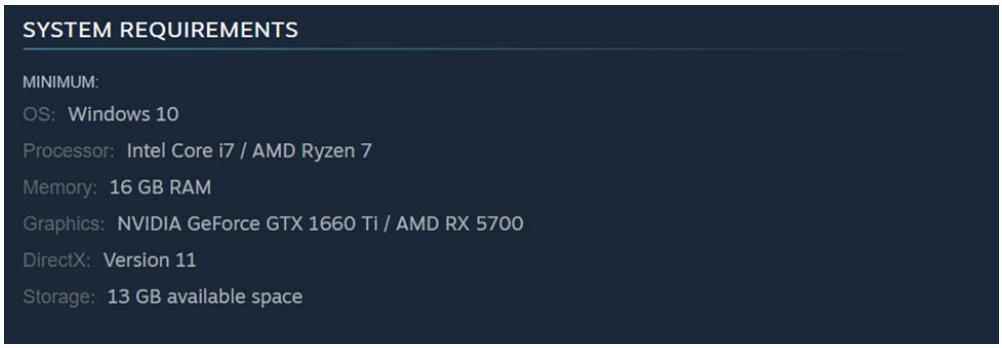


Figure 6 - Example System Specifications for Delivering Experiences

ASPs that provide Cloud compute, on the other hand, may use a single system configuration or a tiered system where a higher monthly rate entitles the subscriber to better system (compute) specifications.

Currently, content stores and ASPs may be separate or "all-in-one". In this example, Steam is the content store that offers downloadable games, while an ASP would be a Cloud gaming/VR provider that runs game executables in its own data centers or in the public Cloud. In this paper, we treat ASPs as an "all-in-one", wherein they have agreements with content publishers and storefronts to present content and, in concert with the CSP, compute for delivering the experience.

The immersive community has been exploring "split rendering", wherein the compute required to serve immersive experiences is shared between a client device and a Cloud compute unit. Qualcomm's "Boundless XR" [21] concept provides details on their implementation of split rendering.

The "split rendering" paradigm, an active area of R&D, divides compute between a client and the Cloud such that highly latency-sensitive rendering is delivered from the client (if the Compute Grade is available) while less latency-sensitive rendering is delivered from the Cloud Compute unit, and these renders are then composited to deliver a seamless experience to the user. Our methodology is easily modified to specify compute for split rendering. To enable this, the client device must send its available compute specifications, and the ASP, once aware of the locally-available compute, may provide a modified Compute Grade entry from its Content Table.

Finally, it is important to note that many latency measurement systems measure latency and jitter by sending a train of packets to a network device/Point of Presence (PoP), and monitoring RTT for each individual packet as well as the inter-packet delay when they are returned to sender. If a jitter value is available, the service provider may also make that a part of the Network Hierarchy – Latency table. If the jitter value from the subscriber to an element in the table exceeds a threshold, the network compute orchestrator shall reject that element as a potential render/compute candidate even if the latency is below the threshold required by the content/experience.

6. Conclusion

In this paper, we explored how a CSP such as a Cable and broadband service provider may work in conjunction with an ASP to commission on-demand compute for delivering immersive experiences to customers. We explained how it will be possible to serve AR/VR content from within the CSP network even if the "motion to photon" latency is published to be as low as 20ms. This has been validated through empirical observation.

Abbreviations

AR	Augmented reality
VR	Virtual reality
HMD	Head-mounted Display
ms	Milliseconds
MSA	Metropolitan Statistical Area
CSP	Communications Service Provider
ASP	Application Service Provider
RTT	Round Trip Time
CMTS	Cable Modem Termination System
CORD	Central Office Rearchitected as a Data center
HERD	HeadEnd Rearchitected as a Data center
CPE	Customer Premise Equipment
DAA	Distributed Access Architecture
R-PHY	Remote-PHY
CDN	Content Delivery Network
SLA	Service-Level Agreement
MSO	Multiple System Operator
SCTE	Society of Cable Telecommunications Engineers
PoP	Point of Prescence

Bibliography & References

[1] What is Motion to Photon Latency, http://www.chioka.in/what-is-motion-to-photon-latency/

[2] The Technology Behind a Low Latency Cloud Gaming Service, Parsec blog, https://parsec.app/blog/description-of-parsec-technology-b2738dcc3842

[3] The Road to Success: How we are defeating Latency, Shadow Blog, https://shadow.tech/en-gb/blog/news/roadmap-cloud-gaming-without-latency

[4] Game on! How broadband providers can monetize ultra-low latency services for gamers, Nokia Blog by Gino Dion, https://www.nokia.com/blog/game-on-how-broadband-providers-can-monetize-ultra-low-latency-services-for-gamers/

[5] theBlu on Steam, https://store.steampowered.com/app/451520/theBlu/

[6] Beat Saber – VR Rhythm Game, https://beatsaber.com/

[7] Power of 10G, SCTE Cable Tec Expo 2020 Keynote Demonstration, https://www.youtube.com/watch?v=I79WMpLrrGU

[8] Charter, partners stream 10G holographic demo at virtual Cable-Tec Expo, Broadband Technology Report, https://www.broadbandtechreport.com/video/article/14185182/charter-partners-stream-10g-holographic-demo-at-virtual-cabletec-expo

[9] The Cable Network, Immersive Experiences and Lightfields, Ip, A. and Lal, D., Broadband Library, https://broadbandlibrary.com/the-cable-network-immersive-experiences-and-lightfields/

[10] How we streamed a Light-field over a 10G Network, IDEA Workshop at 2020 SMPTE Annual Conference, https://www.youtube.com/watch?v=cAg0A9gld5c&t=3s

[11] Immersive Digital Experiences Alliance website, https://www.immersivealliance.org/

[12] Edge computing: current trends, research challenges and future directions, Carvalho, G., Cabral, B., Pereira, V. et al, Computing 103, 993–1023 (2021), https://doi.org/10.1007/s00607-020-00896-5

[13] Fog Computing as an Enabler for Immersive Media: Service Scenarios and Research Opportunities, You, D. et al., IEEE Access, Volume 7, 2019, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8716694

[14] Cabling considerations for CORD networks, https://www.cablinginstall.com/data-center/article/14068510/cabling-considerations-for-cord-networks

[15] Have you Heard About HERD?, Chris Bastian, https://www.cablefax.com/cablefax_viewpoint/have-you-heard-about-herd

[16] HERD for the Gigabit Era, Broadband Technology Report, https://www.broadbandtechreport.com/docsis/article/16449156/herd-for-the-gigabit-era

[17] Evolving to Distributed Access Architectures, Chris Bastian, https://www.cablefax.com/technology/evolving-to-distributed-access-architectures

[18] Distributed Access Architecture Is Now Widely Distributed – And Delivering On Its Promise, Howald, R., Eichenlaub, F., Peck, T., Bonen, A., SCTE Fall Technical Forum 2021, https://www.nctatechnicalpapers.com/Paper/2021/2021-distributed-access-architecture-is-now-widely-distributed

[19] Remote PHY Distributed Access Architecture, Steven Harris, Broadband Library, https://broadbandlibrary.com/remote-phy-distributed-access-architecture/

[20] Steam website, https://store.steampowered.com/

[21] Boundless photorealistic mobile XR over 5G, https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/more_immersive_xr_through_split-rendering_-_web.pdf

[22] QoE-aware dynamic service composition for immersive media-oriented services, Park, J., Lee, H., Yi, D., Kim, J. (2022), https://www.researchgate.net/figure/QoE-aware-dynamic-composition-framework-for-immersive-media-oriented-services_fig2_228930025

[23] Implementation of a Media Aware Network Element for Content Aware Networks, Niculescu, D., Stanciu, M., Vochin, M., Borcoci, E., Zotos, N. (2011). CTRQ 2011 - 4th International Conference on Communication Theory, Reliability, and Quality of Service, https://www.researchgate.net/publication/228948173_Implementation_of_a_Media_Aware_Network_Element_for_Content_Aware_Networks

