



Interactive Applications Powered by Video at the Cloud Edge

The Architecture, Technology and TCO of Interactive Applications at The Edge of the 5G Network

Sponsored by NETINT

Simon Solotko
Senior Media/XR Analyst, TIRIAS Research
simon@tiriasresearch.com

October 2020

Contents

Contents	1
Cloud Edge Architectures for Emerging Applications	6
Network Topology	7
Network Service Providers:	8
Local ISPs and Data Centers	8
Cloud Service Providers	8
Edge Video Requirements	8
High Visual Quality	8
Low Latency Video	9
Economical Scalability	10
Compute Density Creates Viable Edge System Architectures	11
Comparing ASIC and CPU Encoding	11
Edge Encoding with NETINT Dedicated Encoding Processors	11
The Total Operating Cost of Streaming	13
Total Operating Cost Calculation	14
Application Server Operating Costs	15
Total Operating Costs: Cloud Mobile Gaming and Cloud Remote Applications with NETINT Video Encoders	16
Mobile App & Game Streaming	16
Mainstream & High-End Gaming Streaming	17
The Future	17

Overview

This paper describes the future of video streaming and the architecture of emerging cloud edge services utilizing high performance video infrastructures. It examines the technologies and architectures of streaming application services and describes the economics and viability of operating these services. Divided into three main sections, the first covers the drivers for growth in video and the TIRIAS Research streaming forecast; the second discusses architectural considerations for forward looking video services; and the third discusses the operating costs and presents a Total Cost of Ownership (TCO) model for edge based video services and compares technologies. This expands on TIRIAS Research's previous white paper "The Emergence of Cloud Mobile Gaming" which explicates emerging head mounted display and mobile application delivery powered by edge video ASIC processing and the emerging Arm servers.

Video & Interactive Media at the Cloud Edge

Few topics in technology can match the exposure of edge computing and its bold vision for the future. Placing compute resources closer to users and sensors achieves lower latency, allowing new, cloud based interactive services to run with the responsiveness of local applications. In principle, consumer applications can span interactive social video, game streaming, virtual reality, and augmented reality powered by servers that live at the edge. Users engage apps just like they do today, but their input is sent to an application service in the cloud which runs remotely. In the case of interactive applications – like games and productivity apps - video compression handles the job of packing and unpacking the visual output of an application from the edge to the user's screen. Since all the work is done in servers at the edge, client devices can be cheap, low power, wearable, and mobile. New, service-oriented business models, and new cloud interactive applications can flourish.

However, edge deployment lacks the economies of scale of the "hyperscale" data centers that house today's large-scale services. The reality is that cell tower base stations are generally small and expensive, lacking extra space for application servers that can support thousands of concurrent users. So cloud edge services have the dual challenge of demanding workloads (more servers) and more expensive space and operating costs. Since base stations are often run by a single operator, and there are many of operators, scaling services would require scaling to diverse service providers and many locations.

Reducing the cost to deliver low latency is critical to emerging interactive edge services. In this paper we suggest that costs can be managed by placing application servers and video encoding systems together in regional/metropolitan data centers so that the network need only hop from the regional data center to local base stations or network service provider central offices. We further suggest that latency can be significantly reduced, and video quality increased by employing specialized chips – Applications-Specific Integrated Circuits (ASICS) - for video

encoding. Using these specialized video encoders at regional points of presence can achieve the latency required for cell-tower placement without the cost.

Advanced video CODECs must be combined with low-latency processing to deliver visual fidelity for interactive applications and collaborative social interaction. Dedicated video processing including ASIC video encoders and core logic, can radically reduce the hardware required for video processing. And finally, placing compute at regional points of presence creates a balance between the vision of edge computing and the economics of operating consumer services at scale.

The Opportunity: Powering the Exponential Growth of Video

Video technology has demonstrated remarkable versatility, with the potential to transmit real time, interactive experiences. Video can be used to encode social, entertainment, information, or software-based experience and deliver them to ubiquitous devices like PCs, TVs and smartphones. Achieving more fluid collaborative and interactive experiences requires higher resolutions and framerates at lower latency. For services aspiring to scale, challenges have included cost, quality of service, visual quality, and motion-to-photon latency.

New cloud architectures promise to resolve user experience gaps. These architectures distribute compute to the network edge, lowering latency, decreasing backhaul traffic, and enabling new forms of data/sensor/display edge processing. Unfortunately, the cost of deployment and operations increases sharply at the edge of the network. The combination of smaller scale, expensive space and enclosures, remote management and maintenance, multi-site security, dedicated power system, and other factors make base stations and their server capacity expensive relative to large data centers.

To make edge computing economically viable requires breakthroughs in computing performance and hardware density. Simply getting to scale with social streaming services is extremely expensive; increasing the compute workload by running remote applications can drive insurmountable costs. Making edge computing affordable requires realistic network topologies that balance the economics of scale with the need to drive low latency and high visual quality for interactive service delivery.

A new generation of dedicated video transcoders can reduce the compute requirement at the same quality by 10X and performance per watt by 20X improving the viability of cloud edge video services. NETINT, a pioneer in ASIC based encoding has introduced a family of dedicated video processors which combine video encoding, solid state storage, and machine learning. These processors radically reduce the server footprint required for interactive entertainment like gaming and mixed reality, and can scale to economically deliver any native desktop, mobile, or head-mounted display application from the cloud edge.




	ASIC	BEST GPU	SOFTWARE
Impact Factor With 1000 Transcode Channels @ 1080P30	NETINT G4	NVIDIA T4	INTEL SVT
 TCO Annual TCO / 1000 Concurrent Streams	\$52,403	\$110,606	\$580,535
 Environmental Impact Reduce Emissions by ~20X vs. Software	11.7 CO² Metric Tons	41.1 CO² Metric Tons	217 CO² Metric Tons
 Server Density Reduce Footprint by 10X vs. Software	12.5 NETINT Codensity Servers 1000 Streams	25 NVIDIA T4 Servers 1000 Streams	125 Intel Servers 1000 Streams

Figure 1: Summary of this paper's analysis of TCO, environmental impact, and density of today's best in class CPU, GPU, and ASIC encoders. Source: TIRIAS Research.

Forty Million Cloud Encoded HD Streams By 2025

The legacy of internet video was video on demand – stored video delivered as entertainment or instructional content. This content is uploaded, transcoded for multiple levels of quality and resolution, and distributed over the internet. Live streams require real time transcoding for rapid distribution to users. These streams can include a full spectrum of interactivity, from mobile social streams to interactive applications. Cloud based applications can include games and apps that originated on PCs or smartphone platforms, running in the cloud, encoded as video, and experienced like native applications by remote users.

TIRIAS Research estimates that high-quality, high-definition cloud video streams will expand from 4 million encoded concurrent streams today to over 40 million by 2025. It is comprised of the requirement for HD and Ultra HD resolution cloud encoded streams for consumer entertainment, social media, games and apps. Demand for live streaming will be driven by increasing remote work and education, the popularity of social video, and emerging game & application streaming services. Compute and infrastructure will be driven by a drive to higher resolution, framerates, lower latency, and design for scale. These forces are all powerful drivers individually but combined they create the opportunity for 10X growth over 5 years for concurrent streams measured in units of 1080p 30 frame per second (FPS) streams (1080p30).

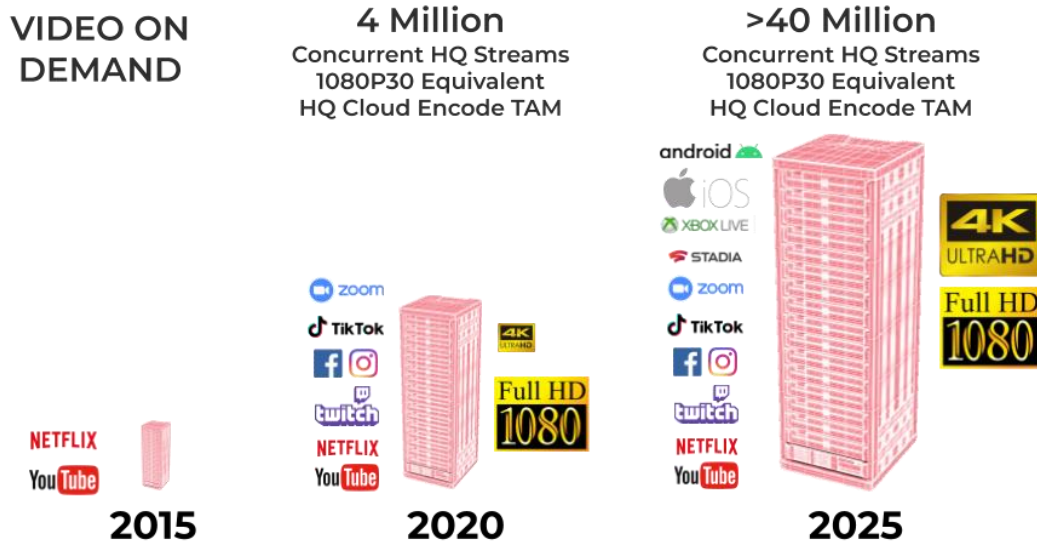


Figure 2: Growth in Video Streams measured in units of concurrent 1080p30 channels through 2025 include demand for high resolution, high quality content, stay/work at home, social collaboration, mobile social media, cloud game and app streaming. Source: TIRIAS Research

Strong growth in social video growth has been powered by 4G wireless networks and fueled by user generated content. Emerging application streaming services require alignment of content providers, service providers, processing technology, and low latency network infrastructure. TIRIAS Research forecasts steady growth in interactive application streaming, with 4 million concurrent streams by the end of 2025 and continued strong growth in the years that follow.



Video Resolution: Flagship smartphones and 4K TVs and PCs are ready for 4K video. Services including Facebook Live, Twitch and YouTube were early 4K streaming adopters. Social mobile video apps such as TikTok, and Instagram are improving stream quality as those services evolve from teen interest to a global entertainment phenomenon.



Virtual Presence: The recent switch to communal, virtual spaces for education and social life will have a lasting impact on user habits and technology infrastructure. Cloud encoding is a force multiplier for video virtual presence services, aiding in video quality and bandwidth optimization. It also drives more complex use cases including multi-stream scenarios, complex visualizations & overlays, and interactive content.



Social Video: Today social applications are driving growth, as users opt to collaborate and socially engage in real time. Successful services operate at a scale where even simple features can have significant costs. Video-centric social media platforms must conserve processing, bandwidth, and storage while providing increasing levels of service and video quality.



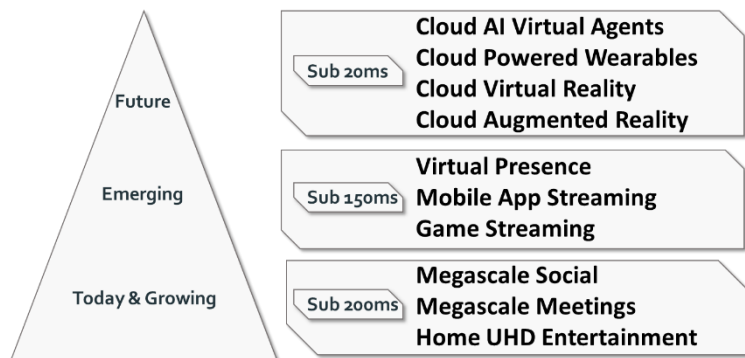
Game Streaming: Game streaming services are now available, but content libraries remain modest and services remain in early stages of development. The expense of deploying cloud game servers, infrastructure, and consumer-facing services has held back scale-out and competition. Nvidia and Google have created new cloud-centric ecosystems, while Microsoft Xbox and Sony PlayStation look to their loyal users and a well cultivated developer ecosystem. Latency and video bitrates remain major complaints for today’s gamers. Improvements in cloud encoding have the potential to significantly decrease latency (as high as 300ms and ideally significantly below 100ms) and lower bitrates for 4K (over 35mbps on VP9).



Mobile & Desktop App Streaming: Cloud mobile streaming employs cloud servers to run mobile apps remotely and then encode graphics to a video stream for transmission over wireless networks to Smartphones. Emerging 5G networks promise to deliver these streams reliably and with low latency. Running these services economically requires a new, virtualized mobile-application server infrastructure, one which employs ARM-based servers, high-performance graphics accelerators, and low-latency video encoding to stream games over the internet. Users continue to criticize the video quality and high latency of these services, with 4K latency creating challenges for players of competitive and fast-action games. Speeding adoption is the use of game streaming by established platforms like Xbox and PlayStation to provide fast game start and increase the value proposition for their monthly social gaming subscriptions.

Cloud Edge Architectures for Emerging Applications

Emerging services require technology that can scale economically to any number of users. Cloud streaming is extremely demanding, presenting a combination of technical challenges. Unlike delivering complex web pages where content can be sourced from many locations, streamed video can be aggregated in the cloud and packaged for rapid delivery at the cloud edge. High video quality and low latency must be achieved together to provide users with a sense that they are experiencing a local, native application.



Source: TIRIAS Research

Figure 3: Latency requirements for emerging cloud streaming applications become increasingly viable as low latency targets can be achieved. Source: TIRIAS Research.

The instantaneous cloud describes a new stage in computing when cloud technologies can evolve to deliver native experiences to any smartphone, PC, or XR display. Delivering a sense of local

application presence, or immersive virtual reality (VR) total sensory presence, requires low latency and high visual quality to ensure users have parity or better in their cloud-based application experience. The cloud can provide more general compute and 3D graphics performance than smartphones or low-performance PC clients. However, driving the visual experience with low latency is a relatively expensive and challenging problem to solve from the cloud. Moving these experiences closer to users can lower latency by avoiding long runs, network hops, and network congestion.

Network Topology

Making edge computing a reality requires servers to be placed near users, within regional points of presence or cell tower base stations, accessing client devices over fast, low latency networks including fiber to the home and 5G. Traditionally servers in the cloud, in large centralized data centers delivered applications with relatively high latency, and these high latencies remain are easily measured today. Today’s real world local networks (~100 miles) can deliver just under 20 ms latency with ideal conditions and latencies of 40 ms to 80 ms are not uncommon for domestic pings. Target network latencies for 5G are below 10 ms nationwide (US) and below 5 ms for regional data centers.

The deployment of 5G infrastructure and placing servers in a regional data center within a fiber run – in the same city – as the central offices or base stations of network service providers can reduce the network latency significantly, virtually identical to base station placement of servers without the associated cost. Partnerships that improve connectivity and network performance between regional data centers and internet service providers are critical to delivering low latency edge applications. Network service providers, local data centers, and cloud service providers must coordinate to reduce latency.

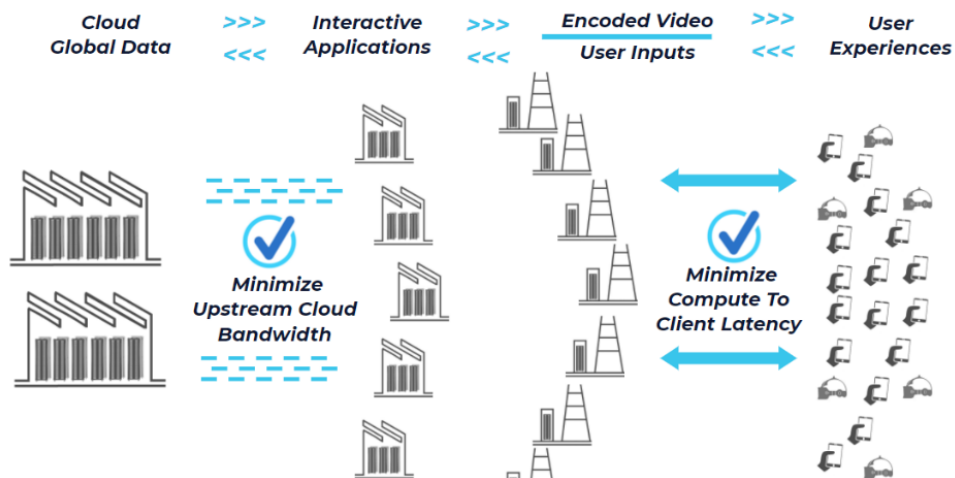


Figure 4: An objective of the instantaneous cloud is to deliver interactive cloud applications while enhancing rather than diminishing the immediacy of local execution. This requires low latency afforded by placing compute near users – at the cloud edge. Source: Tirias Research.

Network Service Providers: Network service providers – which in the US include AT&T, Verizon, T-Mobile, Comcast, Charter Spectrum, and Lumen CenturyLink, - seeking to avoid the expense of base station deployments can employ network central offices for edge servers. Often hosting legacy services and infrastructure, these central offices and points of presence must evolve to support large-scale, high-performance edge-based applications services. Network service providers are in a perfect storm of soaring demand and pressure to upgrade. The deployment of 5G, COVID-19, and high network utilization are putting pressure on resources to deploy additional infrastructure like edge servers and data centers. Solving the broad spectrum of network use cases may distract from edge computing collaborations.

Local ISPs and Data Centers ISPs (Internet service providers) and data centers can create improved connectivity to mobile and home network base stations, lowering the latency of these connections by deploying fiber and tightening network relationships. Agnostic to both network providers and cloud service providers, they are well situated to become a deployment hub for cloud edge servers.

Cloud Service Providers Cloud application providers with a high level of vertical integration, such as Amazon, Google and Microsoft, already have low latency Content Distribution Networks (CDNs) making them ideally situated to provide cloud services at low-latency all the way to the wired/wireless network service providers. Amazon's low latency CDNs and the Twitch service can be utilized to provide edge video services and cloud gaming. Google has created a local presence and run fiber in many cities giving them the opportunity to optimize for latency in their data centers. Microsoft has launched game streaming to augment Xbox subscriptions and provide those subscriptions to any client platform even Android and iOS.

Edge Video Requirements

The emerging cloud streaming experiences poised for growth have common technical requirements. First, they require visuals of high quality and resolution. Second, they require low latency, so that user input can be rapidly incorporated into the visual output. Third, they require scalability such that users can access these services reliably and service providers can offer them economically.

High Visual Quality

Users have become accustomed to crisp, high-quality images on their smartphones and HD screens. Game graphics are particularly sensitive to visual artifacts from compression and feature fast movement diminishing frame to frame compression. Achieving better latency has required sacrificing visual quality and/or bitrate, and the step function improvements offered by emerging video codecs including AV1 come with commensurate increases in latency and compute requirements. Today, game streaming at 1080p utilizing H.265 or VP9 video encoding will require about 3GB/hour or 7Mb/s with peaks over 15 Mb/s. In the future, cutting edge AAA

games threaten to make server hardware obsolete quickly. Running new titles will require switching to more modern graphics hardware, and user expectations shaped by native PC gameplay will require 3D rendering, resolution, and video encoding quality to be at parity with increasingly capable native PC and game console gameplay.

Low Latency Video

The raw latency of video encoding is a function of the complexity of the target video CODEC and the performance of the underlying processor. Today's ASICs are the lowest latency encoders available with 4ms (720P) to 8ms (1080p) using H.265, a complex and computationally expensive CODEC. If this is combined with an ideal network latency of 5ms for local loops, there is plentiful latency budget remaining for complex interactive applications. Game streaming to home and mobile devices has become a beachhead for consumer interactive application streaming, and the network is proving critical to the user experience.

A wide range of low latency use cases is driving the emergence of streaming architectures and video processing technologies. Games that require fast response, quick pointing precision, or easily noticed visual feedback suffer when latencies begin to pass several frames, which at 60FPS is 16.6 ms between frames. For emerging augmented reality (AR) and immersive VR, tracking user movements and translating camera pose into a real time visual experience requires sub-frame motion to photon latency. Movements and head/pose tracking must be incorporated as soon as possible, ideally in the next frame of visual output. This next frame requirement creates the often cited "20 ms motion to photon" requirement for immersive VR – 20 ms being the time required to achieve worst case three frames latency (miss only two frames) at 90 frames per second. In virtual reality, the term presence describes the user's sensation that they are in a particular space as if it were reality. To deliver presence in edge-based VR, every step in the computing and networking architecture must work together to achieve sub 20 ms latency with high-quality, high-resolution, high-framerate visuals.

Variables that normally work against one another must be simultaneously tamed– visual quality vs. bitrate, or latency vs. visual quality can no longer be opposing forces. Looking forward, the demand for lower latency continues far into the future – with everything from immersive VR to cloud-powered virtual agents requiring fast response to user input.

The deployment of 5G front and back ends and utilization of near edge data centers is critical. From a server perspective, dynamic resource allocation should allow better scalability in achieving parity with native experiences, but the high-quality displays in today's smartphones create pressure on both resolution and visual quality. Achieving high quality with low latency remains a gap that high speed, ASIC encoding targets to improve the experience on mobile devices and complementing low latency 5G networks.

Economical Scalability

To achieve the scale and economics required, video processing will need to scale in the cloud and cloud edge – by which we mean everything from city points of presence to cell tower base stations - radically reduces the space and power available for processing. Processing inefficiencies once tolerable now need to be eliminated.

The architecture of systems employing edge processing reduces latency by placing compute closer to users and reduces upstream bandwidth by processing application data and video minimizing backhaul traffic to central data centers. The technologies linking the data center, regional point of presence, wireless or wired network base stations, and users must be tightly integrated. This tight integration requires service providers to work together with application service providers to ensure quality of service for user applications.

Emerging Edge Technologies Driving New User Experiences

Machine learning, mobile infrastructure, and high-density video processing will work together in emerging video centric solutions. The technologies share a common thread of increasing capability and compute density



Machine Learning Processors: Inference at the edge of the network has the potential to create transformative user experiences and automation. Working in tandem with video transcoders and application servers, machine learning can categorize content, monitor and optimize service performance, apply video specific effects such as upscaling or ML based augmented reality, or predict user action to decrease perceived latency.



High-Density Mobile Servers: Arm has enabled the processor ecosystem by developing scalable designs for massively multi-core server processors and enabling a software and core logic ecosystem. These servers simplify the virtualization of mobile operating systems allowing multiple user instances to run on a single chip. Arm-based servers significantly decreases the total cost of operating mobile streaming services with the opportunity to host the massive mobile app code and user base.



High-Density, Low-Latency Encoders: Encoding at the network edge requires high-density and low power to support thousands of concurrent service streams economically in tandem with application processors. ASIC encoders can reduce server footprints from 125 1RU servers with software-based encoding to 12.5 servers per 1000 1080p30 channels. This dramatic improvement places video processing within a viable footprint and operating cost for edge compute applications.

Compute Density Creates Viable Edge System Architectures

Moving video processing to the edge requires a massive improvement in computational efficiency. Locating compute servers at the cloud edge implicates cell towers and local points of presence, removing distance as a factor in network latency. The cloud edge in the broadest sense can describe small data centers operating at local points of presence or within cell tower base stations. Base station deployments are expensive and physical space is extremely limited, creating formidable hurdles for consumer and commercial services. The placement of cloud servers in regional points of presence is the most viable model allowing a major metropolitan area to be serviced by a single data center. The opportunity to tune the server environment and provide best in class systems architecture, networking, and operational efficiency allows high availability with a latency optimized solution.

Comparing ASIC and CPU Encoding

Today, transcoding 1000 concurrent HD video streams using CPU based software encoding at high (but not broadcast) quality requires over 125 1RU servers with a total operating cost of over half a million dollars a year in a standard data center. For services to provide nearly persistent video a radical reduction in cost and footprint is needed. That's where heterogeneous computing and specialized video transcoding ASICs come in. Specialized cloud video encoders – dedicated ASICs that transcode video at high quality with low latency – are extremely efficient and reduce the computation load associated with video processing by almost 10X and reduce the power requirement by about 20X. With NETINT ASICs, only 12.5 1RU servers are required per 1000 concurrent 1080p30 streams. These high-density encoding servers can easily co-locate with application processors in regional point of presence or even cell base stations. The advantage in power consumption has a proportional advantage in environmental impact with lower carbon emissions.

Edge Encoding with NETINT Dedicated Encoding Processors

Encoding video streams with low-latency and high-fidelity graphics is critical to meeting the expectations of users accustomed to high-performance gameplay on their smartphones. High-density ASICs can be deployed in server and system architectures that deliver end to end low latency and economic operation for thousands of concurrent users. NETINT high-speed encoders transform game graphics into H.264/AVC or H.265/HEVC video streams. The encoders run on either U.2 or PCIe modules, enabling up to ten encoders to be installed in a single 1U server, with hundreds of encoders on a single rack. Employing virtualization, a single 1U server with a 64-core processor, multiple GPUs, and ASIC based encoders can be configured to run many concurrent user sessions, and a single rack can run hundreds, or even thousands, of streamed mobile apps and games. NETINT encoders have very low latency and can process typical game graphics in about 8 ms , minimizing the motion to photon latency.

Power Consumption & Environmental Considerations

Data center and large-scale service operators face the challenge of decreasing energy consumption to lower operating costs and emissions. Operating costs are driven by power and the hardware upgrade cycle. Environmental impact is hard to manage. As users embrace new services, computational workloads and usage increase, placing pressure on data centers to increase capacity, driving up power and emissions.

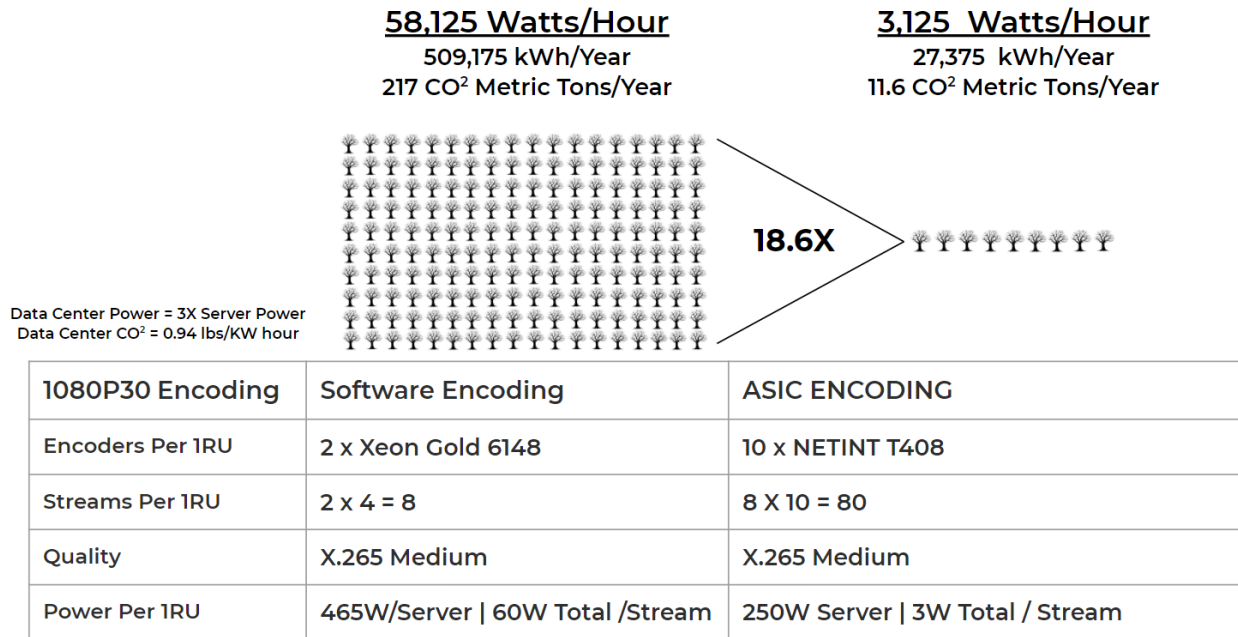


Figure 5: Using ASICs instead of CPU based software encoding reduces emissions from 8,680,000 to 464,000 CO2 metric tons/year at 40,000,000 streams projected by 2025. This is an 18.6X reduction. Source: TIRIAS Research

Rarely does an opportunity arrive to create an order of magnitude reduction in an easily identified workload. Improving processing efficiency in video utilizing heterogeneous computing is such an opportunity. Today, GPUs can increase efficiency over CPUs, and programable logic and ASICs provide the most efficient encoding. ASIC based encoding is over 18X more energy efficient than Intel CPUs using the SVT encoder with the H.265 Codec. NETINT ASICs can achieve x265 medium with 8 streams/encoder or 80 streams/server, whereas x86 servers achieve only 8 streams / server.¹

¹ For power, emissions and TCO calculations the following methodology applies: Calculations for power and emissions and TCO are based on the Tirias Research TCO models which utilize publicly published x265 medium encoder performance for referenced platforms from NVIDIA, Intel and NETINT. These results are checked for equivalent VQ and bit rates based on analysis by Jan Ozer an independent expert. For TCO calculations power costs of \$.08 and a facility, cooling and operations multiplier of 3 is applied. Server costs are calculated based on best available pricing data and amortized over 3 years. With a baseline of equivalent video performance and a uniform cost model comparable power, emissions, and cost data can be calculated. These models have been validated in consultation with video service providers.

We can see the power and emissions impact of video encoding by comparing real world, best-in-class solutions for video encoding. Using 1000 concurrent 1080p30 streams as a baseline, the efficiency of ASIC based encoding results in 27,375 metric tons of CO² emissions per year compared to 509,175 for CPU based encoding. By 2025, with over 40 million concurrent streams, cloud encoding using x86 CPUs and software will produce 8,680,000 Metric Tons/Year of CO₂ emissions equivalent to about 1,000,000 conventional automobiles. Cars are a concrete point of reference; in 2025 using software encoding would produce CO₂ emissions equivalent to about 1,000,000 cars, whereas dedicated video encoding ASICs would reduce the number to about 53,917.2 Today's electric vehicles produce just under half the carbon emissions of traditional gas engine vehicles, on average. ASICs produce about 5% of the emissions of CPU based encoders. Video encoding ASICs bring a 20X improvement in energy efficiency compared to 2X to 3X for electric vehicles.

The Total Operating Cost of Streaming

The major factors which determine the operating cost of operating cloud or edge servers include capital cost of acquiring the servers and their projected useful life, power consumption, cooling, and facilities. The last three are bundled together to get a weighted power cost factor which is then multiplied times power to get an operating expense. In estimating TCO it is most important to use a standard factor to achieve apples to apples comparisons, and as in prior research on video encoding TCO, TIRIAS Research uses a weighted power factor of 3 which comprised power, facility, cooling, and maintenance. The combination of amortization and operating expense yields the operating cost.

In the case of edge computing, the cost to deploy servers increases as you get closer to the edge. Generally, placing servers at scale in base stations is the most expensive – space is almost always seriously limited and operational costs such as rent, site cooling and maintenance can be high. Dedicated rooftop enclosures, remote locations, and tall office building locations boost the cost of adding on premises edge servers. The variance and predicably higher costs and space constraints of base station placement have lead most industry insiders to the conclusion that we must settle for “close to the base station” within regional points of presence that are within a range of one to up to several hundred miles, with the sweet spot within the same metro (generally within 30 miles) in an adjacent metro (30 to 150 miles), or close to the nearest city. In the US,

² For emissions comparisons, the following methodology applies: The average miles per year driven of a car is 13,476 in the US or 21,687KM and Co₂ emissions are about 400g / KM or 8.67 Metric Tons/Year and about 4 Metric Tons/Year for EVs. Emissions for cars estimates from Knoblich, F. (2020) Net emissions reductions from cars and heat pumps in 59 world regions over time. Nature Sustainability, April 2020.

about 83% of the population lives in cities, making for a strong concentration for a single point of presence in a major metropolitan area.

The placement of cloud servers in regional points of presence is the most viable model allowing a major metropolitan area to be serviced by a single data center. The cost to deploy in regional points of presence are very similar to large data centers but subject to slightly higher facility and power costs, and we estimate annual operating costs could be up to 12.5% higher (power + facility + cooling costs at 200% of data center costs) than stated in dense metropolitan areas where even favorable locations could still have significantly higher cost structures.

Total Operating Cost Calculation

Combining the encoding capacity and power estimates already discussed and utilizing the TIRIAS Research TCO model with power estimates for CPU, GPUS and ASIC servers and a three year amortization for server capital costs, we can calculate the TCO of operating cloud video services. Serving 1000 concurrent users with NETINT ASIC encoding a single 1080p30 stream at x265 medium has a total operating cost of \$52,403 compared to GPU encoding at \$110,606 and CPU based encoding at \$580,535. This would require 12.5 NETINT based 1RU servers, 25 GPU based servers, and 125 software-based servers per 1000 streams. Using 1080p60 doubles the number of servers for all platforms. At 1080p60 we have 25 NETINT ASIC servers, 50 GPU servers, and 250 software servers per 1000 concurrent streams.

The application servers that execute client applications could range from mobile app / mobile game servers at 8 to 20 concurrent users per 1RU, to 1-4 users per 1 RU in the case of PC game streaming. Moving to 1080p60 linearly doubles the required number of servers per user due to the demands on the GPU for 3D processing and the massive number of pixels to be delivered using virtualized instances to run multiple users per server.




	ASIC	BEST GPU	SOFTWARE
Impact Factor With 1000 Transcode Channels @ 1080P30	NETINT G4	NVIDIA T4	INTEL SVT
 TCO Annual TCO / 1000 Concurrent Streams	\$52,403	\$110,606	\$580,535
 Environmental Impact Reduce Emissions by ~20X vs. Software	11.7 CO² Metric Tons	41.1 CO² Metric Tons	217 CO² Metric Tons
 Server Density Reduce Footprint by 10X vs. Software	12.5 NETINT Codensity Servers 1000 Streams	25 NVIDIA T4 Servers 1000 Streams	125 Intel Servers 1000 Streams

Figure 6: The operating cost for video encoding alone is \$580,533 with software based and \$110,606 with GPUs per 1000 1080p30 channels at x265 medium. Using ASIC based encoders significantly reduces encode costs, with a TCO of \$52,403. These calculations use \$0.08 per kWh power costs and a 3X multiple to include cooling and facilities. Servers are priced and amortized over 3 years.

Application Server Operating Costs

Looking at just the application servers and not the video encoders, for cloud mobile gaming at 1080p60, the cost to operate application servers for 1000 concurrent users is \$210,000 / year with a 3-year server amortization. For mainstream PC gaming, the cost to operate application servers for 1000 concurrent users is about \$1,045,000 / year with the same 3-year server amortization. Platform components including GPU choice is guided by both performance per watt and total power which could result in the same GPUs being employed in a high-end PC and a highly scaled out mobile application server. While the need for 3D performance is not as high for mobile applications, the total number of pixels rendered is. Whereas in PC gaming we might be shooting for higher framerates and resolutions, in mobile we are aiming for the maximum number of concurrent users with the same total pixels/second count. This drives similarities in power consumption for mobile application and desktop application servers but with much higher user/server count for mobile in most use cases. The cost per server is between \$7,500 and \$10,000 for 1RU mobile and desktop game servers, and power consumption at full load ranges between 650W to 800W.




Gaming/3D Capable Streaming High VQ (H.265 Medium)	Encode Servers	Application Servers	Total Service
1000 Concurrent Users	Infrastructure TCO	Infrastructure TCO	Infrastructure TCO
 Mobile Apps & Gaming 1080P60 Near Flagship Equivalent Performance Edge Implementation	\$52,400	\$210,000	\$265,400 \$265 Per Concurrent User
 Mainstream Gaming 1080P60 PC/Console Equivalent Performance In Edge Implementation	\$52,400	\$1,045,000	\$1,097,400 \$1097 Per Concurrent User
 4K High End/ New Console Gaming Elite PC/ New Console Equivalent Performance In Edge Implementation	\$419,220	\$4,181,200	\$4,600,500 \$4,601 Per Concurrent User

Figure 7: The operating cost for edge services utilizing regional points of presence utilizing high density NETINT ASIC based video encode and contemporary application servers with discussion of calculations below.

Total Operating Costs: Cloud Mobile Gaming and Cloud Remote Applications with NETINT Video Encoders

Mobile App & Game Streaming

Mainstream mobile 3D gaming and apps at 1080p60 with 1000 concurrent users require 50 1RU application servers (20 Users/1RU) and 25 1RU video encoding servers for a total of 75 1RU servers, or two full racks with space for networking and inter-rack cooling. For high fidelity mobile gaming at 1080p60 - Nintendo Switch class gaming - 125 application servers with the same 25 1RU video encoding servers would occupy four full racks with space for networking and cooling. In both cases the cost of encoding would be about \$52,400/year with a 3-year encoder server amortization. The cost to operate the mainstream app servers would be about \$210,000 /

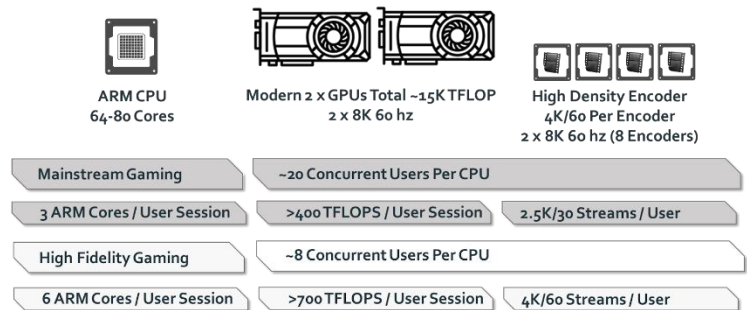


Figure 8: The components of a Cloud Mobile Gaming and Application Server include an Arm based server processor, industry standard GPUs, and low latency encoders capable of encoding multiple virtualized user instances at high visual quality.

1000 users, bringing the total operating cost with video encoding to \$265,400 or \$265 / Year / Concurrent User.

Mainstream & High-End Gaming Streaming

PC/Console gaming and apps at 1080p60 with 1000 concurrent users requires 250 1RU application servers at four Users/1RU and 25 1RU video encoding servers for a total of 275 1RU servers, or 7 full racks with space for networking and inter-rack cooling. At four users and 1080p60 resolution, the operating cost for 1000 users is about \$1,045,000 for the application servers and \$52,400 for the ASIC video encoders bringing total cost to \$1,097,400 or \$1097 / Year / Concurrent User.

With the onset of 7nm and 8nm GPUs, new servers can be deployed delivering an experience more or less on par with the newest consoles such as the Xbox Series X with two to 4 virtualized users per server depending on delivered resolution – 1080p or 4K. The hard step from 1080p60 to 4K/60 complicates the logic of users per application server, but two to four is possible with the latest AAA games at new console speeds, and potentially more than four users can be supported with carefully balanced resolutions and choice of which games run concurrently on which servers. Increasing fidelity to 4K/60 gaming and the highest level of gameplay would limit serving two users with a single 1RU application server and one 1RU video encoder to every 10 users. At 1000 users, that is 500 1RU application servers and 100 1RU video encoders or about 15 racks. This elite use case for 4K gaming would have a total operating cost of 4 X \$1,045,480 and 8 X \$52,403 for a total of \$4,601,144 or \$4,601 / Year / Concurrent User.

The Future

Edge deployment of interactive application services will require high density compute to minimize footprint, maximize reliability, and decrease the costs of fielding and operating servers. Applications at the cloud edge will rely heavily on video encoding technology and upgraded network technology including 5G to transit those experiences to end users with the lowest possible latency. As the network and video encoding improve, the cloud will become almost instantaneous, able to deliver experiences to users which fool their senses and make remote applications seem like they are running locally on high performance clients. Emerging ASIC based encoders are a breakthrough in density, performance, and cost enabling edge applications to scale economically at the edge. By lowering latency, they enable real time technologies including machine learning and sensor data fusion to create entirely new, intelligent, and interactive experiences.

For more information on NETINT edge encoders, visit www.netint.ca online. To read more on the cutting edge of video processing, edge computing, machine learning and more visit www.TIRIASresearch.com/research online.

Copyright © 2020 TIRIAS Research. TIRIAS Research reserves all rights herein.

Reproduction in whole or in part is prohibited without prior written and express permission from TIRIAS Research.

The information contained in this report was believed to be reliable when written but is not guaranteed as to its accuracy or completeness.

Product and company names may be trademarks (™) or registered trademarks (®) of their respective holders.

The contents of this report represent the interpretation and analysis of statistics and information that is either generally available to the public or released by responsible agencies or individuals.

NETINT, Edgefusion, and combinations thereof are registered trademarks of NETINT Technologies, Inc.