凱基亞洲
KGI ASIA

Cloud AI & edge AI

AI moving from cloud to edge will revive the supply chain

Key message

1. AI server shipments will ramp-up in 2024 when CoWoS capacity increases, as CSPs and enterprise all aggressively expands their AI infrastructure in data centers
2. AI PC models will be launched after 2H24, followed by the introduction of Intel's (US) Meteor Lake CPU and Microsoft's (US) Windows 12 that supports AI PC.
3. Key beneficiaries include thermal, power supply, chassis and rail kit, PCB/CCL, and switch, with their product ASP expansion benefiting from computing and transmission performance upgraded for AI servers and AI PC.

Event

We expect cloud AI to take-off from late 2023 into 2024 on CSPs' aggressive expansion and a growing GPU supply. Device brands will launch more edge AI models from 2H24F, reviving the overall AI ecosystem in 2024-25F.

Impact

AI servers main catalyst for server demand growth in 2024F. AI server shipments commenced in 2023, at a limited volume, as they were constrained by CoWoS capacity. With easing supply constraint, we anticipate AI server demand to ramp up in 2024F and forecast training GPU shipments will grow from 1.53mn units in 2023F to 4.57mn units in 2024F, and 5.96mn units in 2025F. If we assume one training server will require eight training GPUs, then total training server shipments will be 191k units in 2023F, rising to 572k units in 2024F, and 993k units in 2025F. Based on the expected shares of training server at around 30% of total AI server shipments in 2024-25F, we maintain our forecast that total AI server (training and inference) shipments will be 578k units this year, 1.91mn units in 2024F, and 3.31mn units in 2025F, comprising a respective 5%, 14%, and 22% of total server market shipments. CoWoS capacity will begin to ramp up from 4Q23F, and as the supply chain requires 2-3 quarters of lead time, we expect overall AI server demand to quickly rise after 2Q24F. As AI servers carry a higher ASP, most of the firms in the AI supply chain have guided their AI server segments to be the main sales and earnings growth driver over the next two years. This includes ODMs, thermal, power supply, chassis, rail kit and PCB companies. Players in the server supply chain also indicated that general server shipments will recover despite mild demand in 2024F, after inventory corrections in 2023. In addition to server growth, we expect networking plays will benefit from rising high-end 400GbE switch adaptation rate, from 18% in 2023 to 21% in 2024F driven by AI/ machine learning (ML) demand uptrend, while 800GbE switches reaching 5% market penetration by 2024F. We expect server sales will be the key driver for these companies' sales and earnings growth in 2024F. Key beneficiaries include GPU module and baseboard suppliers Hon Hai (2317 TT, NT$102, OP) and Wistron (3231 TT, NT$93.3, OP), ODMs Quanta Computer (2382 TT, NT$201, OP), Wiwynn (6669 TT, NT$1780, OP) , Inventec (2356 TT, NT$41.6, OP), and Gigabyte (2376 TT, NT$243.5, OP), and component makers Auras Technology (3324 TT, NT$363, OP), AVC (3017 TT, NT$201, NR), Chenbro Micom (8210 TT, NT$246.5, OP), King Slide (2059 TT, NT$880, OP), Lite-On Tech (2301 TT, NT$109, OP), Accton Technologies (2345 TT, NT$538, OP), and EMC (2383 TT, NT$372, OP).

AI move to edge to enhance PC demand in 2H24-2025F. Intel (US) recently announced its AI PC acceleration program, which targets the enabling of AI functions on more than 100mn PC units, in the next two years, by connecting independent hardware vendors (IHVs) and independent software vendors (ISVs) with Intel resources. It will launch its mobile Meteor Lake CPU (Intel Core Ultra processor) in December 2024, which is compatible with neural processing units (NPU). Next-generation CPUs for both desktop (DT) and NB, including Arrow Lake, Lunar Lake and Panther Lake, will provide more advanced performance and capability. AI-enabled PCs will be able to help users create, edit, optimize and compress videos and audios, improve quality and efficiency, and help users protect their data and privacy, preventing various threats and attacks. AI PC development is focused on 'edge AI' - improving the reasoning capabilities built into PCs - instead of the current AI architecture, which is mainly based on cloud data centers, and requires the addition of more sensors to enable more intuitive operation. We expect more AI PC model launches in 2H24-2025F will fuel PC demand growth. As inventory correction ends in 4Q23-1Q24F, replacement demand should occur alongside an economic recovery, after the last cycle peak in 2020-21. Windows 12 will launch in 2024 with AI function support, which will trigger upgrade demand, and combined with the termination of Windows 10's technical support in October 2025F, we thus forecast for 2024 PC shipments to recover to 5.5% YoY growth (NB and DT both growing 5-6% YoY). The commercial market will replace and upgrade to AI-enabled PCs first. This should cause the NB supply chain's sales, including ODMs, and thermal and power supply manufacturers, to rebound in 2H24-2025F.

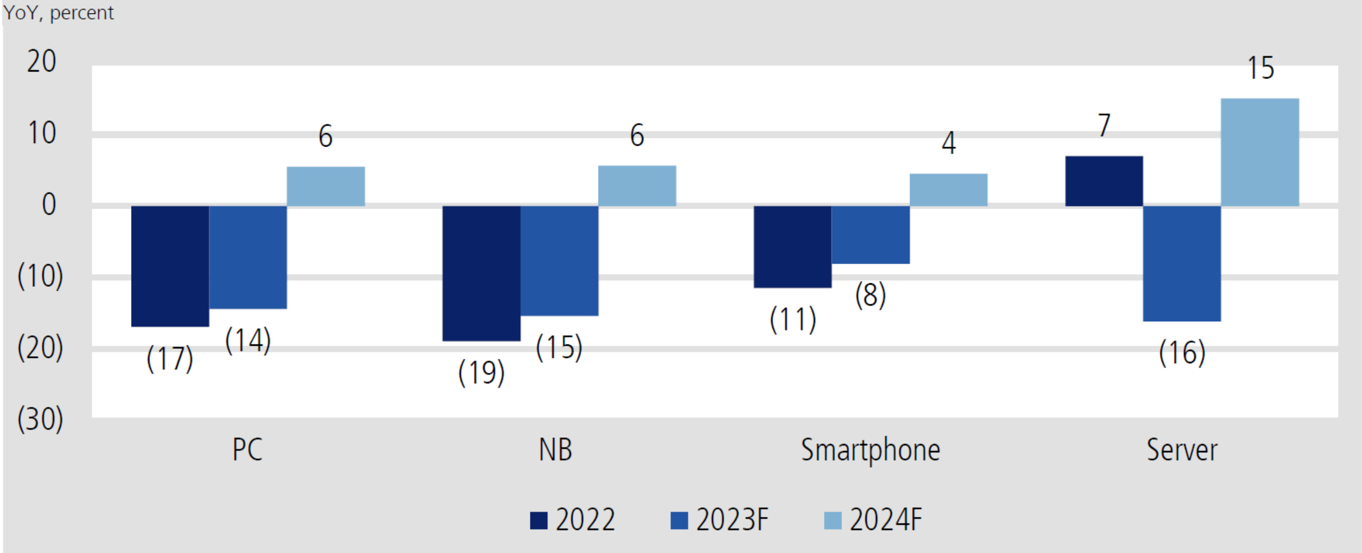
Stocks for Action

We expect beneficiaries of the cloud and edge AI growth wave will be Quanta Computer, Wiwynn, Wistron, Auras Technology, Lite-On Technology, Chenbro Micom, EMC, and Accton Technologies, on strong sales and EPS growth.

Risks

Weak NB PC demand; IT spending constraints; margin dilution for AI servers and PC.

Figure 1: IT hardware devices – NB, PC, smartphone, and server shipments will all grow in 2024F



Source: Gartner, KGI Research estimates

Figure 2: AI server shipments to rise in 2023-25F

GPU shipments (k units)	2022	2023F	2024F	2025F
A100/H100 GPU	875	1,125	3,550	4,615
Others (AMD MI300 / Google TPU)	105	400	1,022	1,341
Total training GPU	980	1,525	4,572	5,956
Server shipments (k units)	**2022**	**2023F**	**2024F**	**2025F**
A100/H100 GPU server	109	141	444	769
Other server (AMD MI300 / Google TPU)	13	50	128	223
Training AI server shipment	**123**	**191**	**572**	**993**
Total AI server (training & inference)	**371**	**578**	**1,905**	**3,309**
Total server (regular & AI server)	**13,815**	**11,532**	**13,262**	**15,251**
YoY growth (%)				
Training AI server	-	56	200	74
AI server (training & inference)	-	56	230	74
Total server (regular & AI server)	**7**	**(17)**	**15**	**15**
Training server weighting of total server (%)	**0.9**	**1.7**	**4.3**	**6.5**
AI server weighitng of total server (%)	**2.7**	**5.0**	**14.4**	**21.7**
Assumptions:				
A100/H100 GPU share of total training GPU (%)	89	74	78	77
GPU units per server	8	8	8	6
Training server share of total AI server (%)	33	33	30	30

Source: Gartner; KGI Research estimates

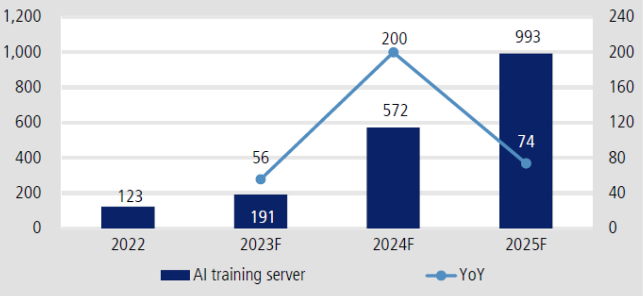
Figure 3: AI server shipment weighting up from 5% in 2023F to 22% in 2025F

Shipments (k units)	2022	2023F	2024F	2025F
AI server	371	578	1,905	3,309
General server	13,444	10,954	11,357	11,942
Total server	13,815	11,532	13,262	15,251
YoY (%)	**2022**	**2023F**	**2024F**	**2025F**
AI server		56	230	74
General server		(19)	4	5
Total server	7	(17)	15	15
Weighting (%)	**2022**	**2023F**	**2024F**	**2025F**
AI server	3	5	14	22
General server	97	95	86	78
Total server	100	100	100	100

Source: Gartner; KGI Research estimates

Figure 4: Training server shipments to grow from 191k units in 2023F to 572k in 2024F

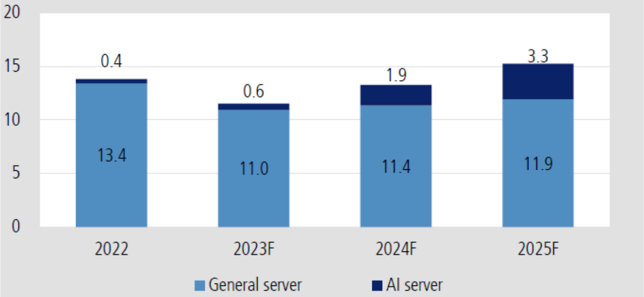
Global training server shipments, k units (LHS); YoY, percent (RHS)



Source: Gartner; KGI Research estimates

Figure 5: AI server training & inference business to boost server revenue uptrend

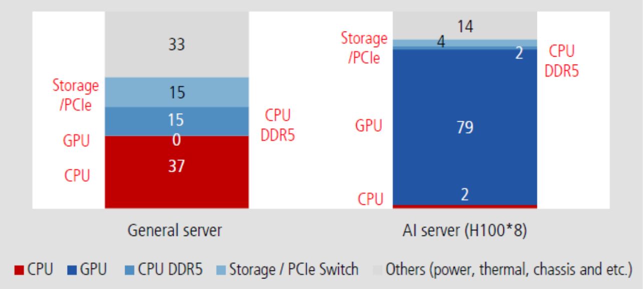
Global server shipments by general-purpose and AI, mn units



Source: Gartner; KGI Research estimates

Figure 6: GPUs account for the bulk of AI server BOM

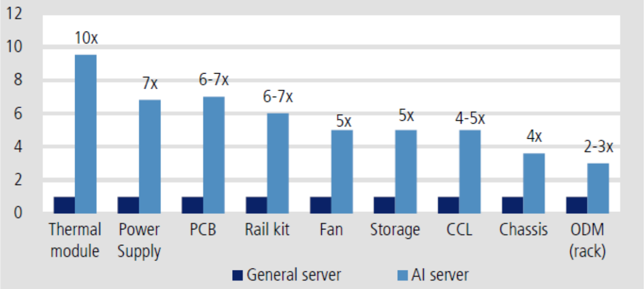
BOM cost weighting, percent



Source: KGI Research estimates

Figure 7: Thermal module & PSU for AI servers have 7-10x higher ASP than for general servers

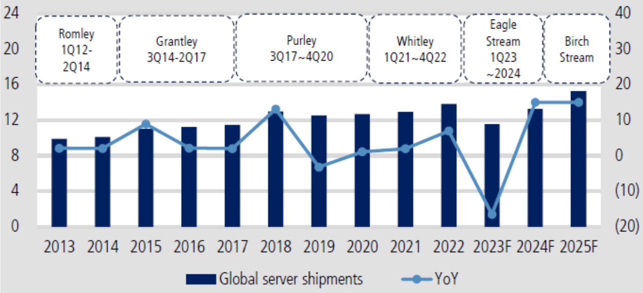
ASP multiples, times (assuming general server ASP is 1x)



Source: Company data; KGI Research estimates

Figure 8: Server demand to decline in 2023F & resume growth in 2024-25F

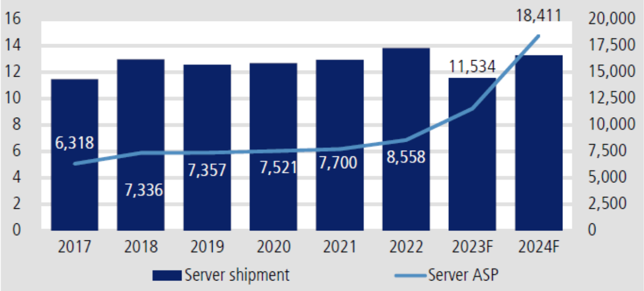
Global server shipments, mn units (LHS); YoY growth, percent (RHS)



Source: Gartner, KGI Research estimates

Figure 9: Server ASP uptrend on computing performance upgrades & high AI server demand

Global server shipments, mn units (LHS); server ASP, US$ (RHS)



Source: Gartner, KGI Research estimates

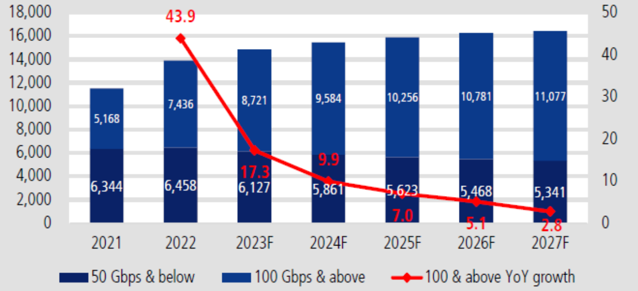
Figure 10: Server platform launches by Intel & AMD in 2023F; shipment ramp-up has experienced delays

Platform	Intel Purley	Intel Purley	Intel Cedar Island	Intel Whitley	Intel Eagle Stream	Intel Eagle Stream	Intel Birch Stream	AMD Zen 2	AMD Zen 3	AMD Zen 4	AMD Zen 4c	AMD Zen 4	AMD Zen 5	AMD Zen 6
Time of launch	3Q17	3Q19	2H20	2Q21	1Q23	4Q23	2024F	2Q19	1Q21	4Q22	1H23F	2023F	2024F	?
CPU	Skylake-EP Cannon Lake-EP	Cascade Lake	Cooper Lake	Ice Lake	Sapphire Rapids (Intel 7)	Emerald Rapids (Intel 7)	Granite Rapids (Intel 3, P-core)	Rome	Milan	Genoa	Bergamo	Siena	Turin	
Process	14nm/ 14nm+	14nm++	14nm	10nm	10nm	10nm++	3nm	7nm	7nm+	5nm	5nm	5nm	3nm / 4nm	2nm
CPU sockets	LGA 3647	LGA 3647	LGA 4189	LGA 4189	LGA 4677	LGA 4677	LGA 7529	FC LGA 4094	FC LGA 4094	FC LGA 6096	FC LGA 6096	FC LGA 4844	FC LGA 6096	
CPU cores	28	28	48	26	60	64	120	64	64	96	128	64	256	
DRAM	6-channel DDR4	6-channel DDR4	8-channel DDR4	8-channel DDR4	8-channel DDR5	DDR5	DDR5	8-channel DDR4	8-channel DDR4	12-channel DDR5	DDR5	DDR5	TBA	
PCIe	PCIe 3.0	PCIe 3.0	PCIe 3.0	PCIe 4.0	PCIe 5.0	PCIe 5.0	PCIe 5.0	PCIe 4.0	PCIe 4.0	PCIe 5.0	PCIe 5.0	PCIe 5.0	TBA	
CPU TDP	45-165W	165-250W	up to 300W	up to 270W	up to 350W	350-400W	400W+	120-225 W	225-280W	320-400W	320-400W	70-225W	480-600W	

Source: Company data, KGI Research

Figure 11: Revenue of enterprise 400Gbps data center switch suppliers

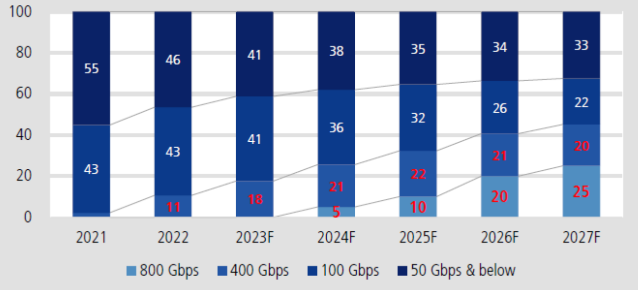
Vendor revenue of enterprise data center switches (by capacity), US$mn (LHS); YoY growth, percent (RHS)



Source：Gartner，KGI estimates

Figure12: Adoption of enterprise 400/800Gbps data center switches trending up

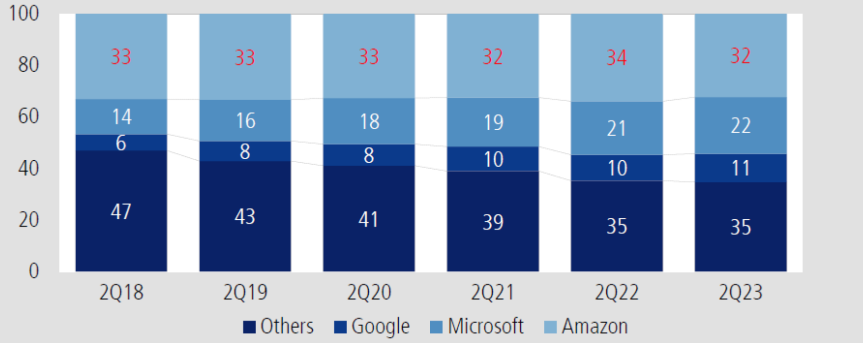
Penetration rate of enterprise data center switches (by capacity), percent



Source：Gartner，KGI estimates

Figure 13: AWS still has largest share of cloud infrastructure services market

2Q23 market share, percent



Source：Synergy Research Group，KGI Research

Figure 14: Decelerating CSP capex growth in 2023F, but will return to YoY growth in 2024F

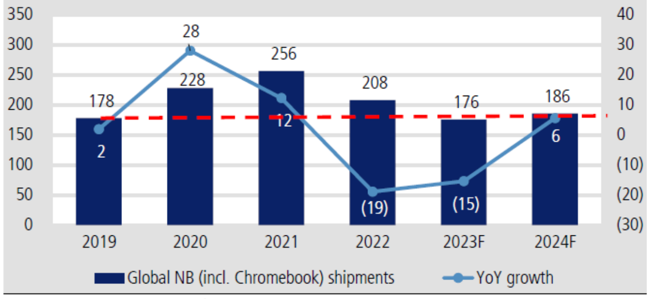
Capex, US$mn	2018	2019	2020	2021	2022	2023F	2024F	2025F
Meta	13,915	15,102	15,115	18,567	31,431	28,021	33,159	35,956
Amazon	13,427	16,861	40,140	61,053	63,645	53,278	59,290	63,843
Microsoft	12,779	13,546	17,592	23,216	24,768	37,009	44,469	44,019
Google	25,139	23,548	22,281	24,640	31,485	31,541	36,267	38,684
Baidu	1,327	931	738	1,689	1,586	1,650	1,708	1,794
Alibaba	7,399	6,517	6,379	8,311	5,014	6,270	7,029	7,192
Tencent	3,356	3,927	5,719	4,808	4,611	4,298	5,012	5,319
Hyperscale subtotal	**77,342**	**80,432**	**107,963**	**142,284**	**162,540**	**162,066**	**186,934**	**196,807**
Apple	12,609	9,247	8,702	10,388	11,692	10,560	11,890	11,975
IBM	3,395	2,286	2,618	2,062	1,346	1,710	1,803	2,055
Oracle	1,468	1,591	1,833	3,118	6,678	8,046	8,854	8,333
Paypal	823	704	866	908	706	730	980	1,087
eBay	651	508	463	444	420	455	494	506
Salesforce	595	643	710	717	798	822	907	1,011
Netflix	174	253	498	525	408	393	465	486
Uber	558	588	616	298	252	244	427	398
Enterprise subtotal	**20,272**	**15,820**	**16,306**	**18,460**	**22,300**	**22,961**	**25,820**	**25,850**
Total	**98,098**	**96,793**	**124,269**	**160,743**	**184,840**	**185,027**	**212,753**	**222,657**

YoY growth, percent	2018	2019	2020	2021	2022	2023F	2024F	2025F
Meta	106.7	8.5	0.1	22.8	69.3	(10.8)	18.3	8.4
Amazon	12.3	25.6	138.1	52.1	4.2	(16.3)	11.3	7.7
Microsoft	29.3	6.0	29.9	32.0	6.7	49.4	20.2	(1.0)
Google	90.7	(6.3)	(5.4)	10.6	27.8	0.2	15.0	6.7
Baidu	87.5	(29.9)	(20.7)	129.1	(6.1)	4.0	3.5	5.0
Alibaba	64.1	(11.9)	(2.1)	30.3	(39.7)	25.0	12.1	2.3
Tencent	86.4	17.0	45.6	(15.9)	(4.1)	(6.8)	16.6	6.1
Hyperscale subtotal	**58.6**	**4.0**	**34.2**	**31.8**	**14.2**	**(0.3)**	**15.3**	**5.3**
Apple	(0.5)	(26.7)	(5.9)	19.4	12.6	(9.7)	12.6	0.7
IBM	5.1	(32.7)	14.5	(21.2)	(34.7)	27.0	5.4	14.0
Oracle	(27.9)	8.4	15.2	70.1	114.2	20.5	10.0	(5.9)
Paypal	23.4	(14.5)	23.0	4.8	(22.2)	3.4	34.3	10.9
eBay	(2.3)	(22.0)	(8.9)	(4.1)	(5.3)	8.2	8.5	2.4
Salesforce	11.4	8.1	10.4	1.0	11.3	3.0	10.4	11.4
Netflix	0.4	45.5	96.8	5.4	(22.3)	(3.6)	18.2	4.7
Uber	(32.0)	5.4	4.8	(51.6)	(15.4)	(3.2)	74.8	(6.8)
Enterprise subtotal	**(2.5)**	**(22.0)**	**3.1**	**13.2**	**20.8**	**3.0**	**12.5**	**0.1**
Total	**40.7**	**(1.3)**	**28.4**	**29.4**	**15.0**	**0.1**	**15.0**	**4.7**

Source: Company data; Bloomberg; KGI Research

Figure 15: Global NB shipments to grow 6% YoY in 2024F; to exceed pre-COVID-19 level of 2019

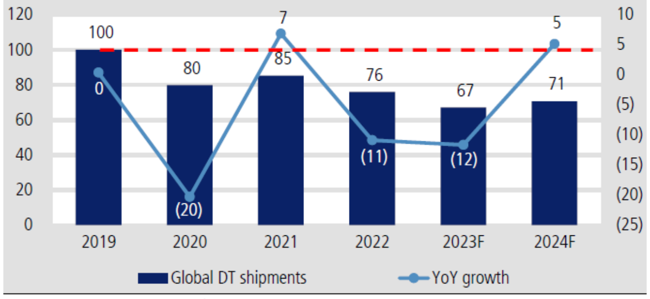
Global NB shipments (including Chromebook), mn units (LHS); YoY growth, percent (RHS)



Source: Gartner, KGI Research estimates

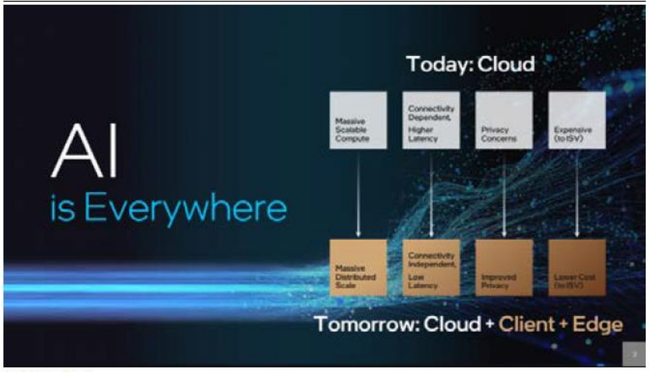
Figure 16: Global DT shipments will also recover in 2024F

Global desktop shipments, mn units (LHS); server ASP, US$ (RHS)



Source: Gartner, KGI Research estimates

Figure 17: Intel’s AI PC acceleration program



AI is Everywhere

Today: Cloud

Massive Scalable Compute

Connectivity Dependent, Higher Latency

Privacy Concerns

Expensive (to ISV)

Massive Distributed Scale

Connectivity Independent, Low Latency

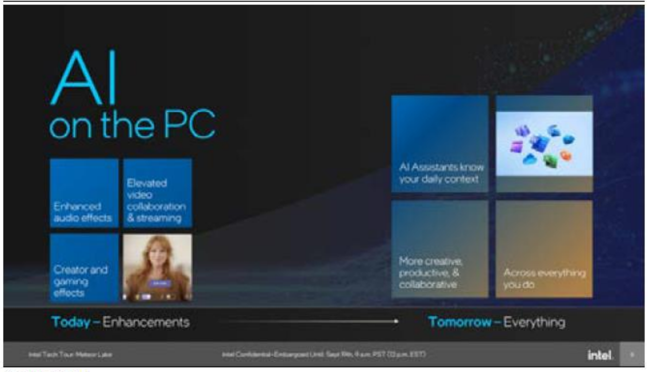
Improved Privacy

Lower Cost (to ISV)

Tomorrow: Cloud + Client + Edge

Source: Intel

Figure 18: Intel’s AI PC acceleration program



AI on the PC

Enhanced audio effects

Elevated video collaboration & streaming

Creator and gaming effects

AI Assistants know your daily context

More creative, productive, & collaborative

Across everything you do

Today – Enhancements

Tomorrow – Everything

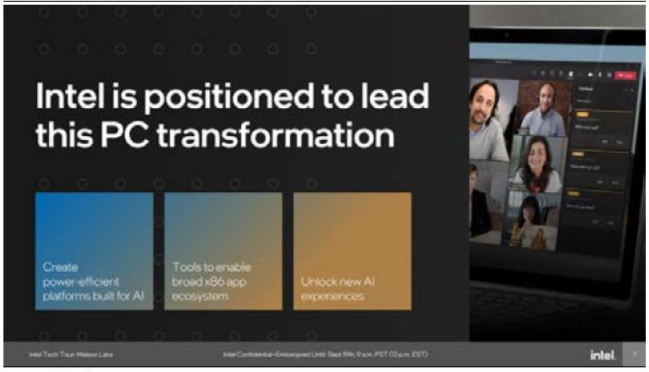
Intel Tech Tour Meteor Lake

Intel Confidential–Embargoed Until Sept 19th, 9 a.m. PST (12 p.m. EST)

intel.

Source: Intel

Figure 19: Intel’s AI PC acceleration program



Intel is positioned to lead this PC transformation

Create power-efficient platforms built for AI

Tools to enable broad x86 app ecosystem

Unlock new AI experiences

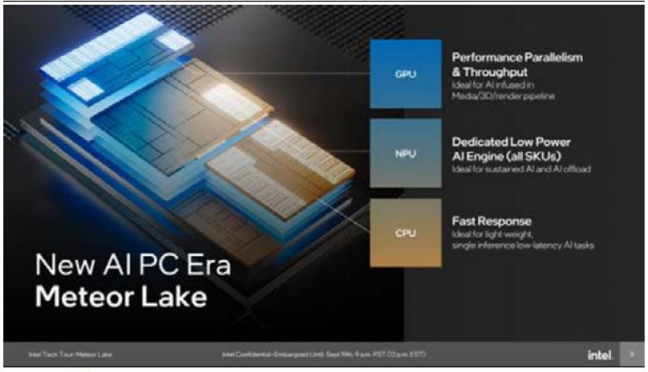
Intel Tech Tour Meteor Lake

Intel Confidential–Embargoed Until Sept 19th, 9 a.m. PST (12 p.m. EST)

intel.

Source: Intel

Figure 20: Meteor Lake CPU to support AI



New AI PC Era Meteor Lake

GPU — Performance Parallelism & Throughput
Ideal for AI infused in Media/3D/render pipeline

NPU — Dedicated Low Power AI Engine (all SKUs)
Ideal for sustained AI and AI offload

CPU — Fast Response
Ideal for light-weight, single inference low-latency AI tasks

Intel Tech Tour Meteor Lake

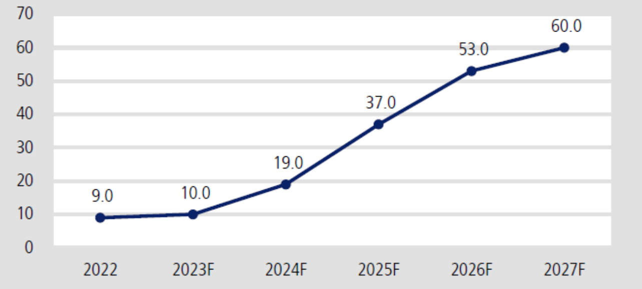
Intel Confidential–Embargoed Until Sept 19th, 9 a.m. PST (12 p.m. EST)

intel.

Source: Intel

Figure 21: Canalys forecasts AI PC penetration of 60% in 2027

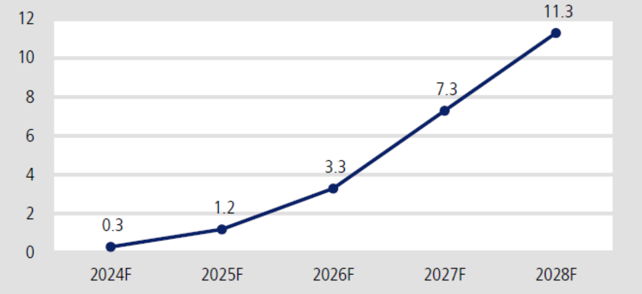
AI PC penetration rate, percent



Source: Canalys

Figure 22: Omdia expects AI NB penetration to be 7.3% in 2027 and over 10% in 2028

AI NB penetration rate, percent



Source: Omdia

Figure 23: Intel/AMD desktop CPU roadmap

	Rocket Lake	Alder Lake	Raptor Lake	Raptor Lake refresh	Meteor Lake	Arrow Lake	Panther Lake	Ryzen 4000 (Renoir)	Ryzen 5000 (Vermeer)	Ryzen 7000 (Raphael)	Ryzen 8000 (Granite Ridge)
Time for launch	1Q21	4Q21	4Q22	3Q23	2024F	2024F	2025F	1Q20	4Q20	3Q22	2024F
Process (node)	14nm+++++	Intel 7 (10nm)	Intel 7 (10nm)	Intel 7 (10nm)	Intel 4 (7nm)	Intel 20A	Intel 18A	TSMC N7	TSMC N7+	TSMC N5	TSMC N3
Microarchitecture (P-Core)	Cypress Cove	Golden Cove	Raptor Cove	Raptor Cove	Redwood Cove	Lion Cove	Cougar Cove	Zen 2	Zen 3	Zen 4	Zen 5
CPU sockets (desktop)	LGA 1200	LGA 1700	LGA 1700	LGA 1700	LGA 1851	LGA 1851	LGA 1851?	AM4 (LGA 1331)	AM4 (LGA 1331)	AM5 (LGA1718)	AM5 (LGA1718)
DRAM	DDR4	DDR4 / DDR5	DDR4 / DDR5	DDR4 / DDR5	DDR5 LPDDR5X	DDR5	TBD	DDR4	DDR4	DDR5	DDR5
PCIe	Gen 4	Gen 5	Gen 5	Gen 5	Gen 5	Gen 5	TBD	Gen 4	Gen 3	Gen 5	Gen 5

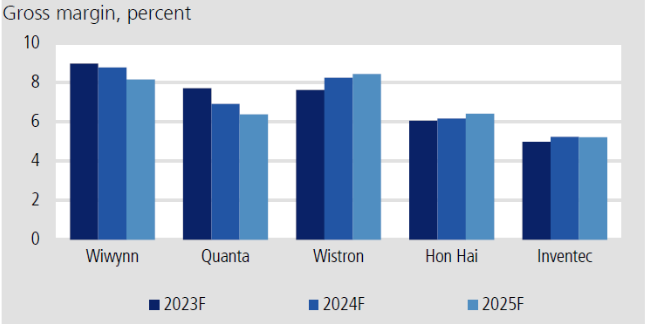
Source: Company data; KGI Research

Figure 24: Intel/AMD NB CPU roadmap

	Rocket Lake	Alder Lake	Raptor Lake	Meteor Lake	Arrow Lake	Lunar Lake	Panther Lake	Ryzen 4000 (Renoir)	Ryzen 5000 (Cezanne)	Ryzen 6000 (Rembrandt)	Ryzen 7000 (Phoenix)	Ryzen 8000 (Strix Point)
Time for launch	1Q21	1H22	1H23	4Q23	2024F	2025F	2025F	1Q20	4Q20	1Q22	1Q23	2024F
Process (node)	14nm+++++	Intel 7 (10nm)	Intel 7 (10nm)	Intel 4 (7nm)	Intel 20A	Intel 18A	Intel 18A	TSMC N7	TSMC N7+	TSMC N6	TSMC N4	TSMC N4
Microarchitecture (P-Core)	Cypress Cove	Golden Cove	Raptor Cove	Redwood Cove	Lion Cove	Lion Cove	TBD	Zen 2	Zen 3	Zen 3+	Zen 4	Zen 5
DRAM	DDR4	DDR4 / DDR5	DDR4 / DDR5	DDR5	TBD	TBD	TBD	DDR4	DDR4	DDR5	DDR5	DDR5
PCIe	Gen 4	Gen 5	Gen 5	Gen 5	Gen 5	TBD	TBD	Gen 4	Gen 3	Gen 4	Gen 5	Gen 5

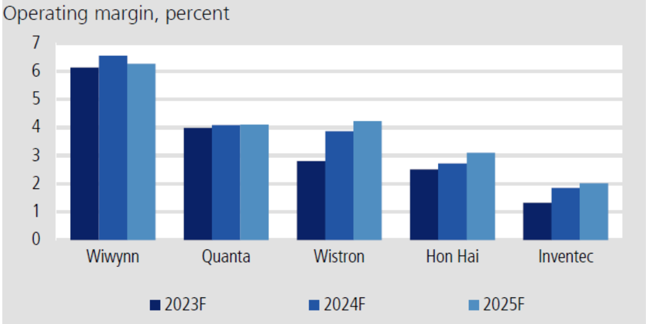
Source: Company data; KGI Research

Figure 25: Increasing AI sales contribution may dilute ODM gross margin



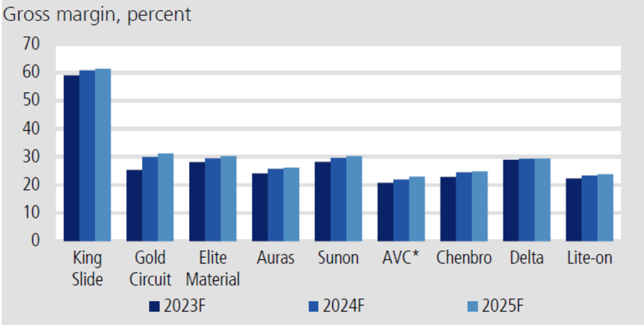
Source: Company data, KGI Research estimates

Figure 26: ODM operating margin to expand YoY in 2024-25F



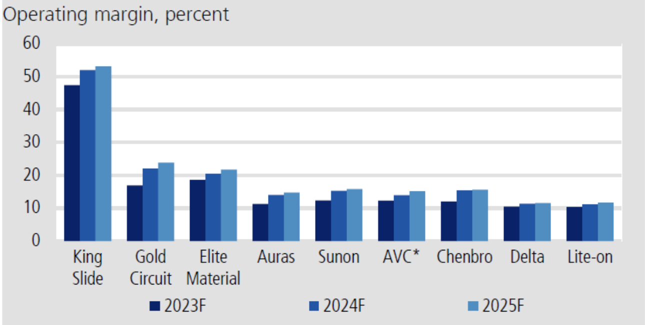
Source: Company data, KGI Research estimates

Figure 27: Component makers' gross margins will benefit from rising AI sales contribution



** Bloomberg consensus*
Source: Company data, KGI Research estimates

Figure 28: Operating margins of component makers to keep expanding



** Bloomberg consensus*
Source: Company data, KGI Research estimates

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.

Disclaimer

All the information contained in this report is not intended for use by persons or entities located in or residing in jurisdictions which restrict the distribution of this information by KGI Asia Limited ("KGI") or an affiliate of KGI. Such information shall not constitute investment advice, or an offer to sell, or an invitation, solicitation or recommendation to subscribe for or invest in any securities or investment products or services nor a distribution of information for any such purpose in any jurisdiction. In particular, the information herein is not for distribution and does not constitute an offer to sell or the solicitation of any offer to buy any securities in the United States of America, or to or for the benefit of United States persons (being residents of the United States of America or partnerships or corporations organised under the laws of the United States of America or any state, territory or possession thereof). All the information contained in this report is for general information and reference purpose only without taking into account of any particular investor's objectives, financial situation or needs. Such information is not intended to provide professional advice and should not be relied upon in that regard.

Some of KGI equity research and earnings estimates are available electronically on www.kgi.com.hk. Please contact your KGI representative for information. The information and opinions in this report are those of KGI internal research activity. KGI does not make any representation or warranty, express or implied, as to the fairness, accuracy, completeness or correctness of the information and opinions contained in this report. The information and opinions contained in this report are subject to change without any notice. No person accepts any liability whatsoever for any loss however arising from any use of this report or its contents. This report is not to be construed as an invitation or offer to buy or sell securities and/or to participate in any investment activity. This report is being supplied solely for informational purposes and may not be redistributed, reproduced or published (in whole or in part) by any means for any purpose without the prior written consent of KGI. Members of the KGI group and their affiliates may provide services to any companies and affiliates of such companies mentioned herein. Members of the KGI group, their affiliates and their directors, officers and employees may from time to time have a position in any securities mentioned herein.