DIGITIMES Research: Complementary report, 2021

Global server shipment forecast, 2022 and beyond

Frank Kung, DIGITIMES Research, September 2021

Introduction

Digitimes Research forecasts a 6.4% increase in 2022 global server shipments, mainly contributed by a double-digit growth in orders from cloud datacenters, IC and component supply gradually returning to normal levels with COVID-19 getting contained around the world and new Intel and AMD CPU platforms spurring some upgrade demand.

In the medium to long term, driven by increasing demand for cloud servers and 5G white-box telecom servers as well as new generations of CPU platforms coming on the market, global server shipments are projected to grow at a CAGR of 6.9% 2021 through 2026.

Going into the next five years, large public cloud service operators will continue to build datacenters while developing hybrid cloud services to expand into the edge computing market. Moreover, with 5G commercialization picking up pace all over the world, telecom carriers are adding core network datacenters. On top of that, O-RAN Alliance strongly promoting Open RAN commercial off the shelf (COTS) devices in place of traditional vendor proprietary solutions will gradually buoy white-box telecom server demand.

On the supply side, chip suppliers including Intel, AMD, Arm and Nvidia aggressively developing and launching new CPUs or accelerators for AI, HPC and edge computing applications will spur CPU demand from cloud datacenters and enterprises.

Chip suppliers are also actively developing open software platforms. For example, Intel highlights that its new CPU launched in 2021 supports the oneAPI toolkit while actively promoting interconnect standards across chips to enhance the performance when CPUs work with AI accelerators through PCIe interfaces in an attempt to compete with Nvidia's robust CUDA ecosystem. Chip suppliers exerting efforts to boost their technology strength is expected to somewhat spur AI and HPC server demand.

Server market

Market trends

Table 1: Server industry trends, 2021-2026

Item	Detail
Demand-side growth drivers	Public cloud service providers continuing to add datacenter infrastructure worldwide
	Cloud service providers developing hybrid cloud boosting edge computing server demand
	O-RAN architecture driving gradual growth in 5G white-box telecom server demand
Supply-side growth drivers	Chip suppliers launching new server CPUs spurring upgrade demand from datacenters and enterprises
	Chip suppliers promoting interconnect standards fueling AI/HPC server shipment growth

Source: Digitimes Research, September 2021

Digitimes Research explores major demand-side and supply-side factors driving server shipment growth 2021 through 2026.

A major demand-side growth driver is large public cloud service providers including Amazon, Microsoft and Google as well as social media like Facebook continuing to build up datacenter infrastructure.

Aside from existing demand, cloud service providers and server brands making continuing efforts toward hybrid cloud or hyper-converged infrastructure (HCI) solutions is expected to spur some edge computing server demand.

In addition to existing core network datacenter demand, telecom carriers will be increasing their purchases of white-box server centralized units (CU) and distributed units (DU) software and hardware to keep up with growingly widespread 5G commercialization around the globe as O-RAN Alliance promotes the Open RAN architecture.

A major supply-side growth driver is CPU and GPU suppliers including Intel, AMD, Arm and Nvidia actively introducing next-generation chips and high-speed interconnect standards.

Intel launching the new Whitley platform in 2021 will spur server upgrade demand and the effect is expected to extend into 2022.

AMD introducing new-generation EPYC CPU solutions will enable it to expand its share in the datacenter and HPC server application markets 2022 and beyond.

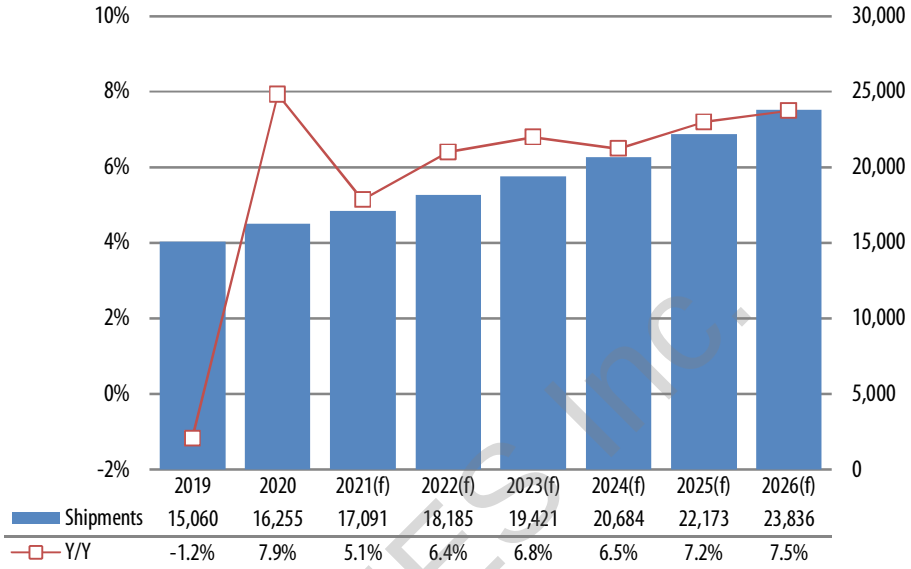
The Arm architectures have a chance of representing 10% of the cloud, HPC and edge computing server market starting 2025 with Amazon and Nvidia aggressively promoting Arm-based solutions.

Chip supplier Ampere has marketed new Arm-based Altra server processors, which are used by server brands (including Gigabyte and Inspur) and China-based cloud service providers (including Tencent and Ucloud). Arm-based processors are a preferred choice in case x86 CPU supply should get disrupted as a result of the US-China trade tensions.

Chip suppliers are actively promoting high-speed interconnect standards, for example, Intel's CXL, AMD's Infinity and Nvidia's NVlink, which enhance the performance when CPUs work with AI accelerators through PCIe interfaces. This is expected to speed up the development of the AI and HPC application server markets.

Global server shipments

Chart 1: Global server shipments, 2019-2026 (k unit)



Note: All figures are based on motherboard shipments.
Source: Digitimes Research, September 2021

Large datacenter demand will remain the main growth driver for server shipments 2021 through 2026. On top of that, 5G white-box telecom datacenter demand will also spur some growth. Along with IC and component supply gradually returning to normal levels, global server shipments are projected to grow at a CAGR of 6.9%.

The four leading North American datacenter operators - Amazon, Microsoft, Google and Facebook – as well as China-based datacenter operators including Baidu, Alibaba, Tencent, ByteDance and BOE continue to purchase servers mainly to provide cloud services and expand their own e-commerce platforms.

Leading chip suppliers keep developing new CPU solutions, which will buoy server shipments 2022 and beyond.

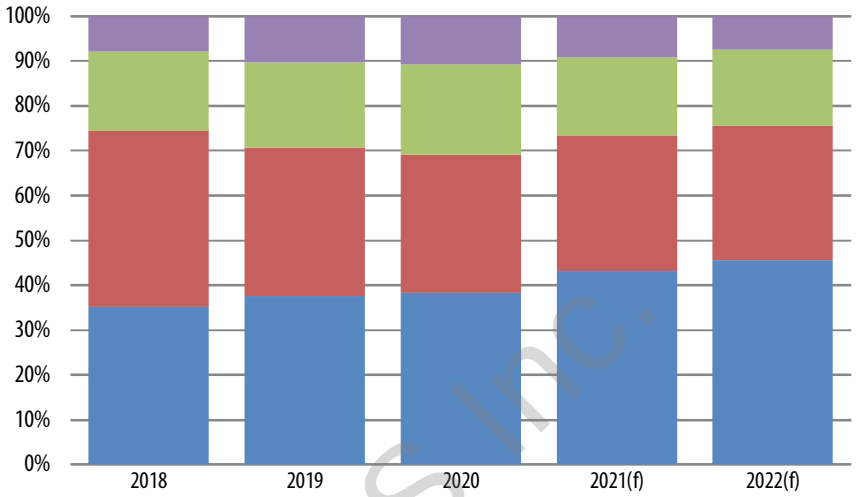
Intel launching the Whitley platform and AMD launching the third-generation EPYC platform in first-half 2021 will spur upgrade demand from cloud datacenters and server brands starting 2022.

Another wave of server upgrade demand will come in 2023 following the debut of Intel's Eagle Stream and AMD's fourth-generation EPYC platforms in second-half 2022.

Nvidia plans to introduce an in-house-developed Arm-based CPU series in 2023, which may further couple with GPU and DPU (Data Processing Unit) to become an integrated solution. This is expected to boost growth in Arm-based servers for cloud HPC and AI applications.

Vendor type

Chart 2: Shipment share by vendor type, 2018-2022 (k unit)



	2018	2019	2020	2021(f)	2022(f)
Others	8.0%	10.3%	10.8%	9.1%	7.5%
China brands	17.5%	19.0%	20.0%	17.5%	16.9%
US brands	39.3%	33.1%	30.9%	30.3%	29.9%
Cloud datacenters	35.2%	37.6%	38.3%	43.1%	45.7%

Source: Digitimes Research, September 2021

The segment of cloud datacenter servers of the global server market shows the strongest growth momentum, which drives white-box server shipments as well. White-box (cloud datacenters) servers represented 35% of the global market in 2018 but their share is estimated to increase to nearly 46% by 2022.

In 2022, the combined server demand from large datacenter operators including Amazon, Microsoft and Google, which continue to expand their datacenters, will grow 14% from the prior year.

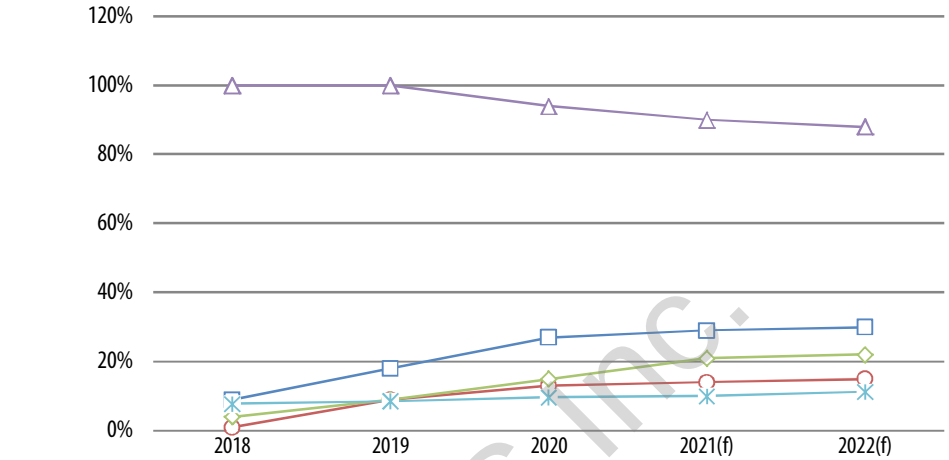
US-based server brands including Dell and HPE will experience a 3% to 4% decline in their server shipments in 2021. Their 2022 server shipments may return to single-digit growth. However, their share among the global server shipments is on the decline year after year from 40% in 2018 to below 30% in 2022.

China-based server brands including Inspur, Lenovo and Huawei benefit from state policy support for datacenter build-up. Their combined share among the global server shipments grew from 19% in 2018 to 20% in 2020 but is expected to fall back to 18% in 2021 and further to 17% in 2022.

This is mainly due to an estimated 34% annual decline in Huawei's 2021 shipments and a double-digit annual decline in its 2022 shipments as the US government prohibits Huawei from gaining access to components amid the trade tensions with China.

In-house production

Chart 3: In-house production shares of vendors, 2018-2022



	2018	2019	2020	2021(f)	2022(f)
Supermicro	9.0%	18.0%	27.0%	29.0%	30.0%
Inspur	1.0%	9.0%	13.0%	14.0%	15.0%
Lenovo	4.0%	9.0%	15.0%	21.0%	22.0%
Huawei	100.0%	100.0%	94.0%	90.0%	88.0%
Overall	7.8%	8.5%	9.7%	10.0%	11.2%

**Note: Overall share includes the four vendors and others with in-house productions.*
Source: Digitimes Research, September 2021

Server brands that engage in in-house production include Supermicro, Inspur and Lenovo. Their ratio of in-house production will continue to increase going forward. It is estimated that more than 11% of servers on the global market will be produced in-house by 2022.

Supermicro will keep expanding capacity at its Bade plant, producing high-end servers. Those made through OEM production by Wistron, OSE and USI are mostly low-end models.

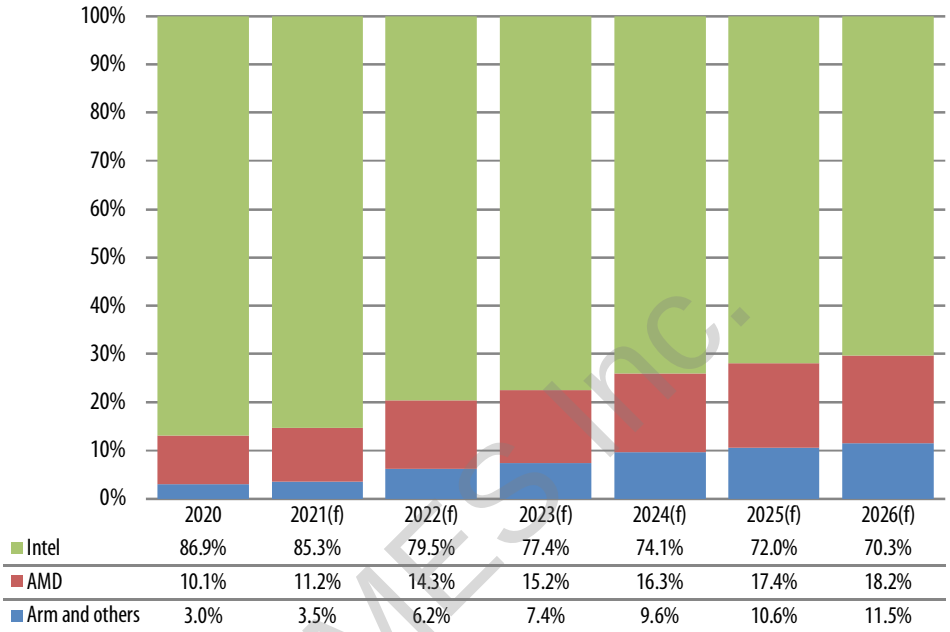
Lenovo will pull together corporate resources and increasingly charge its subsidiary LCFC with parts management, motherboard production and whole-system design.

With its Suzhou plant in operation starting mid-2019, Inspur is expected to ramp up in-house production, whole-system design and motherboard production going forward. Its main customers include datacenter operators Baidu, Alibaba, Tencent and ByteDance in addition to Apple and other enterprises.

Subject to the US government's sanction, Huawei has difficulty obtaining critical components (such as processors and DRAM) so it has changed from 100% in-house production to growingly outsourcing production to OEMs in China and Taiwan.

CPUs/GPUs

Chart 4: Shipment share by CPU, 2020-2026



	2020	2021(f)	2022(f)	2023(f)	2024(f)	2025(f)	2026(f)
Intel	86.9%	85.3%	79.5%	77.4%	74.1%	72.0%	70.3%
AMD	10.1%	11.2%	14.3%	15.2%	16.3%	17.4%	18.2%
Arm and others	3.0%	3.5%	6.2%	7.4%	9.6%	10.6%	11.5%

Source: Digitimes Research, September 2021

Driven by Amazon and Nvidia, Arm-based server CPUs are estimated to represent a more than 10% market share by 2025.

Among large datacenter operators, Amazon undertakes the most aggressive planning toward Arm-based server CPUs. Aside from continuingly expanding its EC2 public cloud service scope, the slimmed-down version of Outposts to debut at year-end 2021 will integrate its self-developed Graviton CPU as part of its efforts to further grow its footprint in the enterprise edge cloud market.

In April 2021, Nvidia announced its plan to introduce an in-house-developed Arm-based server CPU in 2023, which may further couple with GPU and DPU (Data Processing Unit) to become an integrated solution, as part of its efforts to expand into cloud HPC and AI server markets. If Nvidia's Arm-based server CPU enters volume production according to schedule, the market share of Arm-based CPU will show a more significant growth starting 2024.

AMD's EPYC processors feature high computing performance, core count and cost-effectiveness. These compelling advantages will enable their presence in cloud datacenter and HPC application server markets to grow. Their market share is expected to reach 18% by 2026.

In the face of the two strong rivals, Intel may see its market share shrink to 70% by 2026. It is expected that Intel will adjust by expanding its business scope to include AI chips (such as GPU, FPGA and IPU) in order to grow its influence on the AI and HPC server markets.

Table 2: Server CPU roadmaps by supplier, 2020-2022

Supplier	2020	2021(f)	2022(f)
Intel	Cedar Island: support PCIe 3.0 and DDR4	Whitley: support PCIe 4.0 and DDR4	Eagle Stream: support PCIe 5.0 and DDR5
AMD	Second-generation EPYC platform: support PCIe 4.0 and DDR4	Third-generation EPYC platform: datacenters and brand vendors expand adoption.	Fourth-generation EPYC platform: support PCIe 5.0 and DDR5
Arm	Neoverse N1: support PCIe 4.0 and DDR4	Neoverse V1 and N2: support PCIe 5.0 and DDR5	Poseidon: support PCIe 5.0 or 6.0 and DDR5

Source: Digitimes Research, September 2021

Leading chip suppliers including Intel, AMD and Arm will continue to launch new-generation CPU platforms. Although the platforms may replace one another, they will still boost the total volume of server upgrades by cloud datacenters and enterprises.

Intel having unveiled its Whitley platform in first-half 2021 is expected to further hike the volume of server upgrades second-half 2021 through first-half 2022.

Intel's Eagle Stream to debut after second-half 2022 will support PCIe 5.0 and DDR5.

With Whitley and Eagle Stream being launched close to each other and supporting different generations of the PCIe protocol and memory spec, users may not be as eager to upgrade their servers as before. They are more likely to replace some of their servers in consideration of their HPC or AI needs.

Following the launch of its third-generation EPYC platform in 2021, AMD plans to introduce the fourth-generation EPYC in second-half 2022, which is expected to be increasingly adopted by cloud datacenter operators including Microsoft and Google as well as server brands including Supermicro and Gigabyte.

AMD's second-generation 64-core EPYC debuted in 2019 successfully expanded presence in the large datacenter and HPC markets as it got adopted by the top-3 US-based datacenter operators and Alibaba Cloud.

Arm introduced the Neoverse N1 in 2020, which was used by Ampere and Amazon.

Amazon independently developed its Arm-based Graviton processors to keep expanding its EC2 cloud service applications.

Ampere's Arm-based Altra processors are used by China-based Tencent, Inspur and Ucloud as well as Gigabyte.

Ampere plans to launch 128-core Altra Max based on Neoverse N2 by year-end 2021. It has delivered samples to Gigabyte for testing.

CPU/GPU suppliers partnerships

Based on Digitimes Research's observation, chip suppliers including Intel, AMD, Arm and Nvidia making three-tiered planning toward next-generation processors, open software platforms and interconnect standards will strengthen the global HPC and AI server industry ecosystems.

Chip suppliers continue to exert efforts toward next-generation CPU and AI processors.

Intel unveiled its Infrastructure Processing Unit (IPU) in mid-2021, which is a SmartNIC solution similar to Nvidia's BlueField DPU. It is designed to enable cloud datacenter operators to offload infrastructure overhead from the CPU and accelerate network virtualization and security management.

Further to the launch of new-generation CPU and GPU, if everything goes well with AMD's acquisition of leading FPGA supplier Xilinx announced in 2020, it will help AMD expand its presence in AI cloud service and 5G telecom server application markets.

Scheduled to debut in 2023, Nvidia's Arm-based CPU will first target applications in super computers and cloud datacenters.

Chip suppliers make active efforts toward open software toolkits, aimed to offer software development kits (SDK) and application programming interfaces (API) that enable software development across CPU, AI accelerators and AI software framework (TensorFlow and PyTorch) to increase customer interest and adoption.

When launching the Whitley platform in 2021, Intel highlighted it is highly compatible with the oneAPI toolkit as a critical part of its strategy for the Xe GPU. Its intention is to grab a share of the AI ecosystem market that Nvidia has built up with CUDA.

AMD's ROCm and Arm's Arm NN are open software platforms aimed at strengthening the interconnect performance of AI and HPC application software layers.

Chip suppliers make efforts toward industry-wide interconnect standards for AI chips in an attempt to strengthen the AI server market ecosystem.

Intel promotes Compute Express Link (CXL) for the purpose of securing a share of the CPU and PCIe interconnect application market. The new Eagle Stream CPU platform to debut after 2022 is expected to support the higher-speed PCIe 5.0 standard.

Nvidia and AMD work on boosting the interconnect performance between CPU and GPU, respectively introducing their own interconnect interface standard NVlink and Infinity Fabric.

Arm's Nerverse N2 introduced in second-quarter 2021 also supports CXL 2.0 and CCIX 2.0. Its Poseidon server CPU platform to launch after 2022 will support the next-generation CXL and CCIX to enhance its interconnect performance with x86 server CPUs.

5G

With increasing global 5G commercialization, telecom carriers around the world are building 5G datacenters, which will spur growing demand for core network servers and edge network servers.

O-RAN Alliance strongly promoting Open RAN commercial off the shelf (COTS) devices in place of traditional vendor proprietary telecom equipment will drive white-box telecom server shipment growth.

According to a 2021 Dell'Oro report, worldwide investments into Open RAN hardware and software services is projected grow at a double-digit CAGR 2020 through 2025 to approach US$10 billion by 2025.

ABI Research forecasts the revenue generated from Open RAN hardware and software for public networks will reach US$40.7 billion and for enterprise networks US$7.6 billion by 2026. The Open RAN architecture will represent more than 50% of the market by 2028.

Telecom carriers that have announced their plans to build 5G networks based on the Open RAN architecture are mostly in North America and Europe, as well as Japan and South Korea in Asia.

Digitimes Research expects the Open RAN architecture to follow the Open Compute Project's (OCP) development model, wherein white-box server vendors will design, manufacture and directly supply standardized Open RAN servers or customized servers to telecom carriers, bypassing traditional telecom equipment providers or system integrators (SI).

Large-scale conglomerates such as Quanta, Wistron and Foxconn have better potential at becoming white-box server vendors. They can build up technological experience and influence in white-box telecom servers by pulling together their manufacturing, datacenter network equipment and system integration resources.

Cloud datacenter operators

Table 3: Cloud datacenter operator development roadmap, 2018-2022

Operators	2018	2019-2020	2021-2022(f)
AWS	In-house developed Graviton; implemented into EC2 public cloud	Pushing second-gen Graviton; expanding EC2 applications	Integrating Graviton and Nvidia GPUs to expand cloud AI services
	Pushing Outposts 42U hybrid cloud	Pushing Local Zones and Wavelength	Will push 1U (in-house developed CPUs) and 2U (Intel CPUs) Outposts services
Google Cloud	Releasing third-gen in-house developed TPU v3	Releasing Anthos hybrid cloud platform and partnering with server brands to push cross-cloud platform applications	Will push Anthos for Telecom and GMEC platform and will release TPU v4 in 2H21
Microsoft Azure		Releasing Azure Arc with Azure Stack Hub and HCI	Will in-house develop an Arm processor used in Azure cloud or notebook products

Source: Digitimes Research, September 2021

With growing dependence on diverse cloud services, enterprises have transitioned from a single cloud platform to multi-cloud combining several public cloud or hybrid cloud services.

According to a Flexera report released in March 2021, 92% of enterprises have a multi-cloud strategy and 78% have a hybrid cloud strategy, representing the largest share.

Amazon delivered the AWS Outposts hybrid cloud service as a 42U rack in 2018, based on which it further introduced AWS Local Zones for regional datacenters and worked with telecom carriers to promote AWS Wavelength for 5G edge telecom application services in 2019.

Going into 2021, AWS Wavelength is deployed on networks of Verizon, KDDI and SK Telecom in the US, Japan and South Korea. Partnerships with additional telecom carriers such as Vodafone are to follow 2022 and onwards.

Amazon plans to unveil a slimmed-down version of Outposts in 1U and 2U form factors at year-end 2021. It stands a chance of further expanding enterprise edge server applications for mini edge computing datacenters and telecom service deployment, for example.

The 1U form factor features Amazon's self-developed Arm-based CPU while the 2U form factor uses an Intel solution.

Microsoft and Google may follow suit in the case of successful Outposts expansion. Aside from working with server brands, they may also engage in in-house server CPU development to reduce server purchase costs.

Amazon's self-developed Graviton processors based on Arm architectures are estimated to enable 20% to 30% saving compared to the costs of x86 chips.

Microsoft disclosed its plan to develop its own Arm-based processors in 2021. If successfully developed, they will be used in Azure cloud datacenters for application in Internet searches, storage or machine learning while spurring server demand.

Google unveiled the Anthos hybrid cloud in 2019, followed by Anthos for Telecom and Global Mobile Edge Cloud (GMEC) in 2020, targeting telecom carriers' edge cloud applications.

Google and Ericsson partnered to complete 5G network testing and deployment on Anthos in August 2021, which is expected to accelerate autonomous driving, traffic and industrial applications at the edge.