

5G Use Case at the Edge Site

CLOUD GAMING





TABLE OF CONTENTS

1.	EXECUTIVE SUMMARY	1
2.	CLOUD GAMING OVERVIEW.....	2
2.1.	INTRODUCTION.....	2
2.2.	KEY TECHNOLOGY.....	3
3.	HARDWARE CONFIGURATION.....	4
3.1.	HARDWARE CONFIGURATION	4
4.	PERFORMANCE VALIDATION.....	7
4.1.	TEST OVERVIEW.....	7
4.1.1	<i>Resolution and FPS.....</i>	<i>7</i>
4.1.2	<i>Game Source.....</i>	<i>7</i>
4.2.	TEST METHODOLOGY.....	7
4.2.1	<i>Measured Value.....</i>	<i>7</i>
4.2.2	<i>Testing Benchmark Tool.....</i>	<i>8</i>
4.2.3	<i>Testing Architecture.....</i>	<i>10</i>
4.2.4	<i>Testing Process.....</i>	<i>10</i>
4.2.5	<i>Test Case</i>	<i>11</i>
4.3.	TEST CONFIGURATION.....	11
4.3.1	<i>Hardware Configuration for Testing.....</i>	<i>11</i>
4.3.2	SOFTWARE CONFIGURATION FOR TESTING	12
4.3.3	<i>Game server VM Configuration</i>	<i>13</i>
4.4.	TEST RESULT	13
5.	CONCLUSION.....	15
6.	REFERENCE.....	16



TABLES

TABLE 1. SERVER SPECIFICATION OF QUANTAGRID D52G-4U.	4
TABLE 2. SERVER SPECIFICATION OF QUANTAGRID D52BV-2U.	5
TABLE 3. SERVER SPECIFICATION OF QUANTAGRID D52Y-2U.	6
TABLE 4. TARGET RESOLUTION AND FPS FOR TESTING.	7
TABLE 5. TEST CASE WITH DIFFERENT GAME SERVER TOOLS.	11
TABLE 6. HARDWARE CONFIGURATION FOR THE SUT.	12
TABLE 7. HARDWARE CONFIGURATION FOR THE DUT.	12
TABLE 8. SOFTWARE CONFIGURATION.	13
TABLE 9. VCPU AND MEMORY CONFIGURATION IN THE SUT.	13
TABLE 10. SERVER PERFORMANCE WITH DIFFERENT RESOLUTION AND FPS.	14

FIGURES

FIGURE 1. CLOUD GAMING SERVICE ARCHITECTURE WITH 5G INFRASTRUCTURE.	2
FIGURE 2. PERCENTAGE OF GPU UTILIZATION WITH NVIDIA MANAGEMENT SOFTWARE.	9
FIGURE 3. UTILIZATION RATES ADOPTED FROM VMWARE MANAGEMENT SOFTWARE.	9
FIGURE 4. TESTING ARCHITECTURE.	10
FIGURE 5. TESTING PROCESS.	11
FIGURE 6. THE NUMBER OF USERS IN THE SUT.	14

1. Executive Summary

The global game market has been growing steadily throughout the last decade. According to Newzoo's Global Games Market Report (2018) [1], Compound Annual Growth Rate (CAGR) for the three segments, mobile games, PC games, and console games is estimated to improve 11% through 2012 to 2021. Researchers diving into CAGRs for these segments have observed that both PC and console segments have steadily grown 3.1% and 2.3%, respectively, while mobile segment has incredibly grown 26.8%. This indicates the behavioral changes in playing games on mobile devices and the demand of gaming anywhere, anytime, on any device, which stimulates the development of cloud gaming.

Cloud gaming is a scenario that adopts cloud computing to deliver game service by allocating compute and storage resources to the cloud rather than on the game console. It exchanges control command and game video between local gaming devices and remote game servers through the internet. However, video stream from game server to game client requires considerably high bandwidth and low latency that legacy mobile network fails to accomplish. Gamers can only rely on home broadband connections, which results in the flexibility limitation of cloud gaming scenario. With the emergence of 5G technology, cloud gaming development is no longer an obstacle. Enhanced mobile broadband (eMBB) provides high-quality video stream while ultra-reliable and low latency communications (URLLC) provides controller action and game video to be streamed in real time. Both 5G characteristics enable gamers to experience an incredible cloud gaming service without owning a high-performance gaming device.

Mobile operators are keen on developing 5G consumer services to increase the average revenue per user (ARPU). The significant difference in the gaming experience is the key contributor for end users' willingness towards 5G network subscription. Mobile operators can secure more subscribers through providing cloud gaming services. On the other hand, cloud gaming service providers can have a strong distribution channel with mobile operators. One of the key successful factors to provide high-quality cloud gaming service depends highly on the leverage of hardware infrastructure, which determines how flexible the service providers can allocate compute and storage resources to gamers. As a leading global datacenter provider, Quanta Cloud Technology (QCT) offers a full spectrum of hardware infrastructure for cloud gaming all the way from data center to edge site.

2. Cloud Gaming Overview

2.1. Introduction

Importance of edge computing in 5G era

With the 5G network development, mobile operators are urged to set up a high-density network and build central offices and edge sites all over the cities and rural areas. The comprehensive network coverage has stimulated the development of edge computing and enabled diverse applications running close to the end users. Among a variety of applications, cloud gaming as a latency-sensitive service is expected to perform exceptional service quality by minimizing the physical distance between game servers and game clients. To ensure the low-latency requirement for cloud gaming service at the edge sites, sufficient computing resources are inevitable to run heavy workloads, such as rendering, capturing, encoding, and streaming.

Leveraging 5G infrastructure at the edge

Along with rapid global 5G rollout, commercial service of 5G network provided to end users is widely available in 378 cities across 34 countries as of January 2020 according to a VIAVI report [2]. Communication service providers (CoSPs) are looking to differentiate offerings to attract more end users to subscribe 5G network services. Among a wide variety of 5G applications, cloud gaming emerges to be in the spotlight and drives more revenue streams through an ultra-low latency network.

To mitigate the physical distance between service servers and end users, CoSP are suggested to leverage a flexible and scalable software-defined infrastructure with a management and orchestration (MANO) framework deployed at the edge sites. Both exceptional networking performance and computing performance are critical in constructing the 5G infrastructure to fulfill different 5G service workloads. Devoted to the telecommunication industry, QCT is experienced in providing agile, scalable, and reliable network infrastructure in the 5G world. The comprehensive 5G infrastructure, covering from data center to edge, provides an optimized cloud computing environment and extraordinary networking capability for cloud gaming services with fast data transmission and game streaming, as shown in Figure 1.

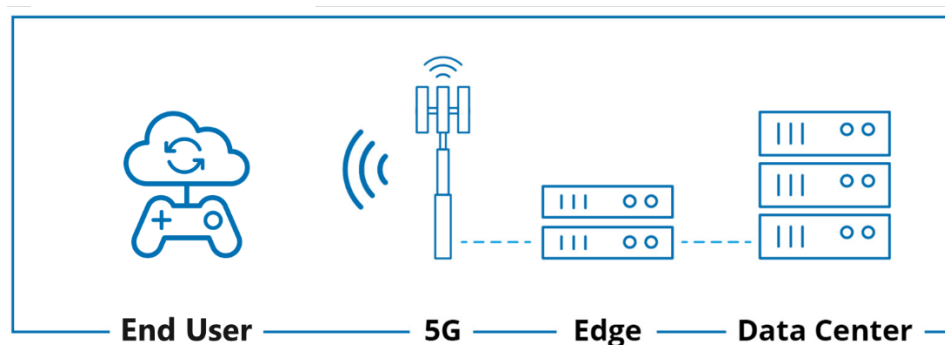


Figure 1. Cloud Gaming Service Architecture with 5G Infrastructure.

On the basis of software-defined network, QCT provides an optimized and commercial-containerized NFVI built on carrier-grade hardware. 5G core network is deployed on this container platform in the data center to achieve the mobile core network evolution, enabling 5G network capabilities of eMBB, URLLC, and mMTC. Apart from the 5G core, the container platform also enables in-housed designed and partners' container network functions (CNFs) such as vBNG and vRAN to ensure network performance with fast data transmission. Nevertheless,

networking capability is not the only factor affecting the latency of data transmission. For some 5G services which require heavy computing capability, such as cloud gaming, the processing time in the cloud computing environment can be a bottleneck of latency. To enhance cloud gaming service quality with high resolution and frame per second (FPS), the hardware infrastructure with strong cloud computing capability is capable of running heavy workloads such as rendering, capturing, streaming, and encoding. To fulfill computing requirement, game servers should be equipped with high-end processor family with high cores and high base frequency as well as the mainstream GPU for cloud gaming.

2.2. Key technology

For cloud gaming use case, rendering, capturing, encoding, and streaming are the most critical processes adopted for running games in cloud computing environment. To implement cloud gaming with low latency, game servers require sufficient computing resources with high-end CPU and GPU to support a great number of users at the same time. Therefore, the performance validation mainly stresses on monitoring the GPU edge server with the following processes and figure out the maximum number of users the server can support.

- Rendering refers to automatically generating a photorealistic image from 3D model using GPU device from vendors such as Intel and NVIDIA. For enhanced visual elements and high-quality gaming image, ray tracing is a rendering technique used to generate incredibly realistic and natural shadowing and lighting effects for gamers.
- Capturing refers to recording the images from the buffering of rendered frames. The accelerators with high-compute capability contribute to maximize the performance of capturing as well as to shorten processing time. Speedy capturing process enables high FPS display, producing smooth performance of gaming video.
- Encoding refers to transforming information to a specific type of encoded data for information processing. The original information converted from one to another can be processed in the server. Common encoding technologies such as H264 (AVC) and H265 (HEVC) are used to transfer the captured frames to specific format by implementing an encoding feature such as NVIDIA NVENC. These encoded data signals are streamed to client devices and displayed after the decoding process in client devices.
- Streaming refers to transmitting the video and audio data to the internet by separating media file into small packages after multimedia data compression. The video frames are transferred into specific format and compressed for streaming. Through video streaming protocol, the encoded data can be streamed to the game clients via the internet.

3. Hardware Configuration

3.1. Hardware Configuration


With the emergence of cloud computing technology, 5G infrastructure provides on-demand availability of resources through the internet. Compute services can be remotely stored and accessed to the cloud or a virtual machine (VM) through the internet-connected devices. By accessing cloud gaming service via 5G end-to-end network, gamers can run cloud games with high performance regardless of local device limitation on compute and storage capacities.

QCT, a global data center and solution provider, is well-known for building x86 servers, storage, switches, and rack system from the data center to edge site. With comprehensive product portfolio, QCT offers 2U/4U GPGPU servers with high density and high performance for cloud gaming. [QuantaGrid D52G-4U](#) and [QuantaGrid D52BV-2U](#) compatible with Intel or NVIDIA GPGPU product family are recommended for heavy workloads in the data center and central office. [Quanta Grid D52Y-2U](#) is the best suited for edge system, which can be closely located to end users and reduce the latency of data transmission.

QuantaGrid D52G-4U

QuantaGrid D52G-4U is the server designed for prevailing over heavy computing challenge with all-in-one GPGPU box. As a purpose-built system, the server supports up to 16 GPUs and run massive workloads with high power efficiency concurrently so that the server can simultaneously support plenty of gamers. Service providers can enjoy a cost-effective hardware infrastructure for cloud gaming. On top of superior computing power, the server featuring with high bandwidth and low-latency networking can expedite communication among GPU nodes. QuantaGrid D52G-4U equipped with two/four additional PCIe slots can provide optional two/four low-latency infiniband or Intel® Omni Path options to implement GPUDirect in the server or RDMA among GPU nodes. With NVLink™, the server can support up to 300GB/s GPU and shorten model-training time frame. Moreover, the server can support up to 8 NVMe SSDs, accelerating both training and inferencing with fast I/O data reading.


Table 1. Server Specification of QuantaGrid D52G-4U.

Product Name: QuantaGrid D52G-4U Dimension: 444x 177.8x 880 (mm)		
Items	Description	Amounts
Processor	Intel® Xeon® Processor Scalable Family	2
Memory	Up to 3TB (128Gx24) of memory for RDIMM/LRDIMM	24
Storage	(24) 2.5" hot-plug SATA/SAS SSD or (20) 2.5" hot-plug SATA/SAS SSD + (4) 2.5" hot-plug NVMe or (8) 2.5" hot-plug NVMe	Up to 24
Expansion Slots	[Option 1] (8) PCIe Gen3 x 16, SXM2 slot (4) PCIe Gen3 x16, LP-MD2 [Option 2] (8) PCIe Gen3 x16, dual-width FHFL + (2) PCIe Gen3 x16, LP-MD2 or (10) PCIe Gen3 x16 dual-width FHFL [Option 3] (20) PCIe Gen3 x16, single-width FHFL or (16) PCIe Gen3 x16, single-width FHFL + (4) PCIe Gen3 x16, LP-MD2	Up to 20
Power Supply	3+1 high efficiency hot-plug 1600W PSU, 80 Plus Platinum (200V-240V)	4

QuantaGrid D52BV-2U

QuantaGrid D52BV-2U is a 2U server tailored to accelerate supercomputing performance based on the compact and optimized architecture. With QCT's system designs, the server breaks hardware configuration limitations in a merely 2U dual-socket system, supporting up to four accelerator cards, 3TBs of memory, and 8 large form factor storage bays. This makes the server a perfect versatile platform for cloud gaming. Besides, equipped with four dual-width accelerator cards and 24 DIMM slots, the server can support cloud gaming with maximized concurrent users.


Table 2. Server Specification of QuantaGrid D52BV-2U.

Product Name: QuantaGrid D52BV-2U Dimension: 440 x 87.5 x 797.9 (mm)		
Items	Description	Amounts
Processor	Intel® Xeon® Processor Scalable Family	2
Memory	Up to 3TB (128Gx24) of memory for RDIMM/LRDIMM	24
Storage	(8) 3.5"/2.5" hot-plug SATA or (8) 3.5"/2.5" hot-plug SATA/SAS or (4) 3.5"/2.5" hot-plug SATA + (4) 2.5" hot-plug NVMe	8
Expansion Slots	SKU1 [8x SATA SKU]: (4) PCIe Gen3 x16 GPU slot (1) PCIe Gen3 x16 OCP 2.0 mezzanine slot or PHY card (2) PCIe Gen3 x 8 LP-MD2 or (1) PCIe Gen3 x16 LP-MD2 SKU2 [8x SATA/SAS SKU]: (4) PCIe Gen3 x16 GPU slot (1) PCIe Gen3 x16 OCP 2.0 mezzanine slot or PHY card (1) PCIe Gen3 x 8 LP-MD2 (1) PCIe Gen3 x 8 LP-MD2 SAS mezzanine slot SKU3 [4x SATA+ 4x NVMe SKU]: (4) PCIe Gen3 x16 GPU slot (1) PCIe Gen3 x16 OCP 2.0 mezzanine slot or PHY card SKU4 [8x SATA SKU]: (4) PCIe Gen3 x16 GPU slot (2) PCIe Gen3 x16 LP-MD2	Up to 7
Power Supply	High efficiency hot-plug 1600W/2200W PSU, 80 Plus Platinum (200V-240V), 1+1 redundant depends on configuration	2

Quanta Grid D52Y-2U

QuantaGrid D52Y-2U is a server with 2 GPGPU cards recommended for cloud gaming at the edge site. Compared to the data center, the rack depth is limited in many locations at the edge site. Edge sites are typically confined to physical space with maximum 600 mm rack depth. The server with 420mm ultra-short chassis fits various environments and offers a small footprint, which is suitable for a variety of media entertainment services happening at the edge site. The NUMA-balanced design with three Ultra-Path Interconnects (UPIs) allows more balance for the I/O transmission, providing optimum computing performance for cloud gaming.

Table 3. Server Specification of QuantaGrid D52Y-2U.

Product Name: QuantaGrid D52Y-2U Dimension: 447.8 x 86.3 x 420 (mm)		
Items	Description	Amounts
Processor	Intel® Xeon® Processor Scalable Family	2
Memory	Up to 2TB (128Gx16) of memory for RDIMM/LRDIMM	16
Storage	(2) 2.5" hot-plug drives	2
Expansion Slots	(2) FHFL PCIe Gen3x16 (1) FHHL PCIe Gen3x16 (1) HHHL PCIe Gen3x16 or (2) HHHL PCIe Gen3x8 or (1) SAS Mezz adapter	Up to 5
Power Supply	1+1 High efficiency redundant hot-plug 1600W PSU	2

4. Performance Validation

4.1. Test Overview

QCT conducted a series of tests to validate the performance of edge server for cloud gaming. To serve more users, the edge server is virtualized into several VMs (hereinafter called game servers). Based on the virtualization architecture, CPU and GPU are virtualized to maximize the resource utilization. The performance of edge server is tested with predefined resolution and FPS. The test results can be provided as a reference for CoSP to ensure the performance of edge server.

4.1.1 Resolution and FPS

Resolution and FPS are the two critical elements commonly discussed for frames rendered in game screen display since these two values can straightforwardly influence gamers' performance and enjoyability. The higher the FPS is, the smoother the streaming can display on the screen. In this test, several common resolution and FPS, including 720P, 1080P, 2K, and 4K with 30 fps/60 fps are set up to be the target to figure out the maximum number of the game servers. The eight test cases are conducted with predefined resolution and FPS, as shown in Table 4.

Table 4. Target Resolution and FPS for Testing.

Test Case	Resolution	Width x Height (px)	FPS
1	720P	1280x720	30
2			60
3	1080P	1920x1080	30
4			60
5	2K	2560x1440	30
6			60
7	4K	3840x2560	30
8			60

4.1.2 Game Source

Cloud gaming has become a new wave in the game market, which emerges diverse cloud game genres to attract a great number of game players. A car-racing game with high-speed genre changes can boost the stimulative and speedy user experiences. Compared with 2D car-racing gaming, 3D car-racing gaming can even bring gamers ultimate immersive experiences. Considering that 3D car-racing game requires the resources with high compute capacity, QCT selected a 3D car-racing game requiring high utilization of 3D GPU to carry out the stress test on QCT server.

4.2. Test Methodology

4.2.1 Measured Value

In this test, the number of game servers is the target value obtained to effectively allocate resources from virtualized hardware and maximize resource utilization on the edge server. Considering smoothness of game screen, FPS is another value used to evaluate the gamer's experience of visual continuity. Besides, the utilization of memory and storage are less likely to influence the gaming performance while the utilization of CPU and GPU

can highly lead to performance bottleneck. Thus, the utilization of CPU and GPU are emphasized for monitoring. The measured values in this test consists of 1) the target FPS, 2) the maximum number of game servers, 3) the percentage of CPU utilization, 4) the percentage of 3D GPU utilization, and 5) the percentage of GPU Encoder utilization in different resolution and FPS.

FPS

Frame per second (FPS), also known as frame rate, indicates the number of frames that the graphics card can render per second or the number of frames that monitor can display per second. This frequency is regarded as a benchmark to display gaming screen, video cameras, etc. Both 30 fps and 60 fps are general benchmarks for acceptable performance used to measure whether the target FPS at a specific resolution is reached in different test cases.

Number of game servers

The number of game servers indicates the maximum number of users that QCT server can support. Based on 3D car-racing game, the number of game servers in different test cases are measured through the benchmark tools when the resources achieve their ultimate utilization.

GPU for 3D

GPU can be adopted as the substitute of CPU to accelerate image rendering, which can free CPU resource to address different sequential serial processing tasks while GPU rendering runs parallel process for assigned applications across multiple cores. GPU rendering process provides on-demand resource for different applications such as graphic processing in gaming and 3D video display, delivering good user experience with smooth game screen. However, with more game servers executed, the utilization rates of 3D GPU can reach to its extreme in the end, as a result, the game fails to run smoothly. Therefore, the utilization rate of 3D GPU is considered to be a critical component for testing server performance.

GPU for Encoder

After capturing the game screen, the encoding technology is adopted to encode the screen into a video format for streaming. The utilization of GPU encoder is closely related to the resolution of game screen. The higher the resolution is, the higher the utilization of GPU encoder is. To accelerate the encoding process, NVIDIA GPU provides NVENC function embedded in Fast Forward Moving Picture Experts Group (FFmpeg) [3] to implement game screen encoding through NVIDIA's in-house API.

CPU

In cloud gaming, CPU can be adopted to execute gaming application, capture game screen, and proceed video streaming. The maximum utilization of CPU stimulates resource competition, which can result in the number of game servers fail to support target FPS in specific resolution and the screen cannot run smoothly. In this test, the CPU is virtualized to different number of vCPUs based on the predefined resolution and FPS.

4.2.2 Testing Benchmark Tool

In this test, QCT adopts several benchmark tools to simulate the implementation of cloud gaming from game server to game client. The tools include game server tools (i.e. QCT Cloud gaming-GSV and OBS), game client tool, and monitor tools (i.e. NVIDIA and VMware management software).

Game server tool

To transmit the video streaming to game client, the two game server tools, including QCT Cloud gaming-GSV and Open Broadcaster Software® (OBS) [4] are adopted to implement capturing, encoding, and video streaming. QCT Cloud gaming-GSV is a tool developed by QCT, which is a cloud gaming streaming software designed based on FFmpeg software and C++ language. The software can be used to capture game screen, encode game screen by H264, and send video streaming to game clients. QCT Cloud gaming-GSV with simplified mode is suitable for testing game screens under or equal to 2K resolution with 30 fps. In order to verify that QCT server can support the game screen with higher resolution and FPS, OBS is popular open source streaming software used to support game screen capturing, HEVC encoding, and video streaming technology.

Game client tool

The game client tool, ffplay [5], is used to simulate the client that receives video streaming information from game servers. The software is a well-known streaming player which can display game screen from game server. The software can not only be adopted to operate the decoding functions through simple command but also support CPU and GPU decoding.

Monitor tool

The two monitor tools, NVIDIA and VMware management software, are used to monitor whether the GPU and CPU resources are efficiently utilized to maximize the number of game servers, as shown in the examples in Figures 2 and 3. In the test, QCT mainly focuses on CPU and GPU resource utilization rather than memory and storage resource utilization since it is found that the utilization of memory and storage remains sufficient regardless of different resolution and FPS. In Figure 2, with the test screen obtained from NVIDIA management software, the two GPUs framed in red indicate the utilization of encoder while the two GPUs framed in yellow indicate the utilization of 3D GPU. In Figure 3, the utilization rates of CPU, memory, and storage are displayed under each bar chart from VMware management software.

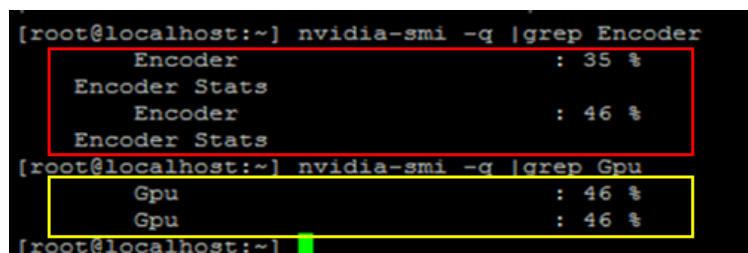


Figure 2. Percentage of GPU Utilization with NVIDIA Management Software.

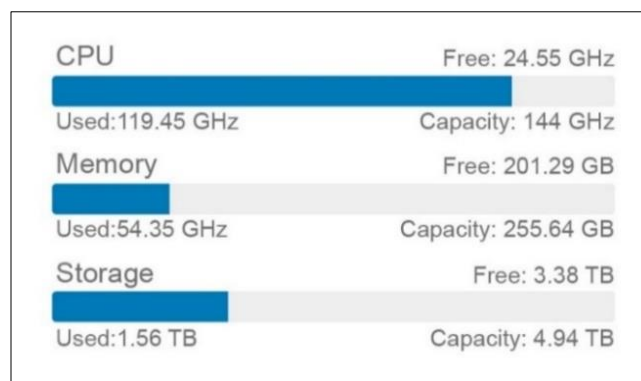


Figure 3. Utilization Rates Adopted from VMware Management Software.

4.2.3 Testing Architecture

The architecture adopts two QCT servers defined as Device Under Test (DUT) and System Under Test (SUT). The SUT and DUT respectively launch their correspondent number of VMs connected through video streaming to be the game servers and game clients. Specifically, a game server is corresponding to a game client. Besides, the SUT is connected to monitor tool using data transfer to monitor GPU and CPU utilization of the SUT, as shown in Figure 4.

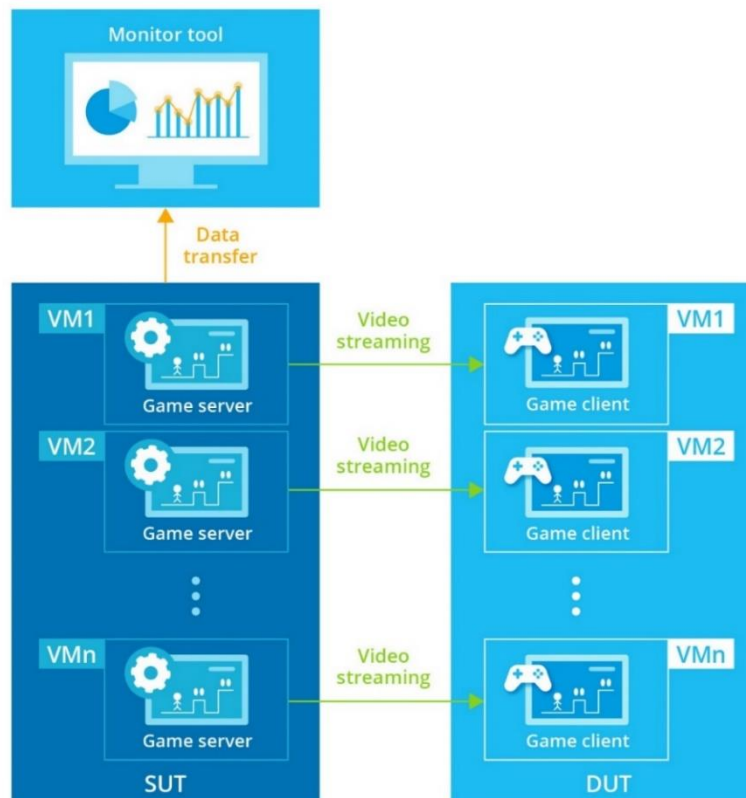


Figure 4. Testing Architecture.

4.2.4 Testing Process

The 3D car-racing game is executed on several game servers with predefined resolution in SUT. When the game servers reach the target FPS (i.e. 30 fps or 60 fps), the resources are supposed to be capable of supporting game servers until the number of game servers reaches to its maximum. In SUT, the game servers repeatedly execute 3D car-racing game, capture game screen, encode the game screen, and send the video streaming to game client. Subsequently, in DUT, the game clients decode and display game screen adopted from game servers through video streaming. During this process, the monitor tools are adopted to monitor and record CPU/GPU utilization and FPS value, as shown in Figure 5.

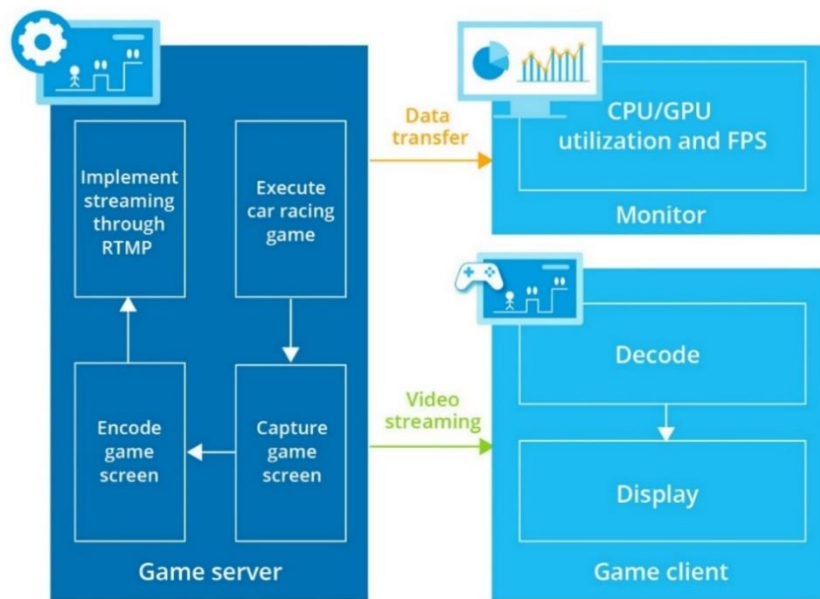


Figure 5. Testing Process.

4.2.5 Test Case

Based on the 16:9 screen ratio, the common resolutions on the screen are 720P, 1080P, 2K, and 4K. The resolution of 720P is so-called standard High Definition (HD), which is composed of 720 pixels of vertical resolution and 1280 pixel of horizontal resolution. Screen resolution above or equal to 720P is defined as high-definition screen. The higher resolution is, the more meticulous and truer the screen displays. In addition to HD, FPS is considered to be an indicator to provide a smooth gaming experience. To produce a continuous visual effect, 24 fps is generally the lowest acceptable value. The higher the FPS is, the smoother gamer experiences. Therefore, 30 fps and 60 fps are set up to be the target FPS for testing. There are eight test cases conducted in this whitepaper based on different resolution and FPS with different game server tools, as shown in Table 5.

Table 5. Test Case with Different Game Server Tools.

Test Case	Resolution and FPS	Game Server Tool
1	720P@30	QCT Cloud gaming-GSV
2	720P@60	
3	1080P@30	
4	1080P@60	
5	2K@30	
6	2K@60	OBS
7	4K@30	
8	4K@60	

4.3. Test Configuration

4.3.1 Hardware Configuration for Testing

Servers for cloud gaming requires to process a great number of compute and storage data, which relies highly on the building of data center. However, building data center requires considerably large space and only small number of data centers can be built in a region. Traditionally, all data should be transmitted to a remote data

center through the network for providing cloud gaming services. To resolve the constraints of space and distance, building edge server is expected to break the limitations beyond the legacy data processing, which can speed up data transmission and greatly reduce data loss and transmission latency.

QuantaGrid D52Y-2U is an edge server with merely 420mm ultra-short chassis, which can be placed extremely close to walls and fits a variety of environments. In addition, the server can support graphic card and accelerator card to implement the functions of 3D GPU and encoder. With up to five PCIe expansion slots, QuantaGrid D52Y-2U can accelerate the compute speed for GPU transmission, ensuring effective data transfer and data processing. Given that powerful CPU and GPU have become indispensable in cloud gaming, the server equipped the CPU with high base frequency and the GPU with high compute power are chosen to be an SUT. Intel® Xeon® Gold 6248R [6] is selected to provide sufficient high-compute power while NVIDIA Quadro RTX 6000 [7] is selected to provide 3D computing module and game screen encoding module, as shown in Table 6. With real-time ray tracing capability, NVIDIA Quadro RTX 6000 can deliver realistic game screen and provides good user experience.

Table 6. Hardware Configuration for the SUT.

Item	Hardware Component	Quantity
Server	QuantaGrid D52Y-2U	1
CPU	Intel® Xeon® Gold 6248R CPU @ 3.00GHz	2
SSD	Intel® S4510 1.92TB SATA SSD	1
RAM	32GB DDR4	6
NIC	Quanta OCP Mezz 82599 10G dual port	1
GPU	NVIDIA Quadro RTX 6000 24GB	2

With the optimized parallel computing architecture, QuantaGrid D52BV-2U can support both Intel® Xeon® scalable process family processor and GPU cards to provide high-performance decoding and high-quality video display. Besides, the server also supports NVIDIA Tesla T4 to enable vGPU and NVIDIA decoder functions. To simulate the user experience of cloud gaming, QuantaGrid D52BV is chosen as a DUT to receive data through video streaming, process decoding, and game screen display after the SUT is established, as shown in Table 7.

Table 7. Hardware Configuration for the DUT.

Item	Hardware Components	Quantity
Server	QuantaGrid D52BV-2U	1
CPU	Intel® Xeon® Gold 6138 CPU @ 2.00GHz	2
SSD	Intel® S4510 1.92TB SATA SSD	1
RAM	32GB DDR4	6
NIC	Quanta OCP Mezz 82599 10G dual port	1
GPU	NVIDIA Tesla T4 16GB	2

4.3.2 Software Configuration for Testing

Server virtualization is the process that abstracts the resources from a physical server, which enables the virtualized platform to run VMs for gamers. By adopting VMware software, an SUT can be virtualized to run several numbers of VMs. Besides, with NVIDIA vGPU technology, GPU can be virtualized to effectively utilize resources. The software configuration is used to run several functions of cloud gaming, including gaming execution, game screen encoding, and video streaming. The software adopted in this test includes VMware vSphere, ESXi, vCenter, and NVIDIA GPU, as shown in Table 8.

Table 8. Software Configuration.

Item	Hardware Components
VMware vSphere®	6.7.0
ESXi™ hypervisor	6.7.0
vCenter Server® Appliance	6.7.0
NVIDIA vGPU	NVIDIA-GRID-vSphere-6.7-430.83-432.33

VMware

VMware is a popular and stable virtualization platform, comprising vSphere and vCenter to manage VM resources, which allows multiple operating systems to run on a single physical computer at the same time. VMware also supports Virtual GPU (vGPU) on virtualization platform, which allows GPU resources to be shared with multiple VMs. VMware can be adopted to deliver highly available, efficient, and centralized infrastructure with flexible and reliable services for cloud gamers.

NVIDIA

NVIDIA featuring Virtual GPU (vGPU) function can be used in VMware virtualization environment to share GPU functions across VMs and address complex graphics. This technology is designed to provide stable performance, flexibly allocate GPU utilization, and ensure the security for any virtual workflow so as to enhance remote users unrivaled experience on any application or device.

4.3.3 Game server VM Configuration

As introduced in section 4.2.5, the eight test cases with predefined resolution and FPS are conducted, including 720P with 30/60 fps, 1080P with 30/60 fps, 2K with 30/60 fps, and 4K with 30/60 fps. The effectiveness of resource allocation can maximize the number of game servers in SUT. The allocation of CPU is closely related to the resolution and FPS factors. As the resolution and FPS of game screen are improved, it is inevitable to increase the resources required. Three different vCPU numbers, comprising 4, 8, and 12 are configured. Since memory resource shows small influential impact on server performance, 8G memory is sufficient for all test cases in the SUT, as shown in Table 9.

Table 9. vCPU and Memory Configuration in the SUT.

No	Test Case	vCPUs	Memory
1	720P@30	4	8G
2	720P@60		
3	1080P@30	8	
4	1080P @60		
5	2K@30	12	
6	2K@60		
7	4K@30		
8	4K@60		

4.4. Test Result

This section shows the maximum number of game servers based on different resolution and FPS using different game server tools (i.e. QCT Cloud gaming-GSV and OBS). The number of game servers is indicated to evaluate how many VMs in SUT can sustain. In Figure 6, the number of game servers is gradually declined when the values of resolution and FPS are increased. The result reveals that since both low resolution and FPS require less compute resource, the SUT can support a greater number of game servers. As the resolution is increased, the number of game servers becomes close, particularly the identical number of game server at 4K resolution.

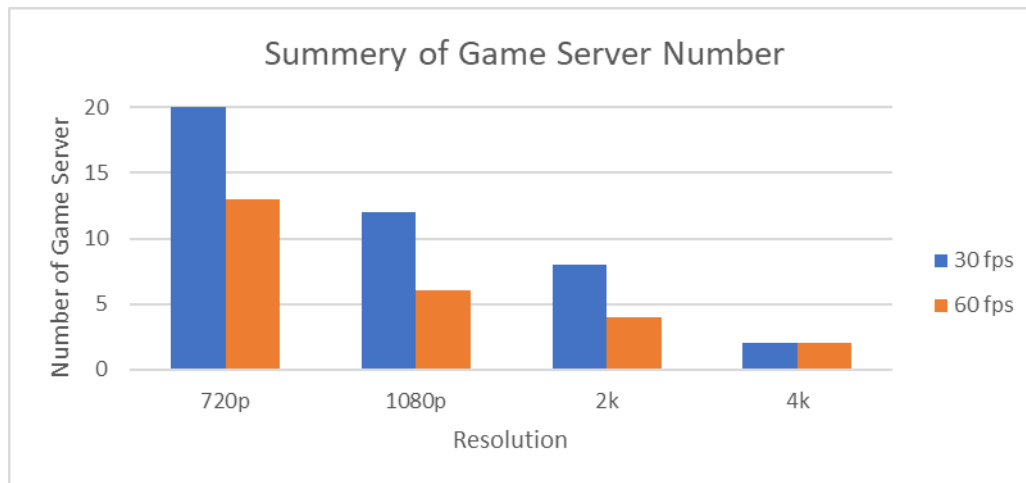


Figure 6. The Number of Users in the SUT.

In Table 10, it is noted that although the number of game servers at 4K resolution is identical when the game screen displays at either 30 fps or 60 fps, the resource utilization still reveals the difference. For example, in the test case of 4K resolution, GPU utilization in encoder is around 50% at 30 fps while the utilization is more than 70% at 60 fps. It is proved that under the same resolution and the number of game servers, the higher FPS can result in higher and busier utilization of server resource. When we deep dive into the results, it is observed that CPU and GPU can mutually influence the performance of edge server. Take 1080P with 30 fps test case as an example, the 3D GPU almost reaches to its maximum utilization rate although CPU still has some resource left. That is, the GPU utilization becomes the bottleneck to prevent the number of game servers from increasing. Conversely, in the 1080P with 60 fps test case, even GPU still has some resource left, the GPU cannot be fully utilized when the utilization of CPU reaches to its extreme.

Table 10. Server performance with Different Resolution and FPS.

Test Case		720P@30	720P@60	1080P@30	1080P@60	2K@30	2K@60	4K@30	4K@60
Number of game servers		20	13	12	6	8	4	2	2
CPU utilization (MAX:144 GHz)		132GHz	126 GHz	114 GHz	123 GHz	104 GHz	31 GHz	48 GHz	22 GHz
3D GPU Utilization	GPU1	99%	92%	97%	38%	100%	93%	56%	46%
	GPU2	99%	99%	99%	42%	86%	91%	61%	64%
GPU Encoder utilization	GPU1	43%	63%	92%	55%	88%	76%	54%	70%
	GPU2	45%	48%	62%	57%	75%	70%	48%	77%

According to the test results, it is found that the server performance can be highly affected by the utilization of GPU and CPU. Considering that GPU and CPU are the two critical elements to maximize the number of game server, the utilization of CPU and GPU requires to achieve balance on the edge server in order to bring gamers optimal user experience. After a series of tests, the results can be provided as useful information with referable number of maximum game servers and a well-validated hardware configuration with the predefined resolution and FPS for cloud gaming.

5. Conclusion

The evolution of mobile network brings a diverse of media entertainment services to the end users via mobile devices. As end users are accustomed to adopting their mobile devices to use entertainment services, particularly for video streaming and cloud gaming, the capability of mobile device has become a bottleneck for service providers to further improve the service quality. Therefore, service providers are moving the computing resource to the cloud computing environment for transformation.

With a comprehensive product portfolio, QCT offers a wide range of servers powered by accelerators including CPU, GPU, VPU, and FPGA, which can provide extraordinary computing performance for cloud gaming. Along with the large-scale 5G network deployment, more services are deployed to the edge sites rather than data center to enable low-latency data transmission to the end-users. To provide latency-sensitive services, cloud gaming services can be deployed to the edge sites to provide high-quality gaming experience. Therefore, QCT has enhanced the computing performance in its edge servers with short chassis depth to fulfill the space requirement at the edge sites.

As a leading telco infrastructure provider, QCT has strong competence in 5G applications enablement and ensure optimized hardware for the service providers. The performance validation test has proven the computing capability for cloud gaming critical workloads such as rendering, encoding, streaming, and decoding under different user scenarios. The validation result shows that QCT edge server can support a smooth cloud gaming experience under different resolutions from 720P to 4K and FPS from 30 fps to 60 fps. In addition, the edge server can support up to 20 users gaming simultaneously. Therefore, CoSPs can enjoy a cost-efficient hardware infrastructure with high stability and computing capability so as to provide good cloud gaming experience to the end users.

6. Reference

[1] Newzoo's Global Games Market Report (2018):

<https://newzoo.com/insights/articles/global-games-market-reaches-137-9-billion-in-2018-mobile-games-take-half/>

[2] VIAVI report:

<https://www.viavisolutions.com/en-us/news-releases/command-5g-network-5g-available-378-cities-globally-according-viavi-report>

[3] Fast Forward Moving Picture Experts Group (FFmpeg):

<https://ffmpeg.org/>

[4] Open Broadcaster Software® (OBS):

<https://obsproject.com/>

[5] ffmpeg:

<https://ffmpeg.org/ffmpeg.html>

[6] Intel® Xeon® Gold 6248R:

<https://ark.intel.com/content/www/us/en/ark/products/199351/intel-xeon-gold-6248r-processor-35-75m-cache-3-00-ghz.html>

[7] NVIDIA Quadro RTX 6000:

<https://www.nvidia.com/en-us/design-visualization/quadro/rtx-6000/>



ABOUT QCT

QCT (Quanta Cloud Technology) is a global datacenter solution provider extending the power of hyperscale datacenter design in standard and open SKUs to all datacenter customers.

Product lines include servers, storage, network switches, integrated rack systems and cloud solutions, all delivering hyperscale efficiency, scalability, reliability, manageability, serviceability and optimized performance for each workload.

QCT offers a full spectrum of datacenter products and services from engineering, integration and optimization to global supply chain support, all under one roof.

The parent of QCT is Quanta Computer Inc., a Fortune Global 500 technology engineering and manufacturing company.

<http://www.QCT.io>



UNITED STATES

QCT LLC., Silicon Valley office
1010 Rincon Circle, San Jose, CA 95131
TOLL-FREE: 1-855-QCT-MUST
TEL: +1-510-270-6111
FAX: +1-510-270-6161
Support: +1-510-270-6216

QCT LLC., Seattle office
13810 SE Eastgate Way, Suite 190, Building 1,
Bellevue, WA 98005
TEL: +1-425-633-1620
FAX: +1-425-633-1621

CHINA

云达科技, 北京办公室 (Quanta Cloud Technology)
北京市朝阳区东大桥路 12 号润诚中心 2 号楼
TEL +86-10-5920-7600
FAX +86-10-5981-7958

云达科技, 杭州办公室 (Quanta Cloud Technology)
浙江省杭州市西湖区古墩路浙商财富中心 4 号楼 501 室
TEL +86-571-2819-8650

JAPAN

Quanta Cloud Technology Japan 株式会社
東京都港区芝大門 2-5-8 芝大門牧田ビル 3F, 105-0012
TEL +81-3-5777-0818
FAX +81-3-5777-0819

GERMANY

Quanta Cloud Technology Germany GmbH
Hamborner Str. 55, 40472 Düsseldorf
TEL +49-211-74077-300

TAIWAN

雲達科技 (Quanta Cloud Technology)
桃園市龜山區文化二路 211 號 1 樓
1F, No. 211 Wenhua 2nd Rd., Guishan Dist., Taoyuan City 33377,
Taiwan
TEL +866-3-286-0707

All specifications and figures are subject to change without prior notice. Actual products may look different from the photos.

QCT, the QCT logo, Rackgo, Quanta, and the Quanta logo are trademarks or registered trademarks of Quanta Computer Inc.

All trademarks and logos are the properties of their representative holders.

Copyright © 2020 Quanta Computer Inc. All rights reserved.