

Android in the cloud on Arm native servers

December 2020

Executive summary

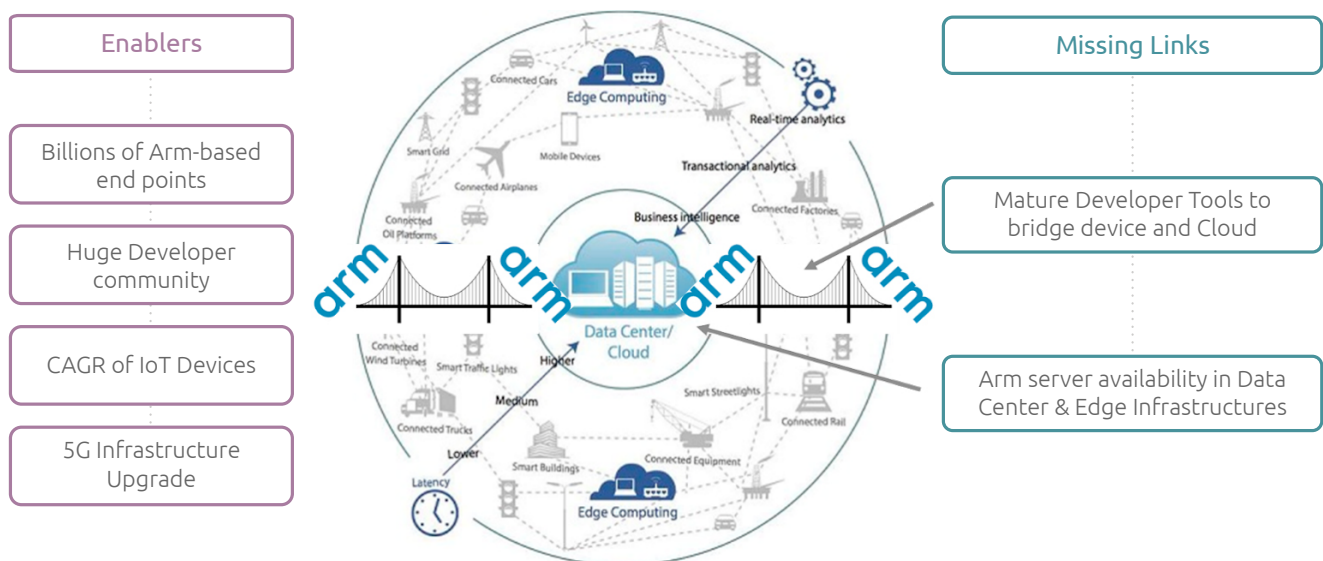
All the chatter about Arm servers and the role of Android in the cloud in recent years generates questions about what these things could mean for a largely mobile based platform like Arm/Android. This whitepaper examines the practical and commercial use cases that are driving a new paradigm linking the primary mobile and embedded computing platform in the world today with its new cousin in cloud infrastructure: Arm servers + virtualisation + native Android execution. While detailing the use cases driving a mobile platform migration to cloud infrastructure, this whitepaper also showcases the solution that Ampere Computing, Canonical, and NETINT Technologies have partnered to produce addressing this migration. The solution is an example of how a service provider or developer can take advantage of Arm native computing in a cloud context to bring together an ecosystem of over 3 million primarily mobile apps integrated with benefits from a cloud enabled infrastructure. A case study around cloud gaming fills in the practical real-world considerations of this new paradigm. Cloud gaming really showcases the advantages of all the solution components the partners have integrated to make a scalable, efficient cloud resource available to a new class of cloud-native applications catering to the billions of users with Arm based devices worldwide. The whitepaper concludes with thoughts and predictions about the evolution of this new paradigm as hardware and software components mature in the coming years.

A new paradigm

The Arm architecture has dominated the mobile processor market with its unrivaled ability to maximise power-efficiency. As a result, there are now billions of Arm-based chips used in mobile phones, laptops, tablets, IoT devices and embedded applications throughout the world. Over the last few decades, Arm's ability to deliver excellent performance at a fraction of the power of existing architectures has driven a new era in mobile computing.

Now, the Arm ecosystem is set to experience another period of innovation with the expansion of the platform to the cloud increasing the reach of the primarily mobile platform as the world rushes to move data and services into the cloud. In fact, the vast majority of what users access on their devices today already lives in the cloud somewhere – from movies, photos and games to enterprise computing and online retail experiences. Facilitating this data migration is an Arm ecosystem supporting a true cloud native computing environment from the edge to the core, enabled by a super-power efficient architectural platform built from the ground up with the ability to scale to previously unimagined core densities. This is where Ampere's market-leading server platform now leads the industry beginning with the Ampere® Altra™ product launch delivering the maximum performance at the lowest possible power.

Ampere's server products are based on the Arm RISC architecture, the very same architecture used by so many devices in our connected world. The combination of server compute density, instruction set compatibility and a virtualisation layer capable of instantiating 100s of Android instances on a server offers a unique platform for new innovation and use cases to lead the next wave of connected applications and services seamlessly integrated between cloud and a wide world of devices. This combination is the new paradigm needed to build the largest developer ecosystem in the world between cloud, client and edge devices built on Arm compatible processors. This paradigm is what we call the Arm Native Cloud.



The Arm Native Cloud Environment

Market drivers

The key use cases driving this new wave of innovation are cloud gaming, developer services such as Continuous Integration and Continuous Development (CI/CD), and secure enterprise application streaming in the cloud including the use of Virtual Mobile Infrastructure (VMI). These segments are driving the need to develop innovative tools and a cloud compute infrastructure running natively on Arm's architecture. All of these use cases have a similar theme; enabling mobile devices to access data or additional programming without having that data or program reside on the actual device. Instead, the application or a portion of it and the relevant data live in the cloud where streamlined applications can run on a high performance and secure infrastructure, allowing users to access it through whatever device they might have. This shift has in turn made compute power, security, cost of maintenance, and ease of use extremely important when enabling new and innovative workloads in the cloud.

Markets Driving Arm Native Cloud

- Real Time Behavior. AI & Analytics
- 5G → Proximity to User/Device
- Guaranteed Service Delivery
- Content Sharing/Streaming
- Autonomous Vehicles
- Smart Everything
- Cloud Gaming
- AR/VR

Gaming:

- Mobile gaming is expected to outpace console growth and, according to [IDC](#), "continues to be a dynamic market with a bright future". They expect global revenue to rise by more than 10% from 2019-2024 with 5G among the key drivers.
- This represents a huge opportunity for cloud based gaming. Transitioning to a cloud platform can transform games from a 'game for one' to a game for many with a single platform that works on many different devices. In addition, game developers are now less constrained by the resources on any given device by taking advantage of server class hardware. This will also take a significant test and compatibility burden off developers because they don't have to support multiple devices and operating systems. The experience for the user is now easier because all they need to start playing a game is access to a web browser. By eliminating the download process, game play, evaluation and organisational models are revolutionised.

Android developers:

- Another large potential cloud computing market is the Android developer infrastructure market. This market is expected to grow significantly as Android continues to expand its market share and supported devices. Traditionally, developers had to purchase or rent many pieces of different hardware to test their applications. The proven method to date is to rent many instances of a single board computer to do large scale testing and development. However, this method of large-scale CI/CD is very onerous to manage and scale inside a typical cloud service provider's infrastructure.
- By moving testing into the cloud, they now can simply access whatever amount of equipment they need virtually – at any time and for whatever length of time they need and, most importantly, only pay for the time they use to get their applications tested. In the future, the addition of new client server

programming tools will allow developers to build their applications with the cloud parts and device pieces in one consolidated environment. This enables seamless integration of the testing and deployment of their applications to both cloud and device infrastructure simultaneously.

Enterprise applications and VMI:

- A third and equally important market is the enterprise, motivated by delivering more information in a secure and efficient manner to employees, partners and customers. This market is very diverse and includes segments such as healthcare and financial services that have data privacy concerns and are more apt to keep information in secure cloud domains.
- Since security in these markets is of the utmost importance, companies adopt a cloud-based model, allowing users to access sensitive and important data without actually putting the data on their phones, tablets or computers. This model is very attractive to enterprises because it does not matter what type of device an employee has – they simply need access to the internet. Enterprises can also reduce their internal application development costs by providing a single workplace application that can be used across different form factors and operating systems. These applications can be easily delivered directly to the employee devices via the cloud, while maintaining the assurance of data privacy and compliance.
- Other examples of VMI include virtual training, live sharing for demos, live gaming projection, interactive customer support and virtual device sharing. These use cases are still early in their lifecycle, but as the infrastructure software and hardware improve, the user experience will get better and better creating all new services and applications.

The Internet of Things (IoT):

- In addition to cell phones and tablets, the expansive ecosystem of embedded Arm devices such as sensors in factories, cameras in smart cities, autonomous vehicles, drones, connected appliances and other industrial or consumer-type appliances drive a continued opportunity for architecturally equivalent cloud-based infrastructure. Looking at one market in particular, it is easy to imagine the autonomous and highly connected vehicle of the future running many instances of a mobile operating system. Separate and secure instances can deliver a rich experience to the occupants of the vehicle while continuously monitoring and operating the critical systems of the vehicle. All local instances being aided by a cloud backed infrastructure for data collection, analysis and entertainment streaming within the vehicle. These use cases are better enabled through a cloud-based computing model.

Arm Native Cloud infrastructure

These types of innovative workloads require a modern architecture built for the cloud with the below requirements.

- **Run natively on Arm-compatible servers** – Ampere processors provide the greatest architectural compatibility to a wide range of devices mentioned above.

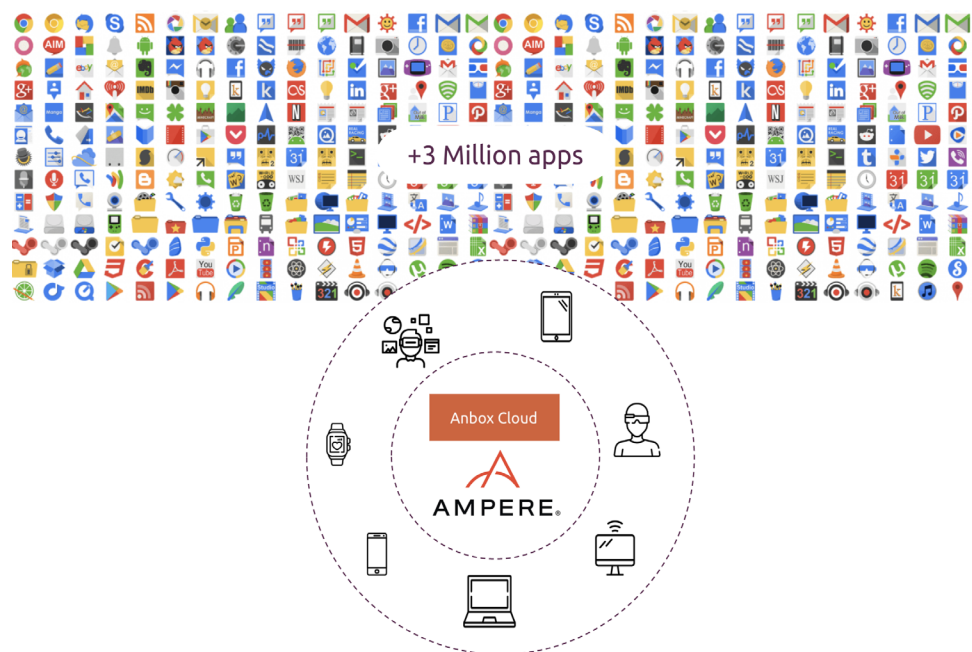
- **High-performance, power efficient processors** – Ampere processors are the first ever cloud-native processors that have been built from the ground up for cloud workloads. Existing server solutions are based on 30 year old technology that can't deliver the scalability, cost, efficiency, power and core density demanded by the cloud and edge of the future.
- **Cloud native compute infrastructure** – Ampere processors support extremely dense containers, microservices, functions as a service (FaaS), and other new programming models that allow scaled services to be built and deployed rapidly with a very consistent and deterministic quality of service.

The solution: Canonical's Anbox Cloud running Ampere Arm servers with NETINT Encoders

For anyone running code at the edge or in any of the Arm-based IoT devices mentioned above, there are significant advantages to running code that is native to that platform. Ampere, Canonical and NETINT Technologies have partnered to produce a solution that caters to the myriad of native applications developed for the Android operating system.

Anbox Cloud

Canonical developed Anbox Cloud on highly efficient containerised workloads using Android as a guest operating system to enable service providers or enterprises to distribute Android applications from the cloud to any device. By running this platform on Ampere server products such as Ampere® Altra™ and Ampere eMAG®, Anbox Cloud customers are able to run all of their applications natively, delivering the performance, low power and cost advantages needed to satisfy consumers demands.



[Anbox Cloud](#) allows enterprises and service providers to deliver mobile applications at scale, more securely and independently of a device's capabilities. Anbox Cloud is highly scalable and offloads compute, storage and energy-intensive applications from devices to any cloud. As a result, developers can deliver mobile applications independently of a device's hardware capabilities. Developers can deliver an on-demand application experience through a platform that provides more control over performance and infrastructure costs, with the flexibility to scale based on several variables that limit device-only applications such as spikes in user demand, device proliferation, bursty data events, and seasonal fluctuation in service demands.

Anbox Cloud can be hosted in the public cloud for near infinite capacity, high reliability and elasticity or on a private cloud or edge infrastructure, where low latency and data privacy are a priority. Public and private cloud service providers can easily integrate Anbox Cloud into their offering to enable the delivery of mobile applications in a PaaS or SaaS-model. Telecommunication providers can also create innovative value-added services based on virtualised mobile devices for their 4G, LTE and 5G mobile network customers.

In order to deliver the backing software infrastructure for running Android in the cloud and delivering high density of individual instances, each providing low latency and high quality video streaming, a stack of different software components comes into play. Anbox Cloud, furthermore, allows scaling across geographical regions in order to bring the actual Android instances as close as possible to their users.

Anbox Cloud is based on top of existing software technologies from Canonical, namely [LXD](#) as a container hypervisor and the Ubuntu operating system. LXD provides secure and performant system containers using a variety of features available from the Linux kernel. The containers themselves are made up of the Anbox runtime environment which abstracts a nested Android container from any direct hardware access and allows mediation and integration with drivers from NVIDIA or AMD for GPU support.

The Anbox runtime environment further integrates WebRTC based streaming of the visual output of the Android to a remote user utilising either software or hardware accelerated video encoding.

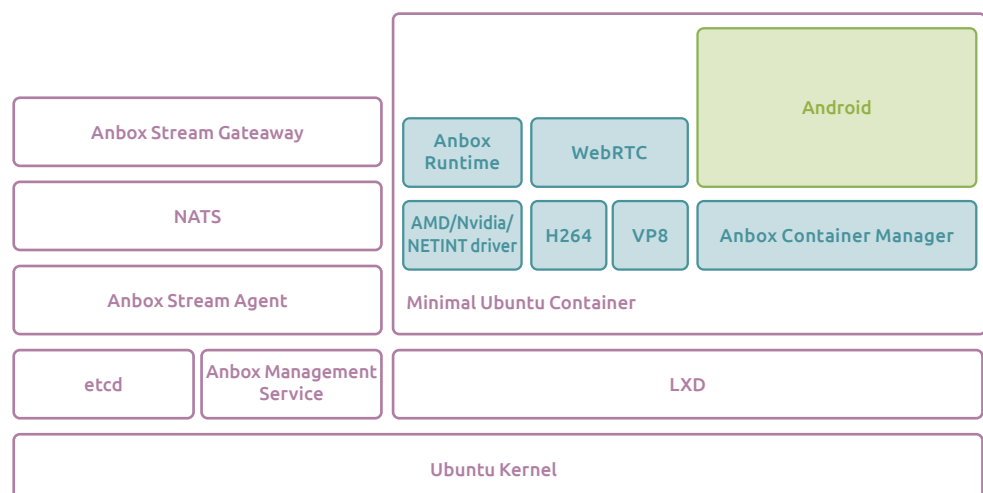
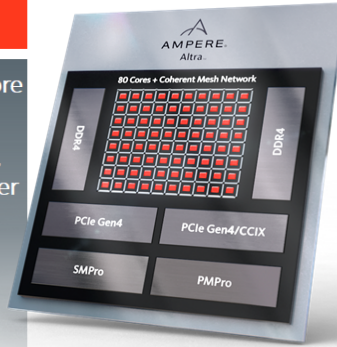


Figure 1: The Anbox Cloud software stack

Outside of the containers, Anbox Cloud comes with additional components to build a control plane which simplifies and abstracts the use of the underlying container platform. The Anbox Management Service (AMS) provides an abstraction layer on top of LXD which adds resource management, container orchestration, application lifecycle management and various other things which are needed to run Android at scale and high density on a single machine.

To allow Anbox Cloud handling the scale out across multiple regions, the Anbox Stream Agent connects a single region powered by AMS and LXD via a NATS message queue to a centralised management service called Anbox Stream Gateway. The Anbox Stream Gateway allows the creation of user specific streaming sessions and routes users to the nearest region.

Predictable High Performance	High Scalability	Power Efficiency
<ul style="list-style-type: none"> • Up to 80 cores • Coherent mesh-based interconnect • High memory bandwidth and density 	<ul style="list-style-type: none"> • Industry leading cores/rack • Cache-coherent multi-socket support • Flexible I/O connectivity 	<ul style="list-style-type: none"> • Leading power/core • Advanced system, security, and power management • Monolithic die on leading 7nm process



The image shows a 3D rendering of the Ampere Altra processor die. It is a square chip with a grid of red squares representing the core array. Labels on the die include 'AMPERE Altra', '80 Cores • Coherent Mesh Network', 'DDR4', 'PCIe Gen4', 'PCIe Gen4/CCIX', 'SMPro', and 'PMPPro'.

Value pillars of the Ampere™ Altra® processor available to the market in Q4 2020

The Ampere server platform

As highlighted above, the [Ampere server platform](#) represents a completely new processor architecture tailored for the emerging growth of cloud computing and next-generation data centers in the core or at the edge. Ampere processors are the industry's most power-efficient server class processors.

Featuring custom high-performance Armv8+ cores, the Ampere eMAG and Altra platforms integrate advanced 64-bit single-threaded Arm Cores with high capacity memory and I/O subsystems. The highly integrated, purpose-built Ampere solution delivers the highest performance, lowest total cost of ownership (TCO) for private and public clouds.

Altra is the world's first cloud native processor, designed to meet the requirements of modern data centers. Altra delivers predictable performance, high scalability and power efficiency for datacenter deployments from hyperscale cloud to the edge cloud. The processors deliver efficiency for workloads including data analytics, artificial intelligence, database, storage, telco stacks, edge computing and web hosting. Key features of the Ampere Altra family of products include:

- Up to 128 Cores per socket, dual socket capable
- Eight, 2DPC, 72-bit DDR4-3200 channels
- Up to 128 PCIe Gen4 lanes in 1P configuration & up to 192 PCIe Gen4 lanes in 2P configuration

- Support for multiple x16 CCIX accelerator slots
- Slots for bulk accelerators, GPU, networking and storage devices
- TSMC 7nm process technology

Predictable behaviour



In a cloud environment, where multiple applications are running on the same processor, problems arise due to the different requirements of various applications. Any time multiple dissimilar workloads are configured to share the same server, the likelihood exists that a small subset will dominate resource usage, compromising the predictability of other workloads sharing the platform. This lack of predictable behaviour is illustrated above. Multiple workloads scale linearly with Ampere Altra yielding the most predictable behaviour in the industry.

Graph above shows stress-ng CPU workload scaled up to the maximum number of threads/cores available per server platform

A winning partnership

The combined solution of Canonical's Anbox Cloud running on Ampere's Arm-based servers and NETINT ASIC based encoders delivers the most efficient platform for processing Arm native code in the industry. The following key attributes define the solution:

- **Predictable high performance:** The Android ecosystem is developed on Arm. To run efficiently and natively, Canonical uses the scalable Arm servers to run code already optimised for Arm, eliminating the inefficiencies of running Android on an x86 server by using technologies such as binary translation. The high core counts in Ampere products allow near linear scaling with very low standard deviation between response times maximising the total number of Android instances while minimising the execution variability on any given Android instance.

- **Hardware-rooted security:** A unique aspect of the Ampere servers is an architecture that utilises a single-thread per core. This protects the Anbox Cloud Android instances against hacking by ensuring there is no sharing of the execution pipeline, data and L1/L2 cache between threads. This feature is key to minimising side-channel attacks that have been shown to occur when running multi-threaded cores. In addition, single threaded cores enable performance predictability that are critical for cloud implementations and microservices based architectures.
- **Best-in-class TCO:** Ampere's Arm-based servers were designed to specifically address scalability needs for cloud computing. This enables Canonical to run more containers simultaneously per server node than anything in the competitive x86 ecosystem. In fact, initial benchmarks show that Anbox Cloud can instantiate dozens of containers running on the current Ampere servers and hundreds of containers on the new Altra processors. This is especially important at run time to dynamically react to variable customer demand. Maximum instance density also lowers the overall unit economics and TCO of the solution for end-users and service providers.

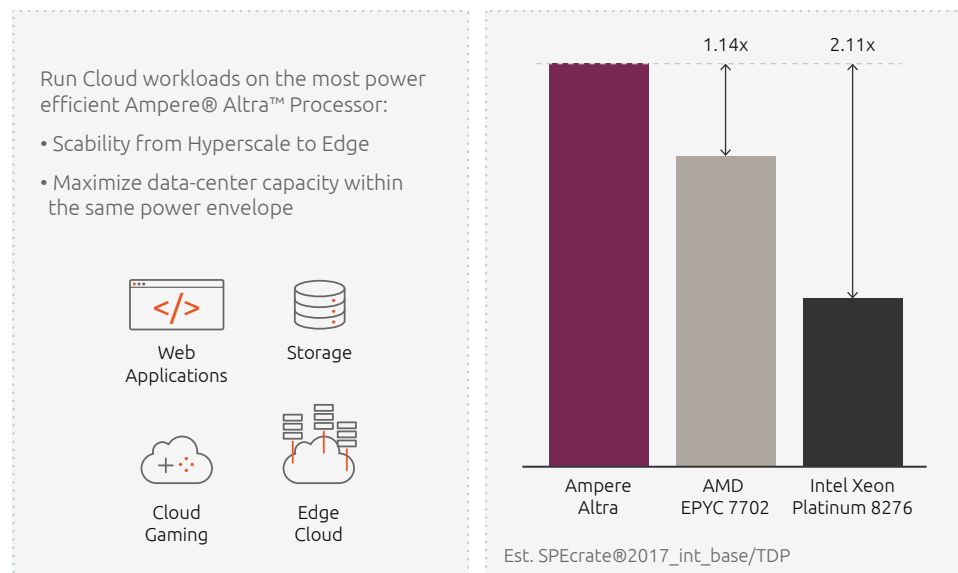


Figure 2: Leadership Power Efficiency: Ampere is the leading server-class processor in performance per watt

- **Best-in-class power efficiency:** The Ampere server platform also enables Canonical to address the growing power consumption challenges in data centers today, particularly those operating at hyperscale or in power constrained edge installations. Scaling up existing power hungry CPUs cannot solve this problem. That is why Ampere designed its processors to provide the highest level of power efficiency.

"It's no secret that TCO is a key driver of cloud provider decisions. And that's what gets Arm excited about the Android in the Cloud solution Ampere, Canonical and NETINT have developed. The Ampere Altra CPU packs 80 high-performance Neoverse N1 cores into a very power-efficient design that should deliver incredible TCO for Android in the Cloud use cases like cloud gaming, Android app development and IoT enablement to cloud providers."

—Robbie Williamson, Arm Senior Director of Solutions Engineering

Featured use case: Android cloud gaming

While streaming games represent an enormous market opportunity, the use case also presents significant challenges with regards to scalability and performance. On one hand, game developers need to ensure excellent user experiences for their customers, which means they need to have low latency and maintain excellent video quality. On the other hand, they also need to be profitable which means they need to have a cost-effective platform that can be easily scaled, and one that can provide gamers with the experience to play any game, any time, on any device, in any location.



NETINT Codensity™ T408/T432 Video Transcoders

- Real-time H.264/H.265 Decoding/Encoding
- Deterministic, Ultra-low-Latency
- 8K/4K/HD support
- HDR format support: HLG/HDR 10/HDR 10+/Dolby Vision
- Flexible NVMe U.2 modular plug & play form factor
- High-Density - 16x 720p30 Streams per U.2 module, 64x 720p30 per T432 card
- Ultra Low Power - 7 watts per U.2 module; 27 watts per T432 card
- Native support for containerization and Virtualization
- Integrated with FFmpeg SDK

The Anbox Cloud solution was designed to meet these requirements in the gaming market. A partnership with both Ampere and NETINT Technologies ensures the solution can run compute intensive and video encoding intensive mobile workloads natively on Arm at scale in the cloud. The combined solution delivers a very cost and power effective streaming platform for dense cloud gaming with the highly efficient software running multiple instances of a graphics intensive game simultaneously. The companies have been tuning the solution beginning on the Ampere eMAG server but continuing on the Altra systems to showcase both efficiency and performance for cloud gaming. Figure 3 shows the density and performance of [BombSquad](https://www.froemling.net/apps/bombsquad)¹, a 3D mobile game, running at extreme density on an eMag server. The game is running in a stress test mode, which simulates actual user playback to provide a realistic test scenario than static or less intensive game scenes. The game runs at various frame rates and a 720p resolution using hardware accelerated vs. software video encoding. Variation is common between game titles, this title was chosen to provide a good representative model of real-world 3D game performance.

¹ <https://www.froemling.net/apps/bombsquad>



Typical scene of the game Bombsquad during the density test

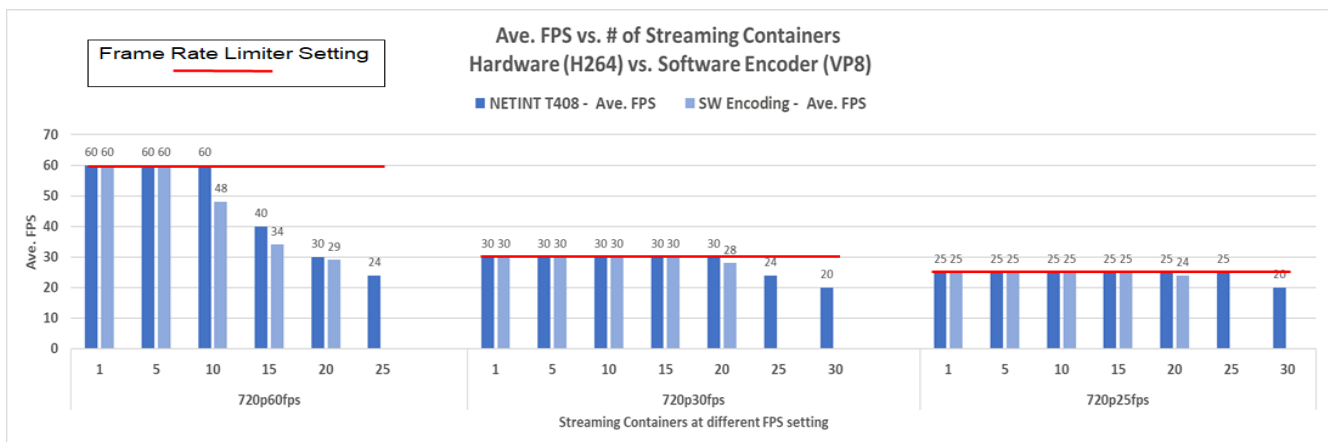
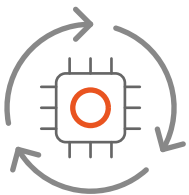


Figure 3: Game Streams 1/container on Anbox Cloud and eMAG + 2 AMD WX5100 GPUs with and without 1 NETINT T408 hardware video encoder

Benefits of using NETINT encoding



Save up to a 60% CPU Cycles



2x Concurrent Streaming

The above figure highlights the benefit of using hardware accelerated video encoding to offload the CPU intensive operations of video streaming. In this case study, the video encoding could use up to 60% of available CPU cycles under high density game stream stress testing. As the density is dialed up, the frame rates of the games significantly degrade until the system saturates. In this figure, no bar indicates the platform was saturated and could not run further instances of the game. Using the Codensity technology from NETINT, the overhead of the encoding is removed from the CPU load delivering roughly 2x performance increase in game density. With the added capacity the platform delivers, more instances per server at more consistent frame rates than the same solution using software based encoding. For details on the full configuration of the system under test see footnote².

Anbox Cloud enables graphic and memory intensive mobile games to be scaled to a vast number of users while retaining the responsiveness and ultra-low latency demanded by gamers, the test was retried with more encoding and GPU hardware to push the platform to its limits. Figure 4 shows the result of adding more video acceleration hardware to the same eMag system to utilise the spare cycles freed up from offloading the encoding of the game streams.

² Server: eMag 32Core 3.3Ghz CPU 256GB DRAM 2667, 2x AMD WX5100 GPUs and 1x NETINT T408 video encoders Software: Anbox Cloud 1.7.3, containers based on Ubuntu 18.04 and Android 10

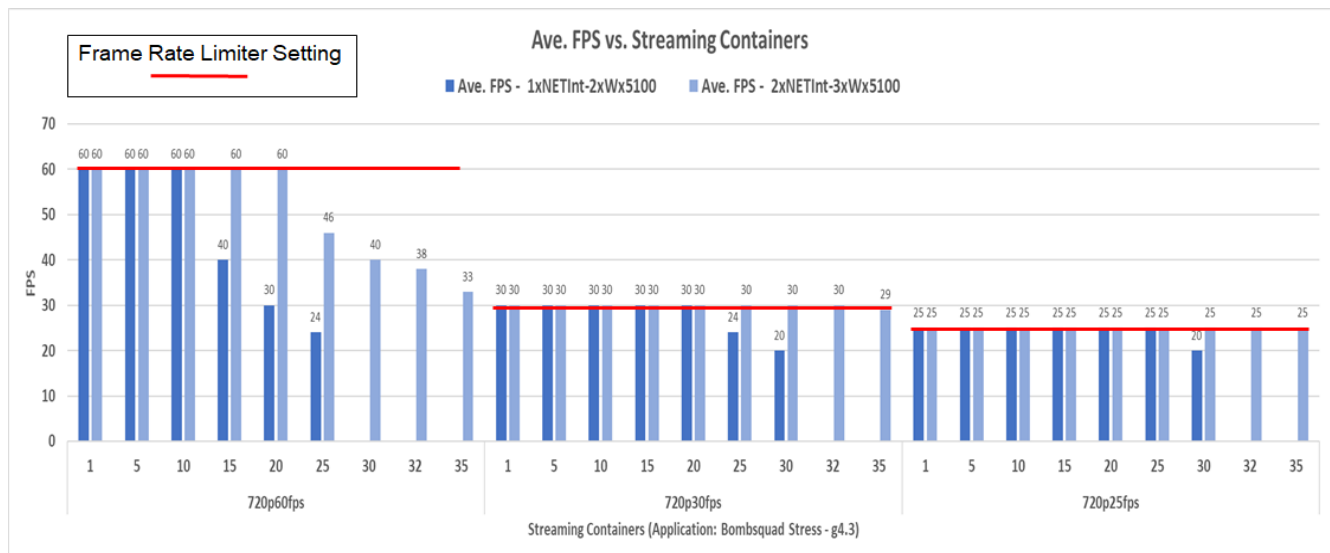


Figure 4: Game streaming - Anbox Cloud on eMAG with 2 vs 3 AMD WX5100 GPUs & 1 vs. 2 NETINT T408 video encoders

As figure 4 illustrates, the value achieved by balancing the system across compute (CPU), video rendering (GPU) and video streaming (Encoders) allows Anbox Cloud to deliver the maximum game instance density for the entire platform reducing the cost per instance of delivering a game service in the cloud. The total cost of ownership (TCO) of the entire platform is a key evaluation criteria for any service provider considering a cloud based service for gaming. The TCO of an Arm based system running native Android instances is by far the most cost effective method of delivering this type of service.

The results obtained in this case study, allow the team to forecast the density of a similar game streaming use case on the forthcoming Ampere Altra platform. The table in figure 5 is derived from projections based on the density of the eMag platform and well known scale factors between the Ampere eMag and Ampere Altra platforms. Based on this analysis, the platform is capable of delivering next generation instance density and a 3x improvement in overall TCO. This type of scalability will fundamentally drive a new era in the cloud gaming market for Android native applications.

System Parameter	Measured	Projected*	
	eMag	Altra 1S	Altra 2S
Cores/Freq (Ghz)	32/3.3	80/3.0	160/3.0
Android Streaming Instances*	32+	80+	160+
GPUs	2-3	3	6+
Video Transcoders Codensity T400 Series	2x T408	1x T432	2x T432

Figure 5: Game Streaming Projection - Anbox Cloud on Altra 1&2 socket systems GPUs and Netint T400 series video transcoders. + Altra: eMag performance projection for this workload is ~4x while 1S:2S scaling is ~1.8x. * Gaming instances are assumed to be 3D games running at 720p and 30 fps nominal

Android and Arm servers: poised for an innovation wave

While games have evolved from consoles and PCs, the mobile gaming market has begun to dominate the industry with more titles and consumers entering the market every year. This is a transformative time in the evolution of mobile applications. In the near future, the touch of an icon on any device can unlock the near limitless capabilities offered by an Android instance running on an Arm native cloud server.

	History	The recent past	Present	The near future
Architecture	Single Board Computers	eMag 32 Cores	Altra 160+ Cores	Altra Max - 256 Cores!
Application Instance Density	1000+	10K+	500K+	Millions
Application Instance Performance	Low	Medium	High	High
Solution Maturity	Inflexible Replica of mobile app	First Generation App ported to Cloud	Scalable Cloud App Built in Cloud	Resilient Cloud App is Transformed

As highlighted in the table above, the industry is on the cusp of an innovation wave. Ampere and Canonical are working together to build the hardware and software that deliver cloud-native environments and dynamic provisioning of Android applications on a stable and highly available Arm native infrastructure. As underlying solutions mature, the applications, whether games or new twists on staid old banking or insurance apps, will begin to employ new and exciting content delivery mechanisms. A fully implemented cloud infrastructure can deliver immense computational power aimed at delivering content with the best possible user experience.

As this innovation wave progresses with significantly more application and remote device support, expanded use cases and new tools for application integration to the cloud become available. New interactive experiences in real-time with unprecedented AI capacity will begin to alter the 'app' as we know it today. At the same time, Altra based platforms will deliver the scale needed inside the data center or at the consumer edge to provide massive compute capacity to serve millions of customers.

All these factors will fundamentally change how the apps are built, delivered and maintained. Cloud infrastructure becomes the basis for data and content delivery. Apps and games designed to be delivered from the cloud will gather content from a variety of sources and AI engines altering content real-time to deliver rich experiences or intelligent data delivery based on myriad conditions. The Android applications of the future will be experienced in AR, VR, autonomous vehicles, smart kiosks, communities and buildings that are context sensitive and able to react or proact with specific users and conditions. Innovative workloads in the cloud will revolutionize how we live, work and play by delivering capabilities beyond our wildest imaginations.

To learn more on how Ampere, Canonical and NETINT Technologies have invented this new future, go to:

- [Ampere Computing](#)
- [Anbox Cloud](#)
- [NETINT Technologies](#)

Resources:

- [Webinar: An introduction to Anbox Cloud](#)
- [Whitepaper: Scale Android applications in the cloud with Anbox Cloud](#)