CANONICAL intel

Cloud gaming for Android: Building a high-performing and scalable platform

How to economically scale Android gaming with Anbox Cloud

January 2020

Executive Summary

To meet the scalability and performance challenges associated with mobile gaming on Android devices--such as lowering latency, maintaining video quality and achieving scalability—Canonical has created Anbox Cloud. This solution is a mobile cloud computing platform for running compute-intensive mobile workloads at scale in the cloud.

This whitepaper introduces Anbox Cloud, a cloud gaming hosting solution for Android games, and how the use of containers and other innovative technologies, including the Intel® Visual Cloud Accelerator Card – Rendering (Intel® VCAC-R), can help service providers deliver a rich mobile gaming experience.

Introduction

The video game industry is changing quickly due to the increasing opportunity and demand to stream games from the cloud, but this poses challenges for video game companies and service providers. Video game companies and game developers are looking for ways to provide a rich experience for gamers while remaining profitable. Such companies want to provide gamers with the experience to play any games, any time, on any device, in any location.

Advances in data centre technologies and the emergence of 5G networks is enabling cloud gaming to become a reality and accessible to more end users. Android is the mobile operating system leader with 76 percent worldwide market share as of September 2019[1]. According to a recent article in TechCrunch, mobile games are on track to reach 60 percent market share of consumer spend in 2019[2]. Several thousands of games have been developed for Android, catering to different types of gamers. With a critical mass building to play the most popular games via the cloud, game server workloads are increasing. This means that gamers may experience a lag in game response time or a service time-out. These issues negatively impact the continuous gaming experience and restrict gaming companies from easily scaling their services.

1 https://gs.statcounter.com/os-market-share/mobile/worldwide
2 https://techcrunch.com/2019/06/11/mobile-games-now-account-for-33-of-installs-10-of-time-and-74-of-consumer-spend/

However, despite proliferation and rapid technology evolution, mobile computing form factors are still constrained. For any given device, battery life is finite, processing power is fixed and data storage capabilities are bounded. These constraints limit the extent to which mobile computing form factors can be seamlessly integrated into our daily activities. One can certainly expect Moore's law to drive computational power increases, but these gains will eventually be capped by the laws of physics.

Computational offloading is an alternative software-defined approach to transcending the constraints of mobile form factors. Computational offloading consists of delegating the execution of compute-intensive workloads to more powerful resources. In this case, mobile cloud computing offloads compute-intensive workloads to centralised hyperscale clouds or edge clouds.
Anbox Cloud is a mobile cloud computing platform that uses Android as an engine in the offloading of mobile workloads through containerisation. High-quality gaming requires advanced hardware performance that is not accessible to most existing mobile devices, despite the fact that gaming enthusiasts increasingly expect this experience on their smartphones. Therefore, offloading compute-intensive workloads such as gaming applications to the cloud, to take advantage of first-class computing capabilities, is an ever-more-attractive opportunity for game developers.

What is the visual cloud?

With visual computing workloads growing at an accelerating pace, cloud service providers (CSPs), communications service providers (CoSPs), and enterprises are rethinking the physical and virtual distribution of compute resources. Visual cloud computing consists of a set of capabilities for remotely consuming content and services that center around the efficient delivery of visual experiences —both live and file-based—as well as applications that add intelligence to video content and tap into machine learning and other artificial intelligence areas, such as object recognition (see Figure 1).

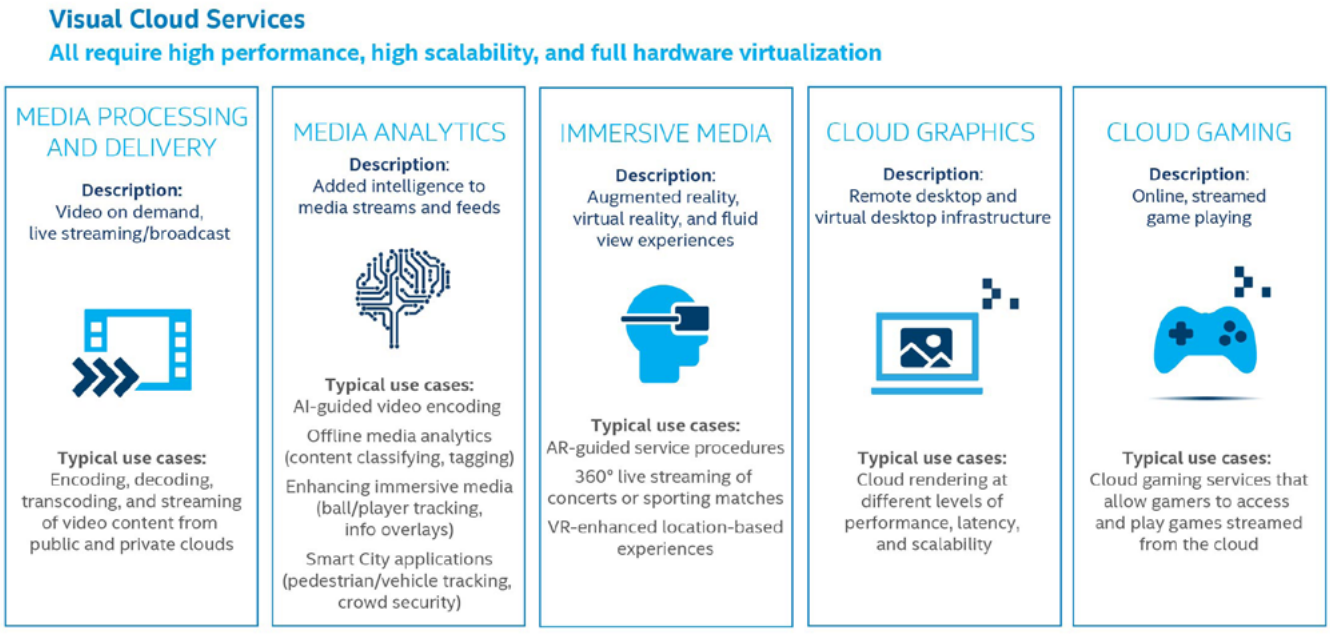

Figure 1. Visual cloud deployments accommodate a diverse range of streaming service workloads.

Challenges in mobile game streaming

Streaming games to mobile devices involves solving a variety of challenges to ensure a positive user experience. Video streaming requires the solution to find the best balance between latency and video quality. On the other hand, a solution needs to ensure it scales easily across multiple regions and brings the server instance as close as possible to the users to avoid unnecessary latency.

Latency and video quality are connected and optimising one will affect the other. Latency is the total delay from the time the player presses a button and then sees a change in a displayed video frame. Every processing and transmission step from the client to server and then back to the client adds some delay. In a game played locally on a PC or gaming console, the delay is caused solely by the CPU and GPU computations. When games are being streamed, network and server responsiveness add to the overall latency. The best video quality would transmit video pictures uncompressed which makes latency worse. Minimising latency, on the other hand, reduces the maximum size of a picture which can be transferred. A game streaming solution must optimise video quality dynamically as network conditions, affecting latency and available bandwidth, change.

The economic viability of game streaming services is also an important challenge. For these services, revenue is typically advertising or subscription-based. However, these sources of revenue are modest at the unit level, especially compared to the sizable cost of operating a game streaming server in the cloud. Furthermore, demand for game streaming is highly variable, which could result in under-utilisation or missed sales in case of server capacity mismatch. Consequently, economically viable game streaming services need to be scalable to spread sizable server and infrastructure costs over many users. Additionally, such services need to be automated in an agile way, to elastically match server capacity and achieve high utilisation of infrastructure.

Anbox Cloud overview

Providing an enjoyable gameplay experience in the cloud requires a platform built for the visual cloud which is capable of handling graphics-intensive workloads, can compress rich media content efficiently and easily scale to accommodate more users without service slowdowns. Anbox Cloud, developed by Canonical, is a mobile cloud computing platform for running Android at hyperscale. It is designed for emerging services in areas such as mobile gaming, telecommunications, advertising, mobile software development and enterprise digital transformation.

Anbox Cloud is a high-density and easy-to-manage containerisation platform that is compatible with the Intel® Visual Cloud Accelerator Card – Rendering (Intel® VCAC-R) to deliver both excellent gaming performance as well as cost-effective unit economics on x86 architectures. The platform can compress rich media content efficiently and easily scale to accommodate more users without service slowdowns.

Building on technologies including Ubuntu, MAAS (Metal-as-a-Service) for remote infrastructure provisioning, Juju for model-driven cloud deployment and LXD for system containers management, Anbox Cloud delivers an easy-to-manage

containersation platform. The long-term enterprise support offering included with Anbox Cloud provides around-the-clock maintenance and continuous security updates, allowing gaming companies to focus on improving the Android applications they deploy at scale in the cloud.

Anbox Cloud reflects the combination of Canonical's deep expertise in cloud-native applications and Intel's industry leadership in providing solutions for 5G, cloud computing and cloud gaming/visual cloud experiences.

How Anbox Cloud works

This section explains the hardware, based on the Intel VCAC-R card, and software solution stack for Anbox Cloud, as well as details regarding orchestration and delivery of containerised Android games in the cloud. Note that the entire stack is performance-optimised for high container density, delivering a highly scalable solution.

The software stack

LXD is at the core of Anbox Cloud's architecture (see Figure 2). LXD, installed in an Ubuntu cloud instance, spins up unprivileged, secure and isolated system containers - themselves running Ubuntu - in the same fashion as a virtual machine (VM). However, system containers have a much lower overhead compared to VMs, thanks to a shared operating system (OS) kernel. This enables a high container density compared to VMs running on top of a hypervisor.

Anbox Cloud runs inside the Ubuntu containers and powers Android (based on the Android Open Source Project (AOSP)). Virtual Android instances contained within system containers can interface with mobile devices via a client application. A custom plugin optimised for specific use cases intermediates signals and data exchanges between the client applications and the emulated Android instances running inside the containers. Device inputs are forwarded to the emulated Android instance, and graphical output is returned to the client to be displayed on the device.

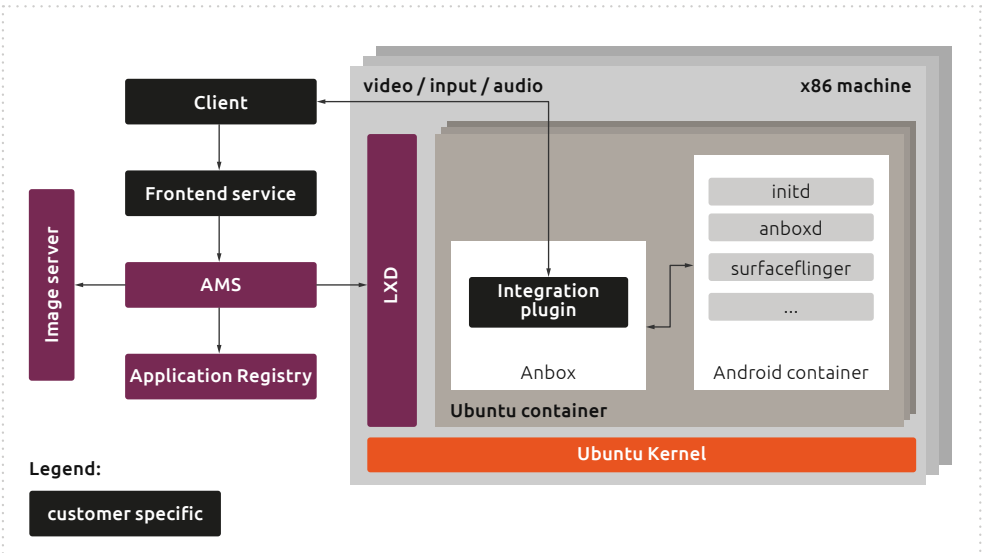

Figure 2. Anbox Cloud solution architecture

CANONICAL intel

The Anbox Management Service (AMS) is the central management component of an Anbox Cloud deployment. It is used to control the containerised Android instances and takes care of resource management and optimal scaleout of the individual containers. A service specific front-end service that can be adapted to various user needs (using a web interface or a simple command-line interface) is deployed along with the AMS. The AMS can pull Android applications, to be run in the virtual instance, from an Application Registry. Images containing Anbox Cloud and Android are pulled from a Canonical-hosted image server, which provides the latest updates and security patches.

The hardware

Online gaming involves both video encoding and image rendering. As shown in Figure 3, the VCAC-R card contains two Intel® Core™ i7-8709G processors, and each system on a chip (SoC) integrates two graphics subsystems: Intel® Ultra-High-Definition (UHD) graphics for low-power, premium media encoding and Radeon™ RX Vega M graphics for rendering gaming and content creation. Both GPUs work in unison to lower the latency as the rendered frames are transferred from the Radeon GPU to the media-optimised Intel GPU for high-quality encoding and streaming. Rendering uses the Mesa driver stack and OpenGL ES 3.1 as shipped with Ubuntu 18.04 LTS; the Intel® Media Driver for Video Acceleration API (VAAPI) is used for encoding the video stream.

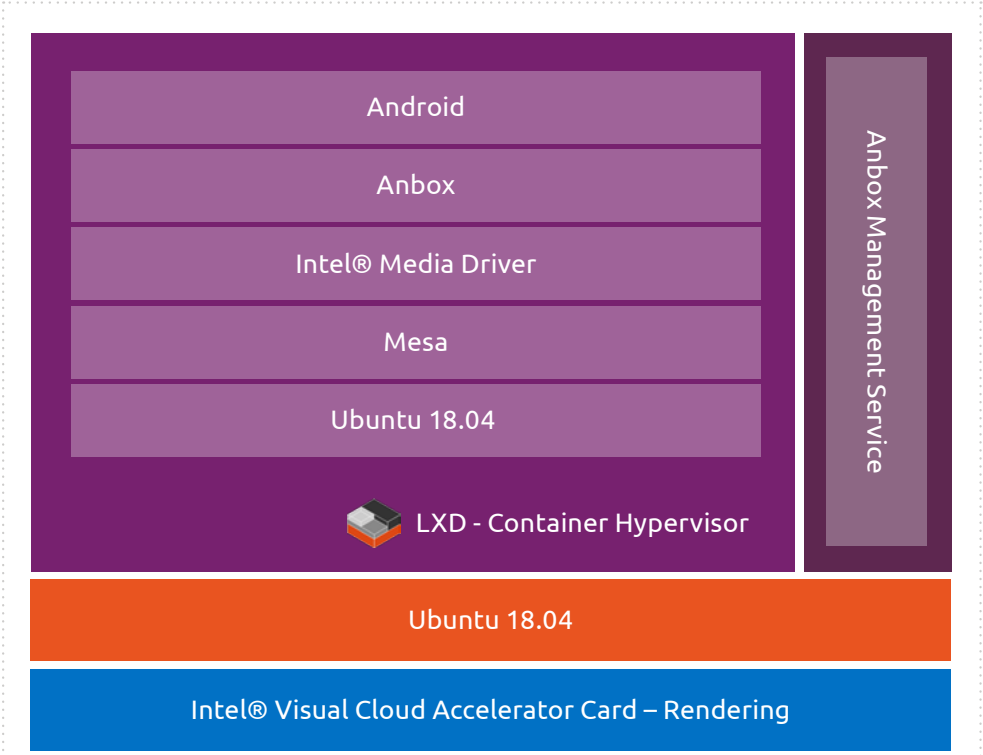

Figure 3. Anbox Cloud is built on the foundation of an Intel® processor specifically built for visual cloud workloads like cloud gaming.

CPU cycles and memory are shared between all Anbox Cloud containers, but CPU-time and maximum-memory usage parameters are used to limit how much of each resource a container is allowed to use. To utilise as many system resources as possible to achieve high density, both the CPU and memory are

overcommitted. Anbox Cloud has advanced monitoring, resource allocation and scheduling logic built-in, to help provide the best use of the available resources. Due to the use of LXD system containers, Anbox Cloud avoids any additional virtualisation overhead, and can allocate extra CPU time to containers that would have been otherwise spent on a hypervisor.

ARM compatibility

Due to its history of being widely deployed on the ARM platform, a large number of Android applications are available only for ARM. This isn't a problem for some use cases, but for those where it is, Anbox Cloud on the Intel VCAC-R card includes binary translation support to allow on-the-fly translation of ARM to x86 instructions. This comes at a small performance overhead, which is acceptable given the other advantages (no hypervisor, full CPU access, and so on). Binary translation opens up Anbox Cloud to run any existing Android application, regardless of the instruction set that its native components use.

Addressing mobile streaming challenges with Anbox Cloud

Canonical, with the integration of Intel's VCAC-R card into Anbox Cloud, has addressed the challenges in streaming mobile games to enable low latency, accelerated graphical processing capabilities and high video quality.

Reducing latency

To reduce the latency of network and streaming responsiveness, Anbox Cloud takes advantage of the capabilities of the Intel VCAC-R card. This card (Figure 4) features dual Intel Core i7-8709G processors with integrated Radeon RX Vega M GH Graphics, providing high frame rates that translate into realistic, smooth-flowing gameplay. This SoC consists of 4 cores/8 threads, 3.1 GHz (up to 4.2 GHz), 8 MB L2 cache, and 24 compute units with a maximum boost frequency of up to 1,190 MHz. Performance is further enhanced by 4G of dedicated high-bandwidth graphics memory (HBM2), which can achieve a memory bandwidth of up to 205 GB per second. Overclocking of the CPU, GPU and HBM2, which is supported by this processor, additionally boosts performance. The result is a capable, cost-effective solution that meets the online gaming requirements of Anbox Cloud[3].

Maintaining video quality

Video quality is another important aspect of the gaming experience. The Intel VCAC-R card provides efficient transcoding, leading to naturally high video quality. But because cloud gaming by definition happens over the internet, video quality must be optimised dynamically as network conditions change. The Anbox Cloud gaming framework applies algorithms that adapt the video quality to best match the current network conditions, by adjusting the video resolution and frame rate when necessary.

Figure 4. The Intel® Visual Cloud Accelerator Card - Rendering provides an excellent gaming experience and cost-efficiency all in one package.

3 Altering clock frequency or voltage may damage or reduce the useful life of the processor and other system components, and may reduce system stability and performance. Product warranties may not apply if the processor is operated beyond its specifications. Check with the manufacturers of system and components for additional details.

Anbox Cloud also uses a technology called frame streaming which helps to protect the user experience from video transmission errors, such as frame skipping and visual artifacts. The combination of Intel technology and Canonical's algorithms help Anbox Cloud to be resilient to errors in the communications channel.

Automation to achieve scale

Canonical's approach to cloud gaming, and Anbox Cloud, focuses on scalability. The key to scalability in an economical and reliable manner is automation. Anbox Cloud has fully abstracted cloud operations, from server hardware provisioning and cloud platform configuration, all the way to application deployment. The intent is to boost the productivity of operations teams, thereby unlocking hyper-scalability for applications, with associated economic benefits. Canonical's MAAS tool can be used to automate server provisioning in a data center for scalability purposes of deploying Android games in the cloud. Building upon automated provisioning, Canonical's Juju, a cloud orchestration tool, can be used to deploy, configure, scale and operate the software infrastructure underlying gaming solutions. Juju can, for instance, take care of launching the LXD container manager, from which system containers will be created to run instances of Android.

The use of specialised Intel® hardware, combined with the scalability of the Anbox Cloud solution, means service providers can offer mobile gamers a great user experience wherever and whenever they want.

Density optimisation for superior economics

Combining Anbox Cloud with the Intel VCAC-R card allows optimised deployments for density (scalability) and computational efficiency. Density is primarily driven by the following key factors:

- CPU load
- Memory capacity
- GPU capacity (rendering and encoding)

As Anbox Cloud opens up access to the CPU without the hypervisor overhead, the extra CPU cycles can be spent on running the actual application workload. To reach optimal distribution and density of containers across an existing deployment, the software components involved must be specifically optimised. From the beginning, Anbox Cloud development has been driven by high density, optimal resource usage and performance, and can achieve its optimal performance when running on hardware platforms like the Intel VCAC-R card.

Benchmarks and additional tests show that it is possible to group 10 to 15 Android containers running in parallel onto the Intel VCAC-R card when encoding the video stream with H.264 at 720p and 30 frames per second (see Figure 5). Testing was conducted with a typical car-racing game for Android[4].

4 Testing by Canonical as of Jan 7, 2020.
Hardware: Hades Canyon NUC – Kaby Lake-G CPU 4 Cores (i7-8809F) together with Radeon™ RX Vega M Graphics 2G/1024 bit HBM memory, 512GB SSD PCIe M.2 512Gb/S (Intel® SSD Pro 7600p Series), 2x 8 GB DDR4-2400 SODIMM, 1x Intel® I219-LM Gigabit LAN
Software: BIOS = HNKBLi70.86A.0054.2019.0214.1350; OS = Ubuntu 18.04.3; kernel = 5.0.0-37-generic #40~18.04.1-Ubuntu SMP; Anbox Cloud 1.3.2; Workload = Beach Buggy Racing 2 (canonical.2019.09.12). This configuration applies to all subsequent performance analyses in this document.

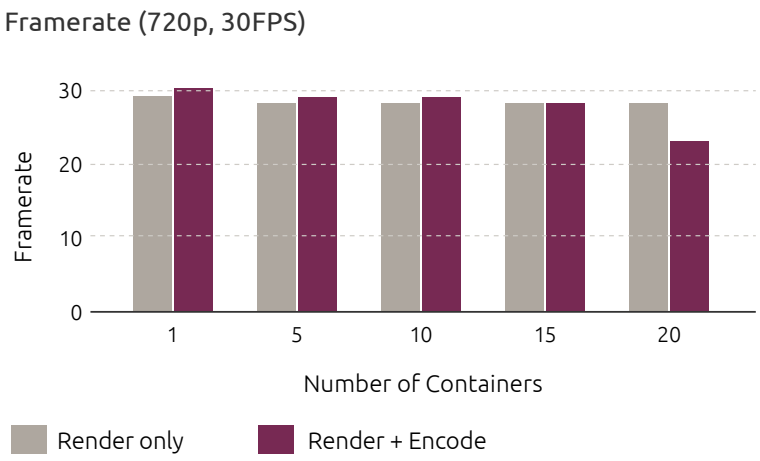

Figure 5. When the framerate is set to 30, Anbox Cloud supports between 10 and 15 Android containers running in parallel on the Intel® Visual Cloud Accelerator Card - Rendering.

Keeping the framerate at 30 has shown the best density results to ensure low latency with the right balance for good video quality. Increasing the framerate to 60 reduces the number of possible containers to run in parallel to 5-10 as shown below in Figure 6.

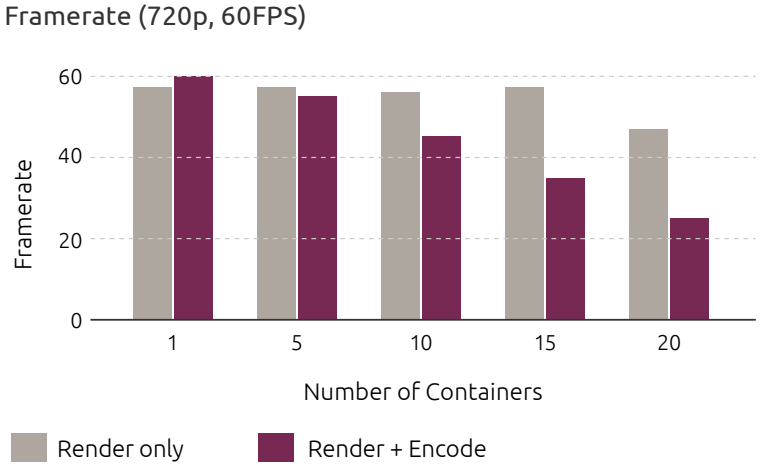

Figure 6. Increasing the frame rate lowers Anbox Cloud container density on the IIntel® Visual Cloud Accelerator Card - Rendering.

Figure 6 shows that at a frame rate of 60, the CPU load starts to increase at 10 containers causing the system to become unreliable, affected by latency spikes and increased visual artifacts in the encoded video. The same applies when the frame rate is set to 30 (Figure 7), but at a higher number of containers - starting around 15 containers - on the same system.

Examining CPU load further confirms that the ideal range is 10 to 15 containers at 30 FPS. When more than 15 containers are hosted, the CPU load starts to increase drastically, due to overloading of the encoder.

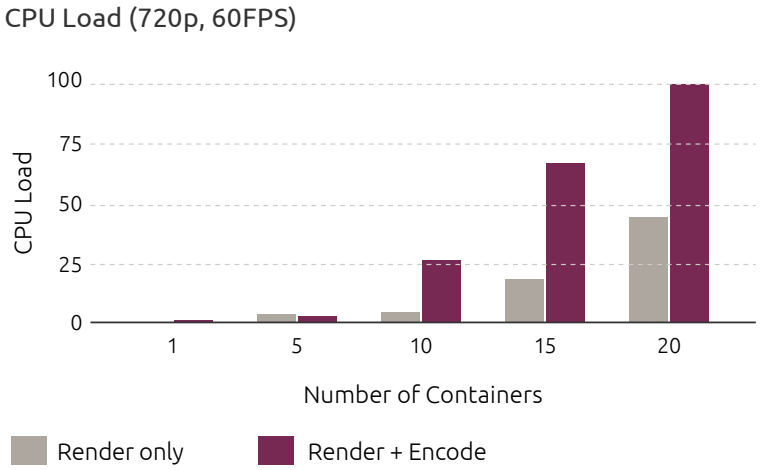

Figure 7. At a framerate of 60, container density of more than 10 containers leads to performance issues such as unreliability and latency spikes.

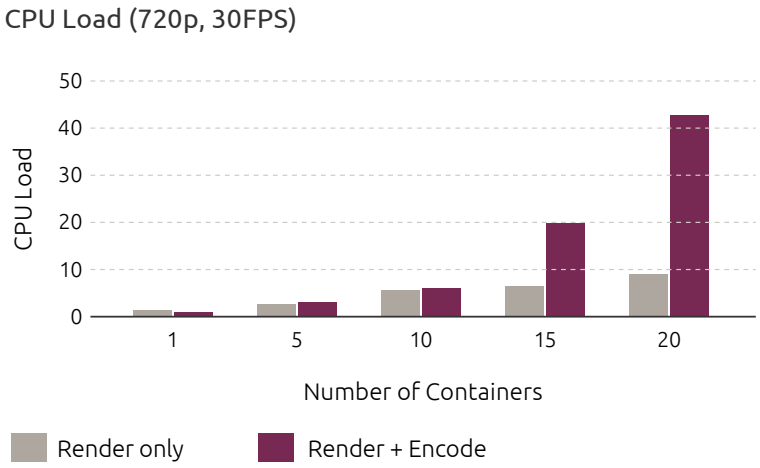

Figure 8. With a framerate of 30, the CPU load starts to affect density at more than 15 containers.

Thanks to overcommitment, the 16GB of memory on the system used for benchmarking is sufficient to run 10 to 15 containers.

Economics and cost structure

As outlined in Table 1, the cost structure for operating Anbox Cloud consists of a mix of fixed and variable costs. Customers will incur fixed costs for licensing, maintenance and long-term support on a per node basis. Additionally, engineering enablement costs associated with integrating Anbox Cloud into the customer's software stack and infrastructure will be incurred. Beyond these fixed costs, variable costs will accrue for cloud usage and operations, based on how much the solution is used (usage fees per hour per node).

	Fixed cost	Variable cost
Direct cost	• License fees per node • Maintenance and support per node	• Cloud usage fees per node per hour
Indirect cost incl. overhead	• Engineering enablement	• Labour cost for operations

Table 1. Fixed and variable costs for Anbox Cloud

Considering this cost structure, the profitability of a commercial solution based on Anbox Cloud can be optimised in two ways:

- **Increasing the unit contribution margins per node, per hour** - A management interface, customised for automating operations pertaining to a particular use case could help improve unit contribution margins, and therefore overall profitability. Another option would be to increase the container density per node (that is, the node utilisation) to spread the hourly cloud usage fees over more containerised Android instances.

- **Reducing the labour cost for operations** - This can be achieved through the automation of operations using Anbox Cloud's management service (AMS).

Anbox Cloud builds upon technology that allows for a higher container density compared to VMs. The cost structure also shows that density is key for profitability margins. The three factors influencing container density are CPU load, memory capacity and GPU capacity. Profitability and user experience will result from clever density optimisation. This can be achieved by choosing the right hardware to match the target workload with the intended rendering performance - taking into consideration the pricing sensitivity of gamers. Finding the optimum combination for these factors and adding a layer of automation is crucial to improve profitability margins and meet service-level agreements (SLAs).

Anbox Cloud - the package

Anbox Cloud is packaged as a complete offering from Canonical for customers wanting to deploy Android at scale and in an economical manner. This offering includes a hardware platform on which to run Anbox Cloud, the option of public or private cloud hosting solutions and long-term maintenance and support.

Canonical offers commercial support for Anbox Cloud via its Ubuntu Advantage support program[5]. The support services that come bundled with Anbox Cloud guarantee SLAs and continuous support for business-critical deployments. Extended Security Maintenance (ESM)[6] helps ensure that Android and the Linux kernel remain patched against security vulnerabilities for up to ten years. These updates are applied through Livepatch[7], which allows for in-situ application of patches without any disruption of operations or downtime. Around-the-clock phone and ticket support complement these services. The commercial support package that is provided with Anbox Cloud ensures a production ready offering that can be relied upon.

5 https://ubuntu.com/support
6 https://ubuntu.com/esm
7 https://ubuntu.com/livepatch

Conclusion

Cloud gaming is growing in popularity, and service providers are challenged to find a solution that can keep up with growth, provide a rich user experience and keep costs affordable.

Anbox Cloud builds upon technologies that allow for a higher container density compared to VMs. The cost structure of the solution also shows that density is key for game publishers' profitability margins. To achieve density optimisation, three factors must be considered: influencing container density (CPU load, memory capacity and GPU capacity), profitability and user experience optimisation. Additional considerations include choosing the right hardware to match the target workload, intended rendering performance and the pricing sensitivity of gamers. Finding the optimum combination for these factors and adding a layer of automation is crucial to improve profitability margins and meet SLAs.

Running Anbox Cloud on Intel® architecture—specifically the VCAC-R card—provides high performance that's economically scalable, due to its inclusion of both discrete and processor graphics, four cores, and a max turbo frequency of 4.10 GHz.

Learn more

- Anbox Cloud
- Canonical Metal as a Service (MAAS)
- Canonical Juju
- Linux Containers (LXD)
- Canonical Extended Security Maintenance (ESM)
- Intel® Visual Cloud Computing
- Intel® Cloud Insider Program

Contact us

Contact Canonical here, or visit https://anbox-cloud.io.

Learn about Intel's Visual Cloud solutions, including white papers, blogs, case studies and videos at www.intel.com/visualcloud.

ubuntu
Delivered by Canonical