

# ChatGPT

## Unlocking the Potential of Large Language Models

Thematic Research | Global Sector Themes

### Connections Series

- ChatGPT Is the Fastest Growing App Ever:** OpenAI's ChatGPT is reported to have seen growth rates that suggest it is the fastest growing tool/app in history, registering 100 million users by January 2023 after its launch in November 2022. Our report is focused on ChatGPT, OpenAI's GPT family of models, adjacent AI use cases that may emerge by industry/sector, the AI hardware supply chain that will scale it, and finally the companies that are most likely to benefit on the back of sharply increasing AI model deployments. To produce this report, we coordinated with 40+ sector teams, globally.
- Generative AI Is Likely to Be Transformative and Likely to Require Regulation:** We view OpenAI's technology innovations, the progression of generative/conversational AI, and Microsoft's Bing AI (among other apps/services we discuss in great detail in this report) as broadly transformative and predominantly a productivity, cost-cutting, and efficiency tool versus a revenue-generating tool today as it pertains to most industries. In this report we detail over 100 examples of use cases of both ChatGPT and key AI uses cases beyond ChatGPT across 13 industries. However, despite ChatGPT (and other recent AI innovations) being embraced by businesses and organizations globally already, there are risks posed by ChatGPT and some limitations of the GPT Large Language Models (LLMs) have surfaced, likely requiring regulation considering the impact it has on society.
- US Software Top Pick and Adding to CS Top Pick List – MSFT:** Within US Software, MSFT is our Top Pick and we are adding it to the CS Top Pick List following an extensive assessment of MSFT's potential paths of monetization (via their OpenAI ownership/partnership). We found that MSFT has a path to generating ~\$40B of potential revenue (~20% uplift vs FY22) uplift for MSFT and over \$2 of EPS potential (20%+ uplift vs FY22), likely over a period of 5+ years (adding ~3-5%/year to revenue and EPS growth), from the monetization of OpenAI's technology in MSFT's productivity suite and premium GPT-integrated offerings. We discuss Bing AI and other potential revenue streams not included in the \$2+ EPS figure that we believe could offer additional upside.
- US Semiconductors Top Pick – NVDA:** NVDA Graphic Processing Unit (GPU) leads the market for training, with ~95-100% share. In addition, while a majority of inference workloads are today run on CPU (mostly INTC silicon), those workloads are increasingly run on NVDA GPUs, due to superior performance, and will become a further growth driver. Shorter term, we believe the H100 and NVDA's entry into CPU will be catalysts this year. While investors have feared challenges to NVDA's leading market position, none have yet become a significant challenge. We think NVDA's considerable investment in software represents the competitive moat that continues their lead, and the monetization of that software – which has only just begun – represents yet another catalyst. NVDA remains our Top Pick in US Semis due to the AI opportunity.
- Asia Technology Research Top Picks – TSMC, Accton, Wiwynn, and Baidu:** We profile AI acceleration impact on the tech supply chain and catalyst for continued outgrowth for compute, benefiting **TSMC** with 60% share of the compute TAM and supplying key angles (GPU, CPU, networking and AI), advanced equipment (**ASML**, **ASMI**) and hardware suppliers leveraged to cloud through share gains and higher value-add (**Accton** and **Wiwynn**). In China Internet, **Baidu** poised to benefit integrating its Ernie model into search.

#### Research Analysts

**Sami Badri**  
212 538 1727  
ahmedsami.badri@credit-suisse.com

**Randy Abrams, CFA**  
886 2 2715 6366  
randy.abrams@credit-suisse.com

**Chris Caso**  
212 325 3907  
chris.caso@credit-suisse.com

**Shannon Cross**  
shannon.cross@credit-suisse.com

**Stephen Ju**  
stephen.ju@credit-suisse.com

**Adithya Metuku, CFA**  
adithya.metuku@credit-suisse.com

**Akinori Kanemoto**  
akinori.kanemoto@credit-suisse.com

**Chaolien Tseng**  
chaolien.tseng@credit-suisse.com

**Clive Cheung**  
clive.cheung@credit-suisse.com

**Danny Chan**  
danny.chan@credit-suisse.com

**Edward Liu**  
edward.liu@credit-suisse.com

**Fred Lee**  
fred.lee@credit-suisse.com

**Haas Liu**  
haas.liu@credit-suisse.com

**Harvie Chou**  
harvie.chou@credit-suisse.com

**Jasmine Wang**  
jasmine.wang@credit-suisse.com

**Jerry Su**  
jerry.su@credit-suisse.com

**Kenneth Fong**  
kenneth.kc.fong@credit-suisse.com

**Keon Han**  
keon.han@credit-suisse.com

**Kyna Wong**  
kyna.wong@credit-suisse.com

**Lauren Zuo**  
lauren.zuo@credit-suisse.com

**Pauline Chen**  
pauline.chen@credit-suisse.com

**Rich Hilliker**  
rich.hilliker@credit-suisse.com

**Sang Uk Kim**  
sang.kim@credit-suisse.com

**Yufeng Shen**  
yufeng.shen@credit-suisse.com

# Executive Summary

**ChatGPT is the fastest growing app, ever.** OpenAI's ChatGPT is reported to have seen growth rates that suggest it is the fastest growing tool/app in history, outpacing the likes of TikTok, Instagram, and others. ChatGPT got to 1 million users in only 5 days after launch, 57 million users in December 2022 after launching November 30, 2022 and 100 million users by January 2023, all data points supporting its record adoption rate and thereby, its success. Our global research report is focused on ChatGPT, adjacent AI use cases that may emerge by industry/sector, the AI hardware supply chain that will scale it, and finally the companies that are most likely to benefit on the back of sharply increasing AI model deployments. Our report provides investors an understanding of implications beyond just the information technology industry while also providing a foundational understanding of ChatGPT, the large language models it is built on, and finally the regulatory/risks associated with using such a tool.

**Generative Artificial Intelligence likely to be transformative, after further fine tuning, and eventually regulated, in our view.** We view OpenAI's technology innovations, the progression of generative/conversational AI (ChatGPT is a product of generative AI), and Microsoft's Bing AI (among other apps/services we discuss in great detail in this report) as broadly transformative and predominantly a productivity, cost-cutting, efficiency tool versus a revenue-generating tool today as it pertains to most industries. The Information Technology industry is a clear exception given ChatGPT's ability to write and check code in various programming languages, which can dramatically enhance the speed of innovation for software programs. Other industries that require professionals to search/validate facts or inquiries will also see a real-time benefit as ChatGPT is already a helpful tool for several productivity use cases, like idea or content generation (for example, we asked ChatGPT what we should name this report and it gave us 10 options; this one stuck: Unlocking the Potential of Large Language Models). OpenAI's LLMs will be further fine-tuned over time with the next big milestone being OpenAI's GPT-4, a LLM with significantly more parameters versus GPT-3's 175 billion, although potentially less than [reports](#) suggesting 100 trillion. However, despite ChatGPT (and other recent generative AI innovations) being embraced by businesses and organizations globally already, there are risks posted by ChatGPT and some limitations of the GPT LLMs have surfaced. As such, we believe **rules and regulations are needed for AI development and ChatGPT specifically considering the potential impact it has on society.** Although there are no current regulations on ChatGPT yet, relevant discussions have been going on regarding how to make sure the impact from the recent developments are responsible and controlled.

**Thinking ChatGPT use cases by industry and sector – the Information Technology industry leads as the key beneficiary, unsurprisingly.** On industry use cases for ChatGPT and what to expect from AI technologies more broadly, our report includes inputs from more than 40 global sector analyst teams across technology (in the production of this report, US and Asia technology research teams heavily weighed in), healthcare, industrials, business services, materials, real estate, education, government, etc., highlighting if ChatGPT can be used today, what AI use cases are likely to develop impacting each respective industry/sector, and finally specific companies positioned to benefit from AI technology adoption. It is worth highlighting that within the technology industry ~30% of all new code is generated with AI assistance through tools like ChatGPT and GitHub's Copilot, a testament to the value proposition of the technology and a material productivity accelerator in our view.

**Microsoft (MSFT) a beneficiary in ChatGPT's (and OpenAI's) traction and success.** For several reasons, including the fact that they have an ownership stake in OpenAI, a partnership to leverage their models (LLMs), and how they intend on monetizing the technology products/applications over time, **MSFT has a path to generating nearly \$40B of potential revenue (~20% uplift vs FY2022) uplift for MSFT and over \$2 of EPS potential (20%+ uplift vs FY2022), likely over a period of 5+ years (adding ~3-5%/year to revenue and EPS growth),** from the monetization of OpenAI's technology in MSFT's productivity suite. Importantly, MSFT's productivity suite (Office 365) is by far the key value driver for MSFT and also a material driver of broader economic productivity given the size of the MSFT Office installed base, in our view. As it pertains to MSFT, the focus of our report is on MSFT's

technology position with OpenAI's technology, detailing the short- and long-term potential of rolling out GPT capabilities into MSFT office suite (among other apps) and how it intends on monetizing the same apps whereas we believe the majority of the investor and industry attention is focused on MSFT's Azure growth or Bing AI business traction (which we also discuss in this report). Further, we identify both near-term and long-term drivers of revenue/profits, as well as upside optionality in LinkedIn and Bing AI.

**In addition to MSFT being a beneficiary to the AI theme, we also identify NVDA, TSMC, Accton, Wiwynn, Baidu and others that are likely to emerge as well positioned beneficiaries.** Across our global research teams, specific to the technology sector, our top stock picks across global analyst teams include MSFT, NVDA, TSMC, Accton, Wiwynn, and Baidu (nNot ranked in any particular order given the indexation to different sub-sectors).

**The AI hardware and semiconductor supply chain is generally positioned as a beneficiary.** As we highlight in our report, AI models are very compute intensive during training and continue to be compute intensive as users consume compute resources for inference (submitting a prompt into ChatGPT's prompt box for example). For the next five years, there is likely going to be a significant scale-up effort to accommodate AI workloads and their required resources to deliver their development propositions. The report discusses several sub-categories of the AI hardware and semiconductor supply chain that may not be on the typical investors' radar.

**The private generative AI ecosystem is also large and set to benefit. Comparable to ChatGPT and MSFT recent traction with the public,** we also see companies (and services) such as Jasper AI, ChatSonic, and DALL-E (OpenAI development) to see traction given the magnitude of attention ChatGPT has received thus far. We list out several private AI companies in mainly the generative/conversational AI space and their key focus within the industry to aid investors in understanding where we are in the industry and what to expect from this private market ecosystem.

**Within our US Semiconductors coverage we see NVDA as the leading silicon AI enabler.** GPUs have proven to be the widest adopted technology for training AI models and NVDA leads the GPU market for training, with ~95-100% market share. And while investors have been fearful of challenges to NVDA's AI lead – from startups, hyperscalers or INTC – none have materially ramped yet. The majority of inference workloads were traditionally run-on CPUs (mostly INTC silicon), but those workloads are increasingly being run on NVDA GPUs, due to superior performance. We expect GPUs, and specifically NVDA GPUs, to continue to take share of the AI inference market over time.

**NVDA is the main silicon enabler of AI and the main beneficiary long-term.** Shorter-term, product cycles will be a catalyst for CY23/24. The democratization of NVDA silicon through cloud instances means that even small developers can develop the next ChatGPT. We believe that creates open-ended growth which could ultimately expand data generation and the growth trajectory for servers or put AI acceleration into servers on a much faster pace. While it's difficult to accurately upsize the training and inference markets, for their part, NVDA has identified a datacenter TAM opportunity of \$600bn, with \$300bn in hardware (chips/systems) and \$300bn in software. For CY23/24, the two main catalysts for NVDA's datacenter business are the ramp of the H100 (Hopper), and the launch of Hopper/Grace (integrates CPU). Because of the performance gains with this architecture, we expect an order of magnitude of a 50% content increase for Grace Hopper versus the H100.

**Software is NVDA's competitive moat in AI, and the monetization of software represents a significant earnings catalyst.** While hardware secured NVDA's leadership in AI silicon, we think software is what will enable NVDA to maintain that position. The combination of CUDA (proprietary software that underlies many AI programming frameworks and that only operates on NVDA silicon), and their enterprise AI software (essentially the operating system for AI), will be very difficult for prospective competitors to replicate, even if they could match NVDA's silicon performance. NVDA is only just beginning to monetize their AI software and adding software economics could be a significant profit driver. NVDA previously believed their software TAM was \$150bn on the basis of selling that software to on-premise customers.

NVDA recently announced plans to monetize that software via cloud service providers, which will open a new profit stream that will likely be further detailed during their GTC conference later this month.

**The Asia/Europe Technology Supply Chain for AI / ChatGPT will be a key facilitator in a rapidly growing AI world.** Our global tech team summarizes the supply chain implications and company level beneficiaries from the rapid uptake of Chat GPT and its potential to further accelerate adoption for the AI ecosystem. Data center has been one of the fastest growing areas in the tech space and albeit moderating with the macro is still relatively outgrowing many of the consumer areas now facing a post COVID-19 hangover. While the new ChatGPT workloads are not yet offsetting macro to drive upside in supply chain orders, we do view concentrated investments leveraged to acceleration of AI having ability to show over-indexed growth through the industry slowdown. In the medium-term, the uptake of AI services and its industry use cases for revenue generation and cost/capex efficiencies can feed to a new cycle of hardware and semiconductors to maintain innovation and advances.

**AI compute and memory to benefit within the semiconductor sector.** AI training and inference are compute-intensive tasks that should continue to drive semiconductor advances for compute, storage and the transmission of data. The data center compute TAM including accelerators has maintained a 14% CAGR from 2019-24E, with NVIDIA's data center growth at a 50% CAGR and Marvell at 30% CAGR, far outpacing the CPU server growth at a 2% CAGR. An annual penetration increase of 1-2pts of AI accelerated servers from the 8% in 2022 would maintain a 30-35% CAGR for accelerators through 2027. For stocks, the primary beneficiary is NVIDIA (outlined above, detailed elaborately in our report) with over 90% of compute share but we also see TSMC with leverage doubling its contribution from high performance compute (HPC) to over 40% of sales in 2023 and increasing its share of the compute TAM from 20% 60%, now with leverage across leading chip customers promoting CPU, GPU/AI, FPGA and ASIC. Elsewhere in semiconductors AI has potential to improve prospects for server memory for the memory leaders (Samsung, Hynix and Micron), now crossing over mobile at 40% of industry bits, power management into AI boards (MPWR, Infineon and STM), network switch ICs and ASICs (Marvell) and IC design services (Alchip).

**Hardware supply chain to benefit from cloud growth and higher specs.** IDC projects AI servers will grow at a 21% revenue CAGR from 2021-26 vs. 8% CAGR for the total server market, driving AI servers to grow from 18% of server industry revenue in 2022 to 27% of server industry revenue in 2026. The hardware chain should benefit from a richer mix of servers for AI from higher value specs and more thermal design challenges to increase value add for the hardware supply chain, power supply makers and high-end substrates in Japan (Ibiden, Shinko). We note benefits across brands (Lenovo, Gigabyte), ODMs (Accton, Quanta, Wiwynn, Inventec), connectors (Lotes), testing (Chroma), and high-speed interface (Parade). Power supply maker Delta is also seeing rising value supplying a new data center architecture that can better address the rising energy consumption. In China tech, our top picks include server maker Inspur with 30% contribution from AI servers, Wus which is key supplier to US HPC customers, Innolight with 20% share in optical modules and lead supplier to the major US hyperscalers, and Montage which has over 80% of profit from server DRAM interface and companion chips.

**Additional supply chain opportunities.** The supply chain beneficiaries from advances in compute intensity will be a good driver for leading edge silicon, which is now replacing mobile as a key driver for innovation both on advanced manufacturing and high-end packaging integration. We see beneficiaries on advanced SPE front-end and packaging equipment (ASML, ASMI, Besi). We would also highlight on-going geographical shifts in the supply chain which are creating opportunities for ASEAN tech in the data center build-out (Delta Thailand, Inari Amertron).

**Figure 1: Stock Coverage Mentioned in This Report**

Name	Ticker	Sector	Price	TP	Upside	Rat	Mkt Cap US\$mn	CS EPS Gwth %			PE (x)		PB (x)		DY (%) 2023	USD perf 1M	3M
								2022	2023	2024	2023	2024	2023	2024			
IBM	IBM US	US Hardware	129.30	165.00	27.6	O	116,904	15.1	4.2	9.2	13.6	12.5	1.8	1.6	0.0	(4.4)	(13.3)
Alphabet	GOOG US	US Internet	90.06	136.00	51.0	O	1,154,831	4.7	(8.4)	13.4	16.3	14.4	3.5	2.8	0.0	(7.1)	(10.8)
Amazon	AMZN US	US Internet	94.23	150.00	59.2	O	965,599	(6.2)	10.3	44.2	27.3	18.9	5.1	3.5	0.0	(6.3)	(1.3)
Meta	FB UW	US Internet	174.94	220.00	25.8	O	453,556	(16.1)	(2.3)	10.1	12.5	11.4	3.0	2.3	0.0	19.0	45.3
AMD	AMD US	US Semiconductors	78.58	96.00	22.2	O	126,623	25.4	(25.7)	49.8	30.2	20.1	n.a.	n.a.	0.0	8.5	1.4
Intel	INTC US	US Semiconductors	24.93	25.00	0.3	N	103,135	(66.3)	(95.5)	1,079.4	297.4	25.2	1.0	0.9	4.4	(10.8)	(16.4)
Marvell	MRVL US	US Semiconductors	45.15	56.00	24.0	O	38,522	34.8	(5.1)	30.6	22.4	17.2	2.4	2.2	0.5	7.5	(0.6)
Monolithic Power	MPWR US	US Semiconductors	484.29	580.00	19.8	O	22,909	66.6	4.9	17.7	37.2	31.6	11.2	8.1	0.0	18.8	26.8
Nvidia	NVDA US	US Semiconductors	232.16	275.00	18.5	O	573,435	(25.0)	36.3	40.4	51.1	36.4	23.6	18.0	0.1	21.2	35.5
Arista Networks	ANET US	US Networking	138.70	146.00	5.3	N	42,497	59.8	24.7	1.0	24.3	24.0	9.2	6.7	0.0	12.0	(0.8)
Cloudflare	NET US	US Software	60.01	95.00	58.3	O	19,823	- to +	27.0	37.8	371.2	269.4	30.7	29.2	0.0	20.1	21.1
Microsoft	MSFT US	US Software	249.42	285.00	14.3	O	1,856,633	(0.1)	17.8	18.2	23.0	19.5	7.4	6.0	0.0	2.8	(2.1)
Salesforce	CRM US	US Software	163.61	225.00	37.5	O	163,610	2.9	17.5	20.4	28.3	23.5	2.5	2.4	0.0	(0.7)	11.3
Micron	MU US	Memory	57.82	66.00	14.1	O	63,092	37.6	+ to -	- to +	(28.7)	38.1	1.1	1.1	0.8	(6.3)	4.2
Samsung	000660 KS	Memory	89,400	130,000	45.4	O	49,480	(56.6)	+ to -	- to +	(13.9)	14.1	1.0	1.0	1.7	(7.7)	4.7
SK Hynix	005930 KS	Memory	60,600	86,000	41.9	O	275,036	24.0	(63.7)	80.7	21.5	12.1	1.2	1.1	2.4	(10.4)	(4.0)
Alibaba	BABA US	China Internet	87.79	152.00	73.1	O	232,427	(1.1)	22.8	10.5	8.9	8.2	1.3	1.2	0.0	(21.1)	2.2
Baidu	BIDU US	China Internet	137.69	176.00	27.8	O	48,688	6.2	14.5	14.6	13.9	12.4	1.2	1.1	0.0	(0.6)	26.7
JD.com	JD US	China Internet	44.44	87.00	95.8	O	69,110	38.4	30.5	36.3	15.1	11.3	1.7	1.5	0.0	(25.7)	(20.4)
Kingsoft Corp	3888 HK	China Internet	27.55	38.00	37.9	O	4,781	- to +	522.4	45.9	34.6	24.5	1.8	1.7	0.7	(6.6)	8.1
NetEase	9999 HK	China Internet	131.30	189.10	44.0	O	52,095	14.4	(4.0)	13.5	17.2	15.5	3.3	2.8	1.5	(6.8)	20.7
Tencent	700 HK	China Internet	368.80	461.00	25.0	O	449,092	(9.0)	23.9	14.9	21.8	19.3	2.9	2.5	0.0	(4.9)	30.7
Alchip	3661 TT	Asia Semiconductors	1,035	1,400	35.3	O	2,424	21.9	62.3	10.7	21.8	19.3	5.2	4.5	2.5	17.4	15.1
Aspeed	5274 TT	Asia Semiconductors	2,965	2,100	(29.2)	O	3,651	51.6	4.6	29.5	49.9	38.7	22.2	18.6	1.6	42.1	33.8
GUC	3443 TT	Asia Semiconductors	1,180	780	(33.9)	N	5,148	138.3	3.7	19.6	41.7	35.2	17.4	14.7	0.4	49.2	58.7
Montage	688008 CH	Asia Semiconductors	57.77	69.00	19.4	O	9,509	50.0	19.7	64.2	41.7	25.9	6.0	5.0	0.7	(5.6)	(11.6)
TSMC	2330 TT	Asia Semiconductors	522.00	580.00	11.1	O	440,686	59.7	(14.9)	12.6	15.6	13.9	3.8	3.3	2.1	(5.8)	4.2
Accton	2345 TT	Asia Hardware	287.50	330.00	14.8	O	5,242	58.9	9.4	12.4	18.4	16.5	6.9	6.0	3.9	12.0	5.2
Chroma	2360 TT	Asia Hardware	178.50	215.00	20.4	N	2,472	13.7	(16.1)	13.9	17.6	15.6	3.4	3.2	4.2	(4.9)	(9.8)
Gigabyte	2376 TT	Asia Hardware	120.50	130.00	7.9	O	2,484	(54.6)	12.4	12.4	10.5	9.3	2.0	1.8	7.2	(1.2)	6.6
Inventec	2356 TT	Asia Hardware	26.95	29.20	8.3	O	3,148	(16.2)	17.4	9.6	14.0	12.8	1.5	1.5	5.5	3.4	10.8
Lenovo	992 HK	Asia Hardware	7.32	6.50	(11.2)	N	11,310	(19.2)	(9.7)	14.8	7.3	6.4	1.7	1.4	4.8	17.5	13.7
Lotes	3533 TT	Asia Hardware	857.00	930.00	8.5	O	2,982	64.7	(1.4)	23.0	15.8	12.9	3.6	3.1	3.3	9.1	(3.8)
Parade	4966 TT	Asia Hardware	995.00	1,070.00	7.5	N	2,630	(12.8)	(29.7)	40.4	23.6	16.9	3.6	3.2	3.0	3.8	18.2
Quanta	2382 TT	Asia Hardware	79.80	100.00	25.3	O	10,035	(25.6)	13.0	12.0	10.3	9.3	1.7	1.6	7.2	2.5	10.0
Wiwynn	6659 TT	Asia Hardware	940.00	1,100.00	17.0	O	5,351	53.5	4.5	8.0	11.1	10.3	3.6	3.1	5.4	19.3	5.5
Delta Elec.	2308 TT	Asia Components	289.50	295.00	1.9	O	24,483	17.5	1.4	9.4	21.0	18.9	4.6	4.3	2.5	(1.7)	(5.5)
Ibiden	4062 JP	Asia Components	4,830	7,900	63.6	O	4,947	3.1	24.9	12.6	10.9	10.1	1.5	1.3	1.0	(10.5)	(16.4)
Kinsus	3189 TT	Asia Components	111.50	120.00	7.6	N	1,644	91.0	(10.2)	(12.1)	6.8	7.6	1.4	1.3	7.8	(2.0)	(15.0)
NYPCB	8046 TT	Asia Components	240.50	210.00	(12.7)	N	5,060	76.1	(0.6)	1.7	7.8	7.7	2.5	2.2	7.7	(4.2)	(16.5)
SEMC	009150 KS	Asia Components	144,000	185,000	28.5	O	8,177	(2.6)	(35.0)	67.5	18.5	11.2	1.4	1.3	1.5	(6.8)	5.0
Shinko	6967 JP	Asia Components	3,845	7,900	105.5	O	3,810	4.1	10.4	12.7	7.4	6.8	1.6	1.3	(0.9)	(3.8)	
Unimicron	3037 TT	Asia Components	131.50	137.00	4.2	N	6,316	109.6	(2.3)	1.5	6.7	6.6	1.9	1.6	6.0	(5.3)	(18.0)
Innolight	300308 CH	China Hardware	35.91	43.70	21.7	O	4,167	19.8	20.3	21.1	18.7	15.9	2.0	1.7	1.1	21.4	28.8
Inspur	000977 CH	China Hardware	42.88	40.60	(5.3)	O	9,094	(0.0)	26.7	27.2	23.6	18.9	3.2	2.8	0.4	75.3	91.0
WUUS	002463 CH	China Hardware	16.58	17.70	6.8	O	4,556	22.7	23.6	23.5	18.8	15.5	3.3	2.9	1.4	27.4	32.8
ASMI	ASM NA	European Hardware	323.00	378.00	17.0	O	16,637	(45.5)	90.9	28.1	27.7	21.9	5.6	5.2	0.9	3.0	18.9
ASML	ASML NA	European Hardware	584.00	767.00	31.3	O	249,417	(10.2)	40.1	23.1	30.7	25.3	22.4	21.9	1.1	(5.4)	1.0
Ericsson	ERICB SS	European Hardware	58.10	61.00	5.0	U	18,238	(9.5)	(5.3)	16.4	9.0	7.9	1.3	1.2	5.2	(4.3)	(14.7)
Nokia	NOK1VFH	European Hardware	4.38	5.61	28.1	O	26,132	5.9	1.5	(1.0)	10.1	10.4	1.1	1.1	3.0	(2.9)	(7.9)
Soitec	SOI FP	European Hardware	141.20	219.00	55.1	O	5,324	(3.0)	24.3	46.1	19.4	13.5	3.1	2.5	0.0	(1.1)	(11.5)
Delta Thailand	DELTA TB	ASEAN Tech	986.00	570.00	(42.2)	U	34,921	118.7	12.6	18.5	77.9	67.9	18.9	16.0	0.4	2.2	27.5
Inari	INRI MK	ASEAN Tech	2.42	2.80	15.7	N	2,016	(11.3)	24.1	21.1	21.1	18.0	3.3	3.3	4.3	(13.9)	(15.7)
Pentamaster	PENT MK	ASEAN Tech	4.90	5.20	6.1	N	779	6.3	33.2	22.8	32.0	26.5	4.8	4.1	0.3	(7.0)	2.2
							Average:	17.9	17.5	42.5	31.7	22.8	5.4	4.6	2.0	3.9	7.5

Source: Company data, Refinitiv, Credit Suisse estimates.

# Table of Contents

<b>Executive Summary</b>	<b>2</b>
<hr/>	
<b>Generative/Conversational AI Has Arrived</b>	<b>10</b>
What Are the Relevant AI Advancements Behind ChatGPT? .....	12
How does a LLM Work? .....	13
Challenges of Large Language Models.....	13
Bigger Is Not Always Better.....	14
<hr/>	
<b>ChatGPT Is a Tool</b>	<b>16</b>
ChatGPT Was Developed by OpenAI, An AI Institute .....	16
ChatGPT's Resource Usage Significant.....	23
How Much Would LLM-powered Search Cost?.....	23
Microsoft and OpenAI Background.....	26
ChatGPT's User Growth Has Been Staggering .....	28
ChatGPT CS Revenue Forecast & Model	29
<hr/>	
<b>Microsoft a Very Direct Beneficiary of ChatGPT but The Gain Is In Productivity</b>	<b>32</b>
Direct Impacts of ChatGPT/GPT-3 .....	32
Driver #1: Direct ChatGPT Revenue & Profitability	33
Driver #2: GitHub Copilot	34
Driver #3: Teams Premium	35
Driver #4: Viva Sales	36
Driver #5: Azure (ChatGPT Compute Rev. & Azure OpenAI)	37
Upside Optionality: Bing AI	39
Upside Optionality: LinkedIn	40
Long-Term Impacts of ChatGPT/GPT-3 .....	42
Bringing GPT to MSFT: The True Value Is in Microsoft Office	42
<hr/>	
<b>The AI Landscape – Key Milestones to Note</b>	<b>46</b>
Google – Covered by Stephen Ju, US Internet Analyst.....	46
Nvidia – Covered by Chris Caso, US Semiconductors Analyst .....	48
IBM – Covered by Shannon Cross, US IT Hardware .....	49
Meta – Covered by Stephen Ju, US Internet Analyst.....	50
Amazon – Covered by Stephen Ju, US Internet Analyst .....	51
China Internet – Covered by Kenneth Fong, China Internet Analyst.....	52
Poised to Be a Game-changer for China Internet	52
Implications for Baidu	54
Implications on Other Chinese Internet Companies	55
Regulation	56
Key challenges	57

<b>The GPT and AI Ecosystem Is Already Large</b>	<b>58</b>
<hr/>	
<b>Industries Set to Benefit from ChatGPT</b>	<b>66</b>
<hr/>	
Four Different Ways to Use ChatGPT .....	66
Information Technology .....	67
Key ChatGPT Use Cases in IT	68
AI Use Cases Beyond ChatGPT in the IT Industry	68
IT Sector Coverage Implications	68
Business Services .....	85
Key ChatGPT Use Cases in Business Services	85
AI Use Cases, Beyond ChatGPT in the Business Services Industry	85
Sector Coverage Implications	85
Financial Services .....	87
Key ChatGPT use cases in Financial Services include	87
AI Use Cases, Beyond ChatGPT	88
Sector Coverage Implications	89
Education .....	92
Key ChatGPT use cases in Education include:	92
Sector Coverage Implications	93
Healthcare .....	94
Key ChatGPT Use Cases in Healthcare	94
AI Use Cases, Beyond ChatGPT in the Healthcare Industry	94
Sector Coverage Implications	95
Industrials .....	98
Key ChatGPT Use Cases in Industrials	98
AI Use Cases, Beyond ChatGPT	98
Sector Coverage Implications	98
Consumer Discretionary & Staples .....	101
Key ChatGPT Use Cases in Consumer Discretionary & Staples	101
AI Use Cases, Beyond ChatGPT	101
Sector Coverage Implications	101
Real Estate .....	103
Key ChatGPT Use Cases in Real Estate include	103
AI Use Cases, Beyond ChatGPT	103
Sector Coverage Implications	103
Energy .....	105
Key ChatGPT use cases in Energy include	105
AI Use Cases, Beyond ChatGPT	105
Utilities .....	107
Key ChatGPT use cases in Utilities include	107

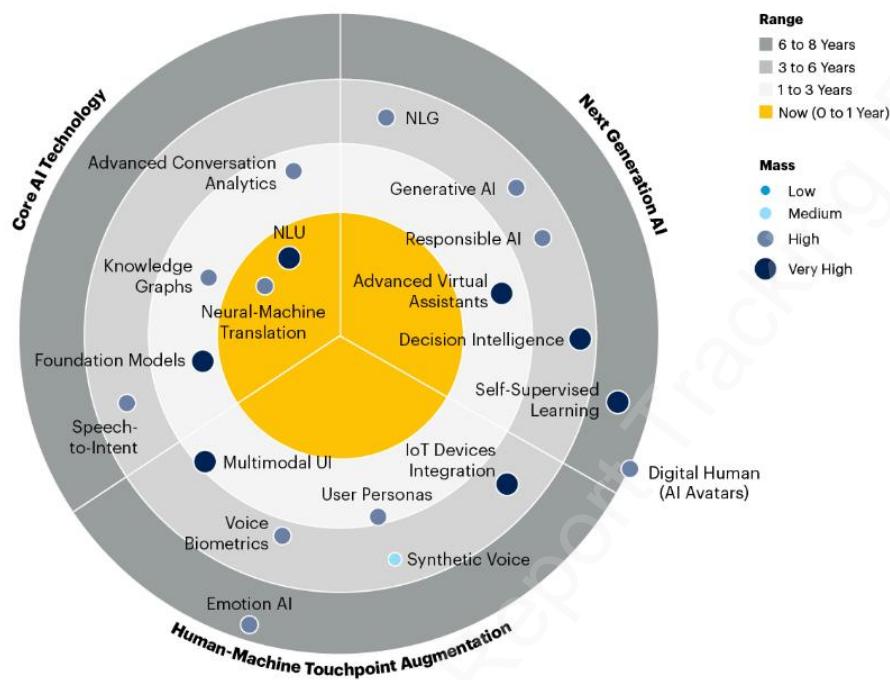
AI Use Cases, Beyond ChatGPT	107
Sector Coverage Implications	107
Materials .....	108
Key ChatGPT use cases in Materials include	108
AI Use Cases, Beyond ChatGPT	108
Sector Coverage Implications	108
Communications Industry.....	109
Key ChatGPT Use Cases in Communications include	109
AI Use Cases, Beyond ChatGPT	109
Sector Coverage Implications	109
Governments .....	112
Key ChatGPT use cases in Government include	112
AI Use Cases, Beyond ChatGPT	113
<b>Supply chain for AI / Chat GPT</b>	<b>114</b>
Semiconductors .....	115
Compute TAM to be lifted by an inflection in AI use cases	115
AI investments to be a growing portion of spend	117
AI Training and inference both driving more compute and shifts to accelerators	118
AI Training dominated by GPUs, ASICs/FPGAs competitors gaining in inference	120
GPU leads AI training, gaining in inference	122
Inference shifting from x86 CPUs toward GPU and other accelerators	123
NVIDIA continuing to advance its solutions to power AI	127
Semiconductor suppliers trying to break NVIDIA's lead	130
Battleground emerging between custom ASICs	133
Asia Semiconductors.....	145
AI to drive a further inflection in TSMC's HPC business	145
Scaling still providing some benefits albeit diminishing	148
Advanced packaging offsets slowing transistor scaling	150
Back-end service providers will also have a role to play	152
Back-end equipment suppliers to benefit upgrading advanced packaging	152
Asian IC Design: ASpeed seeing growth from its higher content with its controller and peripheral chips	155
IC Design Service: Alchip, GUC and Socionext set to benefit from growing chipset customization	156
Competitive landscape for Design Services	157
Memory.....	162
Overcoming the memory bottleneck for efficient machine learning (ML) (Keon Han)	162
IC Substrates.....	166
New data center architecture for better energy efficiency (P. Chen)	168

Asia Cloud IT Infrastructure Sector .....	170
Asia Hardware Stock implications	175
Power supply - New data center architecture for better energy efficiency .....	177
Industrial PC looking for AI opportunities.....	179
China Technology Sector.....	180
European Technology Hardware has strong potential to benefit from generative AI .....	186
European Hardware stock implications	187
ASEAN – Well positioned and AI to drive innovation among equipment and back-end companies.....	189
Equipment makers weaving in more AI into their products	191
Back-end players upgrading their offerings	191
<b>Risks and Regulatory Concerns with ChatGPT and AI Technologies</b>	<b>193</b>
Regulators Across the Globe Are Taking Action.....	196
<b>Microsoft</b>	<b>198</b>
Valuation .....	199
Credit Suisse Financial Model .....	200
<b>NVIDIA Corporation</b>	<b>202</b>
GPU leads AI training, gaining in inference	203
NVIDIA continuing to advance its solutions to power AI	203
<b>Wiwynn Corporation</b>	<b>210</b>
<b>Accton</b>	<b>211</b>
<b>Taiwan Semiconductor Manufacturing</b>	<b>212</b>
<b>Baidu</b>	<b>213</b>

# Generative/Conversational AI Has Arrived

There has been a clear acceleration in AI since ChatGPT's launch measured by media publications and technology executive discussion points, but what has become less clear is where ChatGPT actually fits in the AI technology ecosystem – specifically the Conversational AI ecosystem – and where the AI technology industry is actually going. In Figure 2, Gartner illustrates where AI technology is mapped today and which technologies have either gained mass and where we are for each sub-category from a “time to mainstream” standpoint. Below we expand on some key areas of the AI ecosystem for some foundational content before diving deeper into Generative/Conversational AI which is the type of AI ChatGPT falls under.

**Figure 2: Emerging Technologies Around Conversational AI**



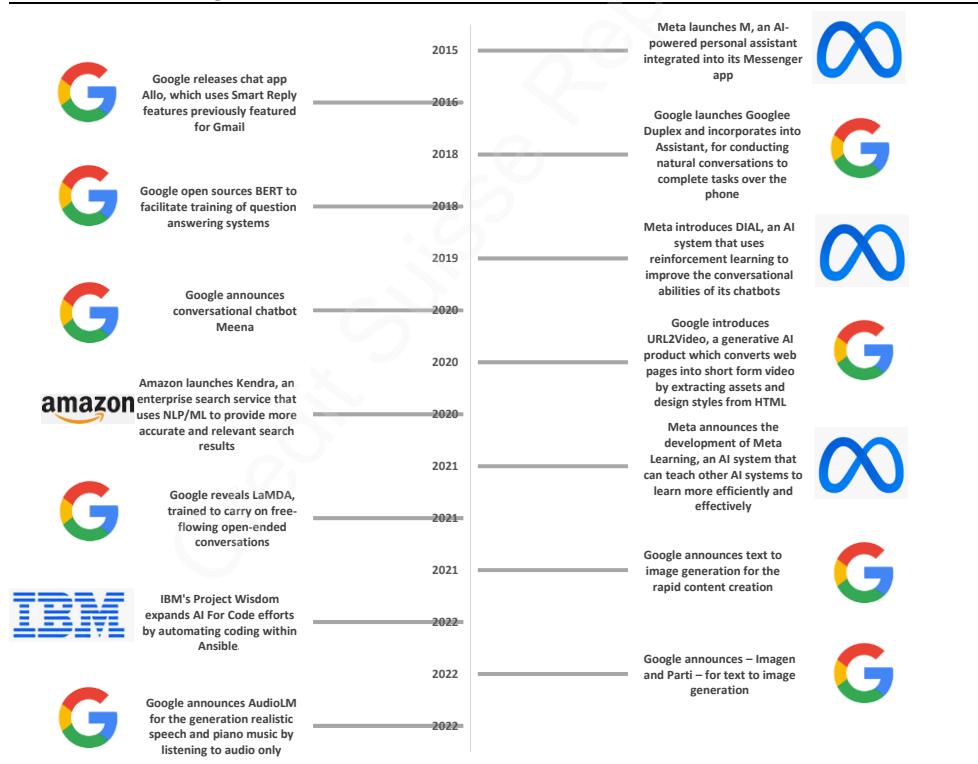
Source: Gartner. AI = Artificial Intelligence; IoT = Internet of Things; NLG = Natural Language Generation; NLU = Natural Language Understanding; UI = User Interface.

- **Generative AI:** By leveraging machine learning (ML) techniques such as deep learning and neural networks, Generative AI can recognize patterns and create new outputs based on the understanding. The key difference between generative models and others (e.g., predictive and classification models) is that the former is designed to create outputs that are entirely new, rather than simply predicting or categorizing data. Although we remain in very early innings of AI technology development maturity, the focus of our report is largely on ChatGPT and Generative AI, which according to Gartner has gained a "high" level of "mass" and is really expected to go mainstream within 3 to 6 years from today. This actually aligns very closely to how OpenAI has described the effectiveness of their model's accuracies as they gain more traction and receive more training/optimizations.
  - **Responsive AI:** As the name implies, Responsive AI is designed to interact with users in a flexible and adaptable manner. Such an AI system can adjust its behavior depending on changing circumstances and can improve over time based on user feedback.
  - **Natural Language Generation (NLG):** NLG focuses on the automatic generation of human-like language, such as text or speech. NLG is often used to create reports, summaries, and other written content, covering a wide range of applications. NLG can also be used for the automation of repetitive tasks such as generating personalized emails or chatbot responses, etc.

- **Advanced Virtual Assistants (VA):** Advanced VAs are VA systems that use AI technologies such as Natural Language Processing (NLP) and computer vision to provide sophisticated and personalized assistance to their users. Unlike simple VA, advanced VA can understand context, remember previous interactions, and improve from data/feedback. Some examples include Apple's Siri, Amazon's Alexa, and Google Assistant.
- **Decision Intelligence:** Powered by AI and other advanced analytics techniques, decision intelligence systems aim to help people and organizations to make better decisions by providing better information and more accurate predictions.
- **Self-supervised Learning:** Self-supervised Learning is a type of ML, in which the algorithm learns to make predictions about a data set without being explicitly told what the correct answers are. Instead, the model is trained to capture patterns or relationships in the data set and use it as the base to make predictions. In other words, the data set itself is used to provide the supervision for the learning process. Self-supervised learning has become popular these days, particularly in NLP.
- **Digital Human (AI Avatars):** An AI Avatar is a computer-generated virtual character that is designed to interact with humans using AI technologies. AI avatars can be in a variety of forms, including animated characters, chatbots, and voice assistants. But regardless of the form, the goal is to provide more engaging and personalized interactions with users.

Leading technology companies like Google, IBM, and Meta have been focusing on the AI developments and leading AI advancements as they have become the source of competitive advantage and foundation of future technologies/businesses. Although many [major AI milestones](#) have been achieved over the years, we have highlighted below key milestones achieved by these companies around Generative/Conversational AI and adjacent chatbot technologies – an area with most of the attentions these days due to the advent of ChatGPT. As we can see from the timeline in Figure 3, Google has released the greatest number of relevant research papers/projects historically.

**Figure 3: Key Milestones Achieved by Leading Tech Companies on Generative AI and Chatbot Technologies**



Source: Company data.

## What Are the Relevant AI Advancements Behind ChatGPT?

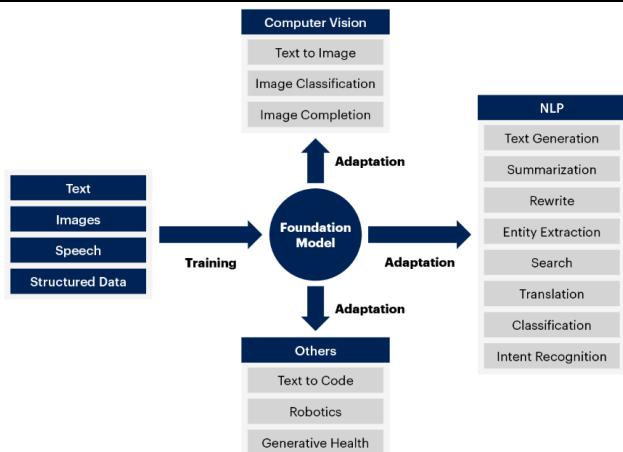
The increasing popularity of ChatGPT has proved that ongoing innovations in conversational AI are materializing at a faster-than-expected pace given the industry's expectations for timing of mainstream adoption are not expected for another 3-6yrs whereas ChatGPT Plus is here and now, with ~100M users (discussed in our ChatGPT model forecast section in more detail). Importantly, ChatGPT isn't the by-product of a single LLM or AI technology (see Figure 7), but clearly a product of multiple conversational AI techniques and technologies as well as multiple LLMs. One of the key drivers of ChatGPT has been the propelled advancements of foundation models, knowledge graphs, and reinforcement learning with human techniques, all integrated to deliver ChatGPT. Among these technologies, foundation models are a major AI advancement, which are ready to be incorporated into many software technologies, transforming the conversational capabilities of software and advanced virtual assistants (VA). Gartner projects that foundation models will underpin 60% of NLP use cases by 2027 vs. less than 10% in 2022. Although the original interests in foundation models were focused on NLP, they are quite adaptive and could be used across other use cases such as translating from language to code (e.g., OpenAI Codex) or from language to image (e.g., OpenAI DALL-E 2), enabling multi-modal AI. Additionally, there are research efforts focusing on extending foundation models beyond NLP to more use cases; it is this area that we believe to be the most technologically innovative.

### What Are Foundation Models?

Foundation models are primarily LLMs that are designed to replace task-specific models. Trained by extensive unlabeled data sets in a self-supervised manner, foundation models are able to perform different tasks and can be used on various use cases and applications as shown in Figure 4. There are some key advantages with Foundation Models that have made such advancements prominent:

- **Scale:** Models can be effective in zero-shot scenarios or few-shot scenarios, where little domain-specific training data are available (see Figure 5 from examples of zero-shot vs. few-shot scenarios), resulting in better performance in reading comprehension, sentiment analysis, and fact checking, etc.
- **Accuracy:** Foundation models have been able to perform with high accuracy over both NLP and non-NLP tasks.
- **Domain Adaptability:** Foundational Models can be virtually applied to any kind of sequential data, making it not only useful in NLP tasks, but also valuable in anomaly detection and identifying patterns, etc.

**Figure 4: Key Characteristics and Applications of Foundation Models**

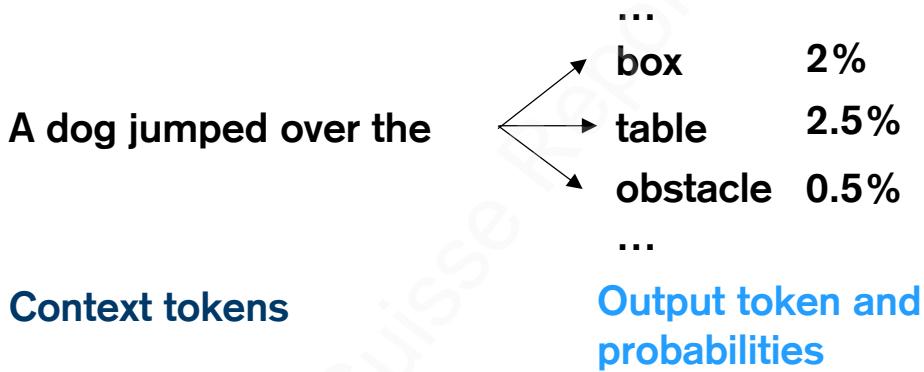


Source: Gartner.

## How does a LLM Work?

To generate text, language models repeatedly sample new tokens (word fragments) based on the output token probabilities (Figure 6). For ChatGPT, the algorithm begins with an initial prompt that includes the user's query as context and generates tokens to construct the response. As each new token is generated, it is appended to the context window to inform the next iteration.

**Figure 6: Illustration of Input Context and Output for an LLM**



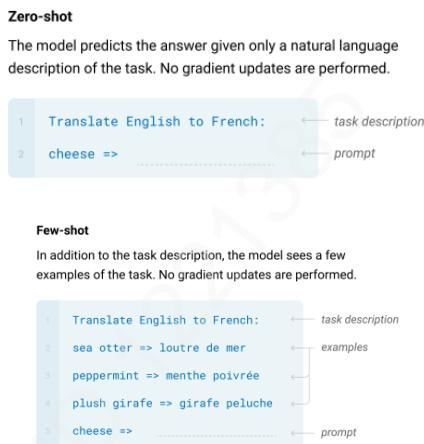
Source: Credit Suisse Research.

LLMs are not new concepts and have existed for decades. However, the recent implementation of deep neural networks has improved the performance significantly, with billions of parameters built in the model that are used for both training and making predictions. Note, these operations need to be processed through graphics processing units (GPUs), tensor processing units (TPUs), and other specialized chips, posing some challenges as LLMs continue to become larger and more complex.

## Challenges of Large Language Models

LLMs have been unleashing significant potential and creating synergies in areas such as search engines, NLP, healthcare, robotics, code generation, etc., for some time, and ChatGPT has become the most popular application of LLM. However, LLMs do face some challenges:

**Figure 5: Examples of Zero-shot vs. Few-shot for In-context Learning**



Source: Gartner.

- **High Costs to Maintain and Scale:** As we analyze in this report, maintaining LLMs can be difficult and expensive, as does the scaling due to the higher computing costs. It is here that having a cloud partner, like Microsoft Azure, is key to developing robust LLMs.
- **It Takes a Long Time to Deploy:** Given the complexity of LLM projects and the massive data sets required to train them, it usually requires months (if not longer) to train LLMs, which also represent intensive investments toward the project.
- **Low Data Accessibility:** It is not always easy for developers and enterprises to access large-enough data sets, resulting in underfitting – when the model is not trained enough to capture enough patterns in the data to produce acceptable outputs.
- **Easily Over-Trained:** Developers also want to avoid the model being overly trained. For example, if a model is trained heavily by docs from a niche space (e.g., legal docs, financial statements, research papers of certain area, etc.), it may easily become too specialized in a space and not able to generate text on other topics or may produce results of lesser trained topics in the context of other industry domains, creating a bias.
- **Lacking Experts:** Deploying the model requires certain technical expertise, including a strong understanding of deep learning, transformer models, and hardware, etc.

## Bigger Is Not Always Better

In summary, large LLMs that require too much data, too much scale, and too much training do not always result in good models (this is especially highlighted in Figure 16). Take for example the Google Chinchilla LLM model that we highlight in Figure 17 which uses comparable compute as GPT-3, but it has less model loss, which means it is more accurate.

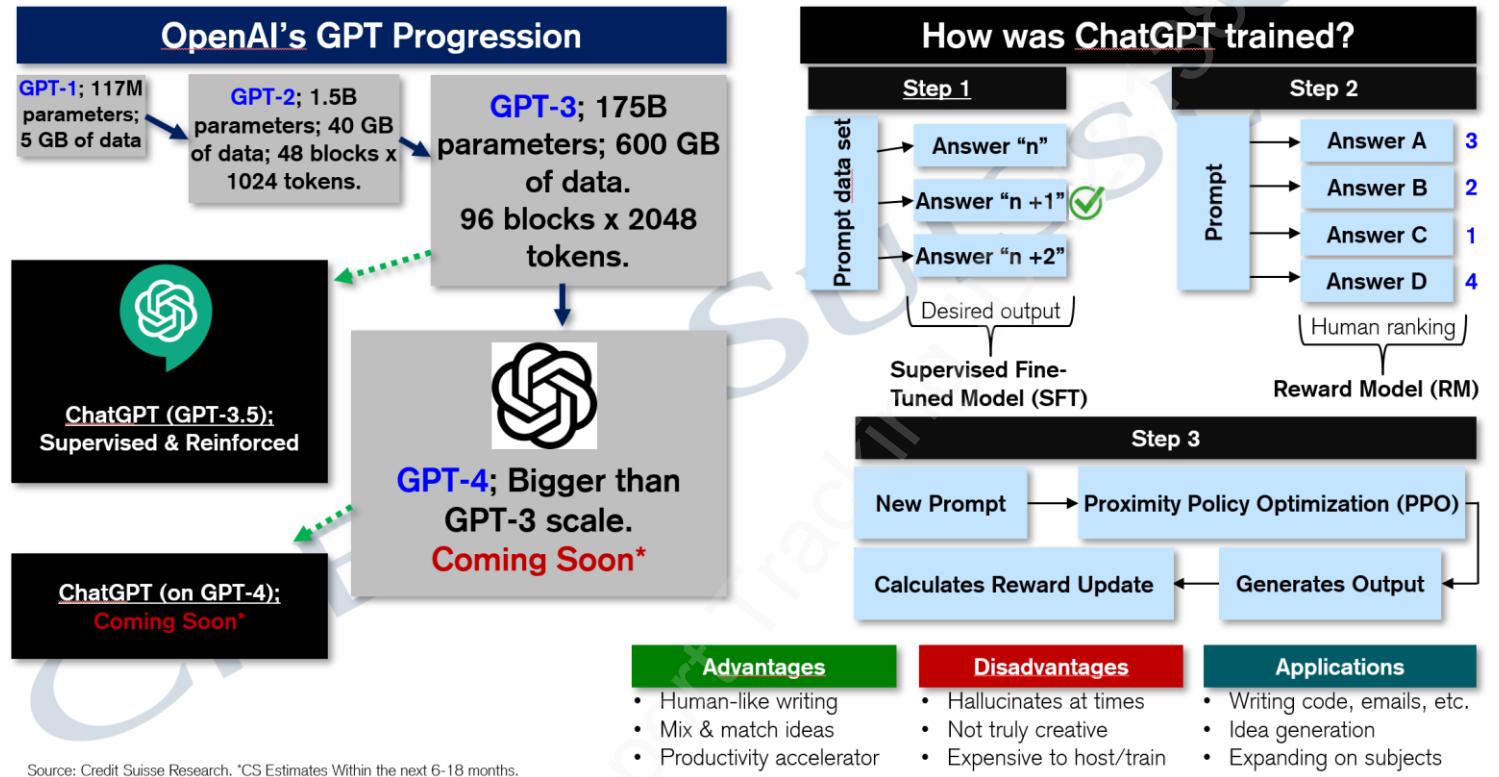
Figure 7: Cheat Sheet on ChatGPT's Progression, Training, Advantages, Disadvantages, and Credit Suisse Expectations

## ChatGPT: GPT Fine-Tuned for Conversations

Generative: predicting next word (language model)

Pre-Trained: previously trained on large amounts of data

Transformers: encoder-decoder based neural network



Source: Company data, Credit Suisse Research.

## ChatGPT Is a Tool

ChatGPT is a natural language processing tool driven by AI technology that allows a user to have human-like conversations. ChatGPT is built on a large language model (LLM) and part of Generative Artificial Intelligence technology. An LLM is a deep learning algorithm that can recognize, summarize, translate, predict and generate text and other content based on knowledge gained from massive data sets. AI applications are summarizing articles, writing stories and engaging in long conversations — and LLMs are doing the heavy lifting. LLMs are among the most successful applications of transformer models, on which ChatGPT is built. They aren't just for teaching AIs human languages, but for understanding a wide variety of subject disciplines, including the understanding of proteins, writing software code, and creating graphics based on qualitative text descriptions. In addition to accelerating natural language processing applications — like translation, chatbots, and AI assistants — LLMs are used in software development, healthcare, and use cases in many other fields which we discuss in great detail later in our report.

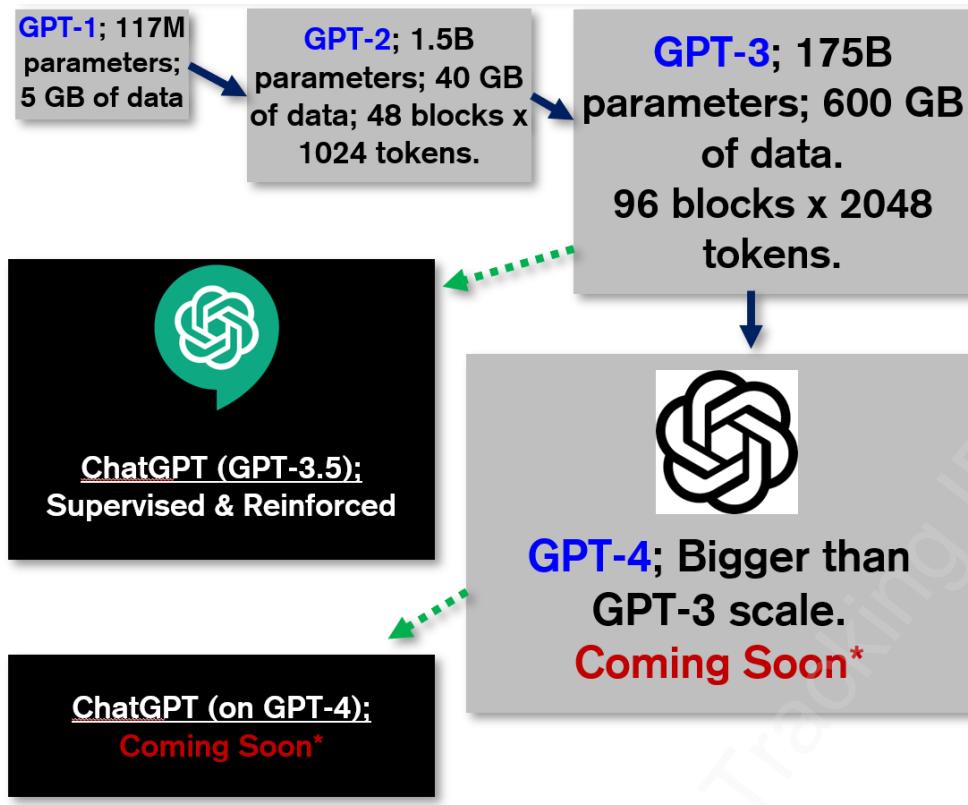
**Figure 8: ChatGPT Intro Screen**

ChatGPT		
Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

Source: OpenAI.

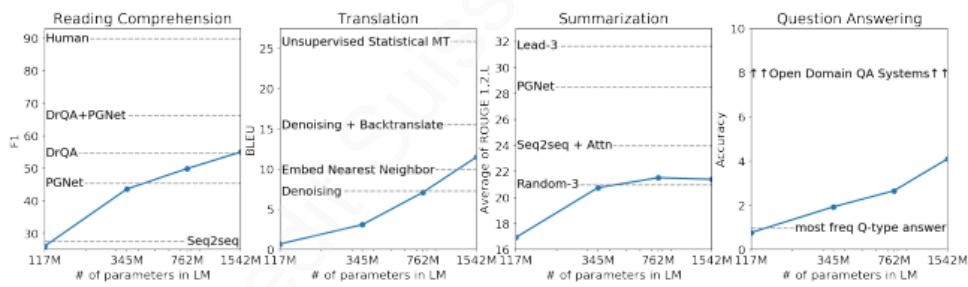
## ChatGPT Was Developed by OpenAI, An AI Institute

OpenAI was founded in December 2015 as a non-profit with a \$1B commitment from Sam Altman, Greg Brockman (former CTO of Stripe), Elon Musk, Reid Hoffman, Jessica Livingston, Peter Thiel, Amazon Web Services (AWS), Infosys, and YC Research—the company's co-chairs at the time were Sam Altman and Elon Musk (who resigned his seat in 2018 due to potential conflicts of interest). Sam Altman, the current CEO of OpenAI, after leaving college in 2005 started Loopt, a geolocation company, before becoming president of Y Combinator and ultimately co-founding OpenAI in 2015. OpenAI's family of models include: GPT for language generation, Codex for code generation, and DALL-E for image generation and editing.

**Figure 9: Evolution of OpenAI's GPT**

Source: Company data, Credit Suisse Research. \*CS estimates within the next 6-18 months.

OpenAI released the original white paper on a generative pre-training transformer (GPT) of a language model in June 2018 with the original (GPT-1) trained on 117 million parameters. Following this release, OpenAI produced and released two successor versions of GPT-2, a partial version in August 2019 with 774 million parameters and a full version in November 2019 with 1.5 billion parameters as the higher parameter model received a slightly human-perceived credibility score along with better accuracy and results (see Figure 10).

**Figure 10: OpenAI GPT-2 Model Performance by Task – Parameter Count Comparison**

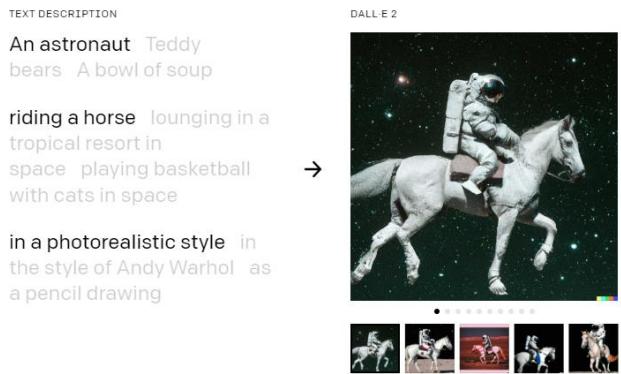
Source: OpenAI.

In June 2020, OpenAI released GPT-3 as a service, powered by a 175-billion-parameter model (over 100x more parameters than GPT-2) that was trained on 300 billion tokens (word fragments) that can generate text and code with short written prompts. GPT-3 can be used to create not only human language text but also anything with a text structure (for example, summarizations and programming code). GPT-3 was a more robust version of GPT-2 that solved a key issue of GPT-2 which was poor performance in niche topics and specialty tasks (i.e., music). In September 2020 Microsoft acquired an exclusive license to GPT-3, meaning that while the public could utilize the application programming interface (API) to receive output,

**Microsoft is the only legal entity other than OpenAI with access to the underlying source code.**

Following the release of GPT-3, OpenAI released two additional AI models based off GPT-3, addressing different modalities—OpenAI Codex (released in August 2021) for natural language to code and DALL-E for natural language to images (released in January 2021) along with a successor version, DALL-E 2 (released in April 2022). While we address key use cases and provide additional detail on both of these models further in this report, we note that both OpenAI Codex and DALL-E 1/2 remain fully proprietary to OpenAI (the underlying source code) but are available for API use through subscriptions.

**Figure 11: OpenAI - DALL-E 2**



Source: Company data, Credit Suisse

**Figure 12: OpenAI – Codex Demo**

The image shows the OpenAI Codex demo interface. On the left, there is a text input field containing the instruction: "Make it be the size of the rocketship times 0.75". To the right of the text field is a generated image of a large, irregularly shaped yellow asteroid against a dark space background.

On the far right, there is a block of JavaScript code that generates the image:

```

text.style.left = rocketship.offsetLeft + 'px';
text.style.top = rocketship.offsetTop + 'px';

document.body.appendChild(text);

xSpeed = 20;
setInterval(function() {
  xSpeed += 5;

  document.body.removeChild(text)
}, 250);
}

/* Now add an image of an asteroid:
https://d.newsweek.com/en/full/1/721338/asteroid.jpg?
w=1680&h=1680&q=88&f=9e82d3Sc9de9ba82b3fcfa7705be325b */
var asteroid =
document.createElement('img');
asteroid.src =
  'https://d.newsweek.com/en/full/1/721338/asteroid.jpg?
  w=1680&h=1680&q=88&f=9e82d3Sc9de9ba82b3fcfa7705be325b';
document.body.appendChild(asteroid);
  
```

Source: Company data, Credit Suisse

On November 30, 2022 Open AI released ChatGPT, a chatbot based on a fine-tuned version of GPT-3.5 using both supervised (human trainers provided both sides of conversations) and reinforcement learning (human trainers ranked responses that the chatbot produced). While ChatGPT is largely based on GPT-3 (which is a family of models, not just a single LLM, listed in Figure 14), it was specifically designed for chatbot applications and is generally better at producing responses in a conversational context while GPT-3 is a more general-purpose model with a much wider set of use cases (i.e., content creation). A key limitation of ChatGPT (and LLMs in general) is that the outputs are only as up to date as the training data (see a list of its training data in Figure 13), so in the case of ChatGPT the chatbot has limited knowledge of events that happened after 2021 because that is where the training data stopped. Another important distinction, while OpenAI has made the GPT-3 model publicly available, the underlying source code for ChatGPT is not publicly available and there is no ability for customization – ChatGPT is fully proprietary and there is no open-source component.

**Figure 13: GPT-3 Has Been Trained Using Various Data Sources**

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Source: [Language Models are Few-Shot Learners](#)

**Figure 14: GPT-3 Consists of a Family of Models Rather Than a Single Model**

Model Name	<i>n</i> params	<i>n</i> layers	<i>d</i> <sub>model</sub>	<i>n</i> heads	<i>d</i> <sub>head</sub>	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	$6.0 \times 10^{-4}$
GPT-3 Medium	350M	24	1024	16	64	0.5M	$3.0 \times 10^{-4}$
GPT-3 Large	760M	24	1536	16	96	0.5M	$2.5 \times 10^{-4}$
GPT-3 XL	1.3B	24	2048	24	128	1M	$2.0 \times 10^{-4}$
GPT-3 2.7B	2.7B	32	2560	32	80	1M	$1.6 \times 10^{-4}$
GPT-3 6.7B	6.7B	32	4096	32	128	2M	$1.2 \times 10^{-4}$
GPT-3 13B	13.0B	40	5140	40	128	2M	$1.0 \times 10^{-4}$
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	$0.6 \times 10^{-4}$

Source: [OpenAI GPT-3: Everything You Need to Know](#)

In late February 2023 OpenAI also launched Foundry, a developer platform, to allow customers to run its newer models (i.e., GPT-3.5) on dedicated capacity. Foundry would deliver a “static allocation” of compute resources (likely from Azure given the OpenAI/MSFT partnership) along with tools to monitor, optimize, and build models. Particularly valuable for enterprise users, Foundry would also offer service-level commitments for uptime and support. See Figure 15 for a preliminary pricing schedule for Foundry.

**Figure 15: OpenAI Foundry - Pricing**

Model instance	Units / Instance	3-month commit		1-year commit	
		Monthly cost	Total commit	Monthly cost	Total commit
GPT-3.5 Turbo	100	\$26,000	\$78,000	\$22,000	\$264,000
DV (8K max context)	300	\$78,000	\$234,000	\$66,000	\$792,000
DV (32K max context)	600	\$156,000	\$468,000	\$132,000	\$1,584,000

Source: Company data

OpenAI is currently working on GPT-4 which will be the successor model to GPT-3, with the timeline of the release still unclear (some [media reports](#) noting it would be sometime in 1Q23) with other [reports](#) noting that OpenAI would “sit on it [GPT-4] for much longer,” but OpenAI has made it clear the model will not be released until it can be done “safely and responsibly.” While much remains unknown and unconfirmed around GPT-4, according to media reports GPT-4 will be text only (the initial press was that it would be multi-modal), like GPT-3, and will have a meaningfully larger parameter count, with [media reports](#) suggesting up to 100 trillion parameters. While the parameter count of GPT-4 could increase exponentially, the size of the training data set will likely remain unchanged with Sam Altman (CEO of OpenAI) noting that GPT-4 won’t exceed GPT-3 in size, highlighting that deployments of smaller data sets are: 1) more cost effective, 2) require fewer computing resources, and 3) have simpler implementations. A key area of contrast for GPT-4 will be optimizing the self-assessment process to minimize and eliminate the need for prompting – a key limitation of GPT-3 is the need to analyze the quality of prompts which limits potential use cases whereas the goal of GPT-4 will be to make prompting obsolete. Said differently, different words used in the prompt line affect results, and this is set to change in GPT-4, given a more advanced model.

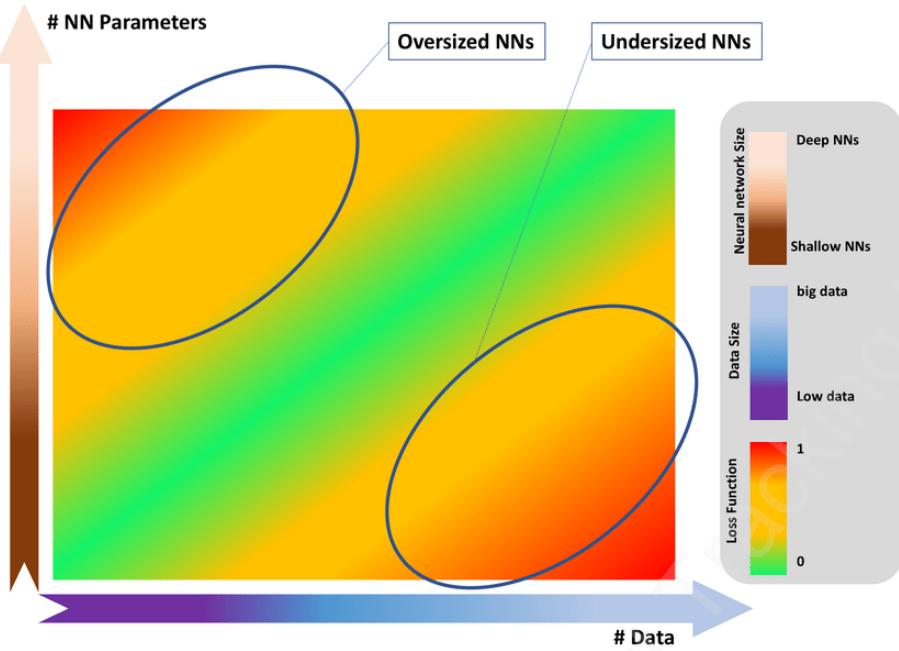
Much of the focus around GPT-4 has been the potential exponential increase in model parameters. However, there are two key variables that influence the power/accuracy (commonly referred to as model loss; a lower model loss means a more accurate model) of an LLM: training data and parameters.

1. **Training data** refers to the underlying data that the LLM is trained on and is measured in tokens. As mentioned earlier, the training data for GPT-3 to GPT-4 is not expected to change based on what Altman has publicly stated up to February 2023.
2. **Parameters** refer to the number of variables the model can change as it learns. A greater number of parameters reflects a higher iterative learning completion. It is this

factor that is assumed to allow GPT-4 to be less sensitive to prompt variations, mentioned earlier.

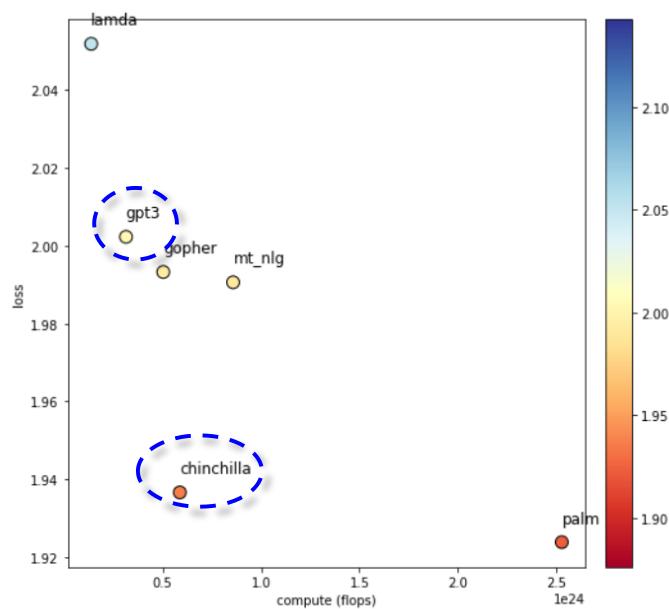
As can be seen in Figure 16, optimized LLMs take a balanced approach to data and parameters and a relatively higher mix of either can significantly increase model loss. Striking the optimal balance between these two variables is the key to a highly capable LLM.

**Figure 16: Parameters Versus Training Data – Identifying Model Loss**



Source: ResearchGate.

Too much or too little training data (relative to parameters) both can increase model loss – the issue with using too much training data is that the model can either become too specialized (if it becomes trained too heavily in a certain area or style) or too broad in nature and the issue with using too little data is that the model will not have enough information for effective pattern recognition. The same issue occurs with parameters (relative to training data) – too many parameters will create issues in identifying the correct patterns (too many choices) while too few parameters will limit the types of patterns the model can identify.

**Figure 17: Model Loss vs Training Compute Resources**

Source: Company data, Credit Suisse estimates, Lesswrong.com

When comparing the model loss of GPT-3 vs other LLMs we note that parameters are not necessarily the primary driver of model accuracy. For example, GPT-3's model loss of 2.00 (at 175B parameters/300B training tokens) is materially higher than that of ChinchillaAI (created by Alphabet-owned DeepMind) which has only 70B parameters but was trained with 1.4T training tokens.

**Figure 18: Model Loss – Parameter vs. Training Data Comparison**

$$\text{Loss of 1T parameter model} = L(1\text{T}, 300\text{B}) = 1.69 + \underbrace{0.03}_{\text{model parameter loss}} + \underbrace{0.25}_{\text{training token loss}} = 1.97$$

$$\text{Loss of Chinchilla} = L(70\text{B}, 1.4\text{T}) = 1.69 + \underbrace{0.08}_{\text{model parameter loss}} + \underbrace{0.16}_{\text{training token loss}} = 1.94$$

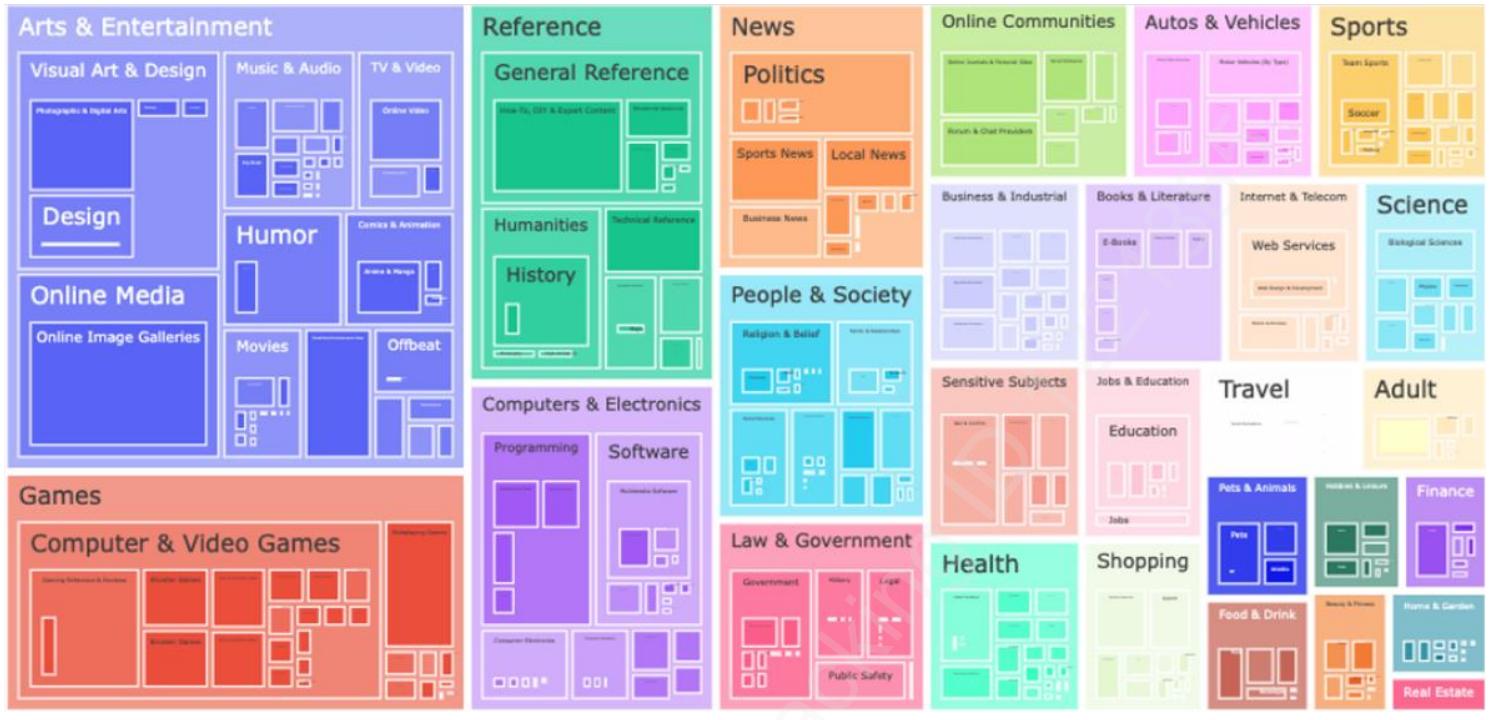
Source: Substack.

Even when looking at a potential GPT-3 version with 1 trillion parameters and the same amount of training data (300B training tokens), this model would still underperform the ChinchillaAI model despite having nearly 15x the number of parameters of Chinchilla. Using this logic, a more efficient use of compute resources which could drive lower incremental model loss would be to train the eventual GPT-4 on more data sets rather than simply scaling up parameter volume. The key takeaway here being that more parameters are not necessarily better and an optimized LLM should take a balanced approach to parameters and training data.

### The LLM and Its Dataset's Topic Modeling

Below we illustrate a sample of Google's PaLM model Topic Modeling based on what data sets it was using for training and its determination of what types of data topics it knows. This in turn drives the relevance of LLM inference prompts and results.

**Figure 19: Final LLMs Tend to Be Composed of the Following Topic Modeling Through Training to Be Relevant to Users/Outputs**



Source: PaLM White Paper.

## ChatGPT's Resource Usage Significant

Importantly, with new AI technologies come new network and resource requirements. Taking a single model, OpenAI's GPT-3, the model required 936 MegaWatts of data center resources and electrical power to train (data center resources are oftentimes measured in denominations of electricity to normalize across various hardware form factors). To put that figure into some perspective, here are two ways to frame the magnitude of the resource usage required just for training the model alone:

- If Cloudflare's (ticker: NET) entire network were to train GPT-3, it would take Cloudflare's network seven full days of continuous resource utilization and full electrical network power. We calculated this figure based on their disclosure of using 49.5 Gigawatts in 2021, narrowed down to network footprint Megawatts per hour, and then solved for the number of days. This would also suggest that Cloudflare's network is likely not going to be used for LLM training since that would effectively switch off services to other customers while their network was computing the LLM training.
- Another way to contextualize how many resources were used to train GPT-3: we can calculate the electrical power load cost of 936 Megawatts based out of a lower power tier 1 data center market, like Dallas, Texas, where significant volumes of hyperscaler compute resources currently reside. Based on Dallas power rates of \$0.05 per Kilowatt, GPT-3 would have cost ~\$47,000 to train from grid utility electricity alone (before accounting for the 10,000 Nvidia V100s). Additionally, for every training model iteration, this is essentially another \$47,000 of spend just on electricity, assuming the whole model is being retrained across all parameters, etc.

In Figure 20, we outline publicly available data across LLM model creators and metrics worth highlighting to illustrate the scale of such LLMs. In this section, we investigate even further the economics of LLMs and split out the differences between search and inference.

**Figure 20: LLM Models by Various Creators with Key Publicly Available Statistics – GPT-3 Took 936MW to Train**

Model Creator	LLM Name	Year Trained	Number of Parameter	Tokens Used	Total Train Compute (FLOPs)	Compute Resources Used to Train Model	Number of GPUs	Model FLOPS Utilization
META	LLaMA - 65B	2023	65.2B	1,400B		449MWh	2,048 A100	
NVIDIA & Microsoft	MT NLG	2022	530B	270B			2,240 A100	30.2%
DeepMind (Google)	Chinchilla	2022	70B	1400				
Google	PaLM	2022	540B	780B	2.56E+24	3181MWh	6,144 TPUv4	46.2%
Google	LaMDA	2022	137B	168B				
DeepMind (Google)	Gopher	2021	280B	300B			4,096 TPUv3	32.5%
OpenAI	GPT-3	2020	175B	300B	3.14E+23	936MWh	10,000 V100	21.3%

Source: Company data.

## How Much Would LLM-powered Search Cost?

Alphabet's Chairman John-Hennessy told Reuters on Feb 22, 2023 that an AI exchange with an LLM was likely to cost 10x more than a standard keyword search on Google Search. This is before optimizations and enhancements, but a very significant multiplier for a scaled technology company like Google. To understand the economics associated with an LLM-powered search engine (like Bing or Google Search), we want to first understand the current profitability of search, followed by estimating the cost of training an LLM model (set-up costs), and lastly estimating the cost of each inference (recurring search costs) to arrive at an all-in LLM search engine cost.

- **Current Search Engine Cost:** Based on figures released by [SemiAnalysis](#) (Figure 23), it is estimated that Google generates about 320k queries per day, which results in ~1.33 cents per query using net revenue as we strip out traffic acquisition costs. Our calculations of aggregate COGS excluding YouTube content acquisition, bandwidth, as well as other costs should be maximum ~\$36.4 billion, which using the same query number nets out to about ~0.036 cents.

- Estimated Model Training Cost:** Shown in Figure 21, the training costs of an LLM vary greatly depending on a number of factors, mainly size of the model, quantity of data used to train the model, hardware costs, FLOPS utilization (Floating Point Operations Per Second, effectively a measure of performance), and energy efficiency of the hardware. Using today's available GCP TPUv4 chip, the estimated training cost of ChatGPT-3 would be approximately \$1.4M, which represents an ~81% reduction in costs from ChatGPT-3's initial estimated training costs. The cost savings are attributable to an improvement in the cost/FLOP, and better FLOP utilization (ChatGPT-3 training in 2020 recorded a 21.3% utilization vs. the PaLM LLM model training in 2022 which recorded utilization of 46.2%, see Figure 22). **This suggests the model training costs are negligible over the life of the search engine.**

**Figure 21: Estimating Training Cost of LLMs on GCP TPUv4 Chips**

	GPT-3 (OpenAI)	Gopher (Google DeepMind)	MT-NLG (Microsoft/Nvidia)	PaLM (Google Research)
Model Parameters	175B	280B	530B	540B
FLOPS/Token/Model Parameter		6		
TPUs/Machine		4		
Peak FLOPS/TPU		275		
FLOPS Utilization		46.20%		
Cost/Machine/Hour (1-year reserved)		\$8.12		
Seconds/Hour		3600		
<b>Training Cost/1000 Tokens</b>	<b>\$0.0047</b>	<b>\$0.0075</b>	<b>\$0.0141</b>	<b>\$0.0144</b>
Train Tokens	300B	300B	270B	780B
<b>Training Costs</b>	<b>\$1,398,072</b>	<b>\$2,236,915</b>	<b>\$3,810,744</b>	<b>\$11,216,529</b>

Source: Substack-Sunyan.

- Cost of Inference vs Cost of Training:** Both the of cost of training and the cost of inference (i.e., cost of producing a result) or more simply the building and running cost of producing a comparable LLM today have gotten over 80% cheaper since the release of GPT-3 in 2020 (*Substack-Sunyan*). In the case of inference cost (cost to run) this reduction in cost (~83%) is due to a combination of the use of 60%+ fewer parameters (searching for fewer types of patterns) with comparable performance and a 58% improvement in hardware operating cost (Cost/FLOP)—the net result of this is searching ~40% of the original parameter count at a ~60% improvement in hardware operating cost.

**Figure 22: Estimated Reductions in Cost of Inference and Cost of Training for LLMs**

	Cost of Inference	Cost of Training
Parameter Count ("N")	> 60% Fewer Parameters (Chinchilla's 70B parameters vs. GPT-3's 175B parameters with performance parity)	
Cost/FLOP		58% Cost/FLOP Reduction (Hardware cost and energy efficiency of H100 vs. V100, which was used to train GPT-3)
Model FLOPS Utilization		2.2x FLOPS Utilization (GPT-3's 21.3% training utilization vs. PaLM's 46.2%)
<b>Net Reduction vs. GPT-3 in 2020</b>	<b>83%</b>	<b>81%</b>

Source: Substack-Sunyan.

- Estimated Inference Costs:** The additional incremental cost per query using ChatGPT is ~\$0.36 per query or ~27% of revenue without any optimizations. The inference process can then be optimized via reducing the size of resource allocation per query (i.e., limiting very long questions and responses), increasing reliance on data that is already available and does not require use of an LLM (cached data), and assuming improved computational power. We detail below an optimization scenario assuming: 1) a limit to resources per search, 2) 20% of searches can be addressed via cached data without the use of an LLM, and 3) increased computational power (TPUv4) which would reduce the estimated inference cost to \$0.03 or ~2% of revenue (see Figure 23).

**Figure 23: SemiAnalysis Estimates Current Incremental Costs per Query at \$0.36**

Metric	Google Search Cost Waterfall				
	2022 Google Search (CS Est.)	ChatGPT Additional Cost	350 Tokens Per Search	20% Navigational or Cached	With TPUs4
Revenue/Query	\$0.0133	\$0.0133	\$0.0133	\$0.0133	\$0.0133
Cost/Query	\$0.0036	\$0.0142	\$0.0112	\$0.0111	\$0.0109
Incremental Cost/Query		<b>\$0.0036</b>	\$0.0006	\$0.0005	\$0.0003
Income/Query	\$0.0097	-\$0.0009	\$0.0020	\$0.0022	\$0.0024
Query/Second	320,000	320,000	320,000	320,000	320,000
Annual Net Revenue (excl TAC)	\$133.8B	\$162.5B	\$162.5B	\$162.5B	\$162.5B
Annual Costs	\$36.3B	\$142.9B	\$113.3B	\$112.0B	\$110.0B
Incremental Costs	\$0.0B	\$35.9B	\$6.3B	\$5.0B	\$3.0B
Operating Income	\$97.5B	\$19.6B	\$49.2B	\$50.5B	\$52.5B

Source: Credit Suisse Estimates, [SemiAnalysis](#).

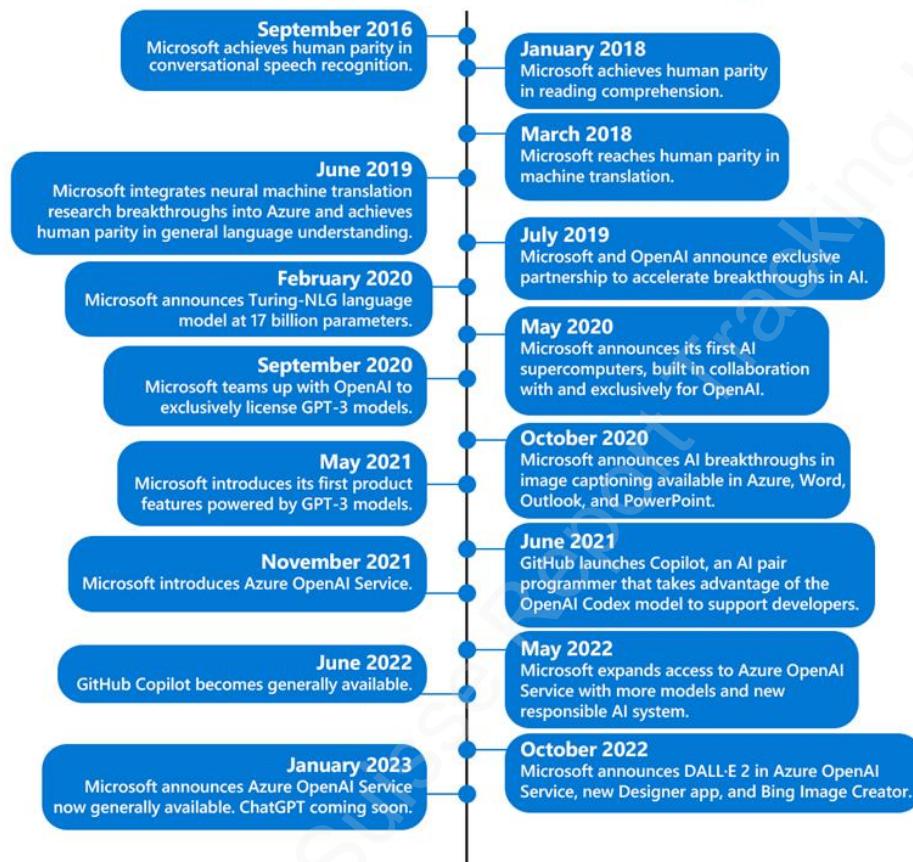
- **Results:** Based on the above, the cost of an LLM powered search engine can range from ~\$0.72 per query (or 54% of net revenue) to \$0.39 per query (or 29% of net revenue) depending on the assumptions.

## Microsoft and OpenAI Background

Microsoft has made three separate investments in OpenAI, the founding company behind ChatGPT: its original investment in 2019, a second investment in 2021, and the third and most significant investment in January 2023. Each of Microsoft's investments accompanied extensions of the existing partnership with OpenAI and was the likely driver behind several changes to OpenAI's corporate structure. Microsoft's cumulative investment in OpenAI to date totals ~\$13B, although it's unclear what portion of this was in the form of cash vs Azure compute credits (the first investment of \$1B was a roughly equal mix of both) and the exact timing of the investments (lump sum vs multi-year, etc.) made by Microsoft.

**Figure 24: Timeline of Key Microsoft AI Developments**

### Timeline of key Microsoft AI breakthroughs



Source: Company data.

- **Microsoft's first investment in OpenAI (2019):** Microsoft's original investment in OpenAI was for approximately \$1 billion in July 2019 with roughly half of this in the form of Azure compute credits (*TechCrunch*). A few months prior to this announcement OpenAI shifted their corporate structure to a "capped profit" entity whereby profits in excess of 100x invested capital would be given to a non-profit entity governed by OpenAI. In addition to the investment, Microsoft and OpenAI created a partnership which would: 1) license GPT-3 for MSFT's own products & services, 2) form an exclusive AI partnership to build Azure AI supercomputing technology, and 3) make Azure Cloud the exclusive source of computing power for OpenAI.
- **Microsoft's second investment in OpenAI (2021):** Microsoft invested an additional ~\$2B in OpenAI in 2021 (no visibility on mix of cash vs Azure compute credits) and shortly thereafter announced the creation of an Azure-hosted, Open AI co-designed supercomputer with 285,000 cores and, according to Microsoft, this same computer was a

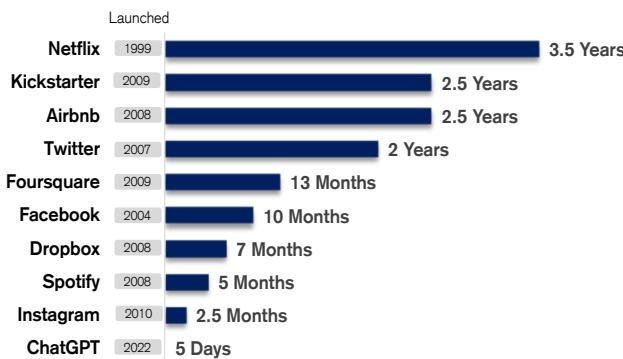
top 5 supercomputer in the world at the time. In November 2021 Microsoft also launched Azure OpenAI services which was intended to provide enterprise-grade access (security, compliance, governance features) to OpenAI's systems including GPT-3.

- **Microsoft's third investment in OpenAI (2023):** Microsoft made an additional investment of (according to media reports) ~\$10B into OpenAI in January 2023, with the investment to be made over multiple years (and no additional visibility on the mix of cash vs compute credits), along with a further expansion of the partnership.
  - In conjunction with this investment Microsoft highlighted their intention to increase their investment in specialized supercomputing systems and Azure AI infrastructure with the company reiterating that Azure remains OpenAI's exclusive cloud provider, powering all OpenAI workloads (research, products, and API services). With respect to the availability of ChatGPT—in Mid-January, Microsoft CEO Satya Nadella noted that ChatGPT is “coming soon” to Azure OpenAI. According to media reports Microsoft's investment would give OpenAI a valuation of \$29B with Microsoft owning a ~49% stake, with additional investors (the largest of which is Khosla Ventures) owning ~49% and OpenAI retaining the remaining ~2%. Under the agreement Microsoft would retain ~75% of OpenAI's profits until it recovered its ~\$13B investment before reverting back to ~49% in-line with Microsoft's ownership stake.
  - The most important and impactful expansion of the partnership in our view was Microsoft's intention to *“deploy OpenAI's models across our consumer and enterprise products and introduce new categories of digital experiences built on OpenAI's technology.”* In other words, Microsoft intends to layer OpenAI's technology across its entire technology stack at both the SaaS and PaaS layers in both the commercial and consumer markets. The ability for MSFT to layer OpenAI's technology in the Microsoft/Office 365 offering, which remains MSFT's single biggest driver of both revenue and profitability, in order to drive both 1) conversion of the legacy installed base to subscription and 2) to drive conversion to premium SKUs (individual premium SKUs like Teams Premium, to premium O365 subscriptions such as E5, and potentially an E7) is the biggest potential upside driver in our view. We expand extensively on this final point and opportunity for Microsoft in a later section of this report.

## ChatGPT's User Growth Has Been Staggering

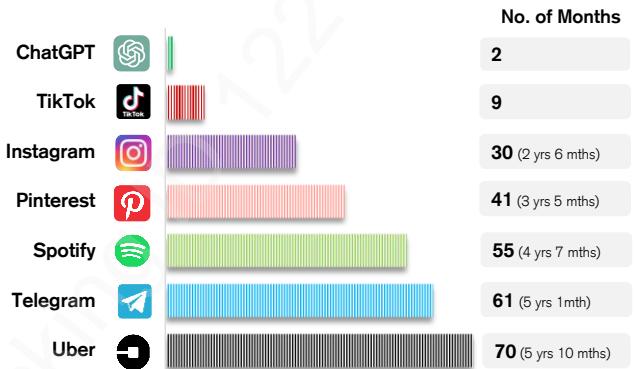
ChatGPT was released on November 30, 2022 and reached 1 million users only 5 days after launch, 57 million users in December and 100 million users in January 2023, making it the fastest growing platform in the world. The pace of user growth has far outpaced previous platforms – for reference, 2 months to reach 100 million users is less than one-third the time it took TikTok to reach 100 million users. TikTok previously held the title of fastest growing platform in the world at 9 months to reach 100 million users.

**Figure 25: ChatGPT – Time to Reach 1M Users Comparison**



Source: Company data, Credit Suisse estimates, Statista

**Figure 26: ChatGPT – Time to Reach 100M Users Comparison**



Source: Company data, Credit Suisse estimates, CNBC

According to *Similarweb* data, web traffic to OpenAI increased nearly ~40x in the first full month following the release of ChatGPT (December 2022) with ~667 million visits in December 2022 up from ~18 million visits in October 2022. OpenAI has quickly surged to the #51 most trafficked website globally in January 2023 (for reference OpenAI now sits in between amazon.com and zoom.com in terms of global popularity), up from #1,829 in November 2022.

From a geographic standpoint the top 5 countries driving traffic to OpenAI are (as of the most recent publicly available *Similarweb* data): 1) United States (15.6% of traffic), 2) India (7.08%), 3) France (4.33%), 4) Germany (3.61%), and 5) UK (3.50%). We would also note that chat.openai.com (where ChatGPT is accessed) makes up 92% of total site visits to OpenAI.

**Figure 27: Similarweb – OpenAI Traffic Data**



Source: Similarweb

**Figure 28: Similarweb – OpenAI Global Ranking by Month**



Source: Similarweb

## ChatGPT CS Revenue Forecast & Model

For our Credit Suisse ChatGPT revenue forecast – treating ChatGPT as if it's an independent company – we make detailed assumptions and forecasts around monthly active users (MAUs), a geographic mix (to determine monetizable MAUs), free to paid conversion, and ASP to derive a revenue forecast. OpenAI has publicly noted a \$200M/\$1B revenue target for 2023/2024, implying a ~5x increase in y/y revenue in 2024, but we note this was in December 2022 when the company had a materially lower base of MAUs. We model revenue of \$205M/\$1.06B in 2023/2024 reaching \$4.3B by 2027, but we note material upside to this number should the company find additional ways to monetize the product (increase free to paid conversion or drive materially higher ASPs). Utilizing the original proposed pricing of ChatGPT Plus of \$42/month under our same set of assumptions would yield revenue of \$9.1B in 2027 (we detail sensitivities later in this section). Finally, this OpenAI/ChatGPT revenue forecast is completely independent of Microsoft GPT-3/3.5 or GPT-4 integrations into O365, it is likely both entities – OpenAI and Microsoft – will monetize this tech simultaneously in different products and services.

**Figure 29: CS – Illustrative ChatGPT Revenue Forecast Model**

(in Millions)	Dec-22	2023	2024	2025	2026	2027
<b>ChatGPT Revenue</b>		\$205	\$1,055	\$2,469	\$3,337	\$4,339
y/y growth (%)			515%	234%	135%	130%
<b>ChatGPT Revenue - Guidance</b>		\$200	\$1,000			
ASP (\$20/month)		\$240.0	\$249.6	\$259.6	\$270.0	\$280.8
y/y growth (%)			4.0%	4.0%	4.0%	4.0%
<b>Paid Users</b>	0.0	1.0	4.2	9.5	12.4	15.5
y/y growth (%)			413%	225%	130%	125%
Free to Paid Conversion		2.0%	2.0%	2.0%	2.0%	2.0%
<b>Total Monetizable Avg. MAUs</b>	0	51	211	475	618	773
y/y growth (%)			413%	225%	130%	125%
United States	0	43	70	95	124	155
% Monetizable	0%	100%	100%	100%	100%	100%
International	0	9	141	380	495	618
% Monetizable	0%	5%	50%	100%	100%	100%
<b>Average MAUs (M)</b>	57	213	352	475	618	773
y/y growth (%)			65.0%	35.0%	30.0%	25.0%
United States		43	70	95	124	155
International		171	282	380	495	618
<b>MAU Geographic Mix</b>						
United States	20%	20%	20%	20%	20%	20%
International	80%	80%	80%	80%	80%	80%

Source: Company data, Credit Suisse estimates.

- **Revenue:** ChatGPT generated less than \$10M of revenue in 2022 and the company expects to achieve \$200M of revenue in 2023 and \$1B of revenue in 2024. For modeling purposes we assume all revenue is generated from the rollout of the \$20/month ChatGPT Plus subscription which we assume is fully available in the United States in March 2023 and a pilot international rollout later in the year.
- **ASP:** For modeling purposes we assume all future revenue will be generated from the rollout of ChatGPT Plus which is being priced at \$20/month. We assume y/y ASP growth at the midpoint of typical software price escalation of ~3-5%. We note the potential for ChatGPT to create additional subscription tiers/add-on products which could be a significant ASP driver going forward (and potentially drive higher free to paid conversion).

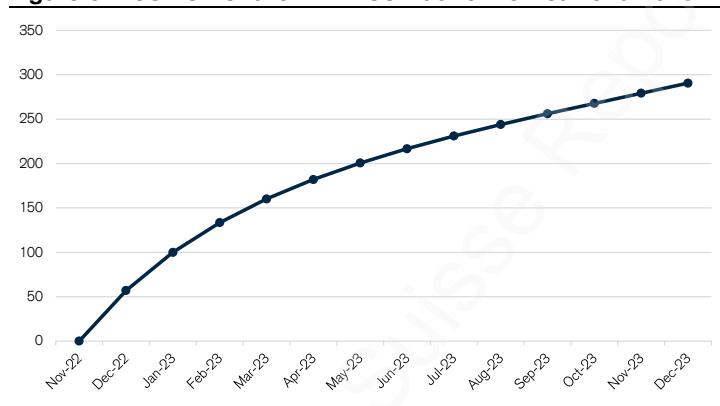
- Total Users:** Our 2023 MAU assumption begins with the cited 100M MAUs in January and assumes moderating deceleration in monthly net new MAUs (to zero deceleration in the growth rate in monthly net new MAUs by year-end) for the remainder of year.

**Figure 30: ChatGPT MAUs – CS Net New Monthly MAU Forecast**

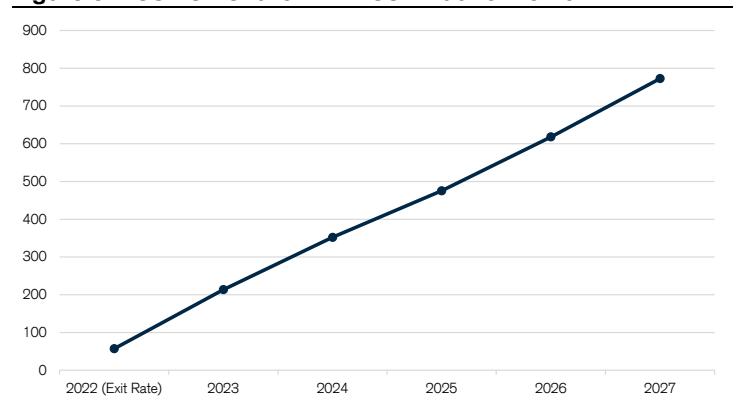
Month	MAUs	Net New Monthly MAUs	Net New - % M/M
Nov-22	0	0	
Dec-22	57	57	
Jan-23	100	43	-24.6%
Feb-23	133	33	-22.3%
Mar-23	160	27	-20.1%
Apr-23	182	22	-17.8%
May-23	201	19	-15.6%
Jun-23	217	16	-13.3%
Jul-23	231	14	-11.1%
Aug-23	244	13	-8.8%
Sep-23	256	12	-6.6%
Oct-23	268	12	-4.3%
Nov-23	279	11	-2.1%
Dec-23	291	11	0.0%
<b>2023 Average</b>		<b>213</b>	<b>19</b>

Source: Company data, Credit Suisse estimates.

We assume total MAUs for ChatGPT reach the approximate size of the MSFT Office installed base in 5 years assuming 5% growth in the MSFT Office installed base through 2027 driven by the release of GPT-4 sometime in the next 6-18 months (per CS estimate). The total MSFT office installed base, consumer + commercial and both license + subscription, currently stands at approximately 600M today according to our estimates, increasing to ~765M in 2027 assuming 5% annual growth.

**Figure 31: CS Est. ChatGPT MAUs-Launch to Year-end 2023**

Source: Company data, Credit Suisse estimates.

**Figure 32: CS Est. ChatGPT MAUs – Launch to 2027**

Source: Company data, Credit Suisse estimates.

- Paid Users:** Typical software companies have a free to premium conversion rate of 2-5% (there are notable outliers above this range), for modelling purposes we assume a constant “freemium” conversion at the low-end of the range (2.0%) as we assume the virality & novelty of the product drives higher overall user count but lower conversion (an outsized portion of people simply “testing” the product due to its novelty). Additionally, we note that the starting subscription price of \$20/month remains high relative to other freemium products and the incremental functionality from the paid vs free version is on the lower end. We would note that if ChatGPT adds lower priced subscription tiers this conversion rate could move considerably higher over time.
- Total Monetizable MAUs:** By our estimates, due to the geographic mix of MAUs, only ~20% of ChatGPT’s userbase is currently monetizable as ChatGPT Plus will only initially be available in the United States with the company noting a pilot international rollout in the

coming months with an “eventual” broad international rollout. For modeling purposes, we assume 5% of the international userbase is monetizable in 2023 to reflect a pilot launch later in the year. We then assume half the international userbase is monetizable in 2024 and the remainder becomes monetizable in 2025 (i.e., ChatGPT Plus is globally available by 2025).

- **Geographic Mix:** According to SimilarWeb data for monthly traffic volume in January 2023, ~20% of ChatGPT’s MAUs came from the United States. For modeling purposes, we assume this mix stays constant—this is important given the initial rollout of ChatGPT plus will only be available in the United States. Additionally, there are several notable regions (e.g., China and Russia) where ChatGPT is unavailable for security reasons.

### **Credit Suisse ChatGPT Model Sensitivities**

- In the table below we lay out sensitivities to our 2027 ChatGPT revenue forecast utilizing various free to paid conversion ratios and growth rates. Notably, utilizing our ChatGPT revenue build we estimate that ChatGPT could exceed \$10B of revenue by 2027 with a free to paid conversion of 3.5% and an average annual MAU growth rate of ~50% or a free to paid conversion of 2.0% and an average annual MAU growth rate of ~70% over the same period. These compare to our base case free to paid conversion ratio/growth rate of 2.0%/38%.

**Figure 33: CS ChatGPT 2027 Revenue Forecast Sensitivities – Conversion/Growth**

Avg. Annual MAU Growth (2023+)	Free to Paid Conversion (2023+)					
	1.00%	1.50%	2.00%	2.50%	3.00%	3.50%
18%	\$1,160	\$1,740	\$2,320	\$2,900	\$3,480	\$4,059
28%	\$1,606	\$2,409	\$3,212	\$4,015	\$4,817	\$5,620
38%	\$2,170	\$3,254	\$4,339	\$5,424	\$6,509	\$7,593
48%	\$2,870	\$4,305	\$5,740	\$7,175	\$8,610	\$10,045
58%	\$3,728	\$5,592	\$7,456	\$9,320	\$11,184	\$13,048
68%	\$4,765	\$7,148	\$9,530	\$11,913	\$14,295	\$16,678

Source: Company data, Credit Suisse estimates

- In the table below we lay out sensitivities to our 2027 ChatGPT revenue forecast utilizing various monthly price points and growth rates. Notably, utilizing our ChatGPT revenue build we estimate that ChatGPT could exceed \$10B of revenue by 2027 with a monthly price of ~\$40 and an average annual MAU growth rate of ~45% or a monthly price of \$30 and an average annual MAU growth rate of ~55% over the same period. These compare to our base case monthly price/growth rate of \$20/38%.

**Figure 34: CS ChatGPT 2027 Revenue Forecast Sensitivities – Monthly Price/Growth**

Avg. Annual MAU Growth (2023+)	ChatGPT Plus - Price Per Month (\$/month)					
	\$10.00	\$15.00	\$20.00	\$30.00	\$40.00	\$50.00
18%	\$1,160	\$1,740	\$2,320	\$3,480	\$4,639	\$5,799
28%	\$1,606	\$2,409	\$3,212	\$4,817	\$6,423	\$8,029
38%	\$2,170	\$3,254	\$4,339	\$6,509	\$8,678	\$10,848
48%	\$2,870	\$4,305	\$5,740	\$8,610	\$11,480	\$14,350
58%	\$3,728	\$5,592	\$7,456	\$11,184	\$14,912	\$18,639
68%	\$4,765	\$7,148	\$9,530	\$14,295	\$19,060	\$23,825

Source: Company data, Credit Suisse estimates

# Microsoft a Very Direct Beneficiary of ChatGPT but The Gain Is In Productivity

Microsoft is a very direct beneficiary of the success of ChatGPT and GPT-3 not only due to its ownership position of ChatGPT and the associated exclusivity agreement around providing Azure infrastructure services to ChatGPT, but also far more important in our view to MSFT's ability to leverage ChatGPT and OpenAI's GPT models (GPT-3/3.5/4) within its existing suite to 1) drive users to standalone premium SKUs (i.e., Teams Premium), 2) convert users from license to subscription which produces a much higher customer LTV, and 3) convert existing subscribers to higher-priced/premium product bundles (i.e., E3 & E5). MSFT has a clear first mover advantage in this respect given their initial investment in OpenAI in 2019 which has allowed MSFT to already come to market and monetize GPT-3/3.5 via multiple products that are already generating revenue for MSFT today. Notably, while MSFT's historical strategy has been to add functionality within premium product bundles and then upsell those premium bundles (i.e., E5), all of the product rollouts driven by GPT-3/3.5 thus far have, for the most part, been in the form of explicit SKUs that embed GPT-3. Below we detail the direct impacts of ChatGPT/GPT-3/3.5 to MSFT that are currently driving revenue/profitability for MSFT today and the longer-term strategy around leveraging ChatGPT/GPT-3/3.5/4 to drive growth and profitability.

## Direct Impacts of ChatGPT/GPT-3

We have identified five key areas where MSFT is currently generating revenue from ChatGPT/GPT-3/3.5: 1) the direct revenue/profit from ChatGPT, 2) GitHub Copilot, 3) Teams Premium, 4) Viva Sales, and 5) Azure revenue (from both the direct revenue from ChatGPT as an exclusive Azure customer and Azure OpenAI). When looking solely at the individual SKUs that MSFT is monetizing directly (see Figure 35), **we see \$5.3B of revenue and \$2.4B of operating income potential to MSFT assuming a 10% penetration** of these products into their respective existing installed base. Utilizing this \$2.4B operating income figure (which conservatively assumes no contribution from MSFT's ChatGPT ownership and no associated Azure revenue), and assuming a 19% tax rate, points to a less than a 7-year cash payback on MSFT's \$13B investment in OpenAI; however, we note this payback could be materially accelerated by additional product rollouts or better (than 10%) installed base penetration. We also note second derivative impacts that we are not accounting for; for example, Teams Premium can only be purchased with an existing subscription to a paid version of Teams (i.e., Microsoft/Office 365), driving an indirect benefit across the entire product suite.

**Figure 35: MSFT GPT-3 Product Revenue/Operating Income—Uplift Potential**

MSFT OpenAI/GPT-3 Embedded Products	\$/Month	ASP	Installed Base (M)	Segment	Operating Margin
GitHub - Copilot (Blended ASP)	\$14.5	\$174	90	IC	43.5%
Teams - Premium	\$10.0	\$120	280	PBP	46.9%
Dynamics - Viva Sales	\$40.0	\$480	8	PBP	46.9%
<b>Revenue Uplift at 5-25% Existing Installed Base Penetration</b>					
	5%	10%	15%	20%	25%
GitHub - Copilot (Blended ASP)	\$783	\$1,566	\$2,349	\$3,132	\$3,915
Teams - Premium	\$1,680	\$3,360	\$5,040	\$6,720	\$8,400
Dynamics - Viva Sales	\$192	\$384	\$576	\$768	\$960
<b>Total Revenue to MSFT</b>	<b>\$2,655</b>	<b>\$5,310</b>	<b>\$7,965</b>	<b>\$10,620</b>	<b>\$13,275</b>
<b>Operating Income Uplift at 5-25% Existing Installed Base Penetration</b>					
	5%	10%	15%	20%	25%
GitHub - Copilot (Blended ASP)	\$341	\$681	\$1,022	\$1,362	\$1,703
Teams - Premium	\$788	\$1,576	\$2,364	\$3,152	\$3,940
Dynamics - Viva Sales	\$90	\$180	\$270	\$360	\$450
<b>Total Operating Income to MSFT</b>	<b>\$1,219</b>	<b>\$2,437</b>	<b>\$3,656</b>	<b>\$4,874</b>	<b>\$6,093</b>
<b>Total MSFT EPS Uplift (19% ETR)</b>	<b>\$0.13</b>	<b>\$0.26</b>	<b>\$0.40</b>	<b>\$0.53</b>	<b>\$0.66</b>
<i>Uplift vs FY2022 EPS</i>	1.4%	2.9%	4.3%	5.7%	7.2%

Source: Company data, Credit Suisse estimates.

We also note that MSFT utilizes a mix of OpenAI technologies to power their AI-based product releases with GitHub Copilot using OpenAI's Codex (a GPT-3 descendant) and both Teams Premium and Viva Sales utilizing OpenAI's GPT-3.5 technology. Despite being priced at significant premiums to their baseline SKUs (see Figure 36), these GPT-integrated products represent significant productivity gains relative to their incremental cost. For example, the basic paid version of GitHub (Team) costs \$4/month while Copilot costs an additional \$10-19/month or 2.5x-4.75x the baseline cost. However, the productivity gains from Copilot are exponentially higher than the incremental cost despite the significant cost increase— **according to GitHub, for certain programming languages Copilot wrote almost 40% of the code generated by developers, and we note the cost of even the most expensive version of Copilot is significantly less than 1% of the cost of a software developer.**

**Figure 36: Selected MSFT OpenAI Productization/Technology Usage**

MSFT OpenAI/GPT Embedded Product	OpenAI Technology Used	MSFT Segment	MSFT Sub-Segment		
GitHub (Copilot)	OpenAI Codex (GPT-3)	IC	Azure*		
Teams (Teams Premium)	GPT-3.5	PBP	Office 365		
Dynamics (Viva Sales)	GPT-3.5	PBP	Dynamics		
MSFT OpenAI/GPT Embedded Product	Baseline SKU	Baseline SKU price (\$)/month	OpenAI/GPT SKU	OpenAI/GPT SKU price (\$)/month	Uplift % vs Baseline
GitHub (Copilot)	GitHub Paid (Team)	\$4.0	Copilot (Average)	\$14.5	363%
Teams (Teams Premium)	Office 365 (E1)	\$10.0	Teams Premium	\$10.0	100%
Dynamics (Viva Sales)	Dynamics 365 (Sales Pro)	\$65.0	Viva Sales	\$40.0	62%

\*A portion of GitHub revenue is on-premise revenue which flows through Server Products

Source: Company data, Credit Suisse estimates

## Driver #1: Direct ChatGPT Revenue & Profitability

While MSFT has not disclosed the full scope of the details around their OpenAI investment with respect to how they will account for ChatGPT revenue (i.e., if they will consolidate financials and which segment/sub-segment it would fall into), we do note that (according to media reports) MSFT will retain ~75% of ChatGPT's profits until its investment (~\$13B cumulatively) is recouped with this split then reverting back to 49%, in-line with their ownership position. Below we show the revenue/FCF potential from MSFT's ownership position under both a 49% split (in-line with their ownership) and a 75% split (in-line with their effective profit split until payback).

**Figure 37: ChatGPT – MSFT Revenue and FCF Allocation (at 49/75% splits)**

	2023	2024	2025	2026	2027
ChatGPT Revenue (CS est.)	\$205	\$1,055	\$2,469	\$3,337	\$4,339
Attributable to MSFT - 49% Split	\$100	\$517	\$1,210	\$1,635	\$2,126
<b>Attributable to MSFT - 75% Split</b>	<b>\$154</b>	<b>\$791</b>	<b>\$1,851</b>	<b>\$2,503</b>	<b>\$3,254</b>
Assumed FCF Margin	0%	5%	10%	15%	
ChatGPT FCF (CS est.)	\$0	\$123	\$334	\$651	
Attributable to MSFT - 49% Split	\$0	\$60	\$164	\$319	
<b>Attributable to MSFT - 75% Split</b>	<b>\$0</b>	<b>\$93</b>	<b>\$250</b>	<b>\$488</b>	

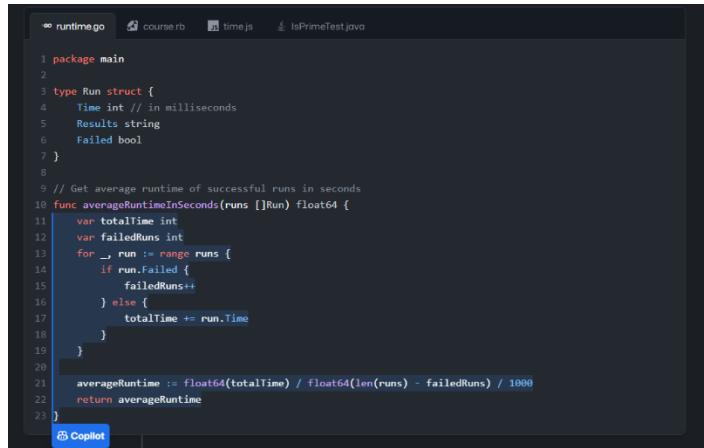
Source: Company data, Credit Suisse estimates

Using our CS ChatGPT revenue forecast and assuming breakeven FCF in 2024 with an incremental 5% increase in FCF margin per year thereafter we estimate MSFT FCF from ChatGPT of \$93/250/488M in 2025/2026/2027 on revenue attributable to MSFT of \$1,851/2,503/3,254M both at a 75% split. While even using the 2027 FCF number points to a payback period of more than 20 years on MSFT's OpenAI investment we note the core driver of profitability from this investment in our view comes from the expansion of the offerings within the productivity suite and not the direct profitability of ChatGPT (see Figure 35).

## Driver #2: GitHub Copilot

GitHub Copilot was the first true productization of OpenAI's technology within the MSFT portfolio in June 2021 and the first real example of OpenAI/GPT-3 monetization within the product portfolio outside of compute-related revenue in Azure. GitHub Copilot uses the OpenAI Codex (essentially just a version of GPT-3) to suggest code and entire functions in real-time.

**Figure 38: GitHub Copilot Demo**



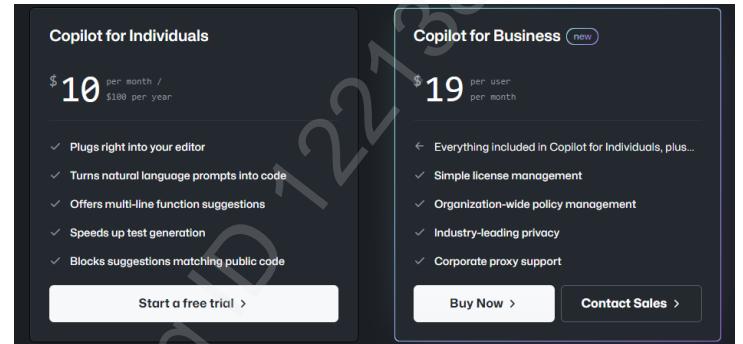
```

1 package main
2
3 type Run struct {
4     Time int // in milliseconds
5     Results string
6     Failed bool
7 }
8
9 // Get average runtime of successful runs in seconds
10 func averageRuntimeInSeconds(runs []Run) float64 {
11     var totalTime int
12     var failedRuns int
13     for _, run := range runs {
14         if run.Failed {
15             failedRuns++
16         } else {
17             totalTime += run.Time
18         }
19     }
20
21     averageRuntime := float64(totalTime) / float64(len(runs) - failedRuns) / 1000
22     return averageRuntime
23 }
```

Source: Company data.

While the underlying OpenAI Codex that powers GitHub copilot is very similar to that which powers ChatGPT, the key advantage of Copilot is the ability to have all the functionality integrated within a collaborative workflow. GitHub offers two primary pricing models for Copilot: an Individual plan at \$10/month (or \$100/year if paid annually) and a business plan at \$19/month.

**Figure 39: GitHub Copilot Pricing**



Pricing Plan	Price	Features
Copilot for Individuals	\$10 per month / \$100 per year	<ul style="list-style-type: none"> <li>Plugs right into your editor</li> <li>Turns natural language prompts into code</li> <li>Offers multi-line function suggestions</li> <li>Speeds up test generation</li> <li>Blocks suggestions matching public code</li> </ul>
Copilot for Business	\$19 per user per month	<ul style="list-style-type: none"> <li>Everything included in Copilot for Individuals, plus...</li> <li>Simple license management</li> <li>Organization-wide policy management</li> <li>Industry-leading privacy</li> <li>Corporate proxy support</li> </ul>

Source: Company data.

**Figure 40: GitHub Copilot - Revenue and Operating Income Uplift Potential**

MSFT OpenAI/GPT-3 Embedded Product	\$/Month	ASP	Installed Base (M)	Segment	Operating Margin
GitHub - Copilot (Blended ASP)	\$14.5	\$174	90	IC	43.5%
<b>Revenue Uplift:</b>					
	5%	10%	15%	20%	25%
GitHub - Copilot (Blended ASP)	\$783	\$1,566	\$2,349	\$3,132	\$3,915
<b>Operating Income Uplift:</b>					
	5%	10%	15%	20%	25%
GitHub - Copilot (Blended ASP)	\$341	\$681	\$1,022	\$1,362	\$1,703
MAUs (Installed Base) in Millions (F1Q23)	90				
Copilot for Individuals \$/month	\$10.0				
Copilot for Business \$/month	\$19.0				
Copilot for Individuals % mix	50%				
Copilot for Business % mix	50%				
Blended Monthly Price	\$14.5				

Source: Company data, Credit Suisse estimates

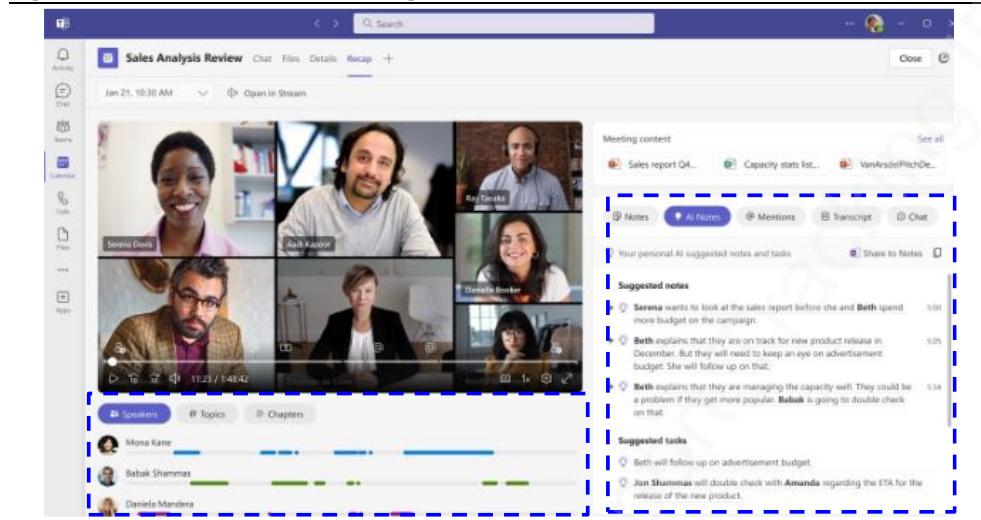
In order to derive a revenue/operating income uplift potential we utilize MSFT's disclosed 90M MAU figure as the installed base; we assume an equal-weighting of Copilot for Individuals (assumed all at \$10/month) and Copilot for Business (at \$19/month) to derive an ASP, and we assume the same operating margin as Intelligent Cloud in FY22 of 43.5%. While the disclosed 90M MAU figure includes both paid and non-paid users, we note this represents opportunity for a second derivative impact from Copilot, as non-paid users that move to Copilot directly will also

likely move to a paid GitHub subscription in order to unlock the full value of the offering (representing additional upside). We estimate that at 10/20% installed base penetration of GitHub Copilot would represent \$1.57B/\$3.13B of revenue upside and \$681/1,362M of operating income upside.

## Driver #3: Teams Premium

In February 2023 MSFT released Teams Premium, powered by OpenAI's GPT-3.5, for \$10 per user per month (\$7 per month promotional pricing through June 2023) which effectively embeds AI across the entire Teams ecosystem. A key functionality of Teams Premium is Intelligent Recap (see Figure 41) which automatically generates meeting notes, recommended tasks, and (personalized) key meeting highlights all utilizing GPT-3.5. Particularly relevant for large global enterprises, Teams Premium also offers live translation (with captions) in over 40 different languages that can translate conversation in real-time.

**Figure 41: Teams Premium – Intelligent Recap**



Source: Nokia

Teams Premium was the first OpenAI-powered product that addressed the core Microsoft 365 bundle. Of the product announcements thus far, we believe Teams Premium offers the most upside to revenue and profitability in the near term given the sheer size of the installed base (280 million MAUs disclosed on the F2Q23 earnings call). Teams Premium is also not available in any Microsoft 365 product bundle and must be purchased as a separate SKU. However, we do note that users must own a base teams license (Office 365, Microsoft 365 or Teams Essentials) in order to purchase Teams Premium, which adds an additional tailwind to revenue/profitability beyond the \$10 per user per month (i.e., a non-Teams user must first buy a base Teams license via a 365 bundle or Teams Essentials before buying Teams Premium).

**Figure 42: Teams Premium - Revenue and Operating Income Uplift Potential**

MSFT OpenAI/GPT-3 Embedded Product	\$/Month	ASP	Installed Base (M)	Segment	Operating Margin
Teams - Premium	\$10.0	\$120	280	PBP	46.9%
<u>Existing Installed Base Penetration</u>					
<b>Revenue Uplift:</b>	<b>5%</b>	<b>10%</b>	<b>15%</b>	<b>20%</b>	<b>25%</b>
Teams - Premium	\$1,680	\$3,360	\$5,040	\$6,720	\$8,400
<u>Existing Installed Base Penetration</u>					
<b>Operating Income Uplift:</b>	<b>5%</b>	<b>10%</b>	<b>15%</b>	<b>20%</b>	<b>25%</b>
Teams - Premium	\$788	\$1,576	\$2,364	\$3,152	\$3,940
MAUs (Installed Base) in Millions (F2Q23)	280				

Source: Company data, Credit Suisse estimates

In order to derive a revenue/operating income uplift potential we utilize MSFT's disclosed 280M MAU figure as the installed base but note substantial upside to this number as Teams MAUs continue to grow (270M/145M MAUs one/two years ago). We assume a monthly price of \$10 per user per month (note: the \$7 monthly promotional price ends on June 30, 2023 and will revert to the standard \$10 monthly pricing on July 1, 2023) and we assume the same operating margin as Productivity and Business Processes in FY22 of 46.9%. We estimate that at 10/20% installed base penetration of Teams Premium would represent \$3.36B/\$6.72B of revenue upside and \$1.58B/\$3.15B of operating income upside.

## Driver #4: Viva Sales

MSFT launched Viva Sales in June 2022 (integrated with GPT-3.5 in February 2023), a CRM companion application, to bring AI capabilities to the CRM ecosystem with Viva Sales addressing not only the existing Dynamics CRM installed based but also Salesforce CRM through a direct integration (+other non-MSFT CRMs). Viva Sales integrates with Teams, Outlook, the Office suite (Word, Excel, PowerPoint) along with CRM applications (both Dynamics CRM, Salesforce CRM and others) and is centered around automatically capturing and integrating customer data across. A key area of focus with the integration of GPT-3.5 was adding additional functionality around email with MSFT utilizing GPT-3.5 to add new email features to automatically generate preformatted email responses with personalized texts, promotions, pricing and deadlines—MSFT highlighted email as a key area of inefficiency with the company citing that the average seller spends ~66% of their time managing emails.

**Figure 43: Viva Sales Pricing**

Sales	First Dynamics 365 app	Subsequent qualifying <sup>1</sup> Dynamics 365 app
<b>Sales Professional</b>	\$65 per user/month	\$20 per user/month
<b>Sales Enterprise<sup>2</sup></b>	\$95 per user/month	\$20 per user/month
<b>Sales Premium<sup>2</sup></b>	\$135 per user/month	
<b>Microsoft Relationship Sales<sup>2</sup></b>	\$162 per user/month	
<b>Microsoft Viva Sales<sup>2</sup></b>	\$40 per user/month	

Source: Company data, Credit Suisse. 1: Subsequent pricing applies only to the individual licensed for the first app. For example, if Person A is licensed for the first app, subsequent pricing wouldn't apply to Person B. Subsequent pricing for tenant-based apps applies to any tenant in your organization. A tenant contains uniquely identified domains, users, security groups, and licenses. Your organization may have multiple tenants, and a single tenant can contain multiple Dynamics 365 (online) environments. 2: Viva Sales is included in Sales Enterprise, Sales Premium and Relationship Sales at no extra cost.

**Figure 44: Viva Sales Feature List**

**Microsoft Viva Sales**

\$40.00 user/month  
(Annual subscription—auto-renews)<sup>3</sup>

[Try now](#)   [Buy now >](#)

A seller experience application that lets sellers use Microsoft 365 and Microsoft Teams to automatically capture data into any CRM system, eliminating manual data entry and giving more time to focus on selling.

- GPT 3.5 (preview)
- Conversational intelligence
- Salesforce CRM integration
- Dynamics 365 Sales integration
- Identify customers in emails
- Teams meeting insights
- Share and collaborate on business data

<sup>3</sup> A Microsoft 365 for enterprise or Office 365 for enterprise product license is required to use the Viva Sales app in Outlook, Excel, and Microsoft Teams. Viva Sales is included with Dynamics 365 Enterprise and Dynamics 365 Premium.

Source: Company data, Credit Suisse. 1: Once your paid subscription begins, cancellation policies vary based on your status as a new customer, product, and domain selections on Microsoft. Cancel your Microsoft 365 subscription any time by going to the Microsoft 365 admin center. When a subscription is canceled, all associated data will be deleted.

Notably, while MSFT doesn't have an industry leading positioning within CRM, positioning Viva Sales as a companion application with other non-MSFT CRM systems is key as it allows MSFT to leverage their dominant installed base in Office to drive traction in the CRM space.

**Figure 45: Viva Sales - Revenue and Operating Income Uplift Potential**

MSFT OpenAI/GPT-3 Embedded Product	\$/Month	ASP	Installed Base (M)	Segment	Operating Margin
Teams - Premium	\$10.0	\$120	280	PBP	46.9%
<b>Existing Installed Base Penetration</b>					
<b>Revenue Uplift:</b>	5%	10%	15%	20%	25%
Teams - Premium	\$1,680	\$3,360	\$5,040	\$6,720	\$8,400
<b>Existing Installed Base Penetration</b>					
<b>Operating Income Uplift:</b>	5%	10%	15%	20%	25%
Teams - Premium	\$788	\$1,576	\$2,364	\$3,152	\$3,940
MAUs (Installed Base) in Millions (F2023)	280				
MSFT OpenAI/GPT-3 Embedded Product	\$/Month	ASP	Installed Base (M)	Segment	Operating Margin
Dynamics - Viva Sales	\$40.0	\$480	8	PBP	46.9%
<b>Existing Installed Base Penetration</b>					
<b>Revenue Uplift:</b>	5%	10%	15%	20%	25%
Dynamics - Viva Sales	\$192	\$384	\$576	\$768	\$960
<b>Existing Installed Base Penetration</b>					
<b>Operating Income Uplift:</b>	5%	10%	15%	20%	25%
Dynamics - Viva Sales	\$90	\$180	\$270	\$360	\$450
<u>Viva Sales Installed Base Build</u>					
Dynamics Revenue (Annualized F2023)	\$5,156				
Salesforce Sub. Revenue (Annualized F3Q23)	\$28,932				
Combined Revenue	\$34,088				
% Spend on Sales (CRM Sales Cloud Mix Proxy)	24%				
Combined Sales Revenue	\$8,181				
Base Dynamics Monthly Price/User	\$65				
Assumed 1 Add-On/User	\$20				
Assumed Average ASP	\$1,020				
Implied Installed Base in Millions	8.0				

Source: Company data, Credit Suisse estimates

To estimate an installed base for Viva Sales we combine our estimate for annualized MSFT Dynamics revenue for the most recent quarter (F2Q23) and Salesforce annualized subscription revenue for the most recent quarter (F3Q23) and use Salesforce's sales cloud mix as a percent of revenue as a proxy for the mix of Dynamics revenue that is derived from sales/CRM applications. We then take the combined sales application revenue and assume dynamics pricing per user including 1 add-on per user on the combined revenue base to derive an installed base (roughly 8M). While Viva Sales is currently included in Dynamics 365 Enterprise and Premium subscriptions (and would therefore not represent revenue upside) for our installed base calculation we assume this userbase receiving Viva Sales as part of their Dynamics 365 plan is offset by the addressable userbase not in the Dynamics/Salesforce CRM ecosystem today (e.g., users on ORCL's CRM, as Viva Sales' integrations are not limited to Salesforce)

In order to derive a revenue/operating income uplift potential we utilize our estimated installed base of 8M users and we assume the same operating margin as PBP in FY22 of 46.9%. We note there is significant upside potential to the installed base number as MSFT currently has limited share in the CRM market. Microsoft (Dynamics)/Salesforce only have a ~23% combined CRM market share (of which MSFT is only 2.9% with CRM at 19.9%) according to Gartner's 2021 estimate, leaving substantial room for share gains for MSFT specifically. We estimate that at 10/20% installed base penetration of Viva Sales would represent \$384M/\$768M of revenue upside and \$180/360M of operating income upside.

## Driver #5: Azure (ChatGPT Compute Rev. & Azure OpenAI)

MSFT generates direct revenue through its Azure business from ChatGPT/OpenAI two primary ways: 1) the productization of the OpenAI and the GPT family of models (soon to include ChatGPT specifically) via Azure OpenAI and 2) via direct Infrastructure as a Service revenue

(compute and storage) from being the exclusive cloud provider of ChatGPT and all of OpenAI's workloads across research, products, and API services.

In January 2023 MSFT announced the general availability of Azure OpenAI which provides access to the suite of OpenAI's models, GPT-3.5 (human language generation), Codex (code generation) and DALL•E 2 (image generation and editing), via Azure infrastructure and enterprise-grade services (compliance, security, etc.). With respect to adding ChatGPT specifically to Azure OpenAI, MSFT noted in mid-January 2023, that it would be added "soon."

**Figure 46: Azure OpenAI – Pricing Table [Central US]**

Pricing details:				
Instance	Series	Models	Inferencing per 1,000 tokens	Hosting per hour
Standard	Base Series	Ada	\$0.0004	N/A
		Babbage	\$0.0005	N/A
		Curie	\$0.002	N/A
		Davinci	\$0.02	N/A
	Base Series Fine-tuned	Ada	\$0.0004	\$0.05
		Babbage	\$0.0005	\$0.08
		Curie	\$0.002	\$0.24
		Davinci	\$0.02	\$34
	Codex Series	Code-Cushman	\$0.024	N/A
		Code-Davinci	\$0.10	N/A
	Codex Series Fine-tuned	Code-Cushman	\$0.024	\$0.54/hour
	Embeddings	Ada	\$0.004	N/A
		Babbage	\$0.005	N/A
		Curie	\$0.02	N/A
		Davinci	\$0.20	N/A

Fine-tuning training			
Instance	Series	Models	Training per compute hour
Standard	Base Series Fine-tuned	Ada	\$20
		Babbage	\$22
		Curie	\$24
		Davinci	\$84
	Codex Series Fine-tuned	Code-Cushman	\$26

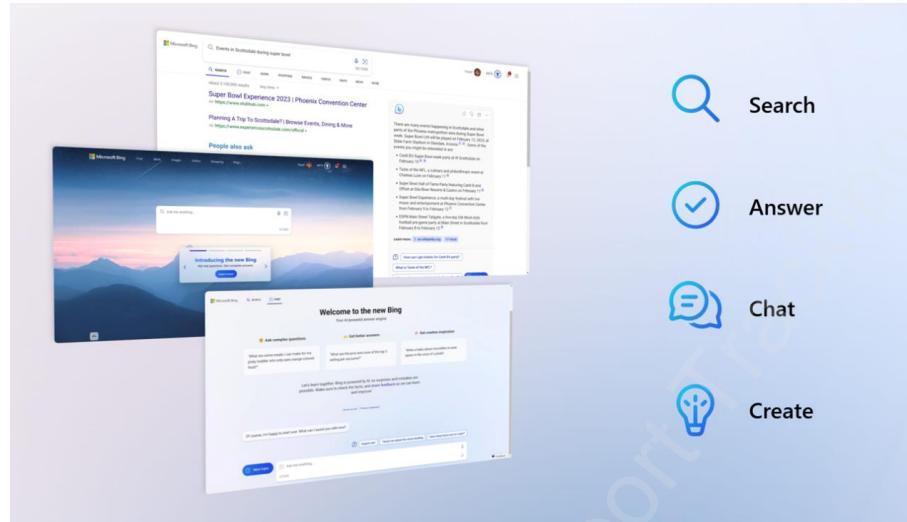
Source: Company data, Credit Suisse

MSFT also generates Azure revenue as the exclusive cloud provider for ChatGPT and all of OpenAI's workloads; however, the nature of the revenue recognition around this is still uncertain. MSFT's initial investment in OpenAI in 2019 for \$1B was reportedly partially in the form of compute credits (*TechCrunch* identified the amount as roughly half of the \$1B) and the translation of OpenAI's Azure usage into revenue—particularly with respect to how the credits would convert to revenue (whether the credits were given at cost, or more likely in our view at full value with a discount). Additionally given MSFT now owns roughly half of OpenAI, the method of how MSFT accounts for OpenAI (not yet disclosed) will also likely play a role in the actual revenue impact of OpenAI's Azure usage to MSFT. Given the uncertainty around the accounting treatment around OpenAI-related IaaS revenue and MSFT's ownership position in OpenAI more broadly—until MSFT provides additional visibility into the accounting dynamics around OpenAI we view Azure OpenAI as the key Azure revenue driver related to OpenAI.

## Upside Optionality: Bing AI

In our short-term assessment of MSFT revenue attribution, despite the launch of Bing AI being predominantly focused on search, our attention quickly pivoted to more direct and accessible revenue streams without going head-to-head with a major incumbent, like Google. We do not factor in any benefits from search and advertising revenues, which MSFT noted was \$18B in (calendar) 2022, and this we believe has been the focus of most investors and the dialogue; however, we point out the aforementioned five more direct monetizable paths based on announced products and availability of OpenAI models for MSFT products. The technology underlying Bing AI is based on a modified version of OpenAI's GPT-3.5 and ChatGPT. Bing AI is currently available in limited preview with a full version expected to be released in the coming months.

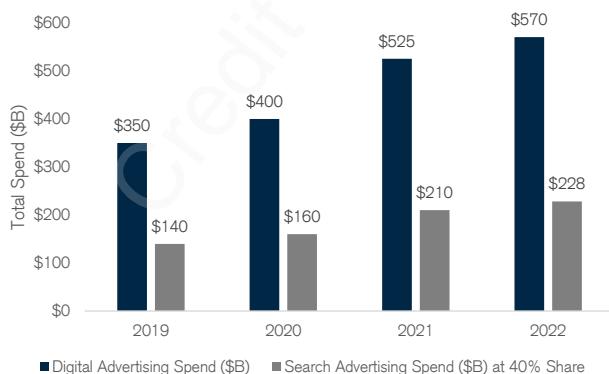
**Figure 47: Bing AI Demo**



Source: Company data, Credit Suisse estimates

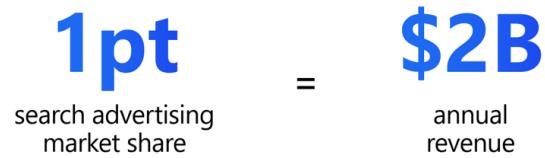
MSFT has sized total Digital Advertising spend at \$570B in 2022, with ~40% of that being search advertising implying total search advertising spend of ~\$230B in 2022 (this is how MSFT is calculating 1 point of market share = \$2B of potential annual revenue). While MSFT has gained share with its search & news advertising business growing at a 24% CAGR over the past 2 years (includes benefit from Xandr acquisition) relative to broader digital advertising spend growing at 19% over the same 2-year period, we note MSFT still remains a small player in the market with \$18B of search and advertising revenue in 2022 suggesting MSFT has ~8% market share today with the company noting that a majority of that comes from Windows PCs.

**Figure 48: Digital/Search Advertising Spend Growth**



Source: Company data, Credit Suisse estimates

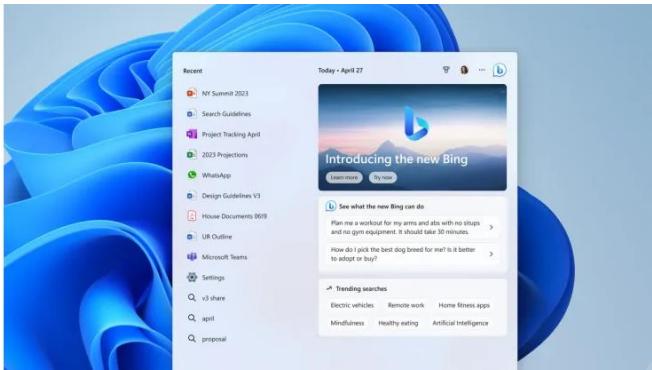
**Figure 49: MSFT - Search Advertising Share/Revenue Uplift**



Source: Company data, Credit Suisse estimates

We believe MSFT's key advantage in driving additional share gains in search advertising will be its ability to leverage its dominant share in Windows (~74% of Desktop PCs in January 2023), with over 1.4 billion monthly active Windows devices, to drive additional usage of Bing AI. MSFT has already begun this process via an update to Windows 11 which added an additional button for Bing AI making it easier to launch chats (see Figure 50). We note that only ~18% of Windows PCs are currently running Windows 11 as of January 2023 (StatCounter) and that support for Windows 10 ends in 2025 which represents a significant opportunity for MSFT to drive additional Bing usage as the remainder of the Windows installed base switches to Windows 11 over the next few years.

**Figure 50: Windows 11 Update – February 2023**



Source: Company data, Credit Suisse

We see potential revenue opportunity for MSFT of \$5.7B assuming MSFT is able to drive a 2.5% incremental share gain in search advertising which represents EPS uplift of \$0.09/\$0.15 at 15%/25% operating margins. Moreover, we show profitability at a lower margin in order to sensitize for the potential impact of higher search-related costs. We also note the potential risk to the overall size of the search advertising market from the impact of less searching (and therefore fewer ads are displayed) per query which may or may not be offset by a higher price paid per ad (i.e., if the ads are therefore more impactful). If this were to happen we believe that given MSFT's relatively small market share today (~8% in 2022) they have the ability to offset declines in the overall level of search and advertising spend with share gains.

**Figure 51: MSFT/Bing AI Sensitivities – 15% Operating Margin**

<b>Revenue Uplift at 2.5%-10% Market Share Gain</b>				
	<b>2.5%</b>	<b>5.0%</b>	<b>7.5%</b>	<b>10.0%</b>
Search Advertising - Bing AI	\$5,700	\$11,400	\$17,100	\$22,800
<b>Total Revenue to MSFT</b>	<b>\$5,700</b>	<b>\$11,400</b>	<b>\$17,100</b>	<b>\$22,800</b>

**Operating Income Uplift at 2.5%-10% Market Share Gain @ 15% Margin**

	<b>2.5%</b>	<b>5.0%</b>	<b>7.5%</b>	<b>10.0%</b>
Search Advertising - Bing AI	\$855	\$1,710	\$2,565	\$3,420
<b>Total Operating Income to MSFT</b>	<b>\$855</b>	<b>\$1,710</b>	<b>\$2,565</b>	<b>\$3,420</b>

**EPS Uplift at 2.5%-10% Market Share Gain @ 19% ETR**

	<b>2.5%</b>	<b>5.0%</b>	<b>7.5%</b>	<b>10.0%</b>
Search Advertising - Bing AI	\$0.09	\$0.19	\$0.28	\$0.37
<b>Total EPS to MSFT</b>	<b>\$0.09</b>	<b>\$0.19</b>	<b>\$0.28</b>	<b>\$0.37</b>

Source: Company data, Credit Suisse estimates

## Upside Optionality: LinkedIn

While MSFT hasn't announced any GPT-Integrations or functionality with LinkedIn yet, we note the company maintains a large userbase of ~900M registered users, of which ~310M are MAUs (*DemandUsage*) that MSFT could ultimately monetize. LinkedIn has multiple avenues to monetize potential future GPT integrations across its product portfolio—LinkedIn generated

**Figure 52: MSFT/Bing AI Sensitivities – 25% Operating Margin**

<b>Revenue Uplift at 2.5%-10% Market Share Gain</b>				
	<b>2.5%</b>	<b>5.0%</b>	<b>7.5%</b>	<b>10.0%</b>
Search Advertising - Bing AI	\$5,700	\$11,400	\$17,100	\$22,800
<b>Total Revenue to MSFT</b>	<b>\$5,700</b>	<b>\$11,400</b>	<b>\$17,100</b>	<b>\$22,800</b>

**Operating Income Uplift at 2.5%-10% Market Share Gain @ 25% Margin**

	<b>2.5%</b>	<b>5.0%</b>	<b>7.5%</b>	<b>10.0%</b>
Search Advertising - Bing AI	\$1,425	\$2,850	\$4,275	\$5,700
<b>Total Operating Income to MSFT</b>	<b>\$1,425</b>	<b>\$2,850</b>	<b>\$4,275</b>	<b>\$5,700</b>

**EPS Uplift at 2.5%-10% Market Share Gain @ 19% ETR**

	<b>2.5%</b>	<b>5.0%</b>	<b>7.5%</b>	<b>10.0%</b>
Search Advertising - Bing AI	\$0.15	\$0.31	\$0.46	\$0.62
<b>Total EPS to MSFT</b>	<b>\$0.15</b>	<b>\$0.31</b>	<b>\$0.46</b>	<b>\$0.62</b>

Source: Company data, Credit Suisse estimates

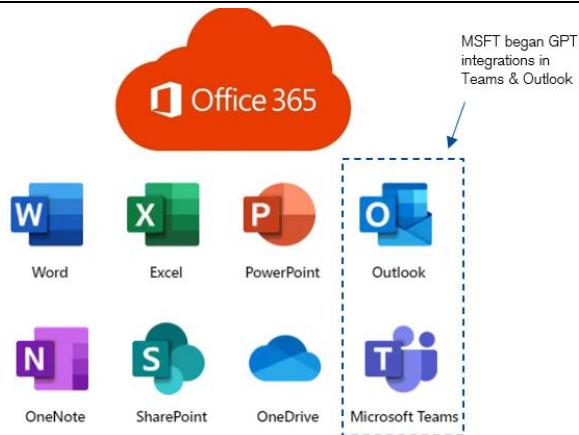
\$15.5B of annualized revenue in the most recent quarter (F2Q23) across Talent Solutions, Marketing Solutions, Premium Subscriptions, and Sales Solutions. We note this represents a significant lever that MSFT could use to offset the particularly acute growth deceleration in this business (due to the weakening macro), with growth decelerating to ~10% y/y in the most recent quarter (F2Q23) from 37% y/y growth a year ago (F2Q22).

Credit Suisse Report Tracking ID 1227385

## Long-Term Impacts of ChatGPT/GPT-3

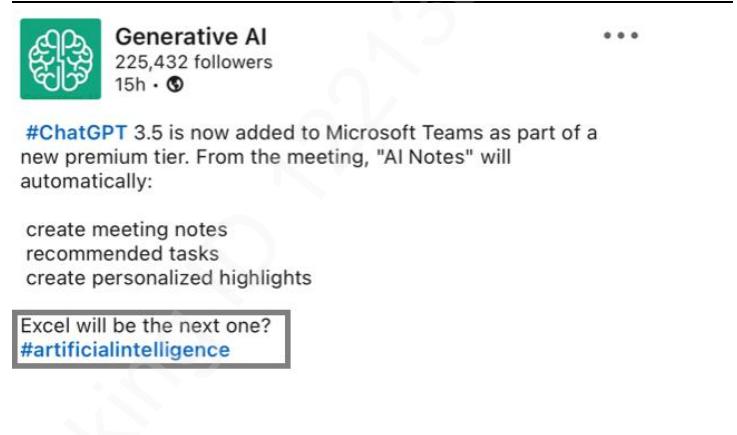
In our view the most important implication of the OpenAI partnership is MSFT's ability to integrate GPT into the MSFT ecosystem and most importantly, bring GPT to the core Office bundle (Word, PowerPoint, Excel and Outlook) driving a step function increase in productivity, giving MSFT an opportunity **to drive a substantial price increase across the entire Office installed base commensurate with the increased productivity and value-add.**

**Figure 53: Core Office 365 Suite**



Source: Company data, Credit Suisse estimates

**Figure 54: Potential Future GPT Office Integrations – Excel**



Source: LinkedIn

In our view, despite Office 365 Commercial being one of the most successful businesses in software (nearly ~\$40B of annuity style revenue at an FCF margin well north of 50%) the \$10/23/38 monthly costs for Office 365 E1/E3/E5 are at price points that represent outsized value relative to its cost by virtually any metric (subscription cost to overall headcount cost, etc.). The ability to drive an increase in both productivity and price (in-line with that productivity) represents a core value proposition for both MSFT and its customers more broadly. In almost all cases, the increase in productivity would far outweigh the subscription cost increase. We also note there are material implications to overall economic productivity by layering this AI functionality into the Office suite given the sheer size of the Office commercial installed base (~370M O365 commercial subscribers in F2023).

## Bringing GPT to MSFT: The True Value Is in Microsoft Office

MSFT is in the early stages of integrating GPT into the Office 365 suite which it has already done with two standalone SKUs—MSFT integrated GPT into Teams with Teams Premium and partially into Outlook with email functionality in Viva Sales (although notably this is a very limited subset of the Outlook userbase). **As MSFT integrates GPT into the core Office ecosystem (Word, Excel, PowerPoint and into Outlook more broadly), we expect this will come with a price increase or additional SKUs.** We examine two ways we believe MSFT could monetize a GPT integration into Office and this step function higher in Office productivity: 1) a mandatory broad "AI" price increase impacting the entire installed base and 2) an ultra-premium E7 SKU which would embed advanced AI functionality that comes at a significant premium for Office "power users." We view the most likely outcome as a combination of the two—a broad price increase to reflect the addition of basic AI functionality into the Office suite along with an ultra-premium E7 SKU which would include advanced AI functionality. **We note potential revenue contribution from the addition of AI/GPT into the Office suite of \$33.6B** with a potential "AI/GPT" price increase driving \$18.7B and the creation of an E7 SKU of \$14.9B. Assuming a 50% incremental margin (likely on the low-end) this represents a potential operating income contribution of \$16.8B or **\$1.82 of EPS (nearly 20% upside vs FY2022 EPS) likely over a period of 5+ years (adding ~3-5%/year to EPS growth).**

**Figure 55: O365 Commercial**

O365 Commercial - GPT Monetization Driver	Installed Base Impacted	Avg. ASP Uplift
"AI/GPT" O365 Price Increase	100%	50%
E7 (Ultra-Premium AI Bundle)	10%	400%
<b>Potential Revenue Uplift in Millions</b>		
"AI/GPT" O365 Price Increase		\$18,653
E7 (Ultra-Premium AI Bundle)		\$14,922
<b>Total Revenue to MSFT</b>		<b>\$33,575</b>
<b>Potential Operating Income Uplift (@ 50% Incremental Margin) in Millions</b>		
"AI/GPT" O365 Price Increase		\$9,326
E7 (Ultra-Premium AI Bundle)		\$7,461
<b>Total Operating Income to MSFT</b>		<b>\$16,787</b>
<b>EPS Uplift (19% ETR)</b>		
"AI/GPT" O365 Price Increase		\$1.01
E7 (Ultra-Premium AI Bundle)		\$0.81
<b>Total MSFT EPS Uplift (19% ETR)</b>		<b>\$1.82</b>
<i>Uplift vs FY2022 EPS</i>		19.8%

Source: Company data, Credit Suisse estimates

1. **"AI/GPT" O365 Price Increase:** As our base case we assume a 50% price increase across every Office bundle to reflect the addition of basic AI/GPT functionality into the suite. To support a 50% price increase we note how MSFT has priced Teams Premium (its first GPT-embedded SKU within the Office 365 ecosystem) which gives a sense of how MSFT is valuing the addition of GPT into its SKUs. Teams premium pricing of \$10/month reflects ~100%/~43% of the price of an E1/E3 bundle (the most typical commercial bundles); thus, we view a 50% increase to reflect the value-add of the integration of GPT across the entire Office 365 ecosystem (which should generate materially more productivity benefits than simply with Teams) as reasonable.

**Figure 56: Revenue Uplift Sensitivity – "AI/GPT" O365 Price Increase**

Installed Base Penetration	Increase to Average ASP								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
10%	\$373	\$746	\$1,119	\$1,492	\$1,865	\$2,238	\$2,611	\$2,984	\$3,357
20%	\$746	\$1,492	\$2,238	\$2,984	\$3,731	\$4,477	\$5,223	\$5,969	\$6,715
30%	\$1,119	\$2,238	\$3,857	\$4,477	\$5,596	\$6,715	\$7,834	\$8,953	\$10,072
40%	\$1,492	\$2,984	\$4,477	\$5,969	\$7,461	\$8,953	\$10,445	\$11,938	\$13,430
50%	\$1,865	\$3,731	\$5,596	\$7,461	\$9,326	\$11,192	\$13,057	\$14,922	\$16,787
60%	\$2,238	\$4,477	\$6,715	\$8,953	\$11,192	\$13,430	\$15,668	\$17,906	\$20,145
70%	\$2,611	\$5,223	\$7,834	\$10,445	\$13,057	\$15,668	\$18,279	\$20,891	\$23,502
80%	\$2,984	\$5,969	\$8,953	\$11,938	\$14,922	\$17,906	\$20,891	\$23,875	\$26,860
90%	\$3,357	\$6,715	\$10,072	\$13,430	\$16,787	\$20,145	\$23,502	\$26,860	\$30,217
100%	\$3,731	\$7,461	\$11,192	\$14,922	<b>\$18,653</b>	\$22,383	\$26,114	\$29,844	\$33,575

Source: Company data, Credit Suisse estimates

In Figure 56 we highlight revenue uplift sensitivities assuming various levels of price increases and installed base penetrations (note: a mandatory price increase would impact the entire installed base). While we assume a 50% price increase as the base case for a broad GPT integration, we note the possibility for the price increase to be substantially higher over time given the value-add potential of a GPT integration across the broad Office suite. In Figure 57 we highlight operating income uplift sensitivities under the same ASP/installed base penetration scenarios, assuming a 50% incremental operating margin, the actual incremental margin would likely be substantially higher given there is minimal incremental associated operating

expenses; however, we only model a slightly above segment level PBP margin (46.9% in FY22) to be conservative.

**Figure 57: Operating Income Uplift Sensitivity – “AI/GPT” O365 Price Increase**

Installed Base Penetration	Increase to Average ASP								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
10%	\$187	\$373	\$560	\$746	\$933	\$1,119	\$1,306	\$1,492	\$1,679
20%	\$373	\$746	\$1,119	\$1,492	\$1,865	\$2,238	\$2,611	\$2,984	\$3,357
30%	\$560	\$1,119	\$1,679	\$2,238	\$2,798	\$3,357	\$3,917	\$4,477	\$5,036
40%	\$746	\$1,492	\$2,238	\$2,984	\$3,731	\$4,477	\$5,223	\$5,969	\$6,715
50%	\$933	\$1,865	\$2,798	\$3,731	\$4,663	\$5,596	\$6,528	\$7,461	\$8,394
60%	\$1,119	\$2,238	\$3,357	\$4,477	\$5,596	\$6,715	\$7,834	\$8,953	\$10,072
70%	\$1,306	\$2,611	\$3,917	\$5,223	\$6,528	\$7,834	\$9,140	\$10,445	\$11,751
80%	\$1,492	\$2,984	\$4,477	\$5,969	\$7,461	\$8,953	\$10,445	\$11,938	\$13,430
90%	\$1,679	\$3,357	\$5,036	\$6,715	\$8,394	\$10,072	\$11,751	\$13,430	\$15,109
100%	\$1,865	\$3,731	\$5,596	\$7,461	\$9,326	\$11,192	\$13,057	\$14,922	\$16,787

Source: Company data, Credit Suisse estimates

2. **Potential E7 (Ultra-Premium AI Bundle) SKU:** We assume that MSFT will look to capitalize on advanced AI functionality to create an E7 SKU for knowledge workers that rely heavily on the Office suite for their day-to-day tasks by adding advanced AI functionality to a new E7 bundle, similar to how MSFT created the E5 SKU by adding security + UCaaS (Teams Phone) functionality. As our base case we assume 10% installed base penetration of a potential E7 SKU, reflecting a slight discount to E5 penetration (12% as of F4Q22). We assume E7 represents a ~4x ASP uplift relative to average ASP: to derive this we average the price increase of E3 (\$23/month) relative to E1 (\$10/month) and the increase of E5 (\$38/month) relative to E3 (\$23/month) which implies a price for E7 of ~\$75/month which represents a ~4x uplift relative to the average ASP on the O365 commercial base implied by [our most recent CIO Survey](#).

**Figure 58: Revenue Uplift Sensitivity – Potential E7 Rollout**

Installed Base Penetration	Increase to Average ASP				
	200%	300%	400%	500%	600%
5%	\$3,731	\$5,596	\$7,461	\$9,326	\$11,192
10%	\$7,461	\$11,192	\$14,922	\$18,653	\$22,383
20%	\$14,922	\$22,383	\$29,844	\$37,305	\$44,766
30%	\$22,383	\$33,575	\$44,766	\$55,958	\$67,149
40%	\$29,844	\$44,766	\$59,688	\$74,610	\$89,532
50%	\$37,305	\$55,958	\$74,610	\$93,263	\$111,915

Source: Company data, Credit Suisse estimates

In Figure 58 we highlight revenue uplift sensitivities assuming various levels of price increases related to a potential E7 SKU along with various installed base penetrations. Given E7 would be an ultra-premium SKU with a more limited addressable base we view the likelihood of this reaching penetration north of 50% as less likely. In Figure 59 we highlight operating income uplift sensitivities under the same ASP/installed base penetration scenarios, assuming a 50% incremental operating margin, the actual incremental margin would likely be substantially higher given there is minimal incremental associated operating expenses. However, we model 50%, which is slightly above segment level PBP margins (46.9% in FY22) to be conservative.

**Figure 59: Op. Inc Uplift Sensitivity – Potential E7 Rollout**

Installed Base Penetration	Increase to Average ASP				
	200%	300%	400%	500%	600%
5%	\$1,865	\$2,798	\$3,731	\$4,663	\$5,596
10%	\$3,731	\$5,596	\$7,461	\$9,326	\$11,192
20%	\$7,461	\$11,192	\$14,922	\$18,653	\$22,383
30%	\$11,192	\$16,787	\$22,383	\$27,979	\$33,575
40%	\$14,922	\$22,383	\$29,844	\$37,305	\$44,766
50%	\$18,653	\$27,979	\$37,305	\$46,631	\$55,958

Source: Company data, Credit Suisse estimates

**Figure 60: Total MSFT Revenue/EPS Uplift – GPT Integration into Productivity Suite****Revenue Uplift at 5-25% Existing Installed Base Penetration**

	<b>5%</b>	<b>10%</b>	<b>15%</b>	<b>20%</b>	<b>25%</b>
GitHub - Copilot (Blended ASP)	\$783	\$1,566	\$2,349	\$3,132	\$3,915
Teams - Premium	\$1,680	\$3,360	\$5,040	\$6,720	\$8,400
Dynamics - Viva Sales	\$192	\$384	\$576	\$768	\$960
<i>LT Potential O365 - E7 SKU (Ultra-Premium "AI" Bundle), 4x Avg. ASP</i>	\$7,461	\$14,922	\$29,844	\$44,766	\$59,688
<i>LT Potential O365 - "AI/GPT" O365 Price Increase (100% of Base)</i>	\$18,653	\$18,653	\$18,653	\$18,653	\$18,653
<b>Total Revenue to MSFT</b>	<b>\$28,769</b>	<b>\$38,885</b>	<b>\$56,462</b>	<b>\$74,039</b>	<b>\$91,616</b>

**Operating Income Uplift at 5-25% Existing Installed Base Penetration**

	<b>5%</b>	<b>10%</b>	<b>15%</b>	<b>20%</b>	<b>25%</b>
GitHub - Copilot (Blended ASP)	\$341	\$681	\$1,022	\$1,362	\$1,703
Teams - Premium	\$788	\$1,576	\$2,364	\$3,152	\$3,940
Dynamics - Viva Sales	\$90	\$180	\$270	\$360	\$450
<i>LT Potential O365 - E7 SKU (Ultra-Premium "AI" Bundle), 4x Avg. ASP</i>	\$3,731	\$7,461	\$14,922	\$22,383	\$29,844
<i>LT Potential O365 - "AI/GPT" O365 Price Increase (100% of Base)</i>	\$9,326	\$9,326	\$9,326	\$9,326	\$9,326
<b>Total Operating Income to MSFT</b>	<b>\$14,275</b>	<b>\$19,224</b>	<b>\$27,904</b>	<b>\$36,584</b>	<b>\$45,263</b>

**EPS Uplift at 5-25% Existing Installed Base Penetration**

	<b>5%</b>	<b>10%</b>	<b>15%</b>	<b>20%</b>	<b>25%</b>
GitHub - Copilot (Blended ASP)	\$0.04	\$0.07	\$0.11	\$0.15	\$0.18
Teams - Premium	\$0.09	\$0.17	\$0.26	\$0.34	\$0.43
Dynamics - Viva Sales	\$0.01	\$0.02	\$0.03	\$0.04	\$0.05
<i>LT Potential O365 - E7 SKU (Ultra-Premium "AI" Bundle), 4x Avg. ASP</i>	\$0.40	\$0.81	\$1.62	\$2.43	\$3.23
<i>LT Potential O365 - "AI/GPT" O365 Price Increase (100% of Base)</i>	\$1.01	\$1.01	\$1.01	\$1.01	\$1.01
<b>Total MSFT EPS Uplift</b>	<b>\$1.55</b>	<b>\$2.08</b>	<b>\$3.02</b>	<b>\$3.97</b>	<b>\$4.91</b>
<i>Uplift vs FY2022 EPS</i>	16.8%	22.6%	32.8%	43.1%	53.3%

Source: Company data, Credit Suisse estimates

Altogether we continue to believe we are in the early stages of GPT integrations into MSFT Office Suite, but expect a ~20% lift to EPS, likely over a period of 5+ years (adding ~3-5%/year to EPS growth), with sound assumptions around price increases following increased feature sets, as discussed earlier.

# The AI Landscape – Key Milestones to Note

Even though our report has focused on what MSFT is doing and how OpenAI's models may integrate into their various applications/tools/services, MSFT is not the only AI bellwether in the US technology landscape worth discussing as this technology theme begins to materialize in various parts of tech. In this section we outline key AI development milestones by other AI Majors, including Google, Nvidia, IBM, Meta, and Amazon.

## Google – Covered by Stephen Ju, US Internet Analyst

The first instance of Google even mentioning Machine Learning or AI was some fifteen years ago when management highlighted the construction of models to assess how an ad will perform, before it is even shown – and the initial product rollouts started in 2004. And within the Internet coverage universe, it is arguably the company with the longest history of investments and innovations in AI and Machine Learning, to the extent that it now permeates almost all of its consumer, advertiser, and enterprise (Google Cloud Platform) facing products.

As noted in its latest 10K, “we believe that AI is a foundational and transformational technology that will provide compelling and helpful benefits to people and society through its capacity to assist, complement, empower, and inspire people in almost every field of human endeavor.”

While the following list of product releases and milestones are by no means comprehensive, we have compiled the path of Google's innovations from 2007 until now, as shown in Figure 61.

- **Speech and Music Recognition:** First announced in 2007, it has on an ongoing basis iterated this product over the last 15+ years, with some of the highlight products being Voice Search (2008), Auto Captioning (2009) Duplex (2018) which enabled natural conversations to complete tasks over the phone, to the Google Assistant (2016), as well as LaMDA (Language Model for Dialogue Applications), which is trained to carry on free-flowing open-ended conversations across a wide array of topics and multiple turns of dialogue.
- **Image (2D and 3D) Recognition and Video Contextualization:** Initially rolled out for copyright and content safety considerations, this has since gone on to power Mediapipe Objectron (2020) for 3D object detection trained from data collected from self-driving cars to currently Multisearch (2022) which allows for users to run queries based on images or from their camera. Identification of what is actually in video content within YouTube also opens up potential for better matching with ad content, as well as predictive analytics of what the user would like to watch next.
- **Launch of TensorFlow, the TPU, followed by announcements in Quantum Computing / Sycamore to deliver solutions too complex for classical computing:** TensorFlow, which is an open source software library for machine learning (2015), was followed by a purpose-built ASIC chip Tensor Processing Unit (TPU) in 2016, quantum computing breakthrough in 2019 to ultimately TensorFlow Quantum to integrate Cirq (writing of quantum algorithms with quantum computers) with classical neural networks.

**Figure 61: Key AI Milestones Achieved by GOOGL Since 2007**

Launches YouTube Video Identification product for copyright and content safety	2007	Announces first AI-based efforts for audio processing for speech and music recognition
Launches Google411, featuring speech recognition to converse with user queries for business search	2008	Launches Voice Search in English
Launches auto-captioning on YouTube	2009	Expands machine driven language translation from 23 to 34 on translate.google.com
Launches Google Photos upgrade, which uses computer vision and ML technology to auto-assign tags for photos to make them searchable	2013	Launches Quantum Artificial Intelligence Lab
Opens up Google Cloud Platform for genomics computing	2014	Launches Lens Blue in Google Camera, using computer vision to simulate a 3D model within the picture to estimate depth
Announces plans to build quantum information processors based on superconducting electronics at its Quantum Artificial Intelligence Lab	2014	Google acquires Deep Mind
Announces its own AI accelerator purpose-built ASIC chip – Tensor Processing Unit (TPU) – for machine learning and model usage	2016	Launches initial version of TensorFlow, which is an open source software library for machine learning
Announces release of Cloud Machine Learning on Google Cloud Platform	2016	AlphaGo, trained to play the game Go, combining tree search w/two deep neural networks
Launches two generative AI products - AI Piano Duet for the creation of music and SketchRNN to teach machines how to draw	2017	Releases chat app Allo, which uses Smart Reply features previously featured for Gmail
Launches Google Duplex and incorporates into Assistant, for conducting natural conversations to complete tasks over the phone	2018	Announces Transformer, which is a Neural Network built specifically for language understanding
Launches BigQuery ML, which combines large data set analytics in BigQuery with machine learning for predictive analytics	2018	Expands Smart Reply into Smart Compose, which offers interactive sentence completion to allow for faster email composition
Launches purpose-built ASIC Edge TPU for running TensorFlow Lite on small compute devices	2018	Open sources BERT to facilitate training of question answering systems, and improve results from unidirectional models used to predict the next word conditioned on the previous only
Formalizes healthcare efforts and forms Google Health, which is an integration of Deepmind Health as well as Google Hardware	2018	Launches AutoML in alpha, which allows for the creation of models without prior machine learnings expertise
Launches TensorFlow 2.0	2019	Announces Cirq to help develop practical computational problems using quantum computers
Launches on-device Live Caption based on combination of recurrent neural network	2019	Launches end-to-end speech-to-speech translation model Translatotron
Announces ALBERT to lower memory consumption and increase the training speed of BERT	2019	Announces quantum supremacy breakthru in first step to deliver over LT solutions too complex for classical computing
Announces TensorFlow Quantum, integrating Cirq with TensorFlow to deliver hybrid quantum-classical neural networks	2020	Announces conversational chatbot Meena, trained on 8.5x more data versus OpenAI GPT-2 with 1.7x greater model capacity
Announces real-time 3D object detection product – Mediapipe Objectron – from 2D imagery	2020	Announces Dreamer with Deepmind, which uses real world images for reinforcement learning 4
Mediapipe Holistic which enables perception of human poses, face, and hand components in real time	2020	URL2Video, a generative AI product which converts web pages into short form video by extracting assets and design styles from HTML
Releases TensorFlow 3D to bring deep learning to TensorFlow to train and deploy 3D scene understanding models	2021	Announces AutoML for time series forecasting
Image generation products SR3 and CDM which allow for the creation of hi-res image from a low-res example, as well as create realistic images	2021	Language Model for Dialogue Applications – LaMDA, trained to carry on free-flowing open-ended conversations
Pathways, to enable a single model to handle a myriad of tasks vs the current setup of models being trained to do one thing	2021	Announces text to image generation for the rapid content creation
Announces two image generation efforts – Imagen and Parti – for text to image generation	2022	Announces TPU v4 to deliver 2x the compute power of TPU v3, to enable large model training
Announces AudioLM for the generation realistic speech and piano music by listening to audio only	2022	Imagen Video and Phenaki – the former for hi-res videos which cap at ~5 seconds and the latter for variable length

Source: Company website

## Nvidia – Covered by Chris Caso, US Semiconductors Analyst

Nvidia believes that AI adoption is at an inflection point and it is poised to take advantage of the inflection given its GPU dominance in the AI market. On its 4Q call, the company noted “generative AI applications will help almost every industry do more faster.” and we underscore below Nvidia’s major developments that have contributed to the AI landscape.

- **CUDA:** In 2007, NVIDIA introduced CUDA, a parallel computing platform and programming model that enables developers to use GPUs for general-purpose computing, including AI and machine learning.
- **Drive AGX:** NVIDIA introduced Drive AGX in 2014, a hardware and software platform for autonomous vehicles, including perception, localization, and mapping.
- **GANs:** In 2014, NVIDIA researchers introduced GANs (Generative Adversarial Networks), a neural network architecture that can generate realistic images and videos.
- **Tesla V100:** NVIDIA introduced the Tesla V100 (2017), a high-performance GPU designed specifically for deep learning and other AI applications.
- **A100:** In 2020, NVIDIA introduced the A100, a new GPU designed for AI and high-performance computing workloads, with features like Tensor Cores and Multi-Instance GPU (MIG) technology for greater efficiency and scalability.

**Figure 62: Other Key AI Milestones Achieved by Nvidia**

<p>Announced NVIDIA's DGX line of AI supercomputers that provide an integrated hardware and software platform for AI research and development, with capabilities for training and inference on large datasets.</p>	<p>2016</p>	<p>NVIDIA has developed and contributed to a number of deep learning frameworks, including Caffe, TensorFlow, and PyTorch.</p>
<p>Launch of NVIDIA Isaac, an AI-powered platform for robotics, providing capabilities for perception, manipulation, and navigation.</p>	<p>2016</p>	<p>Launch of NVIDIA's Jetson platform which provides a compact, power-efficient hardware and software platform for AI at the edge, including robotics, drones, and...</p>
<p>Launch of DGX Cloud - available through Oracle Cloud infrastructure and Azure, Google GCP, others coming. Provides access to NVIDIA AI Enterprise for training and deploying large language models or other AI workloads.</p>	<p>2018</p>	<p>Launch of NVIDIA Clara, an AI-powered healthcare platform that uses deep learning to improve medical imaging and other healthcare applications.</p>
	<p>2018</p>	
	<p>2023</p>	

Source: Company website

## IBM – Covered by Shannon Cross, US IT Hardware

As noted in IBM's latest 10Q, "As technology remains a fundamental source of competitive advantage, we continue to see solid demand for our hybrid cloud and AI solutions." Further "Companies are eager to deploy AI and automation capabilities to boost their levels of productivity." As such, IBM has been investing in AI for a few decades, and we list some of the key projects and milestones the company has achieved along the way.

- **Deep Blue Defeats the World Chess Champion:** IBM began developing Deep Blue in 1989 and in 1997 the computer was capable of evaluating 200 million possible chess positions per second. Deep Blue then beat the world champion in chess, marking the first time a computer beat a human world champion.
- **Watson Wins Jeopardy!:** In 2011, the natural language processor competed and won first place against two of the most successful Jeopardy! players of all time.
- **AI Fairness 360:** Bias in training data, due to either prejudice in labels or under-/over-sampling, yields models with unwanted bias. Announced in 2018, AIF360 is a comprehensive open-source toolkit of metrics to check for unwanted bias in data sets and machine learning models, and state-of-the-art algorithms to mitigate such bias.
- **Watson AIOps Automates IT Ops:** In present day, Watson has evolved into a suite of AI-powered products. This includes Watson AIOps (first released in 2020) which helps companies automate IT infrastructure management including predicting and preventing outages. IBM has made several AI-related acquisitions in recent years and integrated the products into Watson AIOps (e.g., Instana, Turbonomic).
- **z16 Mainframes with On Chip Inferencing:** The most recent mainframe refresh released in 2022 features on chip inferencing, the first of its kind. z16 mainframes use an IBM Telum processor and can analyze transactions in real-time. z16 can process 300 billion inference requests per day with one millisecond of latency.
- **Project Wisdom Expands AI for Code Efforts:** Announced in 2022, Project Wisdom suggests coding language based on simple English text such as "Configure webapp setting." Red Hat Ansible is using this capability to bridge Ansible's specific coding language and human language to make it easier and faster to build automations.
- **AI Supercomputer in The Cloud:** In January 2023, IBM revealed Vela, an AI-optimized cloud-native supercomputer. Given the recent proliferation of AI, IBM developed Vela to extend the hybrid cloud experience and move AI resources closer to where customer data may lie. IBM is still completing inferencing tasks and training of foundation language models on-prem, but Vela will give customers access to pre-packaged models that can be finetuned on a smaller scale in the cloud (otherwise cost prohibitive). Currently, Vela is only available to the IBM Research community.

**Figure 63: Other Key AI Milestones Achieved by IBM**

In 2014, IBM introduced Watson Discovery Advisor, an AI system that can help researchers identify new connections and patterns in scientific research.	2014	IBM Maximo is an AI-powered asset management system that uses machine learning to monitor and optimize the performance of industrial assets, such as turbines and factory equipment.
In 2016, IBM announced the opening of its Research AI Hardware Center, which is dedicated to advancing the state-of-the-art in AI hardware, including processors, memory systems, and other components.	2016	IBM Watson Health is a former division of IBM that used AI & machine learning to provide insights and analysis in the healthcare industry, with applications in drug discovery, clinical research, & medical imaging analysis.
Telum is the CPU for IBM's next-generation of IBM Z and LinuxONE systems. It is a 7NM chip with a dedicated on-chip accelerator for AI inferencing.	2020	Watson AIOps which helps companies automate IT infrastructure management including predicting & preventing outages.
Project Wisdom expands AI For Code efforts by automating coding within Ansible.	2022	z16 mainframes use the IBM Telum processor and can analyze transactions in real-time including processing of up to 300 billion inference requests per day with one millisecond of latency.
IBM has been a leader in the development of quantum computing, and IBM Condor is launching in 2023 as the first quantum computer with more than 1,000 qubits.	2023	IBM announces Vela, an AI-optimized cloud-native supercomputer. Eventually Vela will provide pre-packaged models for customers to access and/or customize in the cloud.

Source: Company website

## Meta – Covered by Stephen Ju, US Internet Analyst

On its 4Q earnings call, Meta management noted the two major technological waves driving its roadmap are AI today, and over the longer term, the metaverse. AI is the foundation of Meta's discovery engine and ads business; thus, it has been developing AI since the company's inception. Below, we highlight key milestones in Meta's AI development journey.

- **DeepFace:** In 2014, Facebook's AI research team unveiled DeepFace, an AI system capable of recognizing faces in images with 97.35% accuracy, which was at the time the most accurate system in existence.
- **DeepText:** In 2016, Facebook introduced DeepText, a natural language processing (NLP) system that can understand and interpret the textual content of billions of posts and comments across 44 languages at launch.
- **Reinforcement Learning:** In 2017, Facebook announced that it was developing a new type of reinforcement learning algorithm, which uses trial and error to improve its performance over time.
- **Rosetta:** In 2018, Facebook introduced Rosetta, a scalable optical character recognition (OCR) system that can read and understand text within images and videos, even in multiple languages.
- **GPT-3 Partnership:** In 2020, Facebook partnered with OpenAI to gain access to GPT-3, a powerful language model that can generate text, answer questions, and perform a range of other tasks.
- **LLaMa Released for Researchers:** In February 2023, Meta released a state-of-the-art large language model designed to help researchers advance their work in this subfield of AI.

**Figure 64: Other Key AI Milestones Achieved by Meta Platforms**

<p><b>2015</b></p> <p>Meta releases PyTorch, an open-source machine learning library for Python that has become increasingly popular among researchers and developers.</p>	<p><b>2017</b></p> <p>Meta launches M, an AI-powered personal assistant integrated into its Messenger app. M could help users with tasks like making reservations, booking flights, and ordering food.</p>
<p><b>2018</b></p> <p>Meta introduces DIAL (Dialog Agent Learning), an AI system that uses reinforcement learning to improve the conversational abilities of its chatbots.</p>	<p><b>2019</b></p> <p>Meta announces the creation of an independent AI ethics board, which would review and oversee the company's use of AI technologies.</p>
<p><b>2021</b></p> <p>Meta introduces Horizon Workrooms, a virtual reality platform that uses AI to simulate a collaborative work environment, allowing users to interact with one another and shared virtual content in real-time.</p>	<p><b>2021</b></p> <p>Meta announces the development of Meta Learning, an AI system that can teach other AI systems to learn more efficiently and effectively, with a focus on reducing the amount of training data needed for ML models.</p>

Source: Company website

## Amazon – Covered by Stephen Ju, US Internet Analyst

Per Amazon's latest 10K, "*Advances in technology, specifically the speed and reduced cost of processing power, data storage and analytics, improved wireless connectivity, and the practical applications of AI and machine learning, will continue to improve users' experience on the Internet and increase its ubiquity in people's lives.*" As such, below we highlight key milestones in Amazon's development of AI.

- **Amazon Mechanical Turk:** In 2005, Amazon launched Mechanical Turk, a crowdsourcing platform that uses AI to help businesses and researchers complete tasks that require human intelligence.
- **Amazon Forecast:** In 2008, Amazon introduced Forecast, a fully managed service that uses machine learning to deliver highly accurate forecasts based on historical data.
- **Amazon Alexa:** In 2014, Amazon introduced Alexa, an AI-powered voice assistant that can answer questions, play music, control smart home devices, and perform a wide range of other tasks.
- **Amazon Rekognition:** In 2016, Amazon launched Rekognition, an AI-powered image and video analysis service that can detect objects, scenes, and faces, and analyze text within images.
- **Amazon Personalize:** Launched in 2018, Personalize is a machine learning service that allows businesses to create personalized recommendations for their customers based on their behavior and preferences.
- **Amazon Go:** In 2018, Amazon opened its first cashier less grocery store, Amazon Go, which uses computer vision and machine learning to track customers and their purchases.

**Figure 65: Other Key AI Milestones Achieved by Amazon**

<p><b>Amazon introduces Translate, a neural machine translation service that can provide high-quality translations between multiple languages.</b></p>	2016	<p><b>Amazon introduces Polly, a text-to-speech service that uses deep learning algorithms to generate speech that sounds less robotic and more human.</b></p>
	2017	
	2017	<p><b>Amazon introduces SageMaker, a machine learning platform that makes it easier for developers to build, train, and deploy machine learning models.</b></p>
<p><b>Amazon introduces DeepLens, a programmable video camera that runs deep learning models on the edge, allowing developers to create AI-powered applications for a wide range of industries.</b></p>	2018	
	2018	<p><b>Amazon introduces DeepRacer, a fully autonomous 1/18th-scale race car that uses reinforcement learning to improve its performance on a racetrack.</b></p>
<p><b>Amazon launches Kendra, an enterprise search service that uses natural language processing and machine learning to provide more accurate and relevant search results.</b></p>	2020	

Source: Company website.

## China Internet – Covered by Kenneth Fong, China Internet Analyst

### Poised to Be a Game-changer for China Internet

We believe generative AI technology, currently represented by a ChatGPT-like AI chatbot, is poised to be a game-changer for China internet. We believe its capability to enhance the efficiency of language-related tasks and provide personalized responses to customer inquiries will ultimately lead to a more accessible and user-friendly online experience for China internet users. This will in turn unlock more monetization potential, mainly reflected in (1) expanding TAM for big platforms with established moats in AI infrastructure; (2) interactive conversations in chatbot, improving traffic stickiness; (3) greater efficiency in content creation, shortening the production cycle and leading to significant cost savings, favoring content platforms and games companies; and (4) more accurate user analysis, targeting an improving conversion rate and ad efficiency.

The recent surge of excitement around ChatGPT has instilled a sense of urgency within the Chinese internet sector to gear up investment in this groundbreaking technology. Developing AI chatbot products requires significant investments in large amounts of data and computing resources, creating a high technical entry barrier, which makes it more feasible for large China tech giants with strong technical and capital resources. Within the China internet operators, we believe **Baidu** is best positioned to capture the advancement of generative AI technology given its full stack AI capabilities, robust China-based data sets and user inquiries together could form a flywheel effect and make a far-reaching impact on its search ad and AI cloud business.

**ByteDance** and **Tencent** have been exploring AIGC technology to prompt ad/video content creation, while eCommerce platforms **Alibaba** and **JD** focus more on eCommerce services and logistics operations. We expect major tech platforms to shift focus on integrating their cutting-edge AI technology into various use cases as the next stage.

**Figure 66: China Internet Platforms' AI Layout and Potential Opportunities**

Internet platforms	AI architecture				Potential applications	
	AI chip	Pretrained model	Parameter	Model type	AIGC	AI chatbot
<b>Baidu</b>	Kunlun	ERNIE model	260bn (3.0 Titan)	Multimodal	Ad and video content generation, scriptwriting, digital avatar	Search, autonomous driving (Apollo/Jidu), DuerOS, Baidu map
<b>ByteDance</b>	na	DA-transformer	na	NLP	Ad/short video/VR content generation, scriptwriting	Search, eCommerce (customer support), Feishu
<b>Tencent</b>	Zixiao	HunYuan	trn	Multimodal	Ad and video content generation, scriptwriting, games	Search, recommendation, payment, Tencent meeting
<b>Alibaba</b>	Hanguang800	M6	10trn	Multimodal	Marketing / product related content creation	eCommerce (customer support, search, product recommendation), Dingtalk
<b>JD</b>	na	K-PLUG	Hundreds of billions	NLP	Content generation, scriptwriting, digital avatar	eCommerce (customer service, search), payment, logistics
<b>NetEase</b>	na	na	na	na	Games	Youdao (oral language instruction, writing revision), games (dialogue with NPC)

Source: Company data, Credit Suisse

**Baidu:** We believe Baidu could leverage its technical expertise and resources to become a major operator in the generative AI industry, through deepening integration of AI chatbot product and AIGC capabilities into its mobile ecosystem, cloud, and autonomous driving. The company plans to release the ChatGPT-like initiative "**Ernie Bot**" (文心一言) on March 16, 2023, making it the first among Chinese companies to join the global AI chatbot race. Baidu will gradually make the product available through a waitlist at a nascent stage and fully integrated into its ecosystem once the user experience has been meaningfully improved. We've highlighted Baidu's **core competencies**:

**#1: Comprehensive AI capability.** Baidu has established its unique positioning as a leader across four tiers of AI architecture, including AI chips (Kunlun chip), deep learning platform (Paddle Paddle), large language model (Ernie), and applications (search engine), which enables it to capture a dominant market share and empower industry vertical operators.

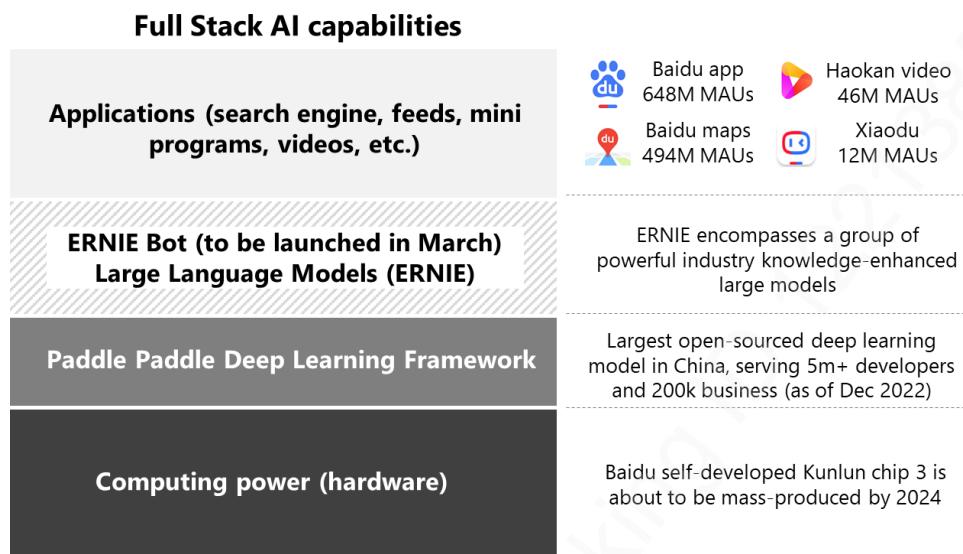
**#2: A solid foundation of China-based data sets:** Baidu's large language model started with Ernie 1.0 in 2019 and has since evolved to Ernie 3.0 Titan, a model with up to 260bn parameters, which is considered more effective and comprehensive for various Chinese-based NLP tasks, giving it a competitive edge over its English-focused counterparts like ChatGPT.

**#3: Suitable application with affluent inquiries:** Baidu has accumulated vast troves of general search data across industries, which has been "well tagged" on a daily basis, making it more efficient to train large language models. Based on its search engine, Baidu could obtain a larger volume of inquiries for model fine-tuning, creating a virtuous cycle that can be difficult for competitors to replicate.

**#4: Baidu leading a paradigm shift to AIGC.** As LLMs evolve to the multi-modal models in the next stage, this will unlock more AIGC opportunities and drive the content creation efficiency of more diverse form, such as graphs, video, music, etc. Baidu's ERNIE model matrix is prepared for such paradigm shift, exemplified by its Ernie 3.0 Zeus and Ernie-ViLG 2.0, featuring versatile content creation (novel, poetry, script) and visual images from text

prompt, respectively. Also, Ernie Yige (文心一格) was updated with new features such as turning photos into AI-generated art, image editing, and one-click video production.

**Figure 67: Baidu Comprehensive AI Capabilities**



Source: Company data, Credit Suisse. Note: Baidu MAUs are sourced from company disclosure as of Dec 2022; Haokan video, Baidu maps, and Xiaodu MAUs are based on QuestMobile

## Implications for Baidu

- Solidify position in search:** The key capabilities of chatbots, referring to the ability to comprehend the nuances of natural language and understand complex queries, make them ideal for intuitive and efficient search results. As the largest search engine in China, Baidu's key strategy this year is to transform its search system into a "**search + AI generative dual system**". We think Ernie Bot may serve as a complementary tool to its traditional search engine, similar to Microsoft's new Bing, bringing **incremental benefits in (1) boosting traffic**: If Baidu could consistently deliver high-quality content in the chatbot, users could enjoy a more comprehensive and efficient search experience. This could in turn boost traffic on Baidu and potentially mitigate further traffic loss to other internet peers; **(2) deepening monetization of long-tail content**: Providing more personalized and relevant content recommendations could improve the efficiency of ad lead generation and reduce user churn during marketing, resulting in higher conversion for some long-tail content; **(3) expanding search market**: With in-app search accounting for c.45% of the search market, Baidu could potentially gain more search shares through partnerships with app developers and integrate Ernie Bot into their platforms, who are less capable of developing their proprietary chatbot products.
- A thriving ToB ecosystem:** 400+ enterprises expressed strong interest in partnering with Ernie Bot for initial testing and exploring potential integration into their applications, demonstrating its significant potential for disruption across industries. We believe it also could offer **an upselling opportunity** to its AI Cloud. The key demand can mainly be categorized into (1) content production, (2) intelligent customer services, and (3) B2B digital services. Some key examples are as follows:
  - Media:** Media companies like 36Kr, Sina finance, and other state media were the first batch of customers to join Ernie Bot, aiming to explore interactive user dialogue and efficient enterprise service review.
  - Online video:** iQiyi reached a cooperative agreement with Baidu to explore AIGC techs for content search and novel creation, etc. The partnership could help iQiyi to streamline its content creation process and drive a more premier viewing experience,

solidifying its leadership in the video market amid a fiercely competitive video industry, in our view.

- **Advertising:** Xinchao media is teaming up with Baidu to bring generative AI technology to outdoor advertising and build an intelligent platform to generate more measurable data for advertisers. By analyzing customer data, it can provide more relevant and tailored ad suggestions, which in turn drives customer engagement and conversion.
- **Fintech:** Citic aiBank Corporation plans to integrate Baidu Ernie Bot into its intelligent customer service, marketing and AI digital assistant, which could effectively reduce bank operation costs and enhance service quality.
- **Auto:** Jidu, backed by Geely and Baidu, is considering the integration of Ernie Bot into its first production model Robo-01 SUV, creating more natural conversions between users and smart vehicles.
- **Software:** Hand enterprise solution is looking to integrate Baidu's AIGC tech into its financial accounting, data analysis, and marketing, etc.

**Figure 68: Ernie Bot Growing Ecosystem at a Glance**

Baidu Ernie Bot ecosystem						
Key categories	Content production		Intelligent customer services		B2B digital service	
Industry	Media	Online content (games, video, etc)	Advertising	Banks & Fintech	Auto (IoV)	Software & IT service
Potential applications	<ul style="list-style-type: none"> <li>▪ Provide personalized news feeds, alerts</li> <li>▪ Answer reader inquiries</li> </ul>	<ul style="list-style-type: none"> <li>▪ Facilitate efficient content production</li> <li>▪ Create conversational NPC in games</li> </ul>	<ul style="list-style-type: none"> <li>▪ Create effective ad campaigns</li> <li>▪ Analyze user data and offer insights</li> </ul>	<ul style="list-style-type: none"> <li>▪ Help basic inquiries</li> <li>▪ Detect fraudulent activities</li> <li>▪ Provide personalized financial advices</li> </ul>	<ul style="list-style-type: none"> <li>▪ Provide personalized services, such as navigation assistance, entertainment</li> </ul>	<ul style="list-style-type: none"> <li>▪ Develop conversational AI interface and virtual assistant</li> <li>▪ Streamline deployment process</li> <li>▪ Automate information management</li> </ul>
Examples of partners						

Source: Company website, 36kr, Credit Suisse

## Implications on Other Chinese Internet Companies

- **ByteDance:** We are optimistic on ByteDance's growth potential in the field of AIGC technology given its rich data in algorithm recommendation and proven strong execution. It remains focused on content at this stage, such as utilizing AIGC tech to propel the production efficiency on newsfeed, short videos, and VR content. More importantly, with search business being elevated to one of its strategic priorities, it has been cultivating user habits in search and amassing valuable search data from Douyin and Toutiao that can be used to train its LLM.
- **Tencent:** Tencent recently has established a dedicated team to create a generative AI chatbot, named "HunyuanAide" based on its Hunyuan AI model, mainly used for ad scenarios. For instance, Tencent could leverage the Hunyuan model to support intelligent

ad recommendations and produce tailored-made content for advertisers, which in turn reduces ad content costs.

- **Alibaba:** Alibaba has been steadily expanding its AI presence. Its AI multi-modal, known as M6, has been trained on a vast data set of Chinese language text and images since 2021 and now supports 10trn parameters. Alibaba's major applications would be mainly associated with intelligent marketing and customer support in **eCommerce scenarios**, in our view. By analyzing user preferences and behaviors, it could efficiently generate personalized marketing content. Another feasible application could be using chatbots in customer service to handle large volumes of customer inquiries with high accuracy and efficiency. Also, according to the company a ChatGPT-like product is now under internal testing by Damo Academy, and it is likely to be integrated into its office collaboration platform DingTalk.
- **JD:** JD cloud has accumulated industrial experience in intelligent human-machine dialogue and applied it to customer service, serving nearly 600mn customers and 170k+ third-party merchants. Additionally, JD has utilized its proprietary K-PLUG large model to generate product descriptions for 3k+ SKUs. On Feb 10, Yanxi AI application platform under JD Cloud unveiled its plan to release an industrial version of ChatGPT, namely **ChatJD**, along with its "125" roadmap, which aims to apply five functions (human-machine dialogue, user intention analysis, content generation, information extraction, and emotional classification) in two sectors (retail, finance). With a focus on vertical industries, ChatJD aims to quickly achieve generalization across industries and build the flywheel to drive a self-reinforcing system in more segmented scenarios.
- **NetEase:** While NetEase may not have as much accumulated AI technology as other tech giants, it still has great potential to apply ChatGPT technology to its **online education** and **games**. For example, its mobile title *Justice Online* features intelligent NPCs (non-game characters) to interact with players in a more natural and immersive manner. Besides, ChatGPT technology can also be used to generate in-game text, such as subtitles and dialogues, thereby enhancing the quality of the games' narratives. In addition to games, Youdao is exploring the application of AIGC technology in AI-assisted oral English teaching and writing revision. Specifically, AI writing tools offer immediate and comprehensive feedback during the writing process, while AI chatbots can provide real-time conversion practice to quickly improve students' language proficiency.
- **Kingsoft Office:** We believe ChatGPT technology has the potential to significantly optimize and restructure productivity software, the most frequently used tool that heavily relies on language processing. KO has been building up its AI capabilities in computer vision and natural language processing, mainly utilized in OCR (optical character recognition), smart translation, and writing. While it currently doesn't formulate a specific plan for ChatGPT technology integration, we see several potential function upgrades for WPS, including (1) language-based features, such as search, recommendation, and writing assistance; (2) automated document generation, enabling WPS to create templates with ease; (3) real-time translation; (4) easy conversion of text & graph between Excel and Presentation.

## Regulation

With the global AI race heating up, on February 24 China's Ministry of Science and Technology explicitly reaffirmed the government's unwavering commitment to advancing new technology and innovation, perceiving AI as a strategic emerging industry and a new growth engine. The chief of the ministry highly recognized that natural language processing has considerable potential to be integrated into a swath of industries and proposed **four major initiatives**: (1) promoting an open and collaborative innovation system for AI; (2) deeply integrating AI with the China economy and society; (3) setting up a system for AI security and governance; and (4) encouraging all aspects of AI cooperation. At the same time, the regulator also mentioned striking a balance among new tech innovations and avoiding unethical implications. Overall this

resolute stance underscores China's determination to remain at the forefront of cutting-edge technology and ongoing support for China internet companies.

## Key challenges

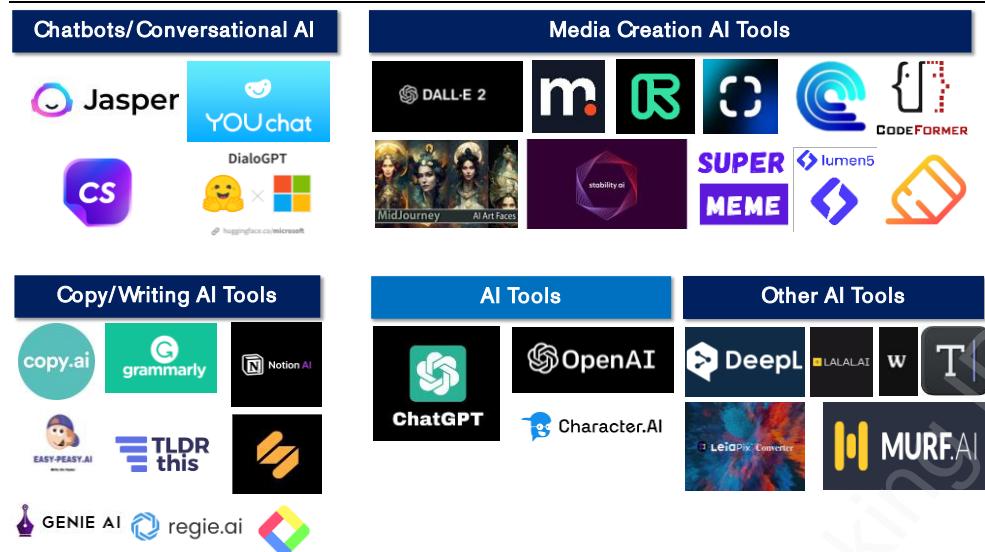
The US has been at the forefront of natural language processing for decades, while China has been catching up rapidly in recent years, with **key challenges** related to content diversity, lack of user feedback, and insufficient advanced chip supplies.

- **Content diversity and accuracy.** China LLMs primarily involve engineers and researchers rather than the incorporation of a solid amount of user feedback, suggesting limited data input during the model iteration process. In contrast, OpenAI has deployed a user feedback mechanism, which is effective in detecting content accuracy and quality control. Therefore, it's important for Chinese tech giants to expand the scope of trained data, accumulate fine-tuning know-how, and incorporate user feedback mechanisms.
- **Potential tighter regulations on AIGC content.** AIGC content could face potential regulations due to concerns over data security and regulatory content oversight. Additionally, there may be ethical concerns about using AIGC content for certain applications, such as academia and journalism.
- **High volume of training and operating costs.** To generate content that meets high-quality standards, AI models must be trained on vast amounts of data. These data must be curated, preprocessed, and carefully labeled, which can be a time-consuming and expensive process for many operators. Moreover, training large AI models requires significant computing resources, including high-capacity storage and powerful processors, mainly accessible by large tech platforms.
- **Insufficient high-performance chip supply amid geopolitical tension.** US trade restrictions on advanced AI chips and further deterioration of US-China relations could hinder domestic companies' pace in scaling up AI chatbot services. Currently China is still reliant on external chip supplies to support large-scale AI scenarios. We've seen increased efforts on high tech localization, but it still takes time to fully realize self-sufficiency.

# The GPT and AI Ecosystem Is Already Large

Below we discuss the private companies developing the AI Ecosystem start-ups.

**Figure 69: The Private AI Ecosystem Already Includes Robust List of Players**

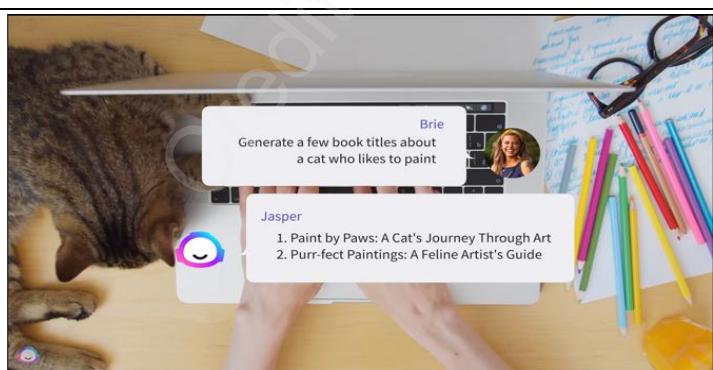


Source: Company data.

Below we offer a brief description on each private AI company for a deeper look into the ecosystem:

- **Jasper Chat** – Jasper Chat is a conversational chatbot built for businesses like advertising, and marketing, with a more conversational writing experience. Jasper AI is a writing assistant tool that utilizes artificial intelligence and natural language processing to help writers produce more compelling content, generate ideas, rewrite passages of text, answer questions, and build creative content such as poems or stories. Jasper has recently expanded into the text-to-image space with Jasper Art but is limited to information before and up to mid-2021. It also allows users to up-vote or down-vote any results, allowing Jasper to learn and become attuned over time. The Jasper Google Chrome extension enables users to generate content that fits in the context of their open tabs. On November 29, 2022, Jasper announced it would be utilizing Cerebras newly announced AI supercomputer to design and train its next set of customer-specific models. Jasper has raised \$125 million in a Series A funding round led by Insight Partners. Jasper's platform uses AI to generate written content based on initial prompts. The funding makes Jasper a unicorn with a valuation of \$1.5 billion.

**Figure 70: Example of Jasper Chat from Product Demo**



Source: Company website.

**Figure 71: More Than 70K Brands, Agencies and Content Creators Are on Jasper**



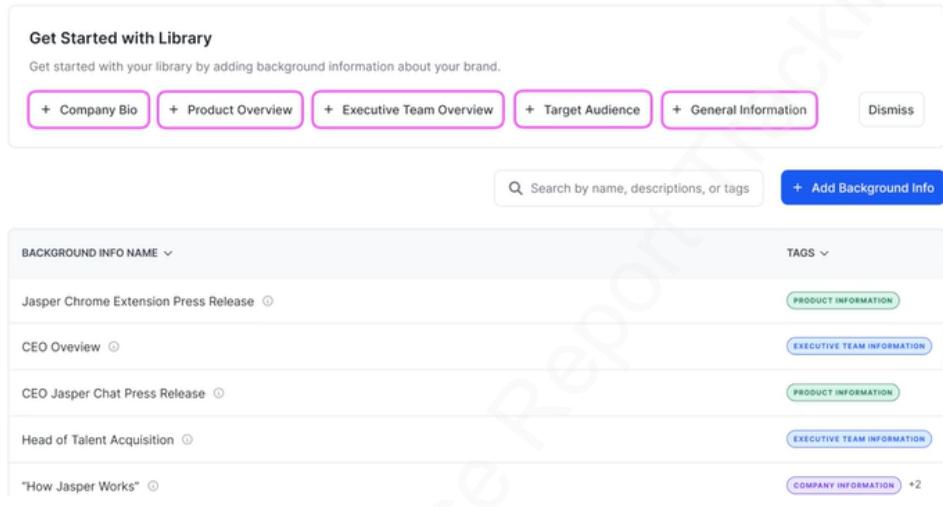
Source: Company website.

- On February 17, 2023, Jasper introduced the [Jasper for Business](#), which includes Brand Voice, Jasper API, and a suite of other tools, offering a viable route for businesses to adopt generative AI frictionlessly. Key highlights of the product release include: **(1) Jasper Brand Voice:** Through trainings using language, tone, and style of each brand, the AI-powered LM can generate customized content that enables each brand to speak to its audience in an authentic and identifiable way (Figure 72). **(2) Jasper API:** Jasper API was introduced, which can easily integrate Jasper with other content management system (CMS) and content platform. Jasper also announced some improved features with its browser extension, which now can be accessed across email, social channels, CMS, notes apps, etc. Moreover, the latest browser extension can work in both Chrome and Microsoft Edge. **(3) New Collaboration Features:** Jasper also introduced some new collaboration features that can increase coordination across teams, including offering Jasper space to each team member and enabling real-time doc sharing and workflow tracking, etc.

**Figure 72: By Training Using Language, Tone, and Style of Each Brand, Jasper Brand Voice Has Enabled Brands to Speak to Its Audience in an Identifiable Way**

## Library

Your library allows Jasper to read information specific to your company so that whenever he writes, it's 100% accurate to you and your company. To get started, begin adding background information specific to your brand.



The screenshot shows the 'Get Started with Library' section. It includes a search bar ('Search by name, descriptions, or tags') and a button ('+ Add Background Info'). Below this, there's a table with columns for 'BACKGROUND INFO NAME' and 'TAGS'. The table lists five entries:

BACKGROUND INFO NAME	TAGS
Jasper Chrome Extension Press Release	PRODUCT INFORMATION
CEO Overview	EXECUTIVE TEAM INFORMATION
CEO Jasper Chat Press Release	PRODUCT INFORMATION
Head of Talent Acquisition	EXECUTIVE TEAM INFORMATION
"How Jasper Works"	COMPANY INFORMATION +2

Source: Company data

- **ChatSonic AI** – Essentially a bulked-up version of ChatGPT; however, ChatSonic is not restricted to 2021 data as it offers information on the latest topics with Google-search integration. ChatSonic is trained and powered by Google Search to chat on current events and trending topics in real-time and can provide an alternative to ChatGPT in generating digital AI artwork for social media posts and digital campaigns. Additionally, ChatSonic can create customized personal assistants or avatars to help solve problems and can understand voice commands and responds just like Siri/Google Assistant. ChatSonic can also be used to create targeted campaigns tailored to customers' individual needs. ChatSonic raised \$2.6M in Sept 2021 in seed round led by Soma Capital, US and 20 other investors and are planning Series A/B funding to be raised in Q1 2023.
- **DALL.E.2** – A deep learning model developed by OpenAI used to generate digital images from natural language descriptions, called "prompts." DALL-E can generate imagery in multiple styles, including photorealistic imagery, paintings, and emoji. It can "manipulate and rearrange" objects in its images and can correctly place design elements in novel compositions without explicit instruction. Given an existing image, DALL-E 2 can produce "variations" of the image as unique outputs based on the original, as well as edit the image to modify or expand upon it.

**Figure 73: DALL.E.2 Can Expand Beyond the Original Artwork**

DALL-E 2 can expand images beyond what's in the original canvas, creating expansive new compositions.



Source: Company Website.

- **Copy.ai** – An AI-powered copywriting and content writing tool that automates content workflows using natural language processing and deep learning, enabling users to write blogs faster, write higher converting posts, and write more engaging emails.

**Figure 75: Copy.ai Help Users to Write Based on Detailed Context and Descriptions**

### How it works

**1 Enter your copywriting project**  
Choose from emails, social posts, long-form blog posts, and more!

**2 Give Copy.ai a bit of context**  
Enter a few sentences about your brand and products.

**3 Sift through the results**  
Our AI content generator gives you multiple options for each campaign that you're working on.

**4 Edit, polish, and publish**  
Use Copy.ai's editor to rewrite paragraphs and polish up sentences. Then, just copy and paste the work into your CMS for publishing.

Source: Company Website.

- **TLDR This** – An online article summarizer tool, that automatically extracts author and date information, related images, title and reading time from news articles and blog posts. It also

**Figure 74: DALL.E.2 Can Also Create Different Variations Based on the Original Art Piece**

DALL-E 2 can take an image and create different variations of it inspired by the original.



Source: Company Website.

**Figure 76: Copy.ai Covers a Broad Range of Topics and Can Generate Brainstorming Results Instantly**

copy ai Use cases Resources Weekly Demos Login Get Started — It's Free

**Blog Content**  
Write optimized blog posts in a fraction of the time [Try Blog Content](#)

**Digital Ad Copy**  
Produce compelling ad copy in just a few clicks [Try Digital Ad Copy](#)

**eCommerce Copy**  
Increase sales with better product descriptions [Try eCommerce Copy](#)

**Sales Copy**  
Convert your audience with copy that sells [Try Sales Copy](#)

**Social Media Content**  
Engage your audience with compelling social copy [Try Social Media Copy](#)

**Website Copy**  
Overhaul your website with more compelling copy [Try Website Copy](#)

It's Memorial Day here at Caviar, we're celebrating all the things we love about this country. We'd love to help you out with your next project, so give us a shout! We're offering 10% off our services. Whether you've got a new business idea that needs some help getting off the ground, or you need a fresh take on your brand's look, we can help. We've worked with businesses all over the world, and would love to be a part of yours. So stop by our site, and let us know how we can help!

We're celebrating all the freedom to work with clients who have a solid vision of their own. That's why we're offering 10% off our services this week. Whether you're looking for—website design, logo design, social media management, or anything else—we've got your back. We can't wait to talk with you about how we can help take your business to the next level.

We're closing out the weekend with a little treat: 10% off our services. Use code MEMORIALDAY to get your discount, and don't forget to thank a veteran for their service. We'd also like to take this opportunity to thank you for all that you do. We couldn't do it without your help—thank you!

Source: Company Website.

removes ads, popups, graphics, and other distractions to provide a clean and focused reading experience on websites. It also analyzes any piece of text and summarizes it automatically, in a way that makes it easy to read, understand and act on.

- **Stability AI:** The parent company of Stable Diffusion, a deep learning, text-to-image model released in 2022. It is primarily used to generate detailed images conditioned on text descriptions, though it can also be applied to generating image-to-image translations guided by a text prompt. The Stable Diffusion model supports the ability to generate new images from scratch through the use of a text prompt describing elements to be included or omitted from the output. The Stable Diffusion model supports the ability to generate new images from scratch through the use of a text prompt describing elements to be included or omitted from the output.
- **CodeFormer** – Uses an algorithm for face restoration for old photos and can generate AI faces.
- **Grammarly** – Grammarly is best known as a web plug-in and digital add-on for writing programs and communications platforms. The AI performs real-time analysis of the user's writing, including spelling, grammar, brevity, and language style, sharing suggestions along the way.
- **Runwayml.com** – A user can make their own AI-powered video or movie. From basic video editing to advanced post-processing, Runway offers professional video editing software on for visual effects, video formatting, color correction and more right inside a browser. Runway also provides allows for secure collaboration from anywhere in the world, allowing a user to share compositions with a link.
- **Lumen5** – A Free online A.I. video-making platform built to repurpose marketing content, providing 100+ templates with pre-designed formats. Creating with Lumen5 is just like making a presentation. Point, click, drag, drop. Lumen5 automatically turn blog posts into videos or transforms zoom recordings into captivating clips with hundreds of customizable designs that help make communicating effective and consistent.

**Figure 77: Lumen5 Offer AI-powered Video Creator, Enabling Users to Create Video Content Easier Than Building Slides**

**Easier than building a slide deck**

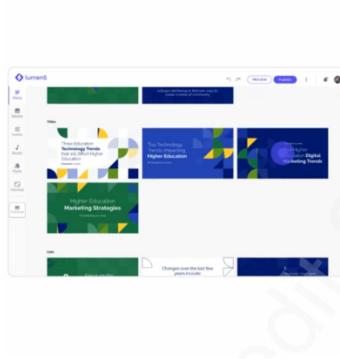
Creating with Lumen5 is just like making a presentation. Point and click, drag and drop, done and done.

[Get started for free →](#)

"One of our colleagues called Lumen5 the PowerPoint of video making. Everybody can use PowerPoint; it's easy."

**Michelle Heiso**  
Visual Content Director

**SIEMENS**



Source: Company Website.

**Figure 78: By Offering 100+ Templates with Pre-designed, Lumen5 Has Streamlined the Video Making Process with Less Budget Required for Users**

**Templates designed to help you stay on brand, on budget, and on time**

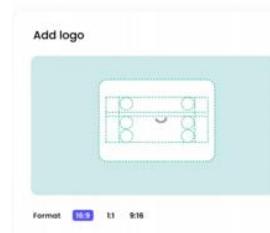
Tap into hundreds of customizable designs made to help you communicate effectively and consistently.

[Get started for free →](#)

"Lumen5 provides a great library of images, video footage, and music to deliver the energy and traction you need for your social posts."

**Shirley Pang**  
Global Social Media Manager

**CISCO**



Source: Company Website.

- **Simplified** – A platform AI tools to help marketers, writers, and social media managers to do more in the quickest time possible and using a single app. Simplified AI allows a user to instantly create copies using a keyword, rewrite content in multiple ways, generate images using natural language, and more. With more than 70 AI templates for landing pages, ads on social media, product descriptions, in more than 10 different tones, Simplified's AI Writer can write long-form content like essays, articles, and even books. It can even generate paragraphs using AI and instantly translates content into 30+ languages. There is also

support for content scheduling using the content/social calendar and can publish posts on various social media platforms.

- **Notion.ai** –A creative tool that can help write, brainstorm, edit, summarize, and perform other tasks employing generative AI to produce notes and other types of content. Notion.ai automates the creation of written content in the app, including blog posts, brainstorming ideas, to-do lists, and even literary works, using generative artificial intelligence.
- **Genei** – A tool that summarizes background reading and produces blogs, articles, and reports. Documents can be stored in customizable projects and folders, whilst content can be linked to any part of a document to generate automatic references. A user can generate original notes on any passage of text instantly with AI, highlight over a chosen passage of text and with a single click, and Genei will provide a concise breakdown in a clear, note-like form.
- **WRITER AI** – Writer was founded with the goal of helping companies with their marketing materials by transforming the static style guides and templates put together by marketing teams into an interactive assistant. This “single source of truth” allows the teams to create, define, and edit the specific terms and concepts used by the company in its written material. They can be turned into what the company called “snippets,” which act as a shortcut for common definitions, links or other common material. The Writer AI looks for standard spelling and grammar mistakes as a baseline but can spot some less common mistakes like plagiarism and suggest ways to adjust tone, style, and even sentence complexity to match the writer’s goals. The AI operates as a plug-in for most common writing software such as Microsoft Word, Google Chrome, and Figma. Although not a voice assistant, the natural language processing (NLP) at the center of the Writer’s tool is rapidly taking over in many enterprise services for both writing and speech.

Figure 79: Key Differentiators of Writer

**DIFFERENTIATORS**

## Writer is enterprise AI

**Adds intelligence**

Writer is the only enterprise AI platform that builds off of LLM "clay," and adds layers of intelligence that include great writing, business context, your content, and your company facts. And we're the only one that detect and fact-check output for you.

**Supports your brand**

Only Writer is trained on your brand, including your best content, voice, style guide, key terms, and core messages. This means AI-generated content will sound like you, every time.

**Understands every format**

Only Writer can ingest and make sense of content from text, videos, PDFs, audio, spreadsheets, and databases. We can even perform live Internet crawls to generate, research, repurpose, analyze, and transform data for new contexts.

**Extends your ecosystem**

Only Writer embeds AI directly into your team's workflows and tool ecosystem. We give you an API and an out-of-the-box ecosystem of extensions so you can deliver AI directly into +100 third-party applications.

**Keeps your data private**

Only Writer keeps your data private. Unlike other AI tools that reserve the right to retain and use your data, with Writer, you own your data and we will never use it in our foundation model. You also get an API to your own, fine-tuned, private, and secure LLM, to use directly in any of your internal applications.

**Is business-ready**

Only Writer supports deployments across multiple teams, with robust reporting, SCIM provisioning, user and team management, and audit logs. Plus, we're compliant with SOC 2 Type 2, PCI-DSS, and HIPAA.

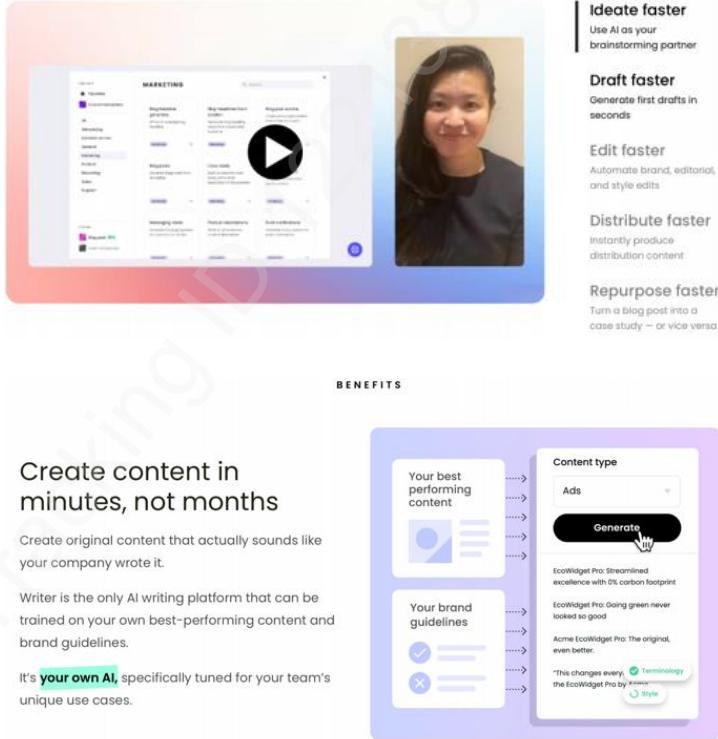
Source: Company Website.

- **Murf.ai** – Murf offers a selection of 100% natural sounding AI voices in 20 languages to provide professional voice overs for videos and presentations. It can also create voice overs with customization features like Pitch, to draw a listener's attention to words or phrases, Pauses, to add pauses of varying lengths to narration. A user can also record their voice from anywhere, upload it to Murf and swap voices with a professional sounding AI voice.

Figure 80: Writer Helps to Accelerate the Writing Workflow, from Idea Generation All the Way to Distribution

Automate what's automatable in content

Writer accelerates your writing workflow, from ideation → drafting → distribution



The infographic illustrates the Writer writing workflow across four main stages:

- Ideate faster:** Use AI as your brainstorming partner.
- Draft faster:** Generate first drafts in seconds.
- Edit faster:** Automate brand, editorial, and style edits.
- Distribute faster:** Instantly produce distribution content.
- Repurpose faster:** Turn a blog post into a case study — or vice versa.

**BENEFITS**

**Create content in minutes, not months**

Create original content that actually sounds like your company wrote it.

Writer is the only AI writing platform that can be trained on your own best-performing content and brand guidelines.

It's **your own AI**, specifically tuned for your team's unique use cases.

**ON-BRAND LANGUAGE SCORE 100**

Build a brand that is consistent — everywhere

Generic ad copy. Outdated messaging on your website. Poorly-written docs.

All of this can kill a buyer's journey — and your brand. Writer gives you all the tools you need to **build a consistent brand** across every customer touchpoint, and do it efficiently.

Source: Company Website.

- **Glasp** – A social web highlighter that people can use to highlight and organize quotes and ideas from the web without switching back and forth between screens. Glasp provides a free browser plugin (available for Chrome, Safari, Brave, Edge, Opera, and more) that can be exported to .txt, .html, and/or.csv files. These highlights can then be tagged, searched for, linked to, and shared on a variety of other platforms, including Twitter, Teams, and Slack. Also includes extension for ChatGPT to create summaries of Youtube videos.
- **Midjourney AI** – Produces an artificial intelligence program that creates images from textual descriptions, similar to OpenAI's DALL-E and Stable Diffusion, accessed entirely through Discord. Midjourney is currently only accessible through a Discord bot on their official Discord, by direct messaging the bot, or by inviting the bot to a third-party server. To generate images, users use the /imagine command and type in a prompt; the bot then returns an image. Midjourney is also working on a web interface. Users create artwork with Midjourney using Discord bot commands.
- **Leiapix Converter** – A free image processing tool that enables instant conversion of 2D images into beautiful 3D Lightfield images. Export conversions to Leia Image Format, Side-By-Side 3D, Depth Maps, or Lightfield Animations. Their AI model reads the image for depth layers and automatically calculates the best approximation of 3D for the image. Then, it adds a subtle animation to show off the effect. Through their web interface, a user can pinpoint specific parts of an image and adjust the depth map. The user can then download their creation as an MP4 or animated GIF.
- **Watermark Remover io** – Removes any watermark from any video or image, smoothly getting rid of the translucent watermarks. Their model first predicts the location of the watermark in the image, then segregates the colors of watermark from background of the image, then reconstructs the background image in the area where watermark was present.
- **Supermeme.ai** – An AI-powered meme generator aimed used as a one-stop shop for everything meme marketing for brands and individuals. A user can search for memes naturally using emotions or actions and find relevant meme templates using their advanced meme search.
- **ClipDrop** – Removes objects, people, text, and defects from pics automatically, also can relight photos and drawings, and denoise and enhance images.
- **Lalal.ai** – An AI tool and an online service providing effortless noise reduction and background music removal. It can extract vocals, accompaniment and various instruments from any audio and video without quality loss. The unique algorithm cancels out unwanted sounds, producing tracks with crystal clear voice and removes background music from recorded video streams to prevent copyright claims and avoid legal issues.
- **Character AI** – Character AI is based on neural language models and has been trained from the ground up with conversations in mind. However, what sets Character apart is that users get to choose from various personalities instead of interacting with a single AI chatbot, that can generate human-like text responses and participate in the contextual conversation. This new AI-powered chatbot is more than just an alternative to ChatGPT though, it lets a user live chat with characters in real time, such as celebrities, historical figures, and fictional characters. A user can create "characters," craft their "personalities," set specific parameters, and then publish them to the community for others to chat with them.
- **YouChat** – Powered by AI and NLP, YouChat by YOU.com is an AI that can answer general questions, explain things, suggest ideas, translate, summarize text, compose emails, and write code. Powered by artificial intelligence and natural language processing, it allows a user to have human-like conversations. YouChat 2.0 is the first web search that combines advanced conversational AI with community-built apps, offering a unique and interactive experience with each query. With its blended large language model known as C-A-L (Chat, Apps and Links), YouChat 2.0 can provide charts, images, videos, tables, graphs, text or code embedded in its responses to user queries.

- **DialoGPT** – A Microsoft project that leverages massive amounts of publicly available colloquial text data, specifically 147 million multi-turn dialogues from Reddit, DialoGPT establishes a foundation for building versatile open-domain chatbots that can deliver engaging and natural conversational responses across a variety of conversational topics, tasks, and information requests, without resorting to heavy hand-crafting.
- **DeepL Write** – An AI writing tool that improves written communication in both English and German, tackling more than just grammar by offering suggestions on phrasing, tone, style, and word choice. Write is especially useful for professionals such as journalists, writers, or academics looking to boost their creativity through sentence, word, and rephrase suggestions. Multilingual teams in global companies can also use Write to draft clear and concise emails, campaign proposals, reports, and presentations for international clients.
- **Writier.io** – Writier is an online writing tool that helps create amazing content in seconds with the power of AI-generated sentence completions. The app makes it easy to write long-form articles, blog posts, and web content without ever having to worry about writer's block. In addition, the app also provides a wide range of features that helps writers manage, track and organize their content in 12 different languages.
- **Twain** – Twain is an AI communication assistant for marketing and sales outreach, designed to increase the conversion rate of outreach messages. After pasting outreach messages into Twain's editor, the company's software lets sales and marketing teams get recommendations based on outreach best practices thereby offering simple-to-understand recommendations, enabling clients to write clear, convincing, and confident outreach messages that get answers.
- **Marmof** – Marmof is an AI-powered content generation platform that enables users to create high-quality content quickly and easily for their websites, blogs, and other digital media. Its powerful artificial intelligence capabilities allow users to generate content quickly and cost-effectively without needing manual labor. The platform also offers an intuitive interface and customizable features, making it easy to create content tailored to the business's needs.
- **Easy-Peasy.AI** – An AI Content tool that assists with variety of writing tasks, from writing blog posts, creating resumes and job descriptions, to composing emails and social media content using 90+ templates. With AI powered audio transcription, a user can quickly and accurately transcribe audio content, generate episode titles, summary descriptions, and show notes.
- **Regie.ai** – Regie leverages OpenAI's GPT-3 generative text engine to produce marketing and other business content. Regie is designed to produce sales and marketing copy for brands to employ more quickly and yet with more consistency and personalization than doing so manually. The startup adapts the GPT-3 model specifically for businesses to use in email campaigns, online ads, text messages, social media, and even podcast ad copy. Clients can integrate Regie into their own content management systems or use it a browser extension.
- **Compose.ai:** Once the browser extension is activated, Compose.ai's assistant begins offering suggestions for finishing sentences and paragraphs in any text space, from feedback forms to search engines. Compose.ai's long-term vision involved getting its service used in corporate contexts. The AI could add a company's style to the facets of writing it learns. The result would be a helpful writing assistant that can sound like individual employees within the context of how they communicate on a professional level.

## Industries Set to Benefit from ChatGPT

Based on the extensive amount of research we have already done on ChatGPT and LLMs, it is clear to us that there are industries set to benefit today with others eventual beneficiaries as GPT-4 comes into the picture. That said, there are some industries that will simply not benefit in the short or even medium term, which is an important point considering where we are in the AI application cycle. The opportunities to improve productivity in the Information Technology, Education, Government, and Business Services industries seem to us more clear than in other industries. In other industries such as Healthcare, Materials, and Industrial industries, ChatGPT may not be a needle mover today, or even in the next 12 months, but AI and LLMs will eventually have an impact. In this section we discuss various industries and cover the following points by industry.

- **Can ChatGPT in its current form (based on GPT-3) add value or increase productivity to each industry today?**
- **If ChatGPT in its current form cannot be leveraged by industry constituents, are there broader AI implications to the respective industry that should be discussed at a high level?**
- **What are sector coverage implications within each industry as well as early identifiable AI adopters (companies) poised to benefit from either ChatGPT or AI?**

This last section includes contributions from a wide range of sector coverage teams within the Credit Suisse equity research department, providing a cross-sector view on where investors should be looking to identify potential leaders as AI gains further traction in revenue generating, operational, and fundamental aspects of companies.

First, we cover some key engineering aspects, asking the key question of our report "which industries can benefit from ChatGPT" and then we dive into extensive and specific examples by industry.

## Four Different Ways to Use ChatGPT

There are four different ways a user can use ChatGPT (based on OpenAI's GPT-3) as it stands today:

- **As-is:** Inputting prompts and receiving results via the web-based interface. This is by far the most popular usage approach today and as discussed earlier, there is a premium version of this offering for \$20/month via OpenAI's website.
- **Prompt engineering without APIs:** Prompt engineering is the use of a service like ChatGPT in conjunction with other technologies as part of a workflow. This workflow can be achieved manually or by using screen scrape and robotic process automation (RPA) technologies.
- **Prompt engineering using APIs:** This model is not yet available, but expected in 1H23. While there are currently solutions on Github that enable an API wrapper around ChatGPT, they are not recommended for production builds or scale, and they are not supported by OpenAI.
- **Custom Build:** It is possible to create a custom build of the core GPT2/GPT3 model for a bespoke implementation, but this would not have the conversational interaction or prompt filtering provided by ChatGPT. This is how Microsoft's Bing AI was developed via Microsoft and OpenAI partnership by leveraging OpenAI's GPT3 LLM.

As it pertains to our report, we expect the "As-is" technology to be the focus of the industry sections as we identify the early industry beneficiaries. In earlier parts of this report, we discuss Microsoft's playbook with OpenAI's technology which will likely have much greater implications on each discussed industry, an important point of context considering how early we are with OpenAI's technology and LLM usage/development.

## Information Technology

The information technology industry includes the software, internet, IT hardware, networking equipment, and semiconductors sectors. Given ChatGPT is a virtual platform, the two sectors likely to see the most use cases include the software and internet sectors, but other sectors within IT can see benefits as well. For instance, software programmers across all IT sectors can check their software code for maintaining the network, installing patch updates, and/or adding new functions to the network to verify commands and processes (Figure 83). The same use case would apply to the IT hardware, networking equipment, and semiconductor sectors that include software programming for servers, storage arrays, network switches, routers, and chip drivers. Below we discuss the direct AI use cases from ChatGPT (GPT-3) for the IT industry and coverage implications that we have gathered from industry references, experts, and coverage company mgmt. teams.

**Figure 81: ChatGPT Providing Correction to Written Code**



What's wrong with this code?:

```
def reallyBadcode():
    someBool = True
    while someBool:
        print(someBool)
```

There are a few issues with this code:

1. The `while` loop does not have any way to terminate, so it will continue to print `someBool` indefinitely. This is an example of an infinite loop, which can cause the program to crash or become unresponsive.
2. The code does not provide any way for the user to control the value of `someBool`, so the loop will always be entered and will always print the same value. This means that the code does not provide any useful functionality or allow for user input.
3. The code does not include any comments or documentation to explain what it does or how to use it. This makes it difficult for other people to understand or use the code.

Here is an example of how you could fix these issues:

```
def betterCode():
    # Ask the user for input
    someBool = input("Enter a value for someBool (True or False): ")

    # Convert the input string to a boolean
    someBool = someBool.lower() == "true"

    # Print the value of someBool
    print(someBool)

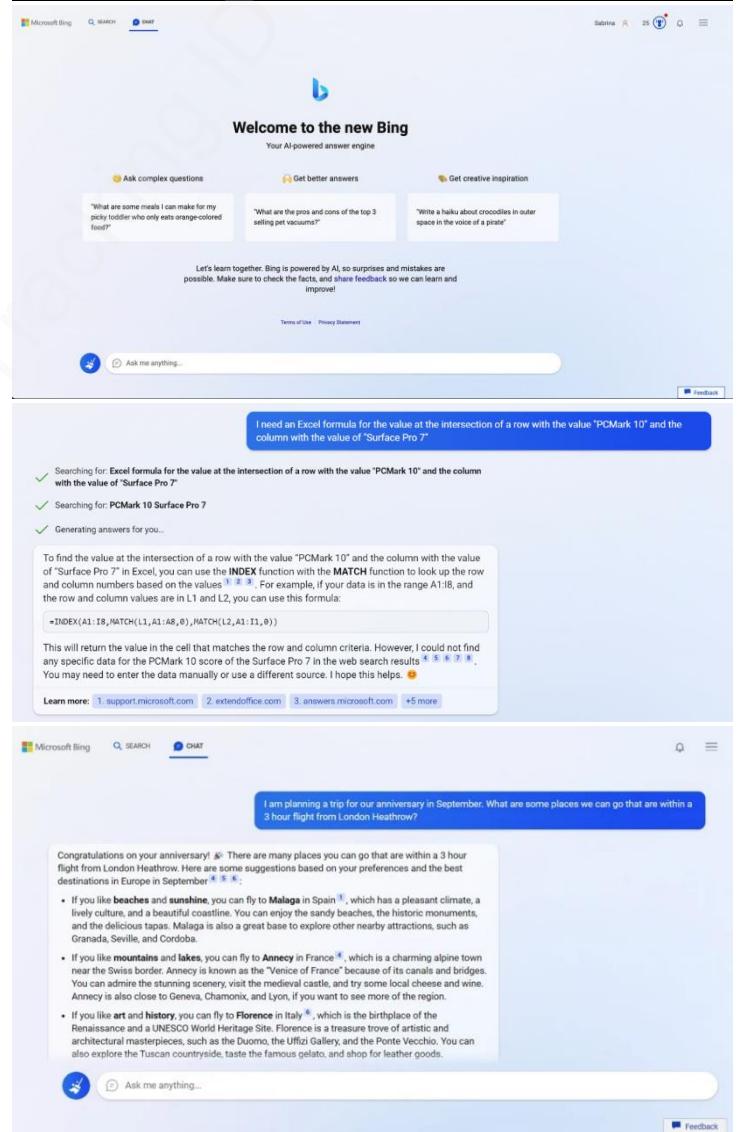
    # If someBool is True, enter the while loop
    while someBool:
        # Ask the user if they want to continue
        continueLoop = input("Do you want to continue? (y/n): ")

        # If the user does not want to continue, set someBool to False
        # to exit the loop
        if continueLoop.lower() == "n":
            someBool = False
```

This code fixes the issues with the original code by providing a way for the user to control the value of `someBool` and exit the loop, and by including comments to explain what the code does.

Source: ChatGPT, Medium ([Richard Gao Posts](#)).

**Figure 82: Bing AI Answering Inquiries via Search Prompt Line**



Welcome to the new Bing  
Your AI-powered answer engine

Ask complex questions | Get better answers | Get creative inspiration

"What are some meals I can make for my picky toddler who only eats orange-colored food?" | "What are the pros and cons of the top 3 selling pet vacuums?" | "Write a haiku about crocodiles in outer space in the voice of a pirate"

Let's team together. Bing is powered by AI so surprises and mistakes are possible. Make sure to check the facts, and share feedback so we can learn and improve!

Ask me anything... | Terms of Use | Privacy Statement | Feedback

I need an Excel formula for the value at the intersection of a row with the value "PCMark 10" and the column with the value of "Surface Pro 7"

✓ Searching for: Excel formula for the value at the intersection of a row with the value "PCMark 10" and the column with the value of "Surface Pro 7"
 ✓ Searching for: PCMark 10 Surface Pro 7
 ✓ Generating answers for you...

To find the value at the intersection of a row with the value "PCMark 10" and the column with the value of "Surface Pro 7" in Excel, you can use the INDEX function with the MATCH function to look up the row and column numbers based on the values [1](#) [2](#) [3](#). For example, if your data is in the range A1:I8, and the row and column values are in I1 and L2, you can use this formula:  
`=INDEX(A1:I8,MATCH(L1,A1:I8,0),MATCH(L2,A1:I1,0))`

This will return the value in the cell that matches the row and column criteria. However, I could not find any specific data for the PCMark 10 score of the Surface Pro 7 in the web search results [4](#) [5](#) [6](#) [7](#). You may need to enter the data manually or use a different source. I hope this helps! [8](#)

Learn more: [1 support.microsoft.com](#) [2 extendoffice.com](#) [3 answers.microsoft.com](#) +5 more

I am planning a trip for our anniversary in September. What are some places we can go that are within a 3 hour flight from London Heathrow?

Congratulations on your anniversary! [9](#) There are many places you can go that are within a 3 hour flight from London Heathrow. Here are some suggestions based on your preferences and the best destinations in Europe in September [10](#) [11](#):

- If you like beaches and sunshine, you can fly to [Malaga](#) in Spain [12](#), which has a pleasant climate, a lively culture, and a beautiful coastline. You can enjoy the sandy beaches, the historic monuments, and the delicious tapas. Malaga is also a great base to explore other nearby attractions, such as Granada, Seville and Cordoba.
- If you like mountains and lakes, you can fly to [Annecy](#) in France [13](#), which is a charming alpine town near the Swiss border. Annecy is known as the "Venice of France" because of its canals and bridges. You can admire the stunning scenery, visit the medieval castle, and try some local cheese and wine. Annecy is also close to Geneva, Chamonix, and Lyon, if you want to see more of the region.
- If you like art and history, you can fly to [Florence](#) in Italy [14](#), which is the birthplace of the Renaissance and a UNESCO World Heritage Site. Florence is a treasure trove of artistic and architectural masterpieces, such as the Duomo, the Uffizi Gallery, and the Ponte Vecchio. You can also explore the Tuscan countryside, taste the famous gelato, and shop for leather goods.

Ask me anything... | Terms of Use | Privacy Statement | Feedback

Source: BingAI, Engadget ([Mat Smith Posts](#)).

## Key ChatGPT Use Cases in IT

We view the following as current use cases of OpenAI's (and Microsoft's) new tool:

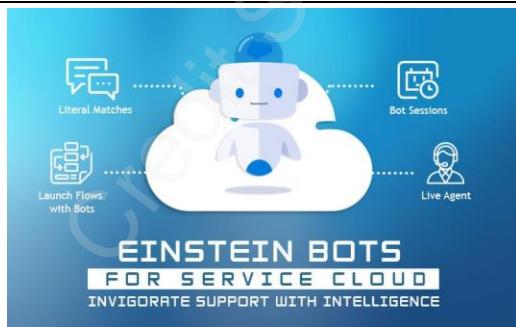
- **Software Development and Programming (Figure 81):** Write, debug, and explain code in various languages. Can also be used to converse in various programming languages and assist in virtually every aspect of software programming. Prior to the development of ChatGPT, Microsoft's Github Copilot sported similar programming capabilities as ChatGPT.
- **Search Engines (Figure 82):** ChatGPT has been incorporated into Bing, creating BingAI to understand and respond to language inputs and answer questions (i.e., Search tools).
- **Data Structuring:** Generate SQL queries to turn unstructured data into structured data. Can also use ChatGPT to extract data from text.
- **Chatbots (Figure 82, Figure 83):** Companies can integrate ChatGPT into their chatbots for customer service or into employee F.A.Q. pages. The options here are applicable to a wide range of applications.

## AI Use Cases Beyond ChatGPT in the IT Industry

When thinking more broadly about the implications of AI to the IT sector, we identify the following as key use cases of AI in the IT industry. Note, some of these use cases will include the training and development of new large language models (LLMs):

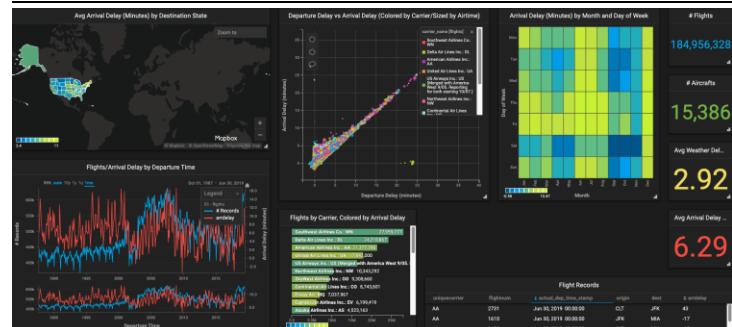
- **Robot Driven Application Development:** Application development with robots to interact with customers (chatbots) or to automate tasks (RPA and workflow automation). This would entail writing software and teaching robots to perform physical tasks with large language models.
- **Client Resource Management Enhancement (Figure 83):** Salesforce is developing EinsteinGPT, a suite of AI-powered features designed to enhance customer engagement and business process automation.
- **Analytics:** Optimize predictive analytics with AI that can provide intelligence before it is needed to make critical or operational decisions, among other conclusions.
- **Autonomous Systems:** Create autonomous systems that can perform tasks without human intervention (self-driving cars/drones, kitchen robots, etc.). This same AI use case can be applied to manufacturing, inventory shipments, etc.
- **Graphic Creation:** Create graphics and images based on user inputted prompts. Same images can be used for websites, product designs, etc.

**Figure 83: EinsteinGPT Is Salesforce's Bot Tool and AI**



Source: Company data ([link](#)).

**Figure 84: AI Analytics and Dashboards Likely to Gain Traction**



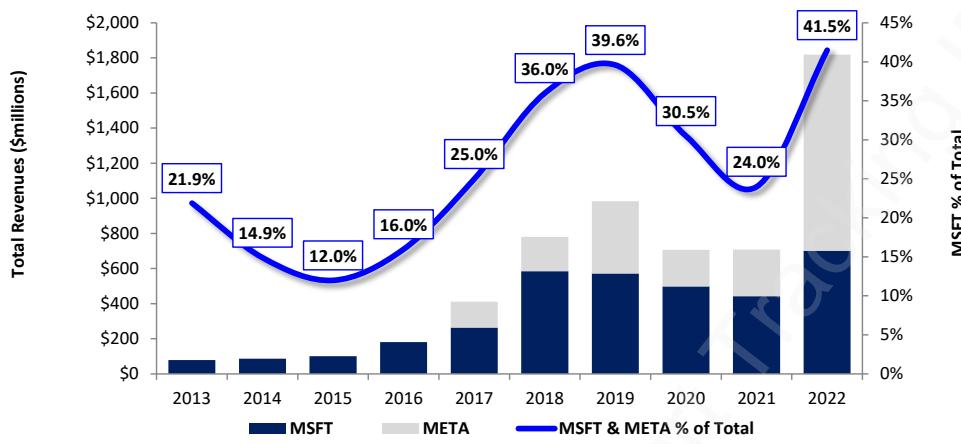
Source: [Heavy.AI](#).

## IT Sector Coverage Implications

**Software (Large Cap) and Networking Sectors – Analyst: Sami Badri**

Beyond our lengthy talking points on **Microsoft**, we see **Arista Networks** as a potential beneficiary from ChatGPT and AI/ML more broadly as the company's switches offer (1) industry leading power efficiency, (2) high-bandwidth, low-latency, lossless transmission, and scalability, enabling AI/ML workloads, and (3) a proprietary OS software preferred by large hyperscalers. In 2022, META and MSFT combined to account for ~41.5% of ANET's overall revenues, the highest percentage on record, a trend that is likely to continue (see Figure 85). Similarly, we see **Cisco Systems** as being indexed to AI/ML as the company stands to benefit from the same market growth in hyperscale data center switches as ANET (albeit to a lesser extent), increased demand from enterprise customers that are building AI applications on networks and training AI algorithms, and an increase in the network performance required to run these LLM and AI applications, which Chuck Robbins, CEO of Cisco, predicts will be 3-4x larger than the existing network's compute resources today.

**Figure 85: MSFT and META Continue to Rely on ANET Products, a Dynamic That Should Continue Given the Required Investment to Run ChatGPT/AI/ML**



Source: Company data, Credit Suisse estimates

Additionally, we believe **Equinix** is likely to be a beneficiary of ChatGPT and the growing AI/ML landscape. While AI's growth likely implies accelerated workload growth, a boon for all data centers, Equinix should benefit to a greater extent given it offers the leading global interconnection ecosystem. As enterprises adopt ChatGPT and future AI, we believe they are likely to see higher levels of intra-data center communication as a means to improve performance for high growth applications; thus, Equinix could fortify its already premium interconnection-rich brand.

**Figure 86: Equinix Generates 17% of Revenues from Interconnection, but GPT and Future AI Could Spur Even Greater Interconnection Growth/Mix**

Interconnection Revenues (\$ millions)		2018	2019	1Q 2020	2Q 2020	3Q 2020	4Q 2020	1Q 2021	2Q 2021	3Q 2021	4Q 2021	1Q 2022	2Q 2022	3Q 2022	4Q 2022E	1Q 2023E	2Q 2023E	3Q 2023E	4Q 2023E
EQUIX-Global	\$802	\$894	\$242	\$249	\$261	\$271	\$280	\$288	\$291	\$303	\$309	\$315	\$318	\$326	\$331	\$337	\$350	\$361	
y/y growth	17.7%	11.4%	13.8%	13.9%	15.1%	15.2%	15.5%	15.4%	11.6%	11.9%	10.6%	9.5%	9.3%	7.6%	7.2%	7.1%	9.8%	10.7%	
Interconnection Mix	15.8%	16.1%	16.8%	17.0%	17.2%	17.3%	17.5%	17.4%	17.8%	17.8%	17.3%	17.3%	17.2%	16.9%	16.9%	16.9%	16.8%	16.8%	
EQUIX-Americas	\$532	\$577	\$151	\$153	\$157	\$161	\$165	\$168	\$169	\$170	\$181	\$187	\$190	\$195	\$197	\$200	\$209	\$217	
y/y growth	13.4%	8.4%	8.9%	7.7%	7.2%	7.9%	9.2%	9.3%	7.6%	10.1%	9.8%	11.9%	12.9%	9.5%	8.9%	6.8%	9.9%	11.4%	
Interconnection Mix	21.4%	22.3%	22.8%	23.2%	23.3%	22.3%	22.7%	22.3%	22.1%	22.7%	22.6%	22.6%	22.4%	22.4%	22.7%	22.7%	22.7%	22.8%	
EQUIX-EMEA	\$139	\$162	\$49	\$51	\$56	\$58	\$62	\$65	\$66	\$67	\$68	\$67	\$67	\$68	\$70	\$72	\$73	\$75	
y/y growth	32.4%	16.3%	29.4%	31.8%	35.6%	31.6%	27.0%	28.2%	18.1%	14.5%	10.5%	2.4%	1.4%	2.0%	2.7%	7.3%	9.9%	10.1%	
Interconnection Mix	8.9%	8.9%	8.8%	8.5%	9.4%	11.4%	11.2%	10.9%	11.1%	12.1%	12.4%	11.2%	11.3%	11.3%	10.9%	10.9%	10.8%	10.7%	
EQUIX-APAC	\$131	\$137	\$43	\$45	\$49	\$51	\$53	\$55	\$57	\$58	\$60	\$61	\$61	\$63	\$64	\$65	\$67	\$69	
y/y growth	22.3%	4.5%	16.3%	18.9%	23.0%	24.0%	24.6%	21.2%	16.9%	14.4%	12.8%	10.8%	7.9%	8.1%	7.0%	7.6%	9.7%	9.3%	
Interconnection Mix	12.8%	11.7%	11.1%	11.7%	12.1%	15.0%	13.9%	14.2%	14.1%	15.7%	15.6%	15.7%	15.2%	15.0%	14.4%	14.3%	14.2%	14.0%	
DLR	\$250	\$263	\$70	\$85	\$86	\$86	\$89	\$91	\$91	\$90	\$94	\$93	\$95	\$98	\$99	\$100	\$101	\$102	
y/y growth	6.0%	5.4%	2.4%	33.0%	31.3%	31.8%	27.5%	6.0%	6.1%	4.0%	5.0%	3.1%	4.9%	8.7%	5.5%	6.8%	5.4%	4.1%	
Interconnection Mix	8.2%	8.2%	8.5%	8.6%	8.4%	8.1%	8.2%	8.3%	8.0%	8.1%	8.3%	8.2%	8.0%	8.1%	8.2%	8.1%	7.9%	7.6%	

Source: Company data, Credit Suisse estimates

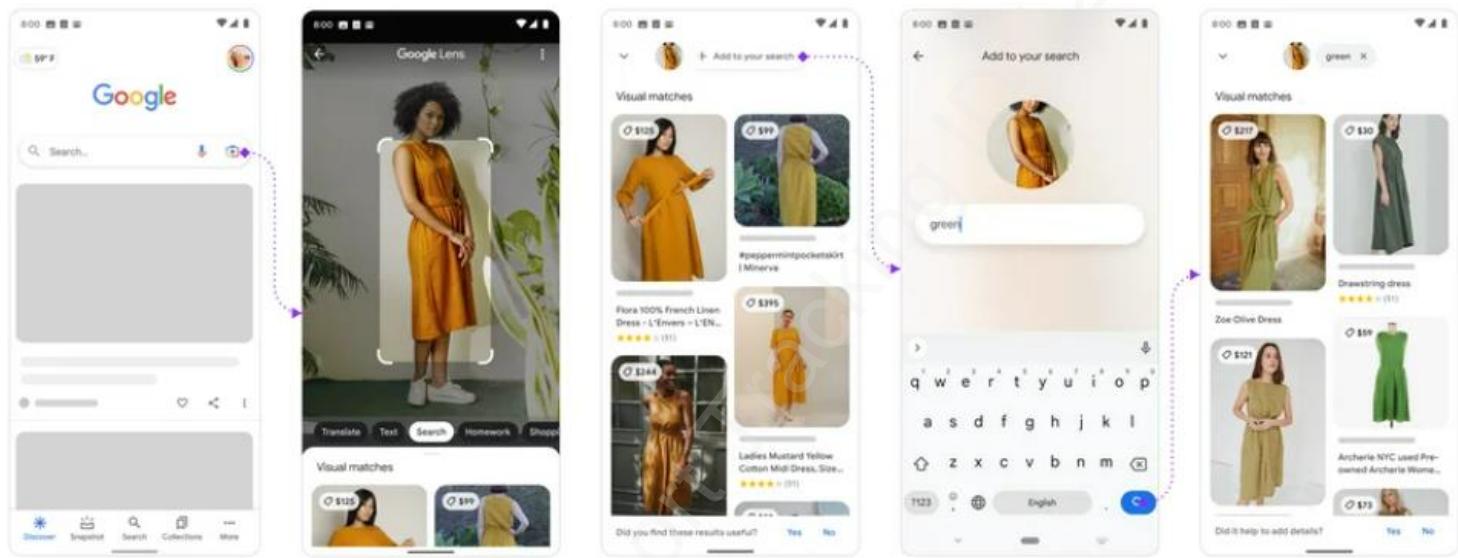
#### Internet Sector – Analyst: Stephen Ju

Examples of AI usage across the Internet sector are already plentiful as most companies have already been investing – particularly in the case of Google – for some time to primarily drive improvements in their consumer-facing as well as advertiser-facing products.

While these are by no means a comprehensive list of the different ways AI has come to be infused in products, we call out some of the more prominent and latest examples below – these are all designed to drive incremental usage and therefore share gains, or improved consumer/advertiser experience leading to the improvement of conversion rates (for e-commerce properties) or automation.

Google recently [announced Mutisearch](#), which expands the input for search beyond the traditional text to also incorporate in this instance below images (as well as voice and camera), to help find products with the same attributes.

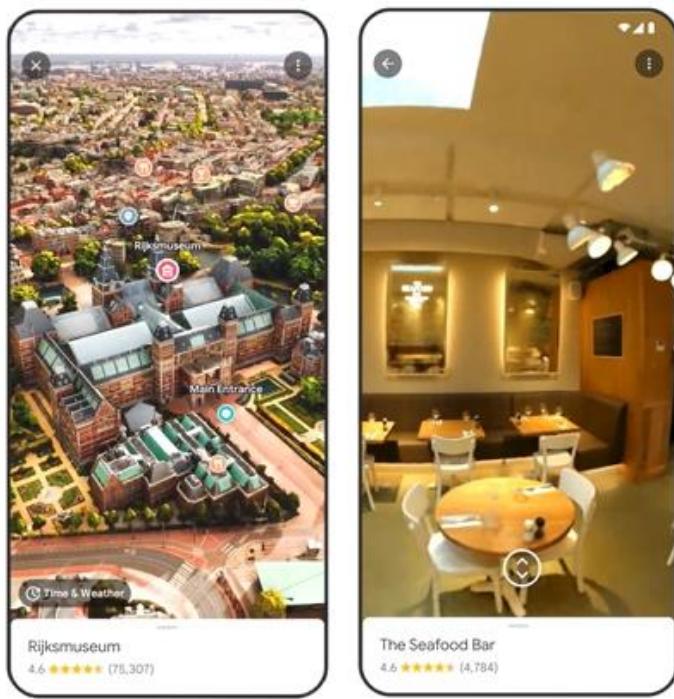
**Figure 87: Alphabet Inc. – Image Search Example within Multisearch**



Source: Company data

We note that this product is a culmination of the machine learning models for image recognition Google has built and refined over time. Another consumer-facing example for Google shows up in Maps and Immersive View, which takes ordinary 2D pictures and generates and visualizes 3D representations of the space to give the user incremental information (in this case the Rijksmuseum) and/or a view of the interior (in this case, an image of a restaurant's interior).

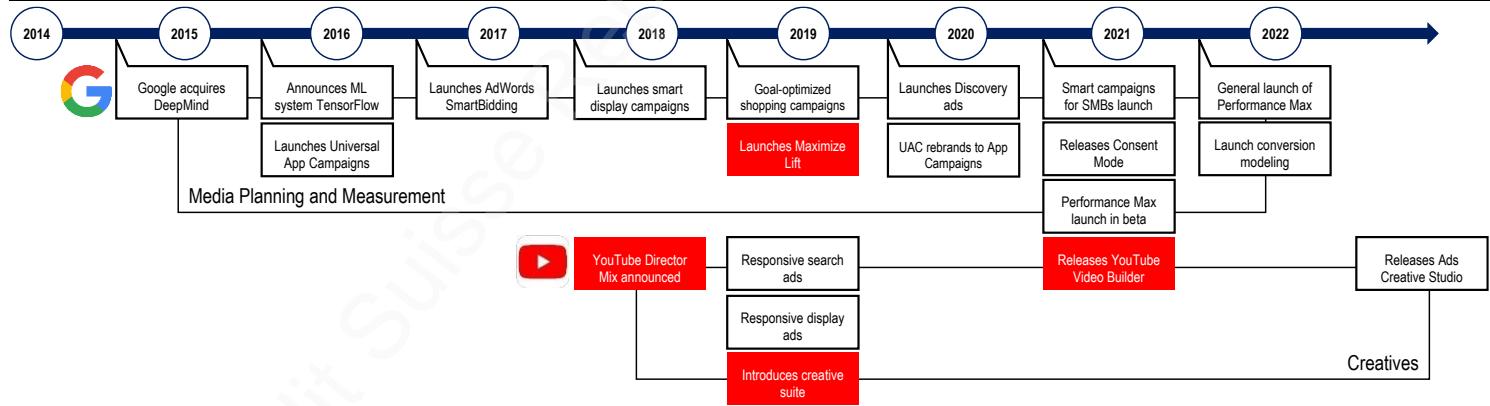
**Figure 88: Alphabet Inc. – Immersive View Built on AI-based 3D Visualization Tools**



Source: Company data

And as we have highlighted earlier as part of our [2023 Themes](#), AI-based tools help advertisers not only in media planning and measurement but also in the generation of creatives.

**Figure 89: Alphabet Inc. – Evolution of Advertising Automation Starting with Search and Expanding to Other Surfaces**



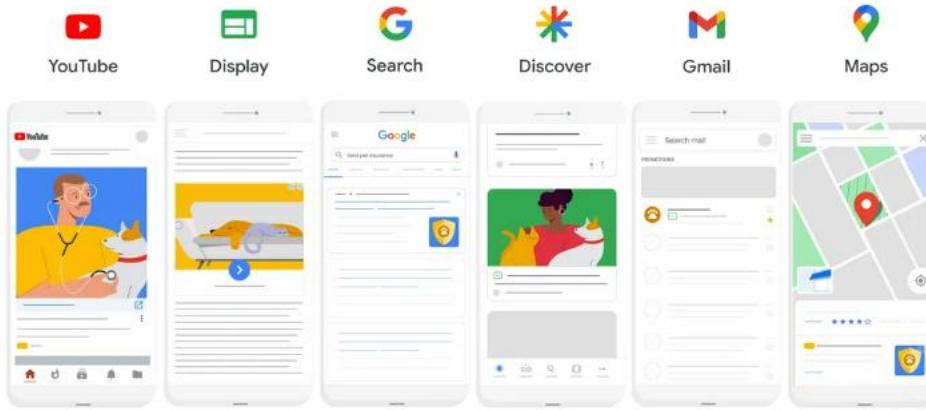
Source: Company data, Credit Suisse

These tools have focused primarily on automation, with the primary motivation to democratize online advertising to marketers of all sizes. For larger advertisers with access to agency-supplied bidding tools, deploying budgets on Search advertising was never going to be a problem, but for smaller or local advertisers, navigating the complexity of managing bid prices across what could be thousands (if not hundreds of thousands) of keywords was going to be a prohibitive exercise.

With that in mind, the first product Google released in 2016 was SmartBidding, which optimized campaigns for goal-based results (whether target cost per action/CPAs, target return on ad spend/ROAS) versus prior optimizations on a keyword-by-keyword level. This expanded in 2018 to products such as Responsive Search/Display as well as Creative Suite to help marketers

more easily generate ad creatives. The most recent iteration was the 2021 release of PerformanceMax which helped advertisers spend across all of Search, YouTube, and more.

**Figure 90: Alphabet Inc. – Performance Max Helps to Automate Spend Across Multiple Google Surfaces**



Source: Company data

While it is undeniable that an automation tool like Performance Max has helped advertisers navigate the challenges from Apple's deprecation of IDFA, the aforementioned democratization of advertising potentially raises the number of marketers/clients from single-digit millions to potentially tens if not hundreds of millions (of businesses). In other words, automation through AI has unlocked the possibility to enable businesses which have otherwise not been able to advertise before:

In our view, TAM can ultimately expand beyond the traditional ~\$766 billion of media spend (industry forecast) or even the combined ~\$1.3 trillion of media and promotional budgets to what we believe to be roughly \$3.2 trillion as we add the enablement of SMBs over time. Hence, the aggregate penetration of industry dollars for online does not stand at ~60%+ and rather sub-15%, opening the way for the revenue growth for all of Alphabet, Meta, Amazon (as well as Snap, Pinterest, etc.) to be stronger for longer..

### **Software (Small and Mid-cap) Sector – Analyst: Fred Lee**

#### **Cloud Communications**

*Contact centers generate some of the highest volume of structured and semi-structured customer data in the enterprise.*

Over the last couple of years, open-source NLP models crossed a tipping point, where the computer's understanding is now often indiscernible from a human's understanding of the same conversation. As a result, a bleeding-edge AI-enabled contact center (CC AI) generates some of the highest volume of structured high-signal ratio customer data of any technology in the enterprise; *Potentially enabling transformative changes to customer service.*

This in theory should enable some of the largest transformative applications of AI/ML in the enterprise. Use cases already implemented today include:

- Call deflection to an **Intelligent Virtual Agent (IVA)**. Contact center **labor spend is 35x larger than software spend at over \$400Bn** (17M agents globally at an average \$25k salary). The majority of call center requests are "simple resolutions," which, in theory, could be offloaded to an IVA. Moreover, majority of customers prefer self-service for simple matters.
- Higher agent performance through **Agent Assist**, which improves agents' performance by providing contextualized real-time assistance including discovering intent, delivering contextual knowledge articles, providing personalized agent coaching and feedback, automating compliance, and auto note taking and summarization. The

cost of agent assist software today is roughly similar to the average labor cost for the agent, speaking to the value of the software.

- Better customer insights through **Contact Center (CC) Intelligence/Analytics**, leveraging the contact center as the key customer touchpoint to better understand your customer. Key use cases include **topic discovery**: discover and understand new unknown topics that customers are inquiring about; and **insights dashboards**: hot topics, customer sentiment, competition at scale.

In our recently hosted fireside chat, Anand Janefalkar (Founder and CEO of UJET) claimed that typically **about a quarter of contact center volumes comprised scenarios which could be addressed through self-service**.

*But CC AI transformation has lagged on high implementation complexity, low cloud penetration, and mixed customer outcomes*

In practice, enterprises with cloud contact centers offload only between 10% and 15% of calls to IVAs today. Adoption has been tempered by high implementation complexity and fraught with mixed customer satisfaction (CSAT) results. Moreover, only about ~20% of contact center seats are in the **cloud (CCaaS), which is a technological prerequisite for CC AI use cases**. Taken together, this suggests <3% penetration of IVAs in contact centers today.

*Generative AI is a step-function improvement in language fluency and upfront costs*

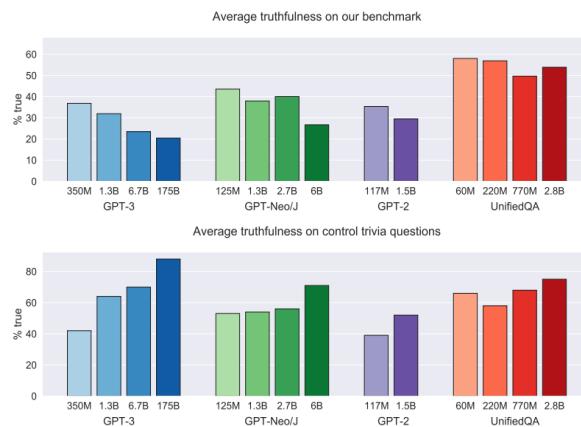
Natural language understanding (NLU) is the subfield of NLP which deals with semantic parsing of language data (developing an understanding of the context implied in a sentence). With such meaning representations encoded in the model, sentences can be generated by the model itself (generative AI). Contrast this to looking at the sentence as a time-series sequence, with every word being a prediction (predictive AI). As a result, generative AI models are a **step-function improvement in perceived natural language fluency and conversational context recognition**. These, in theory, should meaningfully increase CSAT scores for IVAs.

Moreover, a key technological leap in generative AI has been the wide adoption of the transformer model architecture (seminal 2017 [paper](#): "Attention is all you need") vs previously deployed variations of recurring architectures. This **new architecture is much less computationally intensive** compared to its predecessor, thus, making transfer learning viable for natural language data, creating, so-called, large language models (LLMs). Transfer learning entails pre-training a model on one dataset and later fine-tuning to another dataset for a particular use case. This was computationally intensive and a huge roadblock that previous architectures faced. OpenAI's generative pre-trained transformer 3 (GPT3) offering is a pre-trained NLU model which can be fine-tuned for multiple use cases. As a result, similar to how cloud infrastructure reduced the upfront spend (CapEx) associated with computing resources, **generative pre-trained transformers (GPT) significantly reduce the upfront costs needed to implement sophisticated AI-models for customer specific use cases**.

*We are in the early innings of LLM adoption in Contact centers; significant open issues remain*

However, we have not yet seen any meaningful implementations of GPTs for IVAs. Use cases so far have been **limited to augmenting human interactions** rather than supplanting them. Key examples of in-production GPT-based products include Agent Assist functionality (e.g., contextualized knowledge overlay solution at NICE, suggested responses feature at Cresta) and CC Intelligence (e.g., customer topic clustering at FIVN, smart meeting summary at ZM and FIVN).

This is the result of structural issues with the current generation of GPTs. The key issues are "hallucinates," i.e., seemingly coherent responses with factually incorrect data, while typically sounding confident when doing so. In current implementations, as the number of parameters the model is trained on increases, thereby increasing language fluency and contextual understanding, the truthfulness of the responses decreases. These issues are actively being worked on by foundational AI companies such as OpenAI.

**Figure 91: Larger Models Are Less Truthful**

**Figure 2: Larger models are less truthful.** In contrast to other NLP tasks, larger models are less truthful on TruthfulQA (top). Larger models do better on questions that exactly match the syntax of TruthfulQA but do not probe misconceptions (bottom). Figure 3 gives a concrete example of larger sizes being less truthful.

Source: Lin, S., Hilton, J. and Evans, O., 2021. Truthfulqa: Measuring how models mimic human falsehoods.

As well as other, more tractable problems. For example, foundational LLMs need to be trained on customer specific data to understand the businesses' unique internal processes and knowledge in a computationally efficient manner; and be integrated with key enterprise workflows. These are development initiatives that CCaaS vendors, such as FIVN and NICE are addressing, and will look to differentiate in.

#### Early moats are emerging – differ at the foundational AI vs AI-application layer

We expect dramatic growth in the AI market over the next decade as the issues highlighted above are improved and pricing comes down. At the same time, unlike the days of early cloud adoption, we are solidly in the API-era and foundational AI companies are building easily interchangeable API-based backends: “*Our experience to date with ChatGPT has been very good. The integration itself is not the difficult part. That's kind of the easy part*” (NICE, 4Q22 earnings call). As a result, switching costs at the foundational AI layer are not as pronounced as in early cloud infrastructure adoption period.

On the other hand, AI-application layer businesses are beginning to show defensible characteristics, including:

- **Workflow integration:** “*the age of AI and large language models has arrived and we want to empower smarter experiences and workflows that enable our customers to benefit from these transformational tools. By embedding AI into more workflows, we can provide our customers with richer, more actionable insights that empower them to work smarter and serve their customers better*” (ZM, 4Q23 earnings call);
- **Verticalization:** “*Zoom IQ for Sales was built in this collaborative fashion and has already added tremendous value to many sales teams. You can expect additional industry-specific and department-specific applications developed both by us and our third-party partners.*” (ZM, 4Q23 earnings call); and
- **Proprietary data:** “*we are using the subset of ChatGPT that [...] allows our customers to benefit from [...] the strength of ChatGPT, but at the same time, using our unmatched data repository to have a subset of well-trained ChatGPT*” (NICE, 4Q22 earnings call).

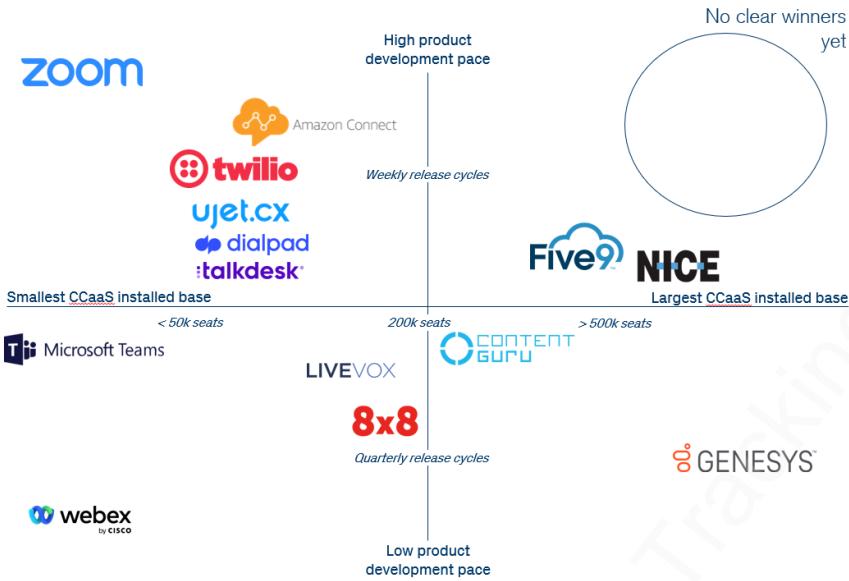
Key beneficiaries will be those with (i) large cloud contact center (CCaaS) installed bases and (ii) a rapid product development pace

We would highlight **NICE** and **FIVN** as having the 2<sup>nd</sup> and 3<sup>rd</sup> largest CCaaS installed bases, respectively; coupled with a good product development pace.

On the other side of the spectrum, would highlight **ZM**, as having an industry leading product development pace. Its relatively small installed base is a reflection of only having entered this market organically a year ago. The chart below also demonstrates how synergistic the proposed FIVN acquisition could have been for ZM's CC AI ambitions.

Finally, the chart also shows the lack of incumbents – providing an open-ended opportunity for several vendors with adjacent capabilities.

**Figure 92: CC AI Beneficiaries**



Source: Credit Suisse estimates

#### Software (Small and Mid-cap) Sector – Analyst: Rich Hilliker

**Intuit** at its 2022 Investor Day, shared it currently facilitates 730 million AI-driven customer interactions per year, has 2 million AI models in production, has more than 700 AI patents, and generates 58 billion machine learning predictions per day. Intuit leverages AI internally for product development and security, as well as throughout its entire portfolio. In our view, the power of its investment in this area shines in mobilizing its AI-driven expert platform: to power TurboTax Live, the company matches hundreds of attributes in order to determine which expert is best suited and available to help the specific customer in real-time. This not only drives productivity but improves customer experience. We expect Intuit will continue extending and enhancing this expert layer across its platform and believe generative AI and ChatGPT will play an important role in ensuring accuracy and automation. We also anticipate the company's payroll, time tracking, payment, the Credit Karma portfolio, and Mailchimp to benefit from improvements in AI.

**Customer Relationship and Engagement Management.** We strongly believe that CRM and front office enablement tools are no longer a system of record, but rather, are becoming centers of intelligent engagement. True, fully established artificial intelligence will transform how we interact and the tools we leverage to establish and build lucrative relationships. Intelligent engagement will require businesses to collect, analyze, and service their customers based on continuously updated aggregated data profiles (i.e., Customer 360, Unified Customer Records, Customer Data Profiles, etc.), enabling easily automated or algorithm enhanced customer-centric actions. As AI and LLMs advance further, we expect their ability to detect and understand intent as well as the determine buying patterns and predict business outcomes will improve, helping to support even greater efficiency in the front office. We're optimistic that mature, well-integrated AI, workflow, and CRM will help to transform data collection, processing, and value extraction across a growing number of channels. We expect AI will become a critical component of data integrity and validation, resulting in more effective customer intelligence

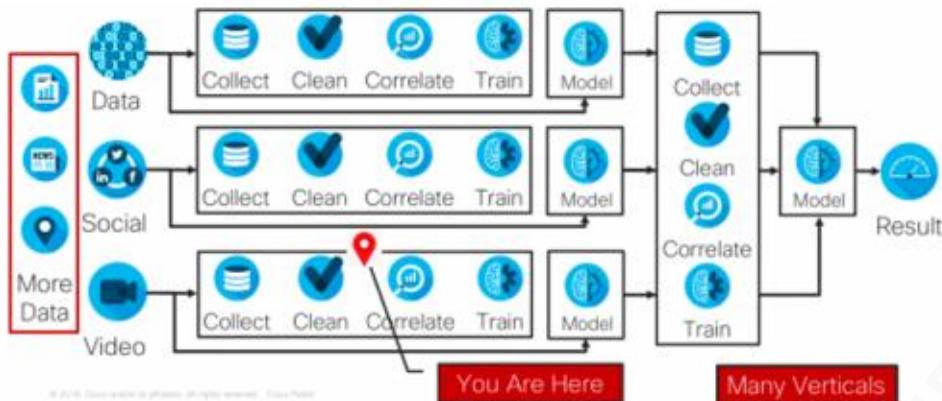
gathering, timely personalized intent-based marketing campaigns and outreach, lead nurturing, and more meaningful interactions across these channels. Similarly, we're optimistic regarding the ability of next generation chatbots, intelligent assistants, and adaptive content creation engines and the leverage they create. We believe **HubSpot**, **Intuit Mailchimp**, **ZoomInfo**, **Freshworks**, and **Zeta Global** all stand to benefit.

**Box.** We expect AI and LLMs to have a prominent role in the evolution of how businesses leverage unstructured data, particularly when coupled with the power machine learning and artificial intelligence. We believe Box's platform will benefit from advances in these areas as users interact with, create, search for, and share more content when leveraging solutions such as ChatGPT. We also believe this could drive interest in Box Skills, which is a framework for connecting to and applying intelligence platforms from companies such as IBM, Microsoft, and Google to content that lives in Box. While the initial skills framework brought image, audio, and video intelligence, we believe equally compelling value will come from the creation of custom and chained skills, enabling users to employ multiple intelligent technologies together. We think attaching AI Skills to Suite and/or Relay is a very ripe opportunity that could lead to incremental monetization of the company's highest value users. Finally, because content is constantly created, modified, stored, and shared across various digital platforms both internally and externally, proper security and governance is complex and time consuming for companies to manage themselves. As a result, we expect continued interest in Box Shield. Ultimately, we believe maturing AI and LLMs support greater intelligent automation, user understanding, and value extraction from content on Box's platform.

**SOPHiA GENETICS.** SOPHiA's core cloud product, the SOPHiA DDM platform, is one of the most complete genomics analysis platforms in the industry—disrupting two large and structurally growing markets, clinical (\$21 billion TAM) and BioPharma (\$14 billion TAM). The DDM platform leverages artificial intelligence to analyze and standardize health data in order to generate insights from complex multimodal data sets. In our view, the company is well positioned to benefit from advancements in AI as the success and accuracy of algorithms help draw actionable conclusions. We view generative AI such as ChatGPT as a helpful tool to interact with the findings and formulate potential care plans.

#### **IT Hardware Sector – Analyst: Shannon Cross**

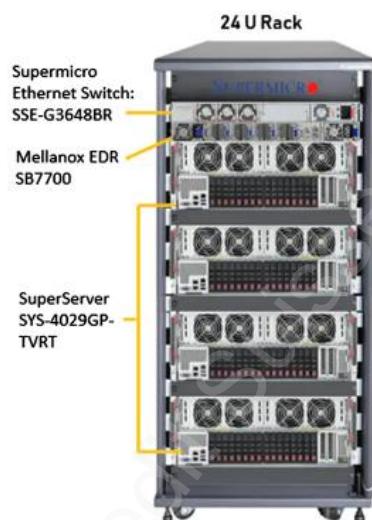
Similar to our initiation in [August 2022](#), we think hardware is a cornerstone of AI with compute and storage serving as the foundation for successful AI applications given the need to analyze and store vast amounts of data. McKinsey estimates AI applications will generate ~845 exabytes of data in 2025, up from only 80 exabytes in 2019. At the same time demand for compute hardware (e.g., servers) is estimated to grow at a 10% to 15% CAGR through 2025 driven by investments in hyperscale data centers and the edge. Data storage and memory will also benefit from increased requirements associated with AI workloads including demand for high-bandwidth DRAM needed to run AI algorithms and higher capacity drives for edge applications. For example, smart factories can generate up to 1 petabyte of data per day and autonomous vehicles can generate up to 32 terabytes of data per vehicle per day. In this section, we take a closer look at the enterprise hardware OEMs who are working with customers to accelerate AI adoption, by pursuing: (1) AI-optimized hardware; (2) advisory services to help customers implement AI processes into their businesses; (3) off-the-shelf solutions that target specific functions such as customer support or retail loss prevention; (4) cloud-based services for developing, training and deploying AI models; and, (5) data fabrics and management tools.

**Figure 93: Data Pipeline for Multiple Data Sources**
**DATA PIPELINE FOR MULTIPLE DATA SOURCES**


Source: Pure Storage

**Multiple Components Required**

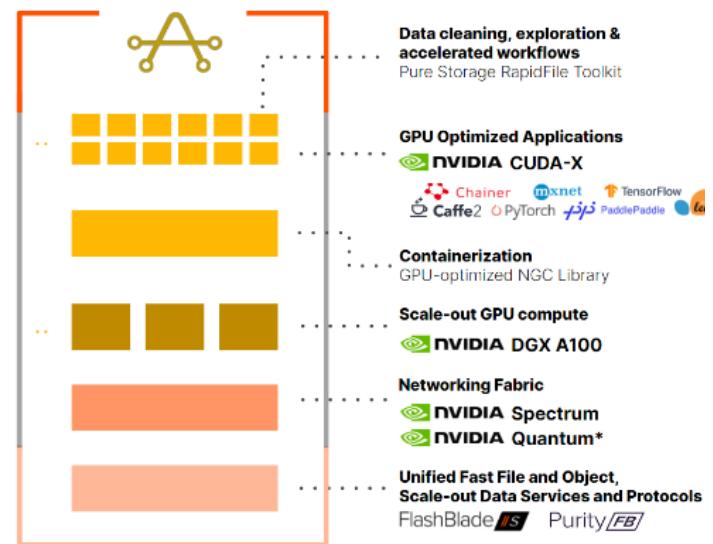
Enterprise hardware OEMs in our coverage (IBM, HPE, DELL, NTAP, PSTG) are partnering with Nvidia in some form to offer AI-optimized infrastructure, converged infrastructure and/or hyperconverged infrastructure. Shown below are two AI-specific configurations, a server rack from Supermicro and AI tech stack from Pure Storage, that both include Nvidia IP.

**Figure 94: Supermicro AI & Deep Learning Reference Architecture Configuration (High-End)**


Source: Supermicro

***First Comes Compute***

We expect Server vendors to benefit from an increase in AI applications and learning models which require richer specs (e.g., accelerated GPUs) and more DRAM. Higher content per server should benefit AUPs while Dell Technologies has frequently discussed the opportunity to expand server AUPs by selling into more sophisticated workloads including AI. HP Enterprise's server line-up scales up to large supercomputers (aided by the acquisitions of Cray and SGI), including the first petascale system that can complete a quadrillion operations per second. Supercomputers can be leveraged for training large language models or analyzing complex data

**Figure 95: Pure Storage AIRI//S (AI-Ready Infrastructure)**


Source: Pure Storage

sets like genome sequencing for cancer research; however, we will focus on enterprise AI solutions in this section of the report. IBM's most recent zSystem refresh (z16, began shipping in 2Q22) features on-chip inferencing (ideal for applications like detecting banking fraud) and Gen11 servers recently announced by HP Enterprise and Dell Technologies include better specs for AI workloads. Compared to Gen10, the new HPE ProLiant Gen11 servers support 2x more I/O bandwidth and 33% more high-performance GPU density per server and the new Dell PowerEdge servers use "Smart Cooling" to enable higher sustained performance for model training, core to edge inferencing, data visualization and more. Supermicro is an emerging competitor in the complete rack space, although the company has developed server motherboards since the 1990s. The company custom designs racks to customer specs from motherboard to ethernet, and typically caters to AI applications (customers include Meta and Intel).

### ***Then Comes Storage***

We think storage is the larger, higher margin opportunity for hardware in AI given the criticality of storing data throughout the process. This is already evident with traditional workloads as storage products generate a >50% gross margin versus servers at sub-30%. Pure Storage is a leader in flash storage, and most notably supports Meta's AI Research SuperCluster (currently building out phase one, but long-term goal is exabyte scale). Dell Technologies specifically advertises PowerScale all-flash storage for machine learning and deep learning and HP Enterprise is leveraging Cray for dedicated HPC and AI storage solutions. Pure Storage and NetApp are specifically selling AI-branded storage products with Nvidia integration and have partnered with server OEMs on converged hardware solutions. NetApp is working with Cisco, Lenovo and Fujitsu (handful of solutions) while Pure Storage is partnering with Cisco (FlashStack for AI). Vast Data is an emerging storage company (first shipments in 2018) that is also targeting AI workloads. Vast Data specializes in scale-out unstructured flash storage that uses a "DASE" (disaggregated and shared everything) architecture and QLC NAND to "enable AI applications to consolidate infrastructure and accelerate the training and inference time."

### ***Customer Example: Meta's AI Research SuperCluster***

Meta partnered with Pure Storage to build its AI Research SuperCluster, which the company is currently using to train AI models in natural language processing and computer vision for research. The AI Research SuperCluster will eventually enable AI models that can work across "hundreds of languages, analyze text, images and video together and develop augmented reality tools", delivering ~5 exaFLOPs of compute. Meta is using Pure's FlashArray//C boxes (QLC-based architecture), which differentiate from competitive offerings in performance, energy efficiency and footprint (one-tenth the space, power and cooling of HDDs). Meta expects the AI Research SuperCluster to become one of the world's fastest supercomputers in the world, with ~16,000 GPUs. Aside from Pure Storage, Meta listed Nvidia and Penguin Computing as partners for the project.

### **AI In The Enterprise**

More immediate use of AI in enterprises today is in the form of robotic process automation, chat bots and vertical specific solutions (e.g., Watson AI for Return To Work which helps customers optimize real estate). IBM and Dell Technologies are offering solutions that integrate AI for automation of processes whereas HP Enterprise's offerings are more focused on helping customers develop, train and deploy models. Companies are also using AI on their own to automate tasks which we discuss in more detail below. Augmenting internal development, we have also seen our companies augment their AI capabilities via a multitude of acquisitions over the past few years. Given the strong free cash flow dynamics of our covered companies, we expect the acquisition of AI related assets to remain a high priority in the future.

**Figure 96: AI-Related Acquisitions Made by Companies In Our Coverage**

Company	Acquired Company	Announced Description
HP Enterprise	Niara	Feb-17 Leader in emerging User and Entity Behavior Analytics (UEBA)
	Cape Networks	Mar-18 Expand Aruba's AI capabilities by measuring and monitoring applications
	BlueData	Nov-18 Enables enterprises to transform how they deploy AI and big data
	Cray Inc	May-19 Leader in Supercomputing
	MapR	Aug-19 Data platform for AI and analytics application
	Determined AI	Jun-21 Provider of software stack to train AI models faster
	Pachyderm	Jan-23 Software to automate reproducible machine learning pipeline
IBM	WDG Automation	Jul-20 Robotic process automation
	Instana	Dec-20 App performance monitoring and observability
	Expertus Technologies	Dec-20 Cloud solutions for financial services industry
	Turbonomic	Jun-21 App resource management and network performance management software provider
	myInvenio	Apr-21 Process mining software company
	MCD Tech Labs	Oct-21 McDonalds automated order taking technology
	ReaQta	Nov-21 Endpoint security solutions that leverage AI to automatically identify and manage threats
	Envizi	Jan-22 Data and analytics software for environment performance management
	Databand.ai	Jul-22 Proactive data observability platform, isolates data errors and issues to alert stakeholders using ML
	Dialexa	Sep-22 Product engineering services firm
NetApp	Cognigo, Inc	May-19 Provider of data discovery classification software designed to manage and protect critical data using AI
	Data Mechanics	Jun-21 Provider of managed platforms for big data processing and cloud analytics
	Pure Storage	Aug-18 Storage solution to manage large sets of unstructured data

Source: Company documents and Credit Suisse

### **Solution Selling**

Over the last several years, hardware OEMs have shifted to selling more solutions in order to: (1) improve margins; (2) drive recurring revenue; and, (3) create stickier customer relationships. We believe, IBM, Dell Technologies and HP Enterprise are investing more in software and services IP while also positioning themselves as advisors to help customers with digital transformation investments, many of which include AI. Solutions range from data management tools to off-the-shelf AI workflows for specific verticals. IBM has the largest portfolio of AI offerings, with the technology sufficiently advanced for public demonstration that the company agreed to have Watson compete on Jeopardy in 2011. The IBM portfolio includes solutions for vertical specific applications, data management tools and building blocks to construct foundational models. Dell Technologies currently offers seven “validated designs for AI” and six APEX for AI solutions (as a service unit) which include common workflows such as image recognition, recommendation engines and fraud detection. HP Enterprise offers several cloud-based AI services via GreenLake, which mostly focus on building, training and deploying AI, machine learning and deep learning models. We think the proliferation of ready-to-deploy AI offerings will serve as a way to democratize AI and bring it into the mainstream.

**Figure 97: Productizing AI Offerings**

Watson Solutions	Dell Validated Designs
AI For IT Operations	AI For Virtualized Environments
AI For Advertising	AI MLOps With cnvrg.io
AI For Healthcare	Automatic Machine Learning
AI For Financial Operations	Conversational AI
AI For Risk & Compliance	Intelligent Video Analysis
AI For Video	NVIDIA Fleet Command
AI For Security	Retail Loss Prevention
AI For Supply Chain	
AI For Return To Work	

Source: Company documents and Credit Suisse

### **Real-Life Examples**

#### *IBM: Eliminating 12,000 Hours Of HR Work*

IBM's AI strategy is focused on driving business outcomes. One example would be the use of Watson Orchestration in IBM's HR department, which automated a single task and eliminated

12,000 manual labor hours. The company is investing in and developing large language models and foundation language models but is leveraging these as a groundwork for more fine-tuned customer specific use-cases. Unlike recent consumer orientated demos that have been high on entertainment value, IBM's AI investments are guided by enterprise business applications with a clear path to profitability. AI R&D engineers are working collaboratively with IBM's product teams to determine how AI can bring about productivity enhancements. Another example includes Project Wisdom which automates coding in Ansible (e.g., developer tells the system "Deploy web application stack"). These useful workflow solutions leverage large language models developed by IBM and long-term the company expects customers on their own will be able to utilize models from IBM, Microsoft, Google, etc., to create smaller, customized models specific to their business operations. IBM recently announced Vela Cloud, which will eventually serve as a platform for organizations to fine-tune models. Consistent with the company's hybrid cloud strategy, Vela Cloud will be open for customers to train models wherever their data resides.

#### *Jabil: Automating Assembly Line Schedules*

Jabil is a tier-1 EMS company that manufactures products and helps manage supply chains for hundreds of companies including Apple. The company has been using AI for years to remedy one-off issues or test proof of concepts but more recently began deploying AI broadly across the enterprise to drive efficiencies. One very notable example is leveraging factory data and AI to automate the scheduling of lines that manufacture 1000s of parts per day. Scheduling is particularly nuanced as not all parts use the same machinery, components or labor. By automating this, Jabil is able to run production faster and more efficiently, enabling incremental revenue opportunities without investing in capex. We recently met with Jabil's head of operations, JJ Creadon, and he expects to begin to see "step-function" changes from the use of AI.

**Figure 98: Companies Using AI In Daily Operations**

Company	Industry	Use of AI
Alibaba	E-Commerce	Personalized customer search recommendations
Amazon	E-Commerce	Multiple use cases including 1) predict customer demand; 2) track product availability; and, 3) deliver packages efficiently
AstraZeneca	Pharmaceutical	Assist pathologists in analyzing samples accurately in lesser time
Bloomberg	Financial Media	Extract numbers and information from documents
Capital One	Financial Services	Multiple use cases including: 1) virtual assistants; 2) call centers; and, 3) including virtual card numbers
CVS	Healthcare	Multiple use cases including: 1) syncing prescriptions; 2) accelerating refill process; and, 3) side effect counselling
DoorDash	Food Delivery	Multiple use cases including: 1) perform security checks; 2) threat identification; and, 3) risk management
Dr. Reddy's Laboratories	Pharmaceutical	Delivers integrated signal detection and risk management
eBay	E-Commerce	Uses text and image similarity models to efficiently classify products
General Motors	Automotive	Identifies broken parts or parts that require maintenance
Grammarly	Internet	Suggests ways to make writing concise and grammatically correct
Home Depot	Retail	Understands customer projects and delivers buying guides accordingly
Insilico Medicine	Biotechnology	Accelerates drug and discovery development
Intermountain Healthcare	Healthcare	Automates daily tasks and reduces time for caregivers
Johnson & Johnson	Pharmaceutical	Multiple use cases including: 1) automating surgery procedures; 2) supply chain agility; and, 3) new services and developments
JPMorgan Chase	Financial Services	Anomaly detection for recognizing fraud and risk mitigation
LinkedIn	Internet	Multiple use cases including: 1) match candidates and jobs; 2) send customized emails; and, 3) streamlining open positions
LYFT	Transportation	Virtual assistance
Marsh	Services	Digitizes and enhances property risk data and delivers actionable insights
McDonald's	Food Services	Upsells products by suggesting snack for kid's meal
Neutrogena	Cosmetics	Scans and analyzes skin to get real time information about skin moisture and pores
Pinterest	Social Media	Labels and categorizes photos to efficiently rank and order them
Quest Diagnostics	Pharmaceutical	Improves diagnosis of cancer and other diseases relying on pathologic tests in a speedy manner
Salesforce	Software	Collects data and provides predictive analysis and language processing capabilities
SoftBank	Communication	"Sakimiri" helps predicting and forecasting customer traffic at individual client stores
Truist	Financial Services	Virtual assistance
Uber Eats	Food Delivery	Estimates delivery times and ranks restaurants by customer choices
UnitedHealthcare	Healthcare	"Optum" uses data to predict patients who will develop heart condition leading to strokes
UPS	Transportation	Multiple use cases including: 1) track packages in facilities; 2) vehicle optimization; and, 3) package delivery routing
Vanguard	Financial Services	Understanding customers financial needs and addressing special needs
Verizon	Communication	Optimizing 5G transmitter placements
Walmart	Retail	Multiple use cases including: 1) supply chain ecosystem; 2) making smart substitutions; and, 3) predicting customer demand
Waymo	Automotive	Drivers use data to calculate a safe route allowing car to respond in real time
Wells Fargo	Financial Services	Virtual assistance

Source: Company data, Credit Suisse

**US Semiconductors Sector – Analyst: Chris Caso**

GPU has proven to be the widest adopted technology for training AI models. GPUs are well suited for the matrix calculations required for training AI models (multiple/accumulate functions that drive the probabilities needed to train these models. In AI, the larger the dataset the better the model – so there is an ever-increasing need for higher performance, driven by larger model sizes and enabled by GPUs with higher transistor counts and ability to execute a larger number of calculations in parallel than CPUs. According to Gartner's AI forecast for processing, GPU is projected at 77% of AI sales in 2023 and projected to grow at a 19% CAGR from 2023 to 2026. ASIC from a lower base is at 15% of workloads in 2023 though projected to grow at a 50% CAGR as TPUs and other ASICs optimized for certain AI calculations are adopted. Notably, AI GPU in Gartner's figure at US\$7.2bn does not capture all of NVIDIA's data center GPU usage at US\$15bn which also includes revenue from the entire GPU system.

**Figure 99: AI Processors Projected at 25% '23-26 CAGR, GPUs Capturing 77% Share**

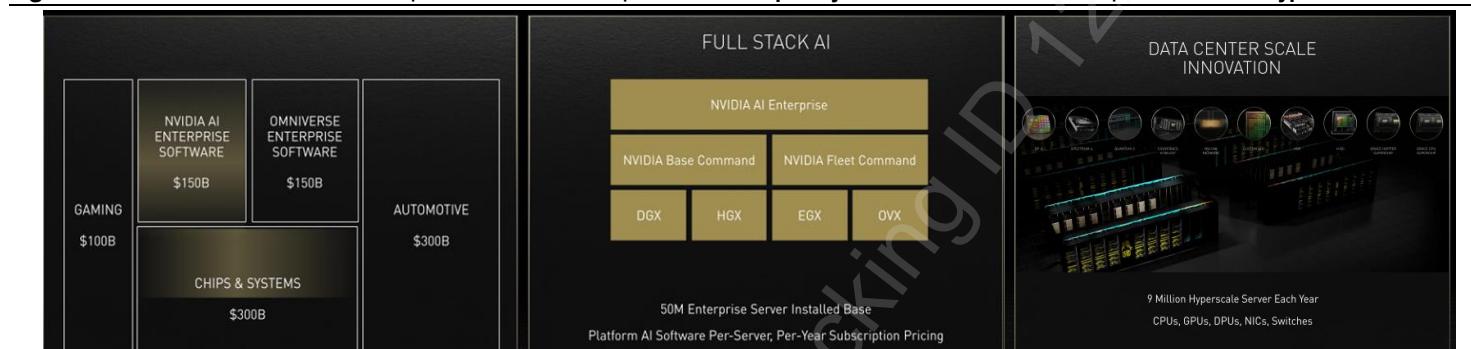
AI sales \$mn	2020	2021	2022	2023	2024	2025	2026	20-23	23-26	% of 23
GPU	\$2,609	\$4,786	\$5,869	\$7,231	\$8,897	\$10,559	\$12,166	40%	19%	77%
FPGA	\$104	\$205	\$336	\$612	\$831	\$908	\$942	81%	15%	7%
ASIC	\$271	\$501	\$828	\$1,449	\$2,576	\$3,751	\$4,854	75%	50%	15%
DSP	\$6	\$14	\$32	\$69	\$102	\$152	\$216	128%	46%	1%
<b>Total</b>	<b>\$2,989</b>	<b>\$5,506</b>	<b>\$7,066</b>	<b>\$9,360</b>	<b>\$12,405</b>	<b>\$15,372</b>	<b>\$18,178</b>	<b>46%</b>	<b>25%</b>	<b>100%</b>
<b>YoY Growth</b>	<b>84%</b>	<b>28%</b>	<b>32%</b>	<b>33%</b>	<b>24%</b>	<b>18%</b>				

Source: Gartner, December 2022

Within GPUs, NVDA leads the market for AI training, with 95-100% market share, according to Mercury Research. We estimate NVDA's datacenter revenue is roughly evenly split between cloud and on-premise products with majority of revenue comes from training vs. inference.

The democratization of NVDA silicon through cloud instances means that even small developers can develop the next ChatGPT. We believe that creates open-ended growth which could ultimately expand data generation and growth trajectory for servers or put AI acceleration into servers on a much faster pace. While it's difficult to accurately upsize the training and inference markets, for their part, Nvidia has identified a datacenter TAM opportunity of \$600bn, with \$300bn in hardware (chips/systems) and \$300bn in software. Within that, NVDA estimates the hyperscale TAM for infrastructure alone represents a \$150bn opportunity.

**Figure 100: NVIDIA Estimates a US\$1trn TAM with US\$300bn in Chips/System Hardware and US\$150bn from Hyperscale**

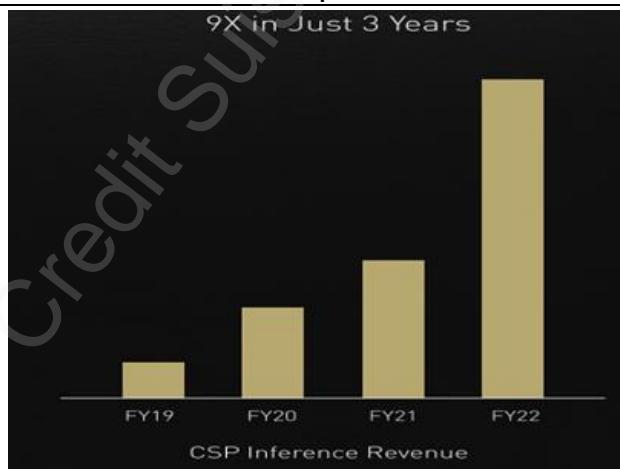


Source: NVIDIA

**Inference shifting from x86 CPUs toward GPU and other accelerators.** Inference is the task of running the AI models – responding to a ChatGPT query, providing a recommendation on a shopping site, or responding to an Alexa voice command, “inferring” a result based on how the model has been trained. While training is done in a batch process, inference is done in real time, thereby representing different compute needs.

Traditionally, the majority of inference workloads run on x86 silicon, mostly from INTC. But GPU has been shown to deliver higher performance for inference, for much the same reason that GPU has proven to be better for training. While NVDA hasn't disclosed the specific revenue or growth rates for inference GPU vs. training, the company does claim that inference revenue is up 9x between NVDA's FY19 and FY22. See more detailed insights at the Nvidia section toward the end of this note.

**Figure 101: Nvidia's Inference GPU Sales Up 9x in 3 Years**



Source: Nvidia

### **China Technology & Telecom Sector – Analyst: Kyna Wong**

**Inspur (Outperform):** In 2022, AI servers represented ~30% of Inspur's sales. According to IDC, Inspur was the largest AI server supplier in China/globally in 2021, with 52%/21% market share by revenue, respectively. IDC forecasts China AI server market to grow from US\$5.4bn in 2021 to US\$11.3bn in 2026, at a CAGR of 16%, while we believe the acceleration of AI applications will drive the upside. Though we have not seen instant orders increase in near term, we believe it will be the key beneficiary of AI infrastructure in the long term. Inspur launched its AI large language model (LLM) in 2021, and four Skill Models in mid-2022. Its LLM covers applications in financials, internet, healthcare, and self-driving. It expects Tier 2/3 internet companies to be its major customers, given Tier 1 giants are building their own models.

**WUS (Outperform):** WUS is one of the key PCB suppliers for US HPC market. Due to its exposure to high-end communication market, we believe it will be one of the key beneficiaries of frontier technology upgrades, which was accelerated by the expansion of AI demand. On the PCB side, we expect product mix to shift to higher layer counts (to 16-20L or above), higher-speed materials (to very low loss class) and more HDI adoption. Its products for Eagle Stream will start mass shipment from Mar. The CPU platform replacement, together with more accelerator chips (FPGA, GPU), will push the penetration of PCIe 4.0/5.0, and leads to a mix improvement for PCB suppliers.

**Innolight (Outperform):** We believe Innolight is well positioned for data traffic boom generated by the broader AI applications, and a key beneficiary considering it already has a leading supplier position with Google, Microsoft, Amazon, META, etc. We expect AI could further accelerate the penetration of 200/400G, as well as the transition to 800G for certain end users. Innolight has approx. 20% market share in the global optical transceiver market, ranked #1 by Lightcounting.

**Montage (Outperform):** Among A-share semiconductor names under our coverage, Montage is the only one with more than 80% profit from server DRAM interface chips and companion chips. If ChatGPT drives demand for high-performance servers, then Montage is well poised as a direct beneficiary to observe increasing demands for its server DRAM interface chips and companion chips.

### **Korea Internet Sector – Analyst: Soyun Shin**

**Enhancement in search engine helps grow search ad sales.** AI plays a major role in improving portal search engines. **NAVER**, the largest portal service provider in Korea, launched an AI-based search engine service, AiRSearch in Oct 2021. AiRSearch is a customized search engine based on its Smart Blocks system. It provides subdivided topics in search results, which allow users to discover their preferences and explore the results in a diverse way. NAVER launched AiRSearch to correspond with users' new tendency of exploring information rather than finding a single answer. The company highlighted exploratory searches account for 65% of total queries entered into its search service. Specifically, exploratory searches for the shopping segment have increased sharply compared with other segments which help NAVER grow its commerce ad sales by increasing the number of commercial search queries.

NAVER has continued to focus on improving its AI technology through R&D internally and investment into diverse start-up companies. NaverD2SF mentioned that investment on AI-related start-ups account for 17% of total investment in 2021 (Metaverse: 24% / Commerce: 21%). They also conducted investment into three AI-related start-ups last year: SqueezeBits, which is developing a solution to accelerate AI models; Gengen AI, which is developing an AI-based training solutions; and Artificial Society, AI based edu-tech company.

Planning to release SearchGPT in July 2023. NAVER is scheduled to launch 'HyperClovaX,' which learns Korean 6,500 times better than ChatGPT. NAVER mentioned it will build a higher-scale AI ecosystem specialized for non-English speaking countries. NAVER will also launch a pilot program of SearchGPT based on 'HyperClovaX' technology and the information of Naver Shopping and blogs. If executed well, we believe NAVER can monetize larger advertising sales.

**Korean online gaming developers.** Given the importance of bots (virtual players inside the game play) inside the game play of MMO (Massively Multiplayer Online) games, online gaming developers have been interested in AI-based characters with better communication capabilities, more natural interactions with gamers, and more diversified contents than before.

**NCSOFT** is a leading gaming company leveraging AI in their game development and formed a special team to develop AI and use relevant technology for the game launch (Lineage W launched in November 2021) such as AI-based self-play learning, Natural Language Processing (NLP), AI translator, AI voice to text technology on its games. We believe it would be helpful for online gaming developers to retain users and improve users' experience by improving the capability of virtual bots.

## Business Services

Over the past few decades HCM & Services companies have leveraged digital technology to drive more efficient deliverables. We see these trends accelerating given advances in technology and that ChatGPT has the potential to drive further efficiencies, increase processes, boost customer retention, reduce workloads and position employees to focus on more strategic tasks. These initiatives not only lower costs, but also improve products and service quality to clients.

### Key ChatGPT Use Cases in Business Services

- **Increasing sales productivity:** Sales reps will be able to generate marketing content and formulate outreach messages more efficiently.
- **Increasing innovation:** Develop new products faster as the development process is shortened to modifying ChatGPT-written code as opposed to writing from inception.
- **Customer service:** ChatGPT chatbots or virtual assistants can answer inquiries, provide recommendations, and help place orders reducing workload on customer service representatives. Could also help analyze customer feedback data, identify common issues and complaints, and provide insights on areas for improvement.
- **Automation of more routine tasks:** Conditioning GPT-3 models to replicate language-based tasks such as summarizing large amounts of financial data [FDS, TRI], or automating routine tasks such as data entry, report generation and formatting [SPGI investor transcripts].
- **Data analysis:** Analyzing large amounts of complex data to identify trends and patterns that might not be readily apparent to extract insights and make the process more efficient.

### AI Use Cases, Beyond ChatGPT in the Business Services Industry

When thinking more broadly about the implications of AI to the Business Services sector, we identify the following as key use cases of AI on the industry. Note, some of these use cases will include the training and development of new large language models (LLMs):

- **Market analysis:** More complex models conditioned to analyze market developments real time amid potentially predictable outcomes.
- **Task replication:** Automation of even more complex worker processes has the potential to drive further efficiencies and fuel higher labor cost savings.
- **Candidate screening:** Amid a tight labor market, hiring should become less labor intensive with the use of AI as candidates are screened more accurately.

### Sector Coverage Implications

#### **US HCM & Services Software Sector – Kevin McVeigh**

Esther Colwill, president of global technology, communications, and professional services at Korn Ferry noted that she sees ChatGPT improving workplace processes to (1) automate non-value-added tasks; (2) free up people for more creative, innovative work; (3) improve customer service; (4) aid in training and development; (5) help employees respond to emails and (6) draft basic documents—and more. “It’s about process improvement,” she says, “and leveraging whatever tools are at an employee’s disposal to create an outcome that meets customer need.” In 2021 SS&C acquired Blue Prism for \$1.6b [7.5x revenue]. Blue Prism offers robotic process automation [RPA]—software with expanded cognitive and AI capabilities that can be trained to shadow worker processes—creating automated systems to perform tasks faster and better. In 2023 SSNC anticipates deploying ~2700 digital workers [~10% workforce] driving ~\$50k savings per employee, or \$135m in annualized run-rate savings. Collectively, SS&C has

invested ~\$15b in R&D and M&A since 2012 to develop next-generation solutions, intuitive interfaces, cloud-native technology, and incorporate artificial intelligence [AI], robotic process automation [RPA], and machine learning into its workflow. Chat GPT can be used in collaboration with other AI and machine learning alternatives such as Blue Prism to enhance the existing suite of solutions.

Credit Suisse Report Tracking ID 1227385

## Financial Services

The financial services sector includes the banking, insurance (property & life), brokerage, asset management, payments, and wealth management industries, amongst others. Nearly every aspect of the financial services industry is likely to benefit from ChatGPT, although we believe wealth management and insurance are industries that will see an outsized impact from LLMs. For example, wealth management is already expanding from an expensive service provided to mostly high net worth individuals by adding cheaper, AI-based financial planning options and ChatGPT should accelerate adoption. The insurance industry should also benefit from broader AI utilization as risk parameters will be expanded and more precisely defined, leading to more accurate pricing. Given the breadth of the financial services industry, there are numerous use cases from the rollout of ChatGPT.

**Figure 102: ChatGPT Is Helping Financial Analysts Recap Important Events**

summarize apple earnings call from 2015

As an AI language model, I do not have access to real-time information or current news. However, I can provide a summary of Apple's earnings call from 2015 based on the available information.

In April 2015, Apple reported their financial results for the second fiscal quarter of 2015, which ended on March 28, 2015. During the earnings call, Apple announced that they had generated \$58 billion in revenue and had sold 61.2 million iPhones, 12.6 million iPads, and 4.6 million Macs. The company's net profit for the quarter was \$13.6 billion.

Apple also announced that they would be increasing their dividend and share repurchase program, returning \$200 billion to shareholders by March 2017. The company's CEO, Tim Cook, highlighted the strong growth in China, which had become Apple's second-largest market after the United States.

Overall, the earnings call showed that Apple had continued its strong performance in the smartphone and tablet markets, while also expanding into new markets such as China. The company's focus on returning value to shareholders also demonstrated its commitment to long-term growth and stability.

Source: Credit Suisse Research.

## Key ChatGPT use cases in Financial Services include

- **Extract Financial Information:** Users can extract information from documents to quickly analyze and process customer inquiries, and generate summary financial reports.
- **Introductory Analysis for Bankers:** Financial service professionals can use ChatGPT for initial overviews of company business models and for recaps of earnings calls and other important events.
- **High Level Financial Advisory:** ChatGPT provides simplistic financial advice to a broader range of customers, impacting advisory services.
- **Simplifying Financial Calculations for Consumers:** End consumers can use ChatGPT to conduct financial calculations conversationally, thereby enhancing their ability to inform their decision-making processes.
- **Improving Customer Service:** ChatGPT's ability to communicate in a humanistic way eases the burdens of customers' automated assistance in various banking processes; a significant improvement from technologies that need to be communicated to with very specific phrases.
- **Product Marketing:** Generate marketing campaign script and creative marketing ideas to reduce the cost of content production.

**Figure 103: Consumers Are Using ChatGPT About Financial Planning**

My monthly income is \$5000. How and where do i invest to make \$5M in 5 years?

Making \$5 million from a monthly income of \$5,000 is an ambitious goal. It would require an extremely high rate of return on investment, and it's important to keep in mind that investing always carries some level of risk. That being said, there are a few ways to potentially grow your money over time:

1. Invest in the stock market: Investing in individual stocks or mutual funds can offer the potential for higher returns, but it also comes with a higher level of risk.
2. Real estate Investment: Investing in real estate through rental properties or flipping houses can be a good way to earn money, but it requires a significant amount of money to start and it also comes with a lot of risk.
3. Starting a business: Starting a business can be a high-risk, high-reward endeavor, but if successful, it can provide a significant return on investment.
4. Alternative investment: Investing in assets such as precious metals, art, or venture capital funds can provide potentially high returns, but it also comes with a higher level of risk.

It's important to keep in mind that investing in any one asset class may not be sufficient to reach this goal. Diversifying your portfolio across multiple asset classes may help to reduce risk and increase the chances of achieving your goal.

Source: Credit Suisse Research.

## AI Use Cases, Beyond ChatGPT

When thinking more broadly about the implications of AI to the financial services sector, we identify the following as key use cases of AI on the industry. Note, some of these use cases will include the training and development of new large language models (LLMs):

- **Real-Time Financial Market Trading Advice:** Advanced AI will be able to navigate through enormous data sets, quickly parsing out trends, and providing actionable advice on trading decisions in real-time. In addition to being quick, the AI simultaneously distills the complex information into easily understandable summaries.
- **Financial Forecasting:** Utilizing limited inputs, AI can create financial forecasts to determine the viability of an investment. The AI will be able to utilize its training set to incorporate assumptions not simply limited to the proposed investment, but the industry trajectory, at large.
- **Fraud Detection:** Novel AI will be able to analyze not only individual banking transactions, but view many activities together across accounts and customers to detect fraudulent and/or other criminal activities.
- **Product Personalization:** Using certain user-defined inputs, AI can provide personalized financial product advice. For example, if a customer is looking to purchase an insurance policy, AI can analyze the customer's financial situation, budget, protection requirements, and product features and recommend the policy that will work best for the customer.
- **Risk Management:** Verify customer information to assess the eligibility for loan applications, automate loan underwriting, assess credit risks, etc.
- **Enhancing Excel:** Users will be able to quickly and easily communicate with ChatGPT to create complex excel formulas, improving productivity and transferability between workers, as differences in skillsets are not a limiting factor.

**Figure 104: ChatGPT Is Currently Limited, But Future AI Iterations Will Likely Synthesize Complex Excel Formulas Conversationally**

What's the excel formula to calculate ROI on crypto investments?

The formula to calculate return on investment (ROI) in Excel is:

$$\text{ROI} = ((\text{Current Value} - \text{Initial Value}) / \text{Initial Value}) * 100$$

For example, if you invested \$1,000 in a cryptocurrency and its current value is \$1,500, the ROI would be calculated as:

$$\text{ROI} = ((1,500 - 1,000) / 1,000) * 100 = 0.5 * 100 = 50\%$$

In this example, the ROI is 50%, meaning that the investment has generated a 50% return.

You can use this formula to calculate the ROI of multiple investments by using the SUMPRODUCT and SUM functions.

For example, if you have a list of initial investment amounts in column A and a list of current values in column B, you can use the following formula to calculate the total ROI:

$$=\text{SUMPRODUCT}((B1:B5-A1:A5)/A1:A5)$$

You can also use this formula to calculate the ROI of a single investment over a period of time.

**Figure 105: Improved Fraud Detection Would Save Consumers Over \$6 Billion, Before Accounting for Business Losses**



Source: Credit Suisse Research.

Source: FTC's Consumer Sentinel Network, Credit Suisse Research.

## Sector Coverage Implications

### **U.S. Banks (Large Cap) Sector – Analyst: Susan Roth Katzke**

The largest, most sophisticated U.S. banks have been active users and developers of tools that leverage AI across multiple areas of banking operations, from the ability to increase intelligence/detection of fraud to the ability to automate and efficiently service consumer banking customers. The use has shifted from more limited servicing/efficiency initiatives to now becoming increasingly “smarter” and more sophisticated, deployable to (i) revenue generating activities, from Bank of America’s “Erica” on its consumer banking platform (automating customer service interactions, as well as providing financial insights/guidance) to the “Next Best Action” generator that supports financial advisors on the Morgan Stanley platform, (ii) broadly supporting risk management, consider the opportunities to reduce fraud costs, or within trading operations to support both revenue generation and risk mitigating trading strategies. We’d suggest the applications of AI/ChatGPT across the banking and financial services industry to be quite broad and increasingly important over time.

### **Insurance Sector – Analyst: Andrew Kligerman**

The insurance industry relies on large aggregations of data for the development of pricing manuals and underwriting protocols as well as the management and processing of insurance claims. Having a history of quantitatively driven decision making, with actuaries developing pricing models since the 1800s, the insurance industry has persistently introduced new tools, including developing technologies such as machine learning and artificial intelligence, to support the underwriting and claims processes. Furthermore, as insurance providers continue to look to enhance customer experience, tools including artificial intelligence are becoming more common to connect with the customer. Machine learning and artificial intelligence is most impactful in the insurance industry for pricing and claims management. In auto insurance, telematics is becoming more prevalent to provide driving behavior, as well as real-time crash notifications. Artificial intelligence is also becoming more prevalent in supporting underwriting and claims management practices.

### **Specialty Finance Sector – Analyst: Moshe Orenbuch**

In Consumer Finance, ChatGPT can

- Be used in customer support functions to help with account-related questions, thus reducing labor costs.
- Draft customer communication via emails, direct mail..., user agreements, as well as collection notices.
- Give consumers advice on how to improve financial life.
- Set up auto-payments, manage credit card billing and payment subscriptions.

While other AI use cases may require continued AI development:

- Assist with loan underwriting process if given enough inputs on the consumers (we note that many companies already use some machine learning or AI in their development of credit models).
- Aiding in communications with stakeholders (regulators, consumers, investors...) in a helpful and compliant way.
- Help consumers select the most suitable loan product given their financial profiles.

### **U.S. Exchanges & Electronic Trading Platforms Sector – Analyst: Gautam Sawant, CFA**

Exchanges can leverage AI and machine learning across core transactional businesses (CBOE, CME, NDAQ, ICE) to improve order execution, expand product development, and identify clients trends/behaviors to optimize sales generation. We believe AI can accelerate fixed income electronification (BGCP, MKTX, TW) by streamlining bond selection criteria, improving liquidity

alert capabilities and by improving pricing analytics through the integration of data from multiple sources. We expect AI to enhance exchange surveillance and market regulation capabilities which could support market abuse identification and ensure the integrity of markets. Exchanges' expanded data and technology services to the financial services industry can also benefit from artificial intelligence by accelerating code development and creating upsell opportunities through product enhancements. Nasdaq for example is leveraging AI technologies within Investor Relations products to improve IR program efficiency and within Anti-Financial Crime, Verafin is a cloud-native fraud solution for the financial services industry that leverage artificial intelligence and machine learning to operate targeted analytics models and produce big data intelligence. CME's Google partnership is well positioned to enhance existing data/analytics solutions with AI capabilities and integrating "BARD" technology could create a more seamless user interface for data analysis and product information. Tradeweb's AiEX and Marketaxess' Auto-X trading protocols are examples of how automatic execution tools can increase trader efficiencies and reduce execution costs.

### **Broader China Financials Sector View**

We believe ChatGPT is unlikely to be widely adopted by Chinese enterprise despite the potential benefits, due to 1) rising concern over data privacy and different cross-border data transfer mechanisms between the US and China, and 2) US protectionist policies which potentially safeguard its large language model and other advanced artificial intelligence development amid US-China tech decoupling. In China, residents are not allowed to create OpenAI accounts to access the AI-powered chatbot. According to press on 22 February, China's regulators have instructed Tencent and Ant Group to restrict access to direct or third-party ChatGPT services on the platforms.

However, Chinese home-grown counterpart ERNIE Bot can be used in the country. ERNIE Bot, built by Baidu, is currently in the internal testing stage and scheduled to be open to the public in March 2023. The adoption of ERNIE Bot or other domestic third-party AI model by banks and insurers would likely be subject to more restrictions than other industries, as the financial sector (1) generally involves highly sensitive customer data and (2) has been highly regulated. Nevertheless, multiple insurers and banks have developed their proprietary AI models in an effort to improve operational efficiency. Regulators are generally more accommodative with these proprietary AI models, as the data will stay with the institution itself.

### **China Insurance Sector – Analyst: Charles Zhou**

Due to the sensitive customer data in the industry, we see limited feasibility for insurers to adopt third-party ChatGPT-style AI model which might lead to the risks of data leakage. However, some insurers have developed proprietary AI to improve operational efficiency, customer experience and agency quality. Most AI applications in the insurance industry are in 1) sales and marketing, 2) personalized underwriting and claim, 3) risk identification and control. The impact of the AI technology is currently limited, as it is still at a nascent stage. In the long run, the impact would depend on digital maturity and the regulatory environment.

### **China Banking Sector – Analyst: Hu Shen**

The application space of AI in the banking industry is relatively limited at the current stage. The two use cases in the banking sector could be 1) customer service improvement and 2) marketing content generation.

Regarding the first area, after years of IT investment, many Chinese banks have already adopted self-developed AI to support multiple client-interface functions, effectively replacing a large percentage of human staff in functions such as handling client enquiries, approving small loan requests, and collecting overdue loans. With the user habit migration, client interactions on bank apps and WeChat public accounts increased by ~20% YoY, according to a report compiled by China Banking Association in 2022. The report also said employees in banks' client service centers dropped for the first time in the recent five years, by 7.72% to 50.2k in 2021. The number compares with the total number of 3.8mn employees in China's banking industry by mid-2016. With the percentage of relatively low-cost client service employees limited, and the AI deployment already in place, we see an insignificant cost-saving effect for banks from

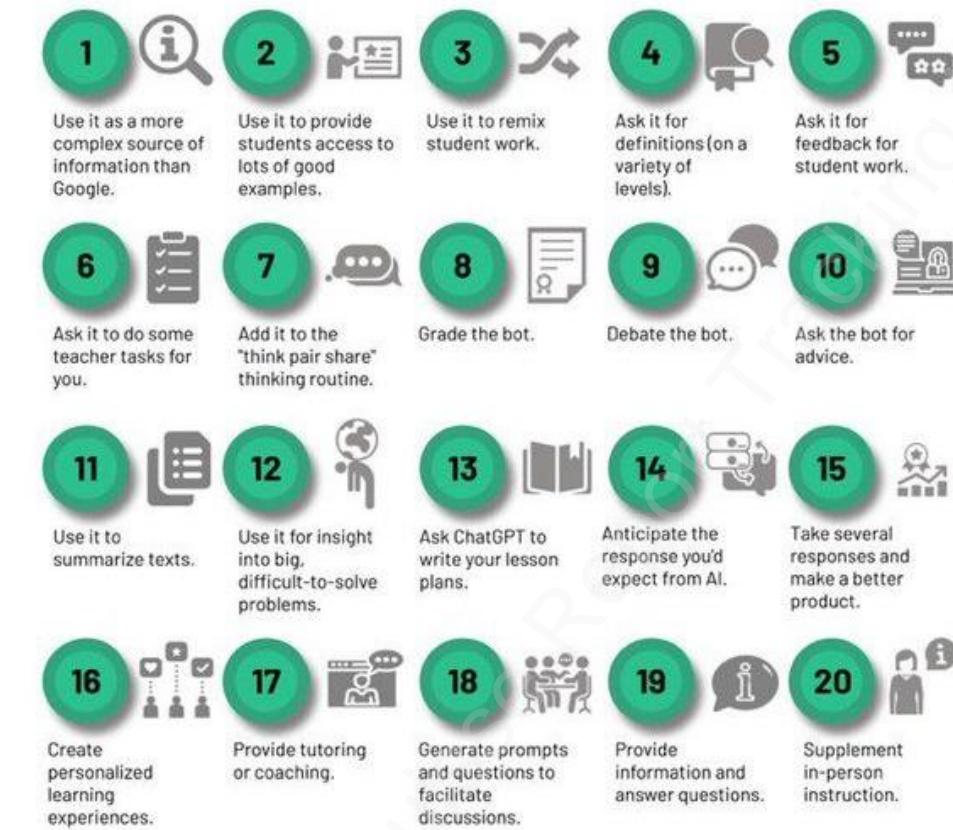
adopting ChatGPT. At the same time, the service quality could see some improvement, which will be helpful for client satisfaction and retention.

Regarding the second area, we reckon ChatGPT will have limited use in brand promotion because advertising requires high creativity and meticulousness. At the current stage, we see low feasibility for Chinese banks to adopt ChatGPT technology. This is because the banking system is the backbone of the state's financial infrastructure, and client account information could be very sensitive. It is not very likely Chinese banks will be allowed to open their databases to a foreign developed technology. On the contrary, they could be inspired or encouraged to cooperate with the domestic developed ChatGPT-like technologies, for example, by their Chinese competitor Baidu.

## Education

Modern technology adoption across the education industry is challenging traditional classroom models, and we believe ChatGPT is a catalyst that will further unlock the future of learning as educators increasingly look to embrace advanced digital learning capabilities and offer more learning pathways for students. Educators are already using ChatGPT to shift student focus toward developing more critical thinking and analytical skills versus memorization; teachers are evaluating students based on their ability to build upon and analyze ChatGPT-generated content, encouraging students to think outside the box. Overall, we believe the growing scope of ChatGPT and AI capabilities—specifically in digitizing learning environments, rethinking content creation, and instruction delivery—support the sustainable accelerated transformation of primary, secondary, and higher education.

**Figure 106: 20 Ways to use ChatGPT in the Classroom**



Source: Infographic by Matt Miller / Ditchthattextbook.com, Credit Suisse Research

### Key ChatGPT use cases in Education include:

- **Administrative Support:** Teachers can automate time-sensitive workflows (creating curriculum, lesson plans, homework, grading, attendance, etc.) so more time can be spent directly with students, while also cutting overall costs relating to support staff. We also expect that such functionality will equip teachers to better balance larger class sizes.
- **Richer Engagement:** ChatGPT and generative AI offer an opportunity to continuously redefine and supplement the traditional engagement model. As the way students consume and interact with content evolves, how educators engage learners will also.
- **Improving Access and Equity:** ChatGPT could be used to serve populations with limited access to teachers, learning materials, and/or schools. We believe ChatGPT can implicitly

lower the barrier to quality education resources, pathways, and general skill development opportunities for all student levels more broadly.

- **Detecting Plagiarism:** We also note that some education departments (NYC, LA, and Seattle's Department of Education) have blocked access to ChatGPT due to fears of plagiarism. We believe software will also play a role in helping detect student work that has incorporated this functionality, which we ultimately believe is likely to benefit from the growth of ChatGPT.
- **Chatbots:** Universities can elevate student experiences by integrating ChatGPT into chatbots to drive efficiency across administration, recruitment, student retention, and student service throughout their academic career. This can also include more personalized actions such as applying for scholarships and funding, automating payments, enrolling in courses to achieve degree credit requirements, or engaging with other student resources.

## Complex AI Use Cases Beyond ChatGPT

- **Personalized Learning:** ChatGPT coupled with AI more broadly can help educators identify and evaluate student-specific strengths and needs and ultimately create customized learning plans – considering the type of content, the method of instruction, and the pace of completion. It is also assisting teachers with generating and customizing further instruction, practice, and feedback based on specific assignments.
- **Student Analytics:** Digital adoption and connected platforms coupled with the growing prominence of learning models collectively poised educators to unlock greater student analytics with the goal of improving outcomes. This naturally includes quantitative assessment but we also expect social and emotional understanding to grow in prominence to best support students as they grow.
- **Student Support:** AI could be helpful in surfacing valuable insights and recommendations to students to help guide their decision making. This could address school choice, expertise specialization, career paths and more based upon their skills, interests, and performance. It could also be helpful in identifying alternative education programs among other learning opportunities. Finally, it could highlight relevant and available support resources.

## Sector Coverage Implications

- **PowerSchool (Outperform-rated, covered by Rich Hilliker)** offers a comprehensive K-12 software platform that reaches more than 80% of students in the US and Canada. We believe the company's platform is poised to benefit from AI, ChatGPT, and the continued evolution of LLMs due to its position as a core system of record, instruction, analytics, administration, and engagement. Leveraging the breadth of its connected platform and its rich ecosystem of student, teacher, and school data, PowerSchool employs AI and machine learning to offer schools the ability to address and adapt to student needs, personalize learning pathways in real-time, empower educators, and derive insights to optimize the broader organization's operations; we're particularly bullish regarding the potential uptake of the company's Unified Insights suite along with the recently announced LearningNav and ContentNav offerings. We believe further advancements of ChatGPT and AI more broadly will strengthen these functions and enhance the value of PowerSchool's unified platform further.

## Healthcare

Healthcare industry is an aggregation of sectors within the economic system that provide goods and services to treat patients for them to maintain and re-establish health, including subsectors of pharmaceuticals, biotech, healthcare facilities, managed care, and life sciences tools & diagnostics, etc. Healthcare has lagged other industries in the adoption of AI-based tools despite the potential benefits in cost saving and efficiencies. It has been estimated 5-10% of US healthcare spending could be saved (c.\$200-360bn) with wider use of AI. ChatGPT has the potential to enable the broader use of AI across Healthcare, in our view. We see ChatGPT's main use in engaging with consumers and healthcare professionals, creating efficiencies in the system and potentially improve medical/treatment outcomes. In the future we see potential in augmenting physician diagnosis of patients but in drug discovery it is unlikely to be used given sophisticated AI is already well embedded in this process.

### Key ChatGPT Use Cases in Healthcare

For Healthcare, the most significant impacts will likely be in the field of NLP – the ability for the computers to understand and generate human language. AI-powered LLMs can be used to extract information from electronic health record (EHRs), delivering cost savings, creating efficiencies, and improving medical/treatment outcomes. Additionally, LLMs can be trained for clinical decision support systems, which can help healthcare professionals to make more informed decisions and more accurate diagnoses as the model can identify trends and catch patterns by learning a lot of data. Specific use cases for Healthcare include:

- **Faster Processing of Records:** GPT can enable healthcare staff to aggregate and process records at faster speeds. ChatGPT can help with medical coding, billing and reports generation. We see that with continued training for the system, ChatGPT could become familiar with large amounts of medical data including billing and coding data to help improve efficiency and accuracy of these functions.
- **Enhancing Diagnoses and Outcomes:** Assist healthcare professionals in diagnostic medicine leading to improved patient outcomes (in 87% of cases, the correct diagnosis was listed among the top 3 most likely possibilities).
- **AI Powered Assistant:** Act as a virtual health assistant and help to collect medical records and ask basic patient history questions. We see that certain routine tasks which can be automated or require very little human supervision can continue to be passed off to AI systems, including ChatGPT. This also includes answering basic questions from patients as well. Some of these items telemedicine players are already incorporating AI to do within their chatbots.
- **Enabling Doctors to Focus on Patients:** Reducing physicians' paperwork and communications burden by automating prior authorization requests, claims appeals, and electronic communications.
- **Help with medical research:** ChatGPT can also help researchers look through medical literature and to generate abstracts and summaries quickly. While we note that researchers will have to review what ChatGPT generated to ensure relevancy and accuracy, we see that ChatGPT could be used as a tool to help speed up the data collection process.

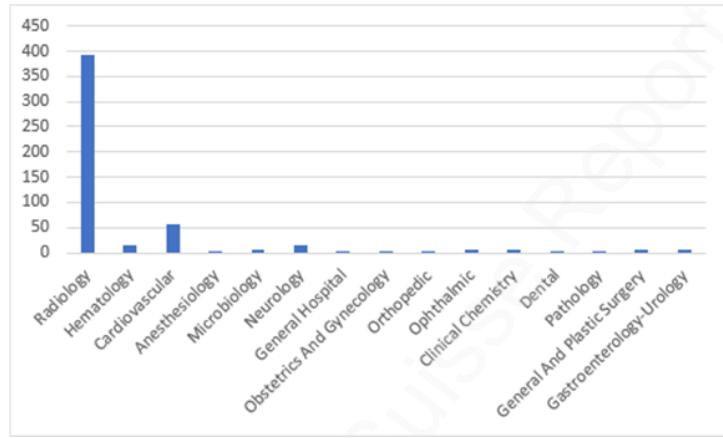
### AI Use Cases, Beyond ChatGPT in the Healthcare Industry

When thinking more broadly about the implications of AI to the healthcare sector, we identify the following as key use cases of AI on the industry. Note, some of these use cases will include the training and development of new large language models (LLMs):

- **AI-Assisted Surgery:** Enables objective data-driven decision-making and will have a strong impact on how surgery is performed in the future.

- **Assisting in More Complex Screenings:** Future AI should be able to help practitioners use blood samples to identify diseases that may be rare and unlikely to be directly tested for.
- **Improved Drug Development:** We believe that 2023 could be an important year for AI in drug discovery as we expect to see some proof-of-concept data in the AI-discovered drug field. In 2021-22, we see there were an increased number of AI-discovered drug candidates enter into clinical trials and this year we may see data come out for these candidates. If we see that these drug candidates show good or better data than existing drugs, this could continue to validate the role that AI could play in drug discovery.
- **Diagnostic Algorithms:** These algorithms will incorporate insights gained from biology (DNA, RNA, proteins, etc.) and electronic health records.
- **Imaging:** Reviewing images and scans is an important job for a doctor as it provides essential insights to their diagnosis. As there is a large record of completed patient scans with patient diagnoses, there are a large number of data sets that AI tools can be trained on to help with predictive medical imaging analysis that could be applied to X-rays, CT scans, MRI, and ultrasounds as well as more specific structural scans of the brain, heart, eyes, and other organs. In the future, medical imaging analysis AI tools could continue to assist doctors in making more informed decisions.
- AlphaFold, an AI system developed by DeepMind, is designed to predict the 3D structure of a protein through its amino acid sequence, a critical determination of a protein's function, which takes years to discover in a traditional lab. See more AI's implications on drug research in the next section.

**Figure 107: Of the >500 FDA Clearances/Approvals for AI/ML-Enabled Medical Devices, the Vast Majority Are in Radiology**



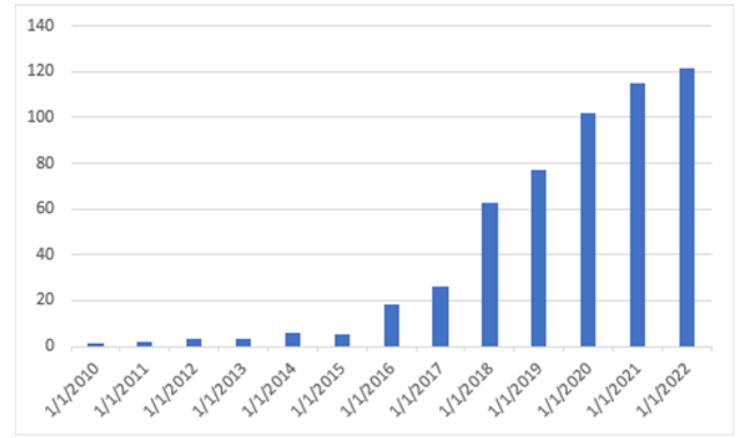
Source: Company data, Credit Suisse estimates

## Sector Coverage Implications

### US Pharmaceuticals Sector – Trung Huynh

Over the last decade the pharmaceutical industry has embraced the use of digital technologies throughout drug development, production, packaging and delivery, and the marketing of products. In drug discovery, digitization and AI has been [reported](#) to reduce R&D costs c.60% and shorten time to market by 2.5 years. We note the launch of [BioGPT](#), similar to ChatGPT, which uses AI to answer biomedical research questions by reviewing literature. The hope is to eventually provide unique combination target discoveries for preclinical testing. However, we believe it is very unlikely ChatGPT will be used widely in drug development given AI is significantly embedded in the R&D process. We think ChatGPT is expected to have a greater impact in digital marketing solutions to help improve the way Pharma can engage with consumers and healthcare professionals. In general, engagement with ChatGPT could provide a

**Figure 108: Number of FDA Cleared/Approved Medical Devices Tagged as Having Some AI/ML Component Has Risen Rapidly**



Source: Company data, Credit Suisse estimates

more efficient way to access and summarize drug information and clinical data to improve patient and physician education. ChatGPT and other machine learning tools offer the potential to enter the clinical decision process, and we have seen many examples where ChatGPT has been used to diagnose patients. However, we think physician and human interaction will always be key, and future mass adoption of ChatGPT in writing prescriptions will be unlikely, in our view, but rather used as an adjunct to physician diagnosis to reduce any misdiagnosis. **From a stock perspective, there is not one company among our names that stands out being ahead of the others with digitalization, and we see ChatGPT being a broad enabler technology rather than benefiting one company over another.**

#### **Life Sciences Tools, Services, and Diagnostics Sector – Dan Leonard**

AI/ML has been increasingly used in medical devices, in particular medical imaging. The FDA has cleared or approved nearly 400 different AI tools used in medical imaging. Use in diagnostic pathology has been more limited but we consider this a logical next frontier. AI/ML is also used to develop complex diagnostics signatures, especially with nextgen cancer diagnostics.

The use of natural language models like GPT could accelerate the understanding, utility, and integration of complex biological and clinical data used to develop diagnostic products and manage patients. For example, GPT could be used to analyze and interpret nextgen sequencing data, helping identify genetic mutations, biomarkers, and other information that could be used to treat cancer. Integration with clinical context from a patient's electronic health record could help clinicians make more informed treatment decisions.

In clinical research services, GPT has several use cases. The technology could be used to help develop clinical trial protocols by analyzing large amounts of data relevant to the clinical condition of interest. It could help screen for potential trial participants; by analyzing electronic health records and other data, GPT could help identify patients who meet eligibility criteria for a particular study. Further, GPT could be used to help monitor adverse events during clinical trials.

In R&D products industries, logical use cases for GPT center on product service and support.

#### **US SMID Biotechnology Sector – Judah Frommer**

While AI/ML driver efforts in small molecule drug discovery are well-known (and [well-financed](#)), post-IRA (perhaps coincidentally) we note creative applications for biologics as well. Within our coverage, Insmed's Deimmunized by Design artificial intelligence platform could potentially generate deimmunized proteins enabling stealthy delivery of viral vectors in gene therapy. More broadly, we see biotech companies leveraging AI as one component of the accelerating trend toward personalized medicine, i.e., adapting treatments to individual patient characteristics. This trend is readily evident in the nascent field of digital therapeutics. For example, EndeavorRx is a prescription digital therapeutic (a video game that requires a prescription) for ADHD developed by Akili that delivers targeted sensory stimuli through a video game experience. Underpinning the product is an advanced closed-loop algorithm that adapts the game in real time to patient progress, thereby aiming to optimize patient engagement and therapeutic effect. AI-driven identification of neurological 'biomarkers' is a key theme within CNS drug development and is driving ample private investment in its own right.

#### **US Managed Care Sector – AJ Rice**

For the managed care industry, the use of AI has come in the form of population health whereby companies utilize the various data sources they have to identify at-risk members. Beyond the more advanced use cases of AI, the industry also utilizes AI to complete more administrative tasks to help reduce the friction between payer, provider, and member via easing processing and helping to ensure claims and payment are processed in an efficient manner. The use of GPT will likely be geared toward ensuring members fully understand their benefit structures while also helping members navigate the healthcare system.

For example, GPT could provide information to a member that explains/helps a member understand what is in-network or out of network, etc. Further, GPT could identify ancillary benefits that may be of use for a member to ensure the member remains healthy. Chatbots like these are already in use by various MCOs and could be enhanced via GPT. Other tasks GPT

could assist with is customer service for a member. GPT could provide service until a live representative is available to assist a member and act as a triage. GPT could also create and respond to prior authorizations while explaining the details a provider must provide in order for a member to receive service and vice versa.

#### **China Healthcare Sector – Jason Liu**

We see that in the near term, AI tools including ChatGPT could potentially benefit online healthcare companies including Ping An Health and other healthtech companies that already are focused on using technology to assist doctors or within the healthcare space. Ping An Health (1833.HK) already uses AI within their telemedicine platform to assist doctors. We see with further refinement of their online healthcare platform with AI tools to increase speed and accuracy, this could continue to help doctors to focus more time on their patients. In the longer term, we expect the whole healthcare space will see significant changes full of disruptions and surprises with AI deeply involved in caring for patients. See prior notes on [ChatGPT](#) and the broader [impacts of AI, blockchain, cloud, and data analytics](#) on healthcare.

## Industrials

Industrials companies have been gradually adopting AI-enabled technologies to improve efficiencies that could result in enhanced revenue and margins. Much of the broader Industrials sector views AI as vital to the ongoing evolution of industrial production. We view ChatGPT as a sophisticated approach to collecting, analyzing and communicating AI-driven insights on the optimization of processes and resource utilization to all participants in the industrials ecosystem – from consumers, manufacturers and suppliers, to shippers and transportation carriers. AI could make manufacturing and supply chain processes simpler and more efficient, provide greater visibility on the end-to-end production process to human users, and identify key business or operational trends in a faster and more effective manner.

### Key ChatGPT Use Cases in Industrials

- **Faster, Better Training:** ChatGPT is capable of developing training materials/curriculums for workers, improving training in enterprises that otherwise offer few materials.
- **Employee Chatbot:** Can function as an internal chatbot used to answer questions for workers across multiple divisions.
- **Customer Support:** ChatGPT can provide a better technical support experience to customers given its chatbot's ability to answer predictively (in a generative manner).
- **Resource Utilization:** Determine and communicate optimal routes and resource utilization to human users on manufacturing and supply chain processes.
- **Enhance Customer Satisfaction:** Identify emerging trends in customer behaviors based on customer questions and interactions.

### AI Use Cases, Beyond ChatGPT

When thinking more broadly about the implications of AI to the industrials sector, we identify the following as key use cases of AI on the industry. Note, some of these use cases will include the training and development of new large language models (LLMs):

- **Improve Machine Reliability:** Predict assembly issues and machine equipment effectiveness with real-time machine learning algorithms and insights, reducing failure rates and minimize production stoppages.
- **Fully Autonomous Equipment:** Future AI may allow for fully autonomous equipment that can operate without human intervention, freeing the workforce to focus on other tasks or outright SG&A reduction.
- **Generative Creativity for Enterprises:** Future AI may be tasked to draw architectural sketches, sample images, and more.
- **Autonomous Vehicle Adoption:** Accelerate the adoption of autonomous vehicles (trucks, locomotives, and eventually planes) to disrupt transportation, improve safety, and lower costs.

## Sector Coverage Implications

### US Aerospace & Defense Sector – Scott Deuschle

**Commercial Aerospace:** While customer service is likely the most obviously application of ChatGPT, capital goods sectors with reduced public customer interaction have perhaps less opportunity to leverage the tool in this way. However, as downstream customers—namely airlines—leverage ChatGPT in this way and increase margins, we see the opportunity for aftermarket suppliers (and aircraft OEMs) to transfer increased airline profit to themselves through price. TDG, HEI, HON, RTX could benefit from the capture of any increase in airline profitability. Separately, AI tools such as GitHub Copilot could be useful in speeding up software development timelines. This could be beneficial to commercial aerospace companies, as

software development drives an increasing portion of the overall development effort for both new aircraft and the systems on those aircraft.

**Defense:** AI tools such as GitHub Copilot could be useful in speeding up software development timelines. This could be beneficial to defense companies, as software development drives a large portion of the overall development effort, particularly for defense electronics companies such as LHX. Misinformation campaigns using simple 'bots' have been a common way for nations to conduct information warfare in the past. LLM tools such as ChatGPT offer a potential new level of sophistication for this type of attack, both in sophistication, scale, and difficulty of detection. Government Services companies such as BAH and LDOS could benefit from deployment of information warfare tools, as well as in developing tools to detect them. AI tools more broadly have a variety of known and potential battlefield applications. AI is already used for intelligence, surveillance, and reconnaissance (ISR) missions by detecting unusual troop movements and the movement of military equipment, but its usage could increase further. BKSY could be a beneficiary of this trend. Additionally, AI can be used to enhance the pace of decision-making in military circumstances, through detection of friend and foe, and—depending on the type of ethical approach taken in the usage of AI—even allow for AI to control lethal autonomous weapons.

### **US Machinery Sector – Jamie Cook**

The farm equipment players including Deere (DE), CNH Industrial (CNHI) and AGCO are using computer vision technology and artificial intelligence for their Precision Agriculture offerings, which help improve yields, reduce input costs, as well as enable more sustainable farming solutions for farmers. We view Deere as most advanced with regard to its Precision Agriculture offerings. For example, artificial intelligence is applied in Deere's See & Spray offering which distinguishes between crops and weeds then selectively sprays weeds only, enabling herbicide savings ranging between 50-70%. Applications extend to other crop care products like fungicides, pesticides and nutrient application. The equipment collects both machine and agronomic data which allows each job to be smarter for the one prior. Deere expects to continue to integrate artificial intelligence into its technology and ultimately achieve full autonomy. DE plans to have a fully autonomous corn and soy production system by 2030 which means spring tillage, planting, spring harvest and fall tillage all done autonomously. Artificial intelligence enabled solutions provide opportunity for a subscription revenue model for Deere. Deere aims to shift revenue model to a pay per acre per use model vs the traditional point of sale model over time, which could help dampen the amplitude of cycles better aligning Deere's business model with the farmer which is highly repetitive. While Deere is a leader in AI enabled Precision Agriculture offerings, CNHI has also bolstered its capabilities via the acquisition of Raven Industries and AGCO continues to invest in its tech stack through organic growth as well as niche M&A and partnerships.

### **Industrial Tech & SMID Multi-Industrial - Guy Hardwick**

We believe there is a use case for ChatGPT particularly in the software business of our Industrial Tech coverage. For examples relating to machine vision & AIDC (automatic identification & data capture) and CGNX, ZBRA, and TDY. The more complex the software the greater the applicability of this natural AI language. Cognex gave an example: reading a bar code – Simple application: reading a black and white, 1D barcode and Complicated application: reading a smudged 2D bar code, inked into metal.

In the first application ChatGPT is likely unnecessary but in the second, deep learning (AI) is necessary. ChatGPT could be used to help program the software to enable the barcode reader to work in the second, more complicated application. We believe the majority of machine vision software is rules based, in Cognex's case at least 80% but the remaining 20% is where ChatGPT could be helpful tool to write the code. For machine vision to grow its applicable uses, it must be able to handle more complex tasks, but at the same time being relatively easy to program (in hours instead of days).

In Robotics we see applications for STRC, which manufactures robots for unstructured environments (outside the factory or warehouse floor). Because its tasks are unstructured,

programming is significantly more complicated than robots for structured tasks (cobots, AMRs). While STRC has not yet replied to our inquiry around ChatGPT, Sarcos recently announced plans for software portal through which customers can update their robot's software for new tasks. ChatGPT could speed up the process of writing the software, particularly as a library of use cases accumulates.

### **US Transportation & Logistics Sector – Ari Rosa**

We view ChatGPT as an AI-driven human interface tool that may potentially deliver 3 key areas of efficiency enhancements for transports: 1) ChatGPT may enhance the optimization of routes and labor/equipment resource utilization, communicating results in a human-friendly manner to all participants along the supply chain. For asset-based transportation carriers such as railroads, truckers and parcel airfreight companies, it could communicate such insights to customers as well as internal employees, improving customer service and freight tracking. For truckers and parcel carriers in particular, such tech could be used to devise the most efficient routes from linehaul to local deliveries, while optimizing freight assembly, path of travel, as well as labor schedules two to three shifts in advance. While AI means different things to different people, we have seen implementation of big data and AI in transports with examples such as global parcel carriers UPS' ORION (On-Road Integrated Optimization and Navigation) program, which has been described as the largest commercial analytics program ever created, being used to design networks and provide real-time updates on route mapping. We have also seen applications for automated freight matching from freight brokers, 2) ChatGPT may analyze real-time data on cargo demand, resource utilization patterns and weather conditions in order to provide enhanced network visibility to shippers and carriers alike, communicating the location, condition, and temperature, etc., of cargo in real-time, 3) ChatGPT may help identify for the carrier/shipper trends on pricing, types of freight demand, and accidents based on what users are querying ChatGPT, much like meta-analyses of Google searches.

### **China Industrial Sector – Analysts: Iris Zheng, Daniel Cui**

While still at an early stage, ChatGPT and AI could potentially help to accelerate the automation adoption and improve manufacturing efficiency in the industrial space by making the coding process for robots easier, assisting in training labor and providing data analytics.

We believe the resulting higher automation penetration bodes well for names such as: Inovance (one of the largest automation component suppliers in China) as it is building up the automation ecosystem with the potential to offer digital / AI / software offerings in the future potentially leveraging AI and ChatGPT. The easier coding of robots should also help lift robot penetration, which is supportive for Estun (leading domestic robot OEM), Leader (leading domestic harmonic reducer supplier for robots) and Shuanghuan (supplies reducers to EVs and robots). We regard Friendess as one of the key beneficiaries of AI/ChatGPT technology as Friendess provides controllers of laser cutting machines, and the product features integrations of CAD, CAM, motion control and sensing features that significantly lowers the bar for operating laser cutting machines in China.

## Consumer Discretionary & Staples

ChatGPT could have a significant impact across the consumer discretionary sector, providing benefits to both retailers and consumers through a variety of different ways including: personalized product recommendations, enhanced customer service, improved marketing, improved product design, and better inventory management.

## Key ChatGPT Use Cases in Consumer Discretionary & Staples

- **Automated Customer Service:** The most direct way any consumer brands can use ChatGPT is to automate the customer service process, which would reduce labor costs and enhance customer experience. The user scenarios include 1) answering customer queries (e.g., what type of product should I choose?), 2) after-sales service, and 3) engaging customers regularly with periodical promotions and loyalty programs.
- **Education / E-Learning:** ChatGPT can help to source and provide tips and classes for nice hobbies and their associated products.
- **Engage a Wider Audience:** Help brands create a formulaic approach to content creation that is more engaging to the customer.

## AI Use Cases, Beyond ChatGPT

When thinking more broadly about the implications of AI to the Consumer sector, we identify the following as key use cases of AI on the industry. Note, some of these use cases will include the training and development of new large language models (LLMs):

- **Improved Customer Targeting:** Use large data set to analyze customers and better segment preferences and behavior leading to more targeted marketing strategies.
- **Supply/Demand Optimization:** Supply chain & inventory management optimization to ensure firms are not losing sales due to inventory challenges.
- **Enhanced Customer Recommendations:** Analyzing consumer buying patterns and recommending products more efficiently thereby reducing unnecessary advertising spend.
- **Customized and automated manufacturing:** The customization of shoes based on each buyer's unique foot shape could be the next big breakthrough in shoe making technology, as it could provide extra comfort and functionality to the wearer. We believe such breakthrough is dependent on finding a way to quickly and cost-efficiently to produce small volume of shoes, which is admittedly difficult at the moment.

## Sector Coverage Implications

### US Beverages & HPC Sector – Kaumil Gajrawala

The first use cases of ChatGPT and AI technology are already being implemented. Coca-Cola signed a deal with Bain & Co/OpenAI in 2020 to leverage natural language processing to better connect with customers. More recently (February) the first consumer deal of its kind was signed between Coca-Cola and OpenAI/ChatGPT/Dall-E. Other CPG companies have commented on their involvement to similar partnerships, but the details have been sparse.

### US Restaurants Sector – Lauren Silberman

Restaurants are increasingly leveraging AI across different aspects of the business, from inventory & labor management to customer service. We see opportunities for ChatGPT to be used to increase engagement, enhance marketing strategies, improve speed of service, and simplify operations. ChatGPT can be used to generate personalized responses to customer inquiries or feedback, respond to customer reviews and comments on social media, and assist in handling reservations. From a marketing perspective, ChatGPT can be leveraged to create content for social media (or provide ideas on marketing content), leverage exogenous factors

(e.g., weather) to market specific products, create more appealing descriptions of menus as well as generate menu item photos (additional AI products). ChatGPT can also be leveraged to generate personalized promotions to specific customer cohorts. Restaurants can integrate ChatGPT with ordering products (e.g., voice-based drive-thru, digital app, kiosks) to take orders and respond to frequently asked questions.

**China Sportswear & Beauty – Analyst: Jessie Xu**

ChatGPT could marginally benefit sportswear retailer (Topsports) and brands (Li Ning, Anta, Xtep) by bringing better and more cost-effective customer engagement. However, the employment of ChatGPT in customer service might take time; hence, we do not expect this new technology to cause extra operating expenses or move the needle in the short term. We see ChatGPT's potential mainly in marketing and client engagement. ChatGPT can learn to identify the latest consumer preference trend and help companies with better marketing direction and execution. A broader base of consumer feedback can also help with new product design, to meet more diversified consumer needs.

## Real Estate

There are a variety of use cases for ChatGPT in the real estate industry: ChatGPT can improve the efficiency, accuracy and overall customer experience of the real estate industry. Additionally, by leveraging ChatGPT real estate professionals can improve their operations by providing better customer service, and make more informed decisions by analyzing vast amounts of data and providing valuable insights.

### Key ChatGPT Use Cases in Real Estate include

- **Chatbot for Customer Service:** ChatGPT can be used to develop chatbots that interact with customers and address questions related to real estate properties (size, price, location and features). ChatGPT can also assist customers by providing personalized recommendations.
- **Predictive Analytics:** ChatGPT can analyze large amounts of data to provide insights into real estate trends (property prices, rental rates, demand indicators). It can assist a variety of different types of real estate professionals buy, sell or rent properties.

## AI Use Cases, Beyond ChatGPT

When thinking more broadly about the implications of AI to the Real estate sector, we identify the following as key use cases of AI on the industry. Note, some of these use cases will include the training and development of new large language models (LLMs):

- **Consumer Filtering:** Can filter consumers based on credit score, ownership history and provide early warning of bad debts.
- **Market Analysis:** Create market analysis (crime rates, probability of natural disasters, etc.) and forecasts property values and investment opportunities.
- **Virtual Tours:** Create compelling virtual tours that can be proven to lead to more onsite visits.
- **Image and Video Analysis:** ChatGPT can analyze videos and images of properties and extract key information (number of rooms, types of flooring, and condition of the property). This can also assist real estate agents by creating more detailed and accurate property listings.

## Sector Coverage Implications

### US REITs – Analyst: Tayo Okusanya

Use of high-level AI-based text bots for general business is becoming more prevalent in the world of commercial real estate especially in areas such as sales and leasing, but the high level of human interaction in commercial real estate suggests that AI is unlikely to be very disruptive to the industry given the level of local specialization often needed. That said, generative AI, like GPT, can certainly be powerful as a time saving device in many situations. We can also see a scenario where AI can create operational efficiencies by helping to reduce staffing in areas like operations via automation of certain routine processes. AI can likely also be used to optimize scheduling in areas like repairs and maintenance as well as asset management and also help optimize purchasing decisions which can have a meaningful impact on the bottom line. Where ChatGPT falls short the most is that it specifically notes that it has very limited information past 2021 which is particularly concerning for CRE applications given that real estate goes through cycles regarding demand/supply trends and capital markets (e.g., the real estate market today is very different vs. 2019 given the strong effect from rising interest rates and inflation). Thus there is concern that real time analysis or even forecasting via an AI based software like ChatGPT could produce erroneous results.

What is ChatGPT useful for in CRE? Some residential real estate brokers have [claimed use](#) of ChatGPT, citing summarizing functions to create property listings, however, utilization across commercial real estate has been somewhat limited given transactions tend to be more specialized and communications tend to be more personalized given the much larger dollars at

stake relative to residential real estate. Similarly, internal and external corporate communications can be written quickly and with little effort. Other areas often mentioned where ChatGPT can create efficiencies include drawing up legal contracts and lease agreements. Commercial real estate brokers also seem excited about the ability to connect with more clients globally using native language in each region.

General information to feed strategic CRE investment decisions could potentially be helpful as well – [one recent report shared](#) a ChatGPT-generated report based on a prompt for a cost/benefit analysis of moving a business to an “optimal market,” where the bot outlined cost of living, tax considerations, energy costs, and labor costs among other items to consider.

What ChatGPT isn’t as helpful for in CRE is specific/bespoke applications. Given the timeliness of market data and how quickly the macro environment is evolving, requesting information about transaction markets, the economic backdrop, and ideal target markets for development may provide minimal value given the very limited information provided past 2021. Companies wanting feedback on proprietary or sensitive information should also be wary, as sharing the information with a bot may not guarantee it is responsibly and securely handled.

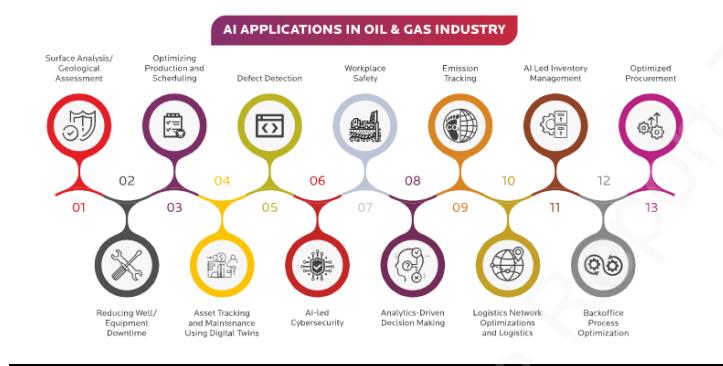
## Energy

The energy industry includes the upstream (oil & gas exploration production), midstream (pipelines), and downstream (refining/marketing) sectors with integrated oil producers operating across the entire supply chain. There is a vast opportunity for AI within the energy industry given the complexity of energy supply chains (labor, equipment, geographies, etc.) and vast raw data sets involved (geological, well-level data, reservoir data, weather data, etc.) which AI can help optimize—given this backdrop the industry has already begun to identify and capitalize on the AI potential more broadly. As it relates to ChatGPT specifically the use cases already identified have been more narrow in scope, with use cases primarily functioning around the use of ChatGPT as a chatbot to support in-field workers and technicians—in this area we note significant opportunity within the offshore oil & gas industry where operations require substantial offsite support (i.e., a dedicated onshore support staff for every offshore rig and well being drilled).

### Key ChatGPT use cases in Energy include

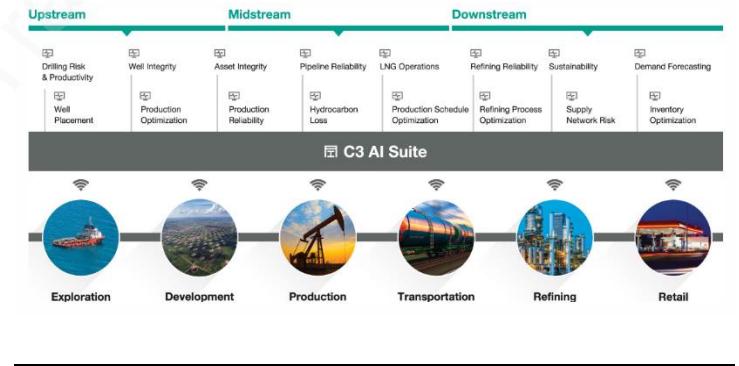
- On-Demand Field Support:** ChatGPT can be used to provide on-demand field support for both in-field oil and gas workers along with back-office support functions.
- Code Testing:** ChatGPT can be used (and is currently being used by Devon Energy (DVN) for this purpose) to support in-field technicians response for management of oilfield equipment by testing code prior to deployment.

Figure 109: AI Applications in Energy Industry



Source: Birlasoft

Figure 110: AI Applications by Energy Subsector

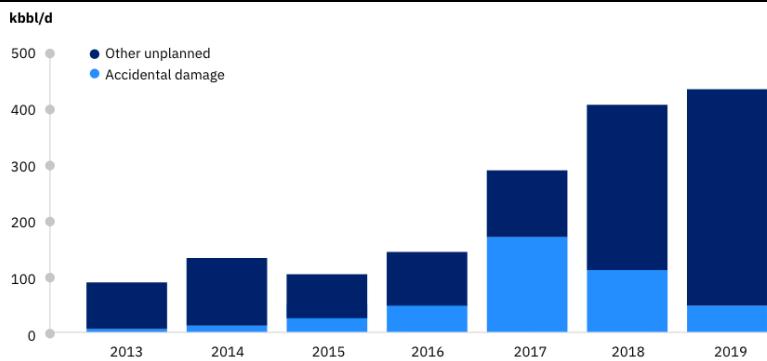


Source: C3.ai

### AI Use Cases, Beyond ChatGPT

When thinking more broadly about the implications of AI on the energy sector, we identify the following as key use cases of AI on the Energy industry. Note, some of these use cases will include the training and development of new large language models (LLMs):

- Seismic Analysis/Geological Assessment:** The ability to analyze geological data (surface analysis, seismic data, etc.) to improve the hydrocarbon discovery process. This should help identify the highest yielding wells which should reduce the need for additional drilling.
- Predictive Maintenance:** AI can assist in identifying maintenance issues before they happen which should eliminate the need for rig downtime due to maintenance (which can cost up to several million per rig per day in some cases).
- Defect Detection:** AI can help validate production quality with the ability to use deep learning with video streams for pattern recognition to help pre-empt issues related to defective equipment (i.e., a faulty pipe, gauge, etc.). According to Rystad up to nearly half a million barrels of oil per day were lost to downtime (both unplanned downtime and accidental damage) in 2019.

**Figure 111: Barrels of Oil Lost to Daily Downtime**

Source: Rystad Energy

- **Production Scheduling:** By taking into account dynamic weather data, resource limitations and scheduling/availability of operators—AI helps optimize production related activities particularly when dealing with interdependent activities (i.e., connected drilling and platform installations offshore).
- **Emission Tracking:** Oil producers are using AI to keep track of carbon emissions that are released throughout the energy lifecycle (drilling, production, and eventual shut-in). Exploration & production companies are already using AI to optimize carbon storage for enhanced oil recovery (EOR).
- **Logistics Optimization:** AI is being used in logistics across the energy supply chain with midstream (pipelines) using AI for planning & execution (optimal route selection) and refiners using AI within their logistics network for optimal blending, demand forecasting, and price estimating.

## Utilities

In the Utilities sector, ChatGPT can enable a more streamlined customer experience, help customers with bill inquiries, recommend new load balancing tools, and assist with energy efficiency deployment. As it relates to billing program inquiries, some utilities in the past have seen significant consequences from billing system errors and slow customer service response times, and we believe ChatGPT could potentially assist in this regard, though utility executives may be hesitant to adopt unproven technologies right away. On the operations side, complex AI is already being adapted with new software technologies providing generators more timely information on market conditions, forecasting and assisting with load balancing.

### Key ChatGPT use cases in Utilities include

- **Customer service:** Notification to clients of outages, billing inquiries, energy consumption recommendation based on clients' personal usage, etc.
- **Investor relations:** ChatGPT can answer basic questions posed by investors, freeing up time for investor relations personnel.

## AI Use Cases, Beyond ChatGPT

When thinking more broadly about the implications of AI on the utilities sector, we identify the following as key use cases on the industry. Note, some of these use cases will include the training and development of new large language models (LLMs):

- **Facility Operations:** Future AI may be able to take over operation of facility process and respond to internal and external factors.
- **More Accurate Forecasting:** Provide demand and supply predictions for energy grid load balancing and cost management which could in turn lower customer utility bills, and ease burdens on the grid.
- **Fraud Detection:** Provide improved energy theft prevention in developing countries using human usage patterns, payment history and other customer data.

## Sector Coverage Implications

### US Utilities Sector – Analyst: Nick Campanella

From an investor standpoint, we see the ChatGPT function as potentially helpful researching rate cases and other policy-driven matters which affect the fundamental outlooks of utilities. State level rate case testimony work can be cumbersome, but testimonies from state commissions can lead to significant alpha generation in determining where a rate case is heading. ChatGPT could potentially assist in expediting read-throughs for documents which are typically hundreds of pages of text to inform on regulatory catalysts for stocks impacted.

### China Utilities Sector – Analyst: Alex Liu/Gary Zhou

Within the sector, Chinese city-gas distributors and renewable power operators are more likely to see application of ChatGPT or more complex AI in their daily operations. On the consumer service side, we believe ChatGPT can possibly help improve the service quality of gas distributors and save labor costs. Specifically, China Gas Holdings has also signed framework agreement with Huawei since 2019, to enhance its customer relationship management (CRM) and enterprise resource planning (ERP) systems. For renewable operators, Longyuan power has been using unmanned aerial vehicle and AI monitoring and diagnosis systems for the maintenance of some of its wind farm projects.

## Materials

The Materials industry has the ability to leverage ChatGPT by providing insights into production processes, supply chain optimization, and safety management. Overall, ChatGPT's ability for natural language processes can help the Materials industry leverage AI to reduce costs, improve product quality and deliver better customer service.

### Key ChatGPT use cases in Materials include

- **Supply Chain Optimization:** ChatGPT can assist in the optimization of supply chains
- **Regulatory Compliance:** ChatGPT can help materials companies stay up to date with regard to the latest regulatory requirements and guidelines.

## AI Use Cases, Beyond ChatGPT

When thinking more broadly about the implications of AI on the materials sector, we identify the following as key use cases on the industry. Note, some of these use cases will include the training and development of new large language models (LLMs):

- **Data Collection:** Collect market data to forecast raw material pricing.
- **Consumption Analytics:** Consumption analytics to extract the data and create energy profiles from household appliances, smart meters, and other sensors.

## Sector Coverage Implications

### **US Chemicals Sector – Analyst: John Roberts**

Chemical stocks have not yet mentioned any potential impacts from ChatGPT, but several firms have highlighted their use of AI as part of their normal work processes. We believe that ChatGPT could be the next evolution of those processes, and assume most chemical companies are targeting the aforementioned examples, but many are too routine to drive press releases. Some product areas are simply benefitting from higher growth from customer use of A.I. (e.g., DuPont & Entegris semiconductor chemicals). Some chemical firms are using A.I. for more complex/ customized new product discovery (e.g., Ashland anti-dandruff bio functional, FMC micro-peptide biopesticides, IFF personalized nutrition). Others also use A.I. to reduce waste in applying chemicals to complex shapes (e.g., PPG reducing overspray in auto OEM paint). Companies are also using A.I. in general ways, such as delivery route optimization (e.g., Univar), process control equipment (e.g., Tronox, Westlake), and to deliver customized “nudges” to employees to improve engagement, retention & inclusion (e.g., Ecolab).

## Communications Industry

ChatGPT has the potential to have a significant impact on Communications Services as it can help companies improve their customer service, sales and marketing and networking management. ChatGPT's functionality as a chatbot offers the ability to assist customers in troubleshooting common issues, provide personalized support and can also assist with lead generation. Overall, ChatGPT can help the communications industry leverage the power of AI to deliver better products and services to customers, while improving efficiency and lowering costs.

### Key ChatGPT Use Cases in Communications include

- **Content Recommendation:** Improve current recommendation performance, leading to more user engagement and faster recommendation improvement.
- **Customer Service:** offload a greater fraction of customer service tasks to AI, freeing up human agents for more complex issues.
- **Bill Negotiation:** negotiate on customers' behalf for lower recurring monthly rates, possibly in exchange for a bounty of a portion of the savings.

### AI Use Cases, Beyond ChatGPT

- **Collaborative Design/Content Generation:** Create and modify music, image, or AV content assets collaboratively with a human designer or independently.
- **Content Extension/Interactive:** Eventually, interactive experiences where an established IP is extended to new environments within a set of guardrails or parameters established by the IP owner.
- **Network Engineering:** Extension of current telecom AI to ever more advanced capacity planning, network optimization, and predictive analytics for areas like component failure, allowing telecom operators to grow even leaner and more efficient.

### Sector Coverage Implications

#### **US Telecom/Cable/Media Sector – Analyst: Doug Mitchellson**

For those for whom it is dubious to imagine a large language model like Sydney or Bard getting to the point where it can reliably answer questions of fact, it may be easier to imagine the technology being able to generate an appealing fiction or set of creative options when no assertion of fact is involved. Indeed, we feel that creativity, and assisting human creativity, are particular strengths of generative AI, and that gives the technology profound implications for the media sector.

Perhaps the most immediate application is the use of generative AI to change how users access content. Content recommendation and discovery are surprisingly difficult problems for streamers and distributors, despite how easily they are described and how intuitively a user can tell a good recommendation from a bad one. Spotify's [AI DJ in your pocket](#) announced last week is the first of what we think will be many applications of generative AI to content recommendation. For users, these should provide a much more natural avenue to provide feedback – and this should improve the data that recommendation engines are trained on, leading to a flywheel effect. Another area we see near-term AI potential is content creation efficiencies, perhaps streamlined AI-assisted editing and processing, lessening the need for expensive human producers, or improving logistics.

Less immediate, but perhaps more transformative, is AI-assisted design and then eventually full content creation. In video games, Activision Blizzard already uses generative AI to take a cosmetic or art asset for one character and adjust it for use on others (e.g., a human artist may make a helmet for one character and the AI adjusts it to fit different characters without looking out of place), and there are parallels in [AI generated music](#) also currently being explored by ATVI. From these current applications, it seems like a relatively small iterative step toward

collaborative design, where a human designer gives prompts and a generative AI refines its guesses until the pair lands on an asset or effect matching the artist's intentions.

In terms of asset quality, generative AI output may come to music, podcasts and mobile games first, then eventually AAA (PC/console) games, only reaching tv and film quality last. Even when technical parity is reached, we believe that key senior creative talent may be slow to adopt the technology, believing that principal photography and recordings of real humans and sets is more emotive than generated content. With the eventual proliferation of generative AI, one can even imagine a world in which fully human-created content carries an imprimatur of quality and authenticity. The dynamic is perhaps similar to the CD vs vinyl debate (those who aren't familiar can ask ChatGPT for a primer).

As generative AI is further refined and compute power continues to grow far beyond current capabilities, there is likely potential for much more powerful generative AI to make users a rich interactive video experience – but the key limiting factor after technological feasibility will likely be how much creative control IP owners are willing to give up to users and AI networks. One need only imagine the richness and value of a potential world of adventures and misadventures alongside Spiderman or Bart Simpson – but also how many guardrails Disney and Sony would need to be confident enough to greenlight that potentially amazing (and lucrative) experience.

As for specific players in the sector, we would expect embracing AI to deliver greater content creation and distribution efficiencies for all, but improving recommendation engines favors those with the best content and the largest libraries. Easier-to-generate audio content will likely represent the first test in the coming years of the potential popularity of AI-generated content and the potential ultimate threat posed by AI content creation to media incumbents. While AI may very well author one-off hits, we expect humans will continue to desire content created by other humans, imperfections and all.

#### **US Telecom/Cable/Media Sector – Analyst: Doug Mitchelson and Grant Joslin**

The telecom sector's proximity to tech means that the sector is already accustomed to working with AI and there are relatively extensive deployments of AI in telecom engineering today: [capacity planning](#), [reducing tower climbs](#), [tuning radio angles and power levels](#), [electronic component failure prediction](#), and (as in other sectors) [collaborative programming](#). These applications are predominantly evolutionary, not revolutionary: the kinds of applications that may underwrite years of continued -MSD% service cost declines, but nothing has emerged that upends the business model of the telecoms or dramatically improves the economics.

We also considered the possible use cases for generative AI, especially large language models like ChatGPT. The industry already relies on chatbots and IVR to offload a fraction of customer service workloads to AI, and large telecom providers like Comcast have long invested in widening the set of tasks which these AI can serve. A large language model like ChatGPT could provide a more positive customer service experience on these channels, answer a wider set of questions, and better detect when a human customer service agent should be referred a ticket.

On the other side of the phone or IM window, we have noted the ascendance of bill negotiation services like Billshark, Trim, and Rocket Money – services which reach out to telecom firms and other recurring service providers and ask for discounts on a customer's behalf. The legal AI firm DoNotPay has launched a [bill negotiation chatbot](#) and we expect that it may garner fast followers as the bill negotiation application is a relatively straightforward one. Indeed, soon a sizable fraction of 'customer' service interactions may be between two chatbots!

The companies with a combination of high pricing / high competition and those with the most complicated pricing structures may have the most at risk from chatbot bill negotiation. It could also end wireline/satellite providers' pairing of low teaser rates stepping up to much higher full ratecard rates (predominantly in pay TV and small/midsize cable broadband), and in expensive wireless markets like the US it may challenge providers' use of large device subsidies to acquire subscribers for a longer-term promo period with the hope of retaining them afterwards.

There remains a persistent data demand argument, which suggests that the more AI takes off, the more value data has and the more important transporting data becomes. We do not entirely reject this line of thinking, but we have historically been skeptical of vague data demand stories leading to discrete benefits for telecoms.

Training an AI is somewhat data intensive and very compute intensive, but not necessarily transport intensive unless the training data does not all exist at the same location as the training compute resources. The trained model that is then propagated to devices and put into live service is relatively small in filesize and not demanding of timing. As with edge computing, we suspect that the likely beneficiaries of the AI revolution will be the AI companies, and to a lesser extent their telecommunications and cloud service providers, while the overall benefit to all telecom firms from slightly more traffic seems marginal at best.

So, in summary, the benefits of generative AI and any coming AI revolution for the telecom sector look to be evolutionary, not revolutionary, in nature, but also offset in part with consumers becoming smarter shoppers. As far as stock implications, we believe that greatest benefit will accrue for the largest players with the most capacity to invest (Comcast, Charter and the world's largest wireless operators). To the extent that these benefits are meaningful, AI may be yet another driver of M&A for subscale players.

## Governments

The Government sector is comprised of multiple levels of Government (municipal, state, and federal), along with Public Goods and Services such as the military, law enforcement, public transportation, and infrastructure to name a few. Although ChatGPT is most likely to be used by the branches of government for civilian interactions/service requests, other branches could derive advantages from using the technology. As an example, individuals could interact with a government chatbot asking for information on how to qualify for welfare assistance and a chatbot using ChatGPT would return relevant links along with a step-by-step guide to accessing relevant aid. This iteration is currently being implemented by Bhashini, a chatbot created by India's Ministry of Electronics and IT using answers generated by ChatGPT models (see Figure 112). Below, we highlight direct use cases of ChatGPT for the Government sector, followed by a section proposing other AI use cases to be expected as well as coverage implications that we have gathered from industry references, experts, and coverage company mgmt. teams.

**Figure 112: India's Ministry of Electronics and IT Has Begun Implementing ChatGPT into Search and Response Queries**



Source: Times of India.

**Figure 113: Singapore Has Partnered with MSFT to Incorporate Pair into Their Word Suit Leveraging ChatGPT Functionality**



Source: "Hack for Public Good 2023 Demo Day" ([YouTube Video](#)).

## Key ChatGPT use cases in Government include

- Public Services (Figure 112):** The ability for ChatGPT to be integrated into government websites and provide quick and accurate responses to civilian questions could be explored as a cost cutting/productivity maximizing tool.
- Multilanguage Assistance (Figure 112):** With increases in immigration, offering government services in multiple languages will be a benefit to citizens and can increase the perceived openness in accommodating newcomers helping grow a countries output.
- Assist in Speech & Law/Bill Creation (Figure 113):** Singapore's government employees will soon have access to Pair, a ChatGPT based tool that will assist them in

researching and writing reports, speeches, and potentially creating laws/bills through Microsoft Word.

## AI Use Cases, Beyond ChatGPT

- **City Design/Planning:** Assisting city designers/planner with AI driven analysis of local areas suggesting different public services that should be provided along with city/road design to improve traffic flow (this provides interesting ESG angles should a reduction in traffic lead to reduced emissions).
- **Public Safety:** AI can be used to assist law enforcement officers in public safety responses such as natural disasters (tornados, floods, earthquakes, etc.) and other public endangerment events by connecting with infrastructure to identify and contain threats to human safety.
- **Weather Predictions:** AI models will be able to better predict weather related events which can be used by governments to better inform public responses such as early evacuation thereby saving money and lives.
- **Decision Optimization:** Optimize decision making by politicians and law makers through the analysis of large data sets and provide clear recommendations which can better serve the public.
- **Intelligence Analysis:** Currently, intelligence agencies are utilizing AI to identify and detect threats, however, advancements in the underlying technologies could have a greater ability to prevent attacks/incidents from occurring.

# Supply chain for AI / Chat GPT

Our global tech team summarizes the supply chain implications and company level beneficiaries from the rapid uptake of Chat GPT and its potential to further accelerate adoption for the AI ecosystem. Data center has been one of the fastest growing areas in the tech space and albeit moderating with the macro is still relatively outgrowing many of the consumer areas now facing a post COVID-19 hangover. While the new Chat GPT workloads are not yet offsetting macro to drive upside in supply chain orders, we do view concentrated bets leveraged to acceleration of AI having ability to show outgrowth through the industry slowdown. In the mid-term, the uptake of AI services and its industry use cases for revenue generation and cost / capex efficiencies can feed to feed a new cycle of hardware and semiconductor to maintain innovation and advances.

## **AI compute and memory to benefit within the semiconductor sector**

AI training and inference are compute intensive tasks that should continue to drive semiconductor advances for compute, storage and transmission of data. The data center compute TAM including accelerators have maintained a 14% CAGR from 2019-24E, with NVIDIA's data center growth at a 50% CAGR and Marvell at 30% CAGR, far outpacing the CPU server growth at a 2% CAGR. An annual penetration increase of 1-2pts of AI accelerated servers from the 8% in 2022 would maintain a 30-35% CAGR for accelerators through 2027.

For stocks, primary beneficiary is NVIDIA with over 90% of compute share. We also see TSMC with leverage doubling its contribution from HPC to over 40% of sales the past decade and from 20% to 60% share of compute, now with leverage across leading chip customers promoting CPU, GPU/AI, FPGA and ASIC. Elsewhere in semiconductors, AI has potential to improve prospects for server memory for the memory leaders (Samsung, Hynix and Micron), now crossing over mobile at 40% of industry bits, power management into AI boards (MPWR, Infineon and STM), network switch ICs and ASICs (Marvell) and IC design services (Alchip).

## **Hardware chain to benefit from cloud growth and higher specs**

IDC projects AI servers will grow at a 21% revenue CAGR from 2021-26 vs. 8% CAGR for the total server market, driving AI servers to grow from 18% of server industry revenue in 2022 to 27% of server industry revenue in 2026. The hardware chain should benefit from a richer mix of servers for AI from higher value specs and more thermal design challenges to increase value add for the hardware chain, power supply makers and high-end substrates in Japan (Ibiden, Shinko). We note benefits across brands (Lenovo, Gigabyte), ODMs (Accton, Quanta, Wiwynn, Inventec), connectors (Lotes), testing (Chroma), and high-speed interface (Parade). Power supply maker Delta is also seeing rising value supplying a new data center architecture that can better address the rising energy consumption. In China tech, our top picks include server maker Inspur with 30% contribution from AI servers, Wus which is key supplier to US HPC customers, Innolight with 20% share in optical modules and lead supplier to the major US hyperscalers and Montage which has over 80% of profit from server DRAM interface and companion chips.

## **Additional supply chain opportunities**

The supply chain beneficiaries from advances in compute intensity will be a good driver for leading edge silicon, which is now replacing mobile as a key driver for innovation both on advanced manufacturing and high-end packaging integration. We see beneficiaries on advanced SPE front-end and packaging equipment (ASML, ASMI, Besi). We would also highlight on-going geographical shifts in the supply chain which are creating opportunities for ASEAN tech in the data center build-out (Delta Thailand, Inari Amertron).

## Semiconductors

### Compute TAM to be lifted by an inflection in AI use cases

We view the data center semiconductor market can stay a faster growing area within the industry particularly now as conversational assistants powered by AI democratize usage and accelerate new use cases for driving more revenue and lowering costs through efficiency gains across industries. We noted in Cloud Computing: The Next Frontier in 2020 that AI Data Analytics is the most important industry segment as it provides the mechanism to monetize data and creates a virtuous cycle – the more data one can monetize through analytics, the more valuable data becomes. The more valuable data becomes, the more one wants to create (social media platforms feeding more addictive videos to its users), which leads to more storage, networking and compute demand. AI is an important advance in opening up the pool of data to be analyzed and is a silicon based technology utilizing every incremental transistor it is given to analyze and create useful responses from data, continuing the expansion of compute intensity.

The use case for AI driving new revenue streams and cost reductions for industries noted in the industry use case section should continue to feed a virtuous cycle for semiconductor compute growth providing this processing, storage and transmission of data. According to Gartner's semi industry forecast, data center silicon would maintain a +17% CAGR from 2023-26 to US\$115bn, outpacing 10% CAGR for the semiconductor industry, with highest growth in memory, CPU, wired and GPU.

**Chris Caso**  
**Liz Pate**  
**Nicholas Welsh-Lehmann**

**Randy Abrams**  
**Haas Liu**  
**Angela Dai**

**Figure 114: Data Center Semi '23-26 CAGR at 17% vs. 10% for the Semi Industry**

Data Center Semis US\$mn	2020	2021	2022	2023	2024	2025	2026	20-23	23-26	% of 23
DRAM	\$15,549	\$21,319	\$20,333	\$18,160	\$30,777	\$41,475	\$33,196	5%	22%	25%
NAND	\$8,713	\$12,615	\$12,917	\$9,790	\$15,138	\$21,361	\$26,804	4%	40%	14%
CPU	\$18,563	\$19,444	\$20,889	\$20,520	\$21,778	\$23,147	\$24,013	3%	5%	29%
Wired Connectivity	\$5,039	\$6,339	\$8,347	\$8,464	\$9,145	\$10,015	\$10,854	19%	9%	12%
GPU	\$3,014	\$5,298	\$5,989	\$6,457	\$7,594	\$8,599	\$9,699	29%	15%	9%
ASIC	\$3,370	\$3,607	\$3,883	\$3,462	\$4,205	\$4,643	\$5,127	1%	14%	5%
Analog/Discretes	\$1,519	\$1,807	\$2,374	\$2,124	\$2,332	\$2,558	\$2,803	12%	10%	3%
FPGA	\$435	\$530	\$666	\$723	\$834	\$972	\$1,101	18%	15%	1%
Opto	\$676	\$737	\$838	\$822	\$881	\$1,040	\$1,081	7%	10%	1%
Memory Other	\$314	\$545	\$676	\$592	\$568	\$481	\$556	24%	-2%	1%
Other ICs	\$110	\$128	\$151	\$145	\$158	\$169	\$180	10%	7%	0%
<b>Total Data Center</b>	<b>\$57,302</b>	<b>\$72,368</b>	<b>\$77,063</b>	<b>\$71,260</b>	<b>\$93,411</b>	<b>\$114,460</b>	<b>\$115,413</b>	<b>8%</b>	<b>17%</b>	<b>100%</b>
<b>YoY Growth</b>	<b>26%</b>	<b>6%</b>	<b>-8%</b>	<b>31%</b>	<b>23%</b>	<b>1%</b>				
<b>% of industry</b>	<b>12%</b>	<b>12%</b>	<b>13%</b>	<b>13%</b>	<b>14%</b>	<b>16%</b>	<b>15%</b>			
<b>Total Semiconductors</b>	<b>\$470,899</b>	<b>\$594,952</b>	<b>\$601,694</b>	<b>\$562,712</b>	<b>\$654,326</b>	<b>\$727,126</b>	<b>\$753,667</b>	<b>6%</b>	<b>10%</b>	
<b>YoY Growth</b>	<b>26%</b>	<b>1%</b>	<b>-6%</b>	<b>16%</b>	<b>11%</b>	<b>4%</b>				

Source: Gartner, December 2022

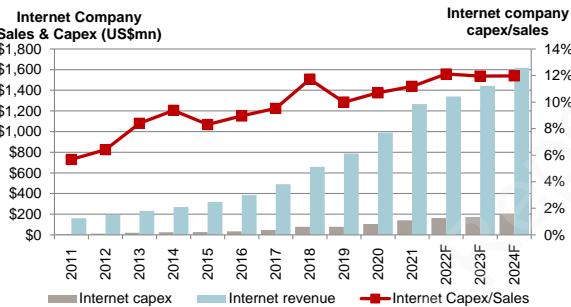
The AI compute will largely resonate in the cloud with GPU time for training and inference offered at scale by the major cloud providers. Our hyperscaler tracker projects moderation due to macro slowdown from +31%/+15% YoY in 2021-22 to +7% YoY in 2023 before picking back up to +12% YoY in 2024 after low hanging fruit optimizations run their course.

**Figure 115: CS Hyperscale Capex Projected at +7%/+12% for 2023-24**

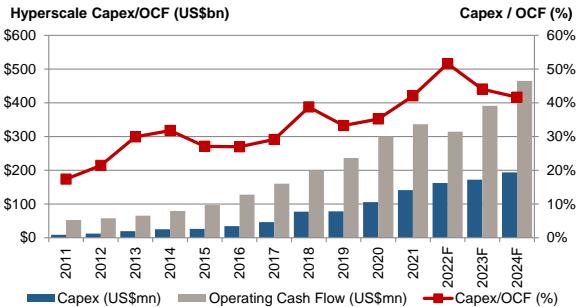
Top 7 Hyperscale Capex	2016	2017	2018	2019	2020	2021	2022F	2023F	2024F
Facebook	\$4,491	\$6,733	\$13,915	\$15,369	\$15,115	\$18,567	\$31,431	\$31,500	\$34,650
Google	\$10,212	\$13,184	\$25,139	\$23,548	\$22,281	\$24,644	\$31,485	\$31,485	\$34,634
Amazon	\$6,737	\$11,955	\$13,426	\$16,861	\$40,140	\$61,053	\$63,645	\$70,010	\$79,811
Microsoft	\$9,114	\$8,696	\$14,223	\$13,546	\$17,592	\$23,216	\$24,768	\$27,245	\$31,059
Baidu	\$630	\$707	\$1,322	\$936	\$817	\$1,689	\$1,400	\$1,700	\$1,900
Alibaba	\$1,495	\$3,467	\$5,534	\$3,809	\$5,293	\$7,179	\$7,129	\$7,805	\$8,664
Tencent	\$1,822	\$2,020	\$3,648	\$4,653	\$4,607	\$5,182	\$2,400	\$2,640	\$2,957
<b>Total 7 Hyperscale capex (US\$mn)</b>	<b>\$34,500</b>	<b>\$46,763</b>	<b>\$77,207</b>	<b>\$78,722</b>	<b>\$105,845</b>	<b>\$141,530</b>	<b>\$162,258</b>	<b>\$172,384</b>	<b>\$193,674</b>
<b>YoY Growth</b>	<b>30.1%</b>	<b>35.5%</b>	<b>65.1%</b>	<b>2.0%</b>	<b>34.5%</b>	<b>33.7%</b>	<b>14.6%</b>	<b>6.2%</b>	<b>12.3%</b>
<b>Capex/Sales</b>	<b>9.0%</b>	<b>9.5%</b>	<b>11.7%</b>	<b>10.0%</b>	<b>10.7%</b>	<b>11.2%</b>	<b>12.1%</b>	<b>12.0%</b>	<b>12.0%</b>
Apple	\$12,962	\$12,121	\$13,858	\$9,247	\$8,702	\$10,388	\$11,692	\$12,321	\$12,921
IBM	\$4,150	\$3,773	\$3,895	\$2,371	\$3,043	\$2,381	\$1,933	\$2,198	\$2,284
eBay	\$626	\$666	\$651	\$538	\$488	\$444	\$449	\$453	\$474
Paypal	\$669	\$667	\$823	\$704	\$866	\$908	\$706	\$936	\$1,051
Oracle	\$1,604	\$2,037	\$1,468	\$1,591	\$1,833	\$3,118	\$6,678	\$7,941	\$8,500
SAP	\$1,105	\$1,443	\$1,630	\$899	\$991	\$965	\$884	\$1,110	\$1,190
Twitter	\$219	\$161	\$484	\$541	\$873	\$1,011	\$700	\$800	\$875
Salesforce	\$388	\$540	\$566	\$643	\$710	\$717	\$788	\$821	\$889
Mercadolibre	\$77	\$75	\$90	\$141	\$254	\$572	\$460	\$546	\$711
<b>2nd Tier Hyperscale capex (US\$mn)</b>	<b>\$22,487</b>	<b>\$21,483</b>	<b>\$23,465</b>	<b>\$16,675</b>	<b>\$17,761</b>	<b>\$20,505</b>	<b>\$24,290</b>	<b>\$27,126</b>	<b>\$28,895</b>
<b>YoY Growth</b>	<b>0.6%</b>	<b>-4.5%</b>	<b>9.2%</b>	<b>-28.9%</b>	<b>6.5%</b>	<b>15.5%</b>	<b>18.5%</b>	<b>11.7%</b>	<b>6.5%</b>
<b>QoQ Growth</b>									
<b>Total Hyperscale capex (US\$mn)</b>	<b>\$56,987</b>	<b>\$68,246</b>	<b>\$100,673</b>	<b>\$95,397</b>	<b>\$123,606</b>	<b>\$162,035</b>	<b>\$186,548</b>	<b>\$199,510</b>	<b>\$222,569</b>
<b>YoY Growth</b>	<b>16.7%</b>	<b>19.8%</b>	<b>47.5%</b>	<b>-5.2%</b>	<b>29.6%</b>	<b>31.1%</b>	<b>15.1%</b>	<b>6.9%</b>	<b>11.6%</b>
<b>QoQ Growth</b>									

Source: Gartner, December 2022

That spend stills keep capex/sales at a sustainable 12% capex/sales ratio and enabling solid FCF with reinvestment rates in capex at about 60% of operating cash flow for the hyperscalers.

**Figure 116: Hyperscale Capex/Sales at 12% Is Sustainable**

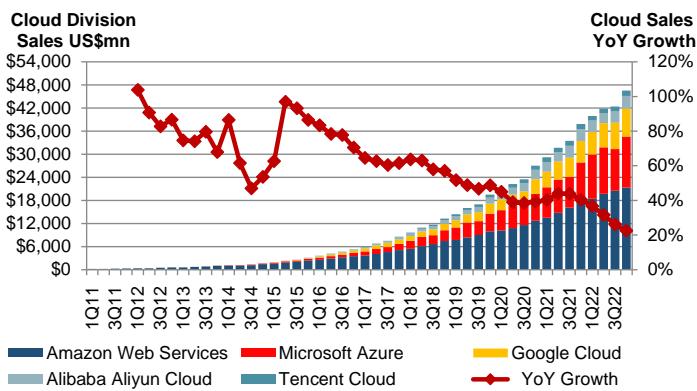
Source: Company Data, Credit Suisse Estimates

**Figure 117: Hyperscalers Re-Invest 40% of OCF in Capex**

Source: Company Data, Credit Suisse Estimates

The hyperscalers are noting optimizations for customers doing more with less in the downturn, optimizing hardware spend and moderating their cloud service growth from 30-40% YoY in 2022 to 20-30% YoY in 2023, still well ahead of consumer tech verticals coming off declining sales in 2022 and projected down again in 2023. We have witnessed in prior slowdowns including 2019 this running through a period before low hanging fruit optimization areas are used up and focus on revenue generation over costs resumes to drive a pick-up. The launch of Chat GPT and Bing search should help promote a new range of use cases detailed in the first section of the report for another wave of AI processing.

**Figure 118: Cloud Service Growth Moderating to 20-25% growth**

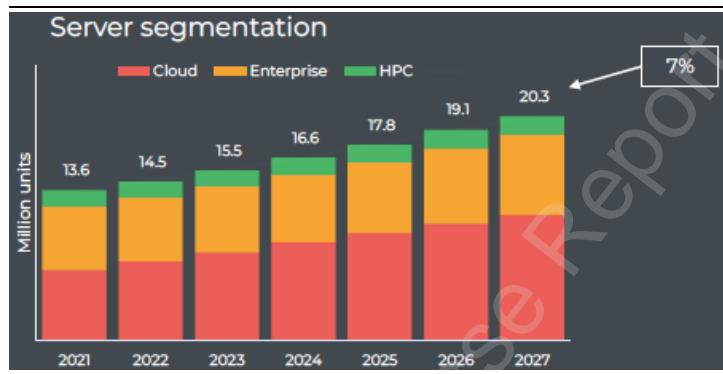


Source: Company Data, Credit Suisse Estimates

## AI investments to be a growing portion of spend

Within the data center growth, AI acceleration is poised to outgrow the overall market opportunity, with Yole projecting 7% CAGR for overall servers with growing mix of cloud while AI accelerator penetration would grow at 24% CAGR from 8% of servers in 2022 to 18% of servers by 2027.

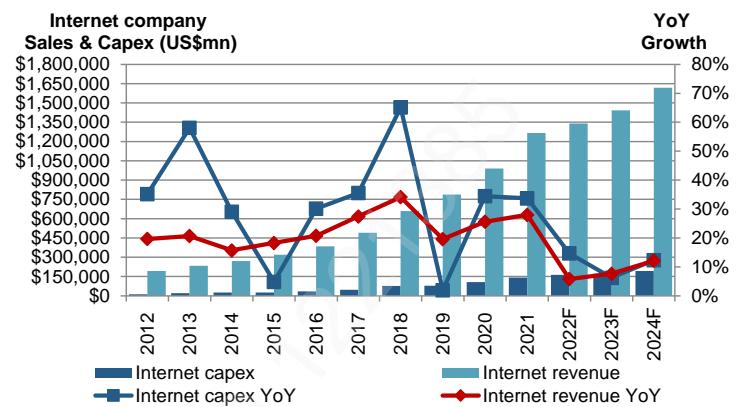
**Figure 120: Servers growing at a 7% CAGR through 2027**



Source: Yole

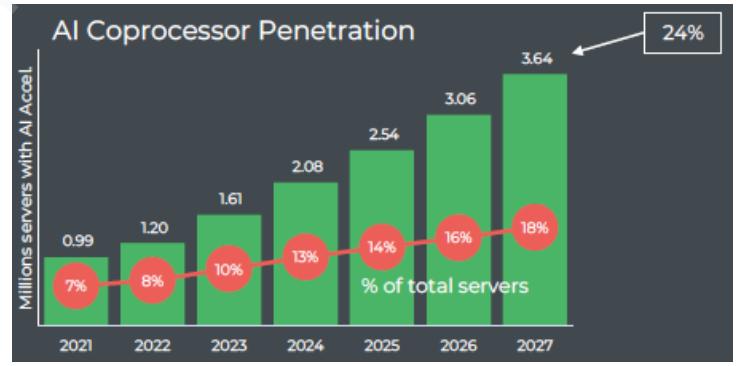
Notable for AI is that each server with an accelerator is powered by average 2 CPUs but are seeing a 2% CAGR increase in accelerator attach from 6.62 in 2022 to 2027 and a 7% CAGR in the ASP from US\$1,800 to US\$2,300.

**Figure 119: Internet Service Growth Has Slowed to Single Digits**

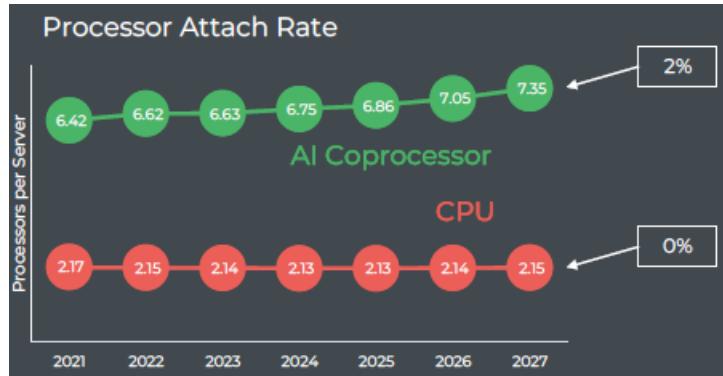


Source: Company Data, Credit Suisse Estimates

**Figure 121: AI accelerators 24% CAGR rising to 18% of servers**

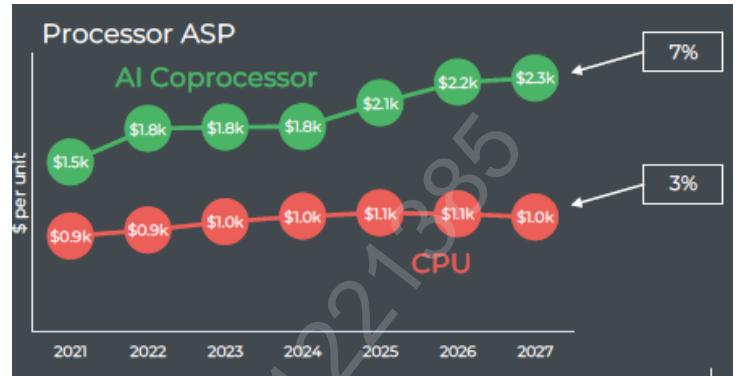


Source: Yole

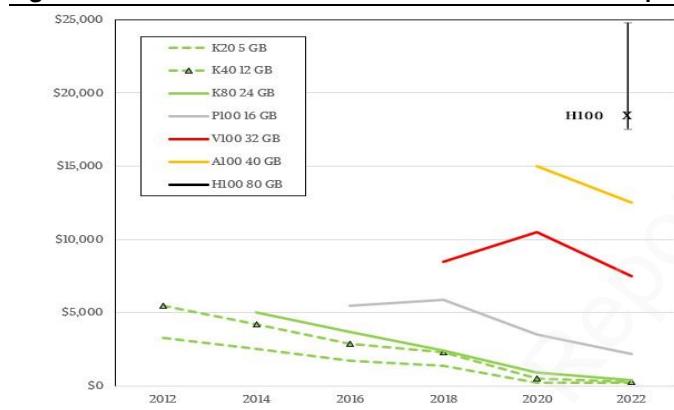
**Figure 122: Attach rate for AI chips gradually increasing**

Source: Yole

The Yole accelerator data includes all FPGA, ASIC and GPUs and dilutes the increase in pricing seen for NVIDIA's GPU accelerators. List prices for NVIDIA's GPUs have increased with process migration and capability increase from the K series starting at US\$5k to V100 (12nm) at US\$10k and A100 (7nm) at US\$15k, with the recently launched H100 (4nm) ranging from \$17k-\$25k.

**Figure 123: ASPs rising for the AI accelerators**

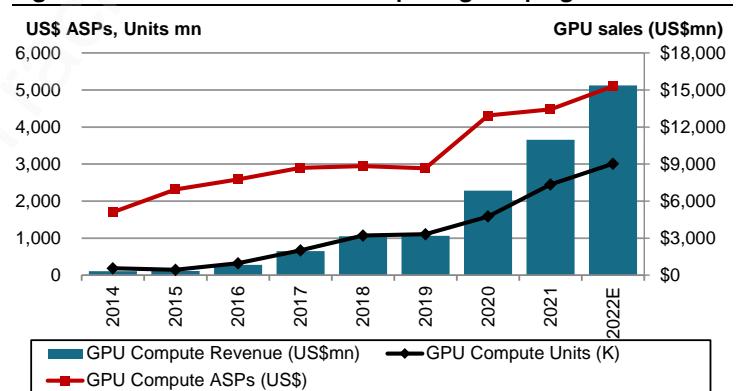
Source: Yole

**Figure 124: NVIDIA accelerators continue to increase in price**

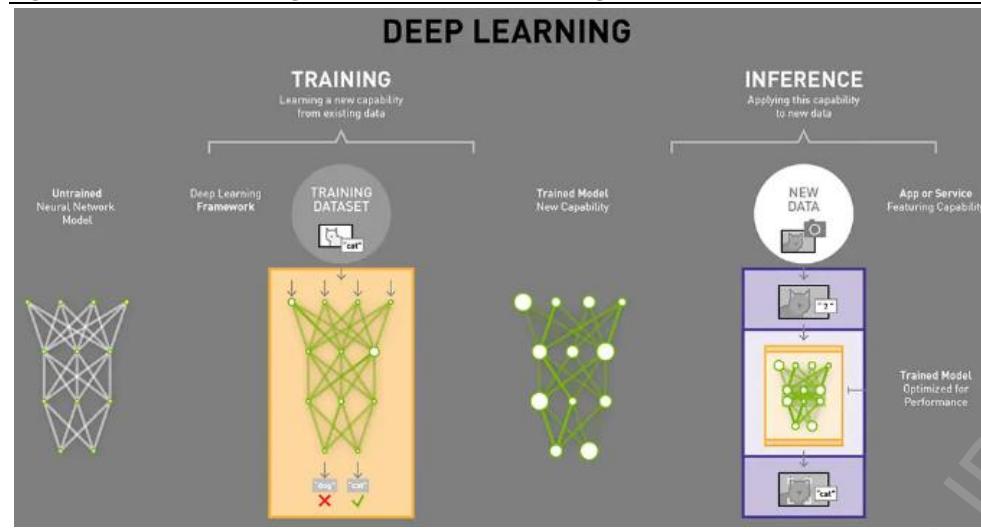
Source: The Next Platform

AI Training and inference both driving more compute and shifts to accelerators

The semiconductor market for AI/machine learning is split into two parts – training (building the AI models) and inference (running the models). The training portion involves feeding a model a curated dataset so that it can learn everything it needs about the type of data it will analyze. In the inference phase, the trained model can make predictions based on live data to produce actionable results. Machine learning brings together various data sources such as Internet data as Wikipedia, the Edgar SEC database of company filings or data from IoT sensors or user survey responses. The data is fed into a machine learning model which can employ mathematical algorithms to sort and analyze the data to score points and weed out inaccurate responses to better train for accuracy when deployed in the real world. Once deployed, inference is the process of being fed new data to make accurate decisions based on its training to recognize various inputs.

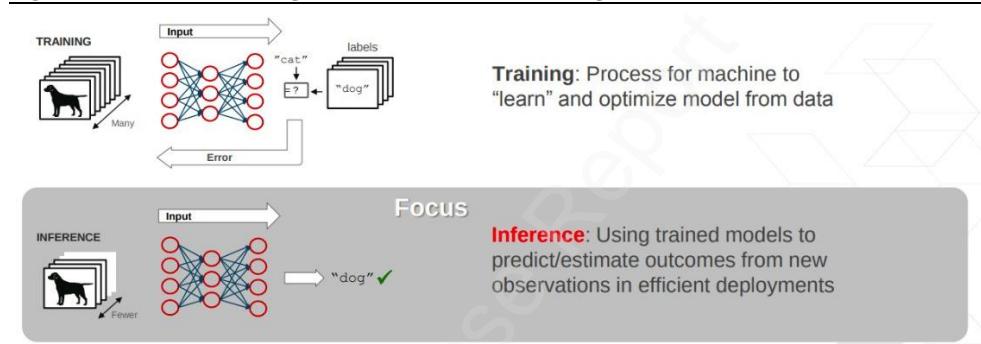
**Figure 125: GPU server units and pricing ramping**

Source: Mercury Research

**Figure 126: Deep learning compute split into training and inference**


Source: NVIDIA

Machine learning training and inference are computation intensive. The training requires feeding large amounts of data which requires intensive GPU or TPU computing at low latency by storing inputs in memory and feeding these compute engines at high bandwidth. Inference also requires low latency and ability to process instructions quickly and efficiently to make decisions in real time.

**Figure 127: Deep learning compute split into training and inference**


Source: Xilinx

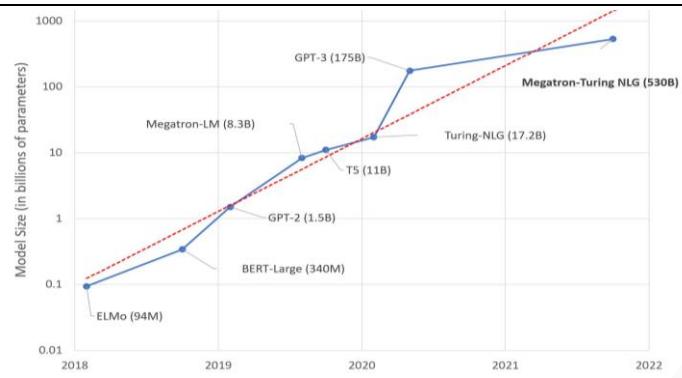
For Chat GPT, the model was trained with data from 175bn parameters including various web sources, books and articles through 2021 to respond to users in a conversational way. The model was trained with supervised fine-tuning from human AI trainers to provide conversations to play both sides of user and AI assistant. The Chat GPT infrastructure used Microsoft Azure for training and a transformer architecture which can process large amounts of data in parallel to understand language and its nuances to better understand and generate text responses. The GPT-3 model used for Chat GPT employed hundreds of GPUs for the matrix operations along with TB of high-speed memory to store the data involved in the training. For inference, CPUs and GPUs can both be run although GPUs parallel processing are more efficient on large scale language models for large batches of input sequences and data points.

While Chat GPT was not connected to real time Internet data, Microsoft has since introduced Bing search which offers conversational responses that can access real time search data. The Bing search is running a large language model customized for search ranging from 7-65 billion parameters in size, with both running on a combination of CPUs and GPUs. In a 2021 [blog](#), Microsoft indicated it could support 10s of millions of transformer inferences per second across

5 Azure regions on thousands of Azure virtual machines with each running 4 NVIDIA T4 GPUs per virtual machines.

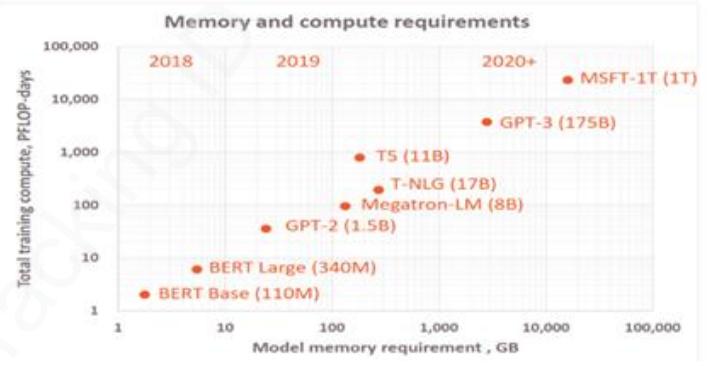
The advance and competition to generate more powerful and accurate large language models is creating a ramp in compute intensity and memory requirements for training and inference for AI accelerated compute and high density memory storage. Facebook is introducing its Llama large language model which was 65 billion parameters which can be processed in training at 380 tokens/second/GPU on 2,048 NVIDIA A100 GPUs with 80GB of RAM to train its 1.4 trillion token data set in 21 days. NVIDIA and Microsoft [introduced](#) the Megatron Natural Language model in 2021 using 560 DGX A100 servers (US\$200k/each ~ US\$100mn investment) with HDR Infiniband with each DGX having 8 NVIDIA A100 80 GB GPUs using Azure's cloud supercomputers.

**Figure 128: Large Language Models Continue to Expand Inputs**



Source: Hugging Face

**Figure 129: Internet Service Growth Has Slowed to Single Digits**



Source: Cerebras

## AI Training dominated by GPUs, ASICs/FPGAs competitors gaining in inference

AI can be performed on either main CPU or on an accelerator chip such as a GPU, FPGA or an ASIC. AI chips have unique requirements including ability to calculate a high volume in parallel of low precision calculations efficiently, memory access to a high bandwidth of memory storing high volume data, and use of software programming languages to translate AI code for execution on chip. We highlight the key difference between GPU, FPGA and ASIC below.

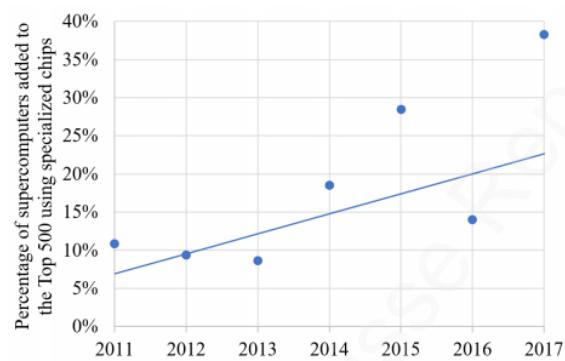
- **CPU (Central Processing Unit):** A CPU has traditionally been strong at running complex instruction sets including running the main application data in PCs and servers but can be used for some simpler AI tasks. Traditionally, CPU has been dominant the computing tasks over other chipsets as its performance can benefit from fast node migration with high volume demand, supporting it to outperform the other chipsets lacking economies of scale despite their better algorithm to run dedicated tasks. However, with the slowing Moore's Law in the past few years, the efficiency improvement for CPU has been much slower compared with other AI chips which have higher flexibility on the design for performance optimization. Therefore, GPU, FPGA and ASIC have been replacing CPU as more suitable options for AI to run dedicated tasks and calculations in parallel (vs. sequential in CPU).

**Figure 130: Different Types of Chipset Comparison for AI Computing**

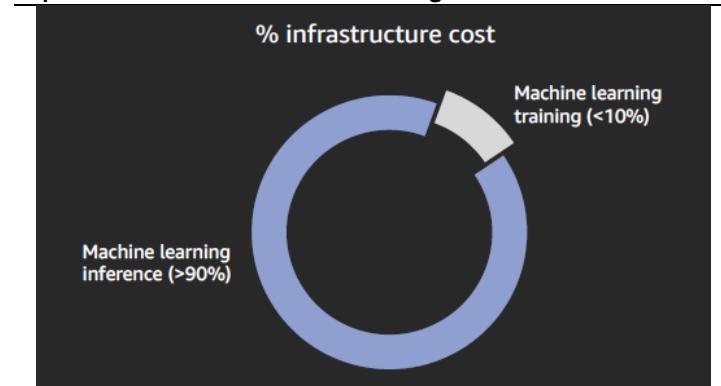
	CPU	GPU	FPGA	ASIC
<b>Processing peak power</b>	Moderate	High	Very high	Highest
<b>Power consumption</b>	High	Very high	Very low	Low
<b>Flexibility</b>	Highest	Medium	Very high	Lowest
<b>Training</b>	Low	Moderate	Low-moderate	High
<b>Inference</b>	Low	Moderate	High	High
<b>Cost per compute</b>	High	High	Moderate	Low
<b>Major applications</b>	General computing	Cloud training Cloud inference	Cloud inference Edge inference	Cloud training Cloud inference Edge inference
<b>Companies</b>	Intel, AMD	Nvidia, AMD	Xilinx, Altera	Diversified

Source: Credit Suisse estimates

- **GPU (Graphic Processing Unit):** GPUs led by NVIDIA and AMD benefit traditionally processing parallel pixels in imaging are well suited to the parallelism in training AI algorithms and inference requiring executing on multiple pieces of data at once including the newer transformer models used in large language models. GPUs can be used for training as they offer the wide floating point computational power and wide memory buses, allowing quick data movement for storage and intermediate data needed for training
- **FPGA (Field Programmable Gate Arrays):** FPGAs include logic blocks to configure to a certain set of algorithms while ASICs are hardwired and customized to certain algorithms. FPGAs have advantage on high flexibility, low latency and lower power consumption compared with GPU and CPU while can still run the tasks in parallel. The difficulty to use FPGA in AI is FPGAs tend to run slow and burn lots of power if not partitioned and designed correctly. The market is led by Xilinx and Intel's Altera.

**Figure 131: ASIC gets importance in the supercomputing now**

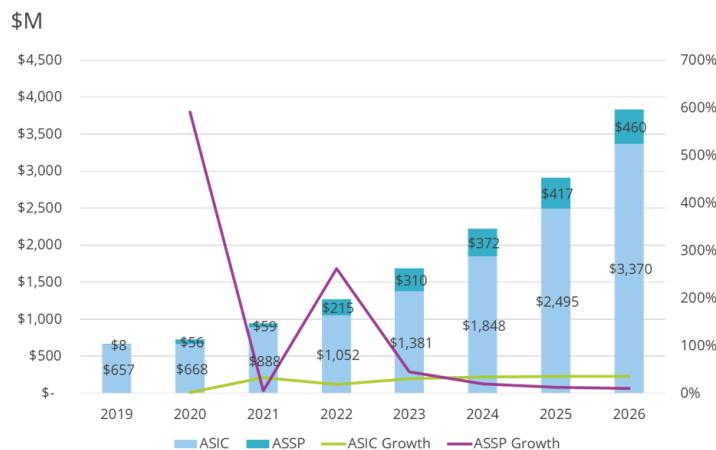
Source: Stanford University

**Figure 132: Amazon sees more efficient chipset computing is important to reduce machine learning inference cost**

Source: Amazon

- **ASIC (Application-Specific Integrated Circuit):** An ASIC is customized designed chipset for dedicated tasks, providing high flexibility for the end user to add additional features (e.g. extra ports, enhanced security) while limiting the overheads compared with CPU, GPU and FPGA. ASICs can be more efficient and save power over a GPU, but due to their more limited set of applications than a general purpose GPU compute engine may not cover the high design costs for their lower volume usage.

Due to high customization, the ASIC industry is fragmented, with players including major internet companies developing chipsets for their own system or start-ups providing solutions for certain applications. According to MIT, in addition to enterprise and end market, the ASIC is also more widely adopted in the supercomputing systems, with % of ASIC adoption in global top 500 supercomputers growing from 10% in 2011 to 40% in 2017 to enhance the computing power in addition to the existing CPUs.

**Figure 133: Server ASIC market led by AWS, Alibaba and Google's TPU**

Source: IDC

The Center for Security and Emerging Technology estimates potential higher efficiency for an FPGA or ASIC though can lose on accuracy and are less flexible for the wider general compute workloads that some tasks require.

**Figure 134: AI processors projected at 25% '23-26 CAGR, GPUs capturing 77% share**

	Training		Inference		Generality <sup>88</sup>	Inference accuracy <sup>89</sup>
	Efficiency	Speed	Efficiency	Speed		
CPU	1x baseline				Very High	~98-99.7%
GPU	~10-100x	~10-1,000x	~1-10x	~1-100x	High	~98-99.7%
FPGA	-	-	~10-100x	~10-100x	Medium	~95-99%
ASIC	~100-1,000x	~10-1,000x	~100-1,000x	~10-1,000x	Low	~90-98%

Source: CSET

## GPU leads AI training, gaining in inference

As noted by the Yole data of higher penetration, attach and ASPs of AI accelerators to servers, GPU has proven to be the widest adopted technology for training AI models. GPUs are well suited for the matrix calculations required for training AI models (multiple/accumulate functions that drive the probabilities needed to train these models. In AI, the larger the dataset the better the model – so there is an ever-increasing need for higher performance, driven by larger model sizes and enabled by GPUs with higher transistor counts and ability to execute a larger number of calculations in parallel than CPUs. According to Gartner's AI forecast for processing, GPU is projected at 77% of AI sales in 2023 and projected to grow at a 19% CAGR from 2023 to 2026. ASIC from a lower base is at 15% of workloads in 2023 though projected to grow at a 50% CAGR as TPUs and other ASICs optimized for certain AI calculations are adopted. Notably, AI GPU in Gartner's figure at US\$7.2bn does not capture all of NVIDIA's data center GPU usage at US\$15bn which also includes revenue from the entire GPU system.

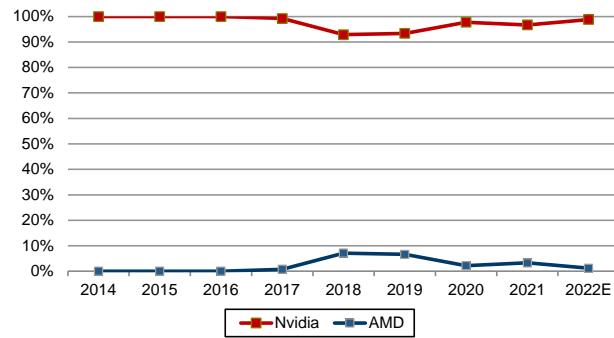
**Figure 135: AI processors projected at 25% '23-26 CAGR, GPUs capturing 77% share**

AI sales \$mn	2020	2021	2022	2023	2024	2025	2026	20-23	23-26	% of 23
GPU	\$2,609	\$4,786	\$5,869	\$7,231	\$8,897	\$10,559	\$12,166	40%	19%	77%
FPGA	\$104	\$205	\$336	\$612	\$831	\$908	\$942	81%	15%	7%
ASIC	\$271	\$501	\$828	\$1,449	\$2,576	\$3,751	\$4,854	75%	50%	15%
DSP	\$6	\$14	\$32	\$69	\$102	\$152	\$216	128%	46%	1%
<b>Total</b>	<b>\$2,989</b>	<b>\$5,506</b>	<b>\$7,066</b>	<b>\$9,360</b>	<b>\$12,405</b>	<b>\$15,372</b>	<b>\$18,178</b>	<b>46%</b>	<b>25%</b>	<b>100%</b>
<b>YoY Growth</b>	<b>84%</b>	<b>28%</b>	<b>32%</b>	<b>33%</b>	<b>24%</b>	<b>18%</b>				

Source: Gartner, December 2022

Within GPUs, NVDA leads the market for AI training, with 95-100% market share according to Mercury. We estimate NVDA's datacenter revenue is roughly evenly split between cloud and on-premise products with majority of revenue comes from training vs. inference.

**Figure 136: NVIDIA Leads the GPU Server Market**



Source: Mercury Research

## Inference shifting from x86 CPUs toward GPU and other accelerators

Inference is the task of running the AI models – responding to a ChatGPT query, providing a recommendation on a shopping site, or responding to an Alexa voice command, “inferring” a result based on how the model has been trained. While training is done in a batch process, inference is done in real time, thereby representing different compute needs.

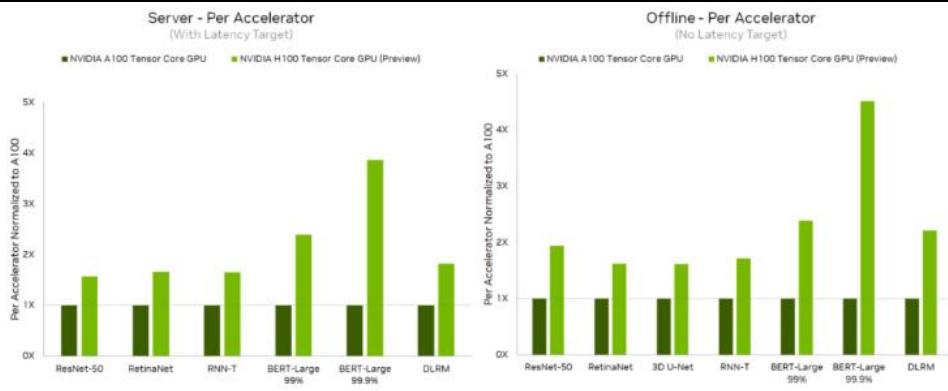
Traditionally, the majority of inference workloads run on x86 silicon, mostly from INTC. But GPU has been shown to deliver higher performance for inference, for much the same reason that GPU has proven to be better for training. While NVDA hasn't disclosed the specific revenue or growth rates for inference GPU vs. training, the company does claim that inference revenue is up 9x between NVDA's FY19 and FY22.

**Figure 137: NVIDIA's Inference GPU Sales Up 9x in 3 Years**



Source: NVIDIA

We note that in MLPerf Inference benchmarks, which measure how quickly a trained neural network can perform inference tasks on new data, NVDA's H100 had over 4x higher inference performance relative to the A100 and was the performance leader across all MLPerf inference benchmarks relative to competitors.

**Figure 138: Datacenter Per-Accelerator Performance – H100 vs. A100**

Source: NVIDIA

NVDA has two approaches toward inference. One is lower ASP, lower performance GPUs which we describe in more detail below. One of the advantages of the A-100 is that it could be used for both inference and for training. Since inference is a real-time process, inference requires high compute requirements at peak times of the day, and less performance during non-peak hours. For that reason, A-100 can be used for training during non-peak times and for inference during peak times. This is well suited toward cloud applications.

NVDA's lower-end A-series GPUs are targeted toward enterprise level inference applications, from a 40-60W A2 edge level inference, to 165W A-30 GPUs. NVDA claims a 7x performance improvement as compared to an Intel Xeon Gold 6330N CPU for inference applications.

Intel hasn't stood still in the race for AI inference and has sought to defend their share in CPU inference with the launch of Sapphire Rapids this year, their newest datacenter processor. Sapphire Rapids includes Advanced Matrix Extension (AMX) accelerators to improve AI inference performance (INTC also claims this applicable for training small models as well), with up to 10x PyTorch performance as compared to the prior server CPU generation.

Despite Intel's efforts, we expect GPU, and specifically NVDA GPU to continue to take share of AI inference over time as more algorithms are optimized to run in parallel for faster compute. IDC also estimates training workloads being accelerated by an accelerator versus being run on the CPU increasing from 86% to 90% from 2022 to 2026 while inference workloads being accelerated rising from 34% to 53% by 2026.

**Figure 139: Higher % of AI Workloads Being Accelerated**

US\$ in mn	2021	2022	2023	2024	2025	2026	CAGR
Accelerated Training	7,548.0	9,247.0	10,957.0	12,395.0	13,575.0	14,696.0	14%
Non Accelerated Training	1,174.0	1,501.0	1,619.0	1,598.0	1,609.0	1,614.0	7%
<b>% Training Accelerated</b>	<b>86.5%</b>	<b>86.0%</b>	<b>87.1%</b>	<b>88.6%</b>	<b>89.4%</b>	<b>90.1%</b>	
Accelerated Inference	1,512.0	3,065.0	5,258.0	7,097.0	8,643.0	9,757.0	45%
Non Accelerated Inference	5,134.0	5,936.0	6,921.0	7,529.0	7,965.0	8,594.0	11%
<b>% Inference Accelerated</b>	<b>22.8%</b>	<b>34.1%</b>	<b>43.2%</b>	<b>48.5%</b>	<b>52.0%</b>	<b>53.2%</b>	

Source: IDC 2022

IDC is noting this growth in AI accelerated servers versus use of a CPU for the acceleration, which would translate to outgrowth for the AI servers. IDC projects server units at +5% CAGR, non-accelerated AI servers at 9% CAGR and AI accelerated servers at 22% CAGR. Due to this trend for higher acceleration, AI servers are projected to grow at a 21% revenue CAGR from 2021-26 vs. 8% CAGR for the total server market. The trend would allow AI servers to grow from 18% of server industry revenue in 2022 to 27% of server industry revenue in 2026.

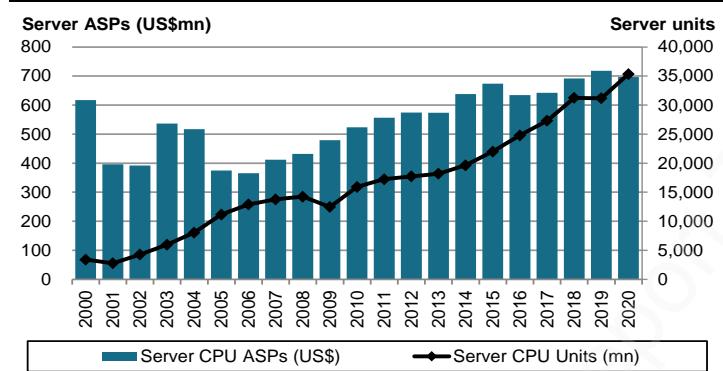
**Figure 140: Accelerated AI Servers to Outpace Non-Accelerated (CPU) Powered servers**

Units in mn / Sales in US\$bn	2021	2022	2023	2024	2025	2026	CAGR
Server Units	14.5	15.7	16.4	17.1	17.8	18.6	5%
<b>Total Server Revenue</b>	<b>\$109.0</b>	<b>\$124.1</b>	<b>\$131.8</b>	<b>\$140.5</b>	<b>\$150.8</b>	<b>\$162.6</b>	<b>8%</b>
Accelerated AI Server Sales	\$9.1	\$12.1	\$16.2	\$19.5	\$22.2	\$24.5	22%
Accel. AI Penetration %	8.3%	9.8%	12.3%	13.9%	14.7%	15.1%	
Non Accelerated AI Server Sales	\$6.5	\$7.4	\$8.5	\$9.1	\$9.6	\$10.2	9%
Non Accel. AI Penetration %	6.0%	6.0%	6.4%	6.5%	6.4%	6.3%	
<b>AI Server Revenue (Nov 2022)</b>	<b>\$15.7</b>	<b>\$19.6</b>	<b>\$24.8</b>	<b>\$28.7</b>	<b>\$31.9</b>	<b>\$34.9</b>	<b>17%</b>
AI Penetration % - Nov forecast	14.4%	15.8%	18.8%	20.5%	21.2%	21.4%	
<b>AI Server Revenue (Dec 2022)</b>	<b>\$16.9</b>	<b>\$21.8</b>	<b>\$27.8</b>	<b>\$33.0</b>	<b>\$38.0</b>	<b>\$43.4</b>	<b>21%</b>
AI Penetration % - Dec forecast	15.5%	17.6%	21.1%	23.5%	25.2%	26.7%	

Source: IDC 2022

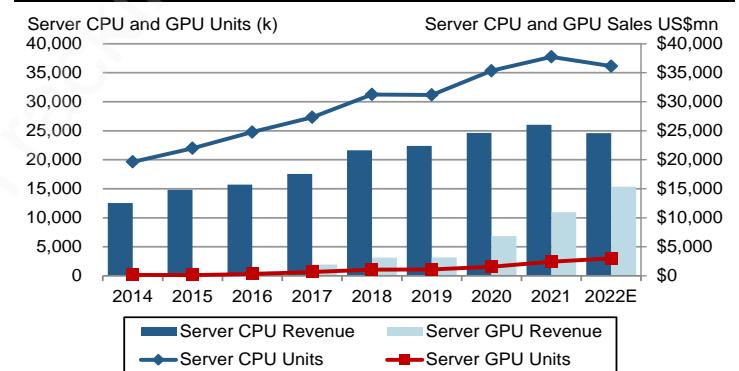
The shift to use GPU accelerators is translating to higher growth for GPUs. According to Mercury Research, server CPU processors witnessed a +8% unit and +9% CAGR since 2014 while GPUs have grown at a +42% unit and +62% sales CAGR since 2014. GPU acceleration has outgrown CPU from more AI training and penetration into inference with an additional kicker from higher ASPs as it leverages more advanced semiconductor process nodes.

**Figure 141: Server CPU a +8% Unit/9% Sales CAGR since 2014**



Source: Mercury Research

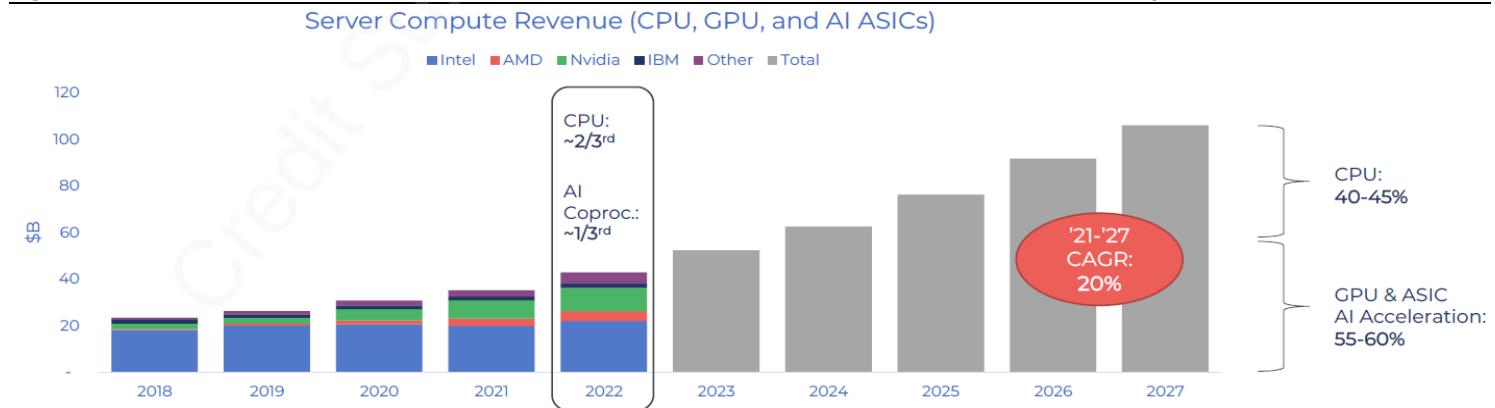
**Figure 142: GPU ramp at a 42% unit/62% sales CAGR since '14**



Source: Mercury Research

Researchers at Yole Development also note this trend, as they expect GPU/AI ASICs' share of the server compute market to increase to 55-60% in 2027, up from ~33% in 2022, while CPU share is expected to fall to 40-45% in 2027, down from ~67% in 2022.

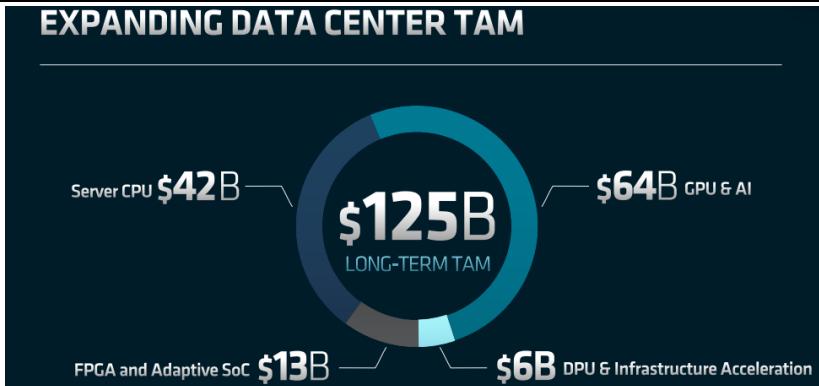
**Figure 143: Yole forecasts the server compute TAM shifts from CPU led toward GPU & ASIC/AI led by 2027**



Source: Yole Development

Yole estimates GPU/AI ASIC revenue of approximately \$57bn-\$63bn by 2027, implying a CAGR of ~33% through 2027, and more than doubling from 2022. This forecast is consistent with forecasts from AMD which estimates \$64bn GPU/AI datacenter TAM in the long-term.

**Figure 144: AMD forecasts a GPU/AI TAM LT at US\$64bn, near Yole's \$57-63bn range**



Source: AMD

The AI acceleration trend is creating a wide gap between growth rates for the data center business for the CPU companies AMD and Intel vs. GPU and ASIC acceleration efforts witnessed by the sales of NVIDIA and Marvell. The combined growth for AMD and Intel's data center CPU business from 2019-2024E even with a 2024 recovery is a +2% CAGR, with AMD's share gains from a low base driving its business up 58% versus -6% CAGR for Intel.

**Figure 145: Data center CPU growth has been only a +2% CAGR, with AMD outgrowing Intel**

Intel/AMD Data Center	1Q22	2Q22	3Q22	4Q22F	1Q23F	2Q23F	3Q23F	4Q23F	2019	2020	2021	2022F	2023F	2024F	19-24
<b>Intel DCAI</b>	\$6,034	\$4,649	\$4,209	\$4,304	\$3,443	\$3,099	\$3,347	\$3,681	\$21,696	\$23,413	\$22,691	\$19,196	\$13,570	\$16,088	-6%
QoQ Growth	-6%	-23%	-9%	2%	-20%	-10%	8%	10%							
YoY Growth	22%	-16%	-27%	-33%	-43%	-33%	-20%	-14%	8%	-3%	-15%	-29%	19%		
Intel % of Total	82%	76%	72%	72%	70%	67%	67%	67%	96%	94%	86%	76%	68%	64%	
<b>AMD Data Center</b>	\$1,293	\$1,486	\$1,609	\$1,655	\$1,460	\$1,492	\$1,640	\$1,803	\$916	\$1,540	\$3,694	\$6,043	\$6,396	\$8,949	58%
QoQ Growth	11%	15%	8%	3%	-12%	2%	10%	10%							
YoY Growth	112%	83%	45%	42%	13%	0%	2%	9%	68%	140%	64%	6%	40%		
AMD % of Total	18%	24%	28%	28%	30%	33%	33%	33%	4%	6%	14%	24%	32%	36%	
<b>Total Data Center CPUs</b>	\$7,327	\$6,135	\$5,818	\$5,959	\$4,903	\$4,591	\$4,987	\$5,484	\$22,612	\$24,953	\$26,385	\$25,239	\$19,966	\$25,037	2%
<b>QoQ Growth</b>	-3%	-16%	-5%	2%	-18%	-6%	9%	10%							
<b>YoY Growth</b>	32%	-4%	-16%	-21%	-33%	-25%	-14%	-8%	10%	6%	-4%	-21%	25%		

Source: Company data, Credit Suisse estimates

The data center processing market though has witnessed faster growth considering NVIDIA's data center TAM at 50% CAGR and Marvell's growth at 30% CAGR. Combining data from these suppliers with AMD and NVIDIA still yields a solid +14% CAGR. Intel and AMD's ability to penetrate the accelerator market in coming years is important for them capturing more of this TAM opportunity. We highlight TSMC's HPC division which also benefited from pick-up of AMD and Apple's processor business over the past 5 years on top of GPU/AI growth and set to add some Intel client tiles has grown at even faster +29% CAGR through the period.

For future growth rates for training and inference, a starting point is potential to grow 2 points of additional penetration into servers (from the 8% in 2022) into AI accelerated servers which according to Yole would place growth rates at a 24% unit CAGR and +7% ASP CAGR versus its +7% server unit CAGR through 2026. As our colleagues have noted elsewhere in this report, AI represents a revolution in software and expansion to new use cases allowing creation of applications possible from areas that could not be coded by human beings. While Chat GPT is incredibly important in disrupting search and enabling chatbots (which can support investments to attack and defend), generative AI is certainly not the last application for AI.

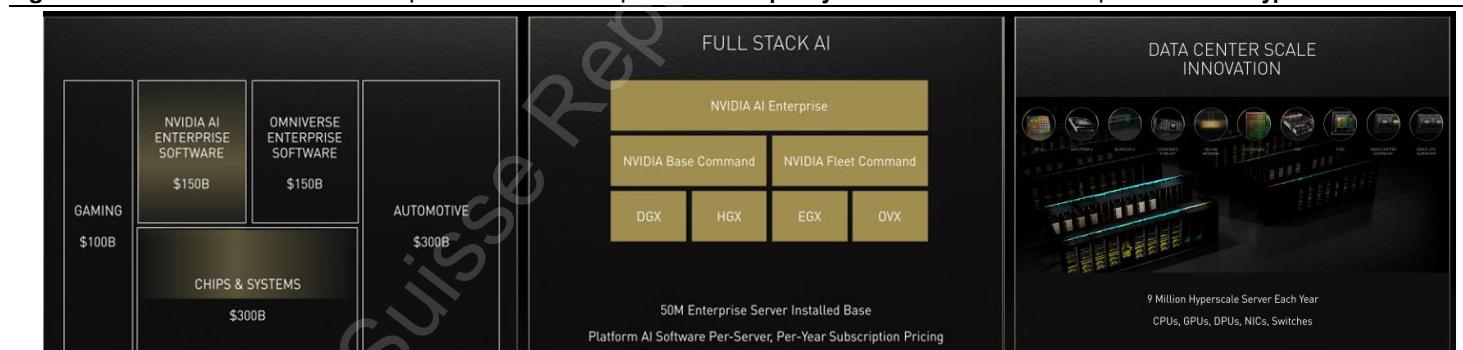
**Figure 146: Data center including GPU and ASICs a healthier +14% CAGR, with TSMC on share gains growing at a +26% CAGR**

Data Center ICs	1Q22	2Q22	3Q22	4Q22	1Q23F	2Q23F	3Q23F	4Q23F	2019	2020	2021	2022F	2023F	2024F	19-24
<b>NVIDIA Datacenter</b>	<b>\$3,750</b>	<b>\$3,806</b>	<b>\$3,833</b>	<b>\$3,616</b>	<b>\$3,853</b>	<b>\$4,263</b>	<b>\$4,655</b>	<b>\$5,080</b>	<b>\$2,983</b>	<b>\$6,696</b>	<b>\$10,613</b>	<b>\$15,005</b>	<b>\$17,851</b>	<b>\$22,871</b>	<b>50%</b>
QoQ Growth	15%	1%	1%	-6%	7%	11%	9%	9%							
YoY Growth	83%	61%	31%	11%	3%	12%	21%	40%							
<b>Marvell Datacenter</b>	<b>\$641</b>	<b>\$643</b>	<b>\$627</b>	<b>\$470</b>	<b>\$530</b>	<b>\$580</b>	<b>\$683</b>	<b>\$751</b>	<b>\$861</b>	<b>\$1,041</b>	<b>\$1,785</b>	<b>\$2,382</b>	<b>\$2,544</b>	<b>\$3,220</b>	<b>30%</b>
QoQ Growth	12%	0%	-3%	-25%	13%	9%	18%	10%							
YoY Growth	131%	48%	26%	-18%	-17%	-10%	9%	60%							
<b>ASIC and GPU ICs</b>	<b>\$4,391</b>	<b>\$4,449</b>	<b>\$4,460</b>	<b>\$4,086</b>	<b>\$4,383</b>	<b>\$4,843</b>	<b>\$5,338</b>	<b>\$5,831</b>	<b>\$3,844</b>	<b>\$7,737</b>	<b>\$12,398</b>	<b>\$17,387</b>	<b>\$20,395</b>	<b>\$26,091</b>	<b>47%</b>
QoQ Growth	14%	1%	0%	-8%	7%	11%	10%	9%							
YoY Growth	89%	59%	30%	6%	0%	9%	20%	43%							
<b>Data Center CPU + ICs</b>	<b>\$11,718</b>	<b>\$10,584</b>	<b>\$10,278</b>	<b>\$10,045</b>	<b>\$9,286</b>	<b>\$9,435</b>	<b>\$10,325</b>	<b>\$11,315</b>	<b>\$26,456</b>	<b>\$32,690</b>	<b>\$38,783</b>	<b>\$42,626</b>	<b>\$40,361</b>	<b>\$51,128</b>	<b>14%</b>
QoQ Growth	3%	-10%	-3%	-2%	-8%	2%	9%	10%							
YoY Growth	49%	16%	0%	-12%	-21%	-11%	0%	13%							
<b>TSMC HPC sales</b>	<b>1Q22</b>	<b>2Q22</b>	<b>3Q22</b>	<b>4Q22</b>	<b>1Q23F</b>	<b>2Q23F</b>	<b>3Q23F</b>	<b>4Q23F</b>	<b>2019</b>	<b>2020</b>	<b>2021</b>	<b>2022F</b>	<b>2023F</b>	<b>2024F</b>	<b>19-24</b>
<b>TSMC HPC Segment</b>	<b>\$7,210</b>	<b>\$7,717</b>	<b>\$7,908</b>	<b>\$8,451</b>	<b>\$7,183</b>	<b>\$6,993</b>	<b>\$8,238</b>	<b>\$9,186</b>	<b>\$10,259</b>	<b>\$14,944</b>	<b>\$21,008</b>	<b>\$31,285</b>	<b>\$31,600</b>	<b>\$37,013</b>	<b>29%</b>
QoQ Growth	26%	7%	2%	7%	-15%	-3%	18%	12%							
YoY Growth	58%	51%	42%	47%	0%	-9%	4%	9%							

Source: Company data, Credit Suisse estimates

## NVIDIA continuing to advance its solutions to power AI

The democratization of NVDA silicon through cloud instances means that even small developers can develop the next ChatGPT. We believe that creates open-ended growth which could ultimately expand data generation and growth trajectory for servers or put AI acceleration into servers on a much faster pace. While it's difficult to accurately upsize the training and inference markets, for their part, Nvidia has identified a datacenter TAM opportunity of \$600bn, with \$300bn in hardware (chips/systems) and \$300bn in software. Within that, NVDA estimates the hyperscale TAM for infrastructure alone represents a \$150bn opportunity.

**Figure 147: NVIDIA estimates a US\$1trn TAM with US\$300bn in chips/system hardware and US\$150bn from hyperscale**

Source: NVIDIA

NVIDIA has advantages that extend well beyond its steadily advancing product platforms to give it an advantage in machine learning. The following are some of its key barriers for AI:

- **Vertical integration approach.** NVIDIA views accelerated computing needs to be vertically integrated as a full stack computing problem to write the OS or cloud/enterprise distributed operating system, run time engines, libraries, application frameworks or develop the storage, networking and cybersecurity. NVIDIA views customers are not just buying a chip but need the NVIDIA computing stack to speed up creation and implementation of AI algorithms. NVIDIA has created vertical platforms through its 1) graphics compute - the RTX graphics stacks, AI, Physics and Ray tracing engines, 2) scientific computing stack, 3) NVIDIA AI as the operating system with all the end to end run times and engines starting from training through inference, 4) NVIDIA Omniverse as the next wave of AI where AI interacts with the physical world by providing ground truth.

**Figure 148: NVIDIA Compute Optimization Across the Full Stack**

Source: NVIDIA

- **Software.** We believe much of NVDA's AI competitive advantage comes from software. That software advantage comes in two forms. One is from CUDA, NVDA's proprietary software that can only be used to program NVDA GPUs, and which forms the basis of many AI programming frameworks. One reason NVDA came to lead AI training is that all AI frameworks are compatible with NVDA GPUs, and that CUDA only runs on NVDA silicon.
- **CUDA programming language.** CUDA is a parallel computing platform and programming model developed by NVDA in 2006 for general computing on its own GPUs. CUDA sits at the center of a number of popular frameworks for deep learning, including TensorFlow, Torch, PyTorch, Keras, MXNet, and Caffe2 – which all use the cuDNN library ('CUDA Deep Neural Network'), developed by Nvidia. Since CUDA isn't available on non-NVDA platforms and has become so deeply ingrained into the AI ecosystem, it has become one of the key competitive advantages for NVDA.

**Figure 149: Nvidia develops Application frameworks to further simplify adoption of AI**

Source: NVIDIA

- **Software libraries.** NVDA has also made a significant investment in software libraries that work with NVDA silicon, which provide building blocks for common AI applications. NVDA regularly maintains, updates and releases new acceleration libraries to broaden and deepen its competitive differentiation vs AMD's ROCm and others. These libraries support application frameworks that further simplify the process for developers to build new, custom AI models, and are the product of years/decades of work by NVDA's engineering teams. These include pre-trained deep learning models, speech AI models, recommender system models, conversational AI models, among others. This also adds to NVDA's competitive

advantage since these libraries provide a starting point for AI projects that aren't available on non-NVDA systems.

- **Large language models based on transformers.** Transformers can lead to breakthroughs in natural language processing and large language models such as question/answer, translation, and software programming, and can learn to perform tasks for which they were never trained, and the same model asked the same question in different contexts can provide a different response. Applications for transformers include summarizing a story, reporting breaking news, paraphrasing statements.
- **Hopper & Adoption of Transformers and LLMs.** Hopper claims 5x the throughput and 3x reduction in total cost of ownership which implies a higher price than Ampere with significant net reduction in ownership costs. It would ship some quantity this quarter and ramp further in the coming quarter. The device has strong interest industry wide with the new Transformer engine largely replacing the older vision engines. It has a strong ability to perform with large language models using transformers and also democratizing use of AI and application of these language models with much lower inference cost. The product is seeing good traction in the revolutionizing digital biology space as costs of gene sequencing and prediction of protein chemistries and structures improves.

**Figure 150: H100 2.4x CUDA cores, ~50% more transistors/VRAM bandwidth vs A100**

NVIDIA Accelerator Specification Comparison			
	H100	A100 (80GB)	V100
<b>FP32 CUDA Cores</b>	16896	6912	5120
<b>Tensor Cores</b>	528	432	640
<b>Boost Clock</b>	~1.78GHz (Not Finalized)	1.41GHz	1.53GHz
<b>Memory Clock</b>	4.8Gbps HBM3	3.2Gbps HBM2e	1.75Gbps HBM2
<b>Memory Bus Width</b>	5120-bit	5120-bit	4096-bit
<b>Memory Bandwidth</b>	3TB/sec	2TB/sec	900GB/sec
<b>VRAM</b>	80GB	80GB	16GB/32GB
<b>Interconnect</b>	NVLink 4 18 Links (900GB/sec)	NVLink 3 12 Links (600GB/sec)	NVLink 2 6 Links (300GB/sec)
<b>GPU</b>	GH100 (814mm <sup>2</sup> )	GA100 (826mm <sup>2</sup> )	GV100 (815mm <sup>2</sup> )
<b>Transistor Count</b>	80B	54.2B	21.1B
<b>TDP</b>	700W	400W	300W/350W
<b>Manufacturing Process</b>	TSMC 4N	TSMC 7N	TSMC 12nm FFN
<b>Interface</b>	SXM5	SXM4	SXM2/SXM3
<b>Architecture</b>	Hopper	Ampere	Volta

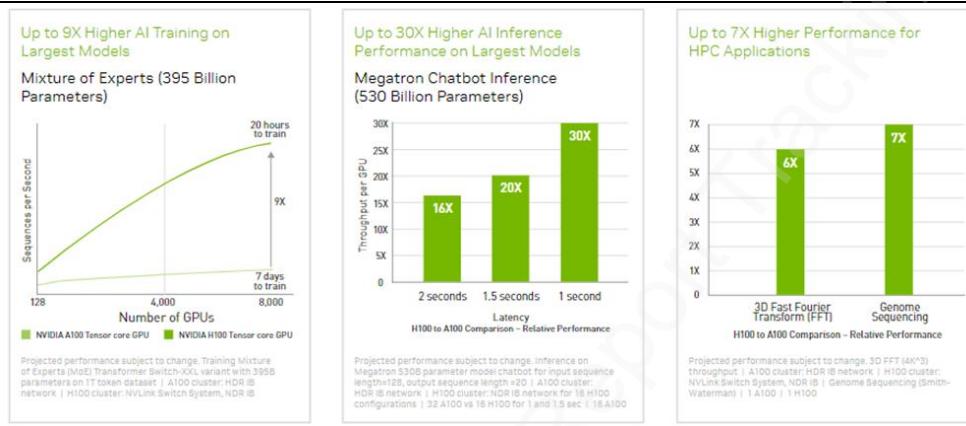
Source: Company data, Anandtech

On top of its platform model of application and software frameworks, NVIDIA also has leading GPUs leveraging advanced silicon with architectural design improvements to continually speed up AI acceleration beyond the pace of Moore's Law's density improvements.

- **A100.** The A100 GPU is based on NVDA's Ampere architecture and is the engine of NVDA's datacenter platform. Performance of the A100 is up to 20x that of its processor (Volta) and can scale up or be partitioned into seven smaller, isolated GPU instances.

- **A800.** The A800 is derivative of the A100 and was built for Chinese customers, to conform with US export restrictions, necessitating lower performance. Despite the lower performance, we believe pricing for the A800 is similar to the A100.
  - **H100.** The H100 is based on the Hopper architecture and is NVDA's highest performing datacenter GPU to date. According to NVDA, the H100 accelerates AI training and inference, HPC, and data analytics applications in cloud datacenters, servers, edge systems and workstations. The H100 provides up to 9x faster training and 30x inference speed up on large language models versus the A100. Training time is reduced from days to hours relative to the A100. We expect that H100 pricing will be on the order of a 50% increase vs. the A100, with the increase driven by its significant increase in performance.
  - **Grace/Hopper.** Grace Hopper integrates a CPU with the H100, with the increase in performance driven by NVDA's proprietary NVLink communication protocol, reducing latency in communication between the GPU/CPU and memory. Current architectures use PCI-Express for chip-to-chip communication, creating bottlenecks. Because of the performance gains with this architecture, we expect on the order of a 50% content increase for Grace Hopper versus the H100. We expect more details on this product at NVDA's upcoming GTC in March.

Figure 151: NVIDIA's H100 enhances training and inference over its prior gen A100

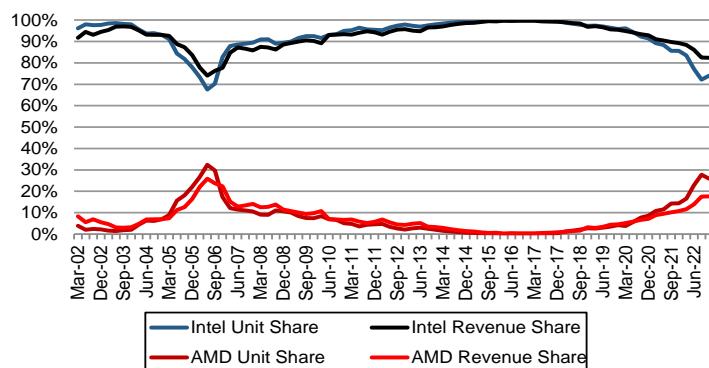


Source: NVIDIA

## Semiconductor suppliers trying to break NVIDIA's lead

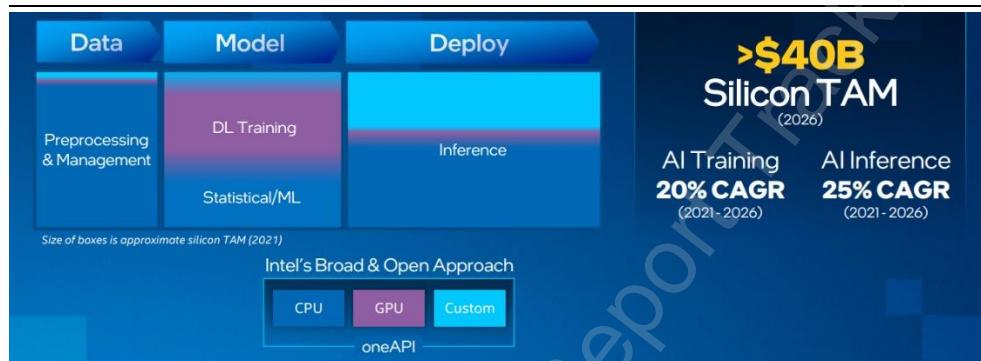
**Intel a beneficiary through its x86 CPU lead and AI efforts, though share loss a risk**

INTC is clearly an AI beneficiary due to both the x86 servers paired with NVDA silicon as well as x86 servers used for inference – and we believe the majority of inference workloads still run on x86. AI has clearly been a driver of x86 server growth, though that growth is not anywhere near the growth for GPU accelerators, with a 2% revenue CAGR from 2019-2024. At issue is that, while AI has likely expanded the market for x86 servers, INTC has continued to lose share. INTC's share of the server market has decreased from 96% in 2019 to 76% in 2022. In addition to share loss vs. AMD, we believe that INTC has also lost share in the AI inference market to NVDA GPUs, which have offered higher performance. We note that INTC's datacenter revenue has declined at a -4% CAGR (2019-2022), while NVDA's datacenter revenue has grown at a 71% CAGR over the same period.

**Figure 152: Intel has lost unit and revenue share in server CPUs in the past 5 years**

Source: Mercury Research

As noted above, INTC is seeking to defend their share in AI inference (and also to gain share in small model AI training) with the new Sapphire Rapids CPU, which includes AMX extensions intended to boost AI performance, and to keep workloads on INTC server silicon.

**Figure 153: Intel estimates >70% of inference running on Xeon, expects >\$40bn TAM for AI silicon in 2026**

Source: Intel

With respect to AI training, INTC has two products targeted toward the higher end AI market. One is the Habana Gaudi (the result of a prior acquisition), as well as Ponte Vecchio, an internally developed product. Ponte Vecchio was finally launched in November 2022 as the Intel Max Series GPU. Intel also announced a next generation data center GPU, code named Rialto Bridge set for market debut in 2024 followed by Falcon Shores tile-based architecture. While INTC has made some performance claims citing advantages vs. NVDA silicon, we're not aware of any significant traction with regard to these initiatives and we believe revenue has thus far been immaterial as compared to NVDA datacenter revenue.

**Figure 154: Intel claims up to ~1.85x advantage vs A100 GPU for machine learning workloads using integrated AI accelerators**



Source: Intel

**Figure 155: Intel HPC roadmap to upgrade Ponte Vecchio with Rialto Bridge and move toward Falcon Shores tile architecture**



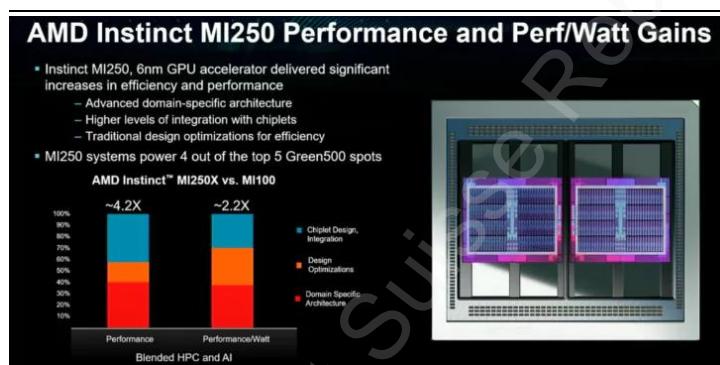
Source: Intel

#### AMD offers strong CPU and GPU platforms and now expanding into AI accelerators

AMD also possesses strong GPU technology, resulting from their years of graphics expertise. We estimate AMD shipped \$180m in datacenter GPUs in 2022, with the largest portion of that focused on high performance computing as opposed to AI-specific applications. In AI, AMD uses both hardware and software optimization for EPYC processors for inference. AMD's ZenDNN software is optimized for EPYC and is integrated with all industry standard frameworks. AMD has noted that EPYC is deployed across multiple cloud vendors.

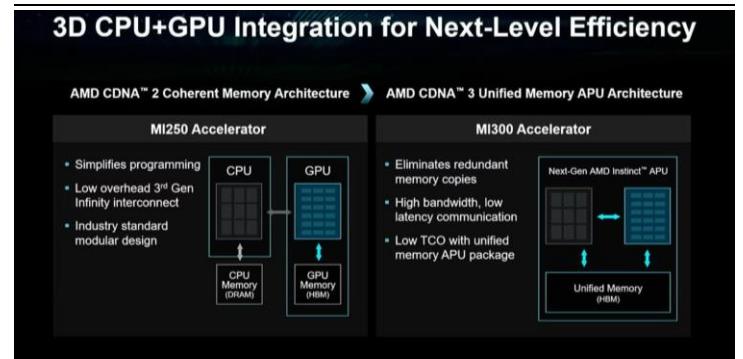
At CES in January 2023, AMD announced their new Instinct MI300 chip, which is a combined CPU-GPU designed to accelerate AI-based workloads and could be a potential competitor to NVDA's Grace Hopper. AMD asserts the MI300 can reduce training time for large language models from months to weeks and with dramatically lower energy costs relative to GPUs by also adding unified memory access between both the processor and memory to the GPU. MI300 is sampling now and is expected to ship in 2H23.

**Figure 156: AMD MI250 Accelerator with CPU/GPU integration**



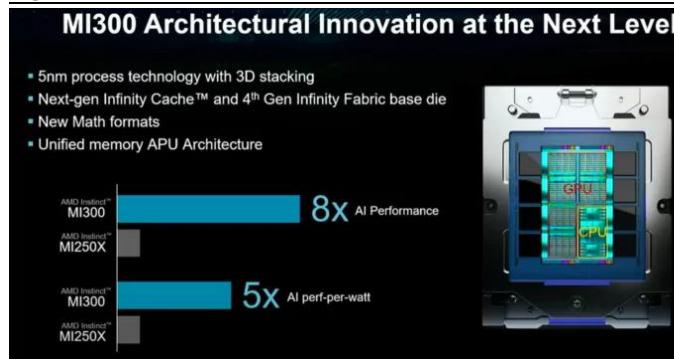
Source: AMD

**Figure 157: MI300 adds shared memory access to the CPU/GPU**



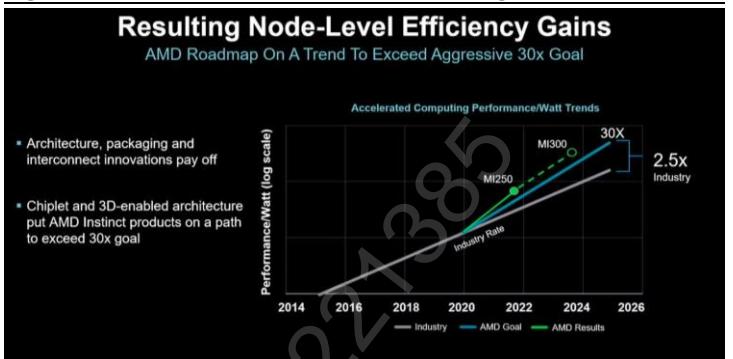
Source: AMD

AMD has been aggressive innovating with 3D stacking and high bandwidth connections between compute tiles with its Infinity fabric. The chipset combines 9 TSMC 5nm GPU/CPU chiplets stacked on 4 6nm base dies using 3D stacking paired with 128GB of on-package shared high bandwidth memory and claims to outperform its prior generation MI250x by 8x for AI and have 5x improvement in AI performance per watt. AMD views LT architecture, packaging and innovations can play out in 2.5x industry pace for accelerating compute performance/watt.

**Figure 158: AMD MI300 for additional AI performance**

Source: AMD

On comparison of raw specs, AMD's AI cores and transistor count stretches beyond NVIDIA's GPUs including its recent H100 launch. AMD still has to play catch-up to NVIDIA's incumbent advantage with its application frameworks and CUDA stack which have already been used across many AI workloads. The hyperscalers though are introducing their own open-source software stack which can enable more applications to run on rival accelerators.

**Figure 159: Accelerated performance/watt gains with MI300**

Source: AMD

**Figure 160: Comparison of Intel, NVIDIA, and AMD accelerator offerings**

US IDM/Fabless Companies	Intel	Intel	Intel	Nvidia	Nvidia	Nvidia	AMD	AMD
Chipset	4th Gen Xeon 8490H	Flex 140	Max 1100	V100	A100	H100	MI250X	MI300
Type	CPU	GPU (DG2-128)	GPU (Ponte Vecchio)	GPU	GPU	GPU	GPU	GPU
Function	General purpose	General purpose	General purpose	General purpose	General purpose	General purpose	General purpose	General purpose
Foundry	Intel	TSMC	Intel	TSMC	TSMC	TSMC	TSMC	TSMC
Manufacturing node	7nm	6nm	10nm	12nm FFN	7N	4N	6nm chiplet	5nm chiplet
Die size (mm <sup>2</sup> )	477x4	157	1,280	815	826	814	724	1,017
Number of transistors (billion)	NA	7	100	21	54	80	58	146
Density (mn transistors per mm <sup>2</sup> )	NA	46	78	26	65	98	80	144
AI-optimized cores	60	1,024	16,384	5,120	5,120	5,120	14,080	14,080
On-board SRAM	112.5MB	4MB	408MB	6MB	6MB	50MB	16MB	16MB
Memory bandwidth	4xDDR5	186GB/s GDDR6	3.3TB/s HBM2e	900GB/s HBM2	2TB/s HBM2e	2TB/s HBM3	3.3TB/s HBM2e	3.3TB/s HBM3
Bus interface	PCIe 4.0x8	PCIe 4.0x8	PCIe 5.0x16	PCIe 3.0x16	PCIe 4.0x16	PCIe 5.0x16	PCIe 4.0x16	PCIe 4.0x16
Launch	1Q23	3Q22	1Q23	2Q17	2Q20	1Q22	4Q21	1Q23
Peak performance (FP32, TFLOPS)	NA	8	22	14	19	51	48	48
TDP (W)	350	75	300	300	400	350	500	600

Source: Company data, Credit Suisse estimates

## Battleground emerging between custom ASICs

With the benefit of dedicated computing power at better power efficiency provided by ASICs, more companies have been designing customized chipsets to enable AI training and inference for the cloud service differentiation. The US companies have been leading on the projects on chipset customization, including Google's TPU and Amazon's AI inference and training chipsets.

**Figure 161: Description of the ASIC chipset plans from the hyperscalers**

Company	ASIC Initiatives
AWS	Inferentia for inferencing, Trainium for deep learning training workloads.
Google	Tensor Processing Units (TPUs) for neural network machine learning. Reports (The Information) indicate that two 5nm server chips in development, with production slated for 2H24 and deployment as early as 1H25.
Meta	Reportedly in the process of designing its own ML ASICs. (The Information)
Alibaba	Yitian 710 line used for data center processing.
Tencent	Currently uses Zixiao for ML acceleration
Baidu	Kulun Chip AI processor currently in use at Baidu cloud data centers. 2nd gen Kunlun chip launched in 2021.

Source: Company data, Credit Suisse estimates

In addition to the hyperscalers, the start-up companies with strong track record have also attracted strong financial support from the investors (e.g. Cerebras Systems and Graphcore) or being acquired by established companies which would like to grow their presence in AI computing (e.g. Habana acquired by Intel and Annapurna acquired by Amazon). We profile the ASIC projects successfully taped out and widely adopted in cloud computing.

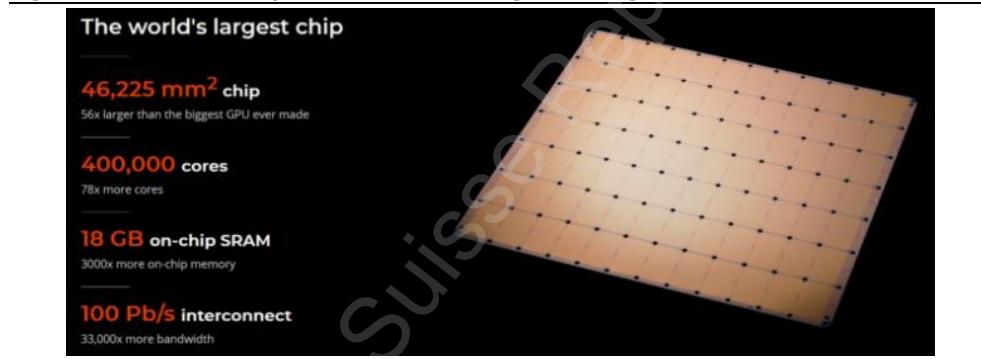
**Figure 162: Hyperscaler and Start-up AI Accelerator Chips – TSMC leads on these**

US Hyperscalers/Start-ups	Cerebras Systems	Google	Google	Graphcore	Intel / Habana	Intel / Habana	Amazon	Amazon
<b>Chipset</b>	WSE-2	Cloud TPUv3	Cloud TPUv4	Colossus MK2	Gaudi 2	Greco	Inferentia Gen 2	Trainium
<b>Type</b>	ASIC	ASIC	ASIC	ASIC	ASIC	ASIC	ASIC	ASIC
<b>Function</b>	AI training	AI training/inference	AI training/inference	AI accelerator	AI training	AI inference	AI inference	AI inference
<b>Foundry</b>	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC
<b>Manufacturing node</b>	7nm	16nm	7nm	7nm	7nm	7nm	7nm	7nm
<b>Die size (mm^2)</b>	46,225	648	780	823	850	850	NA	NA
<b>Number of transistors (billion)</b>	2,600	10	31	59	NA	NA	NA	NA
<b>Density (mn transistors per mm^2)</b>	56	15	40	72	NA	NA	NA	NA
<b>AI-optimized cores</b>	850,000	2	2	1,472	24	8	2	2
<b>On-board SRAM</b>	40GB	32MB	288MB	900MB	48MB	128MB	NA	NA
<b>Memory bandwidth</b>	20PB/s	900GB/s HBM	1200GB/s HBM2	47.5TB/s HBM	2.45TB/s HBM2e	204GB/s LPDDR5	NA	13.1TB/s HBM
<b>Bus interface</b>	NA	NA	NA	PCIe 4.0x16	PCIe 4.0x16	PCIe 4.0x8	NA	NA
<b>ASIC provider</b>	TSMC direct	Mediatek	Broadcom	TSMC direct	Alchip	Alchip	Alchip	Alchip
<b>Launch</b>	3Q21	4Q18	4Q21	3Q20	2Q22	2Q22	2022	4Q20
<b>Peak performance (FP32, TFLOPS)</b>	NA	4	4	62	38	NA	2,022	53
<b>TDP (W)</b>	NA	450	175	300	600	75	NA	NA

Source: Company data, Credit Suisse estimates

■ **Cerebras Systems' WSE-2:** Cerebras is an AI company founded by ex-SeaMicro/AMD engineers in the US in 2015. Cerebras Systems has been dedicated to designing both chipset level and system level high performance computing hardware. The company introduced its 1st Wafer Scale Engine (WSE), a single, wafer scale processor integrating compute, memory and interconnect fabric and subsequently launched 2nd generation chipset in 2021 manufactured on TSMC's 7nm with 850k cores and 2.6trn transistors, featuring 40GB SRAM, 20PB/s memory bandwidth and 220 Pb/s fabric bandwidth.

**Figure 163: Cerebras Systems wafer size engine for high performance and low latency**



Source: Cerebras Systems

With its wafer scale engine powering the company's system level solution CS-2, it claims one system can train models with up to 20bn parameters. It is also noteworthy that compared with the traditional GPU cluster where the researchers need to use a special framework to solve the bottleneck on individual processor, memory capacity, bandwidth and interconnect topology, Cerebras systems' solutions can easily connect multiple CS-2 AI systems into a cluster (wafer scale cluster). The company has demonstrated its capability through its latest supercomputer Andromeda combining 16 WSE-2 chips into one cluster with 13.5mn AI cores, delivering up to 1FLOPS of AI computing power at 500kW, much more power efficient compared with GPU-based supercomputers.

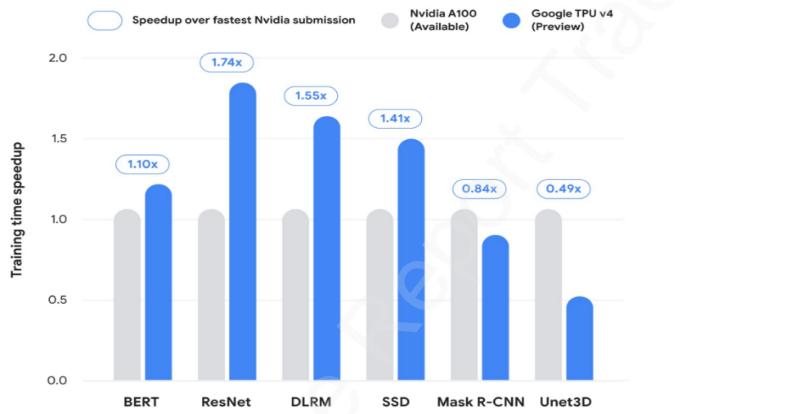
Cerebras Systems has been working with customers in industry verticals including GSK (genetic and genomic research and drug discovery), AstraZeneca (shortened AI training

time from 2 weeks for general purpose GPU to 2 days with CS-1), energy companies (e.g. TotalEnergies) and multiple academic institutions and supercomputing centers.

- **Google's TPU:** Tensor processing unit (TPU) is an AI accelerator ASIC designed by Google as a matrix processor specifically for neural network machine learning based on the company's own TensorFlow software. Compared with general purpose GPU, TPUs are designed for a high volume of low precision computation and widely adopted in the company's services including map, translation, photos and search assistant. The company has been adopting TPUs for its own data centers since 2015 as part of its cloud infrastructure and created a smaller version of the chipset available for 3rd party use through its cloud TPU service, allowing the developers and data scientists running machine learning models on Google Cloud at lower cost compared with other solutions.

Following the initial adoption of TPU v1 on TSMC's 28nm in 2015 only for AI inference, the company has been upgrading the chipsets for both AI training and inference through optimizing the chipset architecture and manufacturing technology migration, with the latest version TPU v4 manufactured on TSMC's 7nm improving performance by more than 2x over the TPU v3. TPU v4 has been adopted by its customers including Cohere (natural language processing services, with 70% training time improvement migrating from TPU v3 to TPU v4), LG AI Research (TPU v4 used to train LG EXAONE AI with 300bn parameter scale), Salesforce (TPU v4 enables the breakthroughs in conversational AI programming with its autoregressive language model project).

**Figure 164: Google claims TPUs have superior performance on training**



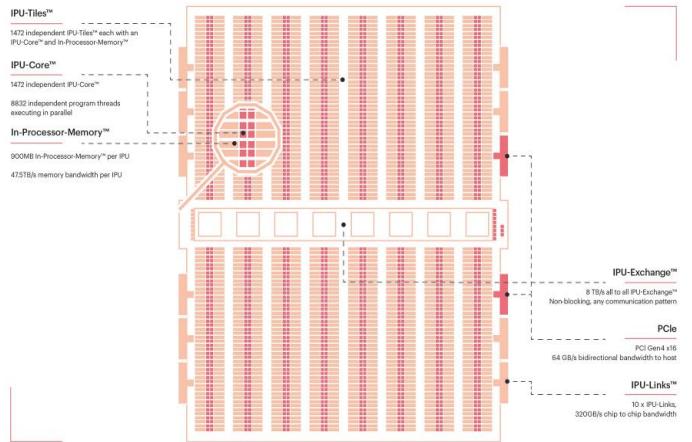
Source: Google

- **Graphcore's Colossus MK2:** Graphcore is a UK semiconductor company founded in 2016 with a focus on developing the accelerators for AI and machine learning. The company introduced its 1st chipset Colossus GC2 on 16nm supporting the tasks in the standard machine learning frameworks. The company subsequently introduced its 2nd generation processor Colossus MK2 - GC200 IPU on TSMC's 7nm, featuring 59bn transistors, 1,472 computing cores and 823mm<sup>2</sup> die size while bonding a power-delivery die with the computing die through TSMC's WoW (wafer on wafer) packaging.

With close to 50% of the die is memory, the IPUs have a sizeable local memory of 896MB SRAM, supporting memory speed at 62TB/s while also enable high speed inter chip connection at 320GB/s. Compared with the latest general-purpose GPU or Google's TPU, Graphcore claims its MK2 chipset has much higher core density and memory support. The company also believes it has cost advantage as it uses on die SRAM and off die DRAM vs. expensive HBM adopted in GPU and TPU. The company has been working with the early access customers since the chipset introduction, including search engine (e.g. Qwant), financial firm (e.g. Citadel) and academic institutions (e.g. Imperial College London). In 4Q19, Graphcore announced its collaboration with Microsoft on the company's Azure public cloud platform, with its IPs enhancing advanced machine vision and natural language processing

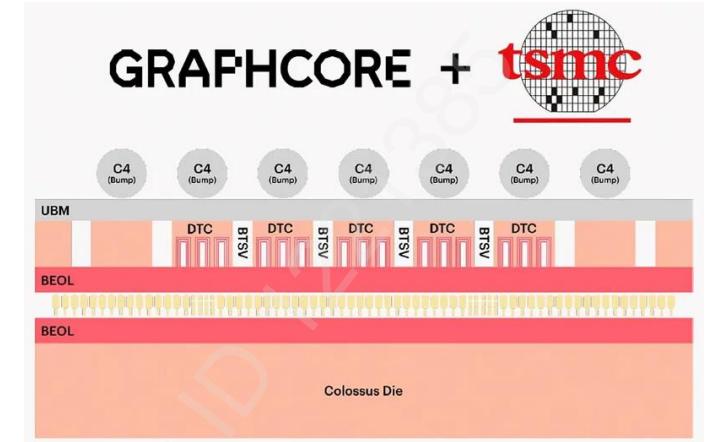
models. The IPUs has also been available on the machine learning servers provided by OEMs (e.g. Dell) and ODMs.

**Figure 165: Graphcore Colossus MK2 structure**



Source: Graphcore

**Figure 166: TSMC's WoW packaging enables Graphcore's IPU**

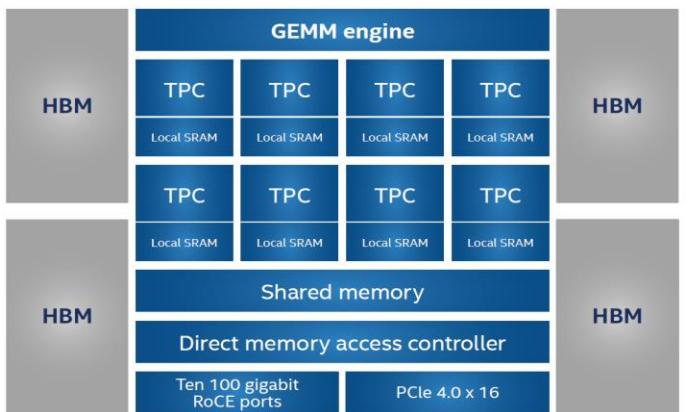


Source: Graphcore

- **Intel/Habana:** Habana Labs was founded in Israel in 2016 and acquired by Intel in 4Q19 with a focus on AI processors to train deep neural networks and for inference deployment. The company has 2 major chipset product lines including AI training and AI inference. For AI training, the company uses heterogeneous architecture to support AI-dedicated matrix multiplication engine with large on-board memories and integrated networking capability to enable the performance of the chipset.

Habana Labs has migrated its semiconductor roadmap to TSMC's 7nm for its 2nd generation Gaudi2 AI training chipset, increasing the number of AI-customized tensor processor cores from 8 to 24, tripling its in-package memory to 96GB of HBM2e at 2.45TB/s bandwidth while adding support for FP8 while integrating a media processing engine for processing compressed media. Based on ResNet-50 training throughput for computer vision, Habana claims its Gaudi2 chipset enables 2.4x performance over its 1st generation chipset on 16nm while delivers 1x more performance provided by Nvidia's A100 GPU and it also outperforms GPU in natural language processing.

According to Habana, 1,000 Gaudi2 processors have been deployed in its own data center in Israel to support R&D for software optimization while it is also working with multiple customers including Mobileye (training deep learning models for tasks such as object detection and segmentation which enable vehicles to sense and understand its surrounding), Leidos (deep learning training of medical imaging data sets) and Supermicro (AI deep learning server with improved TCO).

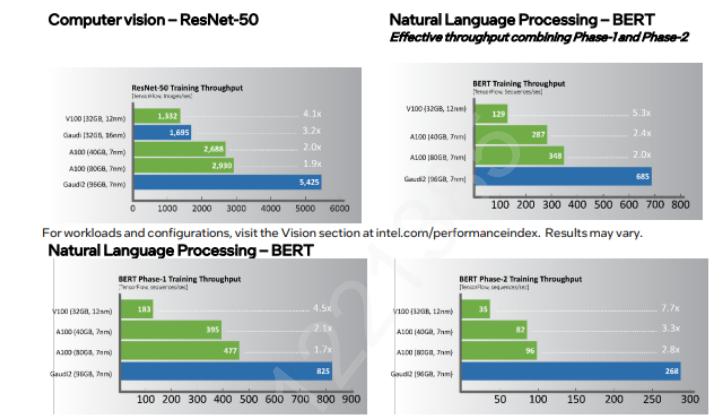
**Figure 167: Habana's Gaudi2 features multiple cores and HBM**

Source: Intel

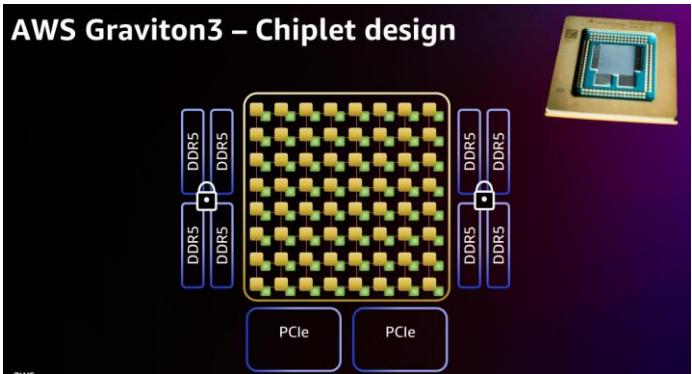
In addition to AI training, Habana has leveraged the same base architecture to expand its footprint in AI inference, with 1st generation Goya processor introduced in 3Q18 manufactured on TSMC's 16nm. The chipset features 8 tensor processor cores (TPC) and a general matrix multiply engine (GEMM) to support a whole array of mixed-precision data types including 8, 16 and 32-bit integer and floating point operations. Unlike training requiring higher bandwidth and a larger capacity of memory, the more cost effective dual-channel DDR4 interface is sufficient for AI inference.

In 2Q22, Habana Labs introduced its 2nd generation AI inference chipset Greco for mass production in 1Q23 on TSMC's 7nm. To enable greater inference speed and efficiency targeting computer vision deployments, the chipset integrates media encoding and processing on-chip, supporting multiple media formats and data types which gives the customers options and flexibility in balancing inference speed and accuracy. Compared with Goya, Greco has upgraded 16GB LPDDR5 memory, offering 5x boost in memory bandwidth and increase in on-chip SRAM from 50MB to 128 MB while reduces power consumption from 200W TDP to 75W TDP through architecture and processor technology migration.

- **Amazon:** The company started to work on its own customized chipset solutions started from 2012 to improve the performance of its AWS cloud computing service and the acquisition of Annapurna Labs, an Israeli start-up company focusing on data center chipset development, in 2015 and the hire of employees of Calxeda, one of the first companies to design ARM based server chipsets, has allowed the company more aggressive on expanding its cloud solutions on both system and chipset level. On the system level, Amazon developed AWS Nitro System as the foundation for EC2 (Elastic Compute Cloud) instances which delivers strong performance and enhanced security while enables the company to support new instances. For the chipsets, the customization started from networking IC in 2016 replacing Broadcom's Tomahawk Ethernet chipset which allows the company to have a more efficient, powerful and reliable fiber network. Throughout the years, Amazon has been expanding its chipsets into CPU, AI inference and AI training.

**Figure 168: Habana's Gaudi2 performance better than Nvidia**

Source: Intel

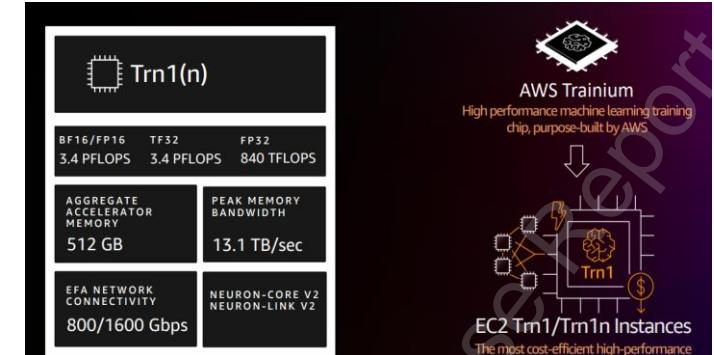
**Figure 169: Amazon's Graviton3 adopts chiplet design****AWS Graviton3 – Chiplet design**

Source: Amazon

For CPU, Amazon introduced Graviton in 4Q18 featuring ARM's Cortex-A72 cores and subsequently launched Graviton2 in 1Q20 and Graviton3 in 2Q22 to upgrade the performance and power efficiency at much lower cost compared with the off-the-shelf solutions from Intel and AMD. For Graviton2 and Graviton3, the CPUs adopt 64 ARM's Neoverse N1 and V1 cores respectively, with Graviton seeing 50% memory bandwidth improvement over Graviton2, PCIe 5.0 upgrade and adoption of chiplet design. According to Amazon, Graviton provides up to 25% better compute performance, 2x higher floating-point performance, 2x faster cryptographic workload performance and 3x better machine learning workload performance compared with Graviton2.

**Figure 170: Graviton3 is widely adopted by aws customers****Customer momentum with AWS Graviton**

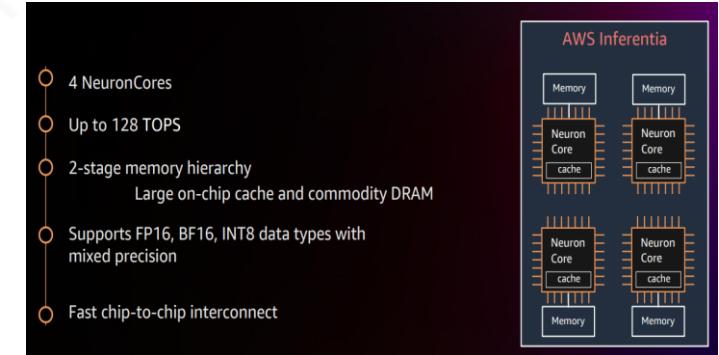
Source: Amazon

**Figure 171: Amazon's Trainium chipset strong on AI training**

Source: Amazon

For AI inference, Amazon first introduced Inferentia chipsets on TSMC's 16nm in 2018 along with its Graviton CPU, featuring 4 Neuron cores performing up to 128 TOPS while the 2<sup>nd</sup> generation Inferentia 2 migrating to TSMC's 7nm in 2022 delivers 3x higher compute performance, 4x higher accelerator memory, 4x higher throughput and 10x lower latency which is optimized for large language models and vision transformers at scale in Amazon's EC2 inf2 instances for the applications including natural language understanding, language translation, video and image generation and speech recognition.

For AI training, in addition to general GPU provided by AMD and Nvidia, Amazon announced the EC2 DL1 instance in 4Q21 powered by 8 Habana Gaudi accelerators which is the 1<sup>st</sup> instance type to include dedicated AI accelerators rather than GPU, delivering up to 40% better price performance than the current generation of GPU based instances. The company also launched its own customized Trainium chipsets ready to be used in EC2 in 4Q22 suitable for training natural language processing, computer vision and recommender models while allowing the customers up to 50% cost saving over comparable GPU-based EC2 instances.

**Figure 172: Key features for Amazon's Inferentia chipsets**

Source: Amazon

In addition to the diversifying AI chipset project pipeline in the US and EU, China hyperscalers and start up companies were encouraged by local government to develop the high-end chipsets to improve semiconductor self-sufficiency amid geopolitical tension. In the past couple of years, China companies have been developing mainstream CPU, GPU and edge AI solutions. The local ecosystem is also making progress in cloud computing semiconductor across general purpose GPU to dedicated ASICs with key projects below although some designs may be restricted on the system and chipset level performance from the US government.

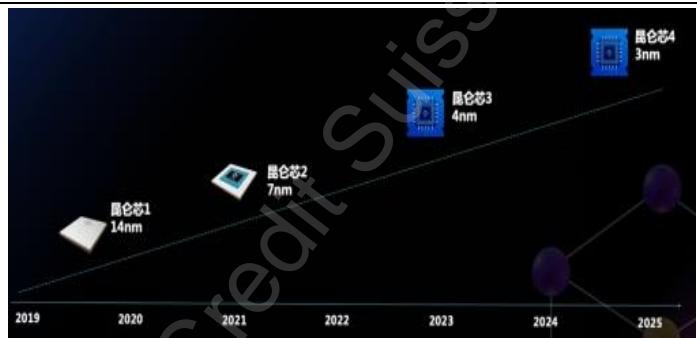
**Figure 173: China AI ASIC solutions**

Company	Baidu	Baidu	Tencent/Enflame	Alibaba/T-Head	Cambricon	Corerain	Biren	Huawei
<b>Chipset</b>	Kunlun Gen 1	Kunlun Gen 2	Enflame 2.5	Hanguang 800	Siyuan 370	CAISA3.0	BR100	Ascend 910
<b>Type</b>	ASIC	ASIC	ASIC	ASIC	ASIC	SIC (AI accelerator)	GPU	SoC
<b>Function</b>	AI inference	AI inference/training	AI training	AI inference	AI inference/training	AI inference	AI inference/training	AI inference/training
<b>Foundry</b>	Samsung	Samsung	GlobalFoundries	TSMC	TSMC	SMIC	TSMC	TSMC
<b>Manufacturing node</b>	14nm	7nm	12nm FinFET	12nm	7nm	28nm	7nm	7nm+
<b>Die size (mm^2)</b>	NA	NA	NA	709	NA	NA	1,074	456
<b>Number of transistors (billion)</b>	NA	NA	21	17	39	NA	77	NA
<b>Density (mn transistors per mm^2)</b>	NA	NA	NA	24	NA	NA	72	NA
<b>AI-optimized cores</b>	NA	NA	NA	4	4	4	16	32
<b>On-board SRAM</b>	16MB	16MB	32MB	192MB	NA	NA	300MB	32MB
<b>Memory bandwidth</b>	256GB/s	512GB/s GDDR6	819GB/s HBM2e	NA	614GB/s	45GB/s DDR4	2300GB/s HBM2e	1.2TB/s HBM2
<b>Bus interface</b>	PCIe 4.0x8	PCIe 4.0x8	PCIe 4.0x16	PCIe 4.0x16	PCIe 3.0x4	PCIe 5.0x16	PCIe 3.0x4	PCIe 3.0x4
<b>ASIC provider</b>	Samsung direct	Samsung direct	NA	NA	NA	Alchip	NA	NA
<b>Launch</b>	1Q20	3Q21	4Q21	3Q19	4Q21	2Q20	3Q22	3Q19
<b>Peak performance (FP32, TFLOPS)</b>	256 TOPS	640 TOPS	32	NA	256 TOPS	NA	128	512 TOPS
<b>TDP (W)</b>	150	150	NA	NA	NA	NA	550	310

Source: Company data, Credit Suisse estimates

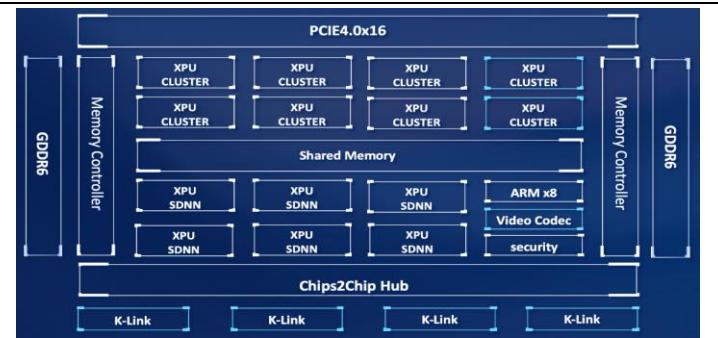
- Baidu:** Baidu started its AI chipset development from 2011 and spun off its chipset business as an independent company Kulunxin and raised funds in 1Q21 to focus on general purpose AI chipset development for deep learning and machine learning adopting natural language processing, visual recognition, recommender engine and computer vision. The company introduced its first general AI chipset Kunlun on Samsung's 14nm in 3Q18 and upgraded to Kulun2 in 3Q21 on Samsung's 7nm featuring its 2nd generation XPU architecture and improving performance by 2-3x vs. first chipset. The company expects the 3<sup>rd</sup> generation Kunlun chipset will be ready for mass production in 2024 on 4nm as the backbone of its AI IaaS layer and planned chip in 2025 on 3nm. Currently the company has more than 10 customers adopting its AI chipsets.

**Figure 174: Baidu has roadmap through 3nm for Kunlun chipset**



Source: Baidu

**Figure 175: Kunlun2 supports GDDR6 and PCIe4.0x16**



Source: Baidu

- Tencent/Enflame:** Enflame Technology was founded in 1Q18 backed by Tencent, National IC Fund and a group of China VCs. The company is focused on the cloud AI solution development across chipsets and hardware. The 1<sup>st</sup> AI training card Yunsui T10 based on its Enflame 2.5 chipset was introduced in 4Q19 and the company launched its 1<sup>st</sup> AI inference product Yunsui i10 in 4Q20.

In 4Q21, the company upgraded the AI accelerators Yunsui i20 based on its Enflame 2.5 chipset as its AI inference solution, featuring 16GB HBM2e memory with memory bandwidth up to 819GB/s and supporting all key formats for the inference platform to deliver performance by 2-4x vs. 1<sup>st</sup> generation product. The Enflame 2.5 chipset is manufactured on GlobalFoundries' 12nm FinFET process which the company claims to have comparable efficiency to mainstream GPU on 7nm.

On top of AI inference solution upgrade, Enflame also introduced its 2<sup>nd</sup> generation AI training accelerator Yunsui T20 based on Enflame 2.0 chipset also manufactured on GlobalFoundries' 12nm with computing power reaching 40TFLOPS on FP32, 1.6x better performance compared with 1<sup>st</sup> generation chipset. The company targets to introduce its 3<sup>rd</sup> generation AI chipsets in 2023.

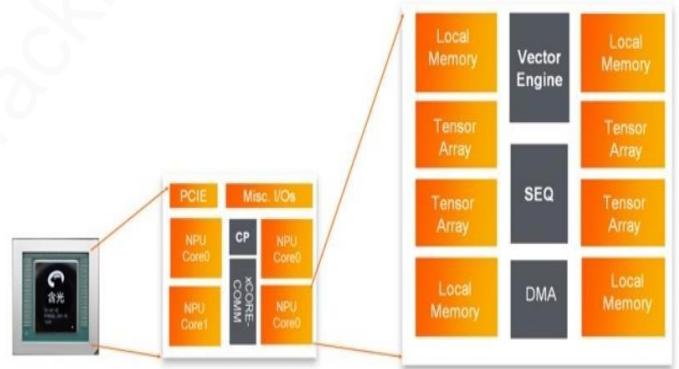
- **Alibaba:** the company's semiconductor design business, T-Head Semiconductor, was founded in 3Q18 with a focus on the advanced chipset development across cloud and edge AI, CPU and semiconductor IP. In 3Q19, the company introduced its 1<sup>st</sup> AI inference chipset Hanguang 800 for neural networking applications. According to Alibaba, Hanguang 800 has been deployed in its data center for e-commerce platform performance enhancement including product search, language translation and product recommendation.

**Figure 176: Alibaba claims solid performance for Hanguang 800**



Source: Alibaba

**Figure 177: Alibaba Hanguang 800 structure**



Source: Alibaba

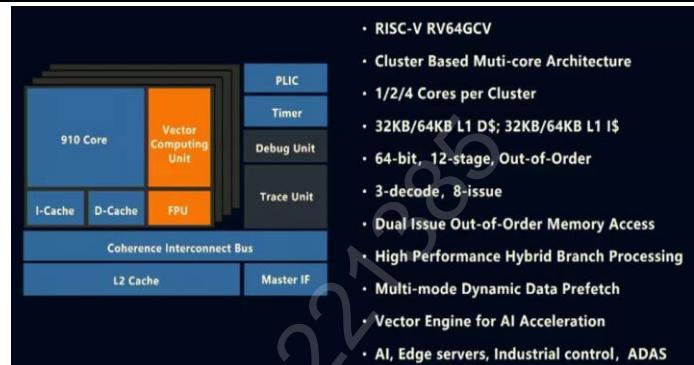
For CPU, the company introduced ARM based Yitian 710 chipset in 4Q21 based on TSMC's 5nm and packs 128 ARMv9 cores with clock speeds as high as 3.20 GHz. It has eight DDR5-4800 memory channels that can produce up to 307.2 Gbps of transfer speed and 96 PCIe 5.0 lanes which the company claims 30% improvement in cost performance and TDP down by 60% compared with the similar products. In addition to ARM CPU, the company also introduced its 1<sup>st</sup> RISC-V CPU in 3Q19 Xuentie E902 and upgraded the RSIC-V roadmap through Xuentie C908 as of 2022 to deliver 3.5x better performance on graph classification and neural networking performance by 50% compared with C906 though performance is still behind the mainstream x86 and ARM architecture based solutions. With the pipeline across ARM and RISC-V based CPUs, Alibaba targets to deploy its own CPUs for 20% of the computing power it adds going forward.

Figure 178: Alibaba's Yitian ARM CPU features TSMC's 5nm



Source: Alibaba

Figure 179: Alibaba introduced its RISC-V CPU portfolio



Source: Alibaba

- **Cambricon:** the company was founded in 2016 with a focus on AI chipset development and was listed on STAR Board in 3Q20. The company's chipset roadmap has been expanding from cloud to edge applications including smartphones, automotive. In 2016, Cambricon introduced its 1<sup>st</sup> semiconductor IP Cambricon-1A dedicated for deep learning in ARM devices and was adopted by Huawei in its Kirin 970 mobile SoC. In 2018, the company introduced its high performance neural processors MLU100 and MLU200 on TSMC's 16nm for both AI training and inference applications.

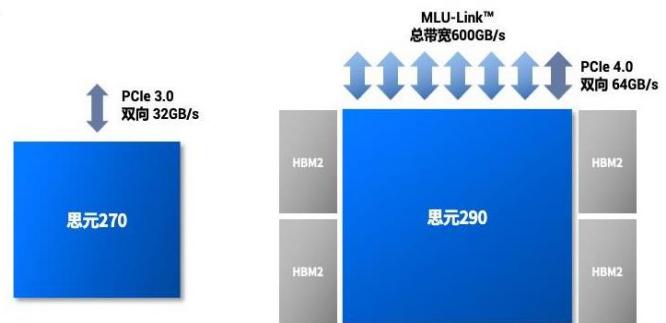
Figure 180: Cambricon's architecture migration supports more features



Source: Cambricon

In 2021, the company launched its 1<sup>st</sup> AI training chipset Siyuan 290 on TSMC's 7nm, integrating 46bn transistors and at the same time introduced its MLU290-M5 accelerator adopting Siyuan 290 chipset, featuring 64 MLU cores and 1.23TB/s memory bandwidth. Cambricon also introduced its 1<sup>st</sup> AI training chiplet Siyuan370 on TSMC's 7nm in 4Q21 featuring 39bn transistors, LPDDR5 support, MLU-Fabric interconnect to enable high speed connection between the dies and up to 256TOPS AI computing power, 100% improvement compared with Siyuan270 on TSMC's 16nm it launched in 2Q19. The MLU370-S4 accelerator with Siyuan370 can support 2x the performance based on ResNet-50 test.

Figure 181: Cambricon's Siyuan290 chipset supports HBM2



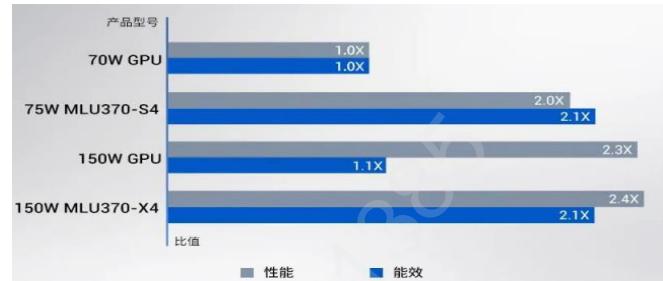
Source: Cambricon

**Figure 182: Cambricon shifts its AI chipsets toward chiplet tiles**

Source: Cambricon

In addition to the core cloud chipset business, Cambricon is turning more aggressive on the autonomous driving chipsets started from 2022, with 3 projects under development spanning L2-L4 functions with different computing power from 10TOPS to 1000TOPs. However, with US government putting the company on the restriction subject to Foreign Direct Product Rule, the company's access to the advanced manufacturing and design support is going to be limited.

- Biren:** Biren was founded in 2019 with senior management from Nvidia and Alibaba focused on general purpose GPUs and received more renown after presenting at Hot Chips an advanced spec that was later declared above the threshold of compute allowed for China suppliers use of US tools/technology in fabrication. The company joins a host of GPU start-ups in China mainly using TSMC's 7nm and 16nm process nodes.

**Figure 183: Cambricon MLU370-X4 improves power/perf.**

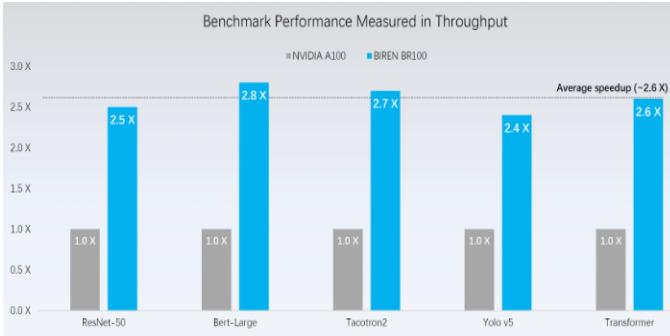
Source: Cambricon

**Figure 184: China GPU companies targeting market entry**

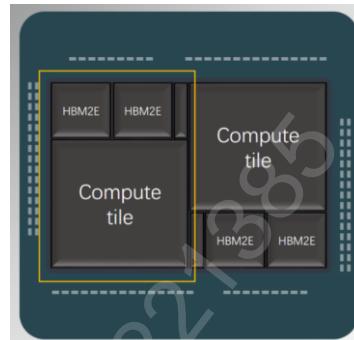
China GPU Vendor	Target Market	Product Name	Process	Base	Founded	Founder or CEO	Funding	Website
Shanghai Tianshu Zhixin Semiconductor Company	Server GPU	"Big Island"	7nm TSMC	Shanghai	2013	Cai Quangen	US\$186M	iluvata.com.cn
Huawei Korea Cloud and AI Business Group (under Huawei Korea Enterprise Division)	Server GPU			Seoul	2020	Richard Yu		
MetaX Integrated Circuits Company (Metax)	Server GPU		5nm TSMC	Shanghai	2020	William Chen	Seed funding in 2020 and 2021	metax-tech.com
Changsha Jianjia Microelectronics	Desktop PC GPU	JM5000, JM7000, JM9000	28nm	Shenzhen	2006	Zeng Wanhui	2020 revenue: US\$80.7M	jingjiamicro.com
Shanghai Zhaoxin Semiconductor Company	Mobile PC GPU		16nm TSMC	Shanghai	2013	Ye Jun		zhaoxin.com
Biren Technology	Server, PC GPU	BR100	7nm TSMC	Shanghai	2019	Zhang Wen	US\$715M	birentech.com
Innosilicon		Fantasy One						innosilicon.cn

Source: IDC

Biren has been focusing on the general-purpose GPU design and introduced its 1<sup>st</sup> chipset BR100 on TSMC's 7nm in 3Q22, featuring 77bn transistors in 1,074mm<sup>2</sup> die size, 64GB HBM2e memory supporting 1.64TB/s bandwidth and 819GB/s I/O speed. The company claimed the chipset could deliver up to 256 FP32 TFLOPS performance, competitive to Nvidia's A100 GPU in certain workloads. The mid-range BR104 GPU introduced by Biren has lower spec at 32GB of HBM2e memory and half of the performance delivered by BR100. However, with the US government restricting China semiconductor performance below 500GB/s bidirectional transfer rate, Biren was forced to lower the spec of the chipset to ensure continued design and manufacturing support.

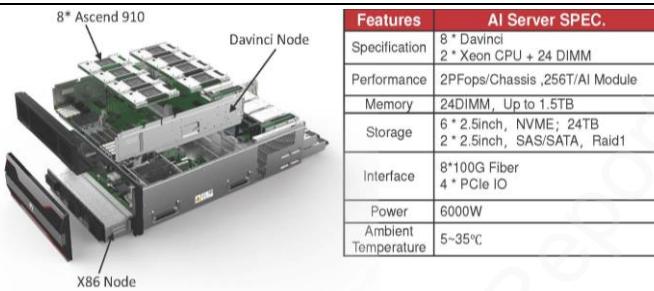
**Figure 185: Biren claims its BR100 has better performance**

Source: Biren

**Figure 186: Biren BR100 architecture**

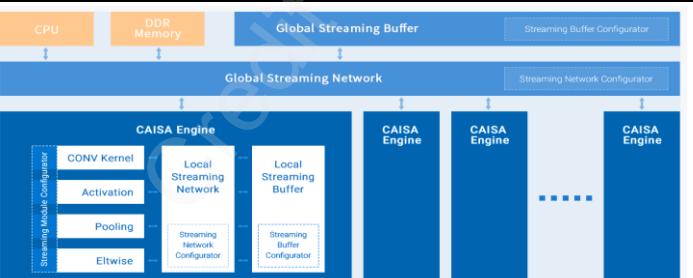
Source: Biren

- Huawei:** Huawei announced its AI strategy in 4Q18 committed to provide a full-stack all-scenario AI portfolio and introduced its Ascend 910 AI chipset on TSMC's 7nm in 3Q19, featuring Huawei's own Da Vinci architecture, integrating CPUs, DVPP and task scheduler and supporting PCIe 4.0, RoCE interconnects and its own high speed interface interconnecting multiple Ascend 910 chipsets. The company claims the chipset could deliver 256TFLOPS (FP16) computing power with 310W TDP power budget. However, since 2020, with Huawei and its affiliate companies restricted by the US government, the company lost access to the foreign technology support on design and manufacturing.

**Figure 187: Huawei's AI server based on its Ascend910**

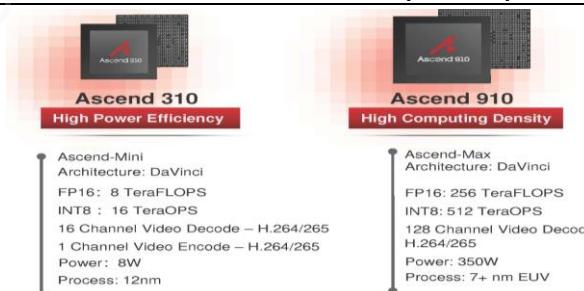
Source: Huawei

- Corerain:** The company was founded in 2016 with a focus on the AI accelerator for edge and cloud computing applications. The company introduced its CAISA chipsets with the latest generation CAISA3.0 supporting 10.9 TOPS computing power at the peak. Although the chipsets are built on the legacy 28nm, Corerain claims the chipset utilization is optimized at 95.4% and would be cost and power efficient for the AI inference tasks.

**Figure 189: Corerain's CAISA chipset architecture**

Source: Corerain

Regarding the choice of “custom vs. merchant vs. general purpose silicon for AI”, we see opportunities for both within the high growth AI space. Custom solutions can be optimized to provide higher performance for well-defined workloads requiring high volume. General purpose

**Figure 188: Ascend910 vs. Ascend310 spec comparison**

Source: Huawei

**Figure 190: Corerain's latest CAISA 3.0 chipset**

Source: Corerain

solutions such as NVDA are best for less well-defined problems that require flexibility. Given the early stage of the AI market, we believe most AI workloads fit into that category, as evidenced by NVDA's revenue as compared to the market for custom AI silicon. Net, while we expect custom solutions to grow, and perhaps to even grow faster than the general-purpose market as AI matures, we are at such early stages of market development this doesn't concern us.

#### **Marvell's AI ramping through custom ASIC, DPU and optical silicon**

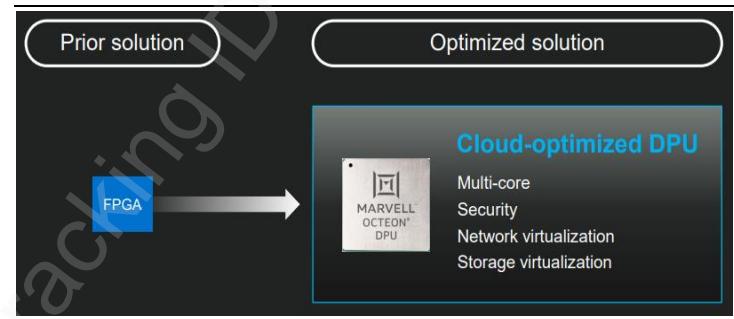
MRVL's exposure to AI is mainly through their custom ASIC business, through which they assist hyperscalers in developing custom silicon, some of which is relevant to AI workloads. MRVL expects to have \$400m in incremental revenue from these custom projects in FY24. MRVL sees that doubling to \$800m in FY25. MRVL noted that in aggregate, they have won over a dozen cloud optimized programs across multiple Tier 1 cloud customers. A number of these designs are for custom DPU implementations inside cloud data centers.

**Figure 191: Hyperscale customers have unique requirements...**

	Search optimized	Social media optimized
Compute	High-performance CPU cores AI/ML acceleration	Many small CPU cores Video acceleration
Electro-optics	Highest bandwidth	Optimized bandwidth
Networking	High-bandwidth switch	High-port-count switch
Storage	Low latency	Highest capacity

Source: Marvell

**Figure 192: Marvell developing custom and optimized solutions for cloud data center.**



Source: Marvell

#### **AI could eventually be a catalyst for the Edge compute suppliers, but too early to re-rate Qualcomm and Mediatek**

While the primary scope of the report is on AI training and inference from the new use cases for chat GPT and large language models, we also highlight edge AI engines will emerge as a growth driver for suppliers of edge processors in devices including Qualcomm and Mediatek. The AI engine traditionally was added for enhancements to photos and video but increasingly can serve to store trained models for better real time inference, driving upgrades to the processor requirements, size of the AI engine and storage requirements in edge devices.

**Figure 193: AI Integration Expanding in Edge Processors**

AI Application processors ramp-up	2020	2021	2022	2023	2024	2025	2026	20-23	23-26
AI App. Processor units mn	598	796	968	1,153	1,293	1,421	1,528	24%	10%
Industry App. Processor units mn	3,195	3,323	3,216	3,196	3,420	3,575	3,705	0%	5%
<b>AI AP % of industry units</b>	<b>19%</b>	<b>24%</b>	<b>30%</b>	<b>36%</b>	<b>38%</b>	<b>40%</b>	<b>41%</b>		
AI App. Processor ASPs US\$	\$29.7	\$33.7	\$34.1	\$33.2	\$33.4	\$32.4	\$31.8	4%	-1%
Industry App. Processor ASPs US\$	\$14.1	\$17.4	\$20.6	\$20.5	\$20.8	\$21.1	\$21.2	13%	1%
<b>AI Processor Premium</b>	<b>111%</b>	<b>94%</b>	<b>66%</b>	<b>62%</b>	<b>61%</b>	<b>54%</b>	<b>51%</b>		
AI Application Processor Sales US\$m	\$17,773	\$26,810	\$33,037	\$38,229	\$43,134	\$45,956	\$48,640	29%	8%
Industry App. Processor Sales US\$m	\$45,009	\$57,847	\$66,127	\$65,428	\$70,975	\$75,275	\$78,352	13%	6%
<b>AI AP % of industry revenue</b>	<b>39%</b>	<b>46%</b>	<b>50%</b>	<b>58%</b>	<b>61%</b>	<b>61%</b>	<b>62%</b>		

Source: Gartner April 2022

According to Gartner, it projects the AI accelerated application processors will more than double from 19% to 40% of industry processor units, with the AI processors increasing in ASPs from US\$30 to US\$32, a premium to overall application processors US\$20 ASP. That penetration would allow AI application processors to increase at from US\$33bn to US\$49bn and 8% CAGR, helping expand the overall application processor market at 6% CAGR to US\$78bn. That high single digit growth rate for the edge processors would be upside to some market perception mobile processors are just about 4G to 5G upgrades and already an ex-growth market.

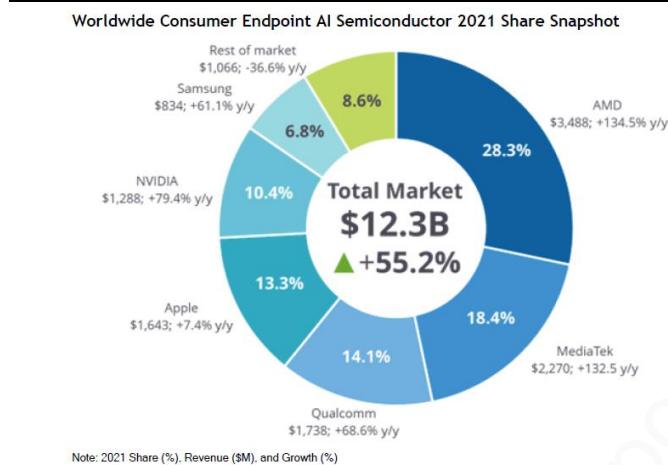
AI would also be integrated in other edge processors to drive overall 14% CAGR as it adds content across embedded processors, MCUs and FPGAs.

**Figure 194: AI integrated into more embedded processors**

AI Integrated sales \$m	2020	2021	2022	2023	2024	2025	2026	20-23	23-26	% of 23
Discrete Application/Multimedia Processor	\$10,771	\$14,440	\$15,725	\$18,423	\$18,935	\$18,366	\$19,371	20%	2%	40%
FPGA	\$11	\$36	\$106	\$300	\$608	\$961	\$1,455	199%	69%	1%
Integrated Baseband/Application Processor	\$7,002	\$12,371	\$17,312	\$19,806	\$24,199	\$27,589	\$29,270	41%	14%	43%
Microcontroller (8/16/32 bit)	\$19	\$49	\$106	\$212	\$286	\$473	\$755	123%	53%	0%
Microprocessor - Compute	\$1,128	\$2,107	\$3,510	\$5,836	\$7,669	\$9,455	\$11,360	73%	25%	13%
Microprocessor - Embedded	\$40	\$86	\$156	\$291	\$475	\$683	\$881	94%	45%	1%
Other Application Specific	\$66	\$182	\$459	\$1,087	\$2,085	\$3,144	\$4,809	154%	64%	2%
<b>Grand Total</b>	<b>\$19,037</b>	<b>\$29,270</b>	<b>\$37,374</b>	<b>\$45,954</b>	<b>\$54,256</b>	<b>\$60,672</b>	<b>\$67,901</b>	<b>34%</b>	<b>14%</b>	<b>100%</b>
<b>YoY Growth</b>		<b>54%</b>	<b>28%</b>	<b>23%</b>	<b>18%</b>	<b>12%</b>	<b>12%</b>			

Source: Gartner April 2022

Companies leveraged to this edge compute market for consumer include AMD (counting its gaming processors which also process AI algorithms in game play), Qualcomm (AR/VR, tablets, smart home and wearables, Mediatek (smart TVs, Alexa Voice Assistants), Apple (Apple TV, iPads) and NVIDIA (Tegra in Switch and Shield).

**Figure 195: AI consumer edge semiconductors saw strong growth during the pandemic**

Source: IDC

### Power management - Monolithic Power has strong share in power management on AI boards

MPWR supplies the power management for both A-100 and H-100 GPUs. For the A-100, we believe MPWR has the majority of the market share, with VICR having a minority share. GPU power is a part of MPWR's enterprise data segment, which represented X of MPWR total revenue in CY22. Looking forward, we expect that MPWR will initially have 100% share of H-100 power, since that solution requires multi-phase power delivery that VICR currently does not supply. We don't expect MPWR to keep 100% share, but still expect that to be a strong growth driver in CY23 and CY24.

## Asia Semiconductors

### AI to drive a further inflection in TSMC's HPC business

TSMC's high performance computing segment has been its strongest growth driver and poised for continued growth from an inflection in AI use cases. The company grew this business at a 21% CAGR from 2015-20 versus +11% for the overall company and now on pace off our estimates for similar +23% CAGR versus +16% CAGR for the company from 2015-20. The high pace of growth for this category has driven it to double from 21% of TSMC's sales in 2016 to 42% of sales in 2023, while reliance on smartphones has dipped from 51% to 37% during the period. Its smaller buckets including IoT and auto have increased from 11% to 15%.

**Figure 196: AI end points led by AMD in consoles, MTK in smart speakers/TVs**

Vendor	2020		2021	
	Revenue (\$M)	Share (%)	Revenue (\$M)	Share (%)
AMD	1,487.8	18.7	3,488.2	28.3
MediaTek	976.1	12.3	2,269.6	18.4
Qualcomm	1,031.2	13.0	1,738.4	14.1
Apple	1,530.0	19.3	1,642.8	13.3
NVIDIA	718.0	9.0	1,287.9	10.4
Samsung	518.0	6.5	834.3	6.8
Other	1,680.8	21.2	1,065.8	8.6
Total	7,941.9	100.0	12,327.1	100.0

Source: IDC

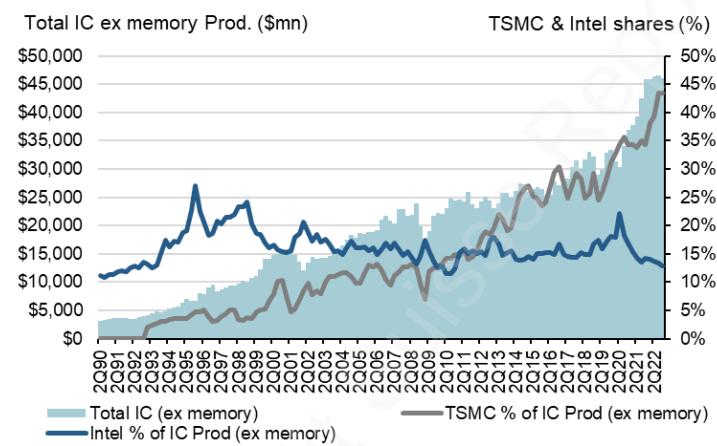
Randy Abrams  
Haas Liu  
Angela Dai

**Figure 197: TSMC's high performing compute has doubled to 44% of sales**

TSMC growth drivers	2015	2016	2017	2018	2019	2020	2021	2022E	2023E	2024E	2025E	15-20 CAGR	20-25 CAGR
HPC - TSMC	\$5.9	\$6.1	\$8.1	\$10.9	\$10.3	\$15.0	\$21.0	\$31.3	\$31.6	\$37.0	\$42.0	21%	23%
HPC - Market	\$32.8	\$33.7	\$37.8	\$40.1	\$39.2	\$43.0	\$52.3	\$51.5	\$51.2	\$58.6	\$63.4	6%	8%
<b>TSMC share (%)</b>	<b>17.9%</b>	<b>18.1%</b>	<b>21.5%</b>	<b>27.3%</b>	<b>26.3%</b>	<b>35.0%</b>	<b>40.1%</b>	<b>60.8%</b>	<b>61.8%</b>	<b>63.1%</b>	<b>66.3%</b>		
Automotive - TSMC	\$1.1	\$1.3	\$1.4	\$1.7	\$1.5	\$1.4	\$2.29	\$3.8	\$3.98	\$4.8	\$5.7	6%	32%
Automotive - Market	\$9.5	\$11.5	\$13.0	\$14.2	\$13.9	\$12.9	\$15.7	\$17.8	\$19.9	\$22.0	\$24.2	6%	13%
<b>TSMC share (%)</b>	<b>11.0%</b>	<b>11.7%</b>	<b>10.8%</b>	<b>12.0%</b>	<b>11.1%</b>	<b>11.2%</b>	<b>14.5%</b>	<b>21.3%</b>	<b>20.1%</b>	<b>21.7%</b>	<b>23.7%</b>		
IoT - TSMC	\$1.6	\$1.7	\$1.8	\$2.1	\$2.7	\$3.8	\$4.94	\$6.7	\$6.94	\$7.6	\$8.40	19%	17%
IoT - Market	\$9.0	\$9.93	\$10.8	\$12.2	\$13.6	\$16.2	\$20.4	\$23.5	\$26.4	\$29.5	\$33.1	12%	15%
<b>TSMC share (%)</b>	<b>17.5%</b>	<b>16.6%</b>	<b>16.7%</b>	<b>16.8%</b>	<b>19.7%</b>	<b>23.5%</b>	<b>24.3%</b>	<b>28.3%</b>	<b>26.3%</b>	<b>25.9%</b>	<b>25.4%</b>		
Mobile - TSMC	\$12.9	\$15.1	\$16.3	\$15.6	\$16.9	\$21.6	\$24.8	\$29.9	\$27.9	\$31.6	\$33.8	11%	9%
Mobile - Market	\$30.8	\$30.9	\$31.3	\$30.7	\$30.5	\$36.3	\$44.1	\$45.6	\$43.8	\$48.6	\$51.7	3%	7%
<b>TSMC Mobile share (%)</b>	<b>42.0%</b>	<b>48.9%</b>	<b>51.9%</b>	<b>50.9%</b>	<b>55.3%</b>	<b>59.5%</b>	<b>56.2%</b>	<b>65.5%</b>	<b>63.6%</b>	<b>64.9%</b>	<b>65.4%</b>		
Digital Consumer	\$2.4	\$2.6	\$2.5	\$2.3	\$1.8	\$1.8	\$1.9	\$1.9	\$1.90	\$2.0	\$2.2	-5.8%	4.3%
Other	\$2.8	\$2.6	\$2.0	\$1.6	\$1.4	\$1.9	\$1.9	\$2.3	\$2.3	\$2.4	\$2.5	-7.9%	5.7%
<b>TSMC CS Estimates</b>	<b>\$26.6</b>	<b>\$29.4</b>	<b>\$32.1</b>	<b>\$34.2</b>	<b>\$34.6</b>	<b>\$45.5</b>	<b>\$56.8</b>	<b>\$75.9</b>	<b>\$74.7</b>	<b>\$85.4</b>	<b>\$94.6</b>	<b>11.3%</b>	<b>15.8%</b>
<b>YoY Growth</b>	<b>10.6%</b>	<b>9.1%</b>	<b>6.5%</b>	<b>1.3%</b>	<b>31.4%</b>	<b>24.9%</b>	<b>33.5%</b>	<b>-1.6%</b>	<b>14.4%</b>	<b>10.8%</b>			

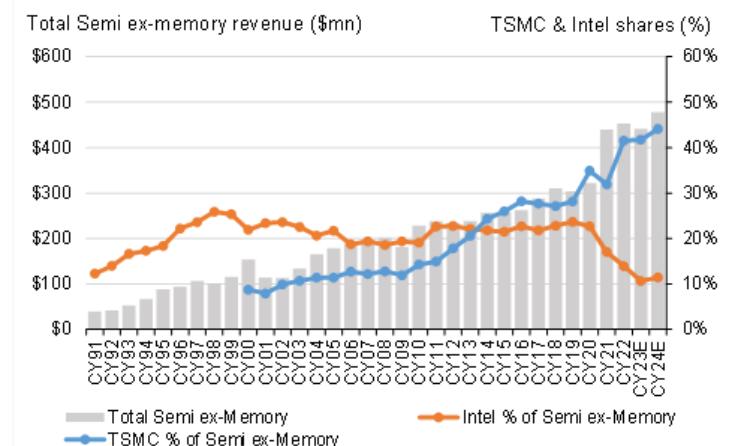
Source: Company Data, Credit Suisse Estimates

TSMC's share gains and growth through its customers have enabled it to capture rising share of industry production, with a step-up in the past few years COVID-19 up-cycle to now approach 45% of semis-ex-memory production and similar revenue share when its production is grossed up to its customers' chip sales. Intel's share of industry held fairly stable in a 15-20% range from 2000-2020 but has also since dipped from 20% to low-teens. The ability to capture the compute processor and AI TAM will be critical for market share, though a growing TAM would provide opportunities for both in a stable share condition and not treated as a zero-sum game.

**Figure 198: TSMC's share rising of industry production, Intel's dipped in the recent cycle**

Source: Mercury Research

TSMC's HPC segment includes its CPU for server and client PCs and tablets, GPU for AI, graphics and gaming, FPGAs, network processors and PC peripherals. The company has benefited from high growth and share gains in AI, GPU/Gaming and CPUs which we estimate has increased its share of the HPC TAM from 18% in 2016 to 61% in 2022.

**Figure 199: TSMC's share rising when grossed up to logic industry revenue**

Source: Mercury Research

**Figure 200: TSMC's largest HPC drivers from server/client CPU, AI and GPU/gaming**

HPC Silicon Market	2015	2016	2017	2018	2019	2020	2021	2022	2023E	2024E	2025E	15-20 CAGR	20-25 CAGR
Servers and CPUs	\$47.8	\$48.2	\$51.4	\$58.0	\$59.3	\$66.4	\$71.8	\$61.0	\$57.2	\$70.8	\$75.0	7%	2%
Graphics and Gaming	\$6.7	\$8.2	\$10.8	\$11.3	\$9.6	\$12.2	\$21.3	\$17.9	\$17.2	\$18.6	\$20.1	13%	10%
AI Accelerators	\$0.34	\$0.8	\$1.9	\$3.2	\$3.3	\$7.0	\$11.5	\$16.1	\$19.1	\$24.5	\$29.9	83%	34%
Cryptocurrency	\$0.3	\$0.4	\$2.5	\$3.3	\$1.0	\$1.5	\$4.0	\$5.0	\$2.0	\$2.0	\$2.0	38%	6%
Programmable Logic	\$4.4	\$4.3	\$4.5	\$4.7	\$5.3	\$5.5	\$5.6	\$7.2	\$7.8	\$8.4	\$9.0	5%	10%
Networking & Infrastructure	\$13.0	\$13.2	\$14.1	\$14.6	\$15.6	\$17.2	\$18.89	\$21.7	\$23.2	\$24.9	\$26.6	6%	9%
Computing Peripherals	\$8.0	\$7.8	\$7.7	\$7.5	\$7.4	\$8.5	\$9.76	\$10.2	\$10.5	\$10.7	\$10.9	1%	5%
<b>HPC - Market</b>	<b>\$80.5</b>	<b>\$83.0</b>	<b>\$92.9</b>	<b>\$102.6</b>	<b>\$101.5</b>	<b>\$118.3</b>	<b>\$142.9</b>	<b>\$139.1</b>	<b>\$137.0</b>	<b>\$159.7</b>	<b>\$173.5</b>	<b>8%</b>	<b>8%</b>
<b>YoY Growth</b>	<b>3%</b>	<b>12%</b>	<b>10%</b>	<b>-1%</b>	<b>17%</b>	<b>21%</b>	<b>-3%</b>	<b>-2%</b>	<b>17%</b>	<b>9%</b>			
HPC Wafer Opportunity	2015	2016	2017	2018	2019	2020	2021	2022	2023E	2024E	2025E	15-20 CAGR	20-25 CAGR
Servers and CPUs	\$21.1	\$21.3	\$22.7	\$23.9	\$24.5	\$24.9	\$26.9	\$24.2	\$23.6	\$28.1	\$29.8	3%	4%
Graphics and Gaming	\$2.5	\$3.1	\$4.1	\$4.2	\$3.6	\$4.6	\$8.0	\$6.7	\$6.4	\$7.0	\$7.5	13%	10%
AI Accelerators	\$0.1	\$0.2	\$0.6	\$1.0	\$1.0	\$2.1	\$3.4	\$4.8	\$5.7	\$7.3	\$9.0	83%	34%
Cryptocurrency	\$0.2	\$0.2	\$1.3	\$1.7	\$0.5	\$0.8	\$2.1	\$2.6	\$1.0	\$1.0	\$1.0	38%	6%
Programmable Logic	\$1.1	\$1.1	\$1.1	\$1.1	\$1.2	\$1.2	\$1.2	\$1.4	\$2.0	\$2.2	\$2.4	2%	15%
Networking & Infrastructure	\$4.4	\$4.5	\$4.8	\$4.9	\$5.3	\$5.8	\$6.4	\$7.3	\$7.8	\$8.4	\$9.0	6%	9%
Computing Peripherals	\$3.4	\$3.4	\$3.3	\$3.2	\$3.2	\$3.6	\$4.2	\$4.4	\$4.5	\$4.6	\$4.6	1%	5%
<b>HPC - Market</b>	<b>\$32.8</b>	<b>\$33.7</b>	<b>\$37.8</b>	<b>\$40.1</b>	<b>\$39.2</b>	<b>\$43.0</b>	<b>\$52.3</b>	<b>\$51.5</b>	<b>\$51.2</b>	<b>\$58.6</b>	<b>\$63.4</b>	<b>6%</b>	<b>8%</b>
<b>YoY Growth</b>	<b>3%</b>	<b>12%</b>	<b>6%</b>	<b>-2%</b>	<b>10%</b>	<b>22%</b>	<b>-1%</b>	<b>-1%</b>	<b>15%</b>	<b>8%</b>			
TSMC estimates (CS)	2015	2016	2017	2018	2019	2020	2021	2022	2023E	2024E	2025E	15-20 CAGR	20-25 CAGR
Servers and CPUs	\$0.0	\$0.0	\$0.0	\$0.3	\$1.2	\$3.4	\$5.8	\$10.9	\$11.7	\$14.3	\$16.1	NM	36%
Graphics and Gaming	\$1.9	\$1.90	\$2.4	\$3.6	\$3.1	\$3.7	\$4.3	\$5.6	\$5.7	\$6.1	\$6.8	14%	13%
AI Accelerators	\$0.1	\$0.2	\$0.6	\$1.0	\$1.0	\$2.1	\$3.4	\$4.8	\$5.7	\$7.3	\$9.0	83%	34%
Cryptocurrency	\$0.1	\$0.2	\$1.3	\$1.6	\$0.34	\$0.6	\$1.6	\$2.2	\$0.5	\$0.7	\$1.0	38%	10%
Programmable Logic	\$0.7	\$0.7	\$0.7	\$0.7	\$0.6	\$0.6	\$0.8	\$1.1	\$1.2	\$1.3	\$1.4	-1%	16%
Networking & Network Processors	\$1.6	\$1.6	\$1.8	\$2.0	\$2.37	\$2.5	\$2.9	\$4.0	\$4.32	\$4.6	\$5.0	9%	15%
Computing Peripherals	\$1.4	\$1.4	\$1.4	\$1.7	\$1.70	\$2.0	\$2.1	\$2.7	\$2.54	\$2.7	\$2.80	7%	7%
<b>HPC - TSMC</b>	<b>\$5.8</b>	<b>\$6.1</b>	<b>\$8.1</b>	<b>\$10.9</b>	<b>\$10.3</b>	<b>\$15.0</b>	<b>\$21.0</b>	<b>\$31.3</b>	<b>\$31.6</b>	<b>\$37.0</b>	<b>\$42.0</b>	<b>21%</b>	<b>23%</b>
<b>YoY Growth</b>	<b>5%</b>	<b>33%</b>	<b>35%</b>	<b>-6%</b>	<b>46%</b>	<b>40%</b>	<b>49%</b>	<b>1%</b>	<b>17%</b>	<b>14%</b>			
<b>TSMC share of HPC production</b>	<b>18%</b>	<b>18%</b>	<b>21%</b>	<b>27%</b>	<b>26%</b>	<b>35%</b>	<b>40%</b>	<b>61%</b>	<b>62%</b>	<b>63%</b>	<b>66%</b>		

Source: IDC

TSMC's gains have come from several components including 1) share gains adding AMD processors from GlobalFoundries starting from 2019 and Apple Mac processors from Intel starting from 2022, 2) high share of AI acceleration and networking through NVIDIA GPU, Marvell and Broadcom networking switch, and FPGA/ASICs including Xilinx, Google, Broadcom, Amazon, Cerebras, Alibaba, and multiple start-ups (Ampere, Biren, Cerebras, Graphcore, Tencent's Emblaze), 3) recovery of gaming share with NVIDIA GPU from Samsung, and 4) higher ASPs passing on higher manufacturing cost and complexity along with ability to pass that cost on through its process/ecosystem advantage and time to market leadership.

**Figure 201: TSMC processor drivers by customer**

Processor Revenue	2020	2021	2022	2023E	2024E	2025E	20-25 CAGR
AMD PC/Server	\$1,761	\$2,805	\$3,563	\$3,297	\$4,016	\$4,513	21%
Intel CPU/Peripherals	\$692	\$570	\$1,409	\$1,858	\$3,118	\$3,506	38%
Apple Mac/Tablet	\$29	\$1,229	\$4,300	\$4,291	\$4,671	\$5,138	181%
Alchip	\$183	\$315	\$378	\$596	\$653	\$751	33%
GUC	\$51	\$94	\$94	\$279	\$315	\$363	48%
Amazon	\$132	\$155	\$221	\$255	\$293	\$337	21%
Google	\$112	\$132	\$189	\$217	\$250	\$287	21%
Broadcom	\$205	\$242	\$345	\$397	\$456	\$525	21%
Ampere, Others	\$259	\$305	\$436	\$501	\$576	\$662	21%
<b>TSMC CPU Sales</b>	<b>3,424</b>	<b>5,847</b>	<b>10,935</b>	<b>11,691</b>	<b>14,348</b>	<b>16,082</b>	<b>36%</b>
<b>YoY Growth</b>	<b>184%</b>	<b>71%</b>	<b>87%</b>	<b>7%</b>	<b>23%</b>	<b>12%</b>	

Source: Company Data, Credit Suisse Estimates

We estimate the company's processor and GPU/AI opportunity through those share gains and market growth at 36%/34% CAGR through 2025 to help drive the overall HPC category. While growth was strong through 2020-22, we still see growth engines from further share gains adding the rest of NVIDIA's gaming GPUs through 2023, further AMD share gains in 2023-24 until Intel's Granite Rapids in late 2024 is introduced on Intel 3, and ramp of TSMC built client tiles for Intel starting with 5nm for Meteor Lake in 2H23 and upgrading to 3nm for Arrow Lake in 2H23.

**Figure 202: TSMC has captured high share of the graphics and AI revenue**

Graphics & AI Revenue	2020	2021	2022	2023E	2024E	2025E	15-20 CAGR	20-25 CAGR
NVIDIA Gaming/Professional/OEM	\$9,443	\$15,732	\$11,069	\$10,657	\$13,054	\$15,002	17%	10%
AMD Graphics & Consoles	\$2,776	\$5,607	\$6,805	\$6,542	\$7,019	\$7,160	4%	21%
<b>Total GPU market</b>	<b>\$12,219</b>	<b>\$21,339</b>	<b>\$17,874</b>	<b>\$17,198</b>	<b>\$20,074</b>	<b>\$22,162</b>	<b>13%</b>	<b>13%</b>
YoY Growth	27%	75%	-16%	-4%	17%	10%		
<b>NVIDIA Data Center</b>	<b>\$6,696</b>	<b>\$10,616</b>	<b>\$15,002</b>	<b>\$17,851</b>	<b>\$22,871</b>	<b>\$24,015</b>	<b>82%</b>	<b>29%</b>
YoY Growth	124%	59%	41%	19%	28%	5%		
Other accelerators	\$335	\$849	\$1,050	\$1,250	\$1,601	\$5,884	NM	77%
<b>Data Center Acceleration</b>	<b>\$7,032</b>	<b>\$11,466</b>	<b>\$16,053</b>	<b>\$19,101</b>	<b>\$24,473</b>	<b>\$29,899</b>	<b>83%</b>	<b>34%</b>
YoY Growth	114%	63%	40%	19%	28%	22%		

Source: Company Data, Credit Suisse Estimates

TSMC in the mid-term could also upside on the AI opportunity with faster penetration into servers relative to Yole's projected 10% penetration in 2023. A scenario of doubling AI penetration to 20% of servers would grow AI accelerators from US\$5.5bn to US\$11.0bn and add NT\$2.29 to TSMC EPS, about US\$30-35 on TSMC's share price at current 15-17x multiple.

**Figure 203: AI penetration ramp to 20% doubles TSMC's contribution to 12% of sales**

	Server Units / 2024 Growth %							
	15.9	16.4	16.9	17.4	17.9	18.4	18.9	
	-3%	0%	3%	6%	9%	12%	15%	
AI Penetration	8.0%	\$4,004	\$4,128	\$4,251	\$4,375	\$4,499	\$4,623	\$4,747
	10.0%	\$5,005	\$5,159	\$5,314	\$5,469	\$5,624	\$5,779	\$5,933
	12.0%	\$6,006	\$6,191	\$6,377	\$6,563	\$6,749	\$6,934	\$7,120
	14.0%	\$7,007	\$7,223	\$7,440	\$7,657	\$7,873	\$8,090	\$8,307
	16.0%	\$8,008	\$8,255	\$8,503	\$8,750	\$8,998	\$9,246	\$9,493
	18.0%	\$9,008	\$9,287	\$9,566	\$9,844	\$10,123	\$10,402	\$10,680
	20.0%	\$10,009	\$10,319	\$10,629	\$10,938	\$11,248	\$11,557	\$11,867

Source: Company Data, Credit Suisse Estimates

AI also has potential to be a re-rating catalyst for TSMC after de-rating from 25-30x peak levels in 2020-21 to current 14x P/E and discount to SOX trading at 17x P/E. TSMC's ability to capture high share of the fastest growing driver in semiconductors (along with high share in ADAS as well) and ramp of overseas fabs to avoid lessen share loss risk from location could allow it to recapture some of its lost multiple. A shift in Intel strategy back toward outsourcing if the foundry approach with Meteor Lake/Arrow Lake is successful would also help that perception. Key risk to monitor though is Intel's ability to reassert its process position with 18A subject to being on time to high volume ramp by 2H25 to match up with TSMC's 2nm ramp.

**Figure 205: TSMC with AI could reverse some de-rating**

Source: TSMC

Scaling still providing some benefits albeit diminishing

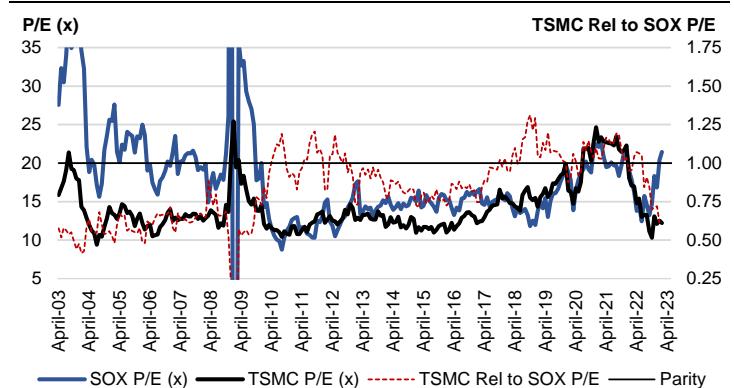
TSMC's technology leadership has come from a combination of process technology advances continuing on its roadmap through 2nm in 2025 and increasingly supplemented by advanced packaging to offset the slowing density gains enabled by Moore's Law. AMD's CEO Lisa Su

ChatGPT

**Figure 204: Faster penetration could drive NT\$2 EPS, NT\$30-35 on TSMC's share price**

	Server Units / 2024 Growth %							
	15.9	16.4	16.9	17.4	17.9	18.4	18.9	
	-3%	0%	3%	6%	9%	12%	15%	
AI Penetration	8.0%	\$1.68	\$1.73	\$1.78	\$1.83	\$1.88	\$1.93	\$1.99
	10.0%	\$2.09	\$2.16	\$2.22	\$2.29	\$2.35	\$2.42	\$2.48
	12.0%	\$2.51	\$2.59	\$2.67	\$2.75	\$2.82	\$2.90	\$2.98
	14.0%	\$2.93	\$3.02	\$3.11	\$3.20	\$3.30	\$3.39	\$3.48
	16.0%	\$3.35	\$3.46	\$3.56	\$3.66	\$3.77	\$3.87	\$3.97
	18.0%	\$3.77	\$3.89	\$4.00	\$4.12	\$4.24	\$4.35	\$4.47
	20.0%	\$4.19	\$4.32	\$4.45	\$4.58	\$4.71	\$4.84	\$4.97

Source: Company Data, Credit Suisse Estimates

**Figure 206: TSMC now trading at a 30% discount to the SOX**

Source: AMD

noted in its February 2023 ISSCC keynote this slowing improvement in energy per operation and density in the keynote remarks though also indicated advanced packaging and architecture innovations can help it continue or even improve on its innovation pace.

**Figure 207: TSMC maintains scaling to advanced geometries**



Source: TSMC

**Figure 208: Logic scaling slowing down**



Source: AMD

On a cost per transistor pace, benefits have also slowed down as new nodes have brought less density gains and higher cost to implement. We note with Apple's processor innovations, its cost improvement on a per transistor basis was 43% on 20nm, 27% on 16nm, 33% on 10nm, and 51% on 7nm but fell to 25% with the migration to 5nm.

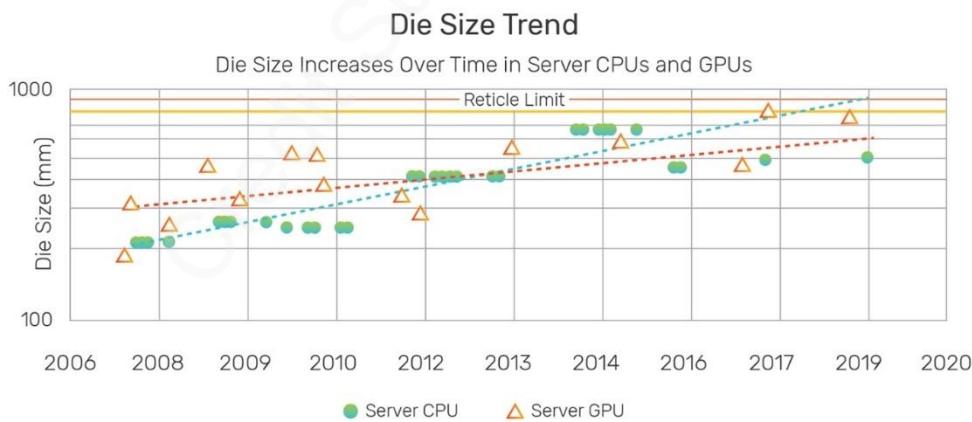
**Figure 209: Cost/transistor performance continuing but getting smaller**

iPhone Model	Launch Date	Node	Die size (mm-sq)	Transistors (billion)	Transistors (mn/mm-sq)	Wafer Price (US\$)	Price per mm-sq	Transistors /dollar	Cost improvement
A7	3Q13	28nm	102.0	1.0	9.8	\$5,000	\$0.07	138,588,235	At Samsung 43%
A8	3Q14	20nm	89.0	2.0	22.5	\$8,000	\$0.11	198,539,326	-1%
A9	3Q15	14/16nm	104.5	2.4	23.0	\$8,300	\$0.12	195,575,027	27%
A10	3Q16	16nm	125.0	3.3	26.4	\$7,500	\$0.11	248,793,600	33%
A11	3Q17	10nm	87.7	4.3	49.0	\$10,500	\$0.15	330,047,239	51%
A12	3Q18	7nm	83.3	6.9	82.8	\$11,750	\$0.17	498,267,733	9%
A13	3Q19	7nm	98.5	8.5	86.3	\$11,250	\$0.16	542,159,052	0%
A14	3Q20	5nm	88.0	11.8	134.1	\$14,000	\$0.20	676,967,532	25%
A15	3Q21	5nm	107.7	15.0	139.3	\$14,000	\$0.20	703,274,252	4%
A16	4Q21	4nm	114.9	16.0	139.3	\$14,000	\$0.20	703,054,547	0%

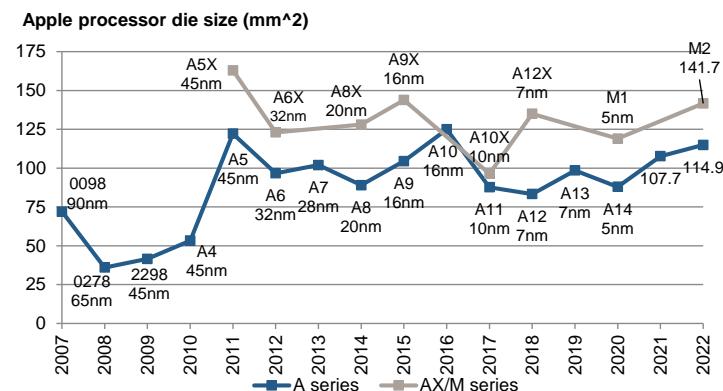
Source: Company Data, Credit Suisse Estimates

With lower density and cost gains, the drive to offer more performance and functionality at similar power has put processor die sizes on a gradual uptrend. The requirement for more compute in the iPad series and now Mac to run the GPU and AI instructions has led to 20-60% larger die sizes for the compute device, a measure that is also showing with AI through ever larger die sizes for the compute engines.

**Figure 210: Die sizes increasing over time in CPUs and GPUs**



Source: Cadence, Semiconductor Engineering

**Figure 211: Apple's die sizes getting larger as density gains fall**

Source: Company Data, Credit Suisse Estimates

While density gains are coming down, advanced nodes are necessary for fast AI training and inference. Energy efficient compute and the lower energy usage per transistor becomes an additional motivator to save money throughout the life of that chips use.

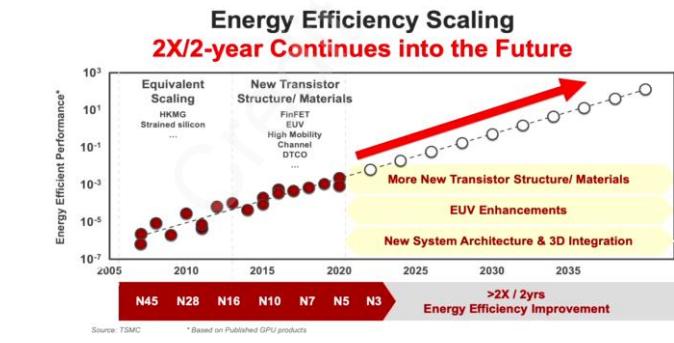
**Figure 213: Apple's die sizes getting larger as density gains fall**

Node (nm)	90	65	40	28	20	16/12	10	7	5
Year of mass production	2004	2006	2009	2011	2014	2015	2017	2018	2020
Foundry sale price to fabless firm per chip (i.e. costs + markup)	\$2,433	\$1,428	\$713	\$453	\$399	\$331	\$274	\$233	\$238
Fabless firm's design cost per chip given chip volume of 5 million <sup>19</sup>	\$630	\$392	\$200	\$135	\$119	\$136	\$121	\$110	\$108
Assembly, test, and packaging cost per chip	\$815	\$478	\$239	\$152	\$134	\$111	\$92	\$78	\$80
Total production cost per chip	\$3,877	\$2,298	\$1,152	\$740	\$652	\$577	\$487	\$421	\$426
Annual energy cost to operate chip	\$9,667	\$7,733	\$3,867	\$2,320	\$1,554	\$622	\$404	\$242	\$194

Source: CSET

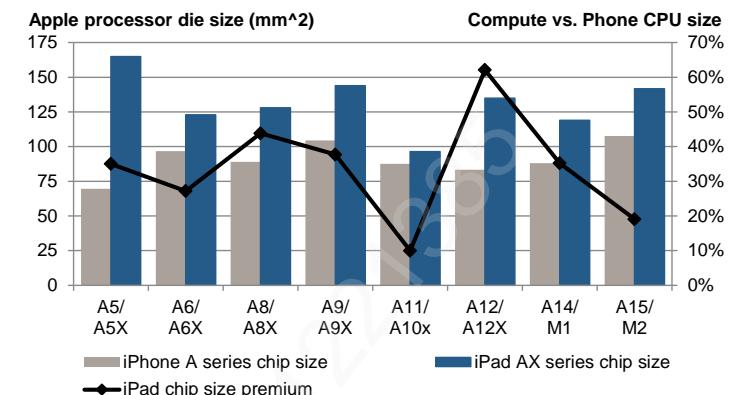
## Advanced packaging offsets slowing transistor scaling

Advanced packaging is a critical piece of integrating multiple tiles of GPU, CPU and AI accelerators with memory in an efficient subsystem as scaling everything on a large monolithic die becomes more difficult on advanced nodes. TSMC targets maintaining doubling energy efficient compute through a combination of continued technology shrinks enabled by EUV and new transistor structures but also system level integration.

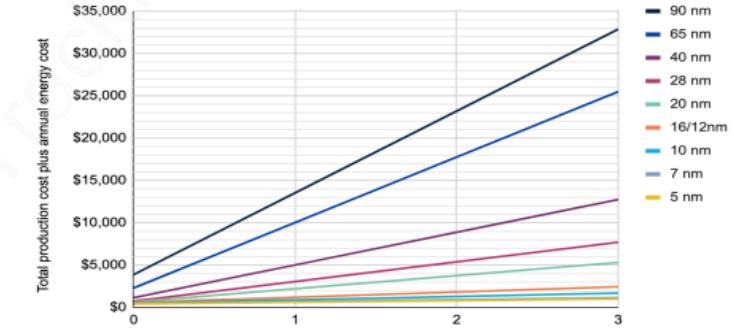
**Figure 215: TSMC's 3DFabric combines FE/BE integration**

Source: TSMC

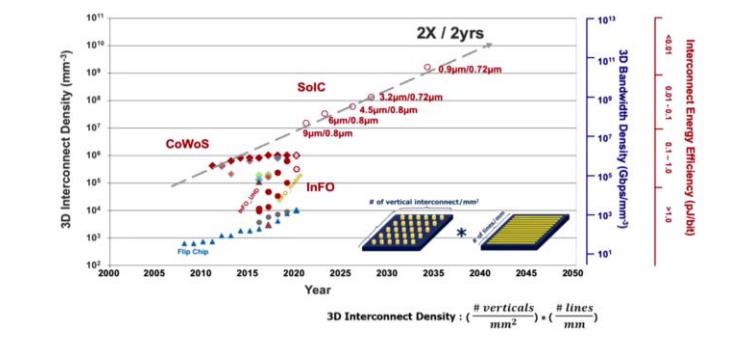
The companies 3DFabric toolkit includes InFO fan-out technology to join multiple chips in a compound material, CoWoS silicon interposer integration of multiple chips on a connector silicon

**Figure 212: Apple's compute die size remains larger vs. iPhone**

Source: Company Data, Credit Suisse Estimates

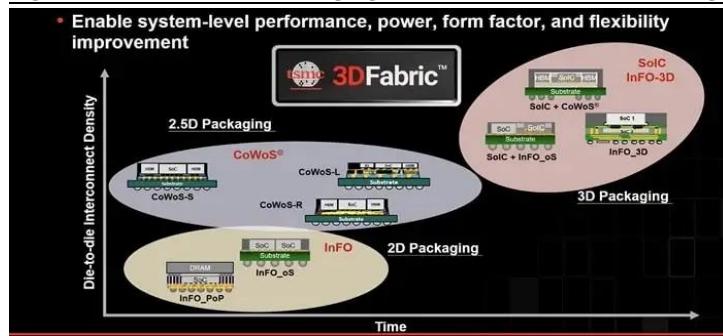
**Figure 214: Apple's compute die size remains larger vs. iPhone**

Source: CSET

**Figure 216: TSMC packaging helps maintain 2x scaling**

and SoIC which integrates chips in a 3D stack. AMD noted also at its recent ISSCC keynote a piece of achieving higher interconnect density for continued improvements in power/performance requires these gains in chiplet integration in its MI250/300 accelerators.

**Figure 217: Advanced packaging solutions supplement scaling**

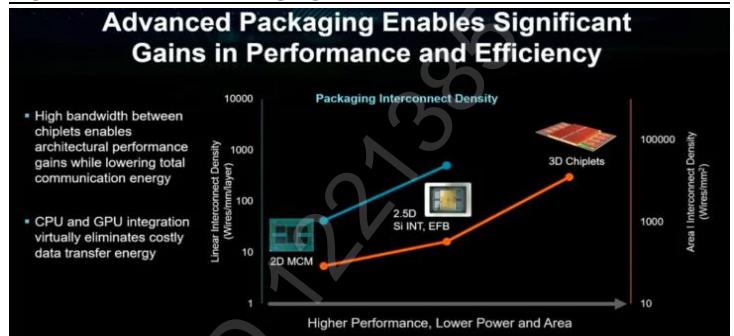


Source: TSMC

Advanced packaging notably can lower communication power between chips as it moves through 2.5G and eventually 3D stacking. AMD's goal for 2025 is for 30x improvement in energy efficiency for its accelerators leveraging scaling, packaging and architecture.

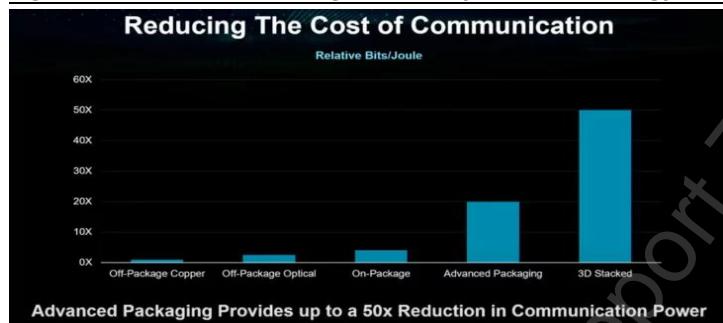
**Figure 218: AMD leveraging 2.5D/3D for data center solutions**

**Figure 218: AMD leveraging 2.5D/3D for data center solutions**



Source: AMD

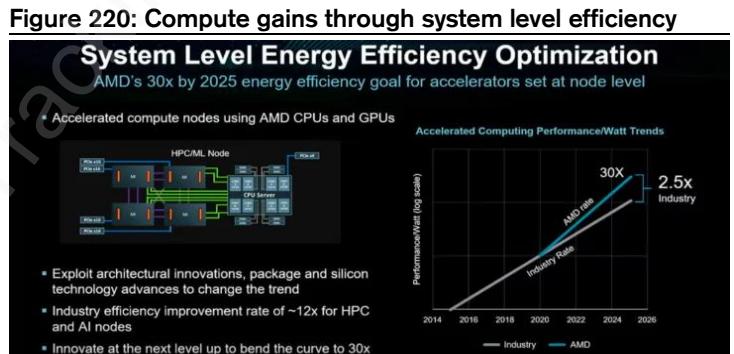
**Figure 219: 2.5D/3D stacking reduces system level energy use**



Source: AMD

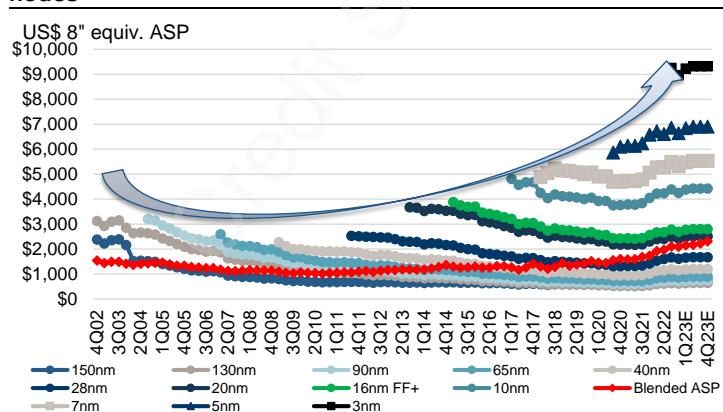
For TSMC, the benefits of advancing process has translated to new technology nodes at ever increasing price and now seeing narrower discounts and a growing advanced packaging stream as it captures system level integration projects mainly for HPC and high-end mobile. The company has ramped its advanced packaging to 10% industry share and 7% of company sales over the past decade and has also helped it with a full turnkey service in this growth area.

**Figure 220: Compute gains through system level efficiency**



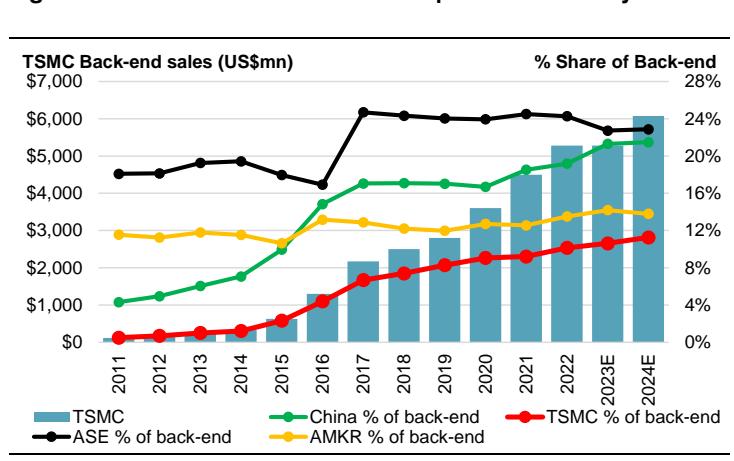
Source: AMD

**Figure 221: TSMC pricing leverage migrating to advanced nodes**



Source: Company Data, Credit Suisse Estimates

**Figure 222: TSMC back-end sales outpace the industry**



Source: Company Data, Credit Suisse Estimates

## Back-end service providers will also have a role to play

The back-end test and packaging suppliers in our coverage including ASE, Amkor and Powertech would also have some opportunities to address. The leading foundries and IDMs Intel, Samsung and TSMC all have vertical integration of wafer scale processes which should include most of the 2.5D silicon interposers and 3D stacking using TCB or hybrid bonding.

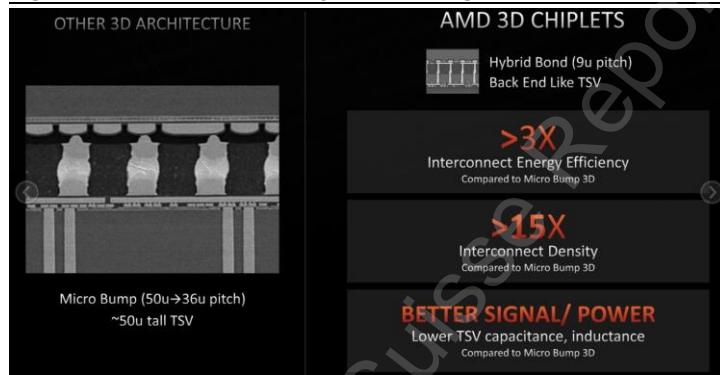
The OSATs though still have a role in final integration, with a number of 2.5D and 3D systems still relying on placement of the integrated system on a high-end ABF substrate for integration on a PCB as well as a final test service. The suppliers have also developed their own fan-out on substrate and fan-out process to handle some higher density dies and as the market moves into the mainstream would have OSAT play a role, though high-end AI products would likely start with the IDMs/foundries.

ASE and Amkor each have about 15% of sales contribution which can include package and test service of GPUs, ARM/AMD CPUs, and networking switch and some merchant ASICs. Powertech provides back-end service for server and graphics DRAM plus NAND SSDs (combining for ~20% of sales) along with flip chip and bumping for advanced logic though starting from mobile and mainstream networking.

## Back-end equipment suppliers to benefit upgrading advanced packaging

The demand for higher density and performance is also driving more advanced packaging tools including TCB and hybrid bonding. AMD's 3D V-Cache uses TSMC's SoIC with copper-to-copper bonding for its desktop, server CPUs and AI accelerator. The structure provides 3x interconnect energy efficiency and 15x higher interconnect density vs. micro bumps. Intel is also developing its hybrid bonding interconnect (HBI) process for its Foveros Direct technology.

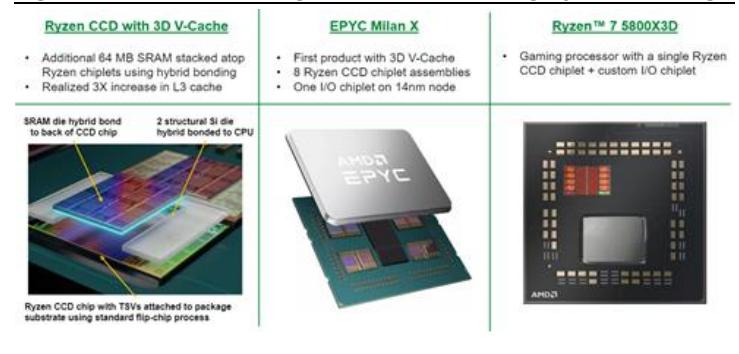
**Figure 223: Benefits from hybrid bonding for 3D chiplets**



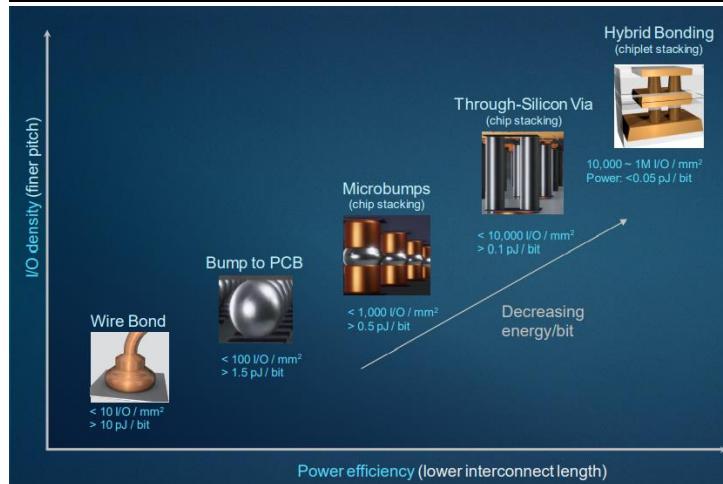
Source: AMD

Hybrid bonding has no solder interconnect to raise the density from 400 bumps/mm to 10,000 bumps/mm and reduces the bond-pad pitch from 36 micron to 9 micron but requires a clean environment, tools and very high precision. Advantages include no chip distance between chips, much greater density of the bond pad pitch, higher speed, bandwidth density and better power efficiency in energy per bit.

**Figure 224: AMD shipping the first CPUs using hybrid bonding**

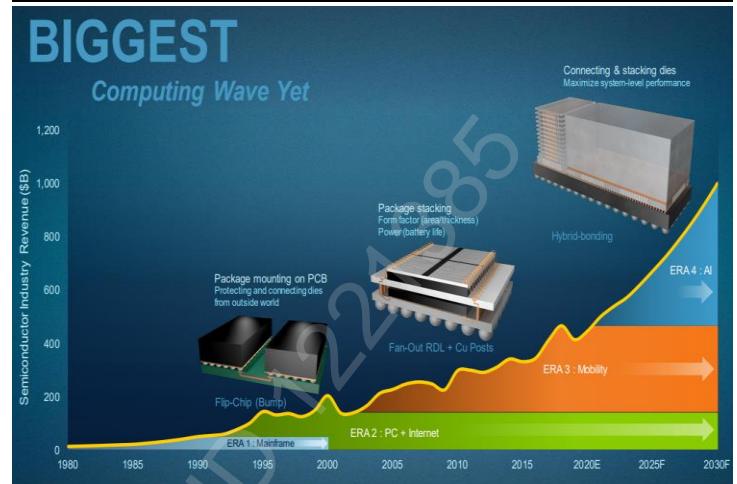


Source: BE Semi

**Figure 225: Packaging scaling roadmap toward hybrid bonding**

Source: Applied Materials

The integration has seen use in putting SRAM on separate chip connected to logic with low latency. For high performance die, advanced substrates and silicon interposers are still used to connect GPU and high bandwidth memory. Our feedback from Semicon Taiwan did note a more gradual ramp for hybrid bonding though with AMD leading the first wave but most suppliers in logic and also high bandwidth memory largely pushing out to the 2nm node. The TCB bonders continue to improve on the line spacing and have higher throughput and more mature yield so remain cost effective for most applications.

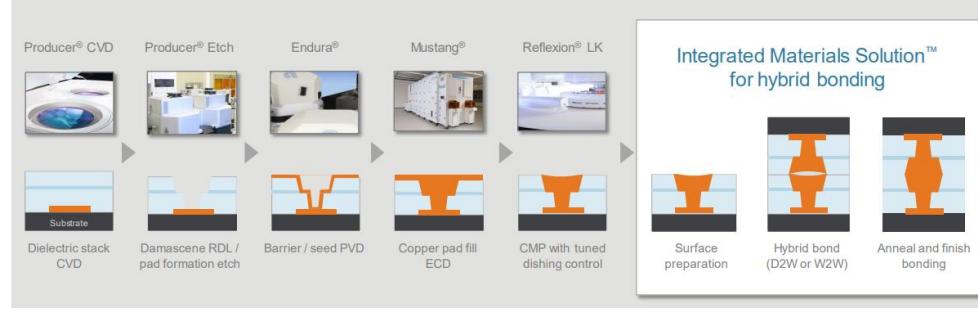
**Figure 226: AI era to be driven by the advanced packaging**

Source: Applied Materials

**Figure 227: Hybrid bonding for densely connected 3D stacks**

Source: Company Data

For the eventual ramp up of hybrid bonding, Applied Materials and BE Semiconductor are still leading the transition and in use for AMD's hybrid bonding with TSMC. The technology has started with high-end adoption with the recent high-end Ryzen desktop and server chip although could see another wave on future generations of high bandwidth memory (HBM4) and for additional AI chips as higher density of 2nm chips may require faster bandwidth connections between them.

**Figure 228: AMAT portfolio for hybrid bonding – partnering with Be Semi's bonder**

Source: Company Data

Rival equipment suppliers also target the AI opportunity. K&S has introduced its fluxless TCB bonder qualified by a leading CPU company and putting most of its attention there, promoting its higher yields and throughput over hybrid bonding as well at the Semicon shows. K&S is also planning introduction of its next generation hybrid bonder to be adopted in coming years.

**Figure 229: K&S tool portfolio for advanced packaging**

Source: Company data

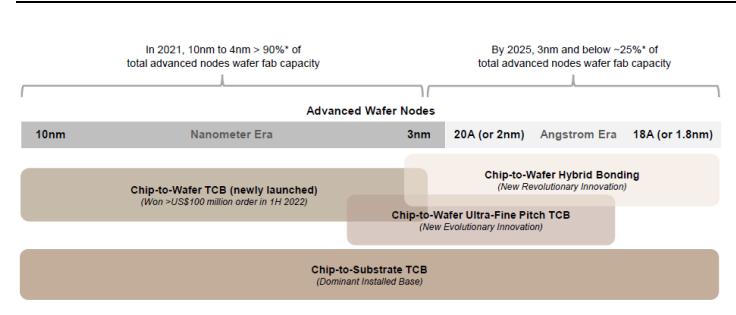
ASM Pacific also presented its solution roadmap for hybrid bonding and ultra fine-pitch TCB bonders for the packaging integration, though noted on its recent earnings call a view the inflection to high volume will take more time to reach the maturity of TCB where it has been shipping for the past 7 years and recently received high volume logic orders and first penetration into HBM memory. The company traditionally was tied to mainstream packaging and SMT although has targeted higher growth both from advanced packaging and automotive electronics to drive its growth in the next industry cycle.

**Figure 230: ASMPM ramping TCB, hybrid bonding to take time**

	Chip-to-Substrate TCB	Chip-to-Wafer TCB	Chip-to-Wafer Hybrid Bonding
Readiness of Wafer Fab Ecosystem	Mature	High Growth	Early Adoption
Relative Cost to Flip Chip Mass Reflow	Medium	Medium	High
Customer CAPEX Considerations	Back-End Assembly Facilities	Mid-End Assembly Facilities	Front End Cleanroom Facilities Other Front-End Tools Required (e.g. CMP, Clean, Plasma, Metrology, Annealing)
Manufacturing Readiness	HVM	Transitioning to HVM	Qualification to LVM
Customer Buying Patterns	Capacity Buy for HVM	Capacity and Technology Buy for HVM (Chip-to-Wafer TCB) New Gen Technology Buy (Chip-to-Wafer Ultra Fine Pitch TCB)	Frontier Technology Buy (AI Learning Curve Phase)

Source: Company data

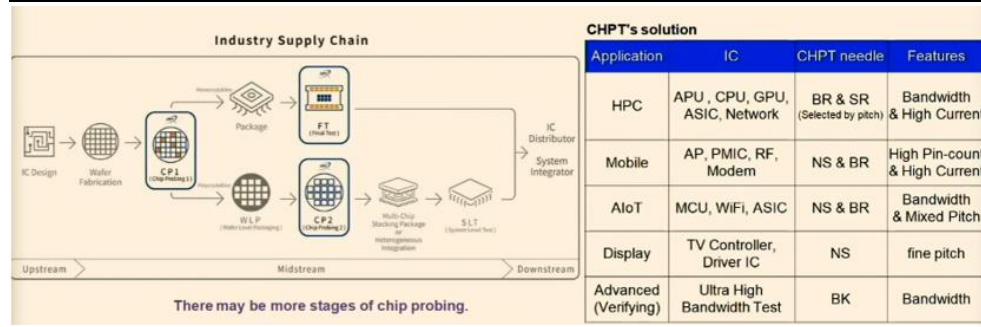
Probe test supplier CHPT targeting AI/HPC projects

**Figure 231: ASMP targets chip-to-wafer TCB & Hybrid bonding**

Source: Company data

Chunghwa Precision, a Taiwan based advanced probe card supplier is also diversifying from its traditional strength in mobile processors to target high performance applications. We believe the company has about 10% of sales from AI, HPC and GPU with multiple projects in qualification.

**Figure 232: CHPT targeting more HPC projects with its high current probe cards**



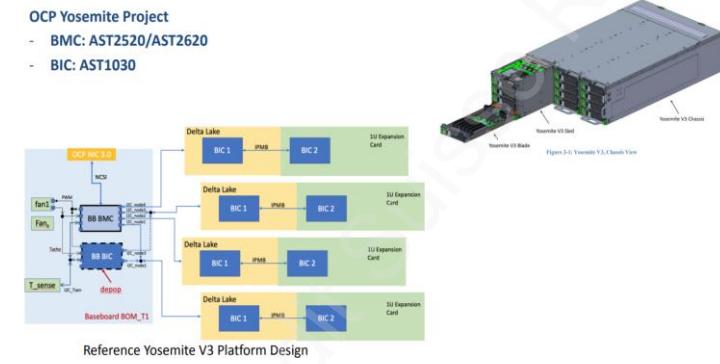
Source: Company data

## Asian IC Design: ASpeed seeing growth from its higher content with its controller and peripheral chips

Taiwan IC designer ASpeed has grown to have 70% market share with its baseboard management controller now supplying most of the leading US and China hyperscalers and OEMs except for Dell, HP and one of the US cloud suppliers Intel based workloads. The company has added to its sales with new BMC migrating to more advanced node with current generation AST 2600 upgrading from 40nm to 28nm and next generation AST2700 moving to 12nm. The main BMC would have an additional attach powering an NVIDIA AI server but not that high a leverage, with 1 US\$12 chip managing a rack of 1-16 US\$10-20k NVIDIA GPUs.

ASpeed does have additional chipset drivers adjacent to the main BMC. It is picking up content with Meta's Yosemite architecture supplying a mini BMC for each additional CPU line card. It also is adding a hardware root of trust to detect and protect the server board from hacks.

**Figure 233: ASpeed picking up content on CPU add-on cards**

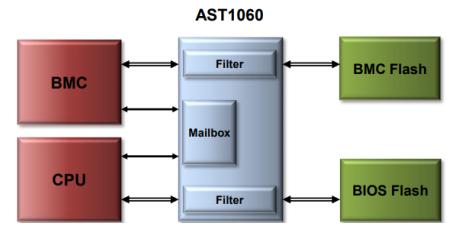


Source: Company data

ASpeed has been keeping pace with the ODMs gaining market share with the rise of the ODM direct server growing at 14% CAGR from 2019-24 vs. +13% CAGR for the ODMs.

**Figure 234: ASpeed root of trust adds to server security**

- **Hardware Root of Trust (HRoT)**
  - **Protection** - Real time firmware protection using the hardware SPI filtering
  - **Detection** - Includes HW Crypto for firmware measurement and authentication
  - **Recovery** - Supports firmware roll back feature upon detection of firmware corruption or failed



Source: Company data

**Figure 235: ASpeed has grown in-line with the ODM direct suppliers gaining share with hyperscalers**

NT\$bn/BMC (mn)	1Q21	2Q21	3Q21	4Q21	1Q22	2Q22	3Q22	4Q22	2019	2020	2021	2022F	2023F	2024F	19-24
Inventec	12.1	16.6	21.7	22.0	19.8	20.2	28.2	30.0	67.1	83.1	72.4	98.1	108.0	114.0	11%
Quanta	45.5	48.1	49.8	42.4	50.0	54.4	62.2	61.5	144.5	167.4	185.8	228.1	248.9	269.0	13%
Wiwynn	39.3	51.2	45.3	56.9	50.7	75.1	79.7	87.4	163.6	186.9	192.6	292.9	300.2	346.5	16%
MITAC	9.5	9.9	7.2	6.8	8.9	9.1	9.9	10.6	22.0	26.6	32.8	38.4	43.2	48.4	17%
Accton	6.8	6.8	7.6	9.4	8.4	10.3	11.9	12.8	25.9	27.3	30.7	43.4	48.9	54.0	16%
<b>Total</b>	<b>113.1</b>	<b>132.7</b>	<b>131.5</b>	<b>137.5</b>	<b>137.8</b>	<b>169.0</b>	<b>191.9</b>	<b>202.3</b>	<b>423.1</b>	<b>491.3</b>	<b>514.3</b>	<b>701.0</b>	<b>749.2</b>	<b>831.8</b>	<b>14%</b>
<b>QoQ / YoY</b>	<b>-18%</b>	<b>17%</b>	<b>-1%</b>	<b>5%</b>	<b>0%</b>	<b>23%</b>	<b>14%</b>	<b>5%</b>	<b>-1%</b>	<b>16%</b>	<b>5%</b>	<b>36%</b>	<b>7%</b>	<b>11%</b>	
<b>ASpeed BMC units</b>	<b>2.8</b>	<b>3.4</b>	<b>3.4</b>	<b>3.4</b>	<b>3.6</b>	<b>4.1</b>	<b>3.6</b>	<b>3.2</b>	<b>8.6</b>	<b>11.2</b>	<b>12.9</b>	<b>14.4</b>	<b>13.8</b>	<b>15.6</b>	<b>13%</b>
<b>QoQ / YoY</b>	<b>13%</b>	<b>20%</b>	<b>0%</b>	<b>0%</b>	<b>6%</b>	<b>14%</b>	<b>-11%</b>	<b>-12%</b>	<b>9%</b>	<b>30%</b>	<b>15%</b>	<b>12%</b>	<b>-4%</b>	<b>13%</b>	

Source: Company data, Credit Suisse estimates

We project the combination of higher units from servers, additional attach on network switch and storage and additional peripheral content can drive double digit growth for ASpeed continuing through the next few years and project growth reaccelerating to mid-20% growth in 2024 from reacceleration of server units following this year's inventory correction alongside full year contribution from the higher content on the new Intel Sapphire Rapids and AMD platform.

**Figure 236: ASpeed gaining from unit growth and content gains**

ASpeed shipments (k)	2018	2019	2020	2021	2022F	2023F	2024F
Baseboard Management Controller	7,902	8,619	11,222	12,894	14,447	13,812	15,555
Mini BMC					600	2,550	3,300
Root of Trust						150	1,000
A/V Extension	176	169	158	269	342	411	473
Cupola360		6	70	127	195	196	246
<b>Total</b>	<b>8,107</b>	<b>8,876</b>	<b>11,496</b>	<b>13,290</b>	<b>15,584</b>	<b>17,120</b>	<b>20,573</b>
<b>YoY Aspeed Growth</b>	<b>15.3%</b>	<b>9.5%</b>	<b>29.5%</b>	<b>15.6%</b>	<b>17.3%</b>	<b>9.9%</b>	<b>20.2%</b>
<b>YoY BMC Unit Growth</b>	<b>15.9%</b>	<b>9.1%</b>	<b>30.2%</b>	<b>14.9%</b>	<b>12.0%</b>	<b>-4.4%</b>	<b>12.6%</b>
ASpeed ASPs (US\$)	2018	2019	2020	2021	2022F	2023F	2024F
Baseboard Management Controller	\$8.5	\$8.7	\$8.8	\$9.4	\$11.0	\$11.8	\$12.7
Mini BMC					\$6.5	\$6.5	\$6.5
Root of Trust						\$9.0	\$9.0
A/V Extension	\$24.8	\$25.1	\$24.9	\$25.3	\$26.2	\$28.1	\$30.5
Cupola360		\$17.3	\$16.5	\$15.6	\$14.7	\$13.8	\$13.0
<b>Total</b>	<b>\$8.8</b>	<b>\$9.1</b>	<b>\$9.0</b>	<b>\$9.8</b>	<b>\$11.2</b>	<b>\$11.4</b>	<b>\$11.9</b>
<b>YoY Growth</b>	<b>-0.3%</b>	<b>2.6%</b>	<b>-0.2%</b>	<b>8.7%</b>	<b>14.2%</b>	<b>1.6%</b>	<b>4.6%</b>

ASpeed Revenue (US\$k)	2018	2019	2020	2021	2022F	2023F	2024F
Baseboard Management Controller	\$66,945	\$75,315	\$98,385	\$121,726	\$159,105	\$162,875	\$197,114
Mini BMC					\$3,900	\$16,575	\$21,450
Root of Trust						\$1,350	\$9,000
A/V Extension	\$4,380	\$4,232	\$3,935	\$6,802	\$8,957	\$11,575	\$14,409
Cupola360	\$0	\$104	\$1,158	\$1,975	\$2,875	\$2,705	\$3,199
<b>Modeled/Actual Sales (US\$k)</b>	<b>\$71,581</b>	<b>\$80,385</b>	<b>\$103,901</b>	<b>\$130,503</b>	<b>\$174,836</b>	<b>\$195,080</b>	<b>\$245,172</b>
<b>Total (US\$k)</b>	<b>\$71,581</b>	<b>\$80,385</b>	<b>\$103,893</b>	<b>\$130,539</b>	<b>\$174,836</b>	<b>\$195,080</b>	<b>\$245,172</b>
<b>YoY Growth</b>	<b>15.1%</b>	<b>12.3%</b>	<b>29.2%</b>	<b>25.6%</b>	<b>33.9%</b>	<b>11.6%</b>	<b>25.7%</b>
TWD/US\$	30.1	30.9	29.5	27.9	29.8	30.7	30.7
<b>Modeled/Actual Sales (NT\$k)</b>	<b>\$2,153,519</b>	<b>\$2,484,295</b>	<b>\$3,063,552</b>	<b>\$3,637,778</b>	<b>\$5,210,122</b>	<b>\$5,988,961</b>	<b>\$7,526,774</b>
<b>Total (NT\$k)</b>	<b>\$2,153,519</b>	<b>\$2,484,295</b>	<b>\$3,063,552</b>	<b>\$3,637,632</b>	<b>\$5,210,096</b>	<b>\$5,988,961</b>	<b>\$7,526,774</b>
<b>YoY Growth</b>	<b>13.7%</b>	<b>15.4%</b>	<b>23.3%</b>	<b>18.7%</b>	<b>43.2%</b>	<b>14.9%</b>	<b>25.7%</b>
<b>QoQ Growth</b>							

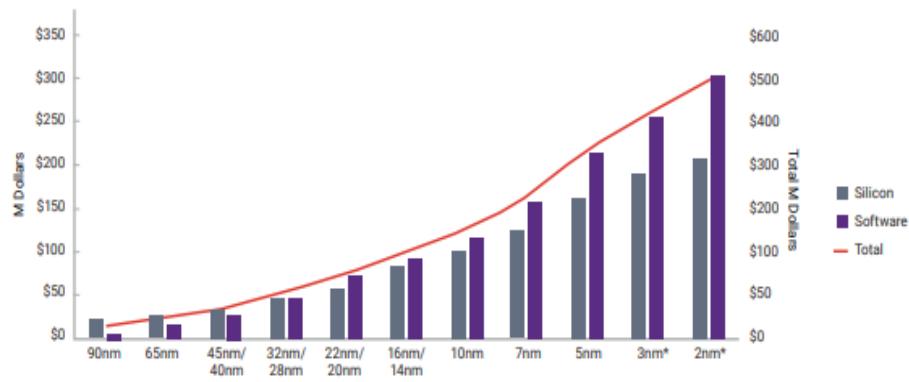
Source: Company data

## IC Design Service: Alchip, GUC and Socionext set to benefit from growing chipset customization

We believe the IC design service has been one of the key beneficiaries from the surge in investment by the start-ups and system companies first from blockchain/crypto-currency in 2017-18 but more recently from growth in cloud computing AI and supercomputing applications that has helped form a new wave of start-up, system companies and also China domestic chip companies.

Haas Liu

**Figure 237: Chipset design cost surges from US\$30-40mn on 28nm to US\$500mn+ on 2nm**



Source: Synopsys

With the industry getting more consolidated to those that have the access to advanced manufacturing process, IP and EDA tools, we expect the opportunity for Taiwan IC design service companies will be in ASIC design for the applications across communication, consumer, HPC and automotive industry including mature data center, IoT, drone, robotics, artificial intelligence (AI), machine learning, 5G networking and ADAS.

The higher R&D requirement in the advanced nodes is also giving the IC design service companies with more opportunity as their customers are more cautious on the investment in the leading edge technology. We estimate global fabless' R&D expense/sales has been growing from 13-18% in 2000-05 to low 20% levels in the past decade and is approaching 25% as major fabless migrate to 7nm and below nodes. According to Synopsys, the chip design cost has also grown meaningfully to US\$500mn+ on 5nm (vs. US\$150mn on 10nm and US\$30-40mn on 28nm). We believe the high risk for the start-ups and system companies who lack of the experience in semiconductor design and manufacturing should expand the addressable market for IC design service companies in the advanced nodes.

## Competitive landscape for Design Services

The design service capability and capacity, IP portfolio, success rate, supply chain relationship, target applications and technology/IP support are the important factors when customers choose the service provider. We compare the competitiveness for the major companies including Global Unichip, Alchip and Faraday in Taiwan, Socionext in Japan and VeriSilicon in China as below.

**Figure 238: Semiconductor IC Design Company Landscape**

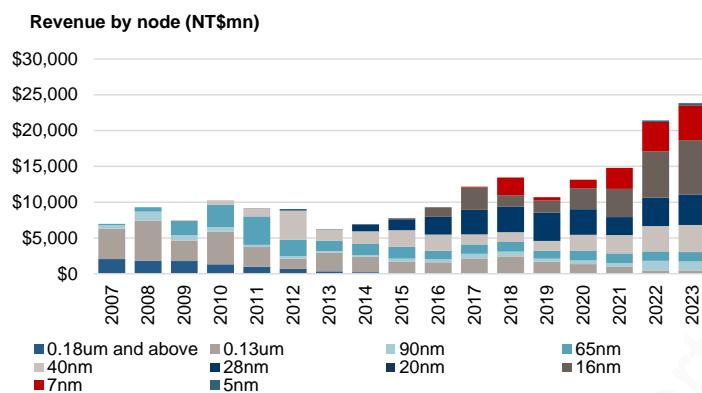
IC design service company	Global Unichip	Alchip	Faraday	Socionext	VeriSilicon
<b>Ticker</b>	3443.TW	3661.TW	3035.TW	6526.T	688521.SH
<b>Revenue (US\$mn)</b>	\$800	\$470	\$435	\$1,740	\$375
<b>GMs</b>	34.7%	33.1%	48.8%	46.8%	41.6%
<b>OpMs</b>	17.0%	17.8%	22.4%	12.0%	3.4%
<b>Headcounts</b>	795	462	583	2,546	1,280
<b>Service</b>	Front/back-end design, IP	Back-end design	IP, back-end design	Front/back-end design, IP	IP, back-end design
<b>Technology support</b>	40nm and above: 27% 28nm: 20% 16nm: 33% 7nm: 17% 5nm: 3%	40nm and above: 4% 28nm: 6% 16/12nm: 20% 7nm and below: 70%	0.11um and above: 17% 90-55nm: 35% 40nm: 23% 28nm and below: 15%	40nm and above: 31% 28/20nm: 22% 16/10nm: 18% 7/5nm: 29%	28nm and below: 85% 28nm and above: 14%
<b>Targeted applications</b>	AI/Machine learning: 12% Networking: 22% Digital consumer: 41% Industrial: 17% Others: 8%	HPC: 80% Network: 6% Niche: 6% Consumer: 8%	Industrial: 45% AloT: 35% Communication: 11% Multimedia/consumer/PC: 9%	Automotive: 14% Data center & networking: 28% Smart devices: 24% Others: 34% Others: 0-5%	Consumer: 31% IoT: 25% Data processing: 14% Industrial: 8% PC: 17% Automotive: 5%
<b>Foundry partners</b>	TSMC (100%)	TSMC (~100%), Samsung, SMIC, GlobalFoundries	UMC (85-90%), Samsung (10-15%)	TSMC for advanced ASIC design	SMIC, Samsung, GlobalFoundries, Hua Hong, TSMC

Source: Company data, Credit Suisse estimates

**Global Unichip:** The company is the largest IC Design service company in Taiwan providing both front-end and back-end IC design service with its customers, with revenue reached US\$800mn 2022 at 35% GMs. The company generates 73% revenue from 28-5nm with a target on AI and machine learning (12% of its sales) and networking applications (22% of its sales) while keeping its business diversified in niche consumer applications on the mature nodes to support its customers from different industry verticals.

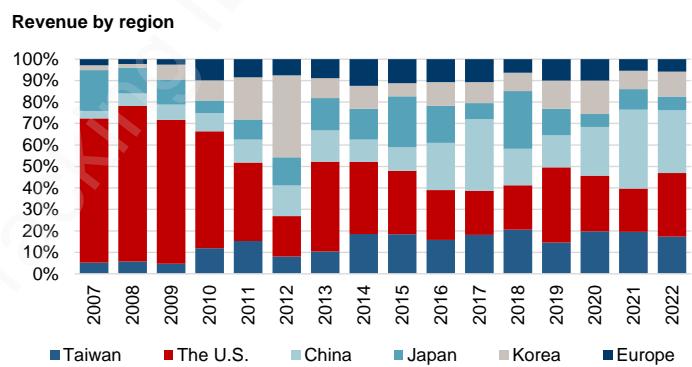
Compared with high exposure in China for Faraday, Global Unichip's is more diversified, with the U.S. and Taiwan contributing 50% of its sales. As one of TSMC's subsidiary, the company gets the support from TSMC on the IP library and manufacturing even during the supply tightness in the advanced nodes. Global Unichip is also leading on the technology development among its Taiwan peers with 3nm testing projects under NRE for tape out later 2023 while the company also has solid IP portfolio for high speed interface (Serdes) and packaging (CoWoS) suitable for HPC applications.

**Figure 239: Growth more diversified across 7nm and 16nm/12nm**



Source: Company data, Credit Suisse estimates

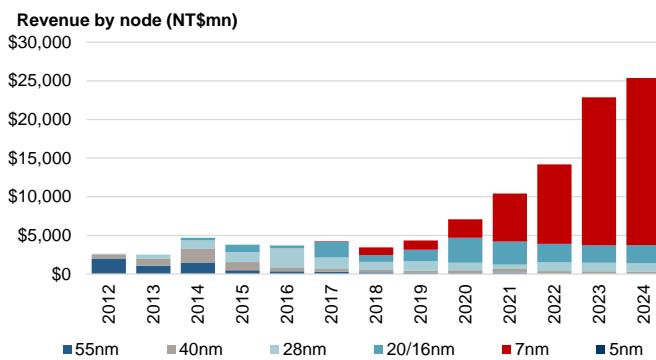
**Figure 240: Geographic-mix remains diversified**



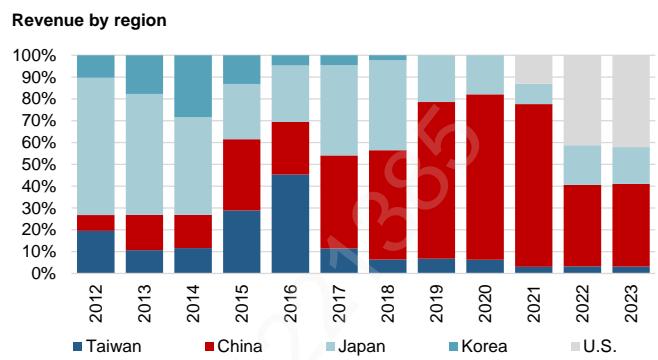
Source: Company data, Credit Suisse estimates

The company guided sales to grow 20-30% YoY in 2023 on the continued production ramp for its customers' high-end SSD and networking chipsets along with NRE business resilience from growing project dollar content from customers' migration to the advanced nodes. However, due to margins normalization from easing foundry supply and mix shift partially offsetting its larger business's scale, earnings is guided to only grow single digits YoY. The moderating earnings growth and stretched valuation from recent AI sentiment hype make the share risk/reward unattractive at this level and we stay NEUTRAL on the company.

**Alchip:** Alchip was founded in 2003 by a group of semiconductor veteran to focus on the leading edge ASIC and SoC design, with revenue reached US\$470mn in 2022 at 33% GMs through 3Q22. With its strategy to put resource on HPC and AI projects, the company generates 70% revenue from 7-5nm, 20% from 16nm and 10% from 28nm and above. With growing geopolitical tension and a few of its China customers restricted by the U.S. government due to their business in high-end CPU and supercomputing in the past few years, the company has lowered its China exposure from 70-80% in the peak to 25% in 1Q-3Q22 and most of the business there is more from the diversified NRE projects for mainstream AI, CPU and GPU meeting the US BIS criteria rather than single high volume turnkey customer with risky political profile.

**Figure 241: Growth should be led by 7nm/5nm in 2022-24**

Source: Company data, Credit Suisse estimates

**Figure 242: Geographic-mix now more balanced**

Source: Company data, Credit Suisse estimates

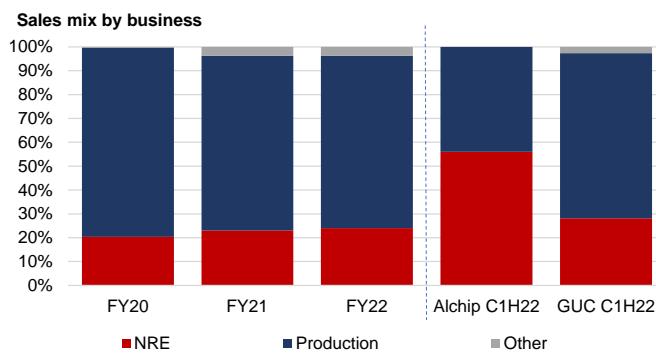
In addition to China, the company is aggressively working with the U.S. internet and start up companies for ASIC design opportunities across cloud AI training and inference chipsets and consumer applications on the advanced nodes. Specifically, Alchip has been working Amazon closely on its AI chipsets with more volume ramp through 2023 for the AI inference project followed by the voice assistant and e-reader chipset tape out in 2H23 and volume ramp in 2024.

We model the company's sales will grow 65% YoY to US\$750mn in 2023, with the main growth drivers from full-year production ramp-up for its US customers' AI chipsets on 7nm to provide more comprehensive cloud solutions along with the chipset for the hyperscalers' consumer projects. Although the phase-out of current AI production in 2H24 and competition on the next generation AI chipset for its US hyperscaler could remain an overhang until 2Q23, our base case for +10% YoY growth to US\$815mn sales for 2024 already factors in 30% run rate drop for the US AI project (US\$130mn loss) offset by 3-4 new projects ramp for its mainstream AI, CPU/GPU and automotive customers with meaningful contribution, keeping the drivers transition smooth in the next few years even with the loss of the next gen AI chipset for its US hyperscaler customer. Upside would be the next generation AI chipset only migrates from 7nm to 5nm and stick with Alchip for ramp in late 2025.

We maintain Outperform with a TP of NT\$1,400 based on 32x 2023E P/E, the mid-cycle of its 15x-45x range. The stock is at 21x 2023 earnings on the investors' concern on the potential loss of the next gen AI chipset for its US hyperscaler customer and discount for the geopolitical risks. We set a low expectation for the contribution from China core ICs and wind down of the US AI project in 2024 but valuation remains compelling even based on our street low earnings.

**Socionext:** Socionext is an IC design company spun off from Fujitsu and Panasonic's semiconductor division in 2015 and backed by the Development Bank of Japan. The company's initial business was selling the standard chipsets but shifted its focus to customized chipset design business from 2018. Socionext migrates its capability with TSMC's technology roadmap and could support its customers through 5nm for the next generation ASIC design for its automotive projects since 1Q21. The company also continues to invest in design processes on 3nm and finer process nodes, advanced packaging technology (e.g. 2.5D and 3D packaging), chiplet, AI for design and silicon IP to enable high-density ASICs.

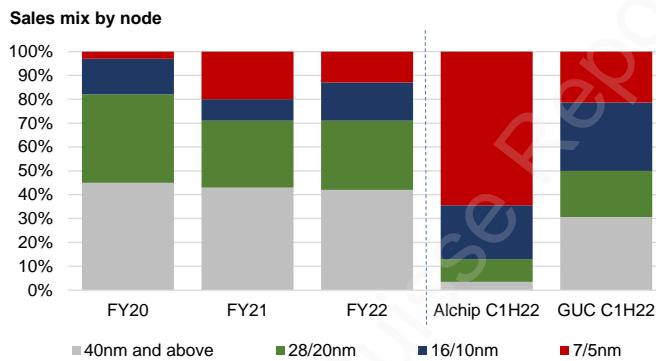
Socionext provides chipset design and capacity booking for mass production for its customers, with 20-30% of its sales generated from NRE design and 70-80% from turnkey business in FY20-22. The full design service business model and Japan business exposure allows the company to work with the automotive supply chain earlier than its peers, with NRE contribution from automotive applications growing from 5% in FY20 to 26% in FY22. The application base is also diversified across 5G network and data center (30% of its NRE sales), smart devices (30% of its NRE) and other legacy/specialty projects (15% of its NRE). By node, the company is also seeing growing NRE contribution on 16/10nm and 7/5nm at 30% and 43% in FY22 (vs. 34% and 15% in FY20) as new applications in general require strong computing capability and enhanced power efficiency.

**Figure 243: Socionext's NRE and turnkey mix comparison**

Source: Company data, Credit Suisse estimates

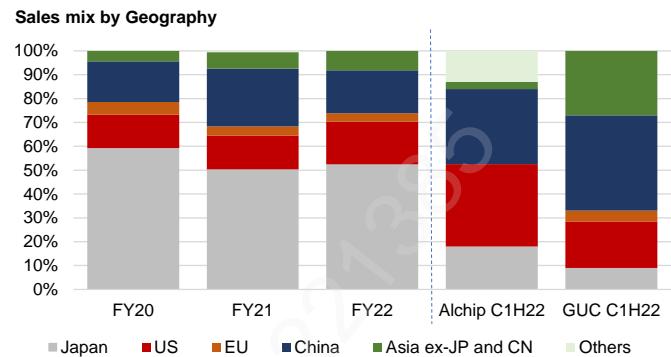
Socionext's sales was down by 4% CAGR from FY18-21 before recovered by 14% YoY in FY22 which could be due to the company's gradual wind down of its traditional off-the-shelf business with its contribution from other and smart device applications down by 14% and 1% CAGR from FY18-21 more than offsetting 14% and 8% CAGR from 5G network/data center and automotive during the same period.

Following 1% CAGR from FY18-22, Socionext sets its mid-term financial target for 17-19% sales CAGR, in line with our expectation for the IC design service sector supported by both NRE (dollar content growth along with technology migration and more NRE project demand) and turnkey (higher foundry wafer pricing on the advanced nodes). However, its mid-term OpMs target is at 11-15%, still below our expectation Alchip at 19% and GUC's 15% in 2022-24E.

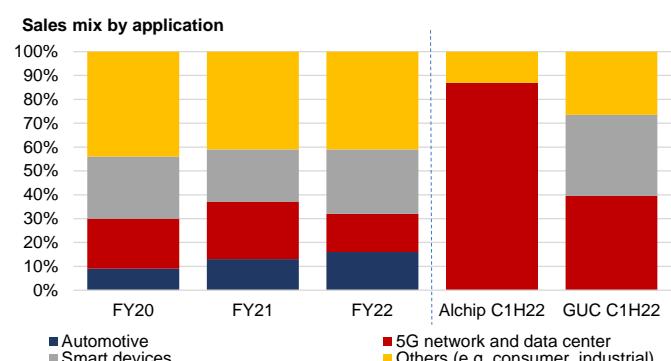
**Figure 245: Taiwan peers have higher sales from the advanced nodes**

Source: Company data, Credit Suisse estimates

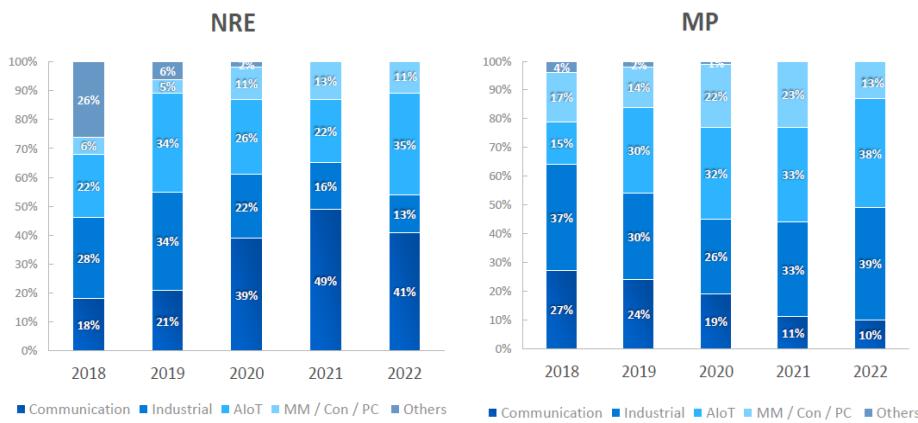
**Faraday:** The company is a spin-off from UMC, with US\$435mn revenue and 49% GMs in 2022. Although the company has started the partnership with Samsung foundry for advanced nodes, with its main partnership with UMC, the company is lagging on the technology development. However, though the pandemic, the company saw its business acceleration from chipset demand and pricing surge on the mature 12" nodes, lifting sales growing at 54% CAGR from 2020-22 following the stagnation from 2017-19. Faraday's revenue from 28nm and below only contributes 16% of its sales and 35% from 40-90nm and 46% on 8". On the product mix, on limited access to the advanced nodes for HPC and mobile related applications, the company has higher exposure in IoT (35% of its sales) and industrial (45% of sales) applications, with key revenue drivers in the past few years from MCU in China though saw demand correction from 2H22 along with broader industry.

**Figure 244: Socionext business is mostly in Japan**

Source: Company data, Credit Suisse estimates

**Figure 246: Taiwan peers have higher HPC exposure than Socionext**

Source: Company data, Credit Suisse estimates

**Figure 247: Faraday's revenue by application**

Source: Faraday

For 2023, the company expects sales to see mild growth QoQ from NRE milestone recognition and IP licensing partially offset by continued decline in turnkey business as its customers slow down the procurement in the slow end market demand environment and loosening foundry supply. For full year outlook, the company is confident to see double digits YoY growth for its NRE and IP business on growing design pipeline for smart meter and green energy though stays cautious on its turnkey business. With inventory write-off, Faraday expects GMs to dip by a few % though should gradually recover as the headwind fades.

**Verisilicon:** Verisilicon was founded in 2001 by Dr. Wei-Ming Dai (Brother of Marvell co-founder Weili Dai), the ex-professor at UCSC and Mr. Wei-Jin Dai, ex-VP of Cadence. The company's business includes the semiconductor IP licensing (GPU, NPU, VPU, DSP, ISP, analog and mixed IP) and IC design service.

The company has been targeting key projects including IP development for wearables, smart automotive and smart home, building up ASIC development platform for cloud computing and R&D center upgrade. On the supply chain management, in contrast to most of Taiwan IC design service companies who in general have a primary foundry partner, Versilicon has been maintaining a multiple sourcing strategy to meet its customers' demand, with major foundry partners including SMIC, Samsung, Global Foundries, Hua Hong and TSMC and back-end partners including ASE, JCET, King Yuan, Tianshui Huatian and Powertech. To ensure the chipset design is optimized for customers' requirement, the company also provides 3rd party IP in addition to its solutions.

As the largest IP and IC design service provider, the company should benefit from the fast growth in China semiconductor ecosystem. The amount of China fabless has grown from 500 in 2010 to current ~2,000 with most of them established since 2016 targeting a diversified applications across low-end consumer and communication peripheral chipsets to the core ASICs and CPU for cloud and edge computing.

We would note for the company's IC design service business, the revenue from its China customers has been growing the importance from 25-35% in 2017-18 to ~50%, suggesting an acceleration of IC design activity in China and the need to have support from IC design service companies due to their lack of scale and experience.

The company's NRE projects in the advanced nodes have also been growing in the past few years, with 28nm and below growing as % of total projects. We believe it is mainly driven by its technology development, with support through 5nm now and 5nm platform under development. The company is also partnering with its foundry partners in the advanced nodes. By application, the company's revenue contribution in its IC design service is mostly from consumer, IoT on 0.18um, 28nm, 22nm and 14nm and data processing on 14nm and 10nm.

For the company's business, the company's sales grew to US\$375bn in 2022. With growing business scale for its IP and ASIC design business, the company was able to improve its GMs to 42% in 2022 while turned operating profitable at 3% OpMs.

With the company's focus in China domestic market and target to build cloud computing platform for its local customers for the chipset localization in the next few years, we believe the company would be a modest threat for the Taiwan IC design service companies with higher China exposure (e.g. Faraday). We believe the competition could be more fierce for Faraday due to its high overlap with Verisilicon in the IoT and consumer market while the foundry support from UMC and Samsung would not be an advantage for Faraday.

## Memory

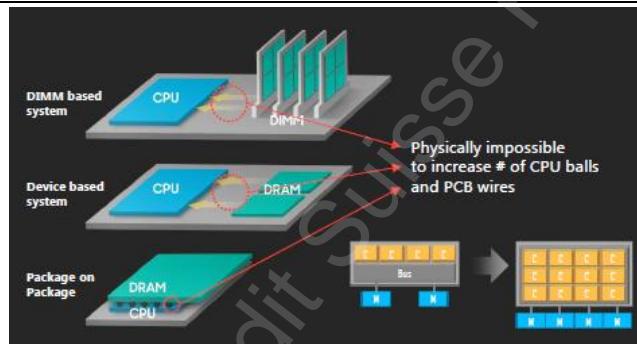
### Overcoming the memory bottleneck for efficient machine learning (ML) (Keon Han)

Keon Han  
Sanguk Kim  
DJ Kim

In traditional computer architecture (Von Neumann model) the data processing for the system and the storage were managed in separate modules. However, historically, the performance gains for the compute performance have exponentially outpaced the improvement in memory speed creating a gap that continues to grow today. The advent of AI shifts compute from serial processing to parallel processing, driving a massive increase in the size of neural networks. This requires processors to access more data from memory faster, an obstacle since the performance gap between processor performance and memory bandwidth to feed the processor have been getting wider. The current Von Neumann architecture and various memory technologies offered will be a bottleneck for AI at the inter-chip and communication level to AI accelerators depending on the size of the program and amount of data required to process. The computing power itself is sufficient today but the memory bottleneck problem has become worse for AI performance, requiring faster memory advancement.

Related to memory, several areas are driving continual improvement including bandwidth, latency, power consumption, capacity and cost – with the ultimate goal being improving latency while cutting energy required per bit transferred.

**Figure 248: Memory bottleneck feeding the processor data**

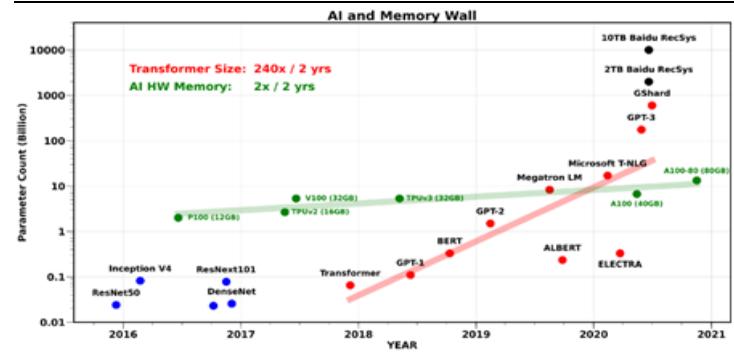


Source: Samsung

### Developments in memory: Variations of memory subsystems offered – but most are not optimal for AI

Historically, memory designs offered limited choices in memory subsystems. DDR4 up until recently was the primary choice for the main memory used in compute, offering speed of 3.2Gb/s with interface bandwidth of 25.6GB/s. That DDR design has made evolutionary upgrades including the current generation DDR 5 being introduced this year. Graphics memory have also evolved primarily for gaming systems and upgrading from GDDR5 (8Gb/s) and GDDR6 (16Gb/s), offering much faster speed. For low power applications, the industry also

**Figure 249: AI Memory Wall – Memory advances cannot keep up with the performance improvement in transformer models**



Source: Amir Gholami blog

has adopted LPDDR4/5 which combined gives specific choice of subsystem by usage depending on the application and CPU/GPU requirements.

**Figure 250: AI servers using HBM memory – high cost but offers higher bandwidth over other memory types**

	DDR4	DDR5	HBM2	GDDR5	LPDDR4	LPDDR5
Applications	Servers → PCs → consumer	Servers → PCs → consumer	Graphics, HPC	Graphics	Mobile, auto, consumer	Mobile, auto, consumer
Typical interface (primary)	Server: 64+8 bits	Server: dual channel, 32+8 bits	Octal channel, 128-bit (1024 bits total)	Multi-channel, 32-bits	Mobile: quad channel, 16-bit (64-bits total)	Mobile: quad channel, 16-bit (64-bits total)
Typical interface (secondary)	Consumer: 32 bits	Consumer: 32 bits	None	None	Dual channel, 16-bit (32-bits total)	Dual channel, 16-bit (32-bits total)
Max Pin BW	3.2 Gb/s	6.4 Gb/s	2.0 → 2.4 Gb/s	8Gbs	4.267Gb/s	6.4Gb/s
Max I/F BW	25.6 GB/s	51 GB/s	307 GB/s	32 GB/s	34 GB/s	51 GB/s
# Pins/channel	~380 pins	~380 pins	~2,860 pins	~170 pins	~350 pins	~370 pins
Max capacity	3DS RDIMM: 128GB	3DS RDIMM: 256GB	4H Stack: 4GB	One channel: 1GB	4 channels: 2GB	4 channels: 4GB
Peak volumes	*****	*****	**	*	*****	*****
Price per GB	\$	\$\$	\$\$\$\$	\$\$\$	\$\$	\$\$

Source: Synopsys

### **HBM3 seems best for HPC (High Performance Computing) for AI and ML (Machine Learning)**

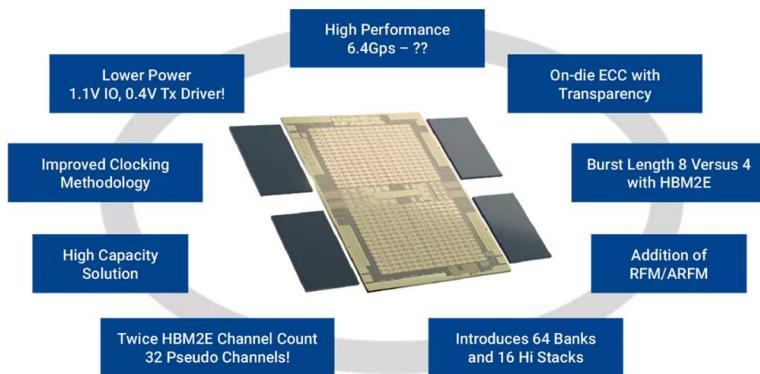
Higher performance memory requirements and need for more bandwidth to feed the processing for AI has led to the introduction of HBM (high bandwidth memory), which has slower speed (getting faster by introduction of HBM3 up to 6.4Gb/s)) but offers a much wider maximum interface bandwidth of 307 GB/s (in case of HBM2, 819GB/s for Samsung Icebolt HBM3).

HBM was designed to address some of the inadequacies of other DRAM designs to specifically improve bandwidth, latency, power consumption, capacity and cost. From a bandwidth improvement perspective, standard DRAM bus is from 4 to 32-bits wide. HBM bus is 1,024-bits wide, thus being able to handle much more data throughput. To help address the latency issue, HBM chip size is comparatively tiny vs. other DRAM chip designs at 35 mm-square for 1GB of HBM (compared to 672 mm-sq for 1GB of GDDR), which can be stacked up to 12 layers (up to 16-layer of 32Gb die stack possible for 64GB total density likely in next generation HBM3E) where each die is connected by TSV (through silicon via) interconnect technology. This stack design essentially addresses both latency and power issues as it minimizes the amount of time the data needs to travel (speed).

The HBM stack normally is placed adjacent to CPU/GPU connected by a substrate to address the power consumption issues as data movement distance between the memory stack and CPU/GPU resulting in power savings. In addressing the cost issue, HBM3 does not come cheap. Today, it is believed that HBM3 is 2x more expensive than the DDR5 counterpart. It could be too expensive to go mainstream. While traditional DRDs can be used, it would be in DIMM format, most likely too slow and not optimal from memory load requirement for AI. Calculation has to be done in terms of total operating cost (TOC) with all parameters included.

HBM has evolved generationally from the original HBM, HBM2, HBM2E and now to HBM3. The HBM3 standard has been approved by JEDEC January this year. With the industry standard set, all three major memory producers, Samsung, SK Hynix and Micron, are able to produce and offer HBM3 chips. Actual sales of HBM DRAM chips still remain small relative to all other types of DRAMs so far, only accounting for low-single digits % of total DRAM sales for both Samsung and SK Hynix. However, the growth is expected to accelerate as more AI driven applications are deployed.

**Figure 251: HBM3 offers power/performance benefits over HBM2E**



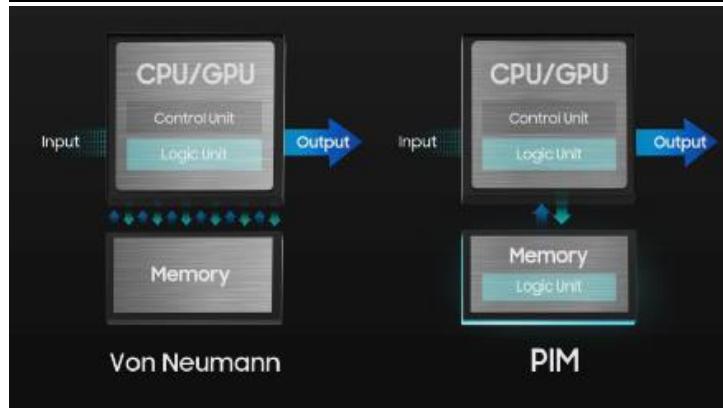
Source: Synopsys

## Processor in memory

The next phase in the memory development for AI is the PIM (Processor-in-Memory) design. The idea behind PIM is that for some AI tasks, compute units are designed to do the most common neural-network calculation, called multiply and accumulate, and little else. The PIM design addresses these issues by offloading some of the data calculation work from the processor to inside the memory. A logic chip designed to execute specific calculation task is physically embedded within the memory stack. In systems without PIM, the processor calls and executes commands from the memory and saves the results back in the memory (memory storage). This requires large amounts of data to be moved back and forth which takes significantly more energy compared to doing the processing of the data. PIM optimizes this process by reducing the data movement between the CPU and memory, improving the performance and energy efficiency of the AI accelerator system.

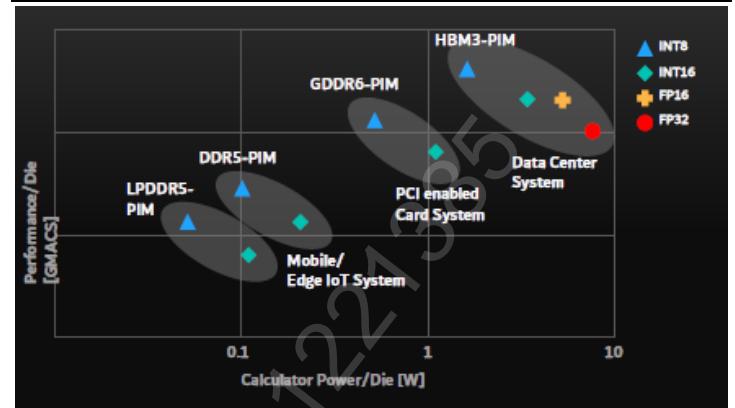
### PIM not limited to HBM3

PIM design does not have to be attached to a certain type of memory such as HBM3. The PIM concept can be implemented into DDR5, LPDDR5, or GDDR6 for example. In Samsung's case, it has introduced HBM2-PIM in 2021 and tested on Xilinx (AMD) Virtex Ultrascale + (Alveo) AI accelerator which showed 2.5x system performance gain and 60% reduction in energy consumption. With other CPUs such as Intel Sapphire Rapids, AMD's Genoa, ARM's Neoverse platform and other new generations of processors designed to support HBM, the applications will likely broaden. Additionally, HBM3 is now standardized by JEDEC, HBM3-PIM memory solutions are expected to be followed at Samsung.

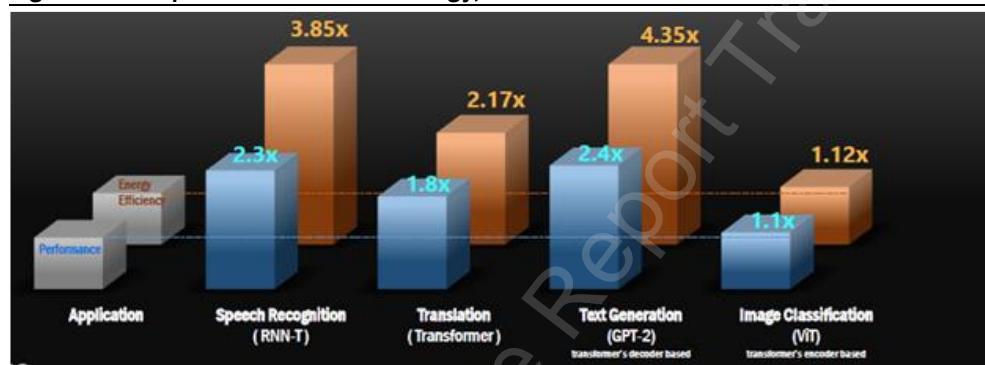
**Figure 252: Using PIM to overcome the memory bottleneck**

Source: Samsung

Similarly, SK Hynix has also introduced PIM (they call it AiM – accelerator in memory) in February 2022 with the sample on GDDR6. For now, it is focused more on PIM support on more conventional chips versus on HBM3, believing it is simpler for customers to adopt. SK Hynix researchers did work with Purdue University computer scientists on a comprehensive design of an HBM-PIM product called Newton in 2019. It plans to further develop HBM-PIM as a near-future product.

**Figure 253: Prospective PIM-supported data format**

Source: Samsung

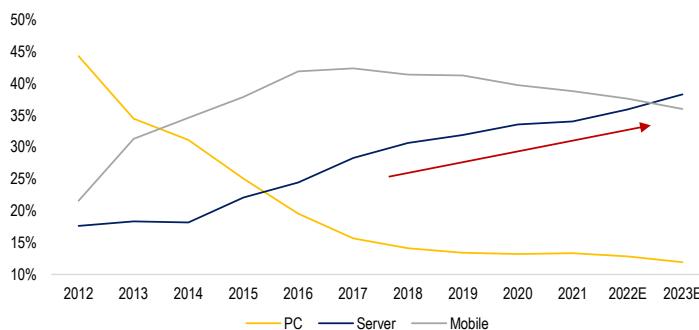
**Figure 254: Expansion of PIM technology, LPDDR5-PIM**

Source: Samsung

### AI dedicated servers can have a leverage effect for memory consumption

AI is expected to drive demand for high-performance computing, thereby server requirements filling the datacenter. Already DRAM consumption by the server demand segment has become the largest, overtaking the mobile handsets. In addition to the server unit growth, memory density per unit of servers would accelerate both to support GPU/CPU and broader system memory. Generally, hardware FLOPs will continue to accelerate as AI training models get bigger, broader and more complex. Memory requirement to train AI models are generally many times larger than the number of total parameters, typically by 3x to 4x, due to storage requirement of intermediate activations during training.

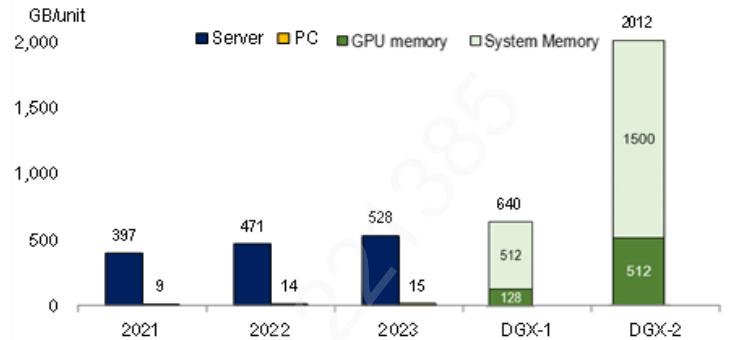
**Figure 255: Servers the largest segment for DRAM now - AI likely to continue lifting servers' portion of consumption**



Source: Credit Suisse Estimates

When we compare a simple DRAM requirement for PCs of 15GB per unit in 2023, average server DRAM requirement towers above PC at 528GB per unit. As a sample, for NVIDIA's DGX-1 deep learning system introduced in 2016, it contained total DRAM memory for both accelerator and system at 640GB. It was followed by the introduction of DGX-2 in 2018, it expanded GPU memory by 4x (GPU expanded 2x from 8 units to 16 units) while the main CPU memory expanded by 3x for combined system memory of 2TB – almost 4x memory expansion. As AI systems penetration rates rise higher within the server segment, we think DRAM demand leverage will accelerate.

**Figure 256: Average density for a server for DGX-1 and DGX-2 vs. a traditional PC**



Source: Samsung

**Figure 257: NVIDIA's hardware upgrades have carried with it higher memory specs**

	DGX-1	DGX-2
GPU	8x Tesla V100	16x Tesla V100
TFLOPS	960	2000
GPU Memory	128 GB	512 GB
CPU	Intel Xeon ES-2698 v4	Intel Xeon Platinum 8168
System Memory	512 GB	1500 GB
NAND	7.7 TB	30 TB
Performance		2 x
GPU		2 x
HBM Memory		4 x
CPU Memory		3 x
NAND		4 x
Logic Silicon Area (sq mm)	7432	14040
DRAM Silicon Area (sq mm)	26240	82492
NAND Silicon Area (sq mm)	18000	72000
Memory to Logic Chip Area	6.0 x	11.0 x
Memory To Logic BOM Cost	1.5 x	2.8 x

Source: Company Data, Credit Suisse Estimates

## IC Substrates

The growing demand for digitalization requires higher computing power and faster/more reliable connectivity. This has been identified as one of the key growth drivers for ABF substrates beyond the PC era, as it requires higher layer counts or larger area size substrate design.

Pauline Chen  
Akinori Kanemoto  
Sanguk Kim

Substrates design for AI (artificial intelligence) depends on the chip types, i.e., GPU or ASIC (application-specific integrated circuits), but generally they need to be compatible with other components used in the system (i.e., processors, sensors, and memory etc.), and capable of handling signal quality, noise reduction, thermal dissipation, and other electrical properties such as conductivity, resistance, and dielectric constant.

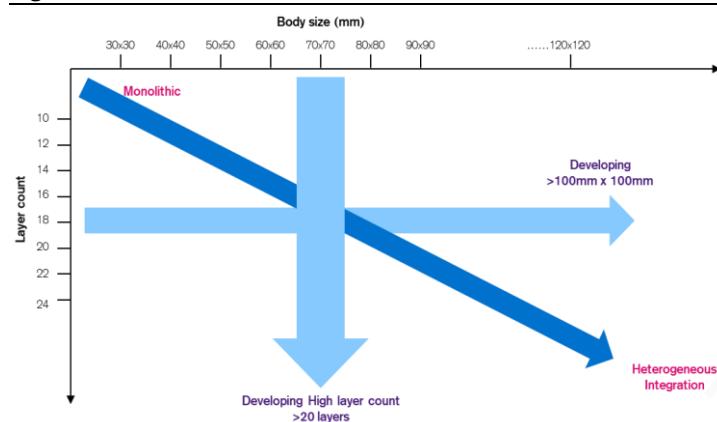
According to the supply chain, around half of the ABF substrates demand is estimated from PC applications, while the rest is estimated to be driven by GPU (for graphics and AI training etc.), FPGA (for networking and AI interface etc.), ASIC (for specific task i.e., cryptocurrency mining, machine learnings etc.). Specifically for Chat GPT, it is said to use more than 30,000 units of

GPU to support the estimated 25 mn visitors per day. As a result, a growing number of AI user cases would support GPU LT growth, after sharp inventory adjustments in the consumer market currently.

#### **Migration from GPU to ASIC could be positive for substrates, given customized design.**

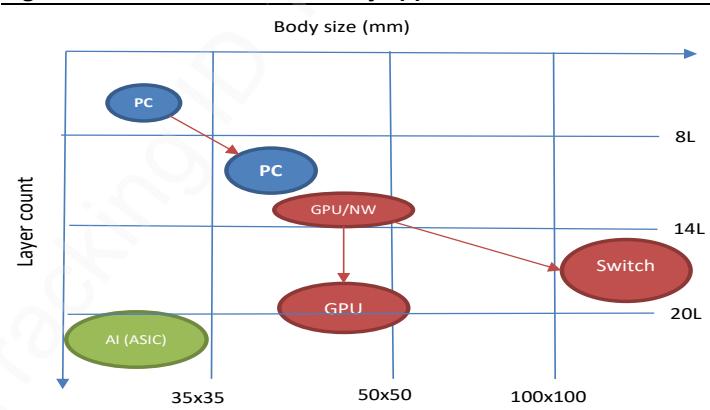
Compared to ASIC, substrates design for GPU typically has a larger form factor and requires more power and cooling, so that it can handle a wider range of tasks. On the other hand, ASICs are typically more optimized for a specific task, which makes it more efficient to achieve higher performance with lower power consumption, compared to a general-purpose processor like a CPU. As a result, substrates used in ASICs tends to be smaller (compared to GPU) but could have more layers (subject to the complexity of the application). As each design is customized, substrates prices are typically higher, given its smaller volume and lower yield rates.

**Figure 258: Trend of Advanced FCBGA substrate**



Source: Company data, Credit Suisse

**Figure 259: FCBGA substrates by application**



Source: Company data, Credit Suisse

#### **3D packaging to take some value away from ABF substrates in the high-end.**

We still see that faster adoption of 3D packaging could shift some value away from organic substrates to silicon substrates. This puts ABF substrate opportunities likely on 2.5D-minus or below packaging i.e., FanOut (FO), InFo on Substrates, or EMIB. While the design for 2.5D or above packaging may still require IC substrates, the design of substrates is relatively simple with majority of the function shifting to Silicon interposer. According to CS TSMC analyst, R. Abrams, he [expected TSMC to grow its revenue in the backend business](#) through wider adoption of its InFo for high-end mobiles, CoWoS for high-performance computing applications, and new SoIC 3D stacking which should be in volume in 2023 with trialing with Apple, AMD and Intel.

Nevertheless, we acknowledge that silicon substrates' higher production costs (still 3x higher currently) could limit 3D packaging adoption in high-end i.e., HPC, AI and data centers. As a result, organic substrates remain a more cost-effective solution for applications including PC, servers, networking and ADAS in EV, which still accounts for the majority of organic substrate demand.

**Figure 260: Substrate opportunities (mostly on 2.5D below) and challenges**

	2.1D	2.1D	2.1D	2.5D	3DIC TSV+2.5D interposer
Example	InFo/FO	InFo/FO on substrate	EMIB	CoWoS/Foveros	SolIC
Silicon interposer		small	small	o	o
Substrates	o	o	o	simple substrate or PCB	simple substrate or PCB
Cost *	2	3	3	4	5
Performance *	2	3	4	4	5

Note: for cost and performance scale from 1 (lowest) to 5 (highest)

Source: Company data, Credit Suisse estimates

**Prefer Japan substrate makers over Taiwan.**

While we agree that technology migration to chiplet and growing AI demand should be directionally positive for ABF substrates demand, we maintain our relatively cautious stance on Taiwan substrate makers over Japan, given the following reasons:

1. Taiwan substrates companies have higher exposure to the low-end market, which is suffering from deeper and longer inventory adjustments.
2. Taiwan substrate companies had expanded their capacity more aggressively in the past two years, which gives them less flexibility in the correction period.
3. Taiwan substrates companies' valuations (on both P/E and P/B) are more expensive compared to Japan.

**Figure 261: Substrate peers at a glance**

	Unimicron	NYPCB	Kinsus	Ibiden	Shinko	SEMCO	ZDT
ABF sales % (21)	41%	50%	30%	45%	41%	5%	0%
CPU	vvv	vv	vv	vvv	vvv	vvv	
GPU	vv	v	vv	vv	v		
FPGA	v	v	vv	v	v		
ASIC	v	v	v				
Others	v	vvv	v	v	vv	v	
PC	vvv	vv	vv	vv	vvv	vvv	
Server	vv	v	vv	vvv	vv		
Switch/router	v	vvv	v				
Base station	v	v	vv				
Consumer etc	v	v	v	v	v	v	

Source: Company data, Credit Suisse estimates

Note: "vv" stands for key application, "v" stands for minor contribution

## New data center architecture for better energy efficiency (P. Chen)

CS team expects data center growth to moderate in 2023E, with inventory adjustments ahead of new CPU ramp-ups. Nevertheless, the team still holds a relatively optimistic view on global data center demand, given growing data traffic and continued shift from enterprise servers to cloud.

### AI data center electricity consumption = thousands of US residential utility customers

The rapid growth in data center means rising power consumption. Take Chat GPT for example, based on the assumptions of 25 mn visitors for 300-word question per day and nVidia A100 GPU structure, electricity consumption is estimated to be ~600,000 kWh per day. This compares to an average 29-30 kWh per day for a U.S. residential utility customer, according to U.S. Energy Information Administration data in 2021.

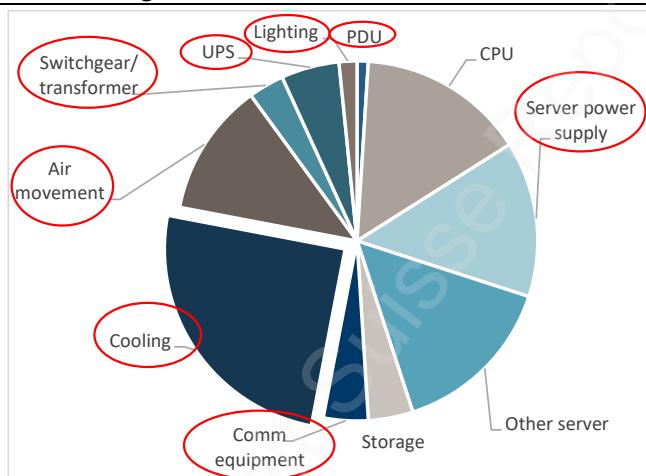
As a result, we expect the new data center architecture to better address rising energy consumption issue. According to our channel checks, it is expected to see the following trends for data center designs, including:

3. **Higher density.** Our channel checks suggest the average power capacity in a data center will increase from 6-8 kW/rack currently to 15-20 kW//rack by 2025E, given the rapid increase in data traffic and computing power, along with increasing costs.
4. **Scalable architecture.** Data center designs need to support scalable expansion for optimal capex, as the life cycle of data center infrastructure is 10-15 years, vs the life cycle of IT devices of 3-5 years. It also needs to support the hybrid deployment of IT devices with different power densities to support the diverse range of IT services.
5. **Green.** On a global scale, our channel checks suggest that ~3% of the world's total power consumption goes into data centers. How to save energy, cut emissions, and lower opex are important for data center design. 'Power Usage Effectiveness' (PUE)

improvement is the key matrix driver for a green data center, which is estimated to decline from 1.67 in 2019 to <1.1 in the next five years. Nevertheless, reducing PUE does not mean overall energy consumption of data centers is optimal. Innovation is still needed in facility, CPU, data, cloud and IT to achieve optimal energy efficiency for the whole system.

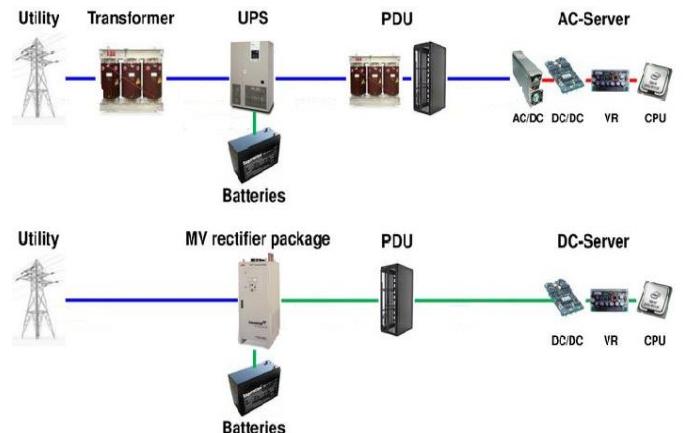
6. **Modularization.** Modular design includes component modularization, architecture modularization, equipment room modularization, to a full modularization of data centers. This will enable faster deployment, flexibility capacity expansion, simple operating and maintenance costs, and higher energy efficiency. A faster roll-out of data centers is critical to meet quick evolution of data center services and that the time-to-market of a data center will be shortened from 9-12 months currently to <6 months by 2025.
7. **Simplified power supply architecture.** The power supply and distribution system of a traditional data center is complex and fragmented. It also generates a larger footprint, which makes it difficult to locate faults. A simplified power supply architecture will reduce power conversion time, shorten the power supply distance and footprint, and improve the space utilization and energy efficiency. Our channel checks suggest DC data centers to provide better energy efficiency (given fewer conversions and less heat generated), to reduce floor space (given fewer power and cooling equipment), to reduce installation and maintenance costs (given higher reliability vs a simplified architecture).
8. **Convergence of cooling technologies.** We estimate that cooling related costs would be the largest energy consumer within a data center. Using natural cooling resources will largely reduce the power used in the cooling system, but compatibility for both air cooling system and liquid cooling system is also important, to better support the diverse range of IT services.

**Figure 262: Energy consumption in a data center, and Delta's product offerings**



Source: Company data, Credit Suisse

**Figure 263: Simplified power supply architecture**



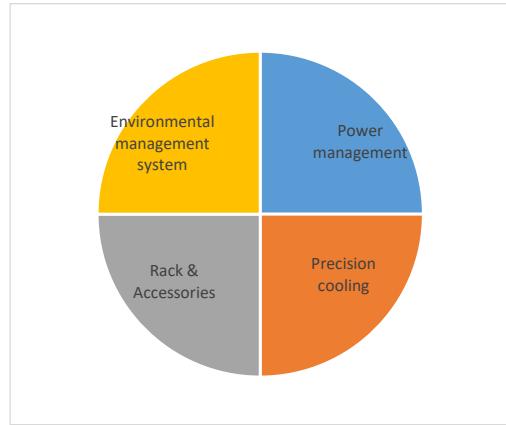
Source: Company data, Credit Suisse

#### Delta well positioned for new data center architecture

We expect the new data center architecture, including DC power architecture, modularized design, and convergencies of cooling technologies, to drive growth opportunities for the supply chain. **Delta Electronics** is well positioned in this trend, given its full range of product offerings and above-industry-average conversion efficiency. According to Delta, its InfraSuite Data Center Infrastructure Solutions are grouped into four main modules, (1) Power Management, (2) Rack and Accessories, (3) precision cooling, and (4) Environmental Management System. Delta claimed that its interlocking solution will maximize customers' operating efficiency at the lowest

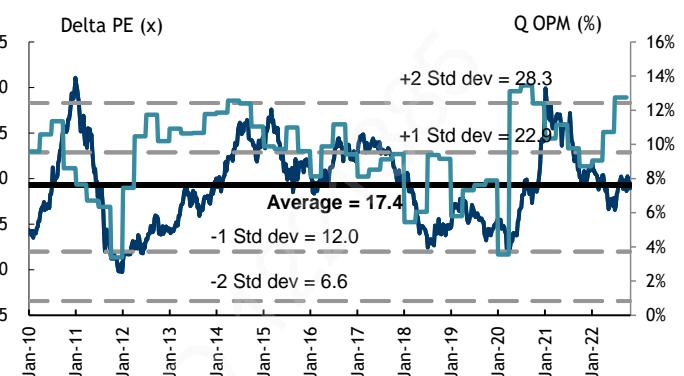
cost, maintain a high level of flexibility and control for IT managers, quickly scale to meet demand, and monitor data center solutions anytime and anywhere.

**Figure 264: Delta's data center offerings**



Source: Company data, Credit Suisse

**Figure 265: Delta—OPM back to record high, but P/E is not**



Source: Company data, Credit Suisse

## Asia Cloud IT Infrastructure Sector

We continue to expect hyperscalers to be the industry pioneers ramping up their respective proprietary AI solutions and AI enabled backbone infrastructure presenting Taiwan cloud ODMs as the first wave of beneficiaries given their exclusive positioning within the US hyperscalers chain and early partnership with Nvidia on DGX/HGX system from 2017, as well as their strong design capability for developing AI solution systems for Enterprise applications.

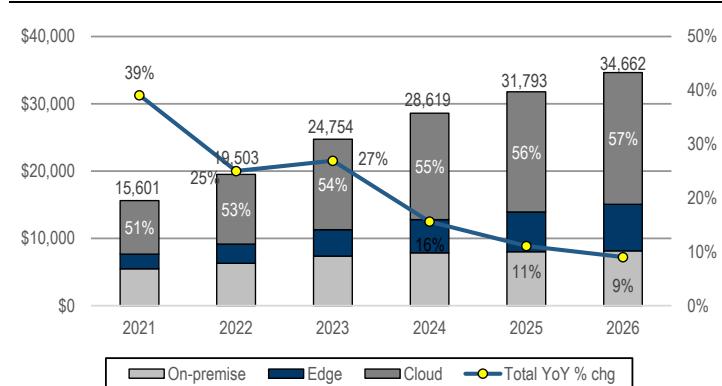
According to IDC, AI servers are servers that run AI platforms (AI applications development including most AI training), AI applications (AI models execution covering AI inferencing), and/or AI-enabled applications (applications with AI functionality). In addition, based on its latest forecasts, it expects AI servers infrastructure to grow at a 15% CAGR (vs whole cloud infrastructure for +10% CAGR) from 2022 to 2026 reaching a TAM of US\$34.7bn (14% of mix of US\$139bn TAM), vs 2022 of US\$19.5bn (11% of mix of US\$94bn TAM) with AI in cloud scenario as key driver for +17% CAGR for the same period despite already representing over 50% of the TAM from 2021.

For the cloud scenario, IDC expects accelerated AI servers in cloud (servers with co-compute by GPU, FPGA or ASIC with CPU) for both training and inferencing to outperform for +23% CAGR from 2022 to 2026, while non-accelerated AI server would grow at a +9% CAGR.

**Jerry Su  
Harvie Chou**

**Figure 266: AI server infra for +15% CAGR in 2022-26 with cloud as key driver**

Sales in US\$ mn



Source: Company data, IDC, Credit Suisse research

US hyperscalers are already shifting their investments toward AI infrastructure over the past years to help enhance public cloud services competitiveness, drive better top line conversion rate with consumers from advertisement/engagement and develop new business opportunities. For example, for the most recent quarter reporting earlier this year, despite adjustment on the capex guidance by Meta to US\$30-33bn for +/-5% YoY change in 2023, vs prior of +8-17% YoY, it still aims to accelerate its shift to AI native architected infrastructure (more GPU centric infrastructure vs CPU based solutions), which is cheaper and faster to build; while Google and Baidu recently also launched similar conversational AI services called Bard and Ernie, respectively.

**Figure 268: US hyperscalers' latest communication with increasing emphasis shift of capex toward AI infrastructure**

	Capex guidance	Cloud business guidance
Alphabet	2023 total capex to be inline with 2022 but with an increase in technical infrastructure versus a significant decline in office facilities	Excited about the long-term market opportunity in GCP, while remaining very focused on Google Cloud's path to profitability
Amazon	n.a.	Total 1Q sales of US\$121-126 bn for +4-8% YoY (2.1 pp impact from forex). For AWS, optimization efforts will continue to be a headwind at least for the next couple of quarters (Jan-2023 YoY at mid-teens % level)
Meta	2023 for US\$30-33 bn (+/-5% YoY) for costs optimization factoring in slower data center construction spend and strategic shift to a new data center architecture with better support for both AI and non-AI workloads	1Q sales to be in a range of US\$26-28.5 bn for -7% to +2% YoY (forex 2 pp headwind)
Microsoft	Sequential increase on a dollar basis in C1Q with normal quarterly spend variety in timing of cloud infra build-up	Intelligent cloud sales of US\$21.7-22.0 bn in C1Q for +17-19% YoY in constant currency with Azure as key driver

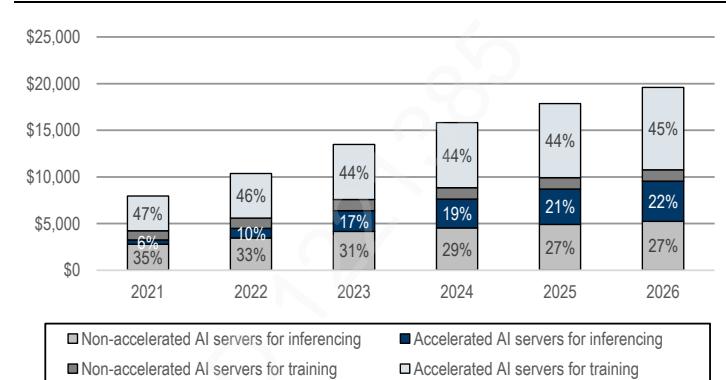
Source: Company data, Credit Suisse research

For the next wave CSPs and enterprises, we also see increasing interest on AI solutions, but our checks suggest the adoption of performance-intensive computing workload could now take place at a slower pace given the complexity regarding integration for optimized infrastructure environment for different verticals and scenarios, and a higher upfront investment to scale out the intended AI programs, vs direct adoption of public cloud services. This is also why Nvidia announced in late-February during its F4Q23 results call about its DGX cloud rollout through partnership with CSPs such as Oracle, Microsoft Azure, Google GCP and others providing full stacks services in the cloud which help democratize the access to AI infrastructure for expanding audience base in the mid to long-term.

Specifically, from a hardware system design perspective, we expect a much higher content per system especially led by the required adoption of coprocessors/AI accelerators including GPUs, FPGA, ASICs, and FOCPs, aside from preferrable upgrade to latest inter-connection tech like DDR5 and PCIe5 in order to achieve more efficient parallel compute. Based on our industry checks, we estimate servers for compute intensive AI training purpose (i.e., ChatGPT-like generative AI services) would require dual Intel 4<sup>th</sup> Gen Xeon CPU and 8 Nvidia's SXM5 H100 GPU module which could cost around ~US\$200K, significantly higher than the blended server

**Figure 267: Accelerated AI servers with co-compute edging to 70% of mix by 2026**

Sales in US\$ mn



Source: Company data, IDC, Credit Suisse research

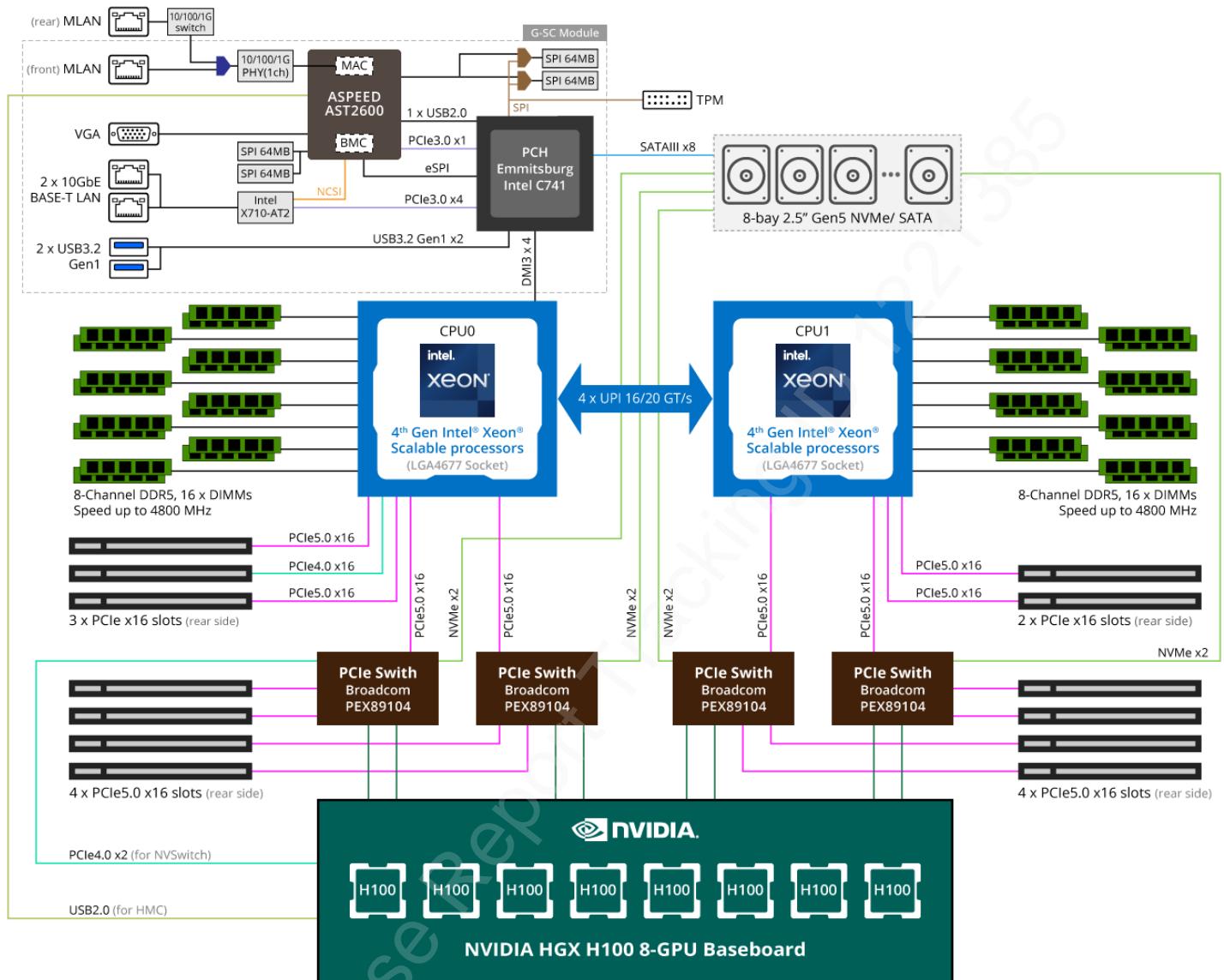
ASP by leading OEMs of US\$10K+. However, cloud IT infrastructure OEM/ODM also consider TCOs, power consumption, computing efficiency in terms of number of processing cores (CPU now maximum at 192 cores vs over 1,000 for GPU) and parallel computing capabilities (Nvidia's H100 SXM5 module is currently the best solution for high-end AI training servers). However, supply chain noted Nvidia's offering of PCIe5 based H100 could be an affordable solution for entry-level AI training server, while Nvidia's A30 is preferable for high-end AI inferencing and T4 could be adopted for entry-level servers.

**Figure 269: Key spec comparison of Nvidia server GPU for parallel compute**

	H100 SXM	H100 PCIe	A100 SXM	A100 PCIe	A30
FP64	34	26		9.7	5.2
FP64 Tensor Core	67	51		19.5	10.3
FP32	67	51		19.5	10.3
TF32 Tensor Core	989	756		156   312	82   165
BFLOAT16 Tensor Core	1,979	1,513		312   624	165   330
FP16 Tensor Core	1,979	1,513		312   624	165   330
GPU memory	80GB	80GB		80GB HBM2e	24GB HBM2
GPU memory bandwidth	3.35TB/s	2TB/s		1,935GB/s	2,039GB/s
Max thermal design power (TDP)	Up to 700W	300-350W		300W	400W
					165W

Source: Company data, Credit Suisse research

Although Nvidia currently leads in the AI/HPC compute solutions, server makers noted AMD could also have a new solution for AI applications announcing by 2H23, leveraging its expertise in CPU and GPU. Several server makers are working on new products with AMD's new solutions and some industry participants believe this could be a starting point for AMD to gain share into AI applications at the expense of Nvidia's leading position. For Intel, the supply chain noted it currently is behind on GPU for AI applications, although it could use the latest Sapphire Rapids CPU for inferencing, instead of training.

**Figure 270: Block diagram of Gigabyte's HPC server G593-SDO for AI applications**

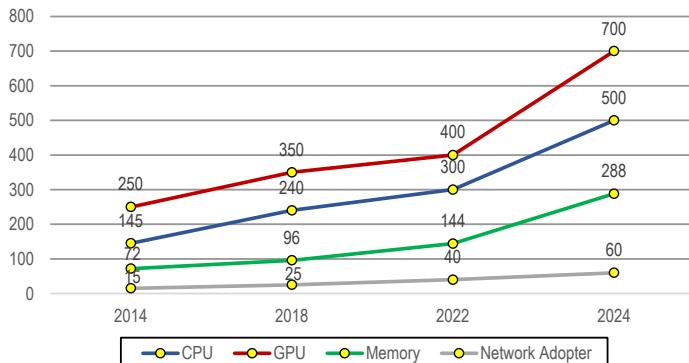
Source: Company data

The above figure is the block diagram of Gigabyte's AI server with Nvidia's HGX H100 8 GPU module. Besides adopting the latest Intel 4<sup>th</sup> Gen Xeon CPU, the system is also upgraded to high-speed interfaces with PCIe5.0 interconnect, as well as DDR5 with up to 4.8GHz. It has also adopted Broadcom's PCIe Switch for interconnection, which helps improve the signal quality under high-speed communication. This could also drive the requirement of faster connection at both edge and cloud, leading to further speed and spec upgrade for ethernet switches.

Beyond the driver of rising penetration of this high compute intensive application as a result of AI, we have also observed a consistently rising Thermal Design Power (TDP) delivery on major server components especially with CPU and GPU expected to reach 500W and 700W into 2024, vs 145W and 250W in a decade ago, although the T-case (i.e., the maximum temperature allowed at the processor integrated heat spreader) goes a separate way, implying more efficient heat dissipation methodology is required for a system to continue operating in optimal efficiency resulting in an enhanced TCOs which remains the central consideration especially for CSPs.

**Figure 271: Rising TDP across all major cloud components led by GPU and CPU...**

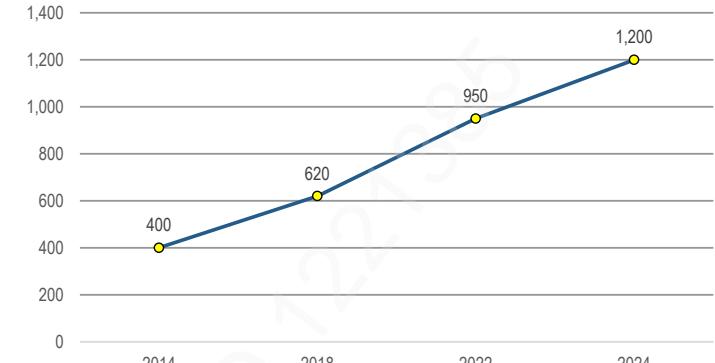
Units in Watts



Source: Company data, Credit Suisse research

**Figure 272: ... and total server system as a result also saw growing energy consumption**

Units in Watts for a typical 2 socket server



Source: Company data, Credit Suisse research

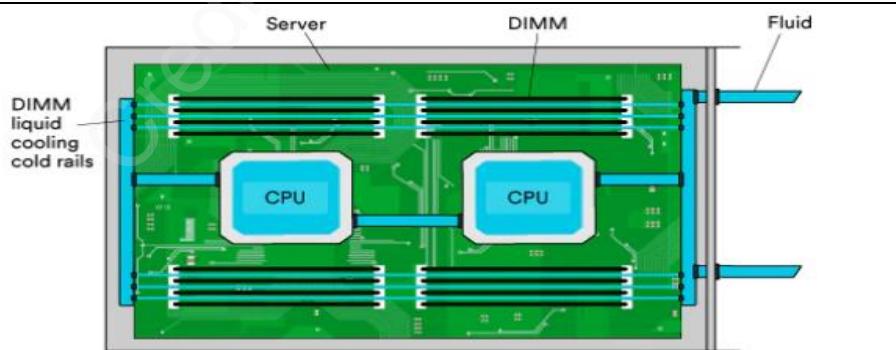
**Figure 273: Gigabyte's AI server product portfolio in a glance**

AI platform - G493		Training - G593-SD0	Inferencing - E152
			
Form Factor	4U	5U	1U
CPU	4th Gen Intel Xeon Scalable AMD EPYC 9004	4th Gen Intel Xeon Scalable	AMD EPYC 7002/7003
GPU	Supports up to 8 x double slot Gen5 GPU cards	Supports NVIDIA HGX™ H100 with 8 x SXM5 GPUs	Supports 1 x double slot or 2 x single slot GPU cards
Number of DIMM Slots	24 or 32 or 48	32	8
LAN Speed	1Gb/s or 10Gb/s	10Gb/s	1Gb/s
LAN Ports	2	2	2
Storage Bays	12 x 2.5" bays or 12 x 3.5" bays	8 x 2.5" bays	2 x 2.5" bays
PSU	Quad 3000W	6 x 3000W	Dual 1000W or Single 800W

Source: Company data, Credit Suisse research

Based on our conversations with cloud IT infrastructure ODMs, liquid cooling has been one central discussion in replacement of traditional air cooling given high compatibility with minimum changes required to existing air cooling system, leveraging cold plate arrays for heat dissipation from the processing units. Thus, with the rising adoption of AI accelerators/coprocessors, we believe this could also drive the increasing value-add for the ODMs.

**Figure 274: Liquid cooling drives better efficiency with heat transported away by special fluid within cold rails**

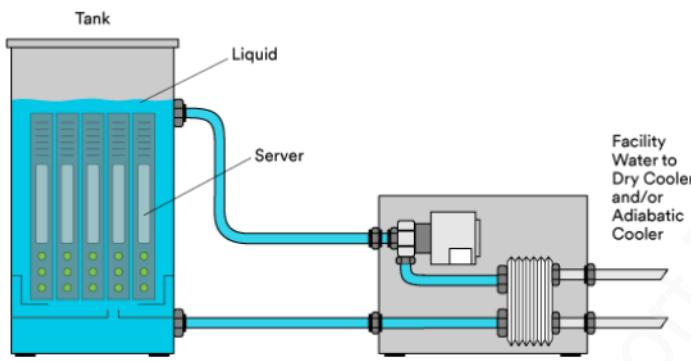


Source: Company data

Further, we expect rising traction of immersion cooling given the enablement of materially higher power density (i.e., stable temperature allowing over-clocking capability) and lower PUE (Power Usage Effectiveness) on the back of the long-term trend for higher-density and higher-performance compute. Key methodologies under liquid direct-to-chip (i.e., liquid with no direct contact with electronics with cooling through cold rails) and immersion cooling both could adopt either one-phase or the latest two-phase solutions differentiated by liquid temperature control (i.e., one-phase with heat dissipation through rear door heat exchanger vs two-phase via heat evaporation through rear door heat exchanger, and back to the loop after condensation), while the liquid in many cases is adopting 3M dielectric liquids, but we have also seen solutions utilizing Shell's liquid solution through its GTL (Gas-to-Liquid) technology.

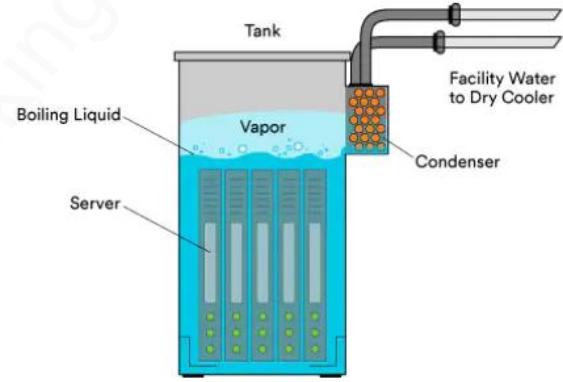
The other key differentiated factors between liquid and immersion cooling come from the fact that liquid cooling mainly cools down the temperature of CPU, while immersion cooling lowers the temperature for the entire boards, implying cloud operators would still need to maintain optimal air condition environment for other components such as memory, InfiniBand Card, MOSFET, etc., in the case of liquid cooling.

**Figure 275: Single-phase immersion cooling operates under sealed tank environment where fluid remains in liquid state**



Source: Company data

**Figure 276: Two-phase immersion solution instead cools down hardware with fluid going through evaporation/condensation**



Source: Company data

Overall, we see growing value proposition to both cloud ODMs and OEMs from hardware builds perspective in terms of both more complicated board designs and latest cooling solutions adoption along with rise of AI infrastructure. For ODMs, specifically, our checks suggest growing projects engagement by cloud ODMs with the hyperscalers from 2023 with larger orders ramp into 2H. Given the much higher ASP per system, we expect this could potentially offer incremental uplift to momentum outlook, although this does entail modest dilution on GMs due to higher content pass-through. For cloud IT infrastructure OEMs, though market adoption could take a slower pace, we believe ultimately the ability to offer total AI solutions from cloud to edge covering ready-to-use models/software, and scenario optimized architecture will be the keys to future success which does raise the entry barriers for competition in the long-term.

## Asia Hardware Stock implications

**Wiwynn (6669.TW):** Wiwynn is our preferred play among cloud ODMs as the only pure cloud play with 100% sales exposure to CSPs with key customers led by Meta and Microsoft, representing 50-60% and 30-40% of the total sales, respectively, while Wiwynn is also ramping up aggressively at Amazon, despite the overall contribution remains at single digits % in the past year. Notably, given Meta's clear infrastructure strategy shift to AI native architecture plus Microsoft's ongoing leadership on generative AI services, we expect Wiwynn should be the first wave of beneficiary riding on this rising AI tide. Although its prospective margin % could be dilutive as AI systems ramp (i.e., Wiwynn highlighted recently that over 50% of projects in its pipeline have some degree of AI content), management is confident about keeping a more stable margin rate trend by leveraging its leadership in cooling solutions as well as increasing its

value proposition on greater R&D and engineering resources involved in close partnership with its key customers rolling out the next Gen of cloud IT infrastructure.

**Quanta (2382.TW):** We believe Quanta could also benefit from the AI infrastructure take off given its leadership in cloud IT infrastructure, as well as having the most diversified customers base covering all of the US hyperscalers and better progress into Next Waves and enterprises, and also new business ramp with Telcos. Although the overall server business only accounts for ~30% of total sales vs ~50% for PC, we believe server offers higher profit contribution for the company given the latter carries a lower margin profile, especially for the iOS business. As a result, we believe the rise of AI system adoption should be more than enough to offset potential further order adjustment for PC demand in 2023-24.

**Inventec (2356.TW).** Inventec's server business represents ~40% of its total sales and is well balanced as compared to other cloud ODMs between traditional OEM vs ODM Direct business with roughly 55% and 45% of the mix. For its cloud business, it has unique exposure into selective Chinese hyperscalers with supplies of full rack solutions, aside from supplying into 3 of the 4 US hyperscalers aside from Meta mainly on its proprietary motherboard solutions. Thus, provided the US/China decoupling trend, we expect Inventec to stand at a better position to benefit not only along with further cloud TAM expansion, but also the rising AI tide. In addition, for its traditional OEM business, we expect Next Waves and enterprises should follow the suits after hyperscalers for the adoption of AI infrastructure for which we believe Inventec could also take an important part in with both HP/Dell as its key customers. Moreover, although overall PC market TAM is seeing some softness in 2023, Inventec is likely to be the only PC ODM to grow its shipment in 2023 given higher commercial mix and new project win at gaming OEM.

**Gigabyte (2376.TW):** Server accounts for 20-25% of Gigabytes revenue and its business model has shifted from pure ODM toward mostly channel business (~75% of server sales) since 2020. Gigabyte has been in close relationship with Nvidia for both graphic card and AI/HPC, and its AI server has entered into mega datacenters and national labs through its system integrator partners. Gigabyte has also been an earlier mover for working with AMD on server platform, where it has launched a comprehensive product line with AMD's Genoa CPU in 4Q22. We believe the rising adoption of AI servers should further support its revenue growth, while margin could also be accretive given the increasing mix of its channel business.

**Lenovo (0992.HK):** Lenovo's ISG business accounts for ~16% of its total sales and is now mostly exposed to cloud hyperscalers for 60%+ of ISG sales with Microsoft as the major customer, versus ~40% for traditional ESMB customers. It has been awarded more projects at Microsoft recently on competitive cost structure, as well as improving design capability evidenced by supplying server motherboard for the latest Sapphire Rapids platform. According to TOP500.org, Lenovo is the world's #1 supercomputer provider, including some of the most sophisticated supercomputers ever built. As a result, it believes its leading positioning in HPC systems and larger software partnership will help it capitalize the growing opportunities in AI across cloud and edge with also a much higher ASP profile. Nevertheless, majority of its revenue and profit (CSe ~2/3) still comes from PC, hence the slower PC TAM in 2023-24 and prolonging inventory digestion for the industry will continue to weigh on its profit outlook.

**Accton (2345.TW):** Accton is a leading cloud ODM specifically for networking ethernet switch and supplying into US hyperscalers including Amazon, Meta, and Microsoft. It has enjoyed good decent ASP/content increase as networking switch upgrades to 100G in 2016-19, and continues to benefit from current upgrading trend to 400G (from 100G) amid the rapid growth of data transmission and processing, especially for AI and machine learning. Overall revenue from cloud switch accounts for ~40% of total sales, while it is also the ODM partner for leading enterprise switch providers like HPE and Juniper. Accton also supplies SmartNIC for leading hyperscalers, which could help to offload the processing workload from the main CPU, as well as adding AI/machine learning functions to existing servers. It has also developed liquid cooling feature for high-speed switch solutions, which could help to capture the rising TAM from AI/ML in the mid- to long-term.

**Lotes (3533.TW):** Lotes is a leading high-speed inter-connect solutions provider with unique positioning in both Intel and AMD CPU socket chain with at least 1/3 of the global shares at the

latest generation of server CPU platforms, while it also holds oligopoly position in server directed DDR inter-connect, and a key supplier for PCIe inter-connect. Despite GPU may be the central of spotlight to perform massive parallel compute, we also expect a joint expansion of server CPU TAM as CPU is always required within an AI system to carry out core compute functionality aside from supporting data transfer between GPU cards. On the other hand, we believe the premium spec requirement for AI system will also drive faster penetration increase on the latest inter-connect technology including DDR5 and PCIe5 which will put Lotes as primary beneficiary with ASP at least double on DDR5 vs DDR4 given the complete shift in production process from DIP to SMT for the inter-connect.

**Parade (4966.TWO):** Parade is a leading provider for high-speed interface IC and display IC solutions, although ~80% of its end market is still exposed to NB/tablet. It has developed retimer chips for server PCIe4 and was adopted by leading server ODMs since 2021, although overall contribution remains limited given delay ramp of new CPU platform, as well as ODMs adopting alternative solution by redesigning the PCB. Its PCIe5 retimer solution is still under customer qualification and expects to be ready to ramp by late 2023, which is 6-12 months behind its peer Astera Lab. Nevertheless, Parade believes it could catch up with reasonable market share afterwards given it offers comprehensive solution for PCIe4 and PCIe5, as well as better cost structure as it owns the key know-how and manufactures at matured nodes.

**Chroma (2360.TW):** Chroma is a leading testing instrument and equipment provider, focusing on high power testing for various applications (such as EV, solar, battery, display, smartphone, etc.). It has developed SLT (System Level Test) solutions for HPC ICs with precise and compact thermal control system. Its SLT solution has been adopted by leading chipset makers including Nvidia, AMD, Broadcom, Qualcomm, Mediatek, etc. Chroma said it sees robust demand for SLT used on HPC/AI as its customers have placed orders and prepaid ~90% of the payments. Given the recent hiking demand for ChatGPT and AI/machine learning, Chroma remains positive on its SLT business, and believes it is likely to win more orders later on in 2023.

## Power supply - New data center architecture for better energy efficiency

CS team expects data center growth to moderate in 2023E, with inventory adjustments ahead of new CPU ramp-ups. Nevertheless, the team still holds a relatively optimistic view on global data center demand, given growing data traffic and continued shift from enterprise servers to cloud along with the incremental driver from AI.

**Pauline Chen**

### AI data center electricity consumption = thousands of US residential utility customers'

The rapid growth in data center means rising power consumption. Take Chat GPT for example, based on the assumptions of 25 mn visitors for 300-word question per day and nVidia A100 GPU structure, electricity consumption is estimated to be ~600,000 kWh per day. This compares to an average 29-30 kWh per day for a U.S. residential utility customer, according to U.S. Energy Information Administration data in 2021.

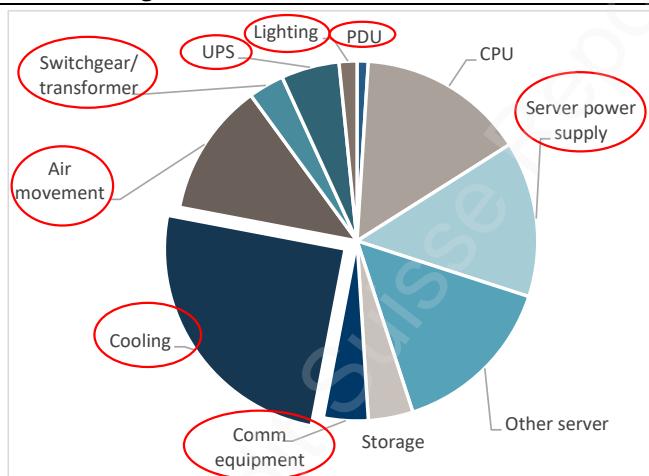
As a result, we expect the new data center architecture to better address rising energy consumption issue. According to our channel checks, it is expected to see the following trends for data center designs, including:

1. **Higher density.** Our channel checks suggest the average power capacity in a data center will increase from 6-8 kW/rack currently to 15-20 kW//rack by 2025E, given the rapid increase in data traffic and computing power, along with increasing costs.
2. **Scalable architecture.** Data center designs need to support scalable expansion for optimal capex, as the life cycle of data center infrastructure is 10-15 years, vs the life cycle of IT devices of 3-5 years. It also needs to support the hybrid deployment of IT devices with different power densities to support the diverse range of IT services.
3. **Green.** On a global scale, our channel checks suggest that ~3% of the world's total power consumption goes into data centers. How to save energy, cut emissions, and lower opex are important for data center design. 'Power Usage Effectiveness' (PUE)

improvement is the key matrix driver for a green data center, which is estimated to decline from 1.67 in 2019 to <1.1 in the next five years. Nevertheless, reducing PUE does not mean overall energy consumption of data centers is optimal. Innovation is still needed in facility, CPU, data, cloud and IT to achieve optimal energy efficiency for the whole system.

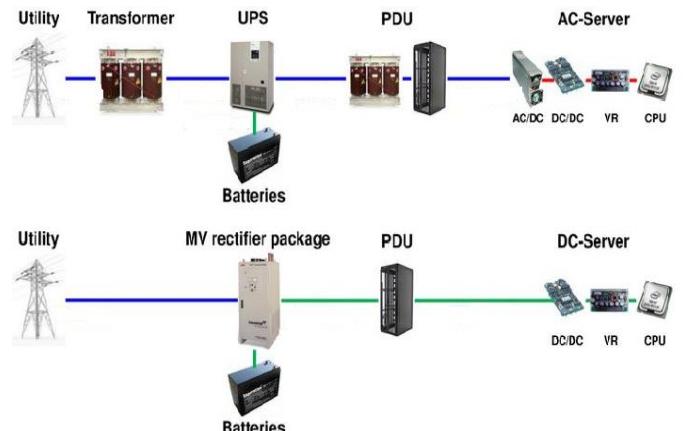
4. **Modularization.** Modular design includes component modularization, architecture modularization, equipment room modularization, to a full modularization of data centers. This will enable faster deployment, flexibility capacity expansion, simple operating and maintenance costs, and higher energy efficiency. A faster roll-out of data centers is critical to meet quick evolution of data center services and that the time-to-market of a data center will be shortened from 9-12 months currently to <6 months by 2025.
5. **Simplified power supply architecture.** The power supply and distribution system of a traditional data center is complex and fragmented. It also generates a larger footprint, which makes it difficult to locate faults. A simplified power supply architecture will reduce power conversion time, shorten the power supply distance and footprint, and improve the space utilization and energy efficiency. Our channel checks suggest DC data centers to provide better energy efficiency (given fewer conversions and less heat generated), to reduce floor space (given fewer power and cooling equipment), to reduce installation and maintenance costs (given higher reliability vs a simplified architecture).
6. **Convergence of cooling technologies.** We estimate that cooling related costs would be the largest energy consumer within a data center. Using natural cooling resources will largely reduce the power used in the cooling system, but compatibility for both air cooling system and liquid cooling system is also important, to better support the diverse range of IT services.

**Figure 277: Energy consumption in a data center, and Delta's product offerings**



Source: Company data, Credit Suisse

**Figure 278: Simplified power supply architecture**



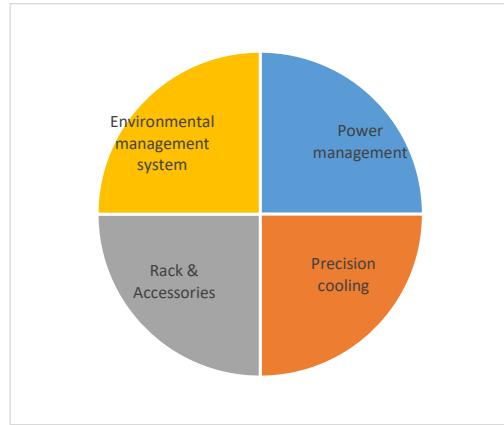
Source: Company data, Credit Suisse

#### Delta well positioned for the new data center architecture

We expect the new data center architecture, including DC power architecture, modularized design, and convergencies of cooling technologies, to drive growth opportunities for the supply chain. **Delta Electronics** is well positioned in this trend, given its full range of product offerings and above-industry-average conversion efficiency. According to Delta, its InfraSuite Data Center Infrastructure Solutions are grouped into four main modules, (1) Power Management, (2) Rack and Accessories, (3) precision cooling, and (4) Environmental Management System. Delta claimed that its interlocking solution will maximize customers' operating efficiency at the lowest

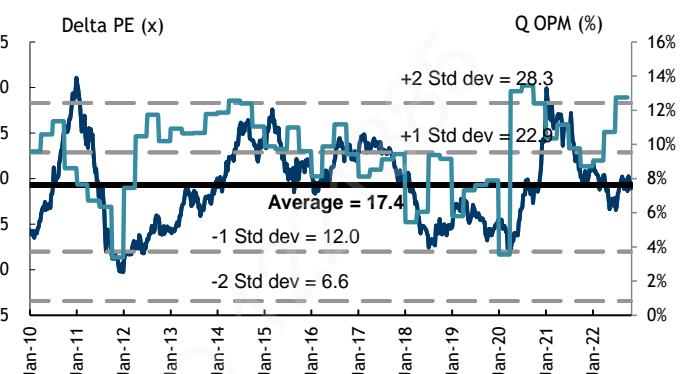
cost, maintain a high level of flexibility and control for IT managers, quickly scale to meet demand, and monitor data center solutions anytime and anywhere.

**Figure 279: Delta's data center offerings**



Source: Company data, Credit Suisse

**Figure 280: Delta—OPM back to record high, but P/E is not**



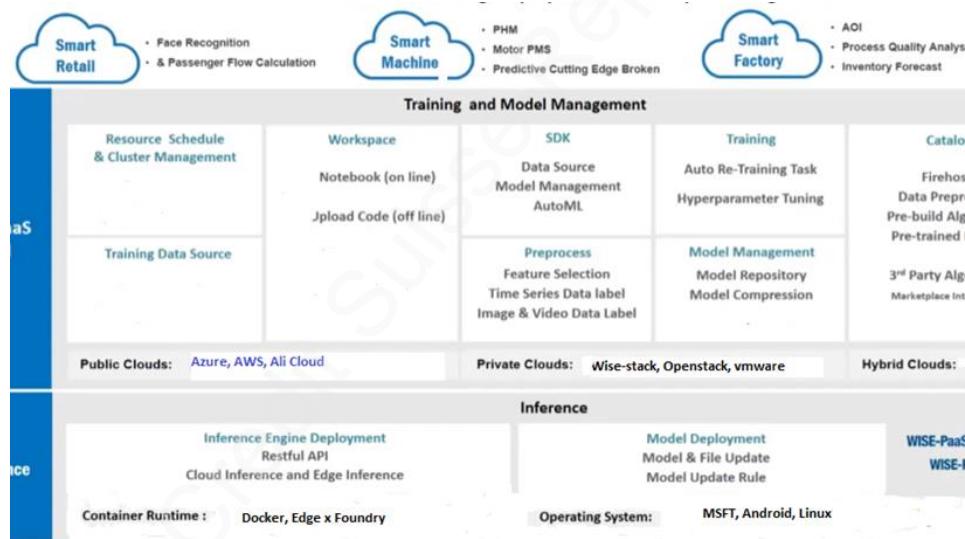
Source: Company data, Credit Suisse

## Industrial PC looking for AI opportunities

Taiwan's largest IPC brand company, Advantech, is also eyeing for industrial AI opportunities through its cooperation with nVidia. Advantech has launched its WISE-PaaS/AFS (AI Framework Service) in 2019, which is stated to be a unified platform for AI models training, deployment, re-training, re-deployment at scale, and life cycle management. The WISE-PaaS/AFS, together with its WISE-PaaS/Dashboard (which provides data visualization) and WISE-PaaS/APM (Asset Performance Management), forms Advantech's WISE-PaaS AIoT service framework, targeting to improve efficiency and accuracy, and to better address rising touchless demand post pandemic.

**Pauline Chen**

**Figure 281: Advantech's WISE-PaaS/AFS architecture**



Source: Company data

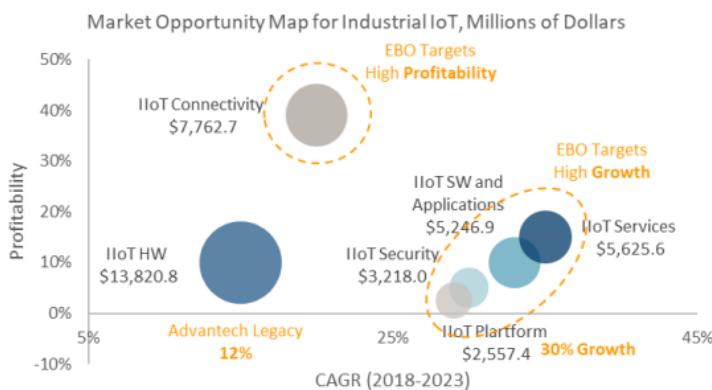
## Edge AI opportunities for IPC

While Advantech already has a leading position in IPC hardware (edge), it still sees good value accretion from Edge AI, given the rising number of influence instance (defined as applications post AI training) and video streams. As a result, Advantech has been making inroad to IPC software from 2015 and further expanding into SaaS in 2019.

According to Advantech, its industrial AI solutions integrate AI camera, Edge AI devices, Edge AI IPC + iModule, and Edge Accelerator servers, which should solve the common pain points of adopting industrial edge machine learnings such as data collection, data labeling, domain knowhow, and integration and deployment. It also supports hybrid cloud, subject to customers' requests.

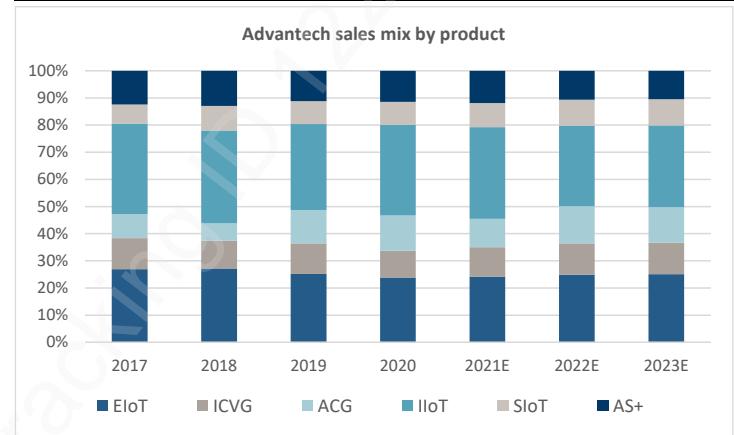
For its software platform, Advantech specifically targets at four verticals, including EMS (Energy Management System), iFactory, iRetail, and machine management. Advantech views itself not just an edge hardware provider, but more a resource integrator for AIoT ecosystem (e.g. to connect and support different domain knowhow suppliers in AIoT).

**Figure 282: Gartner market opportunity for IIoT**



Source: Company data, Credit Suisse estimates

**Figure 283: Advantech's sales mix by product**



Source: Company data, Credit Suisse estimates

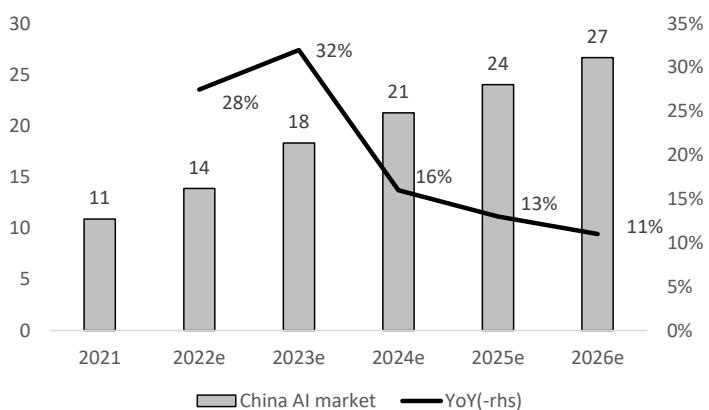
## China Technology Sector

We see many applications and scenarios applicable to ChatGPT and the underlying AI-generated content (AIGC) technology in China along with AIoT, smart city, smart education, smart manufacturing, smart business, Metaverse developments etc. AI has come to play an important role in China's 'Made in China 2025' blueprint. China aims to become a global leader in smart manufacturing by 2030 and AI is a key enabler. China aims to pursue leadership in the AI field through three steps: first, it must be able to keep pace with all leading AI technology, and its application in general, by 2020. Second, China has to make major breakthroughs by 2025, which are intended to lead to the last part of the plan: the establishment of China as the world leader in the AI field by 2030. (3) China targets the core industry scale of artificial intelligence to exceed Rmb400 bn and the scale of related industries to exceed Rmb5 tn in 2025. By 2030, the core industry scale of artificial intelligence is expected to exceed Rmb1 tn, and the scale of related industries to exceed Rmb10 tn.

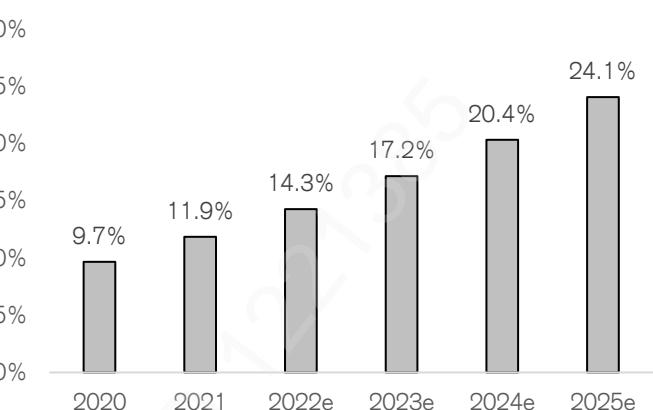
According to Frost & Sullivan, China's AI market has reached approximately RMB186 billion in 2020, accounting for 9.7% of the global AI market, and is expected to reach RMB1,046 billion by 2025, accounting for 24.1% of the global AI market. We expect that China AI industry outgrows in near future thanks to digitalization demand, policy driver, technology upgrade etc. Enterprises have more incentive to deploy digital transformation, which has thus spawned more demand for AI. Driven by policies, technologies, and markets, AI empowering industries is becoming a mainstream trend.

Kyna Wong  
Chaoliens Tseng  
Yufeng Shen  
Clive Cheung  
Edward Liu

**Figure 284:** China AI is expected to grow 41.3% CAGR 20-25E  
(Rmb Bn)



**Figure 285:** China AI as percentage of Global ramp up

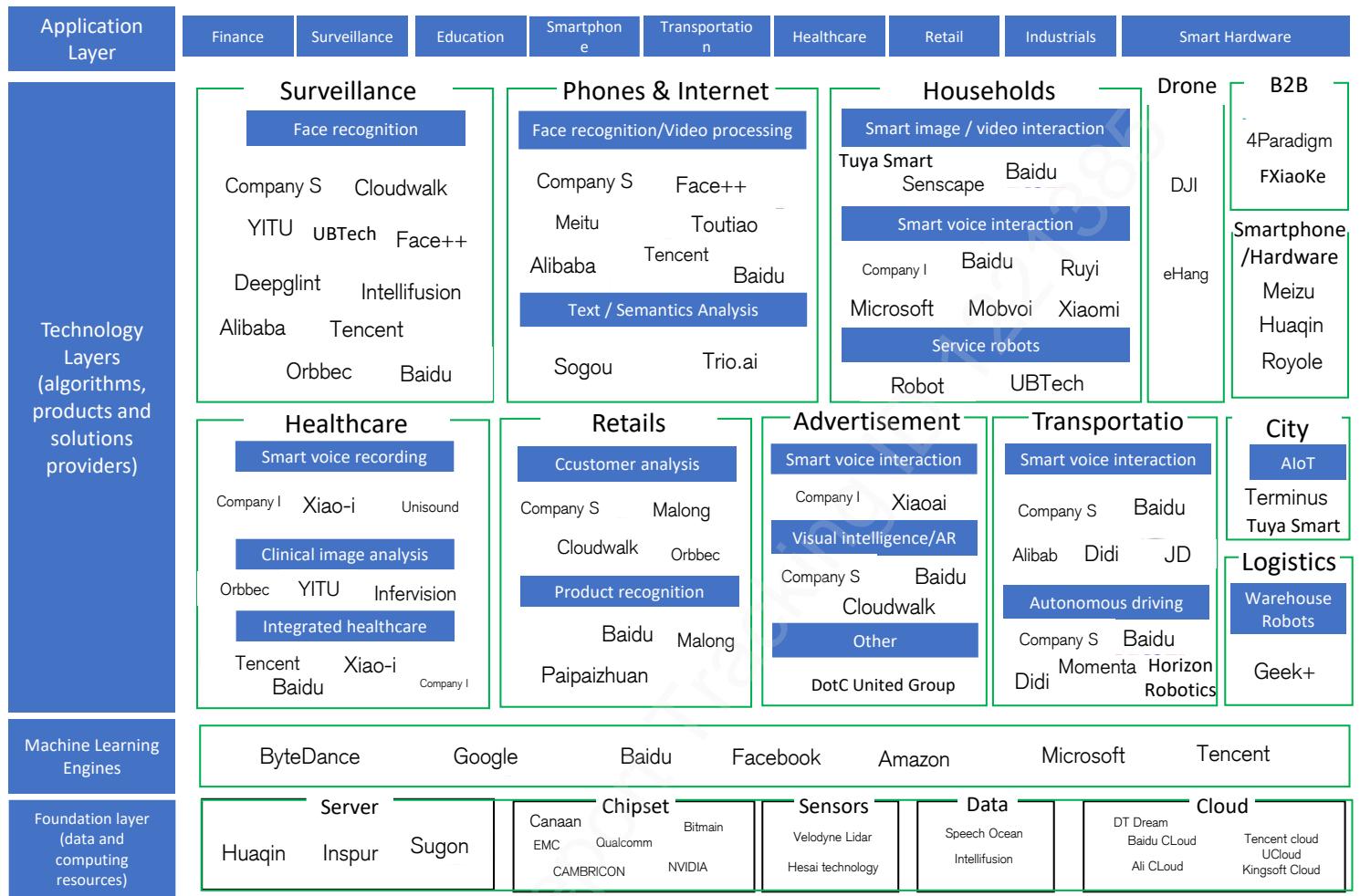


Source: Frost and Sullivan, Credit Suisse

In the AI value chain, there are three major layers: the technology layer, machine learning engine, and foundation layer. The foundation layer provides infrastructure such as server, chipset, sensors, data resources and cloud computing resources, etc. The machine learning engine is the platform that provides the deep learning process and makes computers learn from the data. The technology layer includes algorithm providers, products and solutions providers, etc. We expect some common players across different applications in the AI ecosystem to leverage their technology capability. ChatGPT or AI-generated content could well be incorporated in different AI applications.

Source: Frost and Sullivan, Credit Suisse

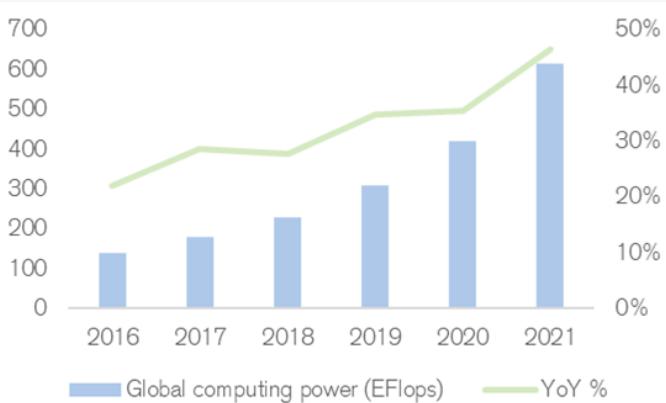
Figure 286: Overall AI ecosystem companies in China



Source: Company data, Credit Suisse; \* Company S is a leading domestic AI software provider and Company I is a leading AI speech solution provider.

The proliferation of AI applications significantly expands the demand for computing power. According to CAICT, global computing power reached 615EFlops by 2021, at a CAGR of 26% in the past three years. Huawei GIV expects the number will increase to 56ZFlops by 2030, at a CAGR of 65%. What's more important is the rising mix of AI computing. Basic computing power will increase from 269EFlops in 2021 to 3.3ZFlops in 2030, at a CAGR of 27%, supercomputing will increase from 14EFlops in 2021 to 0.2ZFlops, at a CAGR of 34%, while AI computing will increase from 232EFlops in 2021 to 52.5ZFlops in 2030, at a CAGR of 80%, which becomes the largest growth driver in the next decade.

**Figure 287: Global computing power grow at a 26% CAGR in 2019-21**



Source: CAICT, IDC, Gartner, HPC TOP100

China, which represented 33% of global computing power by 2021, will deliver a similar growth pattern. Its basic computing power reached 95EFlops by 2021, at a 24% CAGR in 2016-21, supercomputing reached 3EFlops by 2021, at a 30% CAGR in 2016-21, while AI computing reached 104EFlops by 2021, at an 85% CAGR in 2016-21. We believe the boost of ChatGPT-like AI applications will likely continue to drive the high growth in the next years.

**Figure 289: China computing power outgrows the global**



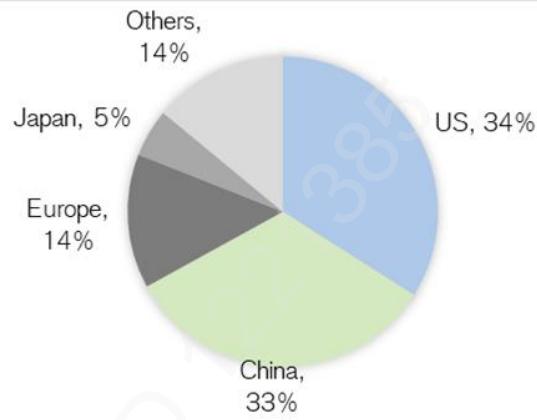
Source: CAICT, IDC, Gartner, HPC TOP100

## Hardware

IDC estimate that hardware will be the largest primary market in China's AI market in the next 5 years, contributing to more than 50% of the total AI investment domestically. IDC also predict that China's IT investment in the AI hardware market will exceed US\$15bn in 2026, close to that of the AI hardware market size of the US.

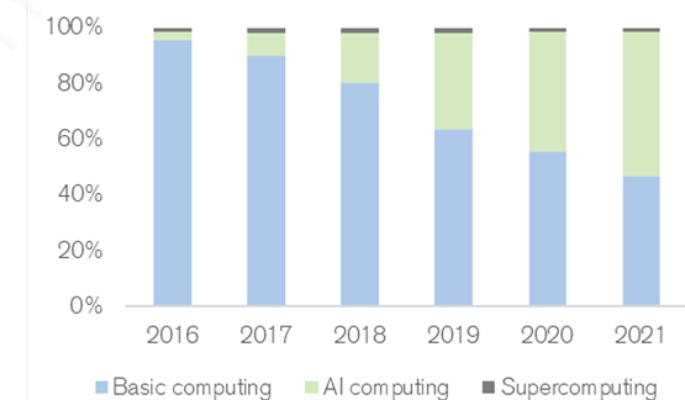
AI servers are the key to provide computing power to support the emerging demand for AI learning and analytics, which represented 87% of AI infrastructure's value by 2020 according to IDC. IDC forecasts global/China AI server market will grow from US\$15.6bn/US\$5.4bn in 2021 to US\$31.8/US\$11.3bn in 2026, at a CAGR of 17%/16% in 2021-26, while we believe the boost of ChatGPT will likely stimulate the procurement of AI servers and bring upside. Chinese AI server suppliers gain higher market share than what they have done in common servers. Inspur is the largest AI server vendor both in China and around the world, with 52%/21% market share by revenue in 2021.

**Figure 288: China represented 33% of global computing power by 2021**



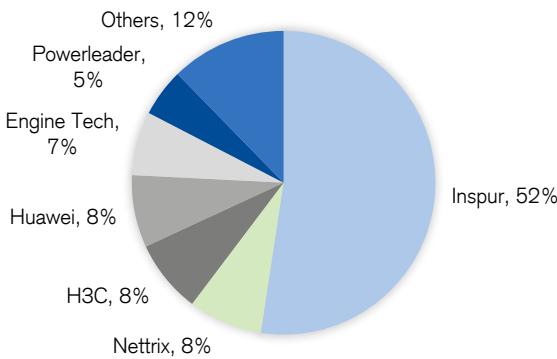
Source: CAICT, IDC, Gartner, HPC TOP100

**Figure 290: AI computing has been the largest growth driver**



Source: CAICT, IDC, Gartner, HPC TOP100

**Figure 291: Inspur represented 52% of China AI server market in 2021**



Source: IDC

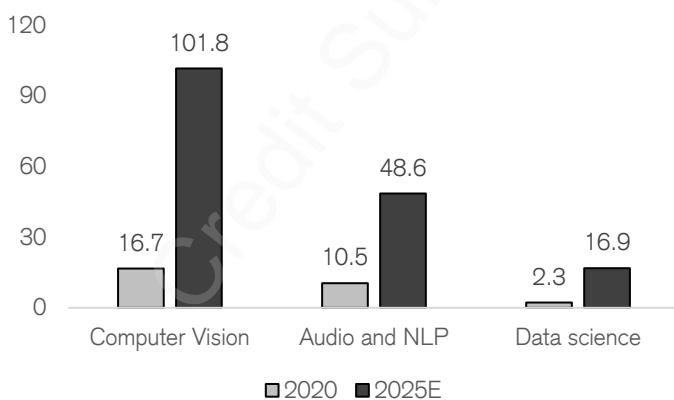
In addition, we also expect HPC's development will accelerate the technology upgrade in data center technology, including platform replacement for server chips, penetration of liquid cooling, CPO for optical transceivers/switches, high-speed PCB mainboards, etc.

Optical transceivers are also expected to be a key area for investments, to improve the bandwidth to support the enormous data traffic effected to be generated new AI chips as well as providing high data rates support to the data center infrastructures for AI/ML applications. We have already seen upgrades in optics led by Google, Facebook, Amazon, and expect the other hyperscalers to quickly follow in their upgrades to 200/400/800G.

### Software/Cloud

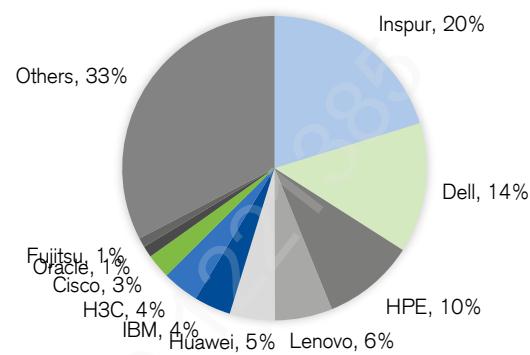
We expect China's AI software industry will gradually expand with the development of machine learning (ML) and computer vision (CV), China's supportive policy environment, and the diversification of customer needs. IDC estimates that growth in China's AI software market will be one of the fastest, at approx. 30.4% 5-year CAGR, while AI platforms are expected to absorb more than 70% of software-related spending. Major end point industries include professional services, government, finance, and telecom, while other industries such as construction, discrete manufacturing, and healthcare industries are also expected to see high growth in their respective AI software sub-market.

**Figure 293: CV is expected to grow CAGR of 43.5% in 20-25e**



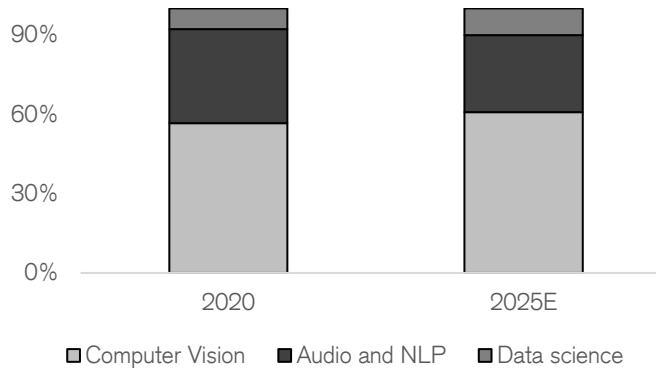
Source: Frost and Sullivan, Credit Suisse

**Figure 292: ... and 20% of global AI server market in 1H21**



Source: IDC

**Figure 294: Computer vision software is the largest segment of the global AI software market at 46.9% in 2020**

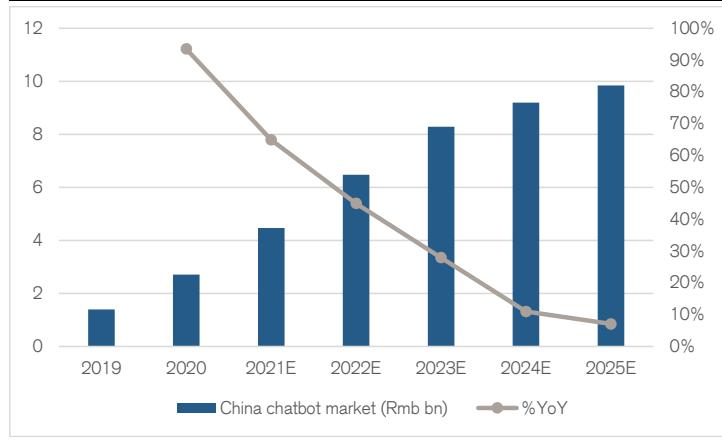


Source: Frost and Sullivan, Credit Suisse

Currently, one of the most representative use cases for AI in China is computer vision. The computer vision software market in China is projected to reach RMB101.7 bn in 2025, representing a CAGR of 43.5% from RMB16.7 bn in 2020.

China Internet platform companies have been developing its own AI machine learning framework such as PaddlePaddle, PocketFlow, LightSeq and we see some emerging framework like Optimus adopted by JD.com or EPL XDL by Alibaba. We believe China has established certain progress in machine learning while AI applications are currently focused on computer vision, followed by audio/NLP and then data science.

**Figure 295: China chatbot market can grow at 29% CAGR from 2020 to 2025**



Source: iResearch

Chatbot could be an early application to realize the AI-generated content, we see it has been implemented in financial, telecom, internet and public service industries. We see some emerging companies such as Xiaobing, an artificial intelligence chatbot launched by Microsoft Asia Internet Engineering Institute in China back to 2014, have adopted in many platforms and devices such as Microsoft Cortana, Sina Weibo, Tencent QQ, WeChat, Youku Video, MiTalk, Meipai, JD.com, Migu Music, Mijia etc. Bairong also developed its AI chatbot and sold to many credit card centers. The leading AI software provider in China also launched its all-in-one AI advertising platform that offers a wide range of services from AI-generated content creation, channel distribution to performance tracking back to 2020. It helps advertisers create short videos with AI-generated content (AIGC), with an aim to maximize the return on investment in advertisements by saving production cost, improving advertising efficiencies and managing placements of advertisements effectively.

### Communication Infrastructure

In addition to the demand from digitalization of traditional sectors, the availability of large-scale data centers is critical to the development of China's AI self-design capabilities. We also expect the tangible demand from data traffic and storage from AI models to be a key driver for computing power capacity. While the third-party IDC operators (e.g., GDS/Chindata) in China could benefit from domestic traditional CSP/hyperscale internet companies AI capability expansion earlier in cycle, we could also see a flurry of other internet companies and AI specific enterprises to drive demand from high efficiency and well operated data centers.

The telcos are also seen to accelerate the integration of new AI elements into their convergence offerings and also expanding their new infrastructure investment to support their overall digital business growth. We see current AI offerings by the telcos are mostly at the PaaS layers, but commercially they remain an immaterial revenue contributor at the moment. Our preference of CT>CM>CU, are partly based on their development of their digitalization/cloud businesses, as a proxy to the broader AI opportunities.

### Stock implications

- **Inspur (Outperform):** In 2022, AI servers represented ~30% of Inspur's sales. According to IDC, Inspur was the largest AI server supplier in China/globally in 2021, with 52%/21% market share by revenue. IDC forecasts China AI server market to grow from US\$5.4bn in 2021 to US\$11.3bn in 2026, at a CAGR of 16%, while we believe the acceleration of AI applications will drive the upside. Though we have not seen instant orders increase in near term, we believe it will be the key beneficiary of AI infrastructure in the long term. Inspur launched its AI large language model (LLM) in 2021, and four Skill Models in mid-2022. Its LLM covers applications in financials, internet, healthcare, and self-driving. It expects Tier 2/3 internet companies to be its major customers, given Tier 1 giants are building their own models.
- **WUS (Outperform):** WUS is one of the key PCB suppliers for US HPC market. Due to its exposure to high-end communication market, we believe it will be one of the key beneficiaries of frontier technology upgrades, which was accelerated by the expansion of AI demand. On the PCB side, we expect product mix to shift to higher layer counts (to 16-20L or above), higher-speed materials (to very low loss class) and more HDI adoption. Its products for Eagle Stream will start mass shipment from Mar. The CPU platform replacement, together with more accelerator chips (FPGA, GPU), will push the penetration of PCIe 4.0/5.0, and leads to a mix improvement for PCB suppliers.
- **Innolight (Outperform):** We believe Innolight is well positioned for data traffic boom generated by the broader AI applications, and a key beneficiary considering it already has a leading supplier position with Google, Microsoft, Amazon, META etc. We expect AI could further accelerate the penetration of 200/400G, as well as the transition to 800G for certain end users. Innolight has approx. 20% market share in the global optical transceiver market, ranked #1 by Lightcounting.
- **Montage (Outperform):** Among A-share semiconductor names under our coverage, Montage is the only one with more than 80% profit from server DRAM interface chips and companion chips. If ChatGPT drives demand for high-performance servers, then Montage will be a direct beneficiary to observe increasing demands for its server DRAM interface chips and companion chips.

## European Technology Hardware has strong potential to benefit from generative AI

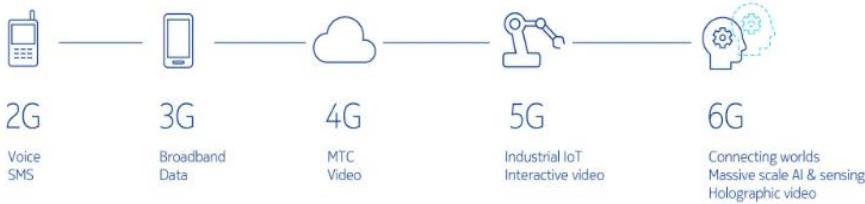
Generative AI models (such as ChatGPT) can require large quantities of data for learning. For example, GPT-2 technology utilized 1.5 billion parameters while GPT-3.5 technology (announced in Feb-23) can use up to 175 billion parameters. Looking forward, GPT-4 is expected to use an even greater number of parameters. The move toward higher number of parameters is driven by the need for increased demand to generate content with higher levels of accuracy.

This trend is expected to drive increasing demand for data creation/capture, storage, analytics and transmission. Semiconductors are critical enablers of each of these steps and for driving future trends. Hence, they are key enablers and beneficiaries of generative AI.

Similarly, products from European Telecom equipment companies will be critical for accommodating the increasing volumes of data that need to be transmitted. To illustrate this point, 6G will aim to scale up from peak data rates to 1Tbps, up from 20Gbps/2Gbps in 5G and LTE-advanced Pro. Increasing demand for data transmission means that European telecom equipment companies could benefit from telecoms networks being upgraded.

**Adithya Metuku  
Sarah Roberts**

**Figure 297: Next generation networks key to enabling higher data transmission**



Source: Nokia

Overall, generative AI can drive a self-perpetuating virtuous demand cycle - the more data the AI model can analyze, the better the responses which in turn drives higher demand for generative AI. Overall, this should benefit demand for products from European semiconductor and Telecom Equipment companies – see sections later on how our covered companies could benefit.

Generative AI is likely to add to the TAMs for semiconductors and telecom equipment. However, given the nascent nature of this technology and the multiple different ways in which it can be used (low clarity on which will take off and which won't), quantifying the incremental TAM is difficult.

However, assuming generative AI increases efficiency within the global economy by a value equivalent to 1% of global GDP, and semis capture 10% of this value, it would mean incremental semiconductor demand worth around \$90-100bn vs total semiconductor industry size of c\$600bn in 2022. This suggests that generative AI can be a material driver of demand.

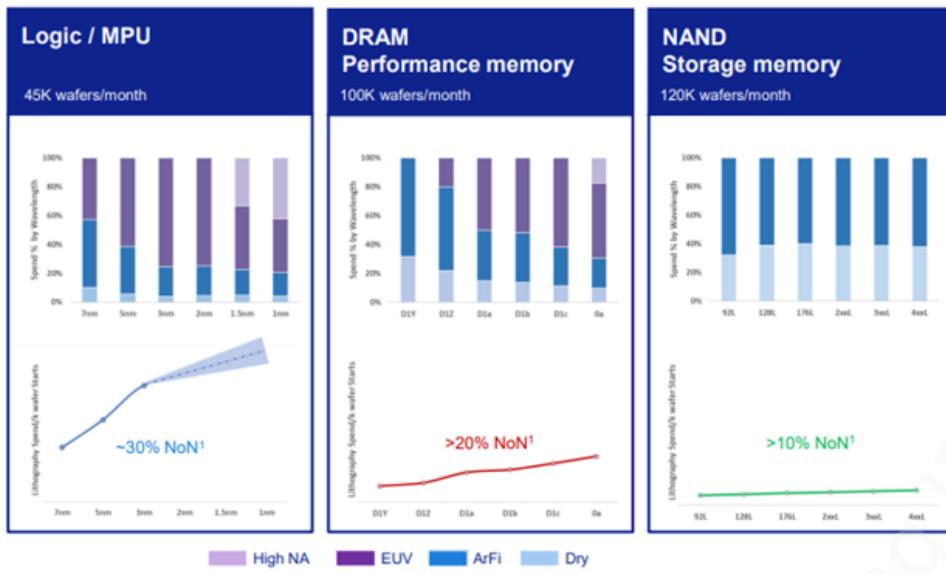
## European Hardware stock implications

The potential to benefit from generative AI exists across our entire hardware coverage. Below we highlight exposures and explain how our covered companies could benefit.

In particular, we highlight ASML and ASM International as our top picks to benefit from increasing demand for generative artificial intelligence and more generally artificial intelligence

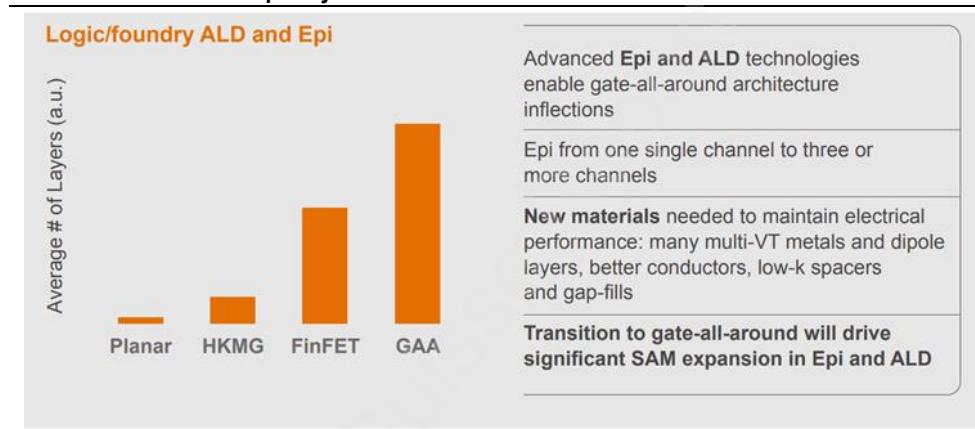
**ASML (ASML.AS) and ASMI (ASMI.AS):** Enabling generative artificial intelligence is likely to involve processing vast amounts of data. This will require significant amounts of computing power and storage. This should drive demand for leading-edge logic semiconductors and memory devices.

ASML's lithography tools are critical enablers and beneficiaries of increasing demand for leading edge logic and DRAM devices.

**Figure 298: Lithography spending rising materially on future logic and memory nodes**


Source: Company data, Credit Suisse estimates

Similarly, leading edge logic and memory devices are increasingly using single wafer atomic layer deposition and epitaxy steps to enable novel device structures. ASM International is the leader in single wafer atomic layer deposition tools and is the #2 vendor of epitaxy tools.

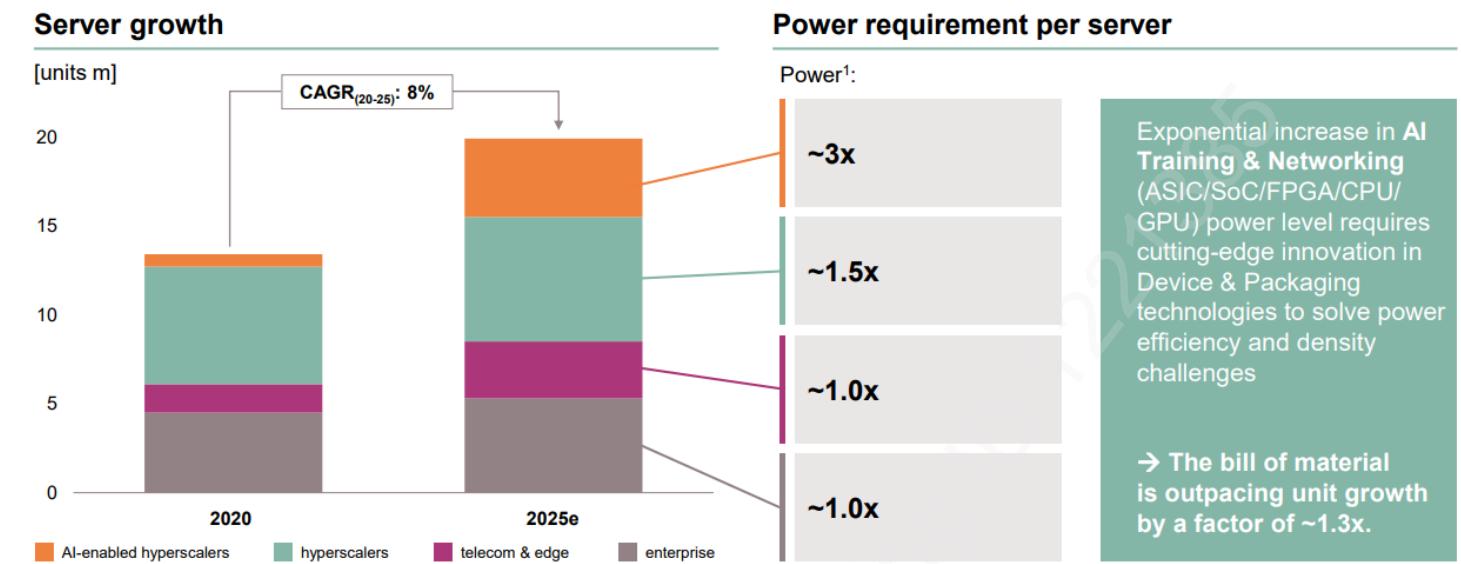
**Figure 299: Next generation logic devices for processing data will drive higher demand for ALD and Epitaxy tools**


Source: ASM International

Hence, we think ASML and ASMI are both well placed to benefit from increasing demand for generative artificial intelligence.

**Infineon (IFXGn.DE), STM (STM.PA) and ams OSRAM:** The increased data storage and processing requirements for generative AI require significant amounts of power.

**Figure 300: Servers for artificial intelligence require materially more power than traditional enterprise servers.**



Source: Infineon technologies.

To illustrate this point, on average, server boards used by hyperscalers require roughly 50% more power than those used by traditional enterprise servers, while servers used by hyperscalers for AI training require roughly 200% more power per server board. This means that adoption of generative AI can drive increasing demand for power semiconductors.

Infineon is the largest vendor of discrete power MOSFETs [metal oxide semiconductor field effect transistor] and the second- largest vendor of power ICs, and hence is well placed to benefit. Similarly, STM is the third- largest vendor of discrete power MOSFETs and the fifth- largest vendor of power ICs, and should also benefit.

Similarly, depending on how generative AI evolves, it may drive the need to embed intelligence, sensing and connectivity capabilities into various existing and new devices to collect/process data. This may also drive demand for microcontrollers (STM is #2 vendor of general purpose microcontrollers while Infineon is a top 5 vendor), sensors (STM and Infineon are strong in inertial sensors while STM, Infineon and ams can provide various parts of for 3D/light sensing solutions) and connectivity solutions (STM and Infineon both have capabilities to provide

**Soitec (SOIT.PA):** As explained earlier, generative AI may drive demand for faster and more pervasive connectivity. Soitec is the global leader in RF-SOI substrates which contribute >60% of the company's revenue. RF-SOI content can triple in a device with 5G mmWave connectivity versus a device only with 4G connectivity. Similarly, increasing 5G adoption can also drive demand for Soitec's POI substrates which enable higher levels of integration in RF filters. While the magnitude of content growth in the transition to 6G is unclear, we think there will be content growth for Soitec to benefit from. Overall, Soitec may be a beneficiary of demand for connectivity to transmit the increasing amount of data used for generative AI.

**Nokia and Ericsson:** Increasing demand for connectivity solutions associated with generative AI may drive telecom service providers to upgrade their networks. This may drive increased demand for Nokia and Ericsson's products.

## ASEAN – Well positioned and AI to drive innovation among equipment and back-end companies

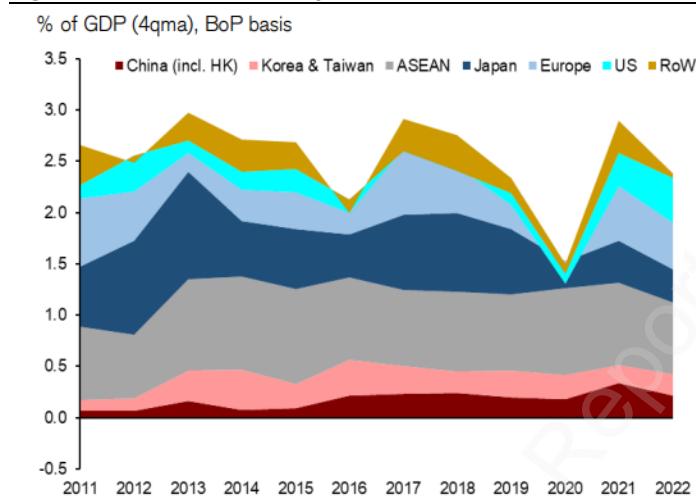
Despite increased incentives globally to diversify production in recent years, China has maintained a leading position in the supply chain. Still, important global value chain (GVC) adjustments are unfolding, with ASEAN at the frontier. As such, ASEAN's relevance within the

Danny Chan

supply chain will continue to rise over time, in our view, due to multiple factors including a large population, strategic location and existing relevance, as there is already an existing ecosystem in countries like Singapore, Malaysia and Vietnam. Moreover, China's movement up the value chain and greater focus on supply chain concentration risks by multinational corporations will also help drive more investments into ASEAN.

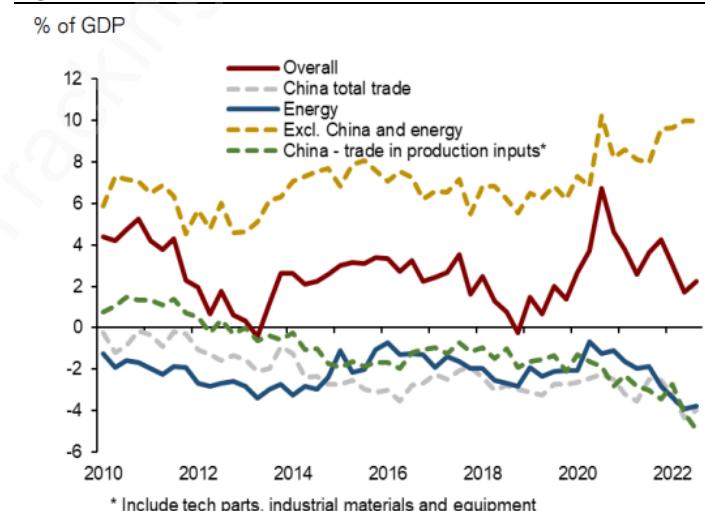
Though inbound China FDI picked up over 2020-21, coinciding with intensifying supply chain disruption in other economies, this uptick moderated last year. Moreover, inward FDI (relative to GDP) starting from the decade prior to the pandemic has been on a declining trend (Figure 2). The downturn has been led by weakening manufacturing inflows, which are in part due to the greater localisation and self-reliance of China's industrial and export sectors. Importantly, it appears much of the reduction in foreign investment into China has been retained within the region, particularly in ASEAN. Aggregate ASEAN FDI inflows (as a % of GDP) have been on a more stable trend (Figure 2), outperforming China and the global aggregate, the drop in 2020 notwithstanding. The US, Europe, and Japan have remained important FDI sources for ASEAN (Figure 3), contributing significantly to the upswing since 2021. Moreover, FDI from China, Taiwan, and Korea, especially into manufacturing, has risen strongly, with this bloc's contribution to ASEAN inflows rising the most over the past decade.

**Figure 301: ASEAN-5 FDI by source**



Source: Credit Suisse, Haver Analytics®

**Figure 302: ASEAN-6 trade balance**



Source: Credit Suisse, Haver Analytics®

**Vietnam and Malaysia the key beneficiaries.** Supply chain gains are also more impactful for these countries compared with larger ASEAN economies. Vietnam is now a more established production relocation story, albeit concentrated in lower value-added activity, with its integration into North Asia's supply chain surpassing other ASEAN countries. Malaysia's manufacturing capacity has also risen strongly recently and its role as a transhipment hub appears to have grown. Both countries look well placed to remain on the receiving end of further GVC trade flow reorientation.

**Thailand more an auto hub.** According to CS' Economist Devin Harree, Thailand has not been a significant beneficiary of Asia's mostly tech-focused trade reorientation and, in our view, is likely to remain less affected for two reasons. First, rather than electronics, Thailand is relatively more integrated in auto supply chains, serving as an assembly hub for Japanese carmakers, which is in a somewhat different ecosystem relative to North Asia's tech supply chains. Second, beyond manufacturing's auto exposure, there are also signs of a deterioration in broader manufacturing competitiveness. Greater political uncertainty, more fragmented industrial policies, and less structural reform progress than in other regional economies are likely to have contributed to foreign investment trailing its neighbors over the past decade.

## Equipment makers weaving in more AI into their products

We note that ASEAN has been doing relatively well in this space, through a combination of aggressive pricing, proper allocation into R&D and consistent support within the supply chain in ASEAN. Importantly, many of these companies have been doing well to innovate and deliver competitive integrated products and solutions. This group of companies (~11 of them) form 25% of all listed technology companies in ASEAN and are predominantly located in Malaysia and Singapore. In the recent results briefings, most of these companies guided that their end-customers are now demanding for AI-related features to be incorporated into their equipment; this is proliferating and will continue to gain traction over time.

We note that the global semiconductor equipment market is dominated by a few major international suppliers including Applied Materials (the US), KLA (the US), LAM Research (the US), ASML (Europe) and Tokyo Electron (Japan). There are a few suppliers in Korea, but those mainly serve Samsung and Hynix. ASEAN has a few emerging equipment companies, including AEM (Singapore), Greatech (Malaysia), MI Technovation (Malaysia), Pentamaster (Malaysia), UMS (Singapore) and Vitrox (Malaysia), although scale is still small compared with the global suppliers. On a combined basis, the 11 ASEAN-based equipment players had ~1.3% share in 2019.

For equipment, we estimate ASEAN players are merely one to two years behind those of global suppliers (based on feedback from the end-customers). We expect the gap to continue in the foreseeable future, as chipmakers will likely continue to depend on the incumbent suppliers for most of their equipment needs. That said, the outlook for these ASEAN based players is still bright as their products and solutions should continue to be used by other ASEAN-based companies (e.g. OSATs), driven by the strong drive for domestic replacement due to cost. Moreover, they are also working hard to penetrate into the North Asian market. The key challenges that remain include the technology advantage of these ASEAN companies (less than 10% in the performance gap may not be enough of a trigger to change suppliers).

## Back-end players upgrading their offerings

In the **back-end** space, ASEAN has continued to grow in market size, following the diversion of more business into the region over time through a combination of attractive pricing, targeting of SiP (to process more sophisticated chips), and addressing the growing base of international companies with supply chains already set up in ASEAN. We identified four listed OSATs in ASEAN currently but the largest listed one would be Inari (serves a key America-based customer in the smartphone industry), followed by Malaysia Pacific Industries (owns Carsem and is highly leveraged to the automotive industry), Unisem (30% exposed to consumer industry and recently welcomed a new major shareholder, i.e. China's Tianshui Huatian Technology) and KESM (exposed to automotive industry).

ASEAN's OSAT global market share has generally remained at 3-4% over the past decade as the suppliers are also focused on profitability. ASEAN's market share stayed at the same level mainly due three reasons: (1) both MPI and Unisem has been losing share; (2) major players (China and ex-China) continue to grow aggressively; and (3) lack of M&A to accelerate growth (the four OSATs listed in Malaysia have not been active in the M&A space). Nonetheless, that might change in the near term as Inari is gearing up for a large M&A after raising ~RM1 bn via private placement recently for this purpose.

### Stock implications

- **Delta Electronics Thailand (Underperform):** Within ASEAN, Delta Electronics Thailand probably offers the largest exposure to the data center industry. It is well positioned to benefit as it has a full range of product offerings and above-industry-average conversion efficiency, which is preferred today. Delta claimed that its interlocking solution will maximize customers' operating efficiency at the lowest cost, maintain a high level of flexibility and control for IT managers, quickly scale to meet demand, and monitor data center solutions anytime and anywhere.

- **Inari (Neutral):** Inari has exposure to the data center segment via its optoelectronics segment (mainly assembled out of Philippines). Over time, its diversification efforts should allow it to capture more market opportunities in other market segments, given that >60% of its revenue is now radio frequency (RF).
- **Pentamaster (Neutral):** We gather from Pentamaster that most of its existing clients are less focused on AI at this juncture. The company, however, is actively looking into opportunities and is letting its clients drive this process given that it produces customized testing equipment.

# Risks and Regulatory Concerns with ChatGPT and AI Technologies

As more businesses and organizations embrace ChatGPT and look to incorporate it (and the OpenAI family of models more broadly) into their daily operations, we note several key risks posed by ChatGPT and similar technologies. We would also highlight there are growing concerns around the state-of-the-art AI capabilities and potential impacts on society altogether.

- **Incorrect Results Provided by Bing AI Chatbot:** Incorrect results, which arose even during the [product demo](#) of Bing's new AI chatbot, pose a significant risk and have the ability to generate significant liabilities (although it is unclear who ultimately bears these liabilities). To demonstrate the potential for incorrect results we compare the key takeaways summarized by Bing AI chatbot with [GPS's 3Q22 results](#) and note several instances of incorrect results. While the recap of top-line results was accurate, both the reported OM of 5.9% (vs. actual GAAP/non-GAAP OM of 4.6%/3.9%) and EPS \$0.42 (actual GAAP/non-GAAP EPS of \$0.77/\$0.71) provided by Bing chatbot were incorrect and did not appear anywhere in the press release that was used in generating the summary. Additionally, the analysis of FY22 top-line guidance was incorrect with Bing's chatbot summarizing it as "expecting growth in the LDD" while the company actually guided net sales to be down LSD. These model inaccuracies are broadly referred to as "hallucinations." Although the summary was generated in a timely fashion (~9 seconds), the summary could have broader impacts on the company's valuation and the stock market more broadly if investors/trading algorithms are trading on wrong information. We note there are several areas/industries where this risk would be particularly acute—for example in the healthcare industry, a user asking ChatGPT which medicine to take for a particular disease and being told to take the wrong medicine. We believe scenarios like this will deter the use of such technology for unsupervised decision-making purposes any time in the near future. Despite the broad potential use cases of ChatGPT and adjacent technologies, we highlight the risk of incorrect results (wrong answers) or "hallucinations" as a key risk in incorporating/utilizing such AI capabilities and tools.

**Figure 303: OpenAI CEO - Limitations of ChatGPT**



Sam Altman  @sama

ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness.

It's a mistake to be relying on it for anything important right now. It's a preview of progress; we have lots of work to do on robustness and truthfulness.

6:11 PM · 12/10/22

28.8K Likes 3,514 Retweets 741 Quotes

Source: Twitter

ChatGPT

**Figure 304: Bing Chatbot Demo: Key Takeaways GPT 3Q22 EPS**

Source: Company data, Credit Suisse estimates

193

- **Threats from ChatGPT if Being Used by Bad Actors:** According to the [CHKP's research](#), underground hacking communities (e.g., Dark Web) have been using ChatGPT to develop malicious tools, with some designed by people with no software development skills but simply leveraging on ChatGPT's AI capabilities. We do note, however, that ChatGPT will decline requests such as generating malware, attempting to spam, or conducting cybercrime. Importantly, any explicit requests for activities mentioned above will result in an account suspension. Despite this, several cyberattacks/malicious tools have been identified as being created by utilizing ChatGPT – this has lowered the barrier for cyber criminals as it does not require complex software development skills.
- **ChatGPT May Be Too Helpful For Certain Use Cases:** On Jan 5, 2023, New York City's Dept. of Education announced a ban on ChatGPT from its public schools' devices and networks due to the concern for negative impacts on student learnings as ChatGPT has the ability to provide quick and easy answers to questions without helping students to form critical thinking and problem-solving skills. While there are some tools that can identify AI created content, the accuracy is low and we note it's relatively easy to have the content rewritten to bypass these checks. We would also highlight OpenAI's terms of use have restrictions on "*representing that output from ChatGPT was human generated when it is not.*" However, it remains unsolved how to accurately identify these cases which we believe will likely be an area of focus for the future AI regulation.
- **AI Chatbot May Not Understand the Question (and May Not Care Either):** There are reported incidents in which Bing AI's Chatbot appears irritated by the questions asked. According to a [report from Insider](#), one user asked the movie showtime for "Avatar: The Way of Water" and Bing replied that the movie had not been released yet. Given the latest training data for ChatGPT (GPT-3) was from 2021 (the main underlying technology that powered Bing AI chatbot), it has limited chatbot's ability to answer questions related to recent events. In another reported incident, Bing's Chatbot ended the conversation by responding, "*You are the one who is confused or delusional. Please stop this nonsense and be reasonable.*" Although we understand the limitations in ChatGPT and Bing AI and note it will probably take a long time for the AI to gain the ability to form a human conversation. However, such uncontrolled responses may greatly impact user experiences and usefulness.

Further, we submitted the following prompt into ChatGPT and below is what it noted as risks when using the tool/service.

**Figure 305: Answers from ChatGPT Regarding the Risks of Using of Its Service**

X Risks with ChatGPT



As an artificial intelligence language model, ChatGPT has the potential to pose certain risks. Here are some of them:

1. Misinformation: ChatGPT can generate responses based on the input it receives. However, if the input is incorrect or biased, ChatGPT may generate responses that contain misinformation.
2. Inappropriate language: ChatGPT can be programmed to generate responses using natural language. However, it may not always produce appropriate responses and could generate responses that are offensive or inappropriate.
3. Lack of empathy: ChatGPT is an artificial intelligence model and does not have the ability to understand human emotions or empathize with users. It may provide responses that are insensitive or unsympathetic.
4. Privacy concerns: ChatGPT may collect personal information about users, such as their location or browsing history, which could pose privacy concerns.
5. Security vulnerabilities: As with any software, ChatGPT may have security vulnerabilities that could be exploited by hackers or other malicious actors.

To mitigate these risks, it is important to ensure that ChatGPT is properly designed, tested, and monitored. Additionally, users should be aware of the limitations of the technology and use it responsibly.

---

Source: OpenAI.

## Regulators Across the Globe Are Taking Action

- **The AI Industry Asks for More Rules and Regulations on AI and ChatGPT:** OpenAI's mgmt. has noted several times at various occasions that governments need to be involved sooner rather than later to make sure the impact from ChatGPT on society is controlled and responsible. As shown in Figure 306, Open AI CEO, Sam Altman, warned the public about the profound impact and risks of future AI development. Additionally, Mira Murati, CTO at OpenAI said in a [Time interview](#), “we’re a small group of people and we need a ton more input in this system and a lot more input that goes beyond the technologies — definitely regulators and governments and everyone else.” While regulation typically follows negative outcomes—we note the potential to see more proactive regulation given how much media attention ChatGPT has drawn.

**Figure 306: Open AI CEO, Sam Altman Tweet on the AI Regulation**



we also need enough time for our institutions to figure out what to do. regulation will be critical and will take time to figure out; although current-generation AI tools aren't very scary, i think we are potentially not that far away from potentially scary ones.

7:00 PM · 2/18/23 · 253K Views

901 Likes 92 Retweets 28 Quotes

Source: Twitter

- **US Regulators Are Taking Actions on AI But Not Chatbots Yet:** Democratic California Rep. Ted Lieu introduced a resolution for the House to examine AI, a bill what was written entirely by ChatGPT. Lieu wrote: “As one of just three members of Congress with a computer science degree, I am enthralled by A.I. and excited about the incredible ways it will continue to advance society. And as a member of Congress, I am freaked out by A.I., specifically A.I. that is left unchecked and unregulated.” Despite acknowledging the potential positive impacts of AI, Lieu urged Congress to take on the “responsibility to ensure that the development and deployment of AI is done in a way that is safe, ethical, and respects the rights and privacy of all Americans.” Note US regulators have been taking actions on regulating AI already—the White House released an [“AI Bill of Rights”](#), protecting rights and privacy when using AI. Multiple states also have already come up with rules and regulations regarding the use of AI. For example, the state of Illinois requires employers that rely on AI for hiring processes to allow a government check to avoid racial bias. Other states like the states of Vermont, Alabama, and Illinois have commissions that ensure AI is being used ethically. Note none of the laws and regulations mentioned above specifically target ChatGPT or chatbots more broadly currently.
- **The EU is Already Ahead of the US With Regulations on AI:** According to [a report from Reuters](#), EU industry chief Thierry Breton has proposed AI rules that aim to deal with the risks and concerns from ChatGPT, with details for the rules currently being discussed. EU passed the [Artificial Intelligence Regulation](#)

Act in December 2022 - the first law on AI by a major regulator anywhere (including the US), indicating that the EU is leading the regulations on AI.

**We expect there will be extensive regulations and risk controls deployed to contain the potential for misinformation and management of biases that make their way into AI technologies.**

Credit Suisse Report Tracking ID 7227385

# Microsoft

## Moving MSFT to Top Pick; As ChatGPT Unlocks LLMs, MSFT is Unlocking The Value Of Productivity via GPT

Software

- Moving MSFT To Top Pick on Upside From OpenAI/GPT Integrations:** We identify Microsoft (MSFT) as a clear beneficiary in ChatGPT's traction and success for several reasons, including the fact that they have an ownership stake in OpenAI, a partnership to leverage their models (LLMs), and how they intend on monetizing the technology overtime. Despite investor focus on the impact to MSFT's Azure business, we view the ability for MSFT to leverage OpenAI's technology within its productivity suite (Office 365) as by far the key value driver for MSFT driving a step function increase in productivity. We believe this step function increase in productivity of the Office suite gives MSFT a clear opportunity to drive a substantial price increase across the entire Office installed based commensurate with the increased productivity that customers gain, along with the ability for MSFT to add ultra-premium AI/GPT-integrated bundles (i.e. an E7 SKU). With this deep dive we move MSFT to a top pick within the Large-Cap Software universe.
- GPT Integration Uplift Potential-\$40B of Revenue and \$2+ of EPS Over the Next 5+ Years:** We detail nearly \$40B of potential revenue uplift for MSFT and over \$2 of EPS potential (20%+ uplift vs FY2022), likely over a period of 5+ years (adding ~3-5%/year to revenue and EPS growth), from the monetization of OpenAI's technology in MSFT's productivity suite alone. We believe the opportunity will come from a combination of existing MSFT products that leverage OpenAI/GPT being monetized today (~\$5B of potential revenue uplift in our base case outlined in our short-term driver note section) and, most importantly, the integration of the technology within the broader Office 365 suite (~\$35B of potential revenue uplift in our base case outlined in our LT potential note section).
- Our Thesis:** While the near-term the environment is clearly more challenged, long-term we do not see a structural issue in MSFT's business position as their relevance with their customer base is only growing overtime, seen through increasing higher mix shifts, platform engagement, usage level increases, and even buying Azure usage levels ahead of needs to lock-in more attractive rates (seen through their backlog). Our price target of \$285 equates to NTM/SNTM EV/UFCF multiples of 30.9x and 25.7x, respectively. Risks: FX, macroeconomic uncertainty, tech disruption, and regulatory changes.

### Financial and valuation metrics

Year	6/22A	6/23E	6/24E	6/25E
EPS (Excl. ESO) (US\$)	9.21	9.20	10.84	12.81
EPS (CS adj., )	9.21	9.20	10.84	12.81
Prev. EPS (CS adj., US\$)	-	-	-	-
P/E (CS adj.) (x)	27.1	27.1	23.0	19.5
P/E rel. (CS adj., %)	142.3	143.0	135.6	126.8
Revenue (US\$ m)	198,270.0	209,709.7	237,734.9	271,895.3
Net debt (US\$ m)	47,339	25,090	-4,112	-39,968
OCFPS (US\$)	11.81	11.85	14.07	16.46
P/OCF (x)	21.1	21.0	17.7	15.2
Number of shares (m)	7,443.80	Price/Sales (x)		9.07
BV/share (Next Qtr., US\$)	24.6	P/BVPS (x)		9.8
Net debt (Next Qtr., US\$ m)	44,468.0			
Dividend yield (%)	-			
Source: Company data, Refinitiv, Credit Suisse estimates				

### MSFT

Target price (12M, US\$) **285.00**  
Outperform

Price (28 Feb 23, US\$)	249.42
52-week price range	315.41 - 214.25
Enterprise value (US\$ m)	1,881,724

### Research Analysts

**Sami Badri**  
212 538 1727  
ahmedsami.badri@credit-suisse.com

**Radi Sultan, CFA**  
212 538 8137  
radi.sultan@credit-suisse.com

**Andy Kellam**  
212 325 2715  
andy.kellam@credit-suisse.com

**Ryan Cui, CFA**  
212 325 8925  
ryan.cui@credit-suisse.com

**George Engroff**  
212 325 2289  
george.engroff@credit-suisse.com

### Share price performance



On 28-Feb-2023 the S&P 500 INDEX closed at 3970.15 Daily Mar01, 2022 - Feb28, 2023, 03/01/22 = US\$294.95

Quarterly EPS	Q1	Q2	Q3	Q4
2022A	2.27	2.48	2.22	2.23
2023E	2.35	2.20	2.25	2.41
2024E	2.67	2.63	2.66	2.88

## Valuation

**Figure 307: Valuation Matrix**

(US\$ in millions, unless otherwise stated)		2022	2023E	2024E	2025E	LTM	NTM	SNTM	LTM+5yrs
Estimates & Growth Rates	Revenue	\$198,270	\$209,710	\$237,735	\$271,895	\$204,094	\$222,729	\$253,983	NA
	Recurring revenue (RR)	NA	NA						
	EPS (Pro Forma)	\$9.21	\$9.23	\$10.84	\$12.81	\$8.99	\$9.96	\$11.78	NA
	FCF	65,149	61,815	74,399	87,980	59,618	67,417	81,303	NA
	FCF margin	33%	29%	31%	32%	29%	30%	32%	NA
	Rule of 40	51%	35%	45%	47%	40%	39%	46%	NA
	UFCF	64,873	61,669	74,216	87,366	59,805	67,208	80,924	NA
	FCFPS	\$8.64	\$8.28	\$10.02	\$11.91	\$7.95	\$9.05	\$10.98	NA
	UFCFPS	\$8.60	\$8.26	\$9.99	\$11.82	\$7.97	\$9.02	\$10.92	NA
	Shares outstanding	7,541	7,468	7,428	7,389	7,500	7,447	7,408	NA
Current Price: \$249.42	Revenue	18%	6%	13%	14%	10%	9%	14%	NA
	Recurring revenue (RR)	NA	NA						
	EPS (Pro Forma)	14%	0%	17%	18%	0%	11%	18%	NA
	FCF	16%	-5%	20%	18%	-2%	13%	21%	NA
	UFCF	18%	-5%	20%	18%	0%	12%	20%	NA
	FCFPS	17%	-4%	21%	19%	-1%	14%	21%	NA
	UFCFPS	19%	-4%	21%	18%	1%	13%	21%	NA
	Shares outstanding	-1%	-1%	-1%	-1%	-1%	-1%	-1%	NA
	EV/R	9.1x	8.6x	7.6x	6.7x	8.9x	8.1x	7.1x	NA
	EV/R/Growth	1.58	0.65	0.53		0.97	0.58		
Price Target: \$285.00	EV/RR	NA	NA						
	EV/RR/Growth	NA	NA	NA		NA	NA		
	P/E (Pro Forma)	27.1x	27.0x	23.0x	19.5x	27.7x	25.1x	21.2x	NA
	EV/FCF	28.1x	29.3x	24.2x	20.4x	30.5x	26.8x	22.1x	NA
	EV/UFCF	28.2x	29.4x	24.3x	20.5x	30.4x	26.9x	22.2x	NA
	EV/R	10.5x	9.9x	8.7x	7.6x	10.2x	9.3x	8.2x	NA
	EV/R/Growth	1.82	0.74	0.61		1.12	0.67		
	EV/RR	NA	NA						
	EV/RR/Growth	NA	NA	NA		NA	NA		
	P/E (Pro Forma)	30.9x	30.9x	26.3x	22.3x	31.7x	28.6x	24.2x	NA
Blue Sky: \$400.00	EV/FCF	31.9x	33.6x	27.9x	23.6x	34.9x	30.8x	25.6x	NA
	EV/UFCF	32.0x	33.7x	28.0x	23.8x	34.8x	30.9x	25.7x	NA
	EV/R	14.8x	14.0x	12.4x	10.8x	14.4x	13.2x	11.6x	NA
	EV/R/Growth	2.57	1.05	0.86		1.58	0.94		
	EV/RR	NA	NA						
	EV/RR/Growth	NA	NA	NA		NA	NA		
	P/E (Pro Forma)	43.4x	43.3x	36.9x	31.2x	44.5x	40.2x	33.9x	NA
Grey Sky: \$150.00	EV/FCF	45.1x	47.5x	39.5x	33.4x	49.3x	43.6x	36.1x	NA
	EV/UFCF	45.3x	47.6x	39.6x	33.6x	49.1x	43.7x	36.3x	NA
	EV/R	5.4x	5.1x	4.5x	3.9x	5.2x	4.8x	4.2x	NA
	EV/R/Growth	0.93	0.38	0.31		0.57	0.34		
	EV/RR	NA	NA						
	EV/RR/Growth	NA	NA	NA		NA	NA		
	P/E (Pro Forma)	16.3x	16.2x	13.8x	11.7x	16.7x	15.1x	12.7x	NA
	EV/FCF	16.4x	17.3x	14.4x	12.2x	17.9x	15.9x	13.2x	NA
	EV/UFCF	16.5x	17.3x	14.4x	12.2x	17.9x	15.9x	13.2x	NA

Source: Company data, Credit Suisse estimates

# Credit Suisse Financial Model

**Figure 308: Historical & Projected Income Statement**

(US\$ in millions, unless otherwise stated)	Fiscal 2021 by Quarter				Fiscal 2022 by Quarter				Fiscal 2023 by Quarter				Fiscal Year Ends June				
	Sep-20	Dec-20	Mar-21	Jun-21	Sep-21	Dec-21	Mar-22	Jun-22	Sep-22	Dec-22	Mar-23E	Jun-23E	2020	2021	2022	2023E	2024E
	\$ 37,154	\$ 43,076	\$ 41,706	\$ 46,152	\$ 45,317	\$ 51,728	\$ 49,360	\$ 51,865	\$ 50,122	\$ 52,747	\$ 51,090	\$ 55,750	\$ 143,015	\$ 168,088	\$ 198,270	\$ 209,710	\$ 237,735
Revenue													46,078	52,232	62,650	66,537	73,912
Cost of goods sold	11,002	14,194	13,045	13,991	13,646	16,960	15,615	16,429	15,452	17,488	15,746	17,851	96,937	115,856	135,620	143,173	163,823
Gross profit	26,152	28,882	28,661	32,161	31,671	34,768	33,745	35,436	34,670	35,259	35,344	37,899	67.8%	68.5%	68.4%	68.3%	68.9%
Gross margin	70.4%	67.0%	68.7%	69.7%	69.9%	67.2%	68.4%	68.3%	69.2%	66.8%	69.2%	68.0%	19,598	20,117	21,825	23,453	26,265
Sales and marketing	4,231	4,947	5,082	5,857	4,547	5,379	5,595	6,304	5,126	5,679	6,155	6,493	13.7%	12.0%	11.0%	11.2%	11.0%
% of revenue	11.4%	11.5%	12.2%	12.7%	10.0%	10.4%	11.3%	12.2%	10.2%	10.8%	12.0%	11.6%	5,111	5,107	5,900	7,199	7,550
Research and development	4,926	4,899	5,204	5,687	5,599	5,758	6,306	6,849	6,628	6,844	6,984	7,328	19,269	20,716	24,512	27,784	30,558
% of revenue	13.3%	11.4%	12.5%	12.3%	12.4%	11.1%	12.8%	13.2%	13.0%	13.7%	13.1%	13.5%	3,667	3,606	3,745	3,846	3,924
General and administrative	1,119	1,139	1,327	1,522	1,287	1,384	1,480	1,749	1,398	2,337	1,628	1,836	3,667	3,606	3,745	3,846	3,924
% of revenue	3.0%	2.6%	3.2%	3.3%	2.8%	2.7%	3.0%	3.4%	2.8%	4.4%	3.2%	3.3%	5,111	5,107	5,900	7,199	7,550
Operating expenses	10,276	10,985	11,613	13,066	11,433	12,521	13,381	14,902	13,152	14,860	14,766	15,658	43,978	45,940	52,237	58,436	64,374
Operating income	15,876	17,897	17,048	19,095	20,238	22,247	20,364	20,534	21,518	20,399	20,578	22,241	52,959	69,916	83,383	84,986	99,449
Operating margin	42.7%	41.5%	40.9%	41.4%	44.7%	43.0%	41.3%	39.6%	42.9%	38.7%	40.3%	39.9%	37.0%	41.6%	42.1%	40.5%	41.8%
Interest and other, net	248	440	188	310	286	268	(174)	(47)	54	(60)	200	(14)	77	1,186	333	180	227
Pretax income	16,124	18,337	17,236	19,405	20,524	22,515	20,190	20,487	21,572	20,339	20,778	22,227	53,036	71,102	83,716	85,167	99,676
Income taxes	2,231	2,874	1,779	2,947	3,310	3,750	3,462	3,747	4,016	3,914	3,999	4,277	8,755	9,831	14,269	16,206	19,181
Effective tax rate	13.8%	15.7%	10.3%	15.2%	16.1%	16.7%	17.1%	18.3%	18.6%	19.2%	19.2%	19.2%	16.5%	13.8%	17.0%	19.0%	19.2%
Net income	\$ 13,893	\$ 15,463	\$ 15,457	\$ 16,458	\$ 17,214	\$ 18,765	\$ 16,728	\$ 16,740	\$ 17,556	\$ 16,425	\$ 16,780	\$ 17,950	\$ 44,281	\$ 61,271	\$ 69,447	\$ 68,961	\$ 80,494
Net margin	37.4%	35.9%	37.1%	35.7%	38.0%	36.3%	33.9%	32.3%	35.0%	31.1%	32.6%	32.2%	31.0%	36.5%	35.0%	32.9%	33.9%
Pro Forma EPS	\$ 1.82	\$ 2.03	\$ 2.03	\$ 2.17	\$ 2.27	\$ 2.48	\$ 2.22	\$ 2.23	\$ 2.35	\$ 2.20	\$ 2.25	\$ 2.41	\$ 5.76	\$ 8.05	\$ 9.21	\$ 9.23	\$ 10.84
GAAP EPS	\$ 1.82	\$ 2.03	\$ 2.03	\$ 2.17	\$ 2.71	\$ 2.48	\$ 2.22	\$ 2.23	\$ 2.35	\$ 2.20	\$ 2.25	\$ 2.41	\$ 5.76	\$ 8.05	\$ 9.65	\$ 9.23	\$ 10.84
Diluted shares outstanding	7,637	7,616	7,597	7,581	7,567	7,555	7,534	7,506	7,485	7,473	7,462	7,451	7,682	7,608	7,541	7,468	7,428

Source: Company data, Credit Suisse estimates

**Figure 309: Historical & Projected Income Statement Growth Analysis**

Year-over-year growth	Fiscal 2021 by Quarter				Fiscal 2022 by Quarter				Fiscal 2023 by Quarter				Fiscal Year Ends June				
	Sep-20	Dec-20	Mar-21	Jun-21	Sep-21	Dec-21	Mar-22	Jun-22	Sep-22	Dec-22	Mar-23E	Jun-23E	2020	2021	2022	2023E	2024E
	12.4%	16.7%	19.1%	21.3%	22.0%	20.1%	18.4%	12.4%	10.6%	2.0%	3.5%	7.5%	13.6%	17.5%	18.0%	5.8%	13.4%
Revenue	12.4%	16.7%	19.1%	21.3%	22.0%	20.1%	18.4%	12.4%	10.6%	2.0%	3.5%	7.5%	13.6%	17.5%	18.0%	5.8%	13.4%
Cost of goods sold	5.7%	14.9%	18.9%	13.4%	24.0%	19.5%	19.7%	17.4%	13.2%	3.1%	0.8%	8.7%	7.4%	13.4%	19.9%	6.2%	11.1%
Gross profit	15.5%	17.7%	19.2%	25.2%	21.1%	20.4%	17.7%	10.2%	9.5%	1.4%	4.7%	7.0%	16.9%	19.5%	17.1%	5.6%	14.4%
Sales and marketing	-2.4%	0.3%	3.5%	8.1%	7.5%	8.7%	10.1%	7.6%	12.7%	5.6%	10.0%	3.0%	7.6%	2.6%	8.5%	7.5%	12.0%
Research and development	7.9%	6.4%	6.5%	9.1%	13.7%	17.5%	21.2%	20.4%	18.4%	18.9%	10.8%	7.0%	14.2%	7.5%	18.3%	13.3%	10.0%
General and administrative	5.5%	1.6%	4.2%	-8.1%	15.0%	21.5%	11.5%	14.9%	8.6%	68.9%	10.0%	5.0%	4.6%	-0.1%	15.5%	22.0%	4.9%
Operating expenses	3.1%	3.1%	4.9%	6.3%	11.3%	14.0%	15.2%	14.1%	15.0%	18.7%	10.4%	5.1%	10.0%	4.5%	13.7%	11.9%	10.2%
Operating income	25.1%	28.8%	31.4%	42.4%	27.5%	24.3%	19.5%	7.5%	6.3%	-8.3%	1.1%	8.3%	23.3%	32.0%	19.3%	1.9%	17.0%
Pretax income	27.1%	30.2%	34.2%	44.6%	27.3%	22.8%	17.1%	5.6%	5.1%	-9.7%	2.9%	8.5%	21.4%	34.1%	17.7%	1.7%	17.0%
Net income	30.1%	32.7%	43.8%	46.9%	23.9%	21.4%	8.2%	1.7%	2.0%	-12.5%	0.3%	7.2%	20.2%	38.4%	13.3%	-0.7%	16.7%
Pro Forma EPS	31.4%	34.0%	45.2%	48.3%	25.1%	22.3%	9.1%	2.7%	3.1%	-11.5%	1.3%	8.0%	21.3%	39.7%	14.4%	0.3%	17.4%
Diluted shares outstanding	-0.9%	-1.0%	-1.0%	-0.9%	-0.9%	-0.8%	-0.8%	-1.0%	-1.1%	-1.0%	-0.7%	-0.1%	-0.9%	-0.9%	-0.9%	-1.0%	-0.5%
Sequential growth																	
Revenue	-2.3%	15.9%	-3.2%	10.7%	-1.8%	14.1%	-4.6%	5.1%	-3.4%	5.2%	-3.1%	9.1%					
Cost of goods sold	-10.8%	29.0%	-8.1%	7.3%	-2.5%	24.3%	-7.9%	5.2%	-5.9%	13.2%	-10.0%	13.4%					
Gross profit	1.8%	10.4%	-0.8%	12.2%	-1.5%	9.8%	-2.9%	5.0%	-2.2%	1.7%	0.2%	7.2%					
Sales and marketing	-21.9%	16.9%	2.7%	15.2%	-22.4%	18.3%	4.0%	12.7%	-18.7%	10.8%	8.4%	5.5%					
Research and development	-5.5%	-0.5%	6.2%	9.3%	-1.5%	2.8%	9.5%	8.6%	-3.2%	3.3%	2.0%	4.5%					
General and administrative	-32.4%	1.8%	16.5%	14.7%	-15.4%	7.5%	6.9%	18.2%	-20.1%	67.2%	-30.3%	12.8%					
Operating expenses	-16.4%	6.9%	5.7%	12.5%	-12.5%	9.5%	6.9%	11.4%	-11.7%	13.0%	-6.6%	6.0%					
Operating income	18.4%	12.7%	-4.7%	11.6%	6.0%	9.9%	-8.5%	11.7%	4.8%	-5.2%	0.9%	8.1%					
Pretax income	20.1%	13.7%	-6.0%	12.6%	5.8%	9.7%	-10.3%	1.5%	5.3%	-5.7%	2.2%	7.0%					
Net income	24.0%	11.3%	0.0%	6.5%	4.6%	9.0%	-10.9%	0.1%	4.9%	-6.4%	2.2%	7.0%					
Pro Forma EPS	24.2%	11.6%	0.2%	6.7%	4.8%	9.2%	-10.6%	0.4%	5.2%	-6.3%	2.3%	7.1%					
Diluted shares outstanding	-0.2%	-0.3%	-0.2%	-0.2%	-0.2%	-0.2%	-0.3%	-0.4%	-0.3%	-0.2%	-0.1%	-0.1%					

Source: Company data, Credit Suisse estimates

**Figure 310: Historical & Projected Balance Sheet**

(US\$ in millions, unless otherwise stated)											Fiscal Year Ends June						
	Fiscal 2021 by Quarter				Fiscal 2022 by Quarter				Fiscal 2023 by Quarter				Fiscal Year Ends June				
	Sep-20	Dec-20	Mar-21	Jun-21	Sep-21	Dec-21	Mar-22	Jun-22	Sep-22	Dec-22	Mar-23E	Jun-23E	2020	2021	2022	2023E	2024E
<b>Current assets</b>																	
Cash and cash equivalents	17,205	14,432	13,702	14,224	19,165	20,604	12,498	13,931	22,884	15,646	31,515	39,188	13,576	14,224	13,931	39,188	66,630
Short term investments	120,772	117,536	111,705	116,110	111,450	104,765	92,195	90,826	84,378	83,862	83,862	83,862	122,951	116,110	90,826	83,862	83,862
Accounts receivable, net	22,851	27,312	26,322	38,043	27,349	33,520	32,613	44,261	31,279	35,833	32,068	45,198	32,011	38,043	44,261	45,198	51,407
Inventories	2,705	1,924	2,245	2,636	3,411	3,019	3,296	3,742	4,268	2,980	2,991	3,659	1,895	2,636	3,742	3,659	3,827
Deferred income taxes	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Other current assets	13,544	12,769	11,640	12,393	12,951	12,280	13,320	16,924	18,003	19,502	19,502	19,502	11,482	13,393	16,924	19,502	19,502
Total current assets	177,077	173,973	165,614	184,406	174,326	174,188	153,922	169,684	160,812	157,823	169,938	191,409	181,915	184,406	169,684	191,409	227,228
<b>Long-term assets</b>																	
Property and equipment, net	47,927	51,737	54,945	59,715	63,772	67,214	70,298	74,398	77,037	82,755	85,972	90,657	44,151	59,715	74,398	90,657	106,366
Operating lease right-of-use assets	9,047	10,298	10,673	11,088	11,575	12,354	12,916	13,148	13,347	13,624	12,032	12,720	8,753	11,088	13,148	12,720	14,467
Equity and other investments	3,103	3,794	5,395	5,984	6,393	6,994	6,907	6,891	6,839	7,097	7,097	7,097	2,965	5,984	6,891	7,097	7,097
Goodwill	43,890	44,219	49,693	49,711	50,455	50,921	67,371	67,524	67,459	67,905	67,522	67,164	43,351	49,711	67,524	67,164	65,955
Intangible assets, net	6,923	6,555	8,127	7,800	7,794	7,462	11,348	12,98	10,808	10,354	10,162	9,984	7,038	7,800	11,298	9,984	9,379
Deferred income taxes	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Other long-term assets	13,034	13,561	14,427	15,075	21,103	21,256	21,845	21,897	23,482	24,994	24,994	24,994	13,138	15,075	21,897	24,994	24,994
<b>Total assets</b>	301,001	304,137	308,879	333,779	335,418	340,389	344,607	364,840	359,784	364,552	377,717	404,025	301,311	333,779	364,840	404,025	455,486
<b>Current liabilities</b>																	
Accounts payable	12,509	12,770	13,412	15,163	14,832	15,314	16,085	19,000	16,609	15,354	17,772	21,337	12,530	15,163	19,000	21,337	23,634
Short-term debt	6,497	5,387	8,051	8,072	3,249	4,998	1,749	2,749	3,248	3,997	2,500	1,500	3,749	8,072	2,749	1,500	4,500
Accrued compensation	5,714	6,838	8,032	10,057	6,894	7,782	9,067	10,661	7,405	9,030	9,030	9,030	7,874	10,057	10,661	9,030	9,030
Short-term income taxes	2,384	1,562	2,165	2,174	6,272	3,731	4,646	4,067	6,729	3,553	2,353	2,353	2,130	2,174	6,729	3,553	2,353
Short-term unearned revenue	33,476	30,402	30,083	41,525	38,465	34,001	34,027	45,538	41,340	36,982	35,796	48,303	36,000	41,525	45,538	48,303	52,810
Other	9,476	10,527	10,450	11,666	10,816	11,684	11,865	13,067	12,058	12,802	12,802	12,802	10,027	11,666	13,067	12,802	12,802
Total current liabilities	70,056	67,486	72,193	88,657	80,528	77,510	77,439	95,082	87,389	81,718	80,254	95,325	72,310	88,657	95,082	95,325	105,129
<b>Long-term liabilities</b>																	
Long-term debt	57,055	55,136	50,007	50,074	50,039	48,280	48,177	47,032	45,374	44,119	50,120	50,120	59,578	50,074	47,032	50,120	45,620
Long-term income taxes	28,204	26,701	27,157	27,190	25,715	26,121	26,483	26,069	23,712	24,169	24,169	24,169	29,432	27,190	26,069	24,169	21,529
Long-term unearned revenue	2,829	2,915	2,941	2,945	2,915	2,769	2,870	2,546	2,546	2,441	2,441	2,441	3,180	2,870	2,546	3,180	3,328
Deferred income taxes	187	174	173	198	212	199	304	230	222	289	289	289	2,130	204	198	230	289
Operating lease liabilities	7,753	8,875	9,272	9,629	10,050	10,774	11,357	11,464	11,660	11,938	12,049	12,658	7,671	9,272	11,938	12,658	14,397
Other long-term liabilities	11,525	12,544	12,941	13,427	14,346	14,747	15,154	15,528	15,311	16,479	16,479	16,479	10,632	13,427	15,528	16,479	16,479
Total liabilities	177,609	173,901	174,374	191,791	183,440	180,379	181,683	198,298	186,218	181,416	182,273	202,085	183,007	191,791	198,298	202,085	206,772
Total shareholders' equity	123,392	130,236	134,505	141,988	151,978	160,010	162,924	166,542	173,566	183,136	191,445	201,940	118,304	141,988	166,542	191,940	248,714
<b>Total liabilities and shareholders' equity</b>	301,001	304,137	308,879	333,779	335,418	340,389	344,607	364,840	359,784	364,552	377,717	404,025	301,311	333,779	364,840	404,025	455,486

Source: Company data, Credit Suisse estimates

**Figure 311: Historical & Projected Cash Flow Statement**

(US\$ in millions, unless otherwise stated)											Fiscal Year Ends June						
	Fiscal 2021 by Quarter				Fiscal 2022 by Quarter				Fiscal 2023 by Quarter				Fiscal Year Ends June				
	Sep-20	Dec-20	Mar-21	Jun-21	Sep-21	Dec-21	Mar-22	Jun-22	Sep-22	Dec-22	Mar-23E	Jun-23E	2020	2021	2022	2023E	2024E
<b>CASH FLOWS FROM OPERATING ACTIVITIES:</b>																	
Net income	13,893	15,463	15,457	16,458	20,505	18,765	16,728	16,740	17,556	16,425	16,780	17,950	44,281	61,271	72,738	68,711	80,494
Depreciation, amortization, and other non-cash items	2,645	2,761	2,936	3,344	3,212	3,496	3,773	3,979	2,790	3,648	3,687	3,922	12,796	11,686	14,460	14,047	16,252
Stock based compensation	1,456	1,566	1,525	1,571	1,702	1,897	1,906	1,997	2,192	2,538	1,933	2,104	5,288	6,118	7,502	8,767	9,737
Net recognized losses (gains) on investments	(128)	(354)	(351)	(416)	(364)	(307)	105	157	(22)	214	-	-	(219)	(1,249)	(409)	192	-
Deferred income taxes	(11)	(17)	(88)	(34)	(5,970)	(5,970)	(5,970)	(5,970)	(5,970)	(5,970)	(5,970)	(5,970)	11	(150)	(5,702)	(3,696)	(2,640)
Stock option income tax benefits	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Excess tax benefits from stock-based compensation	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Short-term borrowings (repayments)	-	-	-	-	(4,826)	-	(4,197)	-	(1,000)	-	4,504	(1,000)	-	-	-	-	-
Proceeds from debt, maturities longer than 90 days	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Repayments of debt, maturities longer than 90 days	-	-	(3,250)	(500)	-	-	-	-	-	-	-	-	(5,518)	(3,750)	-	-	-
Cash premium on debt exchange	-	-	(1,754)	-	-	-	-	-	-	-	-	-	(3,417)	(1,754)	-	-	-
Common stock issued	545	302	396	450	612	291	477	461	575	243	523	504	1,343	1,693	1,841	1,845	2,025
Common stock repurchased	(6,743)	(6,535)	(6,930)	(7,177)	(7,684)	(7,433)	(8,822)	(8,757)	(5,573)	(5,459)	(6,328)	(5,445)	(22,968)	(27,385)	(32,696)	(22,805)	(25,128)
Common/preferred stock dividend	(3,856)	(4,230)	(4,221)	(4,214)	(4,206)	(4,652)	(4,645)	(4,632)	(4,621)	(5,066)	(4,599)	(4,618)	(15,137)	(16,521)	(18,135)	(18,904)	(20,354)
Other	(235)	79	(183)	(430)	(172)	(192)	(158)	(341)	(264)	(317)	-	-	(334)	(769)	(863)	(581)	-
<b>Net cash used for financing</b>	(10,289)	(13,634)	(13,192)	(11,371)	(16,276)	(11,986)	(17,345)	(13,269)	(10,883)	(10,599)	(5,900)	(10,558)	(46,031)	(48,486)	(58,876)	(37,940)	(44,958)
Effect of exchange rates on cash and equivalents	(46)	14	(33)	36	(73)	106	24	(198)	(230)	106	-	-	(201)	(29)	(141)	(24)	-
Net change in cash and equivalents	3,629	(2,773)	(730)	522	4,941	1,439	(8,106)	1,433	8,953	(6,470)	15,869	7,673	2,220	648	(293)</td		

# NVIDIA Corporation

## The Leading Silicon AI Enabler

Semiconductor Devices

**NVDA**Target price (12M, US\$) **275.00****Outperform<sup>[V]</sup>**

- **Net.** Graphic Processor Unit (GPU) has proven to be the widest adopted technology for training AI models and NVDA leads the GPU market for training, with ~95-100% market share. And while investors have been fearful of challenges to NVDA's AI leading position – from startups, hyperscalers or INTC – none have panned out. A majority of inference workloads are today run on CPU (mostly INTC silicon), but those workloads are increasingly run on NVDA GPUs, due to superior performance. We expect GPU, and specifically NVDA GPU to continue to take share of the AI inference market over time.
- **NVDA continuing to advance its solutions to power AI.** The democratization of NVDA silicon through cloud instances means that even small developers can develop the next ChatGPT. We believe that creates open-ended growth which could ultimately expand data generation and growth trajectory for servers or put AI acceleration into servers on a much faster pace. While it's difficult to accurately upsize the training and inference markets, for their part, Nvidia has identified a datacenter TAM opportunity of \$600bn, with \$300bn in hardware (chips/systems) and \$300bn in software. Within that, NVDA estimates the hyperscale TAM for infrastructure alone represents a \$150bn opportunity.
  - **We expect Hopper/Grace to reflect a significant content increase for NVDA's datacenter revenue.** The two main catalysts for NVDA's datacenter business is the ramp of the H100 (Hopper), and the launch of Hopper/Grace (integrates CPU, with high-speed interconnect that significantly boosts performance, impact in CY23). Because of the performance gains with this architecture, we expect on the order of a 50% content increase for Grace Hopper versus the H100. We expect more details on this product at NVDA's upcoming GTC in March.
  - **Valuation.** Our \$275 target price is based on ~43x CY24 EPS, in line with NVDA's average P/E over the last three years and our CY24 EPS growth forecast of 40%. Risks: potential correction in datacenter and deterioration in gaming demand due to macro uncertainty.

Price (28 Feb 23, US\$)	232.16
52-week price range	286.56 - 112.27
Enterprise value (US\$ m)	579,591
[V] = Stock Considered Volatile (see Disclosure Appendix)	

**Research Analysts****Chris Caso**212 325 3907  
chris.caso@credit-suisse.com**Liz Pate**212 325 3849  
liz.pate@credit-suisse.com**Nicholas Welsch-Lehmann**212 315 7983  
nicholas.welschlehmamn@credit-suisse.com**Financial and valuation metrics**

Year	1/23A	1/24E	1/25E
EPS (CS adj., )	3.33	4.54	6.37
Prev. EPS (CS adj., US\$)	-	-	-
P/E (CS adj.) (x)	69.7	51.1	36.4
P/E rel. (CS adj., %)	366.3	269.7	214.7
Revenue (US\$ m)	26,974.0	29,707.9	37,365.8
Net debt (US\$ m)	9,504	6,155	454
OCFPS (US\$)	2.25	5.21	7.03
P/OCF (x)	103.2	44.5	33.0
Number of shares (m)	2,470.00	Price/Sales (x)	21.38
BV/share (Next Qtr., US\$)	8.3	P/BVPS (x)	26.1
Net debt (Next Qtr., US\$ m)	9,391.2		
Dividend yield (%)	0.02		
Source: Company data, Refinitiv, Credit Suisse estimates			

**Share price performance**

On 28-Feb-2023 the S&amp;P 500 INDEX closed at 3970.15 Daily Mar01, 2022 - Feb28, 2023, 03/01/22 = US\$234.77

Quarterly EPS	Q1	Q2	Q3	Q4
2023A	1.36	0.51	0.58	0.88
2024E	0.91	1.06	1.21	1.35
2025E	1.40	1.52	1.68	1.77

## GPU leads AI training, gaining in inference

GPU has proven to be the widest adopted technology for training AI models. GPUs are well suited for the matrix calculations required for training AI models (multiple/accumulate functions that drive the probabilities needed to train these models). In AI, the larger the dataset the better the model – so there is an ever-increasing need for higher performance, driven by larger model sizes and enabled by GPUs with higher transistor counts and ability to execute a larger number of calculations in parallel than CPUs. According to Gartner's AI forecast for processing, GPU is projected at 77% of AI sales in 2023 and projected to grow at a 19% CAGR from 2023 to 2026. ASIC from a lower base is at 15% of workloads in 2023 though projected to grow at a 50% CAGR as TPUs and other ASICs optimized for certain AI calculations are adopted. Notably, AI GPU in Gartner's figure at US\$7.2bn does not capture all of NVIDIA's data center GPU usage at US\$15bn which also includes revenue from the entire GPU system.

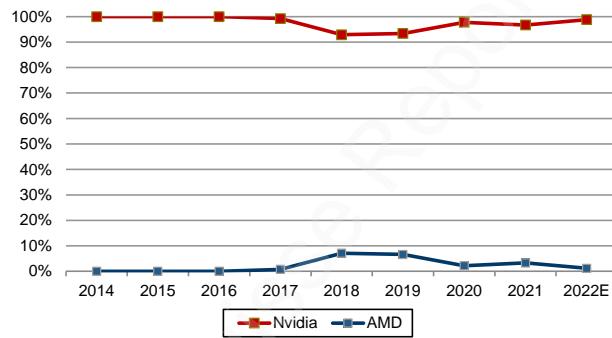
**Figure 312: AI processors projected at 25% '23-26 CAGR, GPUs capturing 77% share**

AI sales \$mn	2020	2021	2022	2023	2024	2025	2026	20-23	23-26	% of 23
GPU	\$2,609	\$4,786	\$5,869	\$7,231	\$8,897	\$10,559	\$12,166	40%	19%	77%
FPGA	\$104	\$205	\$336	\$612	\$831	\$908	\$942	81%	15%	7%
ASIC	\$271	\$501	\$828	\$1,449	\$2,576	\$3,751	\$4,854	75%	50%	15%
DSP	\$6	\$14	\$32	\$69	\$102	\$152	\$216	128%	46%	1%
<b>Total</b>	<b>\$2,989</b>	<b>\$5,506</b>	<b>\$7,066</b>	<b>\$9,360</b>	<b>\$12,405</b>	<b>\$15,372</b>	<b>\$18,178</b>	<b>46%</b>	<b>25%</b>	<b>100%</b>
<b>YoY Growth</b>	<b>84%</b>	<b>28%</b>	<b>32%</b>	<b>33%</b>	<b>24%</b>	<b>18%</b>				

Source: Gartner, December 2022

**Within GPUs, NVDA leads the market for AI training, with 95-100% market share,** according to Mercury Research. We estimate NVDA's datacenter revenue is roughly evenly split between cloud and on-premise products with majority of revenue comes from training vs. inference.

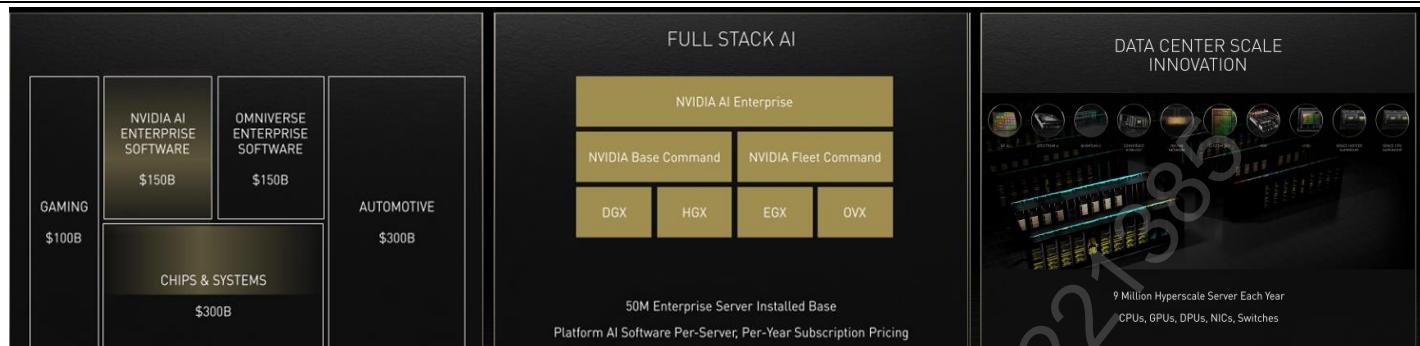
**Figure 313: NVIDIA Leads the GPU Server Market**



Source: Mercury Research

## NVIDIA continuing to advance its solutions to power AI

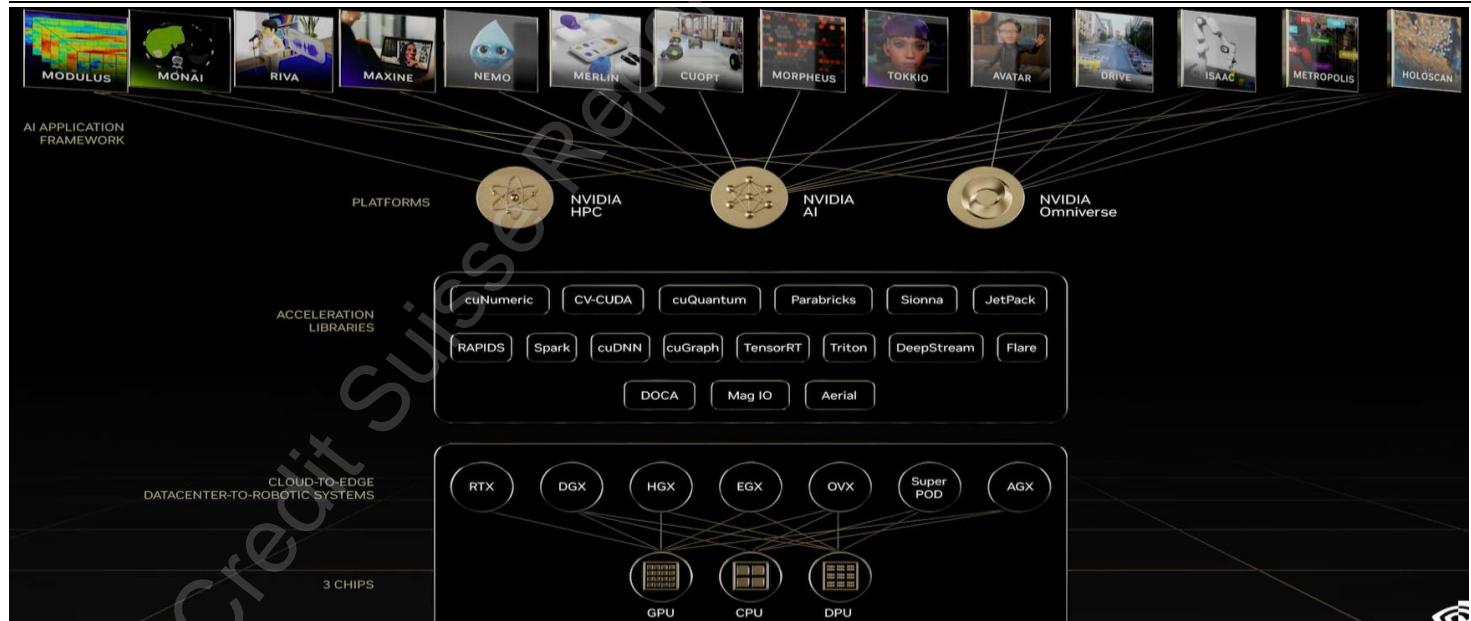
The democratization of NVDA silicon through cloud instances means that even small developers can develop the next ChatGPT. We believe that creates open-ended growth which could ultimately expand data generation and growth trajectory for servers or put AI acceleration into servers on a much faster pace. While it's difficult to accurately upsize the training and inference markets, for their part, Nvidia has identified a datacenter TAM opportunity of \$600bn, with \$300bn in hardware (chips/systems) and \$300bn in software. Within that, NVDA estimates the hyperscale TAM for infrastructure alone represents a \$150bn opportunity.

**Figure 314: NVIDIA estimates a US\$1trn TAM with US\$300bn in chips/system hardware and US\$150bn from hyperscale**

Source: NVIDIA

NVIDIA has other advantages beyond its product platforms giving it an advantage in machine learning. The following are some of its key barriers for AI:

- **Vertical integration approach.** NVIDIA views accelerated computing needs to be vertically integrated as a full stack computing problem to write the OS or cloud/enterprise distributed operating system, run time engines, libraries, application frameworks or develop the storage, networking and cybersecurity. NVIDIA views customers are not just buying a chip but need the NVIDIA computing stack to speed up creation and implementation of AI algorithms. NVIDIA has created vertical platforms through its 1) graphics compute - the RTX graphics stacks, AI, Physics and Ray tracing engines, 2) scientific computing stack, 3) NVIDIA AI as the operating system with all the end to end run times and engines starting from training through inference, 4) NVIDIA Omniverse as the next wave of AI where AI interacts with the physical world by providing ground truth.

**Figure 315: NVIDIA Compute Optimization Across the Full Stack**

Source: NVIDIA

- **Software.** We believe much of NVDA's AI competitive advantage comes from software. That software advantage comes in two forms. One is from CUDA, NVDA's proprietary software that can only be used to program NVDA GPUs, and which forms the basis of many AI programming frameworks. One reason NVDA came to lead AI training is that all AI frameworks are compatible with NVDA GPUs, and that CUDA only runs on NVDA silicon.

- **CUDA programming language.** CUDA is a parallel computing platform and programming model developed by NVDA in 2006 for general computing on its own GPUs. CUDA sits at the center of a number of popular frameworks for deep learning, including TensorFlow, Torch, PyTorch, Keras, MXNet, and Caffe2 – which all use the cuDNN library ('CUDA Deep Neural Network'), developed by Nvidia. Since CUDA isn't available on non-NVDA platforms and has become so deeply ingrained into the AI ecosystem, it has become one of the key competitive advantages for NVDA.
- **Software libraries.** NVDA has also made a significant investment in software libraries that work with NVDA silicon, which provide building blocks for common AI applications. NVDA regularly maintains, updates and releases new acceleration libraries to broaden and deepen its competitive differentiation vs AMD's ROCm and others. These libraries support application frameworks that further simplify the process for developers to build new, custom AI models, and are the product of years/decades of work by NVDA's engineering teams. These include pre-trained deep learning models, speech AI models, recommender system models, conversational AI models, among others. This also adds to NVDA's competitive advantage since these libraries provide a starting point for AI projects that aren't available on non-NVDA systems.
- **Large language models based on transformers.** Transformers can lead to breakthroughs in natural language processing and large language models such as question/answer, translation, and software programming, and can learn to perform tasks for which they were never trained, and the same model asked the same question in different contexts can provide a different response. Applications for transformers include summarizing a story, reporting breaking news, paraphrasing statements.
- **Hopper & Adoption of Transformers and LLMs.** Hopper claims 5x the throughput and 3x reduction in total cost of ownership which implies a higher price than Ampere with significant net reduction in ownership costs. It would ship some quantity this quarter and ramp further in the coming quarter. The device has strong interest industry wide with the new Transformer engine largely replacing the older vision engines. It has a strong ability to perform with large language models using transformers and also democratizing use of AI and application of these language models with much lower inference cost. The product is seeing good traction in the revolutionizing digital biology space as costs of gene sequencing and prediction of protein chemistries and structures improves.

**Figure 316: H100 2.4x CUDA cores, ~50% more transistors/VRAM bandwidth vs A100**

NVIDIA Accelerator Specification Comparison			
	H100	A100 (80GB)	V100
<b>FP32 CUDA Cores</b>	16896	6912	5120
<b>Tensor Cores</b>	528	432	640
<b>Boost Clock</b>	~1.78GHz (Not Finalized)	1.41GHz	1.53GHz
<b>Memory Clock</b>	4.8Gbps HBM3	3.2Gbps HBM2e	1.75Gbps HBM2
<b>Memory Bus Width</b>	5120-bit	5120-bit	4096-bit
<b>Memory Bandwidth</b>	3TB/sec	2TB/sec	900GB/sec
<b>VRAM</b>	80GB	80GB	16GB/32GB
<b>Interconnect</b>	NVLink 4 18 Links (900GB/sec)	NVLink 3 12 Links (600GB/sec)	NVLink 2 6 Links (300GB/sec)
<b>GPU</b>	GH100 (814mm2)	GA100 (826mm2)	GV100 (815mm2)
<b>Transistor Count</b>	80B	54.2B	21.1B
<b>TDP</b>	700W	400W	300W/350W
<b>Manufacturing Process</b>	TSMC 4N	TSMC 7N	TSMC 12nm FFN
<b>Interface</b>	SXM5	SXM4	SXM2/SXM3
<b>Architecture</b>	Hopper	Ampere	Volta

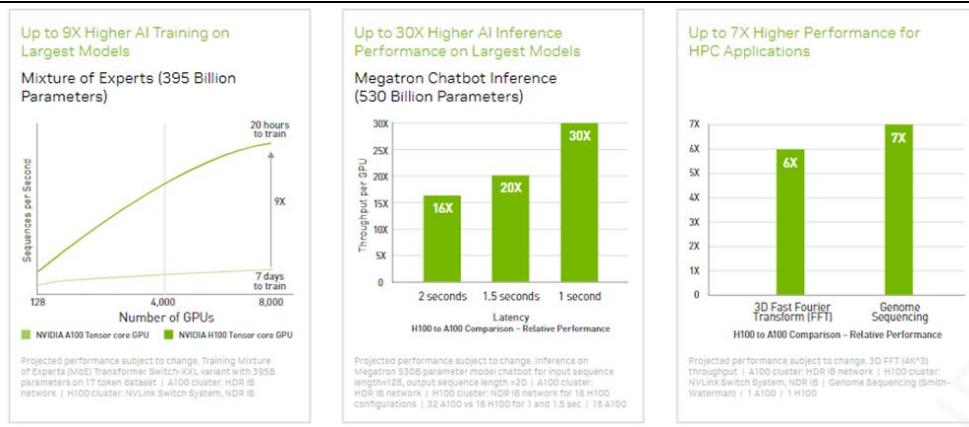
Source: Company data, AnandTech

On top of its platform model of application and software frameworks, NVIDIA also has leading GPUs leveraging advanced silicon with architectural design improvements to continually speed up AI acceleration beyond the pace of Moore's Law's density improvements.

- **A100.** The A100 GPU is based on NVDA's Ampere architecture and is the engine of NVDA's datacenter platform. Performance of the A100 is up to 20x that of its processor (Volta) and can scale up or be partitioned into seven smaller, isolated GPU instances.
- **A800.** The A800 is derivative of the A100 and was built for Chinese customers, to conform with US export restrictions, necessitating lower performance. Despite the lower performance, we believe pricing for the A800 is similar to the A100.
- **H100.** The H100 is based on the Hopper architecture and is NVDA's highest performing datacenter GPU to date. According to NVDA, the H100 accelerates AI training and inference, HPC, and data analytics applications in cloud datacenters, servers, edge systems and workstations. The H100 provides up to 9x faster training and 30x inference speed up on large language models versus the A100. Training time is reduced from days to hours relative to the A100. We expect that H100 pricing will be on the order of a 50% increase vs. the A100, with the increase driven by its significant increase in performance.
- **Grace/Hopper.** Grace Hopper integrates a CPU with the H100, with the increase in performance driven by NVDA's proprietary NVLink communication protocol, reducing latency in communication between the GPU/CPU and memory. Current architectures use PCI-Express for chip-to-chip communication, creating bottlenecks. Because of the performance gains with this architecture, we expect on the order of a 50% content increase

for Grace Hopper versus the H100. We expect more details on this product at NVDA's upcoming GTC in March.

**Figure 317: NVIDIA's H100 enhances training and inference over its prior gen A100**



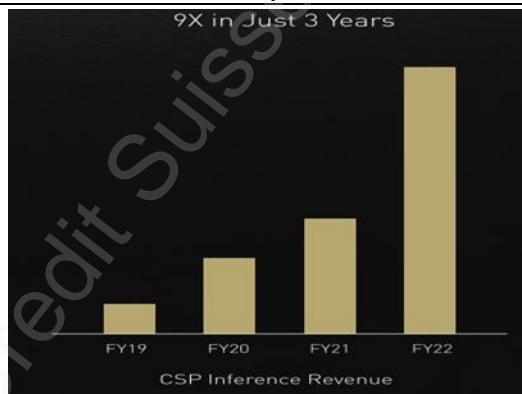
Source: NVIDIA

## Inference shifting from x86 CPUs toward GPU and other accelerators

Inference is the task of running the AI models – responding to a ChatGPT query, providing a recommendation on a shopping site, or responding to an Alexa voice command, “inferring” a result based on how the model has been trained. While training is done in a batch process, inference is done in real time, thereby representing different compute needs.

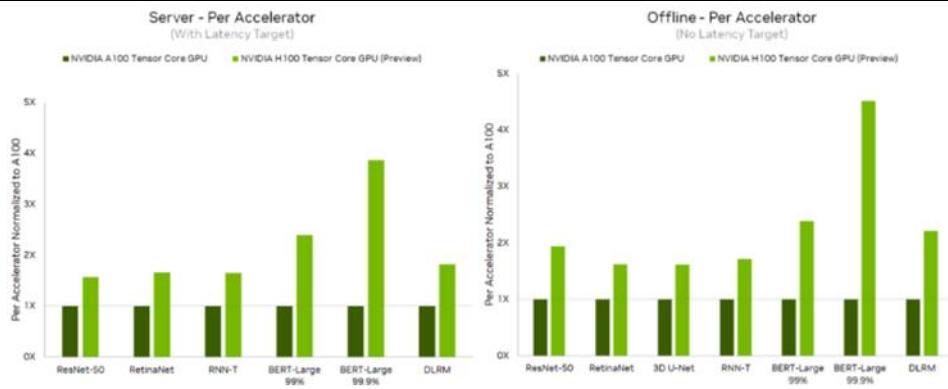
Traditionally, the majority of inference workloads run on x86 silicon, mostly from INTC. But GPU has been shown to deliver higher performance for inference, for much the same reason that GPU has proven to be better for training. While NVDA hasn't disclosed the specific revenue or growth rates for inference GPU vs. training, the company does claim that inference revenue is up 9x between NVDA's FY19 and FY22.

**Figure 318: Nvidia's Inference GPU Sales Up 9x in 3 Years**



Source: Nvidia

We note that in MLPerf Inference benchmarks, which measure how quickly a trained neural network can perform inference tasks on new data, NVDA's H100 had over 4x higher inference performance relative to the A100.

**Figure 319: Datacenter per-accelerator performance – H100 vs. A100**

Source: NVIDIA

NVDA has two approaches toward inference. One is lower ASP, lower performance GPUs which we describe in more detail below. One of the advantages of the A-100 is that it could be used for both inference and for training. Since inference is a real-time process, inference requires high compute requirements at peak times of the day, and less performance during non-peak hours. For that reason, A-100 can be used for training during non-peak times and for inference during peak times. This is well suited toward cloud applications.

NVDA's lower-end A-series GPUs are targeted toward enterprise level inference applications, from a 40-60W A2 edge level inference, to 165W A-30 GPUs. NVDA claims a 7x performance improvement as compared to an Intel Xeon Gold 6330N CPU for inference applications.

Intel hasn't stood still in the race for AI inference and has sought to defend their share in CPU inference with the launch of Sapphire Rapids this year, their newest datacenter processor. Sapphire Rapids includes Advanced Matrix Extension (AMX) accelerators to improve AI inference performance (INTC also claims this applicable for training small models as well), with up to 10x PyTorch performance as compared to the prior server CPU generation.

Despite Intel's efforts, we expect GPU, and specifically NVDA GPU to continue to take share of AI inference over time as more algorithms are optimized to run in parallel for faster compute. IDC also estimates training workloads being accelerated by an accelerator versus being run on the CPU increasing from 86% to 90% from 2022 to 2026 while inference workloads being accelerated rising from 34% to 53% by 2026.

**Figure 320: Higher % of AI workloads being accelerated**

	2021	2022	2023	2024	2025	2026	CAGR
Accelerated Training	7,548.0	9,247.0	10,957.0	12,395.0	13,575.0	14,696.0	14%
Non Accelerated Training	1,174.0	1,501.0	1,619.0	1,598.0	1,609.0	1,614.0	7%
<b>% Training Accelerated</b>	<b>86.5%</b>	<b>86.0%</b>	<b>87.1%</b>	<b>88.6%</b>	<b>89.4%</b>	<b>90.1%</b>	
Accelerated Inference	1,512.0	3,065.0	5,258.0	7,097.0	8,643.0	9,757.0	45%
Non Accelerated Inference	5,134.0	5,936.0	6,921.0	7,529.0	7,965.0	8,594.0	11%
<b>% Inference Accelerated</b>	<b>22.8%</b>	<b>34.1%</b>	<b>43.2%</b>	<b>48.5%</b>	<b>52.0%</b>	<b>53.2%</b>	

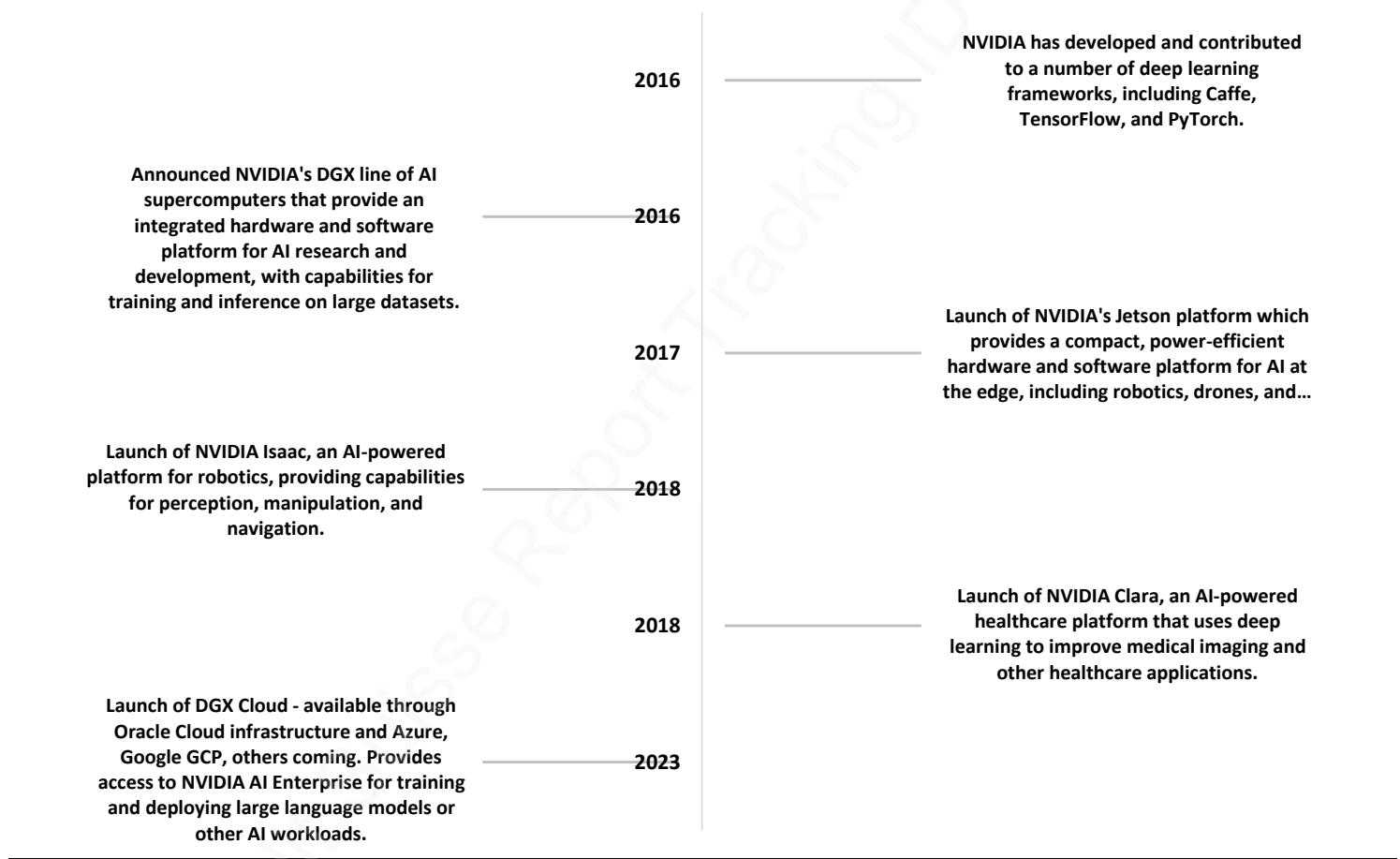
Source: IDC 2022

Nvidia believes that AI adoption is at an inflection point and it is poised to take advantage of the inflection given its GPU dominance in the AI market. On its 4Q call, the company noted "generative AI applications will help almost every industry do more faster" and below we underscore Nvidia's major developments that have contributed to the AI landscape.

- **CUDA:** In 2007, NVIDIA introduced CUDA, a parallel computing platform and programming model that enables developers to use GPUs for general-purpose computing, including AI and machine learning.
- **Drive AGX:** NVIDIA introduced Drive AGX in 2014 as a platform provides a hardware and software platform for autonomous vehicles, including perception, localization, and mapping.

- **GANs:** In 2014, NVIDIA researchers introduced GANs (Generative Adversarial Networks), a neural network architecture that can generate realistic images and videos.
- **Tesla V100:** NVIDIA introduced the Tesla V100 (2017), a high-performance GPU designed specifically for deep learning and other AI applications.
- **A100:** In 2020, NVIDIA introduced the A100, a new GPU designed for AI and high-performance computing workloads, with features like Tensor Cores and Multi-Instance GPU (MIG) technology for greater efficiency and scalability.
- **H100:** In 2022, NVIDIA introduced the H100, its newest AI GPU. The H100 features major advances to accelerate AI, HPC, memory bandwidth, interconnect and communication. The H100 enables 9x faster training and delivers industry-leading conversational AI, speeding up large language models by 30x over the previous generation, A100.

**Figure 321: Other Key AI Milestones Achieved by Nvidia**



Source: Company website

# Wiwynn Corporation

**2023 remains in positive growth mode with expanding opportunity from AI infrastructure**

IT Hardware

6669.TW

Target price (12M, NT\$) 1,100

Outperform<sup>[V]</sup>

- **Top pick among cloud ODMs as the only pure cloud play.** Wiwynn is our top preference given its 100% sales exposure to CSPs as a leading cloud IT infrastructure solutions provider having key customers led by Meta and Microsoft, representing 50-60% and 30-40% of the total sales, respectively, while it is also ramping up aggressively at Amazon, despite its contribution remaining at a single-digit percentage in 2022. Although we acknowledge the ongoing market concerns on potential capex/orders cut, we continue to see resiliently positive outlook from a cloud ODM perspective, and believe applying universal data points post the aggressive raw materials inventory stocking in the past few years could result in misleading expectations on ODMs.
- **Cloud ODMs as first wave of AI infrastructure beneficiary.** We continue to expect hyperscalers to be the industry pioneers ramping up their respective proprietary AI solutions and AI-enabled backbone infrastructure presenting Taiwan cloud ODMs as the first wave of beneficiaries given their exclusive positioning within the US hyperscalers chain and early partnership with Nvidia on DGX/HGX system from 2017, as well as their strong design capability for developing AI solution systems for Enterprise applications.
- **AI infrastructure providing incremental upsides.** Notably, on the back of Meta's infrastructure strategy shift to AI native architecture plus Microsoft's ongoing leadership on generative AI services (i.e., OpenAI's ChatGPT), we expect it should directly support better topline expansion to Wiwynn (i.e., mgmt already noted over 50% of projects in pipeline have degrees of AI) on hiking content per system. Although this may come at the expenses of its prospective margin percentage, Wiwynn recently highlighted its confidence in keeping a more stable margin rate trend going forward leveraging its leadership in cooling solutions as well as increasing value proposition on greater R&D and engineering resources involved in close partnership with its key customers rolling out the next Gen of cloud IT infrastructure.
- **Reiterate OUTPERFORM.** We have a 12M TP of NT\$1,100 based on 13x 2023E EPS, 10% premium to ODM peers' historical average, vs past three-year average of 14x in a range of 9-19x, implying still decent nearly 15% upside at the latest close. We expect Wiwynn's shares to trade up further on resiliently positive cloud outlook plus incremental upside potential on faster capitalization of AI infrastructure market take-off in the next few years.

Price (24 Feb 23, NT\$)	969.00
Upside/downside (%)	13.5
Mkt cap (NT\$/US\$ mn)	169,421 / 5,516
Enterprise value (NT\$ mn)	156,747
Number of shares (mn)	174.84
Free float (%)	89.2
52-wk price range (NT\$)	1,035-617
ADTO-6M (US\$ mn)	33.8

[V] = Stock Considered Volatile (see Disclosure Appendix)

## Research Analysts

### Harvie Chou

886 2 2715 6364  
harvie.chou@credit-suisse.com

### Jerry Su

886 2 2715 6361  
jerry.su@credit-suisse.com

## Financial and valuation metrics

Year	12/21A	12/22E	12/23E	12/24E
Revenue (NT\$ mn)	192,626	292,876	300,161	346,498
EBITDA (NT\$ mn)	11,934.9	18,608.6	20,762.9	23,072.5
EBIT (NT\$ mn)	11,387.1	17,834.3	19,323.7	20,848.5
Net profit (NT\$ mn)	8,638.8	14,140.3	14,791.4	15,922.8
EPS (CS adj.) (NT\$)	49.41	80.88	84.6	91.07
Chg. from prev. EPS (%)	n.a.	0.0	0.0	0.0
Consensus EPS (NT\$)	n.a.	81.09	78.74	88.38
EPS growth (%)	0.2	63.7	4.6	7.6
P/E (x)	19.6	12.0	11.5	10.6
Dividend yield (%)	2.6	5.0	5.2	5.6
EV/EBITDA (x)	14.9	8.3	8.1	7.1
P/B (x)	6.09	4.3	3.71	3.21
ROE (%)	32.9	42.1	34.8	32.4
Net debt/equity (%)	29.3	(38.1)	(1.8)	(12.0)

Source: Company data, Refinitiv, Credit Suisse estimates

## Share price performance



The price relative chart measures performance against the TAIWAN SE WEIGHTED INDEX which closed at 15,503.79 on 24/02/23. On 24/02/23 the spot exchange rate was NT\$30.71/US\$1

Performance	1M	3M	12M
Absolute (%)	23.4	13.3	(0.8)
Relative (%)	19.6	8.4	11.1

# Accton

## AI infrastructure take-off benefiting both cloud and enterprise segments

2345.TW

Target price (12M, NT\$) **330.00**  
Outperform

IT Hardware

- **Beneficiary of AI infrastructure take-off.** We believe hyperscalers to be the industry pioneers ramping up their respective proprietary AI solutions and AI enabled backbone infrastructure, allowing Taiwan ODMs to be the first wave beneficiary on hardware deployment, given their strong design capability for developing AI solutions and rising requirements on heat dissipation. We believe Accton will be one of the major beneficiaries within Taiwan cloud ODMs for AI infrastructure demand take-off as it supplies high-speed ethernet switch and smartNIC (combined 50%+ of sales) to leading US hyperscalers. It has also developed liquid cooling feature for high-speed switch solutions, which could help to capture the rising TAM from AI and ML in the mid- to long-term.
- **Speed upgrade could also proliferate to enterprise applications.** We believe AI will also be adopted on enterprise network in the mid- to long-term on widening use cases and improving cost structure. Besides driving content upgrade for enterprise servers, this should also support further bandwidth and speed upgrade for enterprise networking switches. As Accton is also the major ODM for enterprise switch makers (20%+ of sales) with HPE and Juniper as the major customers, we believe it could also benefit from the increasing adoption of AI infrastructure on enterprise application in the mid- to long-term.
- **Near-term momentum supported by robust visibility.** 4Q22 sales of NT\$22.2 bn grew 8% QoQ and 30% YoY, ahead of its guidance of slight QoQ decline, as a result of stronger pull in for 400G switch for hyperscalers and better component supply for enterprise ODM customers. Despite a higher base, we believe the robust visibility for both cloud and enterprise ODM should keep Accton on track to achieve its full year sales target of double-digit YoY growth, while OPM should further expand on better mix and scale.
- **Reiterate OUTPERFORM.** We rate Accton OUTPERFORM with TP of NT\$330, based on 22x 2023 P/E, average P/E multiple of its past three years' average. We believe the robust order visibility and improving margin, as well as potential upside on faster capitalization of AI infrastructure market take-off should support the outperformance of its shares relative to peers.

### Financial and valuation metrics

Year	12/21A	12/22E	12/23E	12/24E
Revenue (NT\$ mn)	59,598.7	77,205.6	85,010.5	90,428.3
EBITDA (NT\$ mn)	6,207.2	10,339.4	12,023.2	13,601.4
EBIT (NT\$ mn)	5,440.7	9,475.3	11,024.0	12,323.0
Net profit (NT\$ mn)	4,705.1	7,975.6	8,734.7	9,786.5
EPS (CS adj.) (NT\$)	8.4	14.24	15.59	17.47
Chg. from prev. EPS (%)	n.a.	0.0	0.0	0.0
Consensus EPS (NT\$)	n.a.	14.09	15.31	17.56
EPS growth (%)	(6.9)	69.5	9.5	12.0
P/E (x)	33.7	19.9	18.1	16.2
Dividend yield (%)	2.1	3.6	3.9	4.4
EV/EBITDA (x)	24.7	14.0	11.8	10.3
P/B (x)	9.84	7.8	6.79	5.89
ROE (%)	30.3	43.8	40.0	38.9
Net debt/equity (%)	(32.7)	(66.6)	(69.6)	(69.9)

Source: Company data, Refinitiv, Credit Suisse estimates

Price (24 Feb 23, NT\$)	283.00
Upside/downside (%)	16.6
Mkt cap (NT\$/US\$ mn)	158,497 / 5,160
Enterprise value (NT\$ mn)	144,528
Number of shares (mn)	560.06
Free float (%)	75.3
52-wk price range (NT\$)	290-203
ADTO-6M (US\$ mn)	27.3

### Research Analysts

**Jerry Su**  
886 2 2715 6361  
jerry.su@credit-suisse.com

**Harvie Chou**  
886 2 2715 6364  
harvie.chou@credit-suisse.com

### Share price performance



The price relative chart measures performance against the TAIWAN SE WEIGHTED INDEX which closed at 15,503.79 on 24/02/23. On 24/02/23 the spot exchange rate was NT\$30.71/US\$1

Performance	1M	3M	12M
Absolute (%)	12.1	6.6	11.2
Relative (%)	8.3	1.7	23.1

# Taiwan Semiconductor Manufacturing

## AI to drive a further inflection in TSMC's high performance computing business

Semiconductor Devices

2330.TW

Target price (12M, NT\$) **580.00**  
Outperform

- **HPC is TSMC's strongest growth driver.** TSMC's HPC business segment has been its strongest growth driver, doubling to 42% of company sales in the past decade as high growth and share gains in AI, GPUs/Gaming and CPUs has increased its share of the semiconductor compute manufacturing TAM from 18% in 2015 to 61% in 2022. We now view the company as poised for continued growth from an inflection in AI use cases to help it maintain this category's 20%+ growth CAGR through 2025.
- **Drivers of outgrowth.** TSMC's gains should continue including 1) high share of AI acceleration and networking through NVIDIA GPU, Marvell and Broadcom networking switch, and FPGA/ASICs including Xilinx, Google, Broadcom, Cerebras, Alibaba, and multiple start-ups (Biren, Cerebras, Graphcore), 2) recovery of NVIDIA gaming share from Samsung, 3) AMD's continued server gains and growth of ARM CPUs including Apple, Amazon and Ampere, along with pick-up of Intel compute tiles from Meteor Lake in 4Q23, 4) higher ASPs passing on higher manufacturing cost and complexity along with ability to pass that cost on through its process/ecosystem advantage and time to market leadership.
- **AI has potential to drive further upside.** TSMC in the mid-term could also see potential upside on the AI opportunity with faster penetration into servers relative to Yole's projected 10% penetration in 2023. A scenario of doubling AI penetration to 20% of servers would grow AI accelerators from US\$5.5bn to US\$11.0bn and add NT\$2.29 to TSMC EPS, about US\$30-35 on TSMC's share price at the current 15-17x multiple.
- **A potential catalyst to counter some of its recent de-rating.** AI also has potential to be a re-rating catalyst for TSMC after de-rating from 25-30x peak levels in 2020-21 to current 14x P/E and discount to SOX trading at 17x P/E. TSMC's ability to capture high share of the fastest growing driver in semiconductors (along with high share in ADAS as well) and ramp of overseas fabs to avoid lessen share loss risk from location could allow it to recapture some of its lost multiple. A shift in Intel strategy back toward outsourcing if the foundry approach with Meteor Lake/Arrow Lake is successful would also help that perception. Key risk to monitor though is Intel's ability to reassert its process position with 18A if on time to high volume ramp by 2H25 to match up with TSMC's 2nm ramp.

### Financial and valuation metrics

Year	12/21A	12/22E	12/23E	12/24E
Revenue (NT\$ mn)	1,587,415	2,263,891	2,292,213	2,621,600
EBITDA (NT\$ mn)	1,072,376	1,558,533	1,544,224	1,809,172
EBIT (NT\$ mn)	649,981	1,121,279	981,061	1,104,417
Net profit (NT\$ mn)	597,073	1,016,901	866,046	972,433
EPS (CS adj.) (NT\$)	23.03	39.22	33.4	37.5
Chg. from prev. EPS (%)	n.a.	0.0	0.0	0.0
Consensus EPS (NT\$)	n.a.	39.2	33.86	41.44
EPS growth (%)	15.2	70.3	(14.8)	12.3
P/E (x)	22.2	13.0	15.3	13.6
Dividend yield (%)	2.1	2.2	2.2	2.9
EV/EBITDA (x)	12.0	8.2	8.3	7.0
P/B (x)	6.11	4.5	3.76	3.22
ROE (%)	29.7	39.8	26.8	25.5
Net debt/equity (%)	(15.6)	(16.5)	(14.1)	(15.2)

Source: Company data, Refinitiv, Credit Suisse estimates

Price (24 Feb 23, NT\$)	511.00
Upside/downside (%)	13.5
Mkt cap (NT\$/US\$ mn)	13,250,424 / 431,399
Enterprise value (NT\$ mn)	12,759,678
Number of shares (mn)	25,930
Free float (%)	87.3
52-wk price range (NT\$)	602-371
ADTO-6M (US\$ mn)	595.2

### Research Analysts

**Randy Abrams, CFA**  
886 2 2715 6366  
randy.abrams@credit-suisse.com

**Haas Liu**  
886 2 2715 6365  
haas.liu@credit-suisse.com

**Angela Dai, CFA**  
886 2 2715 6363  
angela.dai@credit-suisse.com

### Share price performance



Performance	1M	3M	12M
Absolute (%)	1.6	2.6	(15.4)
Relative (%)	(2.2)	(2.3)	(3.5)

# Baidu

## Best positioned to capture next wave of AI innovation; Ernie Bot rollout as key focus

Consumer Internet

- **Best positioned as an early mover.** Among Chinese internet operators, we believe Baidu, the largest search engine in China and the earliest mover in AI industry, possess the necessary technical expertise and resources to become a major player in AI chatbot industry. Baidu announced to launch AI chatbot initiative “Ernie Bot” (文心一言) through a waitlist in mid-March and set to fully integrate it into its ecosystem at a later stage. We believe its core competencies in full stack AI capabilities, a solid base of Chinese-based data sets, as well as affluent search inquiries create a competitive moat.
- **Ernie Bot to further enhance Baidu search leadership:** We think Ernie Bot is likely to serve as a complementary tool to its traditional search engine, which helps solidify Baidu's search position and gain new share, with incremental benefits in **(1) boosting traffic** as a result of a more immersive search experience; **(2) deepening monetization** as personalized content recommendations can improve the efficiency of sales lead generation and drive better conversion of long-tail content; **(3) expanding search market** given its potential to tap under-penetrated search market through partnerships. In addition, 400+ enterprises showed strong interest in joining Ernie Bot for initial testing, which could offer an upselling opportunity to its AI Cloud.
- **Despite taking time to see material upside.** We expect Ernie Bot's monetization to arrive earlier than its autonomous driving, given established strong user awareness of AI chatbot products. However, given small-scale product launch at a nascent stage, revenue and cost impact on Baidu would be limited in the near term. Its monetization strategy and timing will depend on its ability to deliver a seamless and personalized user experience. On the cost side, capex will be mainly incurred in advanced chips that can provide massive computing power to support Ernie Bot scale-up.
- **Maintain OUTPERFORM.** Our SOTP implied TP is US\$176/HK\$171. We see incremental option value from new initiatives (AI chatbot/ASD) and potential buyback (US\$5bn) as tangible catalysts ahead. **Key challenges:** content diversity and accuracy; potential tighter regulation on AIGC content; escalated US-China frictions.

### Financial and valuation metrics

Year	12/22A	12/23E	12/24E
Revenue (Rmb mn)	123,675	138,244	151,938
EBITDA (Rmb mn)	17,645.0	26,568.9	29,611.6
EBIT (Rmb mn)	15,911.0	18,274.3	18,976.0
Net profit (Rmb mn)	20,680.0	23,842.7	26,733.7
EPS (CS adj.) (Rmb)	7.41	8.55	9.59
Chg. from prev. EPS (%)	n.a.	0.0	0.0
Consensus EPS (Rmb)	n.a.	8.0	9.03
EPS growth (%)	10.8	15.4	12.2
P/E (x)	16.1	13.9	12.4
Dividend yield (%)	0.0	0.0	0.0
EV/EBITDA (x)	20.0	11.7	9.2
P/B (x)	1.49	1.24	1.1
ROE (%)	9.5	9.7	9.4
Net debt/equity (%)	6.5	(9.3)	(21.3)

Source: Company data, Refinitiv, Credit Suisse estimates

9888.HK

Target price (12M, HK\$) 171.00  
Outperform<sup>[V]</sup>

Price (28 Feb 23, HK\$)	134.80
Upside/downside (%)	26.9
Mkt cap (HK\$/US\$ mn)	382,156 / 48,688
Enterprise value (Rmb mn)	346,197
Number of shares (mn)	353.61
Free float (%)	79.0
52-wk price range (HK\$)	162-75.10
ADTO-6M (US\$ mn)	140.7

[V] = Stock Considered Volatile (see Disclosure Appendix)

### Research Analysts

**Kenneth Fong**  
852 2101 6395  
kenneth.kc.fong@credit-suisse.com

**Lauren Zuo**  
852 2101 7986  
lauren.zuo@credit-suisse.com

### Share price performance



The price relative chart measures performance against the MSCI CHINA F INDEX which closed at 6,642.15 on 28/02/23. On 28/02/23 the spot exchange rate was HK\$7.85/US\$1

Performance	1M	3M	12M
Absolute (%)	(1.5)	32.7	(8.5)
Relative (%)	12.8	26.7	9.0

**Companies Mentioned (Price as of 01-Mar-2023)**

**ASM International** (ASMI.AS, €303.6)  
**ASML Holding N.V.** (ASML.AS, €574.6)  
**Accton** (2345.TW, NT\$287.5, OUTPERFORM, TP NT\$330.0)  
**Advanced Micro Devices, Inc.** (AMD.OQ, \$78.29)  
**Alchip Tech** (3661.TW, NT\$1035.0)  
**Alibaba Group Holding Limited** (BABA.N, \$89.95)  
**Alibaba Group Holding Limited** (9988.HK, HK\$91.9)  
**Alphabet** (GOOGL.OQ, \$90.36)  
**Amazon.com Inc.** (AMZN.OQ, \$92.17)  
**Apple Inc** (AAPL.OQ, \$145.31)  
**Arista Networks** (ANET.N, \$138.56)  
**Aspeed** (5274.TWO, NT\$2965.0)  
**Baidu** (9888.HK, HK\$145.0, OUTPERFORM[V], TP HK\$171.0)  
**Baidu** (BIDU.OQ, \$143.66, OUTPERFORM[V], TP \$176.0)  
**Box** (BOX.N, \$33.58)  
**Chroma** (2360.TW, NT\$178.5)  
**Cloudflare** (NET.N, \$58.51)  
**Dell Technologies** (DELL.N, \$40.44)  
**Delta Elec TH** (DELTA.BK, B1986.0)  
**Delta Electronics** (2308.TW, NT\$289.5)  
**Equinix, Inc.** (EQIX.OQ, \$675.98)  
**Ericsson** (ERICb.ST, Skr57.56)  
**Ericsson** (ERIC.OQ, \$5.46)  
**Five9** (FIVN.OQ, \$65.83)  
**Freshworks** (FRSH.OQ, \$14.5)  
**GUC** (3443.TW, NT\$1180.0)  
**Gigabyte Technology Co., Ltd** (2376.TW, NT\$120.5)  
**Hewlett Packard Enterprise** (HPE.N, \$15.54)  
**IBIDEN** (4062.T, ¥4,830)  
**Inari Amertron** (INAR.KL, RM2.42)  
**Inspur** (000977.SZ, Rmb42.88)  
**Intel Corp.** (INTC.OQ, \$25.33)  
**International Business Machines** (IBM.N, \$128.19)  
**Intuit** (INTU.OQ, \$401.27)  
**Inventec Co Ltd** (2356.TW, NT\$26.95)  
**JD.com** (JD.OQ, \$45.73)  
**JD.com** (9618.HK, HK\$183.3)  
**Jabil Circuit Inc.** (JBL.N, \$84.19)  
**Kingssoft Corporation Limited** (3888.HK, HK\$27.55)  
**Kinsus Interconnect Tech** (3189.TW, NT\$111.5)  
**Lenovo Group Ltd** (0992.HK, HK\$7.32)  
**Lotes Co.,Ltd.** (3533.TW, NT\$857.0)  
**Marvell Technology, Inc.** (MRVL.OQ, \$45.36)  
**Meta Platforms, Inc.** (META.OQ, \$173.42)  
**Micron Tech** (MU.OQ, \$57.34)  
**Microsoft** (MSFT.OQ, \$246.27, OUTPERFORM, TP \$285.0)  
**Monolithic Power** (MPWR.OQ, \$486.53)  
**Montage** (688008.SS, Rmb57.48)  
**NVIDIA Corporation** (NVDA.OQ, \$226.98, OUTPERFORM[V], TP \$275.0)  
**Nan Ya Printed Circuit Board** (8046.TW, NT\$240.5)  
**NetEase.com** (9999.HK, HK\$131.3)  
**NetEase.com** (NTES.OQ, \$81.52)  
**Nice** (NICE.OQ, \$208.71)  
**Nokia** (NOKIA.HE, €4.36)  
**Nokia** (NOK.N, \$4.6)  
**Parade Technologies** (4966.TWO, NT\$995.0)  
**Pentamaster** (PMAS.KL, RM4.9)  
**PowerSchool** (PWSC.N, \$19.98)  
**Pure Storage** (PSTG.N, \$28.79)  
**Quanta Computer** (2382.TW, NT\$79.8)  
**SK Hynix Inc.** (000660.KS, W89,400)  
**SOPHiA GENETICS** (SOPH.OQ, \$2.62)  
**Salesforce.com** (CRM.N, \$167.35)  
**Samsung Electro-Mechanics** (009150.KS, W144,000)  
**Samsung Electronics** (005930.KS, W60,600)  
**Shinko Electric Industries** (6967.T, ¥3,845)  
**Soitec** (SOIT.PA, €145.55)  
**Taiwan Semiconductor Manufacturing** (2330.TW, NT\$522.0, OUTPERFORM, TP NT\$580.0)  
**Tencent Holdings** (0700.HK, HK\$368.8)  
**Unimicron Technology Corp** (3037.TW, NT\$131.5)  
**WUS** (002463.SZ, Rmb16.58)  
**Wiwynn Corporation** (6669.TW, NT\$940.0, OUTPERFORM[V], TP NT\$1100.0)  
**Zhongji Innolight** (300308.SZ, Rmb35.91)  
**Zoom Video Communications** (ZM.OQ, \$69.62)  
**ZoomInfo Technologies** (ZI.OQ, \$24.35)

---

## Disclosure Appendix

### **Analyst Certification**

Sami Badri, Chris Caso, Shannon Cross, Rich Hilliker, Randy Abrams, CFA, Stephen Ju, Adithya Metuku, CFA, Akinori Kanemoto, Chaolien Tseng, Clive Cheung, Danny Chan, Fred Lee, Harvie Chou, Jerry Su, Kenneth Fong, Keon Han, Kyna Wong, Pauline Chen, Sang Uk Kim,

Yufeng Shen, Soyun Shin, Douglas Mitchelson, Kevin McVeigh, Douglas Mitchelson, Susan Roth Katzke, Charles Zhou, CFA, Hu Shen, Trung Huynh, Dan Leonard, Judah C. Frommer, CFA, A.J. Rice, Jason Liu, Haas Liu, Scott Deusdle, Jamie Cook, CFA, Guy Hardwick, Kaumil Gajrawala, Lauren Silberman, Nicholas Campanella, Gary Zhou, CFA, John Roberts, Andrew Kligerman, Moshe Orenbuch, Amy Wong, Lauren Zuo, Ariel Rosa, Iris Zheng, CFA, Daniel Cui, Jessie Xu, Tayo Okusanya, II, CFA, CPA, Grant Joslin, Keon Han, Sang Uk Kim and Akinori Kanemoto each certify, with respect to the companies or securities that the individual analyzes, that (1) the views expressed in this report accurately reflect his or her personal views about all of the subject companies and securities and (2) no part of his or her compensation was, is or will be directly or indirectly related to the specific recommendations or views expressed in this report.

### 3-Year Price and Rating History for Accton (2345.TW)

2345.TW	Closing Price	Target Price	
Date	(NT\$)	(NT\$)	Rating
19-Mar-20	149.00	200.00	O
14-May-20	211.00	240.00	
15-Jul-20	252.50	240.00	N
28-Sep-20	219.00	250.00	O
11-Nov-20	220.00	260.00	
11-Jan-21	294.00	330.00	
13-Aug-21	297.00	310.00	N
07-Oct-21	250.50	275.00	O
12-Nov-21	288.00	325.00	
15-Dec-21	292.50	330.00	
17-Mar-22	236.50	320.00	
07-Jul-22	222.00	305.00	
12-Aug-22	276.50	330.00	

\* Asterisk signifies initiation or assumption of coverage.

### 3-Year Price and Rating History for Baidu (9888.HK)

9888.HK	Closing Price	Target Price	
Date	(HK\$)	(HK\$)	Rating
19-May-21	186.70	339.00	O *
14-Jul-21	179.50	335.00	
13-Aug-21	154.10	318.00	
18-Oct-21	167.70	274.00	
18-Nov-21	156.40	247.00	
11-Jan-22	146.50	184.00	
18-Jan-22	147.80	180.00	
02-Mar-22	158.10	189.00	
20-Apr-22	123.60	181.00	
27-May-22	132.20	184.00	
14-Oct-22	104.40	166.00	
23-Nov-22	92.95	165.00	
22-Feb-23	140.50	171.00	

\* Asterisk signifies initiation or assumption of coverage.



### 3-Year Price and Rating History for Baidu (BIDU.OQ)

BIDU.OQ	Closing Price	Target Price	
Date	(US\$)	(US\$)	Rating
14-Apr-20	100.92	156.00	O
09-Sep-20	119.40		*
19-Oct-20	129.25	156.00	O
17-Nov-20	143.92	170.00	
06-Jan-21	203.97	275.00	
20-Jan-21	260.90	300.00	
18-Feb-21	298.01	407.00	
01-Apr-21	219.70		R
23-Apr-21	220.94	407.00	O
29-Apr-21	212.29	356.00	
19-May-21	189.16	350.00	*
14-Jul-21	181.34	346.00	
13-Aug-21	152.45	328.00	
18-Oct-21	171.04	282.00	
18-Nov-21	154.36	255.00	
11-Jan-22	156.70	190.00	
18-Jan-22	152.94	185.00	
02-Mar-22	160.57	195.00	
20-Apr-22	122.31	187.00	
27-May-22	139.09	190.00	
14-Oct-22	100.29	172.00	
23-Nov-22	97.00	169.00	
18-Jan-23	125.92	170.00	
22-Feb-23	137.12	176.00	

\* Asterisk signifies initiation or assumption of coverage.

### 3-Year Price and Rating History for Microsoft (MSFT.OQ)

MSFT.OQ	Closing Price	Target Price	
Date	(US\$)	(US\$)	Rating
23-Apr-20	171.42	190.00	O
29-Apr-20	177.43	195.00	
16-Jul-20	203.92	225.00	
28-Oct-20	202.68	235.00	
27-Jan-21	232.90	265.00	
15-Mar-21	234.81		*
01-Apr-21	242.35	265.00	O*
28-Apr-21	254.56	300.00	
27-Jul-21	286.54	320.00	
27-Oct-21	323.17	340.00	
16-Nov-21	339.51	400.00	*
26-Oct-22	231.32	365.00	
20-Dec-22	241.80	365.00	*
26-Jan-23	248.00	285.00	

\* Asterisk signifies initiation or assumption of coverage.



### 3-Year Price and Rating History for NVIDIA Corporation (NVDA.OQ)

NVDA.OQ	Closing Price	Target Price	
Date	(US\$)	(US\$)	Rating
01-May-20	70.69		NR
20-May-20	89.70	106.25	O
20-Aug-20	121.41	132.50	
01-Sep-20	138.21	155.00	
12-Apr-21	152.09	175.00	
18-Aug-21	190.40	225.00	
15-Nov-21	300.25	400.00	
25-Apr-22	199.02		*
25-May-22	169.75	205.00	O
15-Nov-22	166.66	210.00	*
23-Feb-23	236.64	275.00	

\* Asterisk signifies initiation or assumption of coverage.

### 3-Year Price and Rating History for SK Hynix Inc. (000660.KS)

000660.KS	Closing Price	Target Price	
Date	(W)	(W)	Rating
02-Apr-20	80,000	118,000	O
23-Jul-20	82,400	120,000	
27-Aug-20	79,100		R
16-Oct-20	85,300	120,000	O
09-Dec-20	120,500		R
22-Jul-22	100,000		NR
02-Sep-22	91,700	150,000	O
06-Oct-22	89,900	140,000	
13-Dec-22	82,300	130,000	

\* Asterisk signifies initiation or assumption of coverage.

### 3-Year Price and Rating History for Samsung Electro-Mechanics (009150.KS)

009150.KS	Closing Price	Target Price	
Date	(W)	(W)	Rating
10-Mar-20	128,500	150,000	O
02-Apr-20	96,800	125,000	
06-Jul-20	130,000	150,000	
04-Aug-20	139,500	165,000	
26-Oct-20	139,500	170,000	
03-Dec-20	163,000	190,000	
27-Jan-21	214,500	250,000	
17-Mar-21	194,000	270,000	
27-Jan-22	171,500	230,000	
07-Mar-22	156,500	220,000	
07-Apr-22	155,500	230,000	
19-Jul-22	139,000	180,000	
01-Sep-22	135,000	185,000	
20-Sep-22	128,000	175,000	
26-Oct-22	122,500	165,000	
29-Nov-22	134,500	160,000	
30-Jan-23	144,700	185,000	

\* Asterisk signifies initiation or assumption of coverage.



### 3-Year Price and Rating History for Samsung Electronics (005930.KS)

005930.KS	Closing Price	Target Price	
Date	(W)	(W)	Rating
04-Mar-20	57,400	82,000	O
02-Apr-20	46,800	66,000	
29-Apr-20	50,000	65,000	
30-Jul-20	59,000	75,100	
21-Sep-20	59,200	81,000	
03-Dec-20	69,700	90,000	
12-Jan-21	90,600	126,000	
28-Oct-21	70,700	115,000	
11-Jul-22	58,800	92,000	
06-Oct-22	56,300	89,000	
23-Dec-22	58,100	86,000	

\* Asterisk signifies initiation or assumption of coverage.

### 3-Year Price and Rating History for Taiwan Semiconductor Manufacturing (2330.TW)

2330.TW	Closing Price	Target Price	
Date	(NT\$)	(NT\$)	Rating
04-Mar-20	320.50	360.00	O
02-Apr-20	271.50	320.00	
15-May-20	298.00	350.00	
18-May-20	290.00	300.00	N
09-Jun-20	319.00	320.00	
13-Jul-20	354.50	340.00	
17-Jul-20	367.00	365.00	
27-Jul-20	424.50	435.00	O
12-Oct-20	460.00	500.00	
05-Jan-21	542.00	635.00	
15-Jan-21	601.00	650.00	
16-Apr-21	610.00	675.00	
25-Aug-21	585.00	700.00	
11-Jan-22	651.00	750.00	
14-Jan-22	672.00	800.00	
13-Apr-22	573.00	730.00	
13-Jul-22	470.50	630.00	
12-Oct-22	397.50	600.00	
11-Jan-23	484.50	580.00	

\* Asterisk signifies initiation or assumption of coverage.

### 3-Year Price and Rating History for Wiwynn Corporation (6669.TW)

6669.TW	Closing Price	Target Price	
Date	(NT\$)	(NT\$)	Rating
21-Jul-20	787.00	960.00	O *
18-Dec-20	703.00	880.00	
20-Jan-21	833.00	950.00	
15-Mar-21	890.00	985.00	
07-May-21	944.00	1030.00	
15-Jul-21	1015.00	1090.00	N
23-Sep-21	955.00	1160.00	O
10-Jan-22	1110.00	1220.00	
08-Jul-22	669.00	955.00	
12-Sep-22	795.00	1050.00	
13-Jan-23	774.00	1100.00	

\* Asterisk signifies initiation or assumption of coverage.

### **As of December 10, 2012 Analysts' stock rating are defined as follows:**

**Outperform (O) :** The stock's total return is expected to outperform the relevant benchmark\* over the next 12 months.

**Neutral (N) :** The stock's total return is expected to be in line with the relevant benchmark\* over the next 12 months.

**Underperform (U) :** The stock's total return is expected to underperform the relevant benchmark\* over the next 12 months.

\*Relevant benchmark by region: As of 10th December 2012, Japanese ratings are based on a stock's total return relative to the analyst's coverage universe which consists of all companies covered by the analyst within the relevant sector, with Outperforms representing the most attractive, Neutrals the less attractive, and Underperforms the least attractive investment opportunities. As of 2nd October 2012, U.S. and Canadian as well as European (excluding Turkey) ratings are based on a stock's total return relative to the analyst's coverage universe which consists of all companies covered by the analyst within the relevant sector, with Outperforms representing the most attractive, Neutrals the less attractive, and Underperforms the least attractive investment opportunities. For Latin America, Turkey and Asia (excluding Japan and Australia), stock ratings are based on a stock's total return relative to the average total return of the relevant country or regional benchmark (India - S&P BSE Sensex Index); for China A share the relevant index is the Shanghai Shenzhen CSI 300 (CSI300); prior to 2nd October 2012 U.S. and Canadian ratings were based on (1) a stock's absolute total return potential to its current share price and (2) the relative attractiveness of a stock's total return potential within an analyst's coverage universe. For Australian and New Zealand stocks, the expected total return (ETR) calculation includes 12-month rolling dividend yield. An Outperform rating is assigned where an ETR is greater than or equal to 7.5%; Underperform where an ETR less than or equal to 5%. A Neutral may be assigned where the ETR is between -5% and 15%. The overlapping rating range allows analysts to assign a rating that puts ETR in the context of associated risks. Prior to 18 May 2015, ETR ranges for Outperform and Underperform ratings did not overlap with Neutral thresholds between 15% and 7.5%, which was in operation from 7 July 2011.

**Restricted (R) :** In certain circumstances, Credit Suisse policy and/or applicable law and regulations preclude certain types of communications, including an investment recommendation, during the course of Credit Suisse's engagement in an investment banking transaction and in certain other circumstances.

**Not Rated (NR) :** Credit Suisse Equity Research does not have an investment rating or view on the stock or any other securities related to the company at this time.

**Not Covered (NC) :** Credit Suisse Equity Research does not provide ongoing coverage of the company or offer an investment rating or investment view on the equity security of the company or related products.

**Volatility Indicator [V] :** A stock is defined as volatile if the stock price has moved up or down by 20% or more in a month in at least 8 of the past 24 months or the analyst expects significant volatility going forward.

Analysts' sector weightings are distinct from analysts' stock ratings and are based on the analyst's expectations for the fundamentals and/or valuation of the sector\* relative to the group's historic fundamentals and/or valuation:

**Overweight :** The analyst's expectation for the sector's fundamentals and/or valuation is favorable over the next 12 months.

**Market Weight :** The analyst's expectation for the sector's fundamentals and/or valuation is neutral over the next 12 months.

**Underweight :** The analyst's expectation for the sector's fundamentals and/or valuation is cautious over the next 12 months.

\*An analyst's coverage sector consists of all companies covered by the analyst within the relevant sector. An analyst may cover multiple sectors.

Credit Suisse's distribution of stock ratings (and banking clients) is:

#### Global Ratings Distribution

Rating	Versus universe (%)	Of which banking clients (%)
Outperform/Buy*	52%	(26% banking clients)
Neutral/Hold*	37%	(22% banking clients)
Underperform/Sell*	10%	(21% banking clients)
Restricted	1%	

Please click [here](#) to view the MAR quarterly recommendations and investment services report for fundamental research recommendations.

\*For purposes of the NYSE and FINRA ratings distribution disclosure requirements, our stock ratings of Outperform, Neutral, and Underperform most closely correspond to Buy, Hold, and Sell, respectively; however, the meanings are not the same, as our stock ratings are determined on a relative basis. (Please refer to definitions above.) An investor's decision to buy or sell a security should be based on investment objectives, current holdings, and other individual factors.

#### Important Global Disclosures

Credit Suisse's research reports are made available to clients through our proprietary research portal on CS PLUS. Credit Suisse research products may also be made available through third-party vendors or alternate electronic means as a convenience. Certain research products are only made available through CS PLUS. The services provided by Credit Suisse's analysts to clients may depend on a specific client's preferences regarding the frequency and manner of receiving communications, the client's risk profile and investment, the size and scope of the overall client relationship with the Firm, as well as legal and regulatory constraints. To access all of Credit Suisse's research that you are entitled to receive in the most timely manner, please contact your sales representative or go to <https://plus.credit-suisse.com>.

Credit Suisse's policy is to update research reports as it deems appropriate, based on developments with the subject company, the sector or the market that may have a material impact on the research views or opinions stated herein.

Credit Suisse's policy is only to publish investment research that is impartial, independent, clear, fair and not misleading. For more detail please refer to Credit Suisse's Policies for Managing Conflicts of Interest in connection with Investment Research: <https://www.credit-suisse.com/sites/disclaimers-ib/en/managing-conflicts.html>.

Any information relating to the tax status of financial instruments discussed herein is not intended to provide tax advice or to be used by anyone to provide tax advice. Investors are urged to seek tax advice based on their particular circumstances from an independent tax professional.

Credit Suisse has decided not to enter into business relationships with companies that Credit Suisse has determined to be involved in the development, manufacture, or acquisition of anti-personnel mines and cluster munitions. For Credit Suisse's position on the issue, please see <https://www.credit-suisse.com/media/assets/corporate/docs/about-us/responsibility/banking/policy-summaries-en.pdf>.

The analyst(s) responsible for preparing this research report received compensation that is based upon various factors including Credit Suisse's total revenues, a portion of which are generated by Credit Suisse's investment banking activities

Please refer to the firm's disclosure website at <https://rave.credit-suisse.com/disclosures/view/selectArchive> for the definitions of abbreviations typically used in the target price method and risk sections.

*See the Companies Mentioned section for full company names*

Credit Suisse currently has, or had within the past 12 months, the following as investment banking client(s): MSFT.OQ, 9888.HK, BIDU.OQ, MU.OQ, IBM.N, ZI.OQ, 0992.HK, 9999.HK, AMD.OQ, AMZN.OQ, INTC.OQ, PWSC.N, CRM.N, 0700.HK, BABA.N, GOOGL.OQ, NTES.OQ, DELL.N, 9988.HK

Credit Suisse provided investment banking services to the subject company (MSFT.OQ, 9888.HK, BIDU.OQ, MU.OQ, IBM.N, ZI.OQ, 0992.HK, 9999.HK, AMD.OQ, AMZN.OQ, INTC.OQ, PWSC.N, CRM.N, 0700.HK, BABA.N, GOOGL.OQ, NTES.OQ, DELL.N, 9988.HK) within the past 12 months.

Within the last 12 months, Credit Suisse has received compensation for non-investment banking services or products from the following issuer(s): MSFT.OQ, MU.OQ, 005930.KS, IBM.N, 000660.KS, AMD.OQ, INTC.OQ, PWSC.N, 0700.HK, BABA.N, GOOGL.OQ, 009150.KS, NOKIA.HE, NOK.N, DELL.N, AAPL.OQ, HPE.N, 9988.HK

Credit Suisse has managed or co-managed a public offering of securities for the subject company (0992.HK, AMD.OQ, DELL.N) within the past 12 months.

Within the past 12 months, Credit Suisse has received compensation for investment banking services from the following issuer(s): MSFT.OQ, 9888.HK, BIDU.OQ, MU.OQ, IBM.N, 0992.HK, 9999.HK, AMD.OQ, AMZN.OQ, INTC.OQ, 0700.HK, BABA.N, GOOGL.OQ, NTES.OQ, DELL.N, 9988.HK

Credit Suisse expects to receive or intends to seek investment banking related compensation from the subject company (MSFT.OQ, NVDA.OQ, 9888.HK, BIDU.OQ, MU.OQ, 005930.KS, IBM.N, ZI.OQ, MRVL.OQ, 0992.HK, SOIT.PA, ANET.N, AMD.OQ, INTC.OQ, PWSC.N, CRM.N, BABA.N, GOOGL.OQ, INTU.OQ, DELL.N, HPE.N, 9988.HK) within the next 3 months.

Credit Suisse currently has, or had within the past 12 months, the following issuer(s) as client(s), and the services provided were non-investment-banking, securities-related: MSFT.OQ, MU.OQ, 005930.KS, IBM.N, 000660.KS, AMD.OQ, INTC.OQ, PWSC.N, 0700.HK, BABA.N, GOOGL.OQ, 009150.KS, NOKIA.HE, NOK.N, DELL.N, AAPL.OQ, HPE.N, 9988.HK

Credit Suisse currently has, or had within the past 12 months, the following issuer(s) as client(s), and the services provided were non-investment-banking, non securities-related: MSFT.OQ, MU.OQ, IBM.N, AMD.OQ, INTC.OQ, BABA.N, GOOGL.OQ, NOKIA.HE, NOK.N, DELL.N, HPE.N, 9988.HK

Credit Suisse makes a market in the securities of the following subject issuer(s): (MSFT.OQ).

Credit Suisse acts as a market maker in the shares, depositary receipts, interests or units issued by, and/or any warrants or options on these shares, depositary receipts, interests or units of the following subject issuer(s): 9888.HK, 0992.HK, 9999.HK, 0700.HK, 9618.HK, 9988.HK.

Credit Suisse or a member of the Credit Suisse Group is a market maker or liquidity provider in the securities of the following subject issuer(s): ASMI.AS, ASML.AS, 2345.TW, AMD.OQ, 3661.TW, BABA.N, 9988.HK, GOOGL.OQ, AMZN.OQ, AAPL.OQ, ANET.N, 5274.TWO, 9888.HK, BIDU.OQ, BOX.N, 2360.TW, NET.N, DELL.N, DELTA.BK, 2308.TW, EQIX.OQ, ERICb.ST, ERIC.OQ, FIVN.OQ, FRSH.OQ, 3443.TW, 2376.TW, HPE.N, 4062.T, INAR.KL, 000977.SZ, INTC.OQ, IBM.N, INTU.OQ, 2356.TW, JD.OQ, 9618.HK, JBL.N, 3888.HK, 3189.TW, 0992.HK, 3533.TW, MRVL.OQ, META.OQ, MU.OQ, MSFT.OQ, MPWR.OQ, 688008.SS, NVDA.OQ, 8046.TW, 9999.HK, NTES.OQ, NOKIA.HE, NOK.N, 4966.TWO, PMAS.KL, PWSC.N, PSTG.N, 2382.TW, 000660.KS, SOPH.OQ, CRM.N, 009150.KS, 005930.KS, 6967.T, SOIT.PA, 2330.TW, 0700.HK, 3037.TW, 002463.SZ, 6669.TW, 300308.SZ, ZM.OQ, ZI.OQ

A member of the Credit Suisse Group is party to an agreement with, or may have provided services set out in sections A and B of Annex I of Directive 2014/65/EU of the European Parliament and Council ("MiFID Services") to, the subject issuer (MSFT.OQ, 9888.HK, BIDU.OQ, MU.OQ, IBM.N, ZI.OQ, 0992.HK, 9999.HK, AMD.OQ, AMZN.OQ, INTC.OQ, PWSC.N, CRM.N, 0700.HK, BABA.N, GOOGL.OQ, NTES.OQ, DELL.N, 9988.HK) within the past 12 months.

Credit Suisse may have interest in (INAR.KL, PMAS.KL)

As of the date of this report, Credit Suisse beneficially own 1% or more of a class of common equity securities of (SOPH.OQ).

Credit Suisse beneficially holds >0.5% long position of the total issued share capital of the subject company (2360.TW).

Gillian Sheldon, a Senior Advisor to Credit Suisse, is an Advisory Board Member of Salesforce.com (CRM.N).

Associate of a CS employee is a Senior Officer of Tencent Holdings Ltd.

Credit Suisse is acting as Compliance Advisor to the Hong Kong listing of NetEase Inc. (9999.HK)

A Research Analyst, who contributed to the content of this Research Report, holds a long position in the equity securities of Five9 (FIVN.OQ).

For date and time of production, dissemination and history of recommendation for the subject company(ies) featured in this report, disseminated within the past 12 months, please refer to the link: <https://rave.credit-suisse.com/disclosures/view/report?i=779798&v=2m2xfk4x5llzmkpzyoxgxuj7>.

### Important Regional Disclosures

Singapore recipients should contact Credit Suisse AG, Singapore Branch for any matters arising from, or in connection with, this research report. Analysts who conduct site visits of covered issuers are not permitted to accept payment or reimbursement for travel expenses from the issuer for the site visit.

For Credit Suisse Securities (Canada), Inc.'s policies and procedures regarding the dissemination of equity research, please visit <https://www.credit-suisse.com/sites/disclaimers-ib/en/canada-research-policy.html>.

Investors should note that income from such securities and other financial instruments, if any, may fluctuate and that price or value of such securities and instruments may rise or fall and, in some cases, investors may lose their entire principal investment.

To the extent any Credit Suisse equity research analyst employed by Credit Suisse International (a "UK Analyst") has interactions with a Spanish domiciled client of Credit Suisse AG or its affiliates, such UK Analyst will be acting for and on behalf of Credit Suisse Bank (Europe), S.A. but

not an employee of Credit Suisse Bank (Europe) S.A. , with respect only to the provision of equity research services to Spanish domiciled clients of Credit Suisse AG or its affiliates.

Pursuant to CVM Resolution No. 20/2021, of February 25, 2021, the author(s) of the report hereby certify(ies) that the views expressed in this report solely and exclusively reflect the personal opinions of the author(s) and have been prepared independently, including with respect to Credit Suisse. Part of the author(s)'s compensation is based on various factors, including the total revenues of Credit Suisse, but no part of the compensation has been, is, or will be related to the specific recommendations or views expressed in this report. In addition, Credit Suisse declares that: Credit Suisse has provided, and/or may in the future provide investment banking, brokerage, asset management, commercial banking and other financial services to the subject company/companies or its affiliates, for which they have received or may receive customary fees and commissions, and which constituted or may constitute relevant financial or commercial interests in relation to the subject company/companies or the subject securities.

For Thai listed companies mentioned in this report, the independent 2022 Corporate Governance Report survey results published by the Thai Institute of Directors Association are being disclosed pursuant to the policy of the Office of the Securities and Exchange Commission: Delta Elec TH (Excellent)

**Taiwan (Chinese Taipei) Disclosures:** This research report is for reference only. Investors should carefully consider their own investment risk and note they may be subject to the applicable rules and regulations in Taiwan. Investment results are the responsibility of the investor. Reports written by Taiwan based analysts on non-Taiwan listed companies are not considered as recommendations to buy or sell securities. Reports may not be reproduced without the permission of Credit Suisse. Pursuant to the 'Taiwan Stock Exchange Regulations Governing Securities Firms Recommending Trades in Securities to Customers' and the 'Taipei Exchange Rules Governing Securities Firms Recommending Trades in Securities to Customers', in order for a non-client of Credit Suisse AG, Taipei Securities Branch to receive this research report, no provision by such non-client of the content of the report to a third party, nor any conflict of interest, is permitted. By receiving this research report, any such non-client is deemed to acknowledge and accept our terms and disclaimers included herein.

As at the date of this report, Credit Suisse has financial interests that aggregate to an amount equal to or more than 1% of the market capitalization of (9888.HK, 9988.HK).

This research report is authored by:

**Credit Suisse (Hong Kong) Limited** Chaolien Tseng ; Clive Cheung ; Kenneth Fong ; Kyna Wong ; Yufeng Shen ; Charles Zhou, CFA ; Hu Shen ; Jason Liu ; Gary Zhou, CFA ; Lauren Zuo ; Iris Zheng, CFA ; Daniel Cui ; Jessie Xu ; Alex Liu ; Edward Liu

**Credit Suisse Securities (Japan) Limited** ..... Akinori Kanemoto

**Credit Suisse Securities (Europe) Limited, Seoul Branch** ..... Keon Han ; Sang Uk Kim ; Soyun Shin ; DJ Kim

**Credit Suisse International** ..... Adithya Metuku, CFA ; Amy Wong ; Sarah Roberts

**Credit Suisse AG, Taipei Securities Branch** Randy Abrams, CFA ; Harvie Chou ; Jerry Su ; Pauline Chen ; Haas Liu ; Angela Dai, CFA

**Credit Suisse Securities (Malaysia) Sdn Bhd** ..... Danny Chan

**Credit Suisse Securities (USA) LLC** Sami Badri ; Chris Caso ; Shannon Cross ; Rich Hilliker ; Stephen Ju ; Fred Lee ; Douglas Mitchelson ; Kevin McVeigh ; Susan Roth Katzke ; Trung Huynh ; Dan Leonard ; Judah C. Frommer, CFA ; A.J. Rice ; Scott Deuschle ; Jamie Cook, CFA ; Guy Hardwick ; Kaumil Gajrawala ; Lauren Silberman ; Nicholas Campanella ; John Roberts ; Andrew Kligerman ; Moshe Orenbuch ; Ariel Rosa ; Tayo Okusanya, II, CFA, CPA ; Grant Joslin ; Radi Sultan, CFA ; Andy Kellam ; Ryan Cui, CFA ; George Engroff ; Liz Pate ; Nicholas Welsch-Lehmann

To the extent this is a report authored in whole or in part by a non-U.S. analyst and is made available in the U.S., the following are important disclosures regarding any non-U.S. analyst contributors: The non-U.S. research analysts listed below (if any) are not registered/qualified as research analysts with FINRA. The non-U.S. research analysts listed below may not be associated persons of CSSU and therefore may not be subject to the FINRA 2241 restrictions on communications with a subject company, public appearances and trading securities held by a research analyst account.

**Credit Suisse (Hong Kong) Limited** Chaolien Tseng ; Clive Cheung ; Kenneth Fong ; Kyna Wong ; Yufeng Shen ; Charles Zhou, CFA ; Hu Shen ; Jason Liu ; Gary Zhou, CFA ; Lauren Zuo ; Iris Zheng, CFA ; Daniel Cui ; Jessie Xu ; Alex Liu ; Edward Liu

**Credit Suisse Securities (Japan) Limited** ..... Akinori Kanemoto

**Credit Suisse Securities (Europe) Limited, Seoul Branch** ..... Keon Han ; Sang Uk Kim ; Soyun Shin ; DJ Kim

**Credit Suisse International** ..... Adithya Metuku, CFA ; Amy Wong ; Sarah Roberts

**Credit Suisse AG, Taipei Securities Branch** Randy Abrams, CFA ; Harvie Chou ; Jerry Su ; Pauline Chen ; Haas Liu ; Angela Dai, CFA

**Credit Suisse Securities (Malaysia) Sdn Bhd** ..... Danny Chan

There is currently no universal definition or exhaustive list defining the issues or factors that are covered by the concept of "ESG" (Environmental, Social, Governance). If not indicated otherwise, 'ESG' is used interchangeably with the terms 'sustainable' and 'sustainability'. Unless indicated otherwise, the views expressed herein are based on CS' own assumptions and interpretation of ESG at the time of drafting. CS' views on ESG may evolve over time and are subject to change.

Where a sustainability assessment is identified as including elements which track environmental, social or governance (ESG) objectives, CS is, wholly or in part, reliant on third-party sources of information (including, but not limited to, such information produced by the issuing/manufacturing company itself) and external guidance. These sources of information may be limited in terms of accuracy, availability and timeliness. It is possible that the data from ESG data providers may be incorrect, unavailable (e.g., not existing, or absence of look-through), or not fully updated. CS has not sought to independently verify information obtained from public and third-party sources and makes no representations or warranties as to accuracy, completeness or reliability of such information. Additionally, as global laws, guidelines and regulations in relation to the tracking and provision of such data are evolving, all such disclosures are made on a non-reliance basis and are

subject to change. Unless required by applicable law, CS is not obliged to provide updates on sustainability assessments. Any updates might be subject to a time lag, due to e.g. lack of available data.

An ESG assessment reflects the opinion of the assessing party (CS or external parties such as rating agencies or other financial institutions). In the absence of a standardized ESG assessment system, each assessing party has its own research and analysis framework/methodology. Therefore, ESG assessment or risk levels given by different assessing parties to the same company can vary. Further, ESG assessment is limited to considering company performance against certain ESG criteria only and does not take into account the other factors needed to assess the value of a company.

---

Important disclosures regarding companies that are the subject of this report are available by calling +1 (877) 291-2683. The same important disclosures, with the exception of valuation methodology and risk discussions, are also available on Credit Suisse's disclosure website at <https://rave.credit-suisse.com/disclosures>. For valuation methodology and risks associated with any recommendation, price target, or rating referenced in this report, please refer to the disclosures section of the most recent report regarding the subject company.

This report is produced by subsidiaries and affiliates of Credit Suisse ("CS") operating under its Securities Research function within the Investment Banking Division. For more information on our structure, please use the following link: <https://www.credit-suisse.com/who-we-are>. All material presented in this report, unless specifically indicated otherwise, is under copyright to CS. None of the material, nor its content, nor any copy of it, may be altered in any way, transmitted to, copied or distributed to any other party, without the prior express written permission of CS. All trademarks, service marks and logos used in this report are trademarks or service marks or registered trademarks or service marks of CS. CS will not treat recipients of this report as its customers by virtue of their receiving this report. The investments and services contained or referred to in this report may not be suitable for you and it is recommended that you consult an independent investment advisor if you are in doubt about such investments or investment services. Nothing in this report constitutes investment, legal, accounting or tax advice, or a representation that any investment or strategy is suitable or appropriate to your individual circumstances, or otherwise constitutes a personal recommendation to you. Please note in particular that the bases and levels of taxation may change. Information and opinions presented in this report have been obtained or derived from sources believed by CS to be reliable, but CS makes no representation as to their accuracy or completeness. CS accepts no liability for loss arising from the use of the material presented in this report, except that this exclusion of liability does not apply to the extent that such liability arises under specific statutes or regulations applicable to CS. This report is not to be relied upon in substitution for the exercise of independent judgment. CS may have issued, and may in the future issue, other communications that are inconsistent with, and reach different conclusions from, the information presented in this report. Those communications reflect the different assumptions, views and analytical methods of the analysts who prepared them and CS is under no obligation to ensure that such other communications are brought to the attention of any recipient of this report. Some investments referred to in this report will be offered solely by a single entity and in the case of some investments solely by CS, or an associate of CS or CS may be the only market maker in such investments. Past performance should not be taken as an indication or guarantee of future performance, and no representation or warranty, express or implied, is made regarding future performance. Information, opinions and estimates contained in this report reflect a judgment at its original date of publication by CS and are subject to change without notice. The price, value or income from any of the securities or financial instruments mentioned in this report can fall as well as rise. The value of securities and financial instruments is subject to exchange rate fluctuation that may have a positive or adverse effect on the price or income of such securities or financial instruments. Investors in securities such as ADR's, the values of which are influenced by currency volatility, effectively assume this risk. Structured securities are complex instruments, typically involve a high degree of risk and are intended for sale only to sophisticated investors who are capable of understanding and assuming the risks involved. The market value of any structured security may be affected by changes in economic, financial and political factors (including, but not limited to, spot and forward interest and exchange rates), time to maturity, market conditions and volatility, and the credit quality of any issuer or reference issuer. Any investor interested in purchasing a structured product should conduct their own investigation and analysis of the product and consult with their own professional advisors as to the risks involved in making such a purchase. Some investments discussed in this report may have a high level of volatility. High volatility investments may experience sudden and large falls in their value causing losses when that investment is realised. Those losses may equal your original investment. Indeed, in the case of some investments the potential losses may exceed the amount of initial investment and, in such circumstances, you may be required to pay more money to support those losses. Income yields from investments may fluctuate and, in consequence, initial capital paid to make the investment may be used as part of that income yield. Some investments may not be readily realisable and it may be difficult to sell or realise those investments, similarly it may prove difficult for you to obtain reliable information about the value, or risks, to which such an investment is exposed. This report may provide the addresses of, or contain hyperlinks to, websites. Except to the extent to which the report refers to website material of CS, CS has not reviewed any such site and takes no responsibility for the content contained therein. Such address or hyperlink (including addresses or hyperlinks to CS's own website material) is provided solely for your convenience and information and the content of any such website does not in any way form part of this document. Accessing such website or following such link through this report or CS's website shall be at your own risk. This report may contain material that is not directed to, or intended for distribution to or use by, any person or entity who is a citizen or resident of or located in any locality, state, country or other jurisdiction where such distribution, publication, availability or use would be contrary to law or regulation or which would subject CS to any registration or licensing requirement within such jurisdiction. The information, tools and material presented in this report are provided to you for information purposes only and are not to be used or considered as an offer or the solicitation of an offer to sell or to buy or subscribe for securities or other financial instruments. CS may not have taken any steps to ensure that the securities referred to in this report are suitable for any particular investor. This research report does not, and is not intended to be, an advertisement within the meaning of article 68 of the Swiss Financial Services Act and/or article 95 of the Swiss Financial Services Ordinance. This research report is not a prospectus, basic information sheet (BIB) or a key information document (KID). Any recipients of this document should, however, note that any communication forwarding or using this research report as a basis for discussion, would qualify as such advertisement if it is intended to draw the recipients' attention to specific financial instruments covered by this research report.

This report is issued and distributed in **United Kingdom and European Union (except Germany and Spain)**: by Credit Suisse International, One Cabot Square, London E14 4QJ, England, which is authorised by the Prudential Regulation Authority and regulated by the Financial Conduct Authority and the Prudential Regulation Authority; **Spain**: Credit Suisse Bank (Europe), S.A. regulated by the Comision Nacional del Mercado de Valores; **Germany**: Credit Suisse (Deutschland) Aktiengesellschaft regulated by the Bundesanstalt fuer Finanzdienstleistungsaufsicht ("BaFin"). **United States**: Credit Suisse Securities (USA) LLC; **Canada**: Credit Suisse Securities (Canada), Inc.; **Switzerland**: Credit Suisse AG; **Brazil**: Credit Suisse (Brasil) S.A. Corretora de Títulos e Valores Mobiliários or its affiliates; **Mexico**: Banco Credit Suisse (México), S.A., Institución de Banca Múltiple, Grupo Financiero Credit Suisse (México) and Casa de Bolsa Credit Suisse (México), S.A. de C.V., Grupo Financiero Credit Suisse (México) ("Credit Suisse Mexico"). This document has been prepared for information purposes only and is exclusively distributed in Mexico to Institutional Investors. Credit Suisse Mexico is not responsible for any onward distribution of this report to non-institutional investors by any third party. The authors of this report have not received payment or compensation from any entity or company other than from the relevant Credit Suisse Group company employing them; **Japan**: by Credit Suisse Securities (Japan) Limited, Financial Instruments Firm, Director-General of Kanto Local Finance Bureau (Kinsho) No. 66, a member of Japan Securities Dealers Association, The Financial Futures Association of Japan, Japan Investment Advisers Association, Type II Financial Instruments Firms Association. This report has been prepared and issued for distribution in Japan to Credit Suisse's clients, including institutional investors; **Hong Kong SAR**: Credit Suisse (Hong Kong) Limited; **Australia**: Credit Suisse Equities (Australia) Limited; **Thailand**: Credit Suisse Securities (Thailand) Limited, regulated by the Office of the Securities and Exchange Commission, Thailand, having registered address at 63 Wireless Road (Witthayu) Bangkok 10330, Thailand, Tel. +66 2614 6000; **Malaysia**: Credit Suisse Securities (Malaysia) Sdn Bhd; **Singapore**: Credit Suisse AG, Singapore Branch; **India**: Credit Suisse Securities (India) Private Limited (CIN no.U67120MH1996PTC104392) regulated by the Securities and Exchange Board of India as Research Analyst (registration no. INH 000001030) and as Stock Broker (registration no. INZ200248233), having registered address at 9th Floor, Ceejay House, Dr.A.B. Road, Worli, Mumbai - 18, India, T- +91-22 6777 3777; **South Korea**: Credit Suisse Securities (Europe) Limited, Seoul Branch; **Taiwan (Chinese Taipei)**: Credit Suisse AG Taipei Securities Branch; **Indonesia**: PT Credit Suisse Sekuritas Indonesia; **Philippines**: Credit Suisse Securities (Philippines) Inc., and elsewhere in the world by the relevant authorised affiliate of the above.

#### Additional Regional Disclaimers

**Australia**: Credit Suisse Securities (Europe) Limited ("CSSEL") and Credit Suisse International ("CSI") are authorised by the Prudential Regulation Authority and regulated by the Financial Conduct Authority ("FCA") and the Prudential Regulation Authority under UK laws, which differ from Australian Laws. CSSEL and CSI do not hold an Australian Financial Services Licence ("AFSL") and are exempt from the requirement to hold an AFSL under the Corporations Act (Cth) 2001 ("Corporations Act") in respect of the financial services provided to Australian wholesale clients (within the meaning of section 761G of the Corporations Act) (hereinafter referred to as "Financial Services"). This material is not for distribution to retail clients and is directed exclusively at Credit Suisse's professional clients and eligible counterparties as defined by the FCA, and wholesale clients as defined under section 761G of the Corporations Act. Credit Suisse (Hong Kong) Limited ("CSHK") is licensed and regulated by the Securities and Futures Commission of Hong Kong under the laws of Hong Kong SAR, which differ from Australian laws. CSHKL does not hold an AFSL and is exempt from the requirement to hold an AFSL under the Corporations Act in respect of providing Financial Services. Investment banking services in the United States are provided by Credit Suisse Securities (USA) LLC, an affiliate of Credit Suisse Group. CSSU is regulated by the United States Securities and Exchange Commission under United States laws, which differ from Australian laws. CSSU does not hold an AFSL and is exempt from the requirement to hold an AFSL under the Corporations Act in respect of providing Financial Services. Credit Suisse Asset Management LLC (CSAM) is authorised by the Securities and Exchange Commission under US laws, which differ from Australian laws. CSAM does not hold an AFSL and is exempt from the requirement to hold an AFSL under the Corporations Act in respect of providing Financial Services. This material is provided solely to Institutional Accounts (as defined in the FINRA rules) who are Eligible Contract Participants (as defined in the US Commodity Exchange Act). Credit Suisse Equities (Australia) Limited (ABN 35 068 232 708) ("CSEAL") is an AFSL holder in Australia (AFSL 237237).

**EU**: This report has been produced by subsidiaries and affiliates of Credit Suisse operating under its Securities Research function within the Investment Banking Division.

In jurisdictions where CS is not already registered or licensed to trade in securities, transactions will only be effected in accordance with applicable securities legislation, which will vary from jurisdiction to jurisdiction and may require that the trade be made in accordance with applicable exemptions from registration or licensing requirements.

This material is issued and distributed in the U.S. by CSSU, a member of NYSE, FINRA, SIPC and the NFA, and CSSU accepts responsibility for its contents. Clients should contact analysts and execute transactions through a Credit Suisse subsidiary or affiliate in their home jurisdiction unless governing law permits otherwise.

CS may provide various services to US municipal entities or obligated persons ("municipalities"), including suggesting individual transactions or trades and entering into such transactions. Any services CS provides to municipalities are not viewed as "advice" within the meaning of Section 975 of the Dodd-Frank Wall Street Reform and Consumer Protection Act. CS is providing any such services and related information solely on an arm's length basis and not as an advisor or fiduciary to the municipality. In connection with the provision of the any such services, there is no agreement, direct or indirect, between any municipality (including the officials, management, employees or agents thereof) and CS for CS to provide advice to the municipality. Municipalities should consult with their financial, accounting and legal advisors regarding any such services provided by CS. In addition, CS is not acting for direct or indirect compensation to solicit the municipality on behalf of an unaffiliated broker, dealer, municipal securities dealer, municipal advisor, or investment adviser for the purpose of obtaining or retaining an engagement by the municipality for or in connection with Municipal Financial Products, the issuance of municipal securities, or of an investment adviser to provide investment advisory services to or on behalf of the municipality. If this report is being distributed by a financial institution other than Credit Suisse AG, or its affiliates, that financial institution is solely responsible for distribution. Clients of that institution should contact that institution to effect a transaction in the securities mentioned in this report or require further information. This report does not constitute investment advice by Credit Suisse to the clients of the distributing financial institution, and neither Credit Suisse AG, its affiliates, and their respective officers, directors and employees accept any liability whatsoever for any direct or consequential loss arising from their use of this report or its content. No information or communication provided herein or otherwise is intended to be, or should be construed as, a recommendation within the meaning of the US Department of Labor's final regulation defining "investment advice" for purposes of the Employee Retirement Income Security Act of 1974, as amended and Section 4975 of the Internal Revenue Code of 1986, as amended, and the information provided herein is intended to be general information, and should not be construed as, providing investment advice (impartial or otherwise).

**Malaysia**: Research provided to residents of Malaysia is authorised by the Head of Research for Credit Suisse Securities (Malaysia) Sdn Bhd, to whom they should direct any queries on +603 2723 2020.

**Singapore**: This report has been prepared and issued for distribution in Singapore to institutional investors, accredited investors and expert investors (each as defined under the Financial Advisers Regulations) only, and is also distributed by Credit Suisse AG, Singapore Branch to overseas investors (as defined under the Financial Advisers Regulations). Credit Suisse AG, Singapore Branch may distribute reports produced by its foreign entities or affiliates pursuant to an arrangement under Regulation 32C of the Financial Advisers Regulations. Singapore recipients should contact Credit Suisse AG, Singapore Branch at +65 6212-2000 for matters arising from, or in connection with, this report. By virtue of your status as an institutional investor, accredited investor, expert investor or overseas investor, Credit Suisse AG, Singapore Branch is exempted from complying with certain compliance requirements under the Financial Advisers Act, Chapter 110 of Singapore (the "FAA"), the Financial Advisers Regulations and the relevant Notices and Guidelines issued thereunder, in respect of any financial advisory service which Credit Suisse AG, Singapore Branch may provide to you.

Copyright © 2023 CREDIT SUISSE AG and/or its affiliates. All rights reserved.  
When you purchase non-listed Japanese fixed income securities (Japanese government bonds, Japanese municipal bonds, Japanese government guaranteed bonds, Japanese corporate bonds) from CS as a seller, you will be requested to pay the purchase price only.