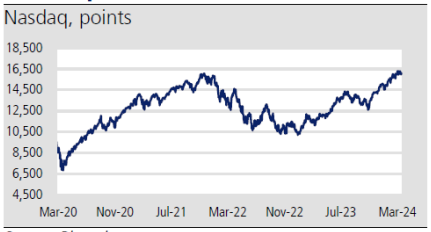
US technology sector

Nvidia GTC 2024: New GPU platforms, software & robots

Key Messages

1. Takeaways from Nvidia's (US) GTC 2024: (1) the Blackwell platform is on schedule for launch; (2) Nvidia NIM & NeMo micro-services announced; (3) Project GR00T and Jetson Thor chip introduced; and (4) Nvidia announced Omniverse Cloud will be available as an application programming interface (API).
2. GB200 NVL72, a liquid-cooled rack hosting 36 GB200 superchips, offers improved performance over the HGX H100 in terms of inferences, training, energy consumption and data processing. US Cloud CSPs including Microsoft , AWS, Google Cloud and Oracle Cloud Infrastructure intend to deploy the GB200.
3. We expect Quanta Computer (2382 TT, NT$257, OP), Hon Hai Precision (2317 TT, NT$136, OP), Wistron (3231 TT, NT122.5, OP), and Wiwynn (6669 TT, NT$2,175, OP) to be the key beneficiaries of the announcements. We expect liquid cooling penetration to grow in 2025-26F, which will benefit Cooler Master (TW; unlisted), Asia Vital Components (3017 TT, NT$569, OP), and Auras Technology (3324 TT, NT$655, OP).

Nasdaq



Source: Bloomberg

SOX Industrial Average

Source: Bloomberg

Event

Nvidia (US) held the 2024 GPU Technology Conference (GTC) and introduced the next-generation Blackwell platform of data center GPUs, as well as AI Foundry services. Our takeaways are as follows: (1) the Blackwell platform includes B100 chips, GB200 superchips, GB200 NVL72 racks, and NVlink switch chips for connecting multiple GPUs; (2) the firm announced NeMo and Nvidia Inference **micro**-services (NIM), which, together with DGX Cloud, fall under the brand of AI Foundry services. These are designed to help streamline the development, deployment, customization and operation of generative AI for enterprises; (3) Project GR00T, a general-purpose foundation model for humanoid robots, was introduced during the event, alongside Jetson Thor chips; and (4) Nvidia announced Omniverse Cloud will be available as an application programming interface (API) for creating digital twin applications by software developers.

Launch of Blackwell GPU platform on schedule. The Blackwell GPU consists of two dies, etched with TSMC's (2330 TT, NT$762, OP) N4P process, with both carrying 30% more transistors than a Hopper die, featured with the 2nd generation Transformer engine and 5th generation NVLink. The Blackwell GPU has a 2.5x increase in TFLOPS FP8 computing power for AI training, or 5x more FP4 computing power for AI inferencing, compared to the Hopper GPUs. Blackwell and Hopper operate on the same HGX standard, so clients may easily upgrade. Blackwell will hit the market later this year, and Nvidia indicated it will be adopted by major US CSPs, namely Amazon Web Services (AWS), Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure, and by tier-2 US CSPs like Applied Digital, CoreWeave, Crusoe, IBM Cloud and Lambda, as well as sovereign AI cloud operators.

GB200 Superchip the next flagship GPU for data centers. The GB200 Superchip will come with two B200 GPUs and one Grace CPU. In light of rising demand for computing power from trillion-parameter multimodal AI models, Nvidia also unveiled GB200 NVL72, a DGX GB200 system featuring a liquid-cooled rack which carries 36 GB200 Superchips, and offers better performance than HGX H100 in terms of inferences, training, energy consumption and data processing. GB200 NVL72 will also become available later this year. The adoption rate of GH200 on the Hopper platform is low, as it is meant to help clients become familiar with the Grace system, but many cloud CSPs, including Microsoft, AWS, Google Cloud and Oracle Cloud Infrastructure, revealed intentions to deploy the GB200.

GB200 Superchip to drive AI rack-scale demand. We expect Quanta Computer (2382 TT, NT$257, OP), Hon Hai Precision (2317 TT, NT$136, OP), Wistron (3231 TT, NT122.5, OP), and Wiwynn (6669 TT, NT$2,175, OP) to be the key beneficiaries of the announcements. In addition, Nvidia's main US clients, including AWS, Microsoft, Google, Oracle, Meta, OpenAI, and Tesla, will be buyers of DGX computers powered by the GB200 Superchip, so Taiwanese ODMs serving these clients will benefit. Despite the GB200 having no base board design, Wistron will not suffer as it will offer several boards (i.e. NVLink switch and others) with content value higher than the current value of boards for the H100. However, as the GB200 will launch later this year, we expect system and rack production to commence after 2Q25F. In addition, the forthcoming GB200 Superchip-powered DGX computers and GB200 NVL72 server rack will utilize liquid-cooled MGX modular designs, according to Nvidia, so we expect liquid cooling penetration to grow in 2025-26F, which will benefit Cooler Master (TW; unlisted), Asia Vital Components (AVC; 3017 TT, NT$569, OP), and Auras Technology (3324 TT, NT$655, OP), and create business opportunities due to rising demand for coolant distribution units (CDU), manifolds, and quick connectors.

Stocks for Action

We maintain Outperform on Nvidia, with a target price of US$950. Within the firm's Taiwan supply chain, we think Quanta Computer, Hon Hai, Wistron, Wiwynn, AVC, and Auras Technology will be the key beneficiaries of the GB200 product cycle.

Risks

Weak ICT demand; low adoption of new products.

Figure 1: Performance of major indices

Index	Recent close (pts)	1W (change, %)	2W	1M	3M	6M	12M	YTD
Dow Jones	38,790	0.1	12.0	0.4	4.0	12.0	21.7	2.9
Nasdaq	16,103	0.5	17.5	2.1	8.0	17.5	38.5	7.3
SOX	4,758	(2.7)	36.2	5.1	15.9	36.2	54.3	13.9

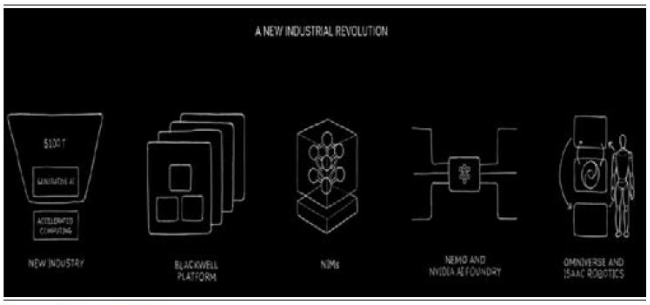
Source: Bloomberg

Figure 2: Peer valuation comparison

Company	Ticker	Market cap (US$bn)	Turnover (30 days moving avg.) (US$mn)	Rating	Target price (US$)	Share price (US$)	Upsides (%)	PE (x) 2023	PE (x) 2024F	PE (x) 2025F	EPS (US$) 2023	EPS (US$) 2024F	EPS (US$) 2025F	EPS YoY (%) 2023F	EPS YoY (%) 2024F	EPS YoY (%) 2025F
NVIDIA	NVDA	2,211	46,790	OP	950.0	884.6	7.4	68.2	28.7	22.9	12.96	30.78	38.61	283.8	137.5	25.4
Broadcom	AVGO	573	4,265	OP	1,435.0	1,237.2	16.0	29.3	22.2	21.9	42.25	55.78	56.39	12.3	32.0	1.1
AMD	AMD	308	12,911	OP	200.0	190.7	4.9	72.0	46.9	36.1	2.65	4.06	5.27	(25.1)	53.4	29.8
Intel	INTC	181	1,913	OP	53.0	42.7	24.1	40.9	30.8	20.3	1.04	1.39	2.10	(43.4)	32.7	51.6
Marvell	MRVL	58	1,114	OP	85.0	67.2	26.5	44.7	29.3	24.0	1.50	2.29	2.80	(29.1)	52.3	22.4
Dell	DELL	76	850	OP	140.0	106.6	31.3	15.0	13.4	10.8	7.12	7.97	9.88	(6.4)	11.9	24.0
HPE	HPE	22	291	N	18.0	17.1	5.5	7.9	8.6	8.3	2.15	1.97	2.06	6.8	(8.3)	4.4

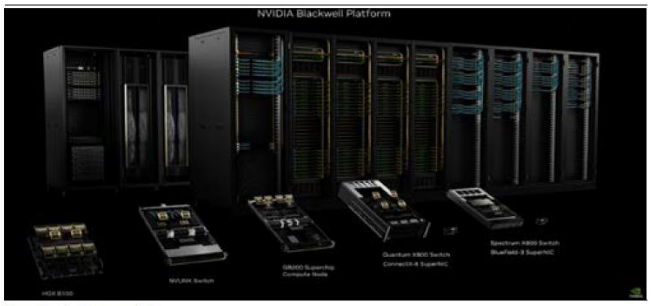
Source: Bloomberg; KGI Research

Figure 3:Major announcements at Nvidia GTC



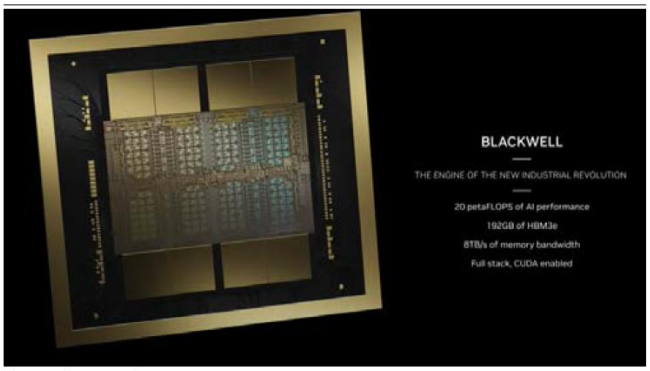
Source: Company data

Figure 4: Overview of Nvidia's Blackwell platform



Source: Company data

Figure 5: Features of the Blackwell GPU



Source: Company data

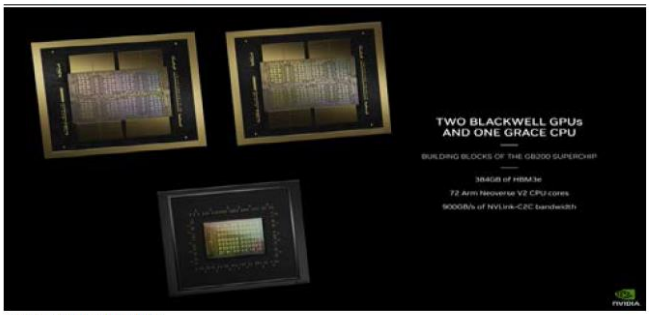
Figure 6: Blackwell to boost AI performance by 5x compared to Hopper

Source: Company data

Figure 7: Overview of the GB200 Superchip

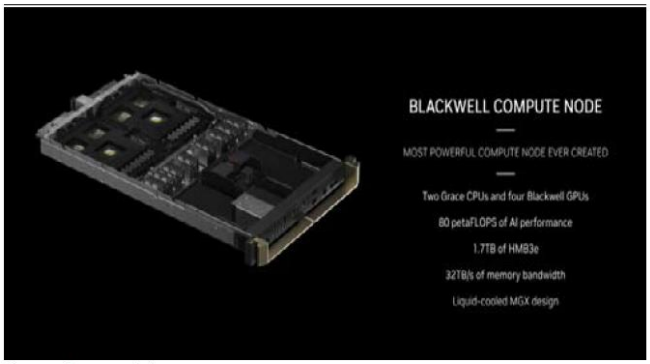
Source: Company data

Figure 8: GB200 Superchip consists of two Blackwell GPUs and one Grace CPU



Source: Company data

Figure 9: A Blackwell compute node consists of two CPUs and four GPUs



Source: Company data

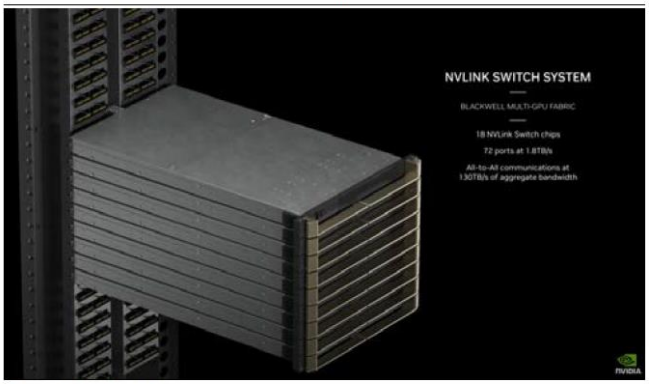
Figure 10: A single GB200 NVL72 provides computing power of 1.4 exaFLOPS

Source: Company data

Figure 11: A GB200 NVL72 consists of eighteen Blackwell compute nodes

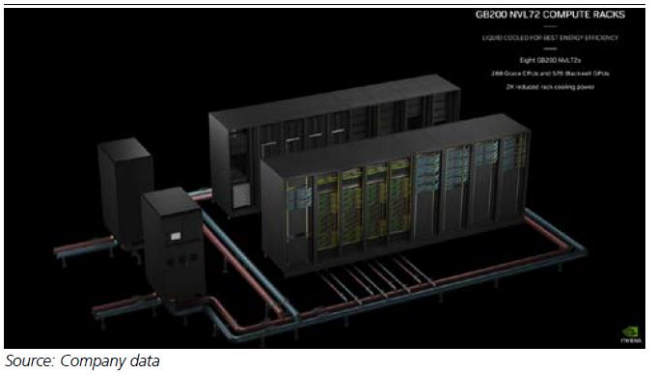
Source: Company data

Figure 12: GB200 NVL72 consists of nine switch nodes, which carry eighteen switch chips



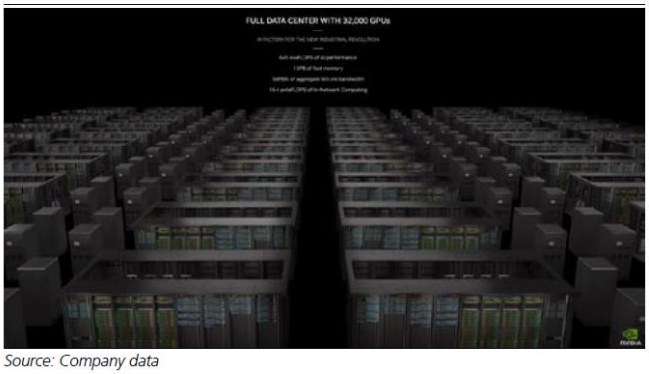
Source: Company data

Figure 13: DGX SuperPOD consists of eight GB200 NVL72 with liquid cooling for optimal efficiency



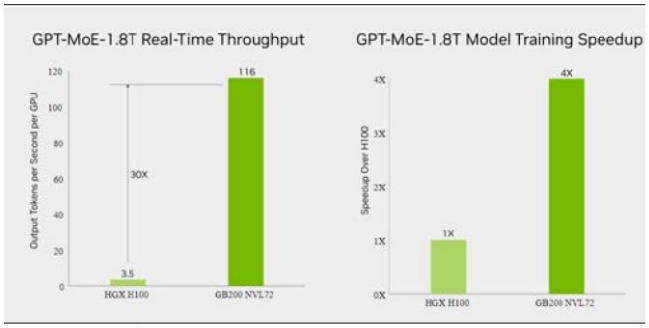
Source: Company data

Figure 14: A data center may host up to 32,000 GPUs



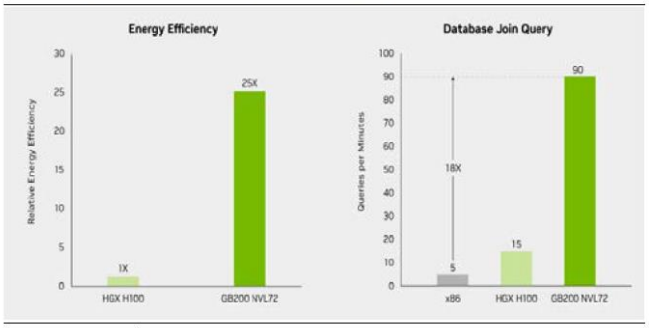
Source: Company data

Figure 15: GB200 NVL72 will allow faster AI inferences and training



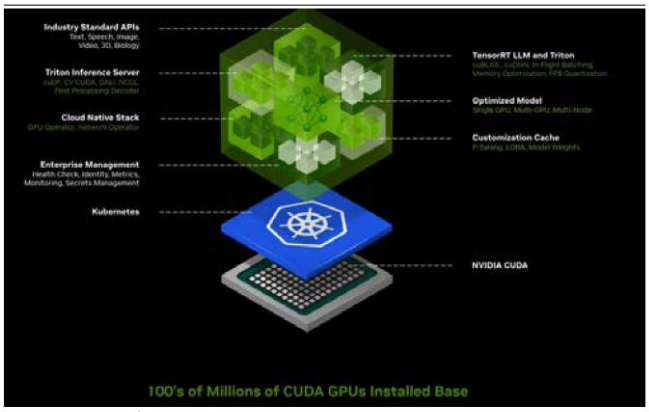
Source: Company data

Figure 16: GB200 NVL72 offers more efficient energy consumption and data processing capabilities



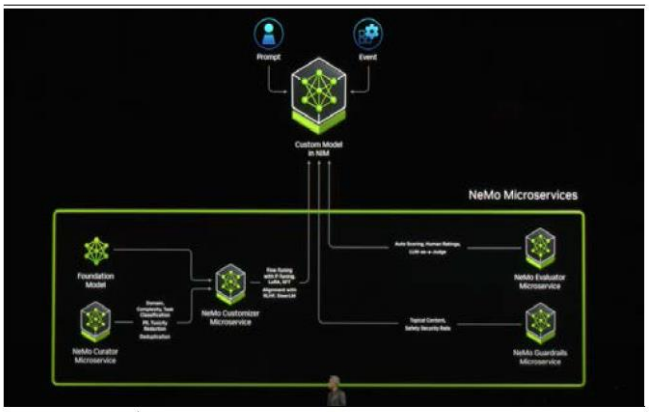
Source: Company data

Figure 17: Nvidia NIM



Source: Company data

Figure 18: Nvidia NeMo microservice



Source: Company data

Figure 19: GB200 specs versus H100 series

	GB200 Superchip	GH200 Superchip	H200 SXM	H100 SXM	H100 PCIe
Configuration	1 Grace CPU : 2 Blackwell GPUs	1 Grace CPU : 1 Hopper GPU	1 Hopper GPU	1 Hopper GPU	1 Hopper GPU
FP64 Tensor Core	90 teraFLOPS	67 teraFLOPS	67 teraFLOPS	67 teraFLOPS	51 teraFLOPS
TF32 Tensor Core	5,000 teraFLOPS	989 teraFLOPS	989 teraFLOPS	989 teraFLOPS	756 teraFLOPS
BFLOAT16 Tensor Core	10,000 teraFLOPS	1,979 teraFLOPS	1,979 teraFLOPS	1,979 teraFLOPS	1,513 teraFLOPS
FP16 Tensor Core	10,000 teraFLOPS	1,979 teraFLOPS	1,979 teraFLOPS	1,979 teraFLOPS	1,513 teraFLOPS
FP8 Tensor Core	20,000 teraFLOPS	3,958 teraFLOPS	3,958 teraFLOPS	3,958 teraFLOPS	3,026 teraFLOPS
INT8 Tensor Core	20,000 teraFLOPS	3,958 TOPS	3,958 TOPS	3,958 TOPS	3,026 TOPS
FP4 Tensor Core	40,000 teraFLOPS				
GPU Memory	Up to 384 GB HBM3e	Up to 96GB HBM3 Up to 144 GB HBM3e	HBM3e 141GB	HBM3 80GB	HBM3 80GB
GPU Memory Bandwidth	16 TB/s	Up to 4TB/s Up to 4.9TB/s HBM3e	4.8TB/s	3.35TB/s	2TB/s
Max Thermal Design Power (TDP)		Programmable from 450W to 1000W (CPU+GPU+memory)	Up to 700W (configurable)	Up to 700W (configurable)	300-350W (configurable)
Form Factor	Superchip module	Superchip module	SXM	SXM	PCIe dual-slot air-cooled
Interconnect	NVLink: 3,600GB/s	NVIDIA NVLink: 900GB/s	NVIDIA NVLink: 900GB/s PCIe Gen5: 128GB/s	NVIDIA NVLink: 900GB/s PCIe Gen5: 128GB/s	NVLink: 600GB/s PCIe Gen5: 128GB/s
Server Options			NVIDIA HGX™ H200 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs	NVIDIA HGX H100 Partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs NVIDIA DGX H100 with 8 GPUs	Partner and NVIDIA-Certified Systems with 1–8 GPUs

Source: Company data

Figure 20: H100/ GB200 suppliers

GPU model	GB200	H100
Platform	MGX	HGX
Major components	Major supplier	
GPU module (OAM)	Hon Hai (FII)	Hon Hai (FII)
GPU baseboard (UBB)	x	Wistron (major) Hon Hai (FII)
NVLink switch board	Wistron	
Motherboard	Inventec Wistron	Inventec Wistron Mitac Quanta Hon Hai (FII)
Thermal	- Liquid cooling Cooler Master CoolIT AVC Auras Vertiv (CDU) CPC , Staubli (manifold)	- Air cooling (3D VC) Cooler Master AVC
Server (L10)	Quanta	Quanta
Rack (L11)	Wiwynn Hon Hai (FII) ZT System	Hon Hai (FII) Wistron group (Wiwynn) ZT System Supemicro Gigabyte Mitac Dell Asustek Oracle

Source: Company, dataKGI Research

All the above named KGI analyst(s) is SFC licensed person accredited to KGI Asia Ltd to carry on the relevant regulated activities. Each of them and/or his/her associate(s) does not have any financial interest in the respectively covered stock, issuer and/or new listing applicant.