TIRIAS RESEARCH

The Emergence of Cloud Mobile Gaming

A New Solution Utilizing Arm-Based Infrastructure & High Performance, Low Latency Video Encoding & Game Graphics on Virtualized Sessions to Stream Mobile Games to Smartphones

Whitepaper sponsored by NETINT

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

April 2020

Contents

Introduction

Cloud streaming employs servers to run applications remotely, using video streaming technology to deliver those applications to client devices. The potential advantages include a uniform experience regardless of client platform, improved manageability of a single cloud-based code base regardless of client platform, instant start for easy movement into games enabling subscription offers, and low latency server-side interactions for massively multiuser experiences. The initial forays into cloud streaming at scale have been in PC and console gaming, driven by high financial potential and user engagement. Today PC and console gaming comprise 55% of the global $150B[1] gaming market and converting even a tiny portion to streaming is a multi-billion-dollar opportunity.

Very little has been said regarding the other 45% of the market, the mobile game streaming opportunity. The technical challenges and economics of mobile gaming are different – the scale is massive, the network is cellular, and the distribution of mobile apps is virtually impossible outside of established app stores. The existing code base for mobile apps is centric to the iOS and Android platforms, encouraging server-side implementations of the mobile technology stack. Powering mobile game streaming requires a transformational user experience and superior economics to break the lock hold of the app store establishment.

Technical innovation is required in the cloud, the network, and the user interface. The user experience will need to embrace the technical advantages of cloud streaming while overcoming the economic burden of procuring and operating cloud-based game servers and content networks. Service providers will need to develop and attract enough game content to entice users into a new service.

Tirias Research has been engaged in the analysis of the technologies required to advance cloud streaming for gaming and XR. This paper focuses for the first time on mobile streaming – the experience requirements and the technologies necessary to deliver cloud streaming to millions of smartphones and mobile head mounted displays. Delivering mobile streaming at scale requires technology to achieve an unprecedented level of compute density. The paper's sponsor, NETINT, has developed high density video encoders with integrated solid-state storage designed to deliver streaming entertainment at scale with extremely low latency and low TCO. These products fit into the architecture of a new, mobile streaming cloud that can power a transformative experience for mobile gaming.

[1] https://newzoo.com/insights/articles/the-global-games-market-will-generate-152-1-billion-in-2019-as-the-u-s-overtakes-china-as-the-biggest-market/

Nirvana

Today, the app store is the status quo, ingrained in the habits of smartphone and tablet users. Not unlike the shift to music streaming, a transformation in the user experience and the economics of the current service model are required for any new service offering to succeed. Users were willing to trade a small library of purchased content for instant access to virtually every song. If mobile games are to experience a similar transition the experience must deliver better economics and a new experience – something akin to a mobile gaming Nirvana.

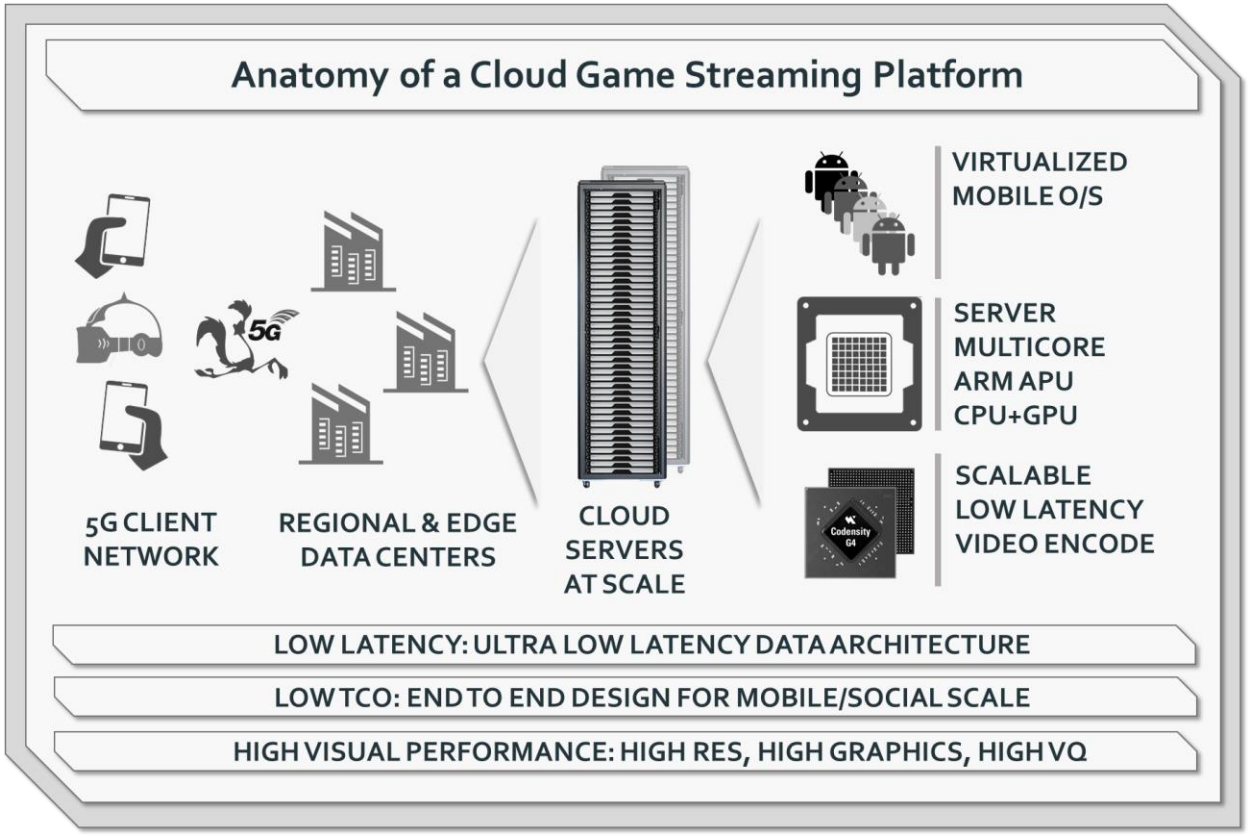
NIRVANA - A World of Games Awaits

> *Nirvana is a social game streaming service where every game start, every social experience, every interaction, and every transition is instant. Players can step from Narnia into Middle Earth into 2077 instantly and seamlessly. Game worlds contain arenas where thousands of players simultaneously interact. When players launch the Nirvana app all games share your identity, prestige, badges, and friends. In game purchases are all available with Nirvana Karma that can be used across all game experiences. New games appear in Nirvana and players can instantly click to play and enter the game without friction. A player's goods, clan, and avatar can be carried throughout Nirvana. Massive interactive experiences engage thousands of players with zero latency in the most social and detailed online worlds ever created. Gamers, able to move from title to title, are never bored, never feel locked in, and are always able to seek out new experiences and engagements with their friends. And anyone can come to Nirvana - you can play on any smartphone or head mounted display. Nirvana - a world of games awaits!*

NIRVANA sounds great for gamers and with opportunities for new business models for service providers – but it will not build itself, and challenges stand in the way. Streaming adds costly cloud infrastructure that service providers must procure and maintain. It requires customers to be - and to stay - connected to the network. Failures in the network or by the streaming service provider break the user experience and create swift dissatisfaction. And the technology must scale, providing a high-quality experience to users who might be anywhere. Finally, the entire solution must attract developers and publishers seeking strong returns.

Cloud Mobile Gaming

Mobile streaming is complicated by the many combinations of infrastructures that run games today, and the opportunity for those platforms to stream to mobile. Much has been written on the PC and console gaming worlds – and focus on streaming mobile games to mobile platforms – MG2M for short. In MG2M we are streaming games designed to run on today's smartphones and tablets using a cloud infrastructure. The most efficient implementation is to run Arm-based servers and cloud ready instances of iOS or Android. Gaming grade graphics acceleration and low latency video encoding are required for each user session and scale 1:1 with server resources and active users. High density servers are needed to run large numbers of user session, and regional data centers are required to lower user experience latency over wireless networks. Let's look at the anatomy of a mobile to mobile game streaming platform:



Virtualized User Instances

The virtualization of user instances including real time video output of 3D accelerated game graphics is required to provide flexibility in the architecture of dense cloud computing platforms. Today's mobile platforms integrate an Arm-based processor and an integrated GPU. In the high density, multi-user cloud compute server node, virtualizing the user instance and separation of these functions into dedicated, high performance ASICs provides the greatest flexibility and highest density.

Arm Processors

New Arm-based servers that can deliver today's native mobile apps - and low latency cloud-based video encoders are lowering the operating cost and increasing the scalability of prospective services. Arm has enabled the processor ecosystem by developing scalable designs for massively mutli-core server processors and enabling a software and core logic ecosystem. The rise of processors with dozens of CPU cores and paired GPUs capable of running today's mobile apps natively will simplify migration and enable game streaming on emerging cloud-based services. These servers simplify the virtualization of mobile operating systems allowing multiple user instances to run on a single chip. Together the combination of virtualization and native, lower power Arm-based servers is a tipping point that significantly decreases the total cost of operating mobile streaming services.

Ampere™ Altra™ processor is an 80 core Arm Neoverse N1-based server processor running at up to 3Ghz announced in February of 2020 supporting eight DDR4-3200 channels and 128 lanes of PCIe Gen4 per socket (up to 192 for 2P).

Huawei/HiSense Kunpeng 920 is a 64 core Arm-based processor developed for demanding cloud workloads announced in January of 2019. The processor has 64 cores running at 2.6 GHz and features eight memory controllers with DDR4 memory at 2.93 GHz for a total memory bandwidth of 1.5 Tb/sec.

Marvel Thunder X3 fearues up to 96 Arm v8.3+ custom cores running at 3GHz. Mutlithreading allows the 96-core CPU to run up to 384 threads. Announced in March 2020, the processor is positioned to excel at emerging cloud and supercomputing workloads.

Scalable Graphics Accelerators

The capability of modern GPUs vastly exceeds the compute requirement of even the most demanding mobile games. However, today's smartphones are increasingly capable of rendering demanding 3D experiences, and emerging mobile applications – including augmented and virtual reality – bring compute requirements exceeding the requirements of even the most powerful of today's gaming PCs. This range of experience drives the need to support multiple levels of user experience on the same server infrastructure.

From a software perspective, delivering the visual experience of a 2D or 3D game requires the visual performance of a contemporary graphics accelerator. The use of standard graphics APIs, namely OpenCL, Vulcan and Metal provide gateways to the development of virtualized graphics drivers able to support existing games with minimal code changes. Some code changes including standardizing the handling of graphics such as resolution handling and user interface to make the game more seamless with the streaming platform seem likely if not essential. The development of games that fully take advantage of the low latency backend will require dedicated development.

Real Time Video Streaming Encoders

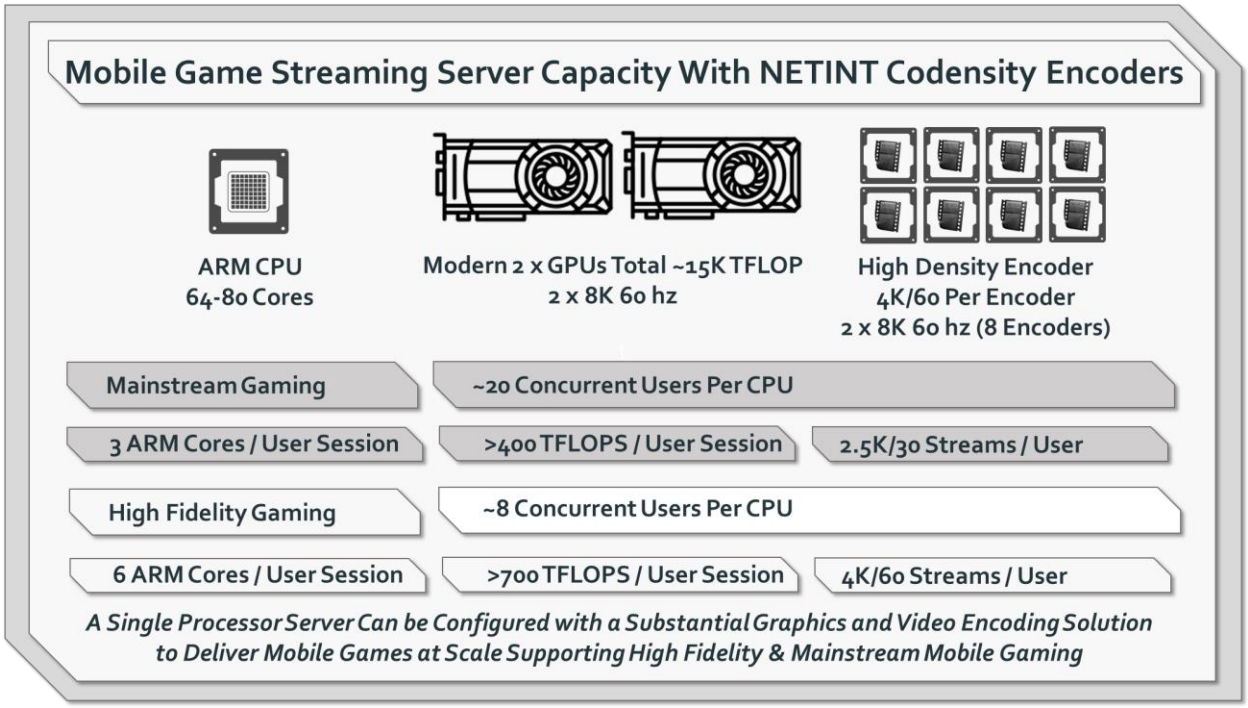
Today's integrated video encode/transcode logic in GPUs and Mobile APUs is designed for the acceleration of video playback and offloading for social video uploading and consumer video editing. In contrast, cloud-based video encoders require real time performance, ultra-low latency, high visual quality, multi-stream output and high density.

Today's cloud-based video encoders, featuring high visual quality and real-time encoder performance, are making it possible to deliver even the most visually demanding 2K to 4K resolution games over bandwidth constrained networks. Modern encoders have very low latency while delivering the visual quality of a native application experience. They are able to meet the visual quality expectations of consumers and render at the native resolution of today's 2K to 4K mobile displays. The utilization of high performance CODECs, including the current state of the art class of H.265 and VP9, can stream 4K gaming content today. The advent of next class, one

where industry players have centered on the royalty free AV1 CODEC, promises to bring high framerate and ultra-resolution for emerging XR applications.

Delivering Mobile Streaming Video with High Density Encoders

The delivery of game streams requires scalable video encoding unavailable in today's consumer-grade GPU encoders. To fill this gap and enable streaming solutions to come to market, NETINT has developed a series of dedicated encoders and plans to introduce a hybrid encoder/high speed storage module. The NETINT high density encoder solution can be employed in combination with a single or dual processor Arm-based processor and high-performance GPUs to deliver high density mobile cloud gaming solutions. Multiple levels of user experience can be achieved by allocating an appropriate number of processor cores and setting the resolution appropriate to the game experience and client screen. Users demanding the highest fidelity games can be supported concurrently with mainstream games with smart allocation of server resources.



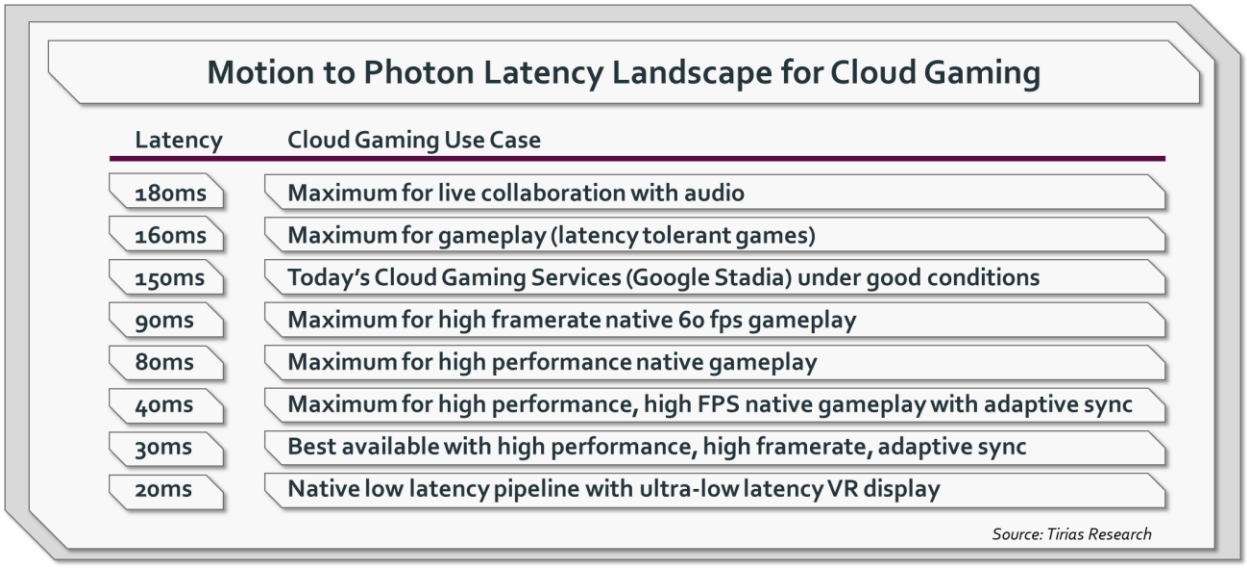
The NETINT video encoders support high speed H.264/AVC and H.265/HEVC CODECS on PCIe add-in cards and U.2 form factors for cloud streaming and social video delivery. The company offers stand-alone encoders and is introducing the EdgeFusion line that combines SSD storage and video encoding in the same U.2 and PCIe form factors utilizing the company's high performance Codensity ASIC. All offer the high quality required for game streaming with the high density and low cost necessary for deployment at scale. For easy integration into existing cloud encoding workflows these solutions use standard FFMPEG software interfaces. Today NETINT encoders deliver a 1080p30 stream at less than 1W with a standard 40X resolution ladder and supports up to 4Kp60 with H.264 or H.265 encoding. The low power design enables high density encoding using a U.2 form factor – the standard SSD interface module – with up to 10 encoding modules in a 1RU form factor. NETINT's encoders can achieve very high density

while still keeping deterministic ultra-low latency. With 10 U.2 moduels installed in a 1RU single server, 40 1080p60 real-time ultra-low latency streams can be encoded – more than 1000 encoded steams in a single high-density rack.

The company has successfully demonstrated and qualified its solution on both Arm and x86 servers utilizing a number of operating environments. With high density and industry standard interfaces the NETINT encoders are positioned to provide the cost, visual quality, and performance necessary to enable mobile streaming solutions at scale.

Latency

Low latency is critical to multi-player gaming and has been frequently cited by gamers as a major point of resistance with cloud gaming. We estimate the ideal motion-to-photon latency for smartphone gaming is 70ms, or about 2 full frames at 30FPS, 4 full frames at 60FPS. For future head mounted display solutions such as virtual reality, head tracking creates particularly demanding motion to photon requirements, and latency can create motion sickness above 20ms. It is useful to survey the latency landscape of today's native gaming platforms and cloud services against experience thresholds.



Cloud encoders must achieve sub-frame latencies minimizing their contribution to the overall motion to photon latency. NETINT encoder latency for 1080p gaming is 8ms with 8 concurrent encoded streams at 30FPS, and 4 concurrent streams at 60FPS. At 720p, latency is 4ms latency with 16-stream 720p30 encoding. Breaking down the latency budget internal to the encoder, at 1080p, NETINT T408 achieves a YUV transfer of 1.8ms, encoding time of 4.8ms and bitstream transfer time of about 0.03ms. At 720p, NETINT T408 achieves a YUV transfer of 1.3ms, encoding time of 1.9ms. and bitstream transfer time of the same 0.03ms. At 30FPS frame to frame time is 150 milliseconds, therefore encoding latency is at one quarter of a frame, whereas

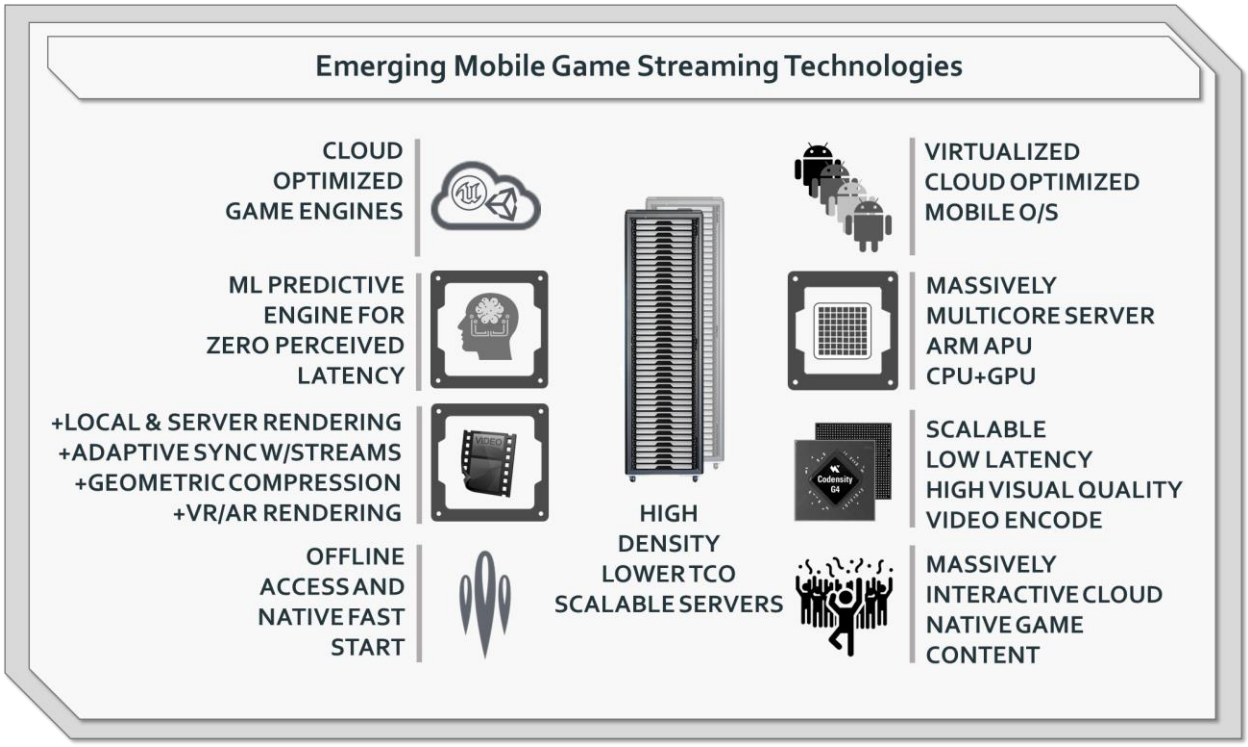
current total latency for encoding services can exceed 150 milliseconds. The NETINT encoder's low latency helps cloud gaming solutions achieve ideal targets for mobile streaming.

Visual Quality

Achieving high visual quality (VQ) has proven important to meeting user expectation for gaming on today's high-resolution smartphone displays. The following images show the using NETINT T408 and best available GPU video encoders, both running on "low latency mode" encoding the CSGO output stream sequence. The T408 encoded average bitrate is 4701.84 kb/s while the best available GPU encoder average bitrate is 5531.01 kb/s. Detail is maintained in textures and the outline of game characters and targets is maintained at a level suitable for competitive, multi-player gameplay.

The Future

Emerging technology will continue the evolution of mobile game streaming. Optimization of game engines and games to take advantage of low cloud latency will provide a differentiated experience unique to cloud gaming. Improvements in the scalability and performance of processors, GPUs and video encoders will achieve 4K resolution and beyond for high visual fidelity and AR/VR. The use of machine learning will accelerate game response and remove the perception of latency from user input, removing lag in demanding multi-player games. And the development of video encoding technologies such as AV1 will permit further reduction in bitrates for high resolution, high fidelity gameplay.



Delivering a transformative experience for cloud mobile gaming will require the simultaneous development of these technologies and the ability to maintain them in virtualized, cloud native instances. The advantage of hosting these services on low power Arm-based servers is a smoother transition for developers today, and a low performance per watt, mobile scalable solution tomorrow.

High density video encoding technology is ready to power the first wave of cloud mobile game services at scale. Delivering high visual quality for high motion graphics at low latencies, NETINT video encoders have the performance necessary to power these services at scale. Gamers can expect that these emerging mobile game services will offer nearly native visual quality and high responsiveness. These advantages will attract gamers, games, and game tech investments to create a virtuous cycle for mobile cloud gaming.