DIGITIMES Research: Complementary report, 2022

Global server shipment forecast, 2023 and beyond

Frank Kung, DIGITIMES Research, September 2022

Introduction

Digitimes Research forecasts a 5.2% increase in 2023 global server shipments, mainly driven by large US-based cloud service providers actively expanding their datacenter infrastructure. On top of that, the IC and component supply will gradually return to normal with COVID-19 becoming like the flu and Intel as well as AMD producing new-generation CPUs in volume is expected to spur some upgrade demand. In the mid-to-long term, driven by growing cloud, HPC and edge server demand and continuing launches of new CPUs by chip suppliers, global server shipments are projected to grow at a CAGR of 6.1% from 2022 through 2027.

Going into the next five years, global server shipment growth will still mainly be fueled by large public cloud service providers continuing to expand their datacenter infrastructure worldwide to satisfy cloud service needs as well as demand for media streaming platforms, e-commerce and other online services. Moreover, enterprise users having to process a rapidly growing amount of data on the cloud and at the edge for a variety of application scenarios is also boosting HPC and AI server demand.

Cloud service providers and server brands are set to shift their focus toward supporting enterprise hybrid cloud services closer to the edge to meet edge AI inferencing or instant massive data processing or storage needs. Furthermore, 5G datacenter infrastructure being built up all over the world will drive core and edge telecom server development. On top of that, O-RAN Alliance strongly promotes the use of Open RAN commercial off-the-shelf (COTS) devices as edge RAN equipment will gradually buoy white-box telecom server shipments from 2023 onward.

The main supply-side factor fueling global server shipments will be leading chip suppliers continuing to launch new CPUs or AI accelerators, which will spur server purchases or upgrades by cloud datacenter operators and enterprises. More than that, chip suppliers are also actively developing next-generation accelerated data processing solutions such as Nvidia's and AMD's Data Processing Unit (DPU) as well as Intel's Infrastructure Processing Unit (IPU). These are expected to drive datacenters toward 400Gb/s data transmission, thus elevating their computing, networking, storage and security control performances.

Server market

Market trends

Table 1: Server industry trends, 2022-2027

Item	Detail
Demand-side growth drivers	Public cloud service providers adding datacenter infrastructure worldwide
	Growing demand for the computing performance of HPC/AI servers
	Edge AIoT and telecom applications buoying enterprise hybrid cloud demand
Supply-side growth drivers	Chip suppliers launching new solutions spurring HPC/AI server market growth
	DPU raising cloud and HPC server performance

Source: Digitimes Research, September 2022

On the demand side, large public cloud service providers such as Amazon, Microsoft and Google will continue to expand datacenter infrastructure worldwide to meet the needs for cloud services, media platforms, e-commerce and remote collaboration.

Rapidly increasing needs to process massive data and accelerate AI computing for different application scenarios will boost HPC server shipment growth.

The potential growth drivers for edge servers will include enterprise users' AIoT computing and storage applications as well as increasing purchases of white-box server centralized units (CU) and distributed units (DU) on the part of telecom carriers with O-RAN Alliance aggressively promoting the Open RAN architecture.

On the supply side, leading chip suppliers including Intel, AMD and Arm actively introducing next-generation CPUs will drive datacenter operators and server brands to replace old servers or buy new ones based on their actual HPC and AI needs.

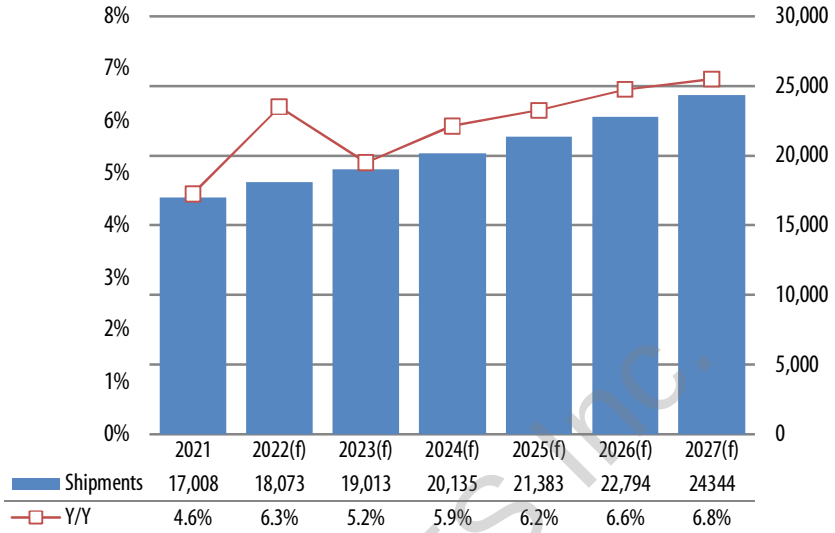
AMD and Arm are expected to aggressively expand their footprint in the server CPU market, poised to market next-generation CPUs with enhanced performance and cost-performance (C/P) ratios or competitive new solutions targeting niche segments every year. This has a chance of speeding up server upgrades by some cloud service providers and enterprises.

Nvidia, Intel and AMD scrambling to develop accelerator chips for datacenters will help push cloud and HPC server performance higher.

Nvidia plans to unveil its next-generation DPU BlueField-4 in 2024, supporting up to 800Gb/s data transmission to hoist datacenter storage and networking performance.

Global server shipments

Chart 1: Global server shipments, 2021-2027 (k unit)



Note: All figures are based on motherboard shipments.
Source: Digitimes Research, September 2022

The demand from large US-based cloud datacenter operators is expected to remain the major growth driver for server shipments from 2022 through 2027. On top of that, cloud service providers and server brands will be scrambling for hybrid cloud developments targeting different usage scenarios driven by edge AIoT and 5G telecom applications. As such, global server shipments are projected to grow at a CAGR of 6.1%.

Large US-based datacenter operators including Amazon, Microsoft, Google and Meta will still be the main pillars supporting the server market growth going into 2023. Their server purchases will be made to satisfy the needs of their cloud services or media platforms. With the problem of mixed shortage levels of some ICs and components mitigating, 2023 server shipments are projected to grow by 5.2%.

The shipments of Intel's next-generation Eagle Stream chips and AMD's fourth-generation EPYC platforms (codenamed Genoa) are set to ramp up in the first half of 2023, which is expected to spur some server upgrade demand in 2023.

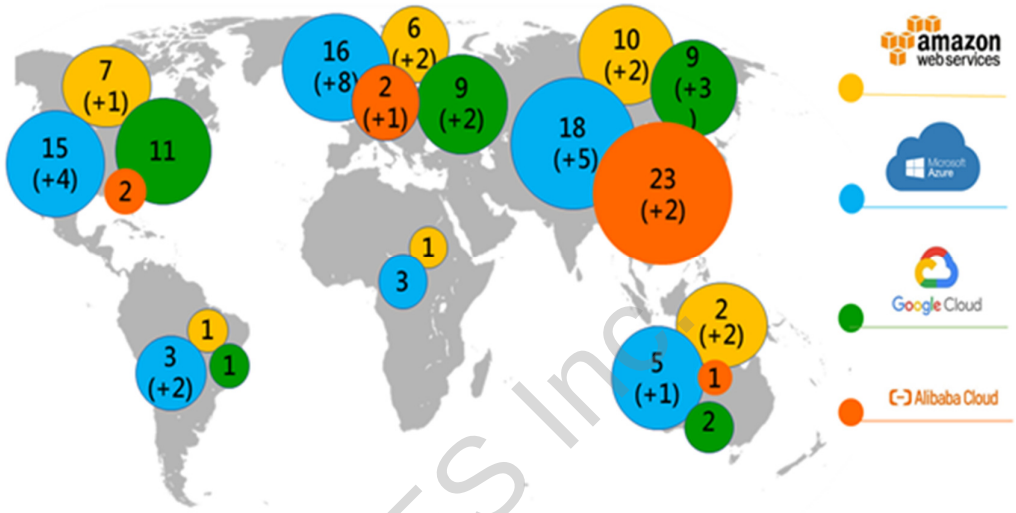
As the economy falls under the influence of rising inflation and the Russia-Ukraine war and therefore weakens, enterprises may cut back their capital expenditures and place conservative server orders in 2023. Global server shipment growth 2023 through 2024 is thus projected to come short of 6%. Going into 2025, with the server demand from large cloud service providers remaining strong and the server demand from enterprises building up their digitization infrastructure gradually ramping up, global server shipment on-year growth is expected to rebound above 6%.

Intel and AMD are set to launch new CPU solutions targeting cloud, HPC and edge server applications in 2023. This is expected to fuel some server upgrades from 2024 onward.

Arm-based solutions will further erode the market share of x86-based servers as large cloud service providers (such as Amazon) and chip suppliers including Nvidia and Ampere aggressively expand their cloud and HPC offerings in 2023 and beyond.

Public cloud datacenter deployment

Table 2: First-tier public cloud datacenter operators capacity and expansion for next five years



**Note: The numbers indicate datacenters operators currently have in each region, while numbers in the bracket are planned new capacity.*
**Note: Datacenter operators are building datacenters in each region to provide services there, while each datacenter has independent infrastructure including servers, network and backup mechanism.*
Source: Companies, compiled by Digitimes Research, September 2022

Leading cloud service providers including Amazon, Microsoft, Google and Alibaba continuing to build up datacenter infrastructure worldwide will sustain global server shipment growth over the next five years.

Breaking down by geography, the top-4 public cloud service providers will make the most active efforts toward expanding datacenters in Europe and Asia, including 13 regions in Europe and 12 regions in Asia.

The public cloud service providers have plans to build datacenter infrastructure throughout different geographical locations, each of which is called a region. Every region includes three to four availability zones (AZ) and every AZ is furnished with thousands of servers, network devices and power supply systems.

As of 2022, Amazon has built up 27 regions and 87 AZs, with plans to add seven regions and 21 AZs in Australia, Canada, India, Israel, New Zealand, Spain and Switzerland.

Amazon is adding its datacenters to meet not only external public cloud needs but also internal e-commerce and media streaming needs.

Continuing to add datacenters throughout the world, Microsoft plans to build new facilities in the US, Asian countries including Taiwan, Malaysia, India, Indonesia and Israel, European countries including Denmark, Belgium, Spain, Greece, Poland, Finland, Austria and Italy, South American countries including Chile and Mexico as well as New Zealand.

Microsoft aims to use the added datacenters to support its public cloud service Azure in addition to its other offerings including the productivity tool Office Suites, the collaboration platform Microsoft Teams, the search engine Bing, and the ERP solution Dynamics 365.

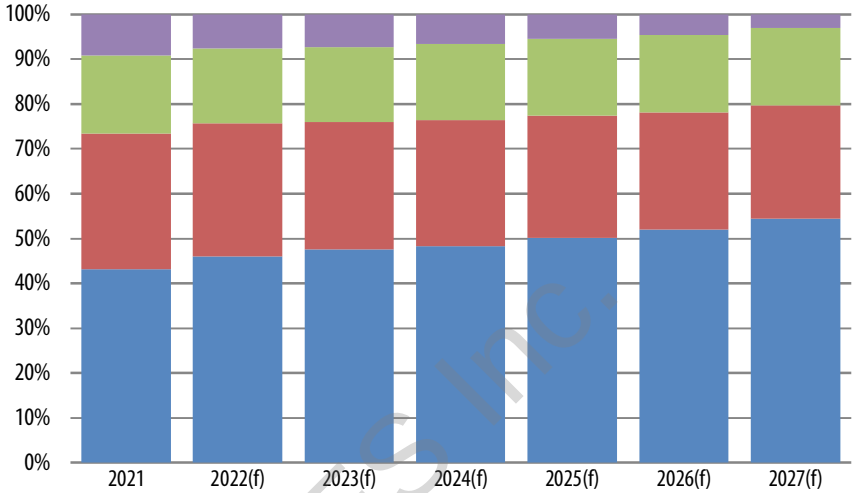
Google operates 34 regions and keeps expanding into additional parts of the world including Italy, Germany, Qatar, Israel and Arabia.

Google is adding datacenters to cope with public cloud service and media platform needs. Specifically, public cloud services encompass cloud computing and storage, container services, cloud SQL database, virtual private cloud service and security management services.

Focusing on building datacenters locally in China, Alibaba has 13 regions running in China, representing more than half of them. It has plans to add datacenters in Europe and Asia.

Vendor type

Chart 2: Shipment share by vendor type, 2021-2027



	2021	2022(f)	2023(f)	2024(f)	2025(f)	2026(f)	2027(f)
Others	9.1%	7.5%	7.2%	6.6%	5.5%	4.6%	3.0%
China brands	17.5%	16.8%	16.8%	17.0%	17.1%	17.3%	17.3%
US brands	30.3%	29.7%	28.5%	28.0%	27.2%	26.1%	25.2%
Cloud datacenters	43.1%	46.0%	47.5%	48.4%	50.2%	52.0%	54.5%

Source: Digitimes Research, September 2022

In the global server market, the cloud datacenter segment shows the strongest growth momentum, which drives the growth of white-box server shipments. White-box servers represented only 43% of the global server market in 2021. The share is expected to rise to nearly 55% by 2027, buoyed by the robust growth in large cloud datacenters.

White-box server shipments, which are mainly for large cloud datacenters, are projected to grow at a CAGR of close to 10% from 2022 through 2027.

Server shipments to US-based server brands are projected to grow at a CAGR of only 3% from 2022 through 2027 and their market share will lower to nearly 25% by 2027 as their customers continue to feel pressure from cloud service providers.

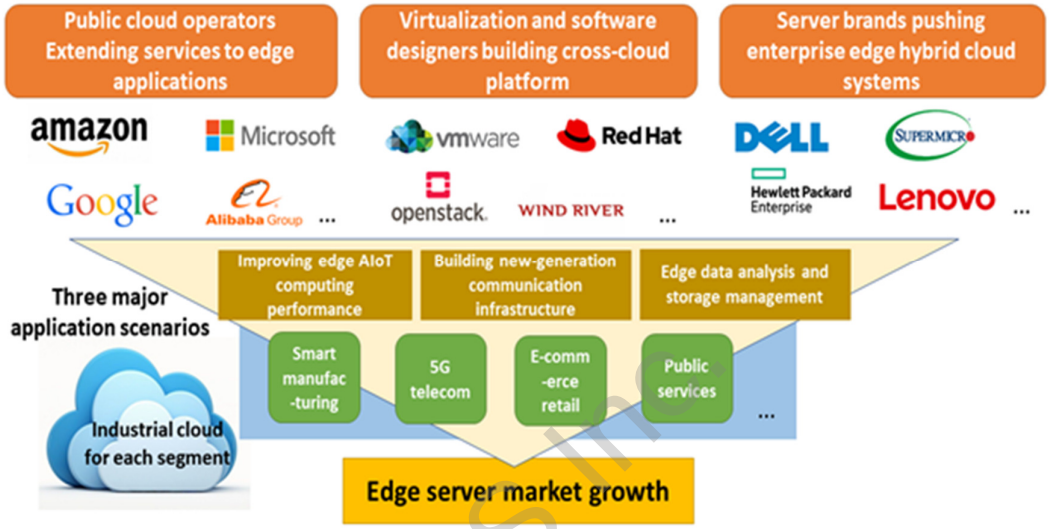
The share of shipments to US-based server brands will fall short of 29% in 2023 as their enterprise customers reduce their capital expenditures toward server purchases amid the weak global economy.

China-based server brands (including Inspur, Lenovo and Huawei) target local datacenter operators and brands. The US-China trade conflict and geopolitical uncertainties blocking their access to CPU and GPU may affect their server shipment growth. Nevertheless, with the China government's support, 17.3% of global server shipments will go to China-based server brands in 2027, up 0.5pp from the 2022 level.

Going into 2023, large China-based datacenter operators (including Alibaba, Tencent and Baidu) will show flat growth as they place conservative server orders in response to information security control measures. In 2024 and beyond, the growth momentum will gradually resume due to lower base periods. However, with a focus on the local China market, which is expected to approach saturation, their on-year growth will slow down and their shipment share will maintain around 17%-18%.

Hybrid cloud deployment

Table 3: Cloud and server brands hybrid cloud deployment



Source: Digitimes Research, September 2022

Digitimes Research has observed that in the post-pandemic era, public cloud service providers and server brands are enjoying growing multi-cloud platform and edge hybrid cloud demand as their enterprise customers show increasing needs for cloud or remote online services.

Amazon is expected to expand its collaborations with telecom carriers (including Verizon, Vodafone and Dish) on AWS Wavelength for integrated 5G telecom services while leveraging the AWS Outposts hybrid cloud to keep building up its presence in the Open RAN white-box telecom solution market.

From 2023 onward, Amazon plans to enhance its lightweight version of Outposts (in 1U or 2U form factors) hybrid cloud services for enterprise customers' edge server applications including edge AI inferences or deployment of on-premise edge telecom functionalities.

Server brands will work with virtualization platform developers (including VMware and RedHat) to expand their edge hybrid cloud services.

From 2022 onward, Dell and HPE are expanding their subscription-based offerings of integrated server hardware and various application software to secure their foothold in the edge hybrid cloud market.

Their subscription-based offerings are copied from public service providers, where instead of one-time capital expenditure, their enterprise customers pay regularly according to their actual needs.

Dell, HPE and Supermicro will also make additional efforts toward the Open RAN telecom equipment market and work with telecom carriers (including Vodafone, Dish, KDDI and NTT Docomo) to introduce CU and DU servers or target opportunities in the private 5G network market.

Applications

Table 4: Cloud services providers infrastructure establishment for each different application segment

Segment	Key development focus	Growth potential
Financial services	Focusing on government regulatory compliance and financial data security	↑ ↑ ↑ ↑
Telecom infrastructure	Augmenting cloud network integrated services and O-RAN cloud service deployment	↑ ↑ ↑ ↑
Smart manufacturing	Increasing edge cloud AIoT applications by the high-tech and traditional manufacturing industries	↑ ↑ ↑
E-commerce platforms	Using AI to optimize consumer experience and accelerate workflow	↑ ↑
Healthcare	Strengthening hospital infrastructure and raising healthcare quality	↑ ↑
Public infrastructure services	Meeting governments' digitization initiatives or efforts to make people's life more convenient	↑ ↑ ↑

Source: Digitimes Research, September 2022

Public cloud service providers will keep expanding their industry cloud services across different vertical applications and sectors, which will drive growth in edge server shipments.

Continuing to add datacenters worldwide, public cloud service providers aim to offer all-purpose cloud services to enable cloud computing, remote data backup, container management and network security control while tailoring special-purpose industry cloud for specific industries to help enterprises complete their digital transformation or enhance the efficiency of their daily operations or special tasks.

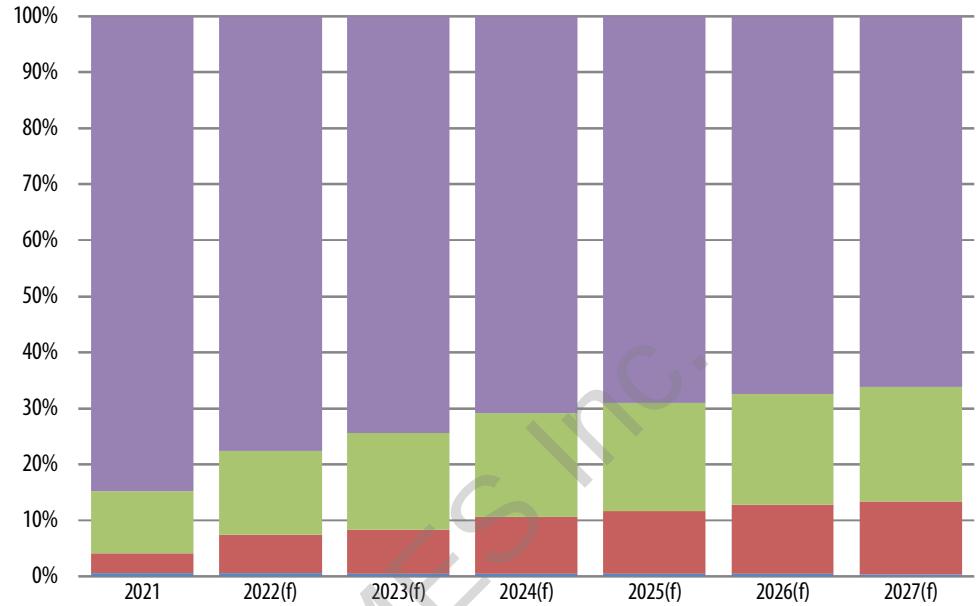
According to Digitimes Research's observations, the industries that public cloud service providers target their special-purpose industry cloud for currently include financial services, telecom infrastructure, smart manufacturing, e-commerce platforms, healthcare, and public infrastructure services. In particular, the financial services and telecom infrastructure markets will show the most growth potential.

The financial industry cloud is built for regulatory compliance concerning sensitive financial data and client confidentiality, which prompts public cloud service providers to establish on-premise datacenters.

Telecom industry cloud opportunities arise from public cloud service providers' abilities to leverage their software-defined everything (SDx) functionalities, such as integrating software-defined networking (SDN) technology in Open RAN compliant servers, to provide all types of telecom cloud application solutions for edge telecom base stations.

CPUs/GPUs

Chart 3: Shipment share by CPU, 2021-2027



	2021	2022(f)	2023(f)	2024(f)	2025(f)	2026(f)	2027(f)
Intel	84.7%	77.5%	74.4%	70.9%	69.1%	67.4%	66.2%
AMD	11.2%	15.1%	17.2%	18.5%	19.2%	19.8%	20.5%
Arm	3.5%	6.8%	7.9%	10.1%	11.2%	12.4%	13.0%
Others	0.6%	0.6%	0.5%	0.5%	0.5%	0.4%	0.3%

**Note: Others include IBM's RISC-based Power CPUs that are used in a few specific workstation systems and super computers.*
Source: Digitimes Research, September 2022

AMD's EPYC CPUs will represent more than 20% of processors used in datacenters and HPC servers by 2027.

As of 2022, the four leading US-based datacenter operators including Amazon, Microsoft, Google and Meta as well as their China-based counterparts including Alibaba, Tencent and Baidu all use AMD's EPYC chips.

Microsoft shows the most aggressive use of EPYC CPUs. It was the first to adopt the first-generation EPYC CPU to support its Azure cloud services in 2017. AMD's Milan-X, launched in first-half 2022, was also first used by Microsoft to entirely replace the previous generation Milan chips powering Azure cloud services.

AMD completing the acquisition of leading FPGA supplier Xilinx in first-half 2022 will help it expand its presence in the edge server market, encompassing AI inferencing for various applications as well as edge telecom servers.

Growing adoption by US- and China-based datacenter operators will buoy the market share of Arm-based CPUs to 13% by 2027.

Among large cloud service providers, Amazon makes the most aggressive use of Arm-based CPUs. Aside from the continuing use of its chips to expand the EC2 public cloud services, the lightweight version of Outposts, launched at year-end 2021, also used its own Graviton 2. It is expected to become the main growth driver for Amazon's footprint in the enterprise edge cloud market.

Nvidia is poised to unveil its server CPU Grace in 2023, which will be based on Arm Neoverse V2 to target HPC and AI applications requiring high performance.

US-based datacenter operators including Microsoft and Google, China-based firms including Tencent and Inspur as well as server vendors including HPE, Gigabyte and Foxconn began to use Ampere's Arm-based CPUs in 2022.

The competition from the two strong rivals will cause Intel's market share to fall below 70% starting in 2025. In response, Intel will strengthen the development of AI chips such as GPU and FPGA or datacenter infrastructure processing units (IPU).

Table 5: Server CPU roadmaps by supplier, 2021-2024

Supplier	2021	2022(f)	2023(f)	2024(f)
Intel	Whitley: supports PCIe 4.0; 10+nm node	Sapphire Rapids: supports PCIe 5.0; Intel 7 node	Emerald Rapids: Intel 7 node	Granite Rapids: Intel 3 node
AMD	Milan: supports PCIe 4.0; 7nm node	Milan-X Genoa: supports PCIe 5.0, 5nm node	Bergamo, Genoa-X and Siena: 5nm node	Turin: 3nm node
Arm	Neoverse N1	V2	New N series	
Ampere	Altra Max: 7nm node	AmpereOne-1: supports PCIe 5.0; 5nm node	AmpereOne-2	AmpereOne-3: 3nm node
Nvidia			Grace: 5nm node	

Source: Digitimes Research, September 2022

Intel's Eagle Stream platforms are scheduled to enter volume production in first-half 2023, to be followed by the new Emerald Rapids CPU in 2023 and Granite Rapids CPU in 2024.

Due to the delay in Eagle Stream volume production and Emerald Rapids being an interim product, server manufacturers expect Granite Rapids, featuring a significant boost in performance, to spur major upgrades. Granite Rapids will initially be used in HPC or AI servers.

AMD plans to market the Genoa platforms in fourth-quarter 2022, which will be largely used by cloud datacenter operators including Microsoft and Google starting 2023.

AMD is set to further introduce the Bergamo, Genoa-X and Siena platforms in 2023, respectively targeting cloud, HPC and edge telecom servers. This is expected to drive some server upgrades on the part of datacenters and enterprises from 2024 onward.

Arm launched Neoverse V2, targeting the HPC market, in September 2022. It is also the architecture that Nvidia's self-developed chip Grace, to debut in 2023, will be based on.

Arm's server CPU product family Neoverse includes the V, N and E series. The V series is designed for HPC application scenarios, the N series is often adopted by cloud datacenter operators or enterprise users and the E series comprises edge servers, DPU and networking processors featuring compelling CP ratios.

Ampere, which develops Arm-based chips, continues to work on new-generation processors that are seeing growing use by US-based and China-based cloud service providers as well as server brands.

Leading server CPU suppliers are developing in the following directions going forward:

Server CPUs are trending toward being application specific. For example, in 2023, AMD will market new-generation CPUs respectively for large cloud datacenters, HPC/AI servers and edge servers.

Server manufacturers will move toward multi-core systems (for example, AMD and Ampere are developing CPUs with more than 100 cores) to target cloud datacenters or accelerated parallel computing applications.

Going into 2024 and beyond, leading chip suppliers will have their CPUs made on the 3-nm process, which will bring the computing performance of new-generation server systems to the next level.

Arm-based CPU

Table 6: Cloud datacenter operators and server companies' plans for Arm-based products

Source: Digitimes Research, September 2022

With US-based cloud service providers building bigger and bigger datacenters and end market demand becoming more and more diverse, it is a growing trend that cloud service providers are strengthening their in-house chip development.

Amazon makes the most aggressive efforts toward developing an Arm-based CPU. Unlike Graviton 2, which was based on the Neoverse N1 architecture, Graviton 3, launched at year-end 2021, adopted the Neoverse V1 architecture, which features higher performance, to support Amazon's EC2 high-speed cloud computing services.

According to supply chain sources, Apple's use of Arm-based CPUs largely in MacBook generated positive results. To meet all kinds of needs for online services, iCloud and gaming platforms, Apple is set to add self-built cloud datacenters while undertaking the development of Arm-based server solutions.

Microsoft announced in September 2022 that it plans to use Ampere's Arm-based CPUs to support Azure cloud computing services, which will be initially offered in the US, Europe, Asia and Australia. Subsequently, Microsoft will launch Kubernetes cloud container services powered by Arm-based chips.

In July 2022, Google announced its plan to adopt Ampere's Arm-based chips to support the computing capacities of Google Cloud.

Alibaba unveiled its own Arm-based CPU Yitian 710 at year-end 2021, which will be largely used in Alibaba's cloud datacenters. Its peers are expected to increasingly develop Arm-based CPUs or AI accelerators in response to growing geopolitical uncertainties including the US-China trade war.

Ampere continues to develop new-generation processors. Aside from US-based cloud service providers Microsoft and Google, Ampere's chips are also embraced by Chinese firms including Tencent and Inspur as well as server brands including HPE and Gigabyte.

To server brands, servers running on Arm-based processors feature better C/P ratios or allow them to offer a more diverse product portfolio to customers.

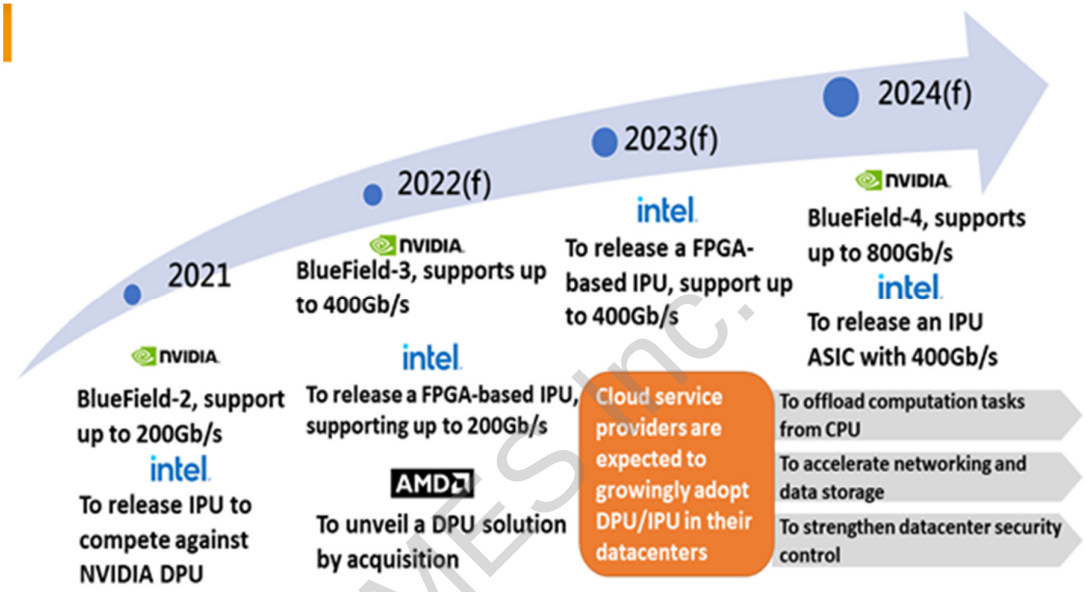
HPE used Arm-based processors in its Apollo server product line, which was more for specific projects. Starting second-half 2022, HPE will use Ampere's Arm-based chips in ProLiant, its major and standard server product line.

In the face of strong competition from Arm, Intel is expected to adopt a more open licensing strategy for its x86 architecture, allowing in-house customization based on customer needs.

To obtain licenses for Intel's IP, customers must outsource the production of their self-developed chips to Intel.

DPU/IPU

Table 7: Nvidia, Intel, AMD DPU/IPU roadmap, 2021-2024



Source: Digitimes Research, September 2022

As CPUs or AI accelerators (such as GPUs) feature higher and higher processing speed and performance, chip suppliers including Intel, Nvidia and AMD are expected to emphasize smart network interface card (NIC) (SmartNIC) technologies to address the interconnect bottlenecks between CPU or AI chips and networking or storage devices and to strengthen security control on remote access to datacenters.

SmartNIC chips are designed to process networking, storage or security control in datacenters at faster speeds to help offload tasks from the CPU and thereby improve server system performance. Such solutions offered by chip suppliers may have different names. For example, Nvidia and AMD call their SmartNIC chips DPU while Intel dubs its solution IPU.

As part of its aggressive campaign to promote its NPU product line, Nvidia has acquired high-speed network device manufacturer Mellanox and launches a new generation of products every 1.5 to 2 years. Its BlueField-4, to debut in 2024, will support data transmission up to 800Gb/s, doubling the speed of BlueField-3. It is aimed to accelerate the interconnect with its CPU Grace and GPU.

Intel announced its IPU in 2021, mainly to compete against Nvidia's DPU.

Intel plans to introduce Mount Evans, an IPU featuring 400Gb/s data transmission, in 2024. Instead of the FPGA architecture, Mount Evans will be Intel's first ASIC-based IPU, featuring higher performance for specific accelerated computing or networking tasks.

Intel also makes active efforts toward promoting Compute Express Link (CXL), an industry-supported standard based on the PCIe high-speed interconnect interface to accelerate the connections between CPUs or AI accelerators and memory modules.

By completing the acquisition of Xilinx and programmable DPU accelerator developer Pensando in 2022, AMD began its foray into the DPU market with an ambition to further increase its presence in the datacenter infrastructure market.

VMware announced in second-half 2022 that the company officially supports Nvidia's and AMD's DPUs and Intel's IPUs with its new vSphere virtualization platform to accelerate computation for its services such as NSX network virtualization and security management as well as vSAN storage virtualization.