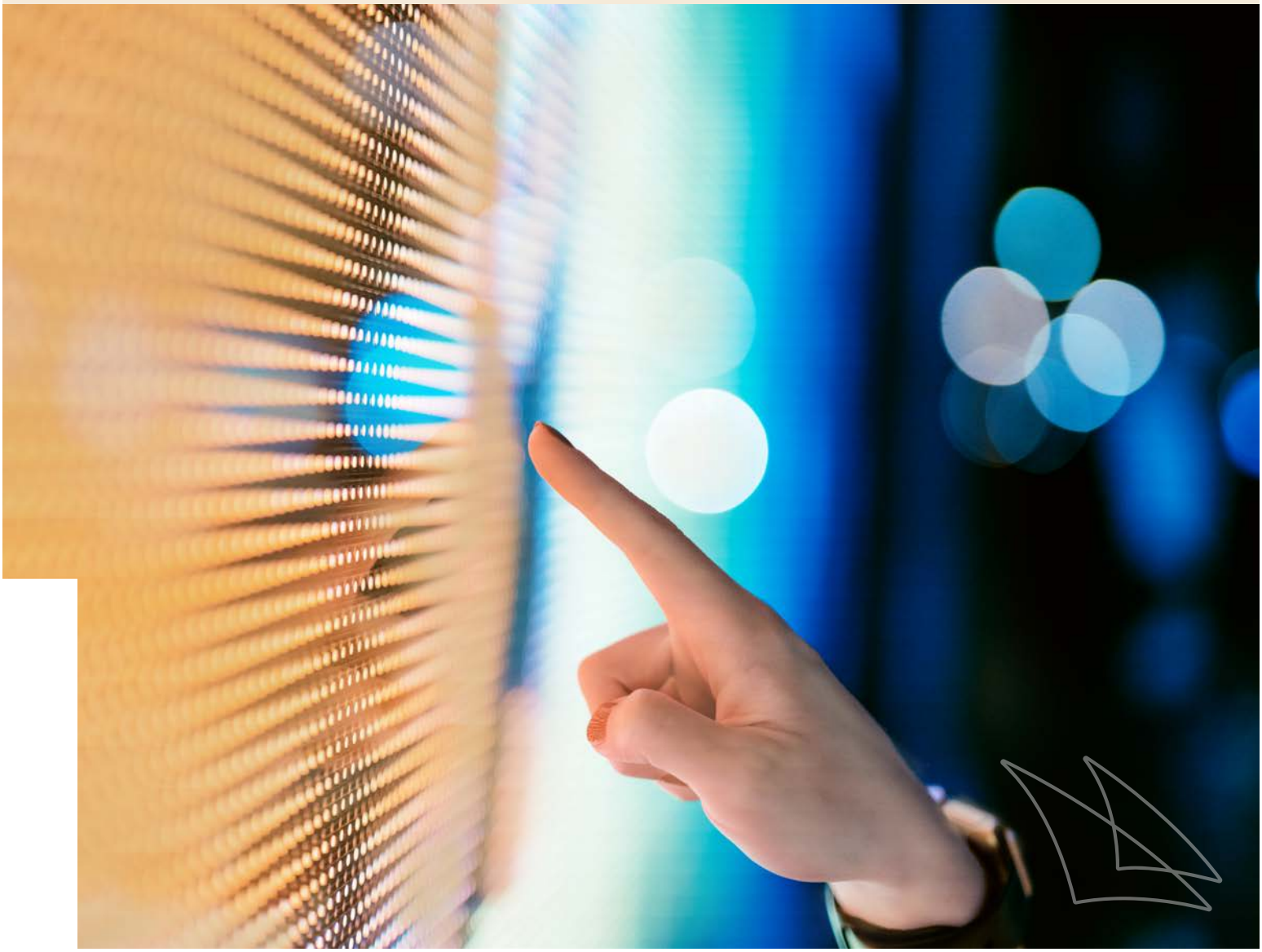
Connections Series

CREDIT SUISSE

ChatGPT: Unlocking the Potential of Large Language Models



Authors

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
212 325 8003
shannon.cross@credit-suisse.com

Stephen Ju
212 325 8662
stephen.ju@credit-suisse.com

Adithya Metuku, CFA
44 20 788 3 0384
adithya.metuku@credit-suisse.com

Akinori Kanemoto
81 3 4550 7363
akinori.kanemoto@credit-suisse.com

Chaolien Tseng
852 2101 6795
chaolien.tseng@credit-suisse.com

Clive Cheung
852 2101 7069
clive.cheung@credit-suisse.com

Danny Chan
60 3 2723 2082
danny.chan@credit-suisse.com

Edward Liu
852 2101 6734
edward.liu@credit-suisse.com

Fred Lee
212 325 2724
fred.lee@credit-suisse.com

Haas Liu
886 2 2715 6365
haas.liu@credit-suisse.com

Harvie Chou
886 2 2715 6364
harvie.chou@credit-suisse.com

Jasmine Wang
852 2101 7181
jasmine.wang@credit-suisse.com

Jerry Su
886 2 2715 6361
jerry.su@credit-suisse.com

Kenneth Fong
852 2101 6395
kenneth.kc.fong@credit-suisse.com

Keon Han
82 2 3707 3740
keon.han@credit-suisse.com

Kyna Wong
852 2101 6950
kyna.wong@credit-suisse.com

Lauren Zuo
852 2101 7986
lauren.zuo@credit-suisse.com

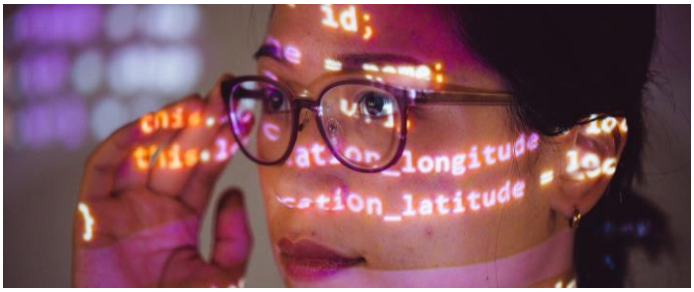
Pauline Chen
886 2 2715 6323
pauline.chen@credit-suisse.com

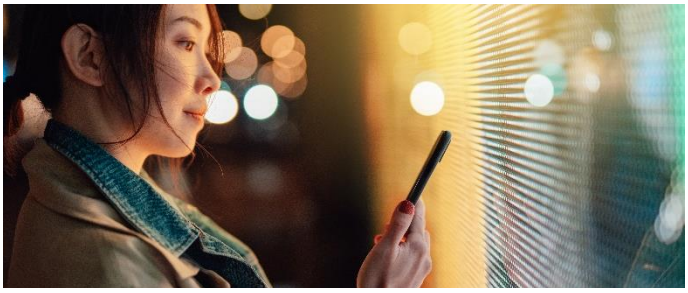
Rich Hilliker
212 325 5603
rich.hilliker@credit-suisse.com

Sang Uk Kim
82 2 3707 3795
sang.kim@credit-suisse.com

Yufeng Shen
852 3748 2582
yufeng.shen@credit-suisse.com

Contents





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 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 have been further fine-tuned over time with the next big milestone being OpenAI's recently released GPT-4, a LLM with significantly more functionality (reliability, creativity, sophistication of instruction) versus GPT-3. 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 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.

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 2019-2024E, 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.

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.

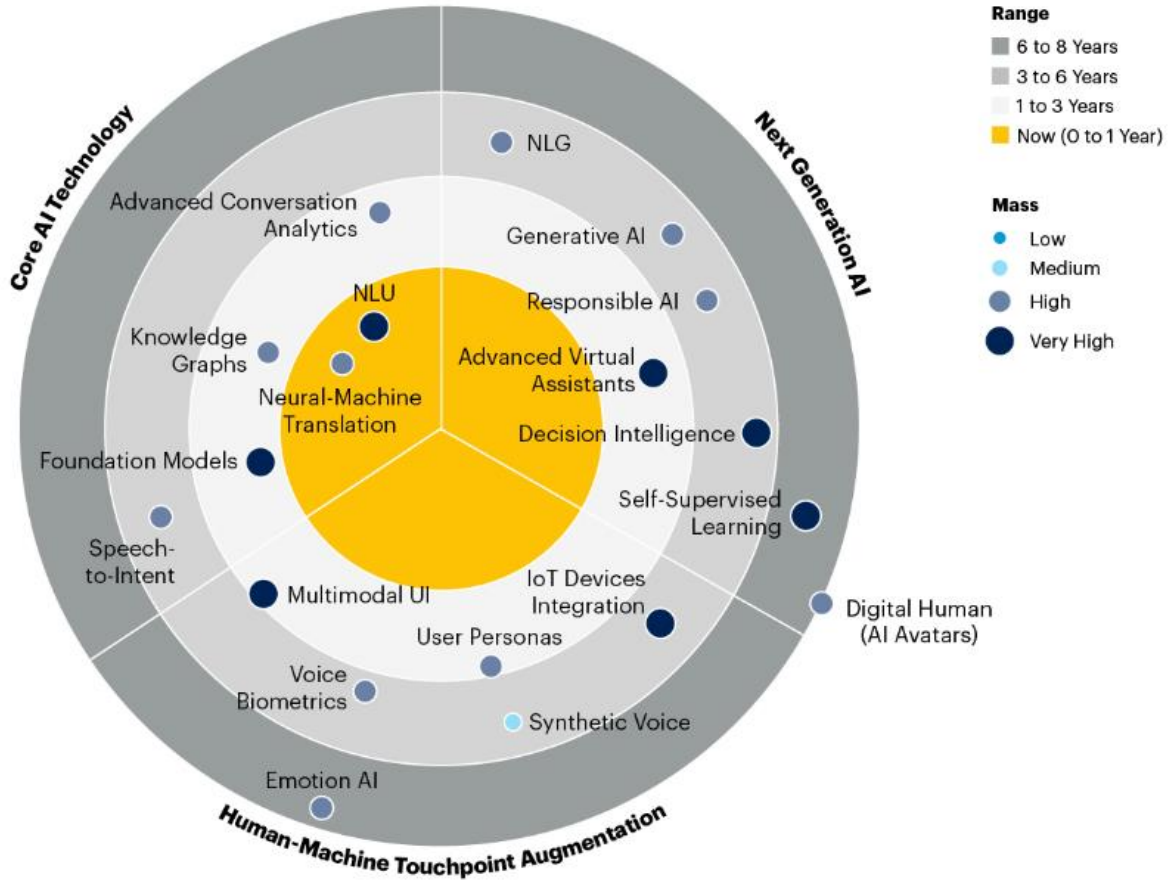
“The increasing popularity of ChatGPT has proved that ongoing innovations in conversational AI are materializing at a faster-than-expected pace… ChatGPT Plus is here and now.”

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.

- **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 three to six years from today. This actually aligns very closely to how OpenAI has described the effectivity 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.

Figure 1: Emerging Technologies Around Conversational AI

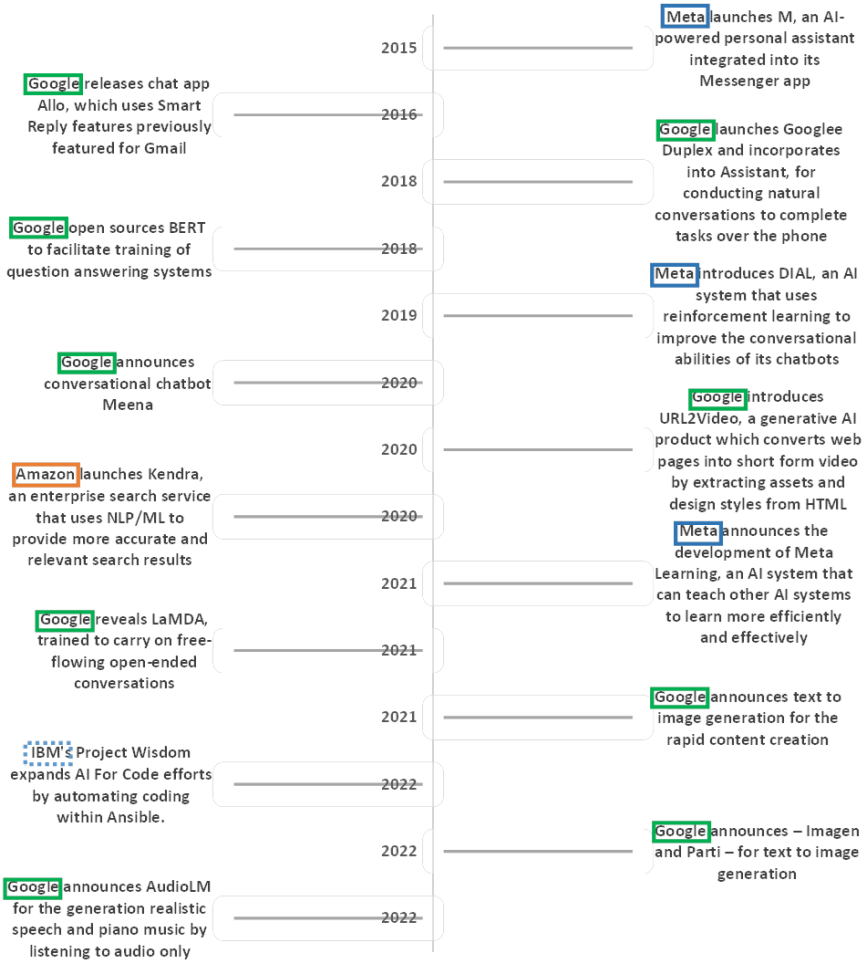


Source: Gartner. AI = Artificial Intelligence; IoT = Internet of Things; NLG = Natural Language Generation; NLU = Natural Language Understanding; UI = User Interface.

- **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 2, Google has released the greatest number of relevant papers/projects historically.

Figure 2: Key Milestones Achieved by Leading Tech Companies on Generative AI and Chatbot Technologies



Source: Company data.

What Are the Relevant AI Advancements Behind ChatGPT?

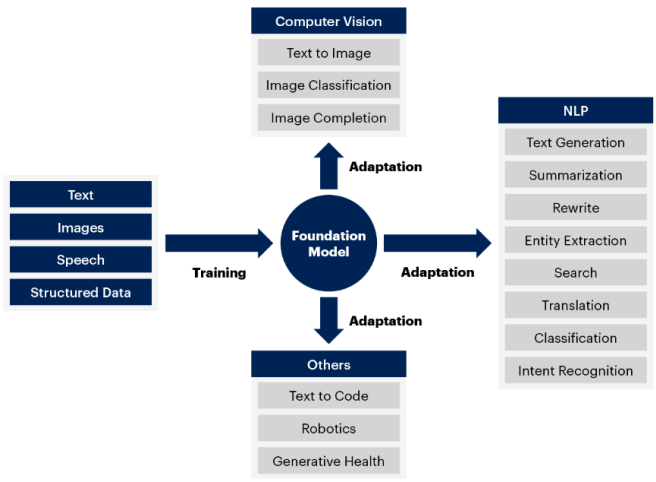
The increasing popularity of ChatGPT has proved that ongoing innovations in conversational AI are materializing at a faster-than-expected pace given the industry's expectations for timing of mainstream adoption are not expected for another 3-6yrs whereas ChatGPT Plus is here and now, with ~100M users (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 6), 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 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 3. 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 4 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 3: Key Characteristics and Applications of Foundation Models



Source: Gartner.

Figure 4: Examples of Zero-shot vs. Few-shot for In-context Learning

Source: Gartner.

How does an LLM Work?

To generate text, language models repeatedly sample new tokens (word fragments) based on the output token probabilities (Figure 5). 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.

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, posting some challenges as LLMs continue to become larger and more complex.

Challenges of Large Language Models

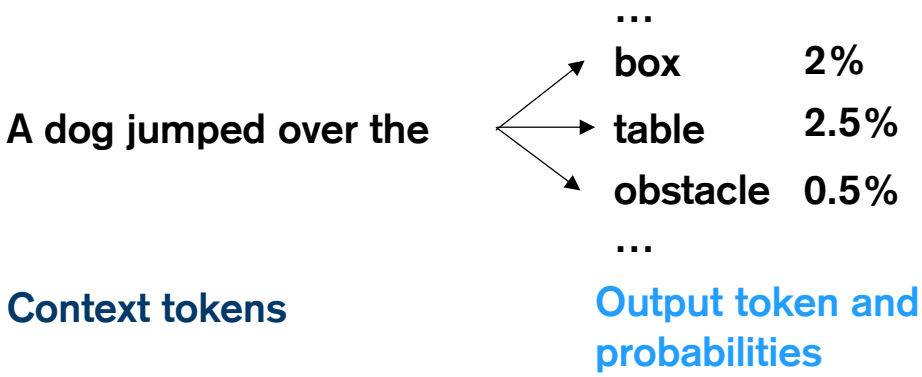
LLMs have been unleashing significant potential and creating synergies in areas such as search engines, NLP, healthcare, robotics, code generation, etc., for some time, and ChatGPT has become the most popular application of LLM. However, LLMs do face some challenges:

- **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

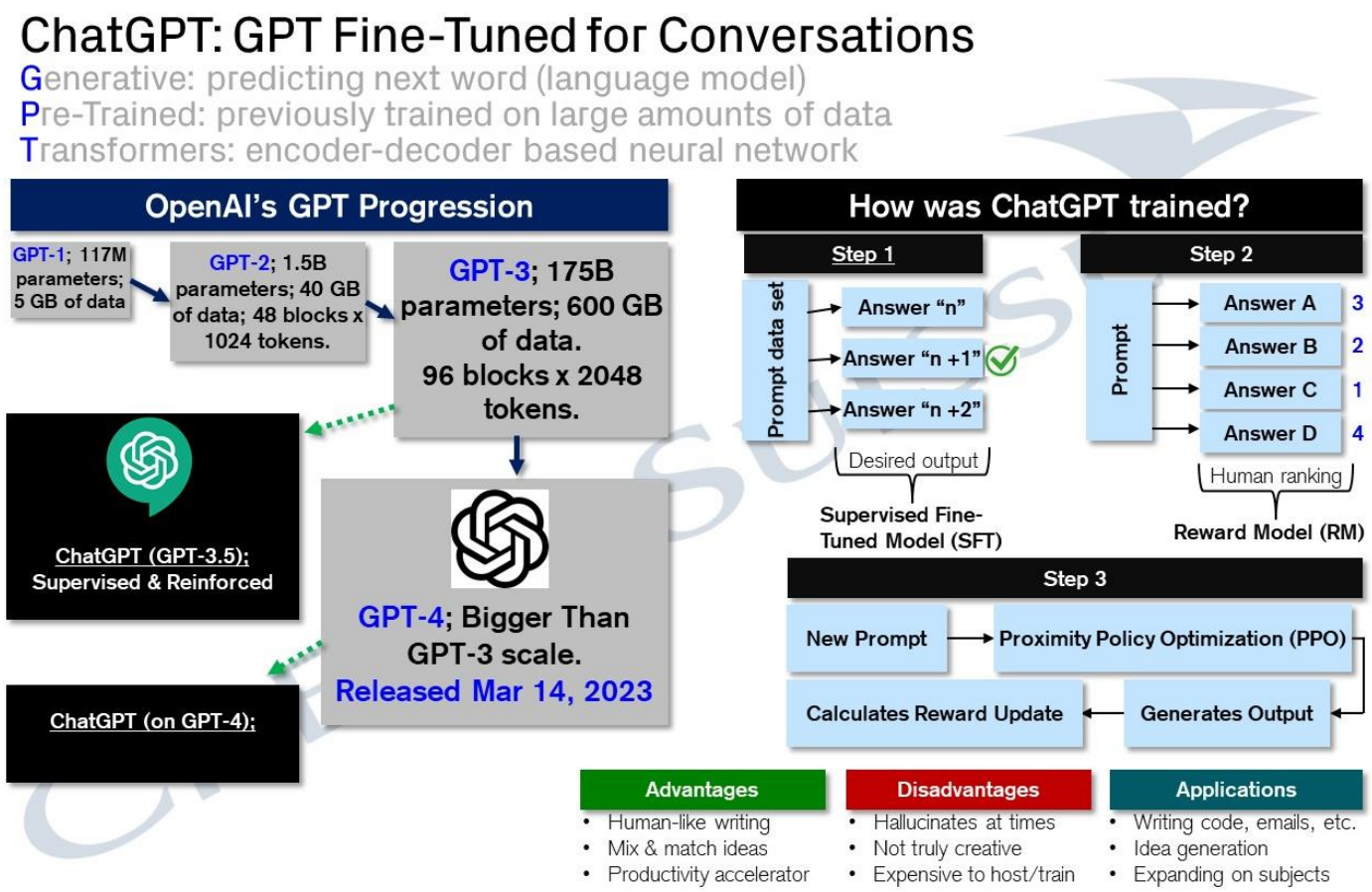
In summary, large LLMs that require too much data, too much scale, and too much training do not always result in good models (this is especially highlighted in Figure 15). Take for example the Google Chinchilla LLM model that we highlight in Figure 16 which uses comparable compute as GPT-3, but it has less model loss, which means it is more accurate.

Figure 5: Illustration of Input Context and Output for an LLM

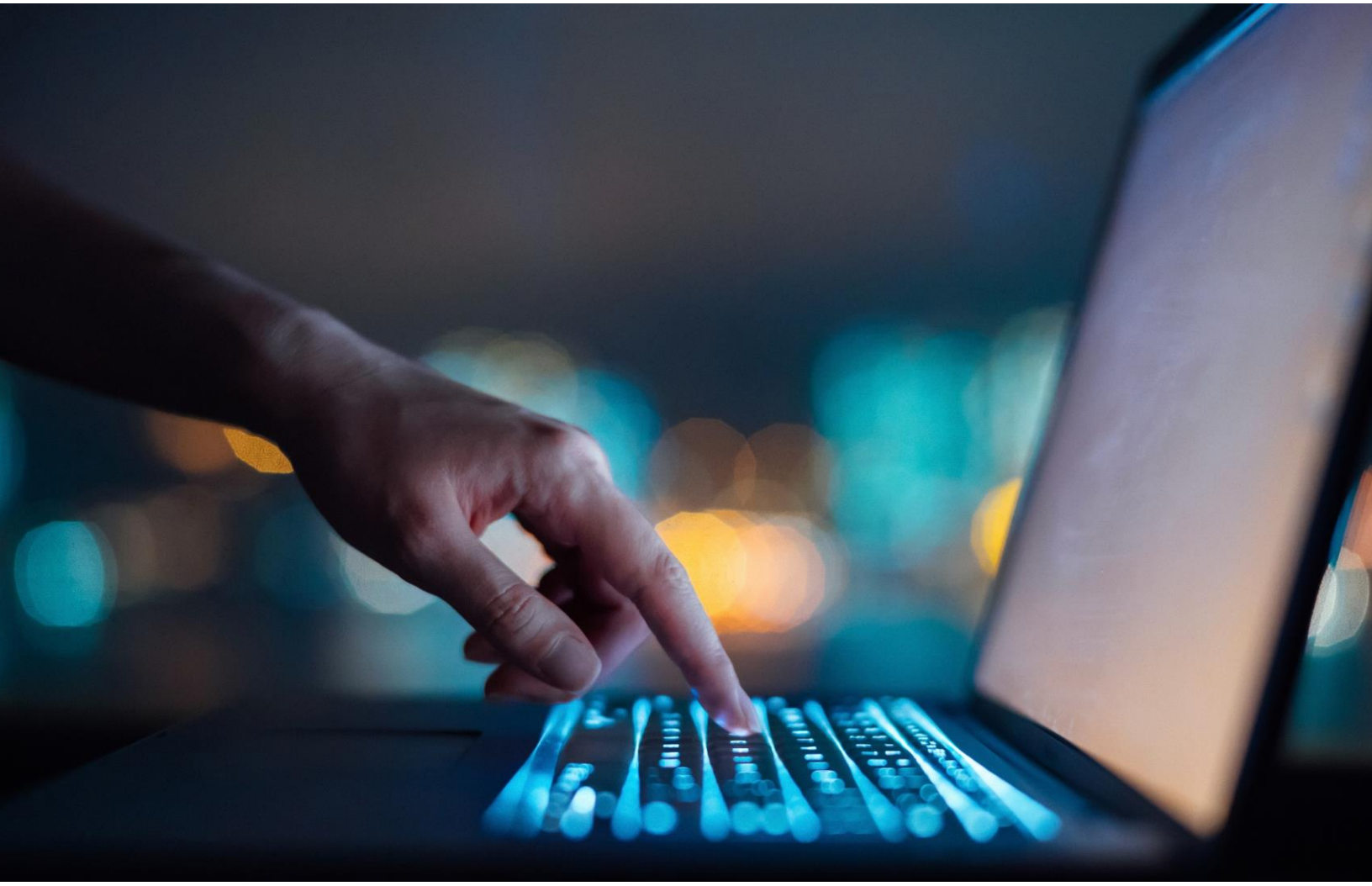


Source: Credit Suisse.

Figure 6: Cheat Sheet on ChatGPT's Progression, Training, Advantages, Disadvantages, and Credit Suisse Expectations



Source: Company data, Credit Suisse.



“ChatGPT is the fastest growing app, ever.”

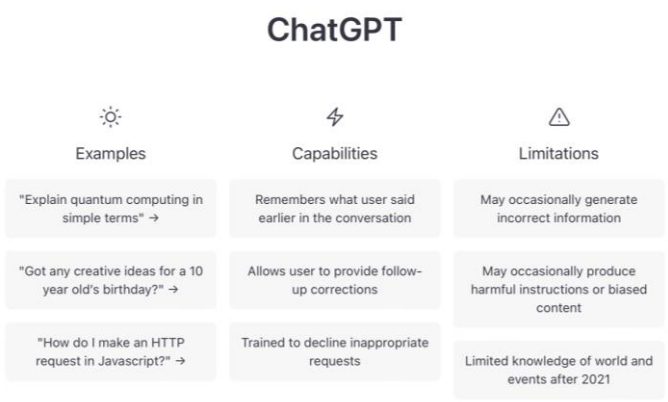
ChatGPT Is a Tool

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.

ChatGPT Was Developed by OpenAI, An AI Institute

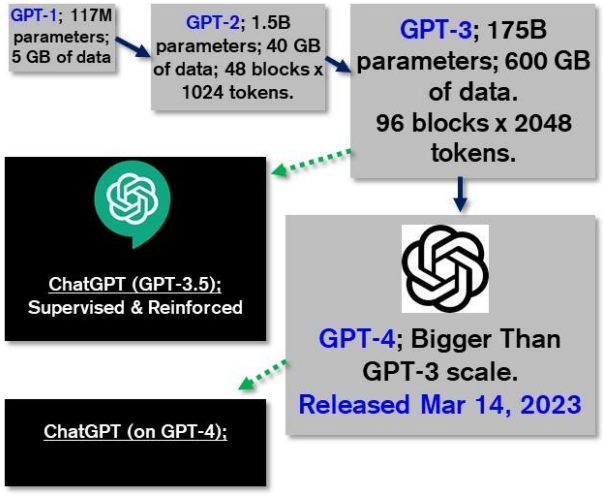
OpenAI was founded in December 2015 as a non-profit with a $1B commitment from Sam Altman, Greg Brockman (former CTO of Stripe), Elon Musk, Reid Hoffman, Jessica Livingston, Peter Thiel, Amazon Web Services (AWS), Infosys, and YC Research–the company's co-chairs at the time were Sam Altman and Elon Musk (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 7: ChatGPT Intro Screen



Source: OpenAI.

Figure 8: Evolution of OpenAI's GPT



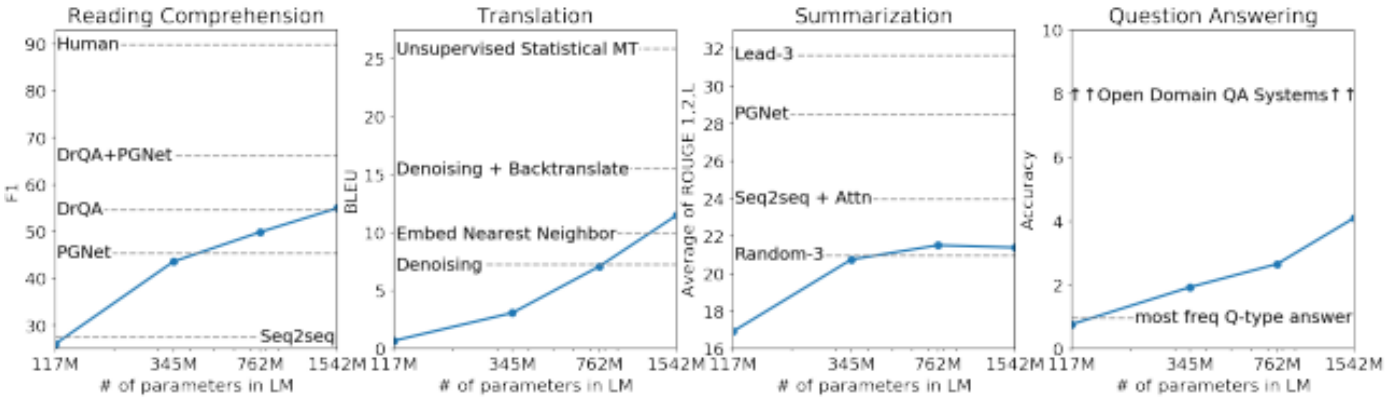
Source: Company data, Credit Suisse. *CS estimates within the next 6-18 months.

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 9).

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 9: OpenAI GPT-2 Model Performance by Task – Parameter Count Comparison

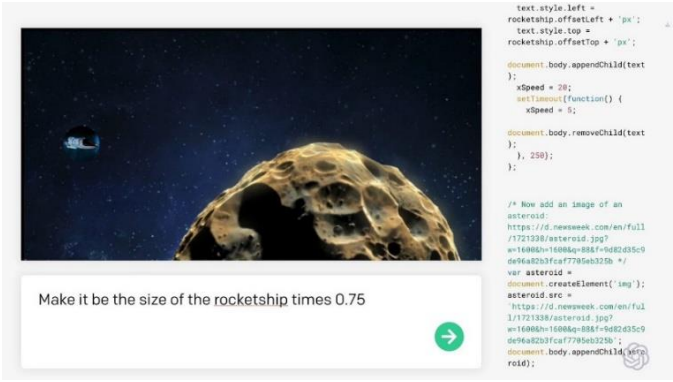


Source: OpenAI.

Figure 10: OpenAI - DALL·E 2

Source: Company data, Credit Suisse

Figure 11: OpenAI – Codex Demo



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 13), 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 12), 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.

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 14 for a preliminary pricing schedule for Foundry.

OpenAI recently released GPT-4, the successor model to GPT-3, with improved performance and multimodality (the ability to handle not only text but also images) with GPT-4 more reliable, creative, and able to handle more nuanced instructions compared to predecessor versions. Additionally, GPT-4 was built with a heavy emphasis on safety and reliability, according to OpenAI's analysis GPT-4 is 82% less likely to respond to requests for disallowed content and 40% more likely to produce factual responses than GPT-3.5.

Figure 12: 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 13: GPT-3 Consists of a Family of Models Rather Than a Single Model

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

Source: OpenAI GPT-3: Everything You Need to Know

Figure 14: 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

In addition to the parameter count of GPT-4, the size of the training data set is also a key variable where size is not necessarily correlated with a higher quality LLM. Sam Altman (CEO of OpenAI) noted that deployments of smaller data sets can be superior because they are: 1) more cost effective, 2) require fewer computing resources, and 3) have simpler implementations. While much of the focus around GPT-4 has been around model parameters, 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, using additional training data adds cost, complexity and requires additional compute resources.
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 the model to be less sensitive to prompt variations, mentioned earlier.

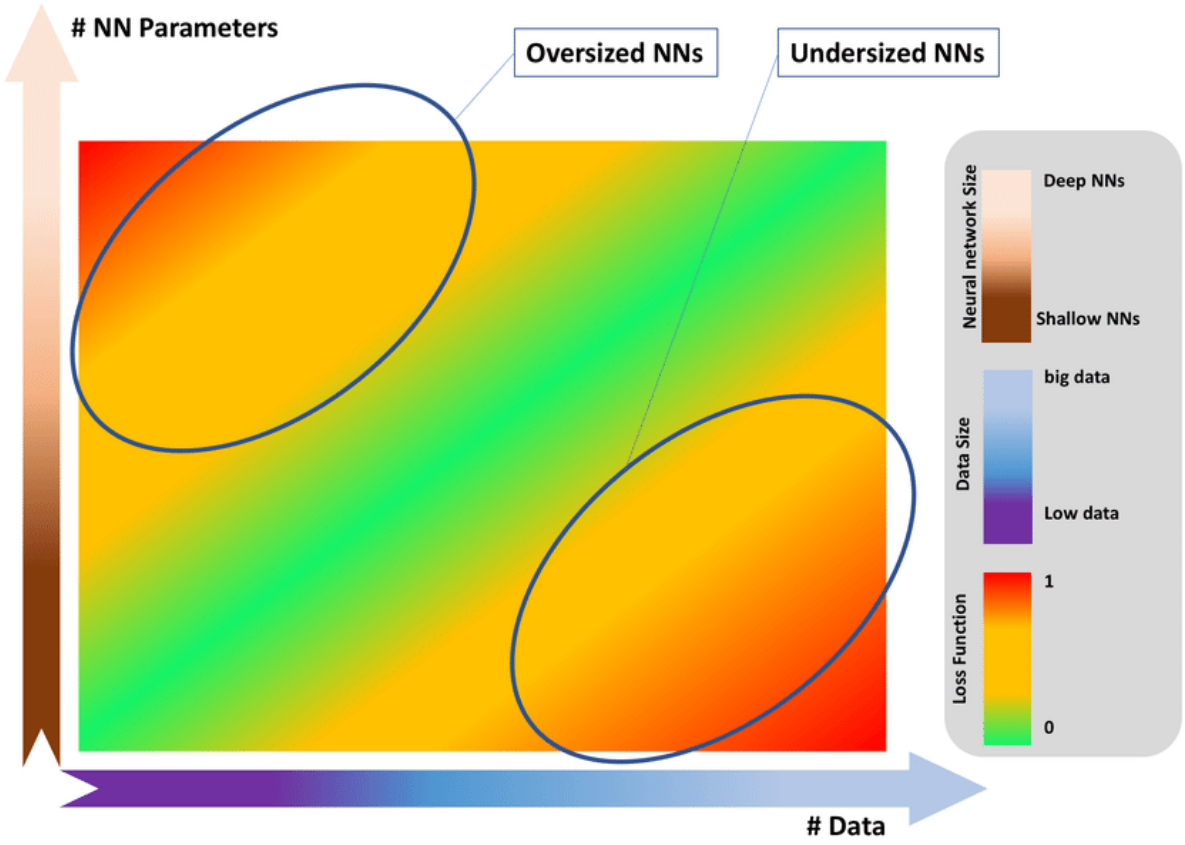
As can be seen in Figure 15, optimized LLMs take a balanced approach to data and parameters and a relatively higher mix of either can significantly increase model loss. Striking the optimal balance between these two variables is the key to a highly capable LLM.

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.

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.

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

Figure 15: Parameters Versus Training Data – Identifying Model Loss



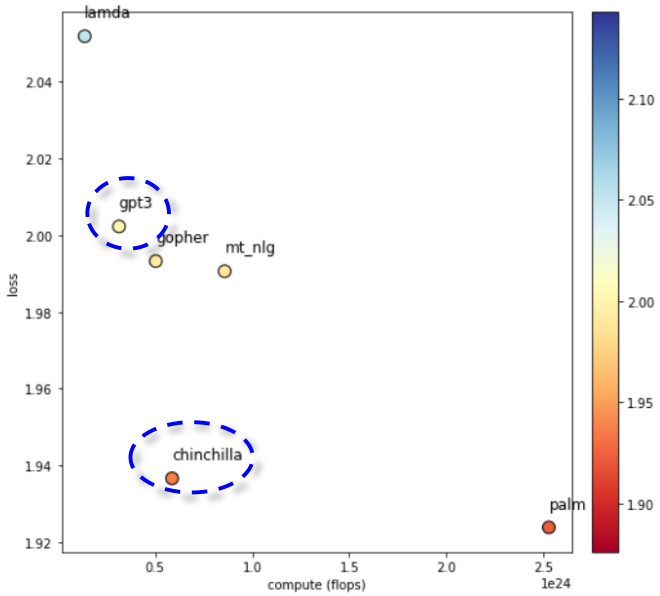
Source: ResearchGate.

model loss would be to train the model 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 16: Model Loss vs Training Compute Resources



Source: Company data, Credit Suisse, Lesswrong.com

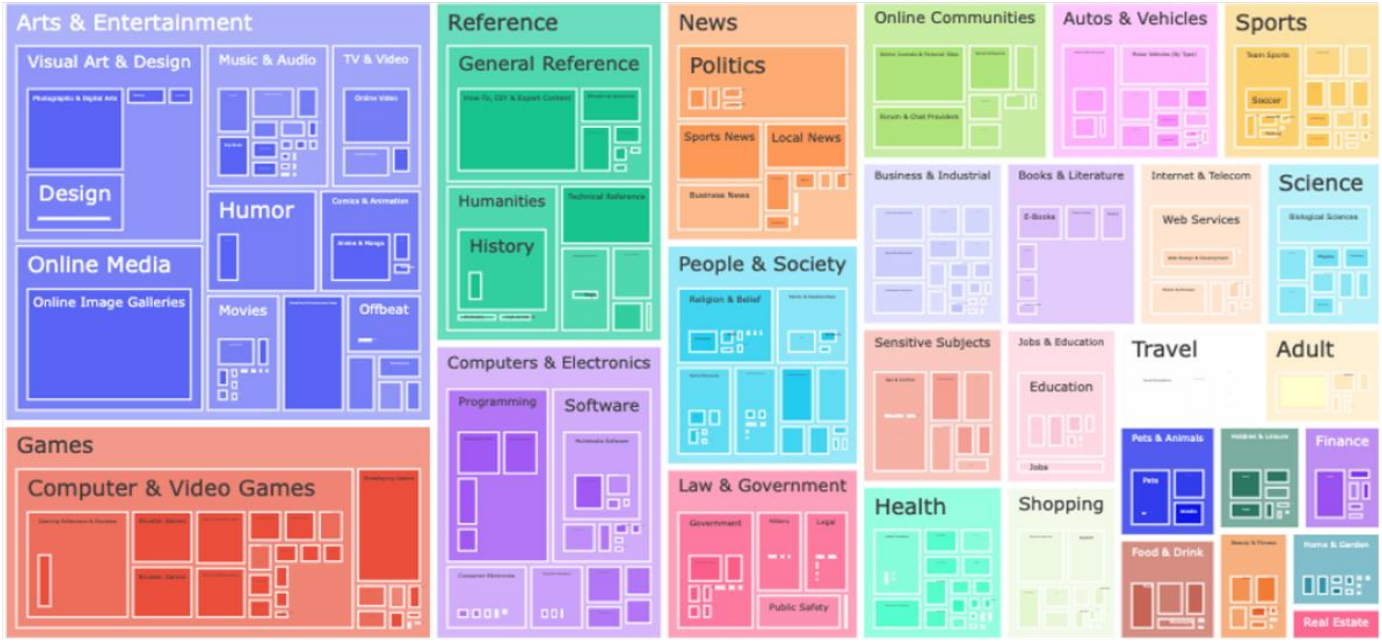
Figure 17: Model Loss – Parameter vs. Training Data Comparison

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

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

Source: Substack.

Figure 18: 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 19, 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.

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 22), 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 20, 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 21). **This suggests the model training costs are negligible over the life of the search engine.**

Figure 19: 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.

- **Cost of Inference vs Cost of Training:** Both the 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.
- **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 22).
- **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.

Figure 20: 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		
Training Cost/1000 Tokens	**$0.0047**	**$0.0075**	**$0.0141**	**$0.0144**
Train Tokens	300B	300B	270B	780B
Training Costs	**$1,398,072**	**$2,236,915**	**$3,810,744**	**$11,216,529**

Source: Substack-Sunyan.

Figure 21: 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%)*
Net Reduction vs. GPT-3 in 2020	**83%**	**81%**

Source: Substack-Sunyan.

Figure 22: SemiAnalysis Estimates Current Incremental Costs per Query at $0.36

Google Search Cost Waterfall					
Metric	**2022 Google Search (CS Est.)**	**ChatGPT Additional Cost**	**350 Tokens Per Search**	**20% Navigational or Cached**	**With TPUv4**
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		**$0.0036**	$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

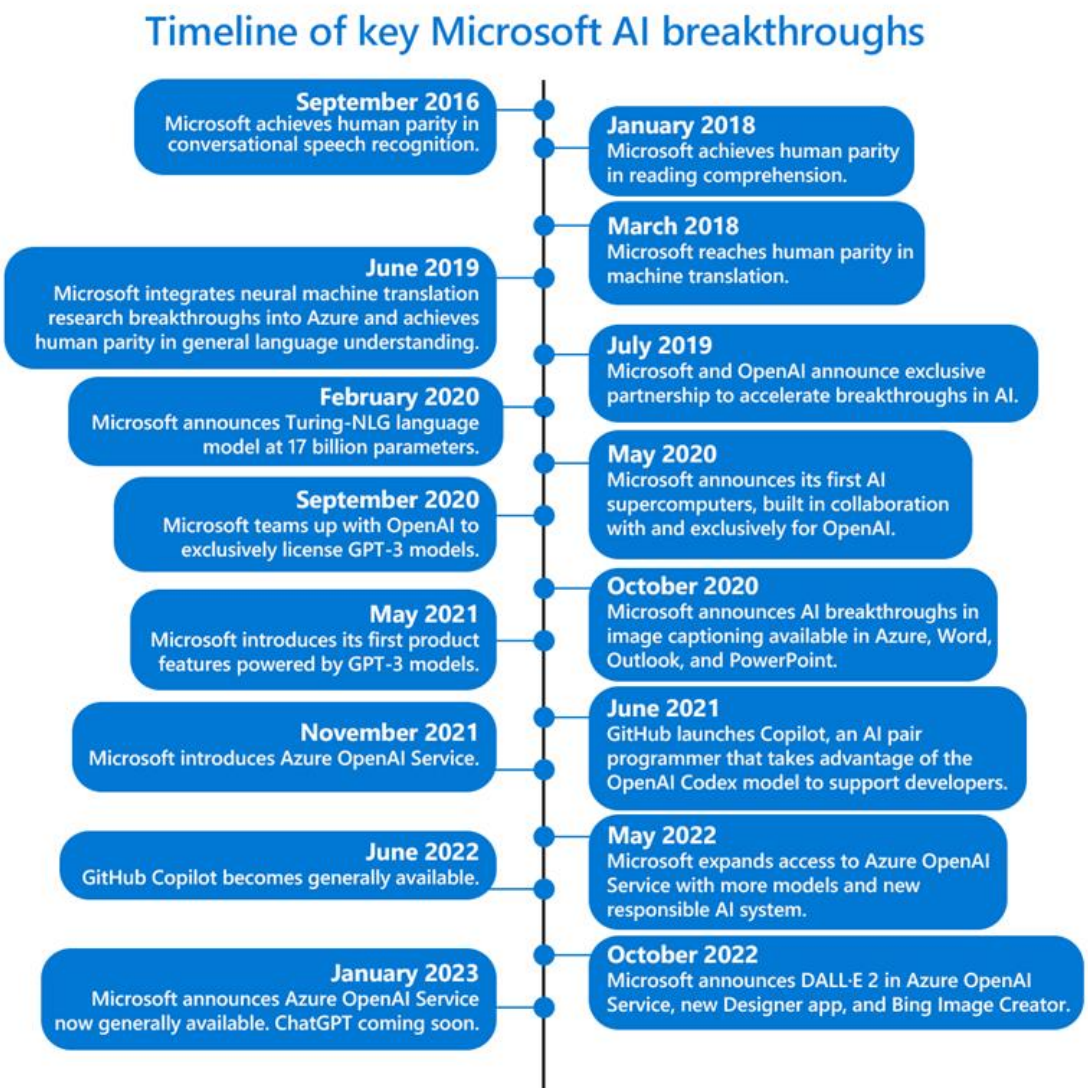
Source: Credit Suisse Estimates, SemiAnalysis.

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.

- **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.

Figure 23: Timeline of Key Microsoft AI Developments



Source: Company data.

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.

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 #1,879 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.

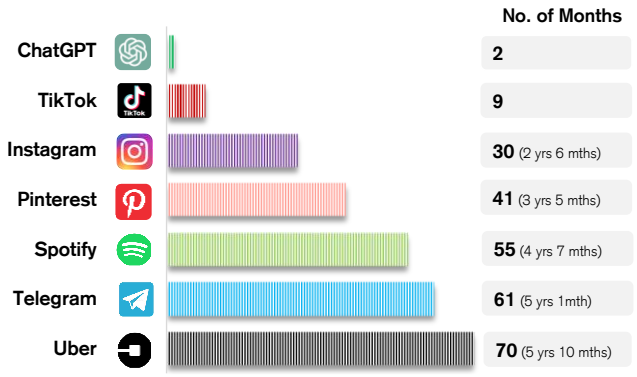
ChatGPT CS Revenue Forecast & Model

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

Figure 24: ChatGPT – Time to Reach 1M Users Comparison

Source: Company data, Statista, Credit Suisse

Figure 25: ChatGPT – Time to Reach 100M Users Comparison



Source: Company data, CNBC, Credit Suisse

Figure 26: Similarweb – OpenAI Traffic Data

Source: Similarweb

Figure 27: Similarweb – OpenAI Global Ranking by Month

Source: Similarweb

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.

- **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 28: CS – Illustrative ChatGPT Revenue Forecast Model

(in Millions)	Dec-22	2023	2024	2025	2026	2027
ChatGPT Revenue		**$205**	**$1,055**	**$2,469**	**$3,337**	**$4,339**
y/y growth (%)			*515%*	*234%*	*135%*	*130%*
ChatGPT Revenue - Guidance		***$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%
Paid Users	**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%
Total Monetizable Avg. MAUs	**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%
Average MAUs (M)	**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
MAU Geographic Mix						
United States	**20%**	20%	20%	20%	20%	20%
International	**80%**	80%	80%	80%	80%	80%

Source: Company data, Credit Suisse.

Figure 29: ChatGPT MAUs – CS Net New Monthly MAU Forecast

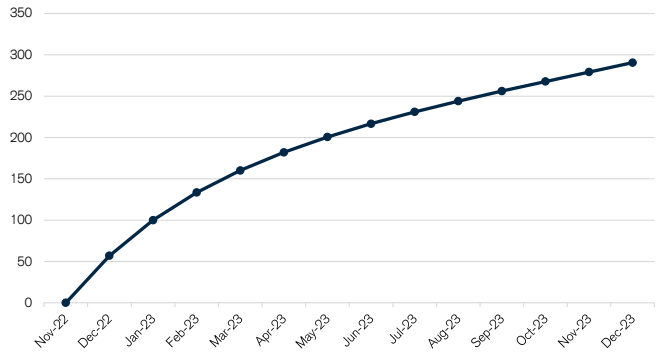
Month	MAUs	Net New Monthly MAUs	Net New - % M/M
Nov-22	0	0	
Dec-22	57	57	
Jan-23	100	43	-24.6%
Feb-23	133	33	-22.3%
Mar-23	160	27	-20.1%
Apr-23	182	22	-17.8%
May-23	201	19	-15.6%
Jun-23	217	16	-13.3%
Jul-23	231	14	-11.1%
Aug-23	244	13	-8.8%
Sep-23	256	12	-6.6%
Oct-23	268	12	-4.3%
Nov-23	279	11	-2.1%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

Source: Company data, Credit Suisse.

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.

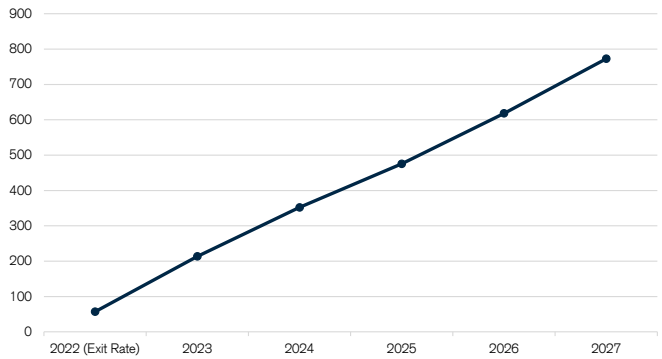
- **Paid Users:** Typical software companies have a free to premium conversion rate of 2-5% (there are notable outliers above this range), for modelling purposes we assume a constant "freemium" conversion at the low-end of the range (2.0%) as we assume the virality & novelty of the product drives higher 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.

Figure 30: CS Est. ChatGPT MAUs–Launch to Year-end 2023



Source: Company data, Credit Suisse.

Figure 31: CS Est. ChatGPT MAUs – Launch to 2027



Source: Company data, Credit Suisse.

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%.
- 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 32: CS ChatGPT 2027 Revenue Forecast Sensitivities – Conversion/Growth

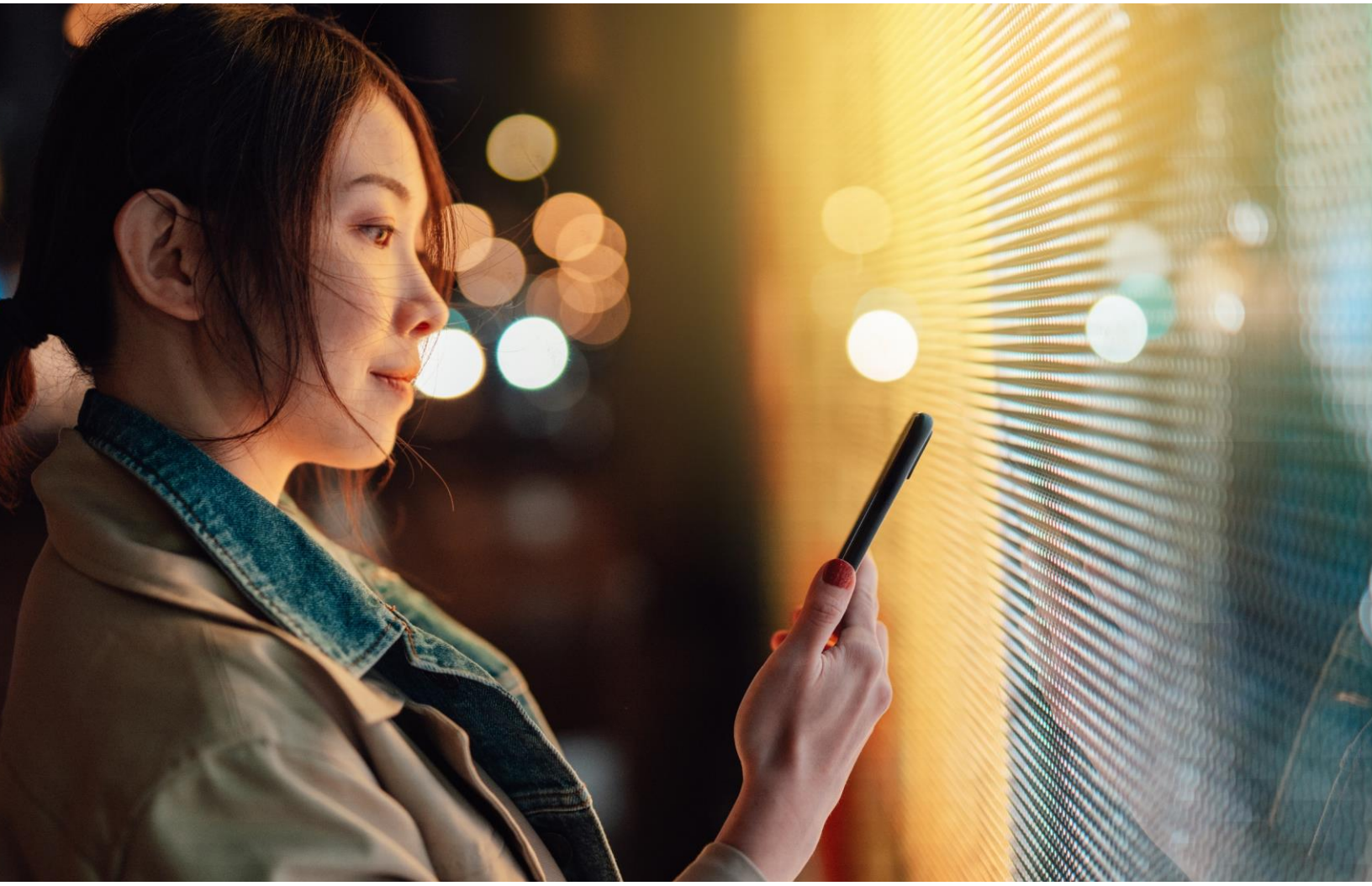
Avg. Annual MAU Growth (2023+)	Free to Paid Conversion (2023+)					
	1.00%	**1.50%**	**2.00%**	**2.50%**	**3.00%**	**3.50%**
18%	$1,160	$1,740	$2,320	$2,900	$3,480	$4,059
28%	$1,606	$2,409	$3,212	$4,015	$4,817	$5,620
38%	$2,170	$3,254	**$4,339**	$5,424	$6,509	$7,593
48%	$2,870	$4,305	$5,740	$7,175	$8,610	$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

Figure 33: 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



“The GPT and AI ecosystem is already large.”

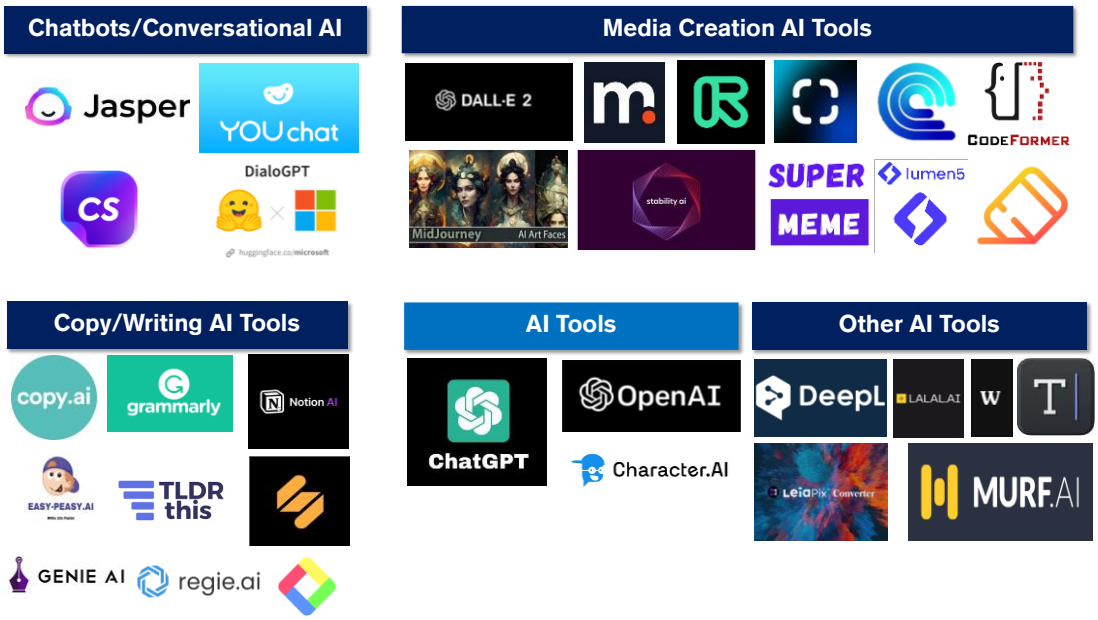
The GPT and AI Ecosystem Is Already Large

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

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

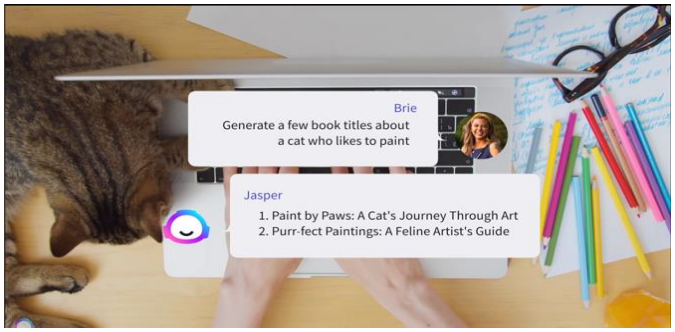
- **Jasper Chat** – Jasper Chat is a conversational chatbot built for businesses like advertising, and marketing, with a more conversational writing experience. Jasper AI is a writing assistant tool that utilizes artificial intelligence and natural language processing to help writers produce more compelling 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.

Figure 34: The Private AI Ecosystem Already Includes a Robust List of Players



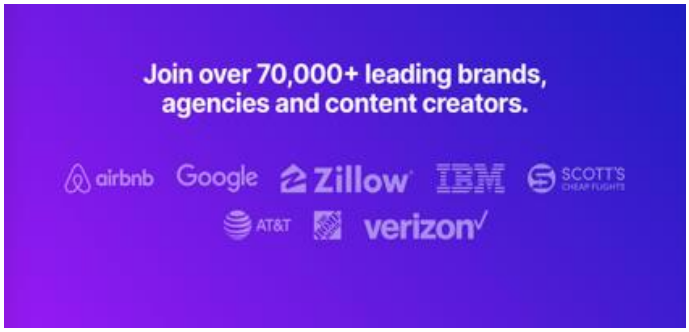
Source: Company data.

Figure 35: Example of Jasper Chat from Product Demo



Source: Company website.

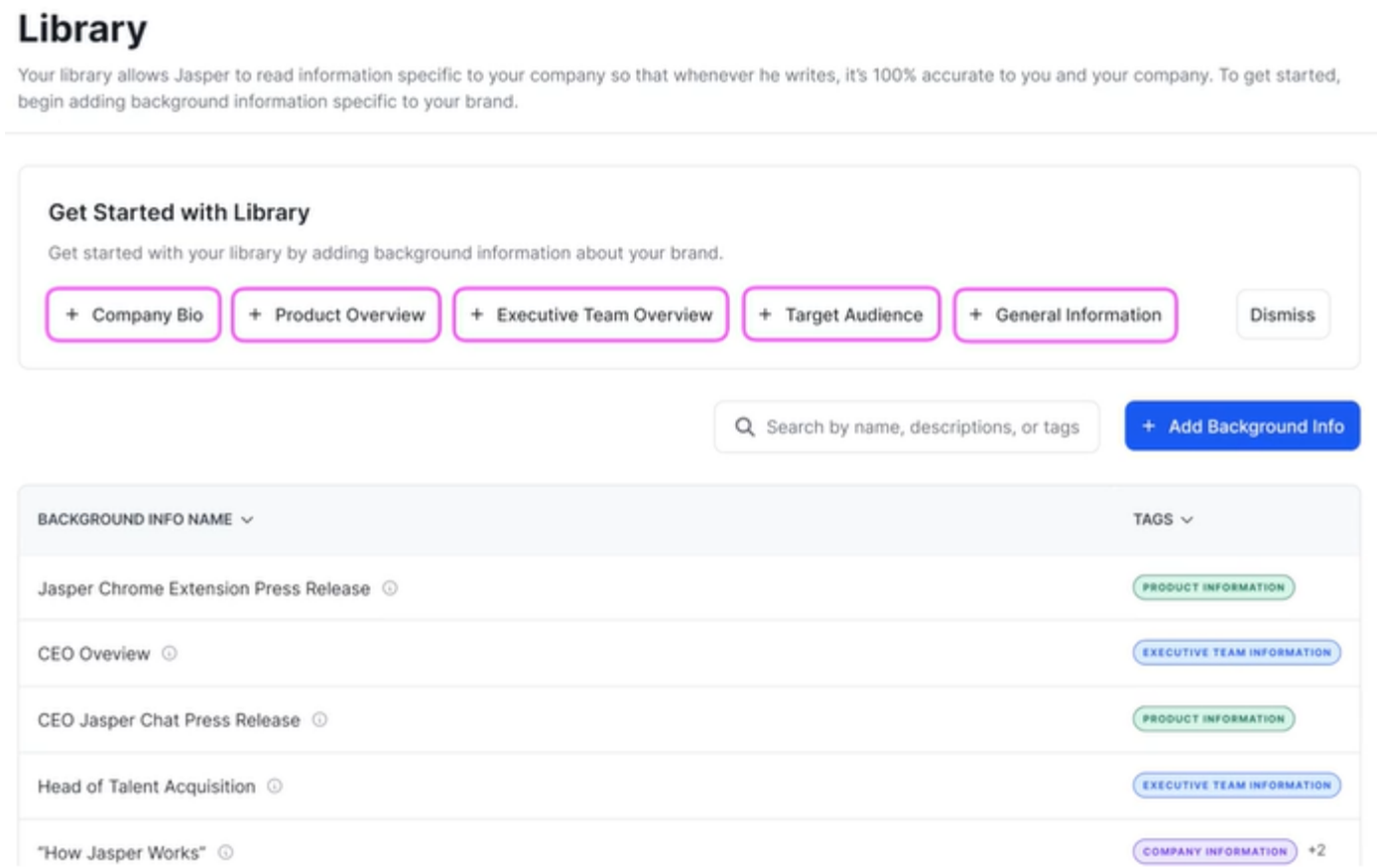
Figure 36: More Than 70K Brands, Agencies and Content Creators Are on Jasper



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 37). **(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.
- **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.

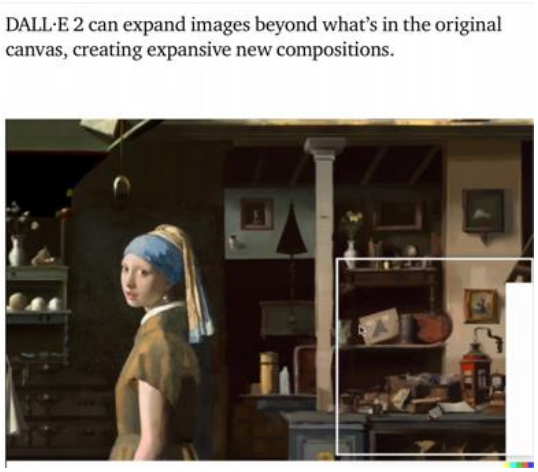
Figure 37: By Training Using Language, Tone, and Style of Each Brand, Jasper Brand Voice Has Enabled Brands to Speak to Its Audience in an Identifiable Way



Source: Company data

- **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.
- **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 38: DALL.E.2 Can Expand Beyond the Original Artwork

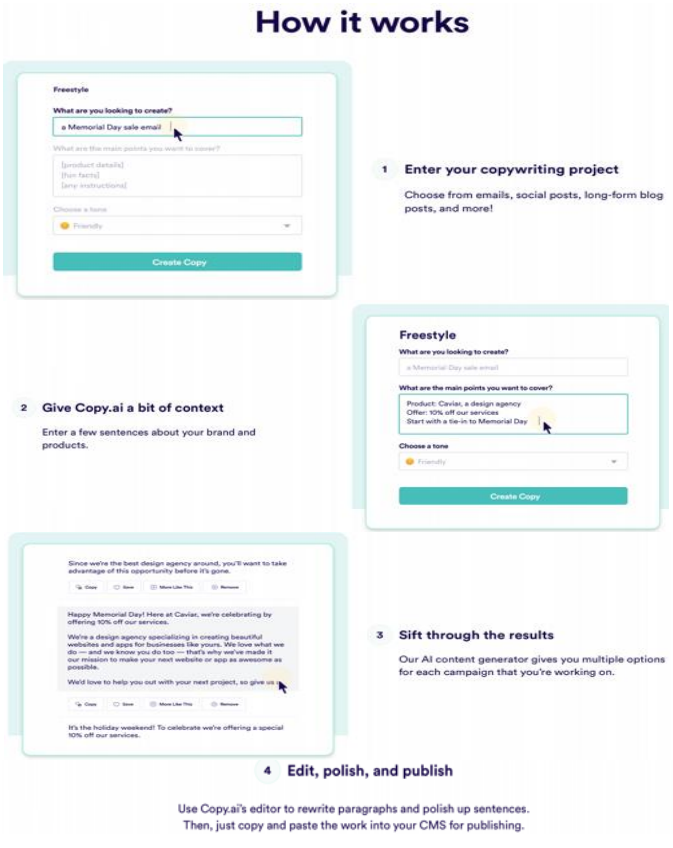


Source: Company Website.

Figure 39: DALL.E.2 Can Also Create Different Variations Based on the Original Art Piece

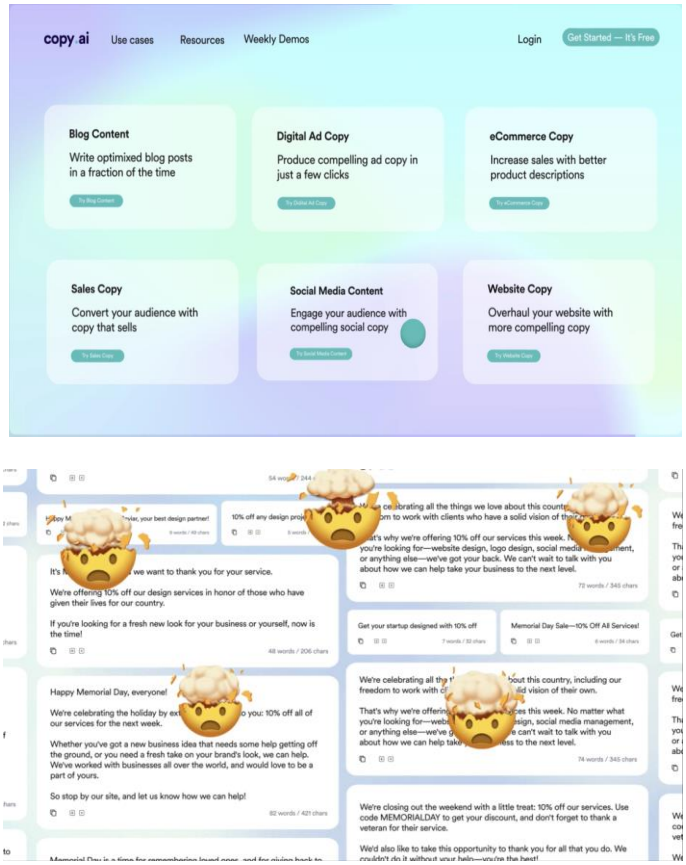
Source: Company Website.

Figure 40: Copy.ai Helps Users to Write Based on Detailed Context and Descriptions



Source: Company Website.

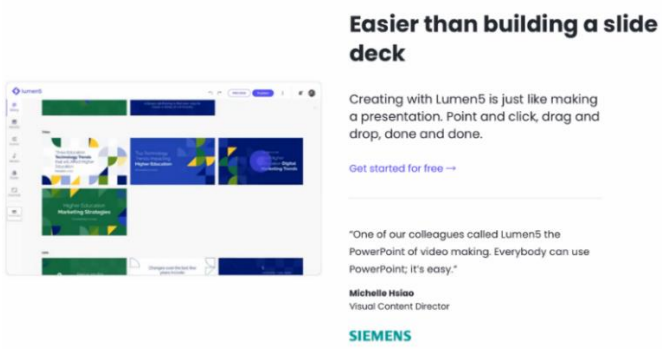
Figure 41: Copy.ai Covers a Broad Range of Topics and Can Generate Brainstorming Results Instantly



Source: Company Website.

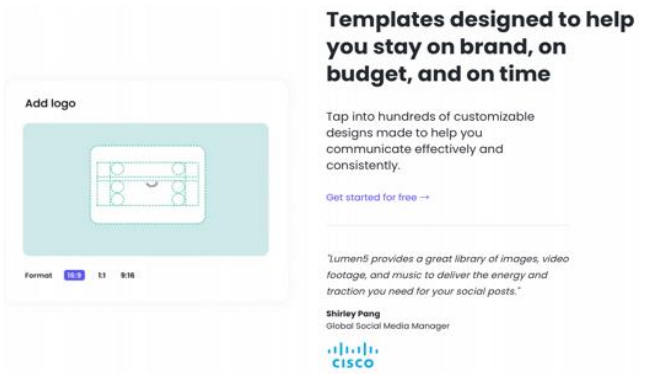
- **TLDR This** – An online article summarizer tool, that automatically extracts author and date information, related images, title and reading time from news articles and blog posts. It also removes ads, popups, graphics, and other distractions to provide a clean and focused reading experience on websites. 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 generated 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.
- **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

Figure 42: Lumen5 Offer AI-powered Video Creator, Enabling Users to Create Video Content Easier Than Building Slides



Source: Company Website.

Figure 43: By Offering 100+ Templates with Pre-designed, Lumen5 Has Streamlined the Video Making Process with Less Budget Required for Users

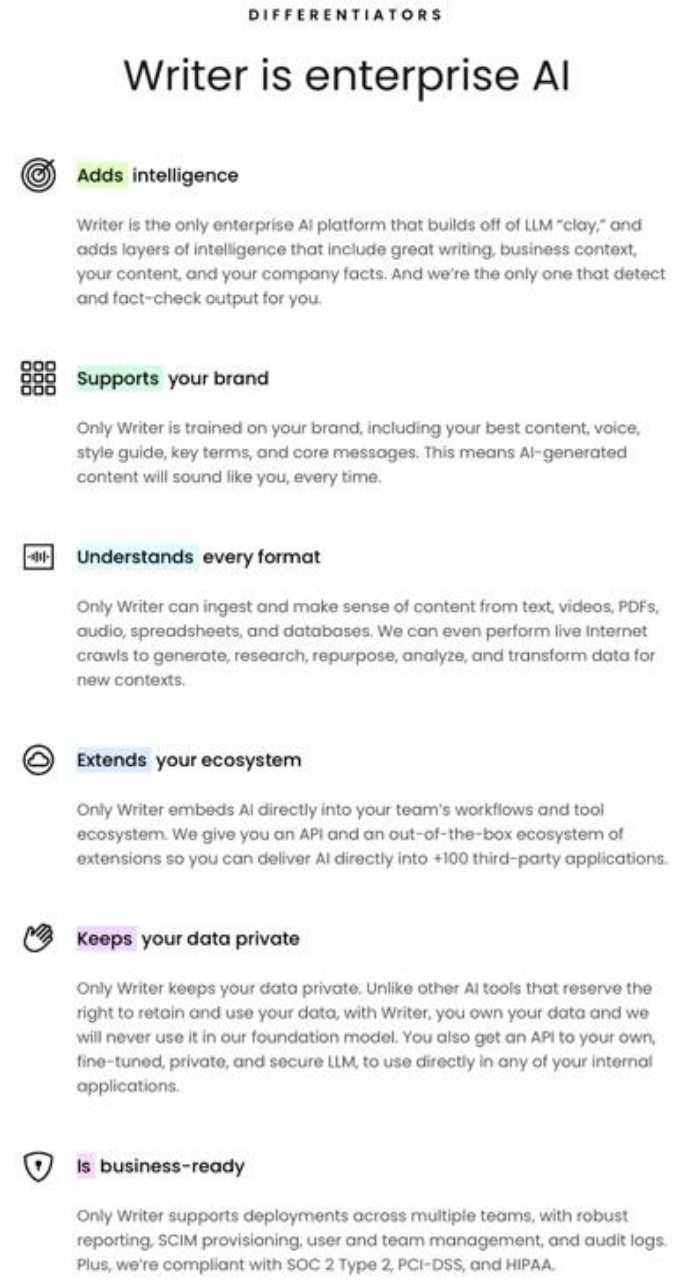


Source: Company Website.

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.

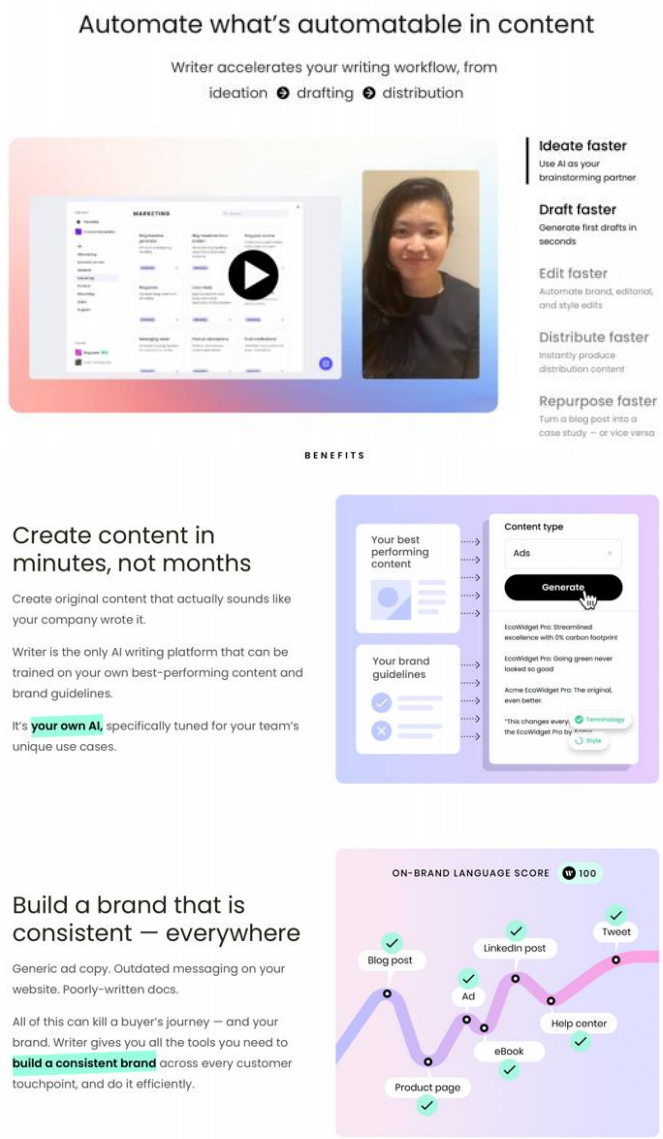
- **Murf.ai** – Murf offers a selection of 100% natural sounding AI voices in 20 languages to provide professional voice overs for videos and presentations. It can also create voice overs with customization features like Pitch, to draw a listener's attention to words or phrases, Pauses, to add pauses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

Figure 44: Key Differentiators of Writer



Source: Company Website.

Figure 45: Writer Helps to Accelerate the Writing Workflow, from Idea Generation All the Way to Distribution



Source: Company Website.

- **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.

“There are some industries that will simply not benefit in the short or even medium term [from ChatGPT]… The opportunities to improve productivity in the Information Technology, Education, Government, and Business Services industries seem to us more clear than in other industries.”

Industries Set to Benefit from ChatGPT

Based on the extensive amount of work 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 teams within the Credit Suisse, providing a cross-sector view on 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

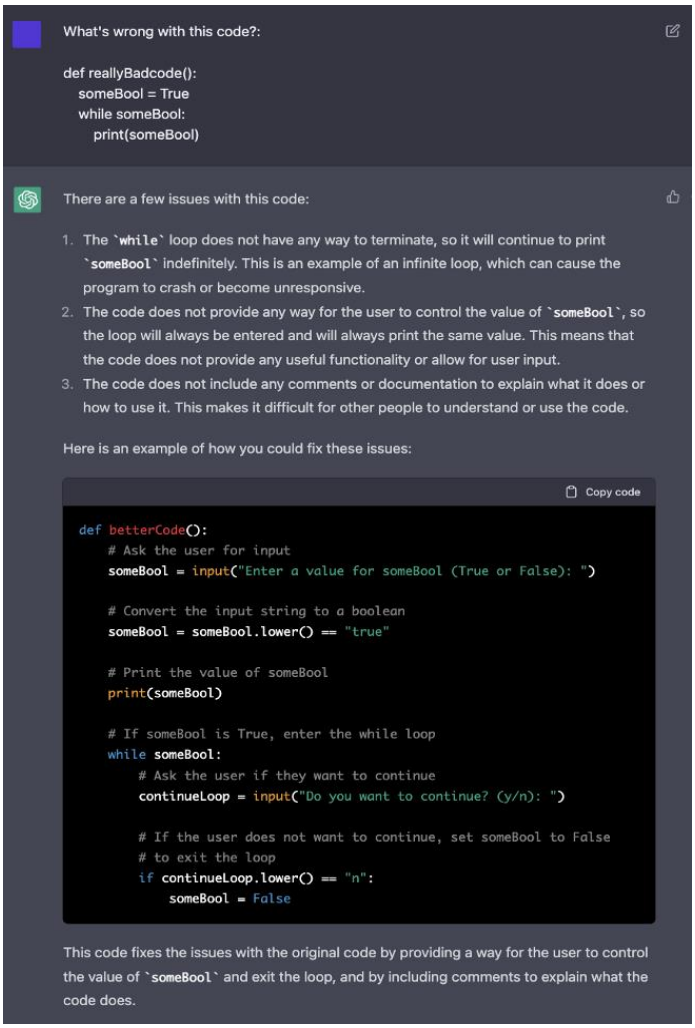
The information technology industry includes the software, internet, IT hardware, networking equipment, and semiconductors sectors. Given ChatGPT is a virtual platform, the two sectors likely to see the most use cases include the software and internet sectors, but other sectors within IT can see benefits 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 48). 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.

Key ChatGPT Use Cases in IT

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

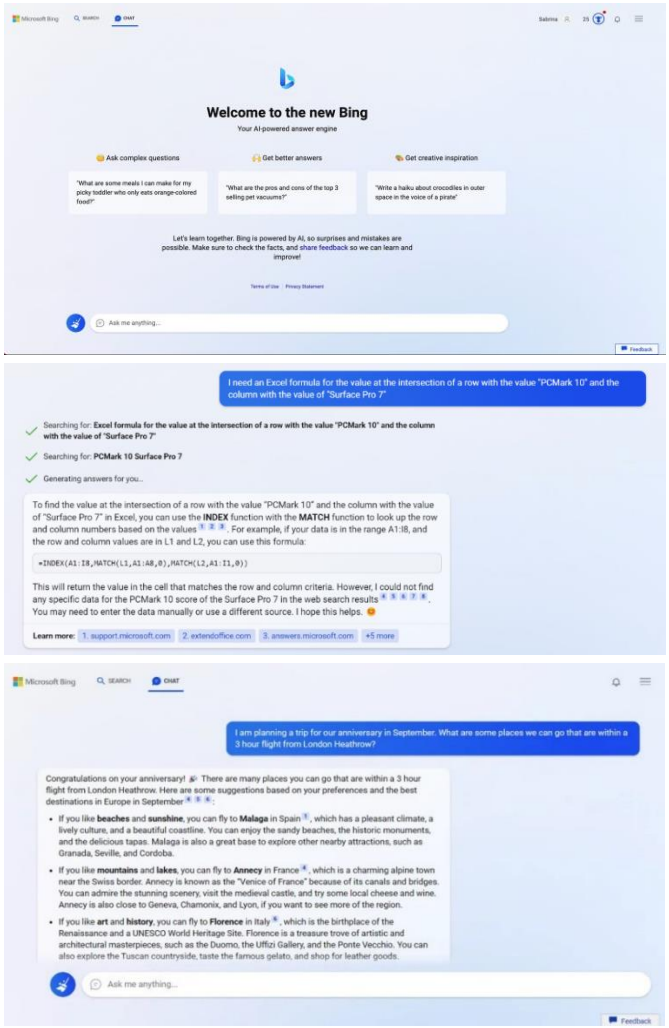
- **Software Development and Programming (Figure 46):** Write, debug, and explain code in various languages. Can also be used to converse in various programming languages and assist in virtually every aspect of software programming. Prior to the development of ChatGPT, Microsoft's Github Copilot sported similar programming capabilities as ChatGPT.
- **Search Engines (Figure 47):** 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 47, Figure 48):** 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.

Figure 46: ChatGPT Providing Correction to Written Code



Source: ChatGPT, Medium (Richard Gao Posts).

Figure 47: Bing AI Answering Inquiries via Search Prompt Line



Source: BingAI, Endgadget (Mat Smith Posts).

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 48):** 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.

Business Services

In recent 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

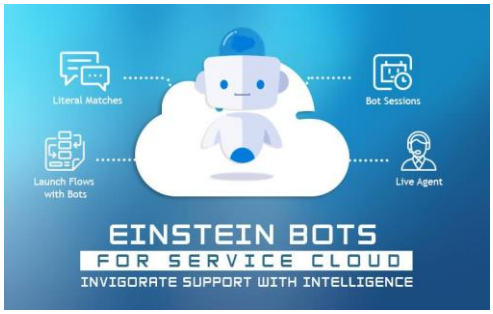
- **Increasing sales productivity:** Sales reps will be able to generate marketing content and formulate outreach messages more efficiently.
- **Increasing innovation:** Develop new products faster as the development process is shortened to modifying ChatGPT-written code as opposed to writing from inception.
- **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):

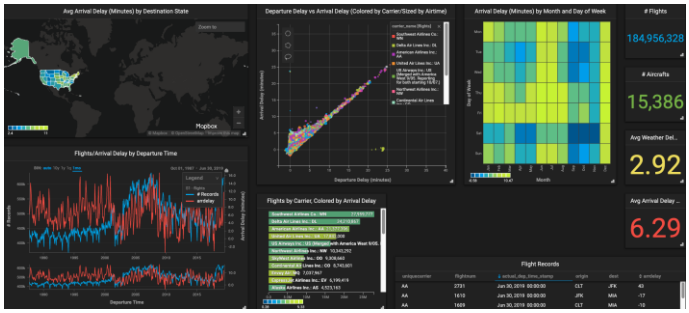
- **Market analysis:** More complex models conditioned to analyze market developments real time amid potentially predictable outcomes.
- **Task replication:** Automation of even more complex worker processes has the potential to drive further efficiencies and fuel higher labor cost savings.
- **Candidate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

Figure 48: EinsteinGPT Is Salesforce's Bot Tool and AI



Source: Company data (link).

Figure 49: AI Analytics and Dashboards Likely to Gain Traction



Source: Heavy.AI.

Key ChatGPT use cases in Financial Services include

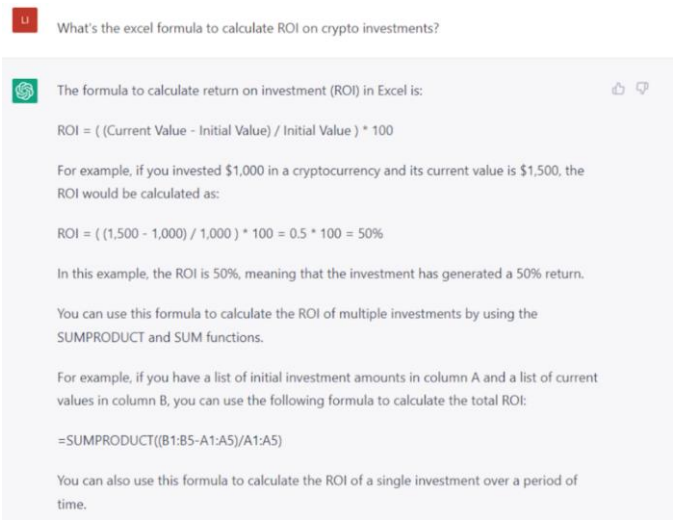
- **Extract Financial Information:** Users can extract information from documents to quickly analyze and process customer inquiries, and generate summary financial reports.
- **Introductory Analysis for Bankers:** Financial service professionals can use ChatGPT for initial overviews of company business 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.

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



Source: Credit Suisse.

Figure 51: Improved Fraud Detection Would Save Consumers Over $6 Billion, Before Accounting for Business Losses

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

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.

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.

Figure 52: 20 Ways to use ChatGPT in the Classroom

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

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 poise 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.

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.

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

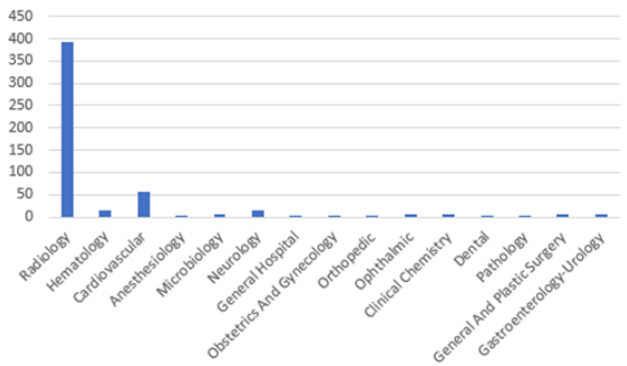
- **Faster, Better Training:** ChatGPT is capable of developing training materials/curriculums for workers, improving training in enterprises that otherwise offer few materials.
- **Employee Chatbot:** Can function as an internal chatbot used to answer questions for workers across multiple divisions.
- **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):

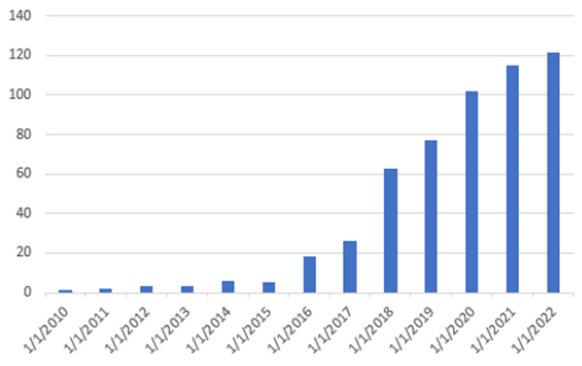
- **Improve Machine Reliability:** Predict assembly issues and machine equipment effectiveness with real-time machine learning algorithms and insights, reducing failure rates and minimize production stoppages.
- **Fully Autonomous Equipment:** Future AI may allow for fully autonomous equipment that can 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.

Figure 53: Of the >500 FDA Clearances/Approvals for AI/ML-Enabled Medical Devices, the Vast Majority Are in Radiology



Source: Company data, Credit Suisse

Figure 54: Number of FDA Cleared/Approved Medical Devices Tagged as Having Some AI/ML Component Has Risen Rapidly



Source: Company data, Credit Suisse

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.

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.

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 identity 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

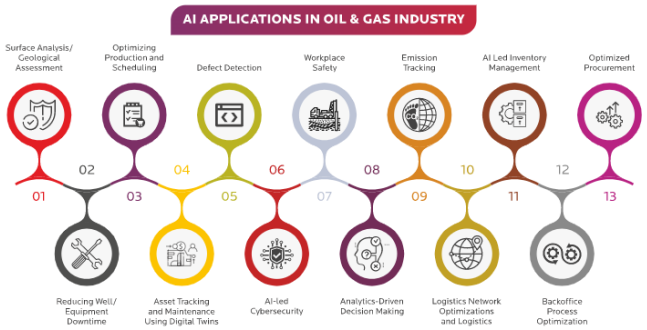
- **On-Demand Field Support:** ChatGPT can be used to provide on-demand field support for both in-field oil and gas workers along with back-office support functions.
- **Code Testing:** ChatGPT can be used (and is currently being used by Devon Energy (DVN) for this purpose) to support in-field technicians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

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):

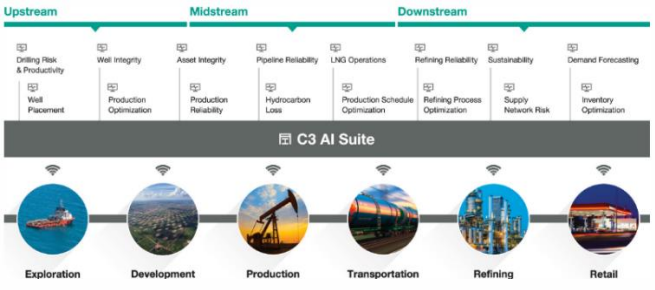
- **Seismic Analysis/Geological Assessment:** The ability to analyze geological data (surface analysis, seismic data, etc.) to improve the hydrocarbon discovery process. This should help identify the highest yielding wells which should reduce the need for additional drilling.
- **Predictive Maintenance:** 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.
- **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.

Figure 55: AI Applications in Energy Industry



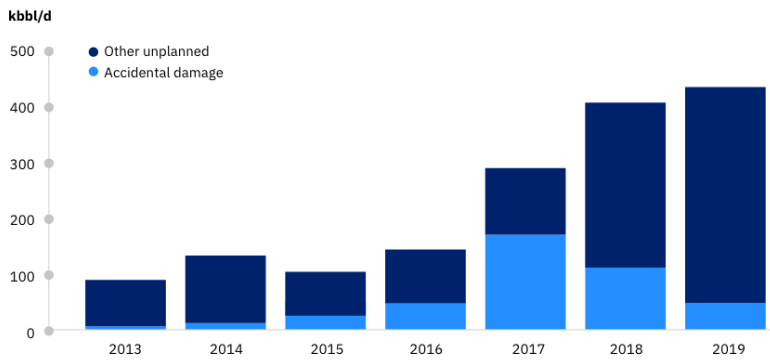
Source: Birlasoft

Figure 56: AI Applications by Energy Subsector



Source: C3.ai

Figure 57: Barrels of Oil Lost to Daily Downtime



Source: Rystad Energy

Utilities

In the Utilities sector, ChatGPT can enable a more streamlined customer experience, help customers with bill inquires, 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.

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 ceate energy profiles from household appliances, smart meters, and other sensors.

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.

Governments

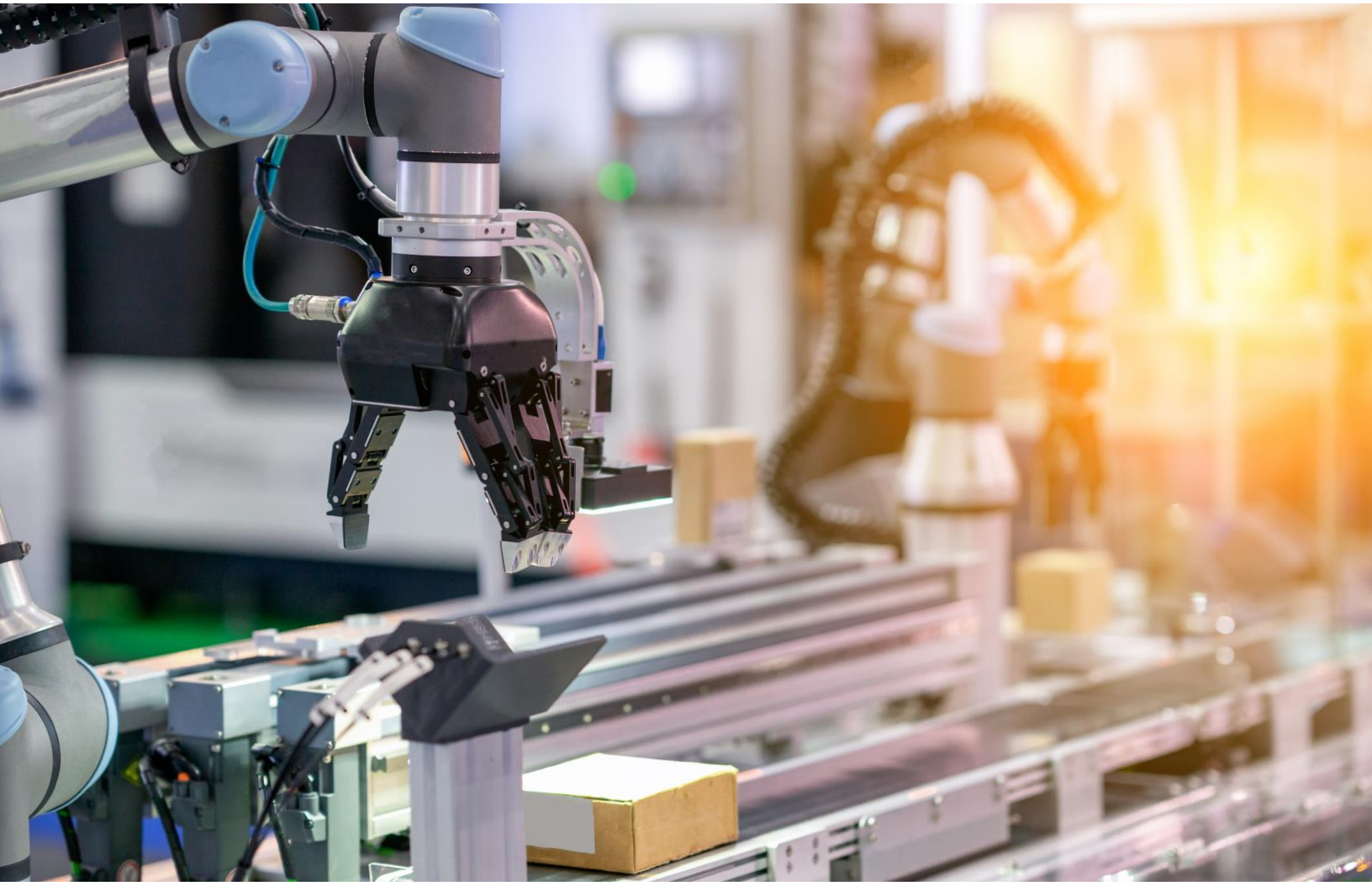
The Government sector is comprised of multiple levels of Government (municipal, state, and federal), along with Public Goods and Services such as the military, law enforcement, public transportation, and infrastructure to name a few. Although ChatGPT is most likely to be used by the branches of government 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. 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.

Key ChatGPT Use Cases in Government Include

- **Public Services:** 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:** 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:** 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.



“AI training and inference are compute intensive tasks that should continue to drive semiconductor advances for compute, storage and transmission of data.”

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

(Chris Caso, Liz Pate, Nicholas Welsh-Lehmann, Randy Abrams, Haas Liu, Angela Dai)

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.

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.

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 58: 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%
Total Data Center	**$57,302**	**$72,368**	**$77,063**	**$71,260**	**$93,411**	**$114,460**	**$115,413**	**8%**	**17%**	**100%**
YoY Growth		**26%**	**6%**	**-8%**	**31%**	**23%**	**1%**			
% of industry	**12%**	**12%**	**13%**	**13%**	**14%**	**16%**	**15%**			
Total Semiconductors	**$470,899**	**$594,952**	**$601,694**	**$562,712**	**$654,326**	**$727,126**	**$753,667**	**6%**	**10%**	
YoY Growth		**26%**	**1%**	**-6%**	**16%**	**11%**	**4%**			

Source: Gartner, December 2022

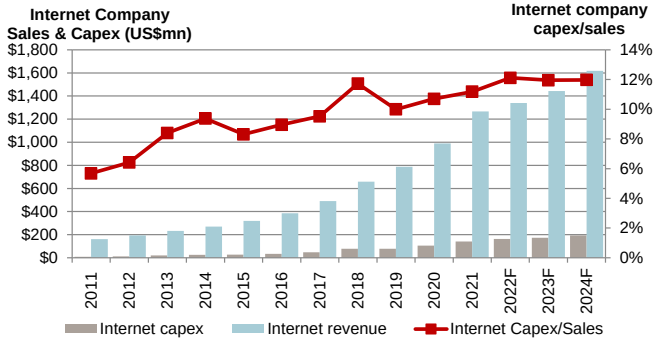
Figure 59: CS Hyperscale Capex Projected at +7%/+12% for 2023-24

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	$2,957
Top 7 Hyperscale capex (US$mn)	**$34,500**	**$46,763**	**$77,207**	**$78,722**	**$105,845**	**$141,530**	**$162,258**	**$172,384**	**$193,674**
YoY Growth	**30.1%**	**35.5%**	**65.1%**	**2.0%**	**34.5%**	**33.7%**	**14.6%**	**6.2%**	**12.3%**
Capex/Sales	**9.0%**	**9.5%**	**11.7%**	**10.0%**	**10.7%**	**11.2%**	**12.1%**	**12.0%**	**12.0%**
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	$711
2nd Tier Hyperscale capex (US$mn)	**$22,487**	**$21,483**	**$23,465**	**$16,675**	**$17,761**	**$20,505**	**$24,290**	**$27,126**	**$28,895**
YoY Growth	**0.6%**	**-4.5%**	**9.2%**	**-28.9%**	**6.5%**	**15.5%**	**18.5%**	**11.7%**	**6.5%**
QoQ Growth									
Total Hyperscale capex (US$mn)	**$56,987**	**$68,246**	**$100,673**	**$95,397**	**$123,606**	**$162,035**	**$186,548**	**$199,510**	**$222,569**
YoY Growth	**16.7%**	**19.8%**	**47.5%**	**-5.2%**	**29.6%**	**31.1%**	**15.1%**	**6.9%**	**11.6%**
QoQ Growth									

Source: Gartner, December 2022

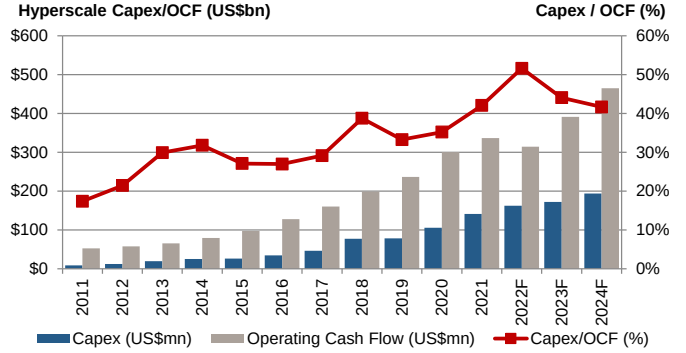
The hyperscalers are noting optimizations for customers doing more with less in the downturn, optimizing hardware spend and moderating their cloud service growth from 30-40% YoY in 2022 to 20-30% YoY in 2023, still well ahead of consumer tech verticals coming off declining sales in 2022 and projected 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 60: Hyperscale Capex/Sales at 12% Is Sustainable



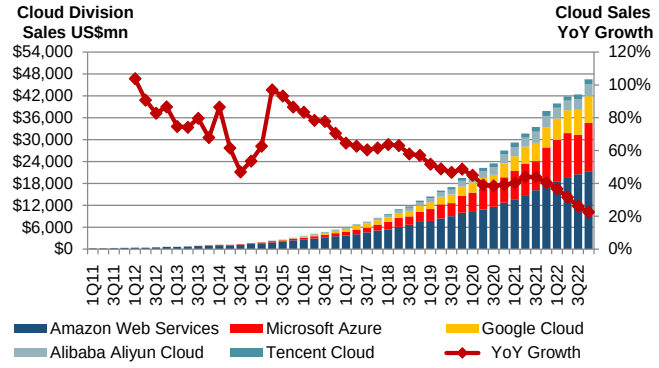
Source: Company Data, Credit Suisse Estimates

Figure 61: Hyperscalers Re-Invest 40% of OCF in Capex



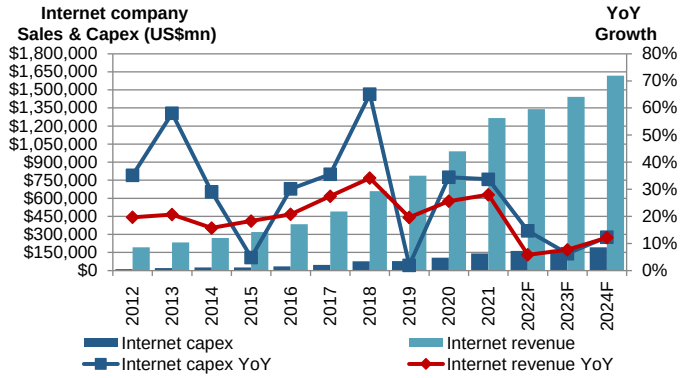
Source: Company Data, Credit Suisse Estimates

Figure 62: Cloud Service Growth Moderating to 20-25% growth



Source: Company Data, Credit Suisse Estimates

Figure 63: Internet Service Growth Has Slowed to Single Digits



Source: Company Data, Credit Suisse Estimates

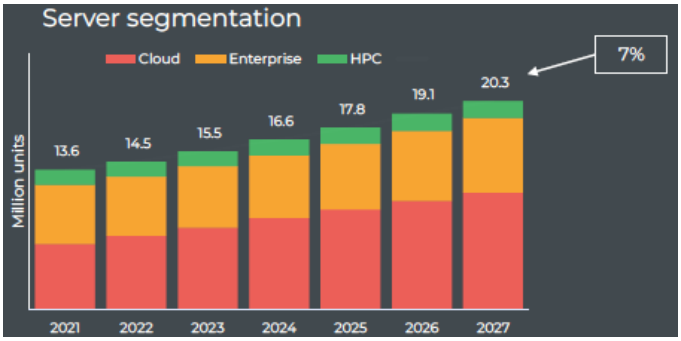
AI investments to be a growing portion of spend

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.

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.

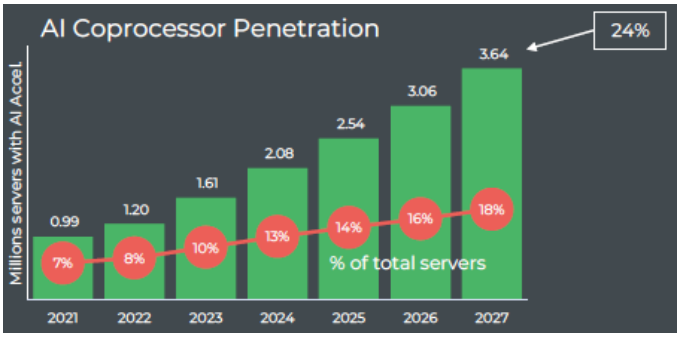
The Yole accelerator data includes all FPGA, ASIC and GPUs and dilutes the increase in pricing seen for NVIDIA's GPU accelerators. List prices for NVIDIA's GPUs have increased with process migration and capability increase from the K series starting at US$5k to V100 (12nm) at US$10k and A100 (7nm) at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

Figure 64: Servers growing at a 7% CAGR through 2027



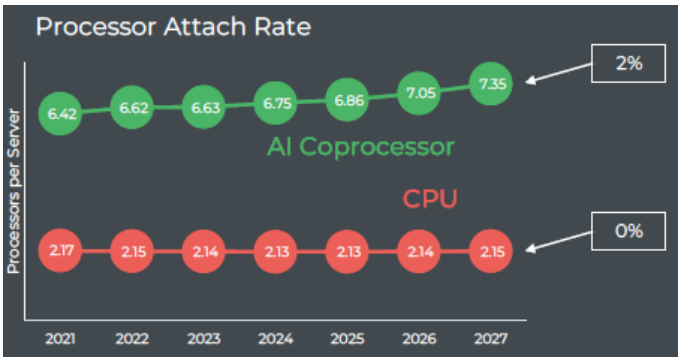
Source: Yole

Figure 65: AI accelerators 24% CAGR rising to 18% of servers



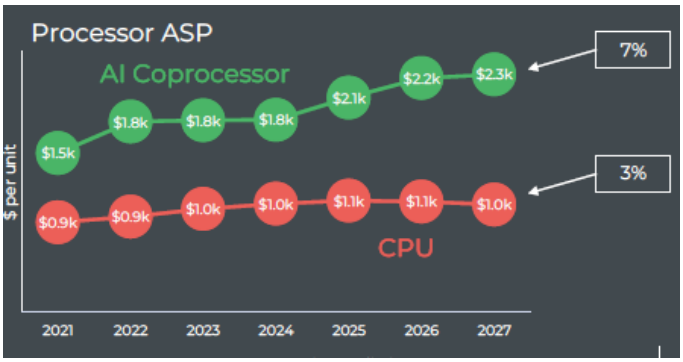
Source: Yole

Figure 66: Attach rate for AI chips gradually increasing



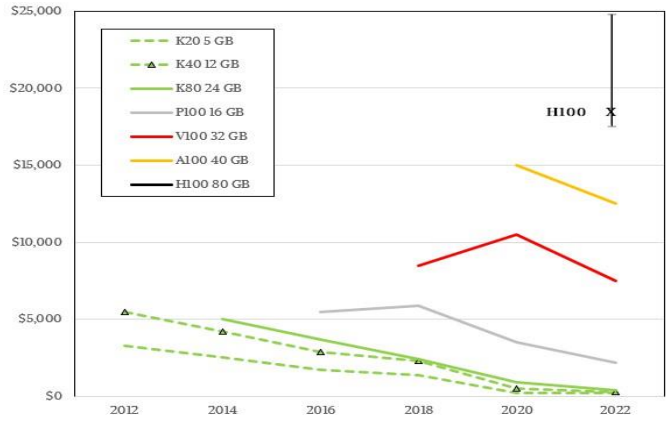
Source: Yole

Figure 67: ASPs rising for the AI accelerators



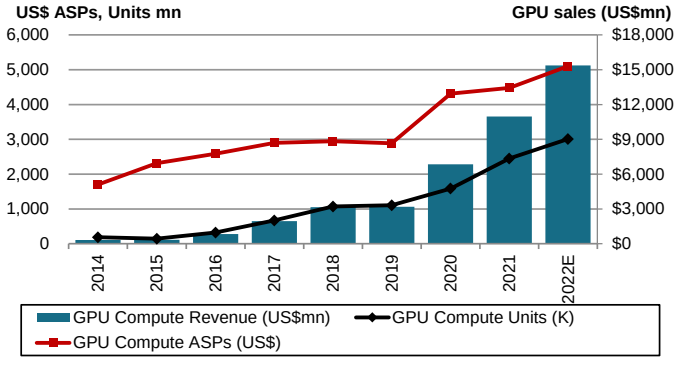
Source: Yole

Figure 68: NVIDIA accelerators continue to increase in price



Source: The Next Platform

Figure 69: GPU server units and pricing ramping



Source: Mercury Research

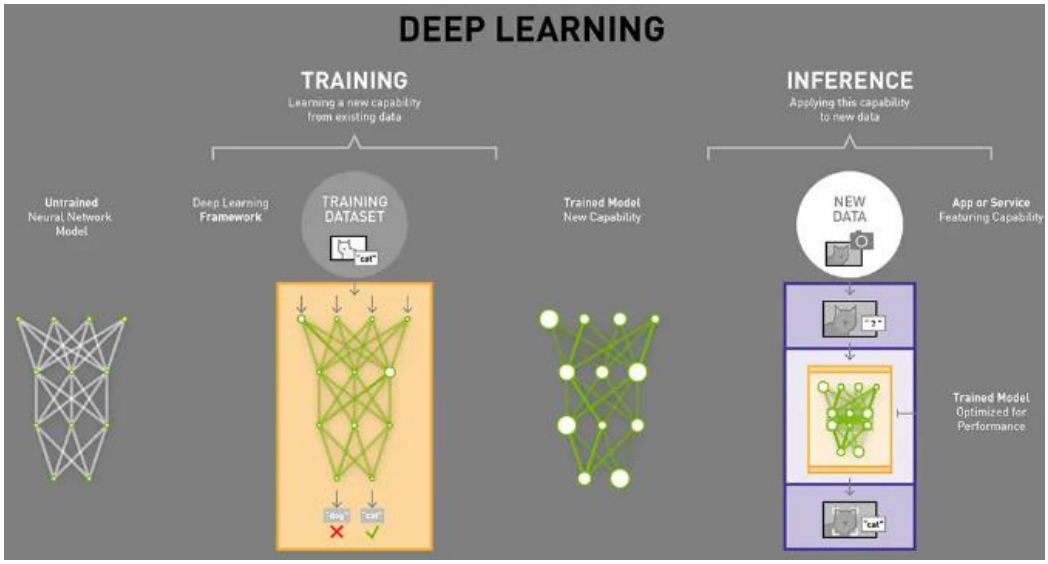
AI Training and inference both driving more compute and shifts to accelerators

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 beings 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.

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 efficiency to make decisions in real time.

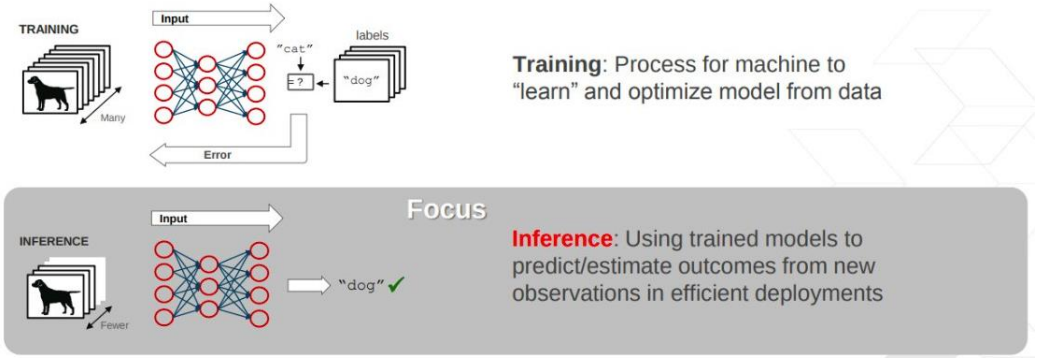
For Chat GPT, the model was trained with data from 175bn parameters including various web sources, books and articles through 2021 to respond to users in a conversational way. The model was trained with supervised fine-tuning from human AI trainers to provide conversations to play both sides of user 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 CPT 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.

Figure 70: Deep learning compute split into training and inference



Source: NVIDIA

Figure 71: Training learns and optimizes an AI model, inference deploys the trained model into action



Source: Xilinx

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.

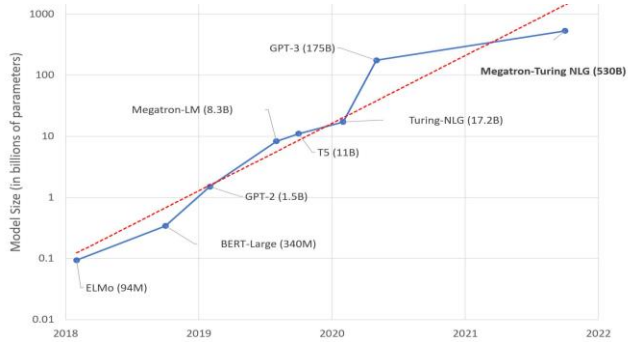
The advance and competition to generate more powerful and accurate large language models is creating a ramp in compute intensity and memory requirements for training and inference for AI accelerated compute and high density memory storage. Facebook in introducing its Llama large language model which 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.

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.

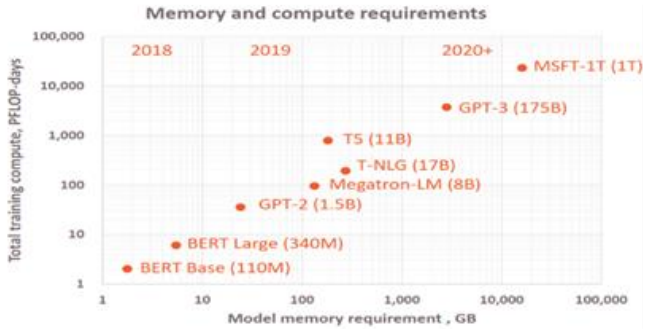
- **CPU (Central Processing Unit):** A CPU has traditionally been strong at running complex instruction sets including running the main application data in PCs and serversbut can be used for some simpler AI tasks. Traditionally, CPU has been dominant the computing tasks over other chipsets as its performance 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 72: Large Language Models Continue to Expand Inputs



Source: Hugging Face

Figure 73: Internet Service Growth Has Slowed to Single Digits



Source: Cerebras

Figure 74: Different Types of Chipset Comparison for AI Computing

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

Source: Credit Suisse

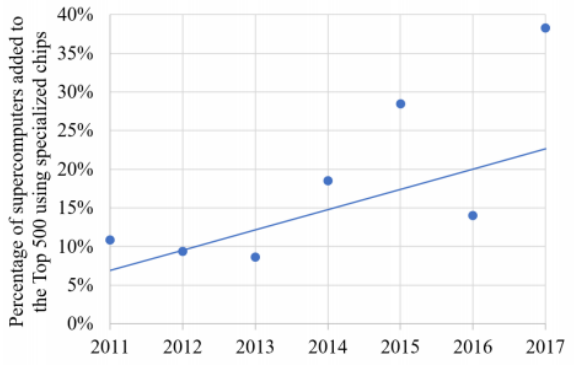
- **GPU (Graphic Processing Unit):** GPUs led by NVIDIA and AMD benefit traditionally processing parallel pixels in imaging are well suited to the parallelism in training AI algorithms and inference requiring executing on multiple pieces of data at once including the newer transformer models used in 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 has 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.
- **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.

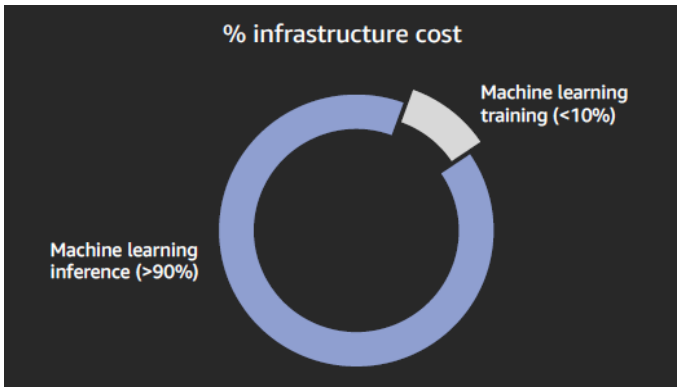
The Center for Security and Emerging Technology estimates potential higher efficiency for an FPGA or ASIC though can lose on accuracy and are less flexible for the wider general compute workloads that some tasks require.

Figure 75: ASIC gets importance in supercomputing now



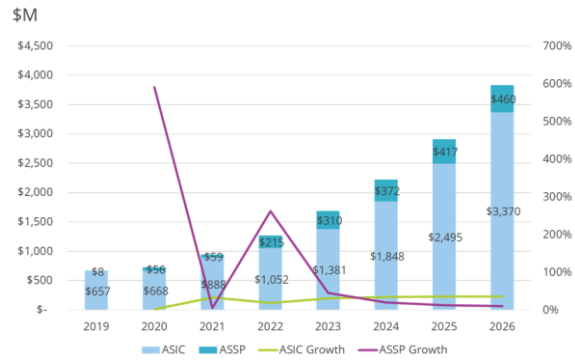
Source: Stanford University

Figure 76: Amazon sees more efficient chipset computing is important to reduce machine learning inference cost



Source: Amazon

Figure 77: Server ASIC market led by AWS, Alibaba and Google's TPU



Source: IDC

Figure 78: AI processors projected at 25% '23-26 CAGR, GPUs capturing 77% share

	Training		Inference		Generality[88]	Inference accuracy[89]
	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.

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.

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.

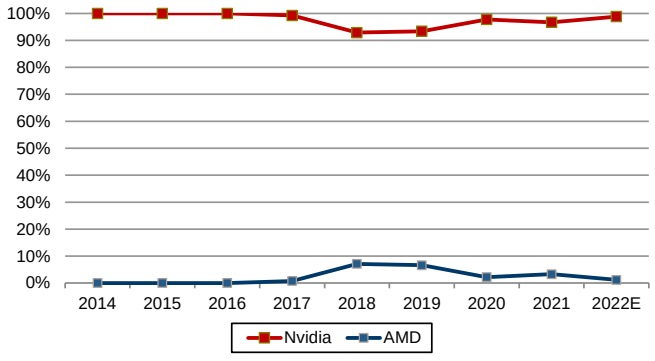
We note that in MLPerf Inference benchmarks, which measure how quickly a trained neural network can perform inference tasks on new data, NVDA's H100 had over 4x higher inference performance relative to the A100 and was the performance leader across all MLPerf inference benchmarks relative to competitors.

Figure 79: 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%
Total	**$2,989**	**$5,506**	**$7,066**	**$9,360**	**$12,405**	**$15,372**	**$18,178**	**46%**	**25%**	**100%**
YoY Growth		**84%**	**28%**	**32%**	**33%**	**24%**	**18%**			

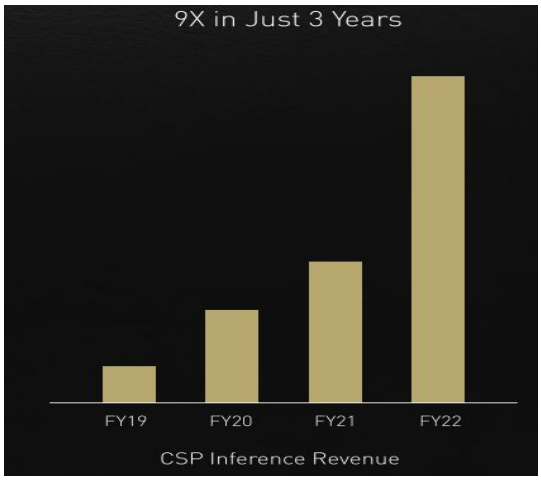
Source: Gartner, December 2022

Figure 80: NVIDIA Leads the GPU Server Market



Source: Mercury Research

Figure 81: NVIDIA's Inference GPU Sales Up 9x in 3 Years



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.

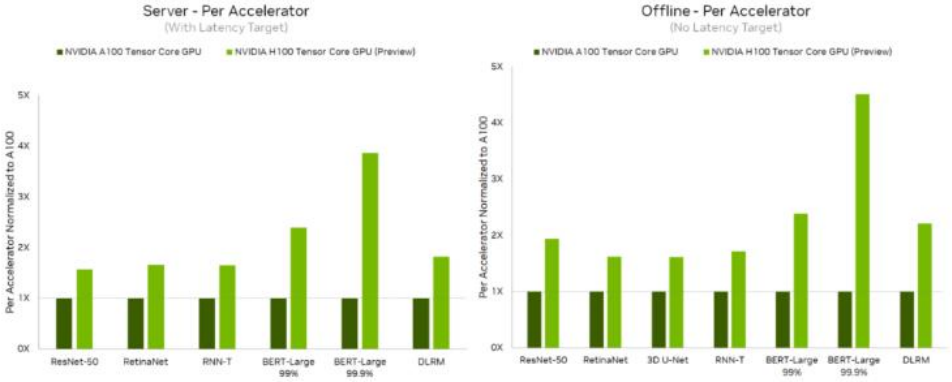
NVDA's lower-end A-series GPUs are targeted toward enterprise level inference applications, from a 40-60W A2 edge level inference, to 165W A-30 GPUs. NVDA claims a 7x performance improvement as compared to an Intel Xeon Gold 6330N CPU for inference applications.

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.

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 82: Datacenter Per-Accelerator Performance – H100 vs. A100



Source: NVIDIA

Figure 83: 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%
% Training Accelerated	**86.5%**	**86.0%**	**87.1%**	**88.6%**	**89.4%**	**90.1%**	
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%
% Inference Accelerated	**22.8%**	**34.1%**	**43.2%**	**48.5%**	**52.0%**	**53.2%**	

Source: IDC 2022

Figure 84: 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%
Total Server Revenue	**$109.0**	**$124.1**	**$131.8**	**$140.5**	**$150.8**	**$162.6**	**8%**
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%**	
AI Server Revenue (Nov 2022)	**$15.7**	**$19.6**	**$24.8**	**$28.7**	**$31.9**	**$34.9**	**17%**
AI Penetration % - Nov forecast	**14.4%**	**15.8%**	**18.8%**	**20.5%**	**21.2%**	**21.4%**	
AI Server Revenue (Dec 2022)	**$16.9**	**$21.8**	**$27.8**	**$33.0**	**$38.0**	**$43.4**	**21%**
AI Penetration % - Dec forecast	**15.5%**	**17.6%**	**21.1%**	**23.5%**	**25.2%**	**26.7%**	

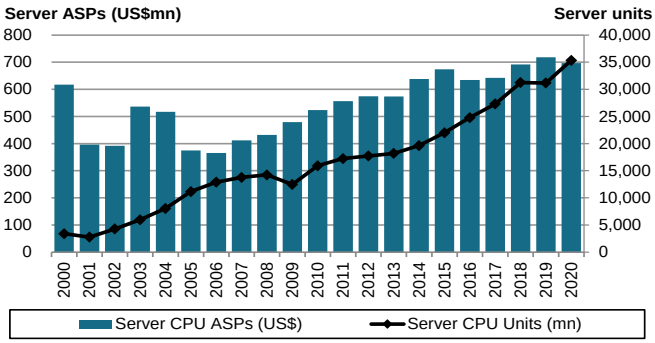
Source: IDC 2022

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.

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.

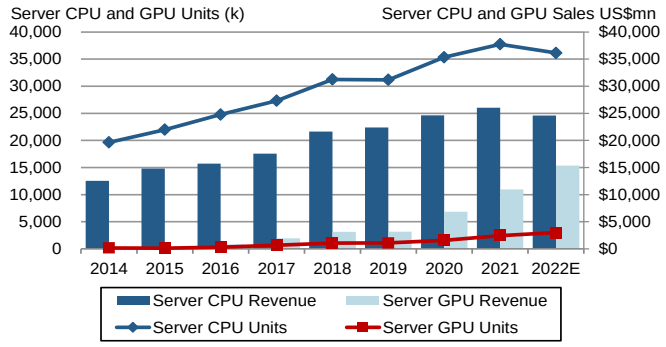
Yole estimates GPU/AI ASIC revenue of approximately $57bn-$63bn by 2027, implying a CAGR of ~33% through 2027, and more than doubling from 2022. This forecast is consistent with forecasts from AMD which estimates $64bn GPU/AI datacenter TAM in the long-term.

Figure 85: Server CPU a +8% Unit/9% Sales CAGR since 2014



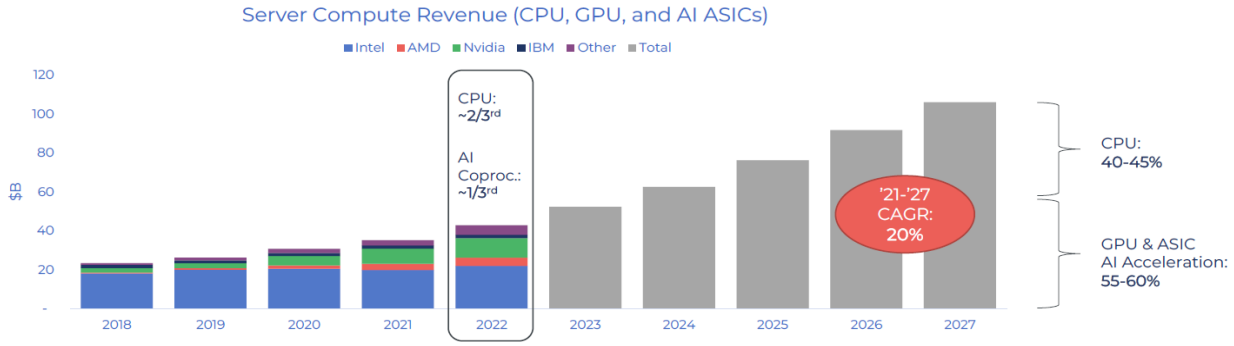
Source: Mercury Research

Figure 86: GPU ramp at a 42% unit/62% sales CAGR since '14



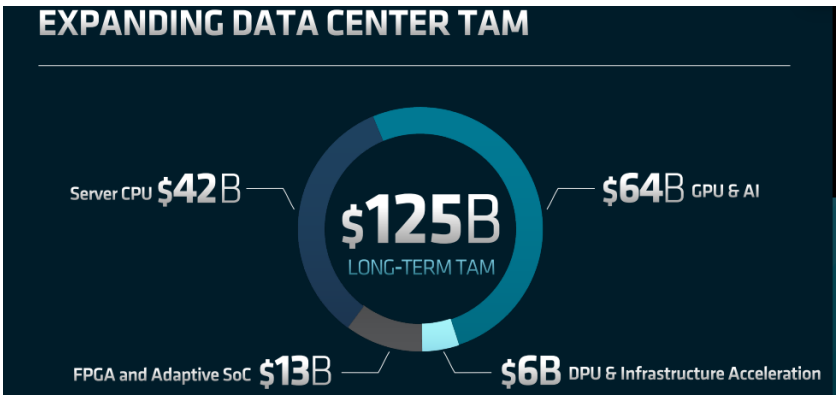
Source: Mercury Research

Figure 87: Yole forecasts the server compute TAM shifts from CPU led toward GPU & ASIC/AI led by 2027



Source: Yole Development

Figure 88: 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.

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 89: 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
Intel DCAI	**$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%	
AMD Data Center	**$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%	
Total Data Center CPUs	**$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%**
QoQ Growth	**-3%**	**-16%**	**-5%**	**2%**	**-18%**	**-6%**	**9%**	**10%**							
YoY Growth	**32%**	**-4%**	**-16%**	**-21%**	**-33%**	**-25%**	**-14%**	**-8%**		**10%**	**6%**	**-4%**	**-21%**	**25%**	

Source: Company data, Credit Suisse

Figure 90: Accelerated AI Servers to Outpace Non-Accelerated (CPU) Powered servers

Data Center ICs	1Q22	2Q22	3Q22	4Q22	1Q23F	2Q23F	3Q23F	4Q23F	2019	2020	2021	2022F	2023F	2024F	19-24
NVIDIA Datacenter	**$3,750**	**$3,806**	**$3,833**	**$3,616**	**$3,853**	**$4,263**	**$4,655**	**$5,080**	**$2,983**	**$6,696**	**$10,613**	**$15,005**	**$17,851**	**$22,871**	**50%**
QoQ Growth	15%	1%	1%	-6%	7%	11%	9%	9%							
YoY Growth	83%	61%	31%	11%	3%	12%	21%	40%	2%	124%	58%	41%	19%	28%	
Marvell Datacenter	**$641**	**$643**	**$627**	**$470**	**$530**	**$580**	**$683**	**$751**	**$861**	**$1,041**	**$1,785**	**$2,382**	**$2,544**	**$3,220**	**30%**
QoQ Growth	12%	0%	-3%	-25%	13%	9%	18%	10%							
YoY Growth	131%	48%	26%	-18%	-17%	-10%	9%	60%		21%	71%	33%	7%	27%	
ASIC and GPU ICs	**$4,391**	**$4,449**	**$4,460**	**$4,086**	**$4,383**	**$4,843**	**$5,338**	**$5,831**	**$3,844**	**$7,737**	**$12,398**	**$17,387**	**$20,395**	**$26,091**	**47%**
QoQ Growth	**14%**	**1%**	**0%**	**-8%**	**7%**	**11%**	**10%**	**9%**							
YoY Growth	**89%**	**59%**	**30%**	**6%**	**0%**	**9%**	**20%**	**43%**		**101%**	**60%**	**40%**	**17%**	**28%**	
Data Center CPU + ICs	**$11,718**	**$10,584**	**$10,278**	**$10,045**	**$9,286**	**$9,435**	**$10,325**	**$11,315**	**$26,456**	**$32,690**	**$38,783**	**$42,626**	**$40,361**	**$51,128**	**14%**
QoQ Growth	**3%**	**-10%**	**-3%**	**-2%**	**-8%**	**2%**	**9%**	**10%**							
YoY Growth	**49%**	**16%**	**0%**	**-12%**	**-21%**	**-11%**	**0%**	**13%**		**24%**	**19%**	**10%**	**-5%**	**27%**	

TSMC HPC sales	1Q22	2Q22	3Q22	4Q22	1Q23F	2Q23F	3Q23F	4Q23F	2019	2020	2021	2022F	2023F	2024F	19-24
TSMC HPC Segment	**$7,210**	**$7,717**	**$7,908**	**$8,451**	**$7,183**	**$6,993**	**$8,238**	**$9,186**	**$10,259**	**$14,944**	**$21,008**	**$31,285**	**$31,600**	**$37,013**	**29%**
QoQ Growth	**26%**	**7%**	**2%**	**7%**	**-15%**	**-3%**	**18%**	**12%**							
YoY Growth	**58%**	**51%**	**42%**	**47%**	**0%**	**-9%**	**4%**	**9%**		**46%**	**41%**	**49%**	**1%**	**17%**	

Source: Company data, Credit Suisse

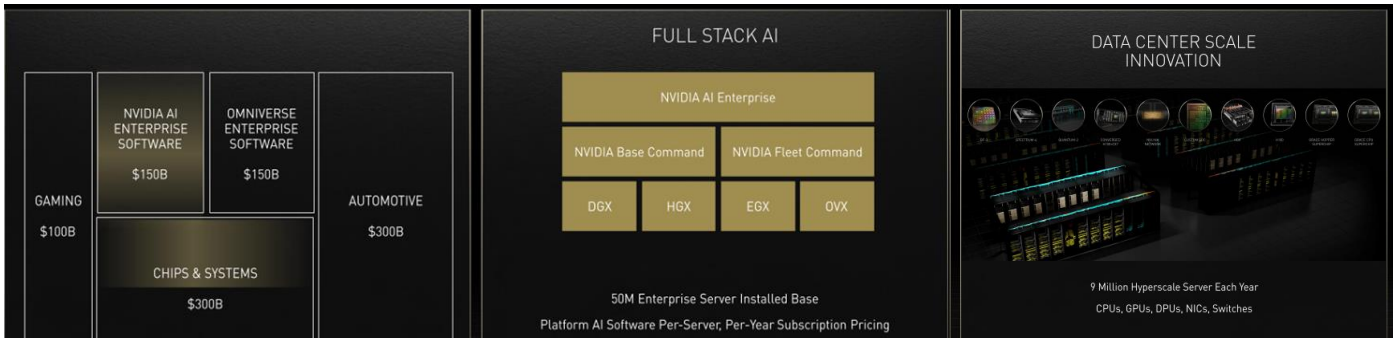
NVIDIA continuing to advance its solutions to power AI

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.

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.
- **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'),

Figure 91: NVIDIA estimates a US$1trn TAM with US$300bn in chips/system hardware and US$150bn from hyperscale



Source: NVIDIA

Figure 92: NVIDIA Compute Optimization Across the Full Stack

Source: NVIDIA

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

Figure 93: Nvidia develops Application frameworks to further simplify adoption of AI

Source: NVIDIA

Figure 94: H100 2.4x CUDA cores, ~50% more transistors/VRAM bandwidth vs A100

NVIDIA Accelerator Specification Comparison			
	H100	A100 (80GB)	V100
FP32 CUDA Cores	16896	6912	5120
Tensor Cores	528	432	640
Boost Clock	~1.78GHz (Not Finalized)	1.41GHz	1.53GHz
Memory Clock	4.8Gbps HBM3	3.2Gbps HBM2e	1.75Gbps HBM2
Memory Bus Width	5120-bit	5120-bit	4096-bit
Memory Bandwidth	3TB/sec	2TB/sec	900GB/sec
VRAM	80GB	80GB	16GB/32GB
Interconnect	NVLink 4 18 Links (900GB/sec)	NVLink 3 12 Links (600GB/sec)	NVLink 2 6 Links (300GB/sec)
GPU	GH100 (814mm2)	GA100 (826mm2)	GV100 (815mm2)
Transistor Count	80B	54.2B	21.1B
TDP	700W	400W	300W/350W
Manufacturing Process	TSMC 4N	TSMC 7N	TSMC 12nm FFN
Interface	SXM5	SXM4	SXM2/SXM3
Architecture	Hopper	Ampere	Volta

Source: Company data, Anandtech

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.

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 95: 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.

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.

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.

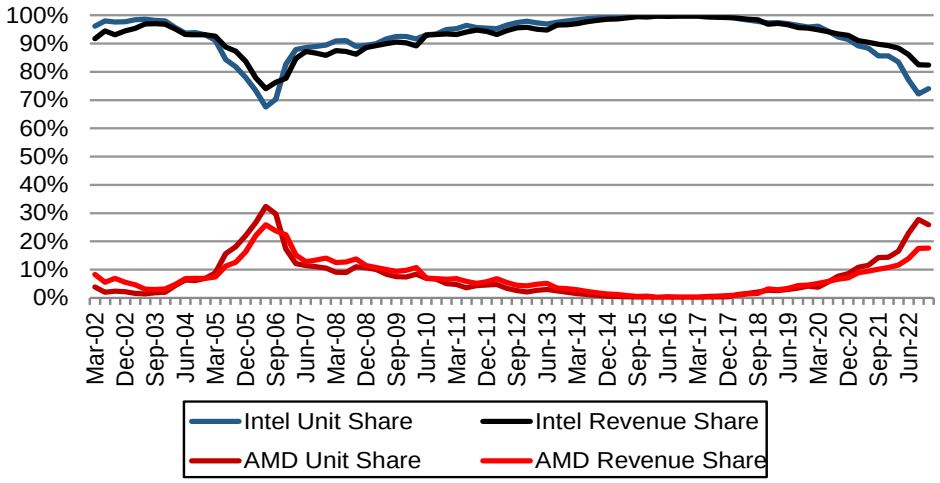
Intel also announced a next generation data center GPU, code named Rialto Bridge set for market debut in 2024 followed by Falcon Shores tile-based architecture. While INTC has made some performance claims citing advantages vs. NVDA silicon, we're not aware of any significant traction with regard to these initiatives and we believe revenue has thus far been immaterial as compared to NVDA datacenter revenue.

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 96: Intel has lost unit and revenue share in server CPUs in the past 5 years

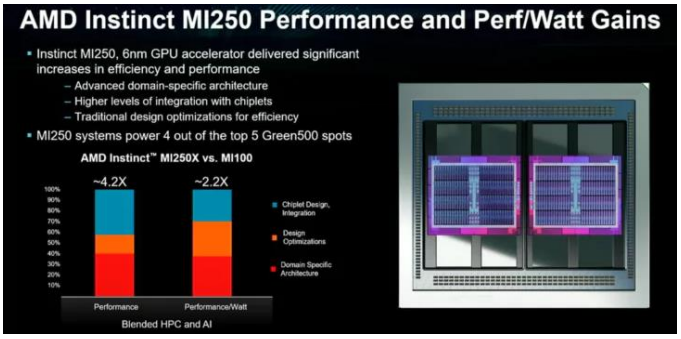


Source: Mercury Research

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.

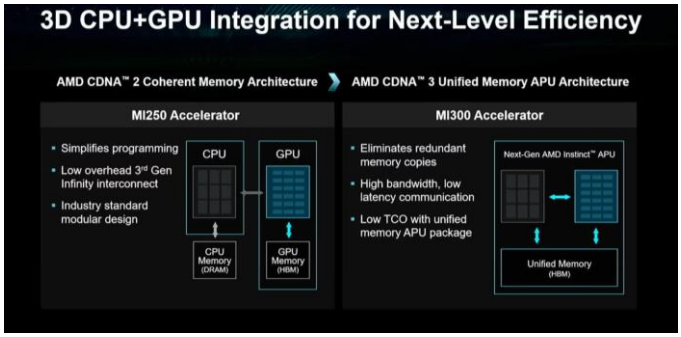
On comparison of raw specs, AMD's AI cores and transistor count stretches beyond NVIDIA's GPUs including its recent H100 launch. AMD still has to play catch-up to NVIDIA's incumbent advantage with its application frameworks and CUDA stack which have already been used across many AI workloads. The hyperscalers though are introducing their own open-source software stack which can enable more applications to run on rival accelerators.

Figure 97: AMD MI250 Accelerator with CPU/GPU integration



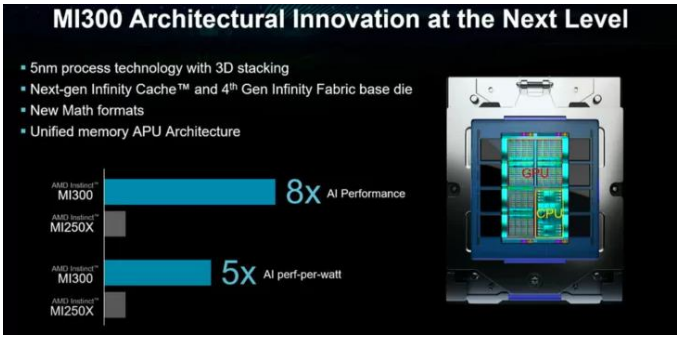
Source: AMD

Figure 98: MI300 adds shared memory access to the CPU/GPU



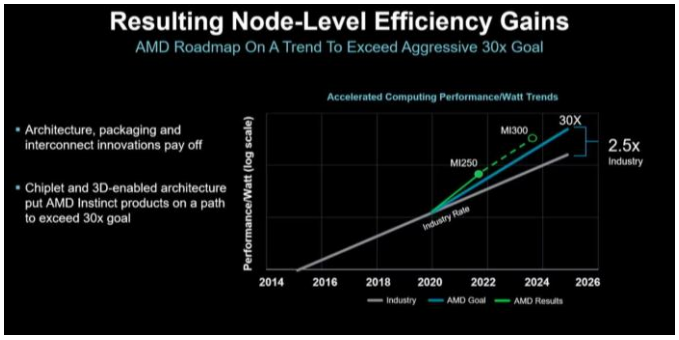
Source: AMD

Figure 99: AMD MI300 for additional AI performance



Source: AMD

Figure 100: Accelerated performance/watt gains with MI300



Source: AMD

Figure 101: 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^2)	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^2)	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

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.

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 102: 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

Figure 103: 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
Chipset	WSE-2	Cloud TPUv3	Cloud TPUv4	Colossus MK2	Gaudi 2	Greco	Inferentia Gen 2	Trainium
Type	ASIC	ASIC	ASIC	ASIC	ASIC	ASIC	ASIC	ASIC
Function	AI training	AI training/inference	AI training/inference	AI accelerator	AI training	AI inference	AI inference	AI inference
Foundry	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC
Manufacturing node	7nm	16nm	7nm	7nm	7nm	7nm	7nm	7nm
Die size (mm^2)	46,225	648	780	823	850	850	NA	NA
Number of transistors (billion)	2,600	10	31	59	NA	NA	NA	NA
Density (mn transistors per mm^2)	56	15	40	72	NA	NA	NA	NA
AI-optimized cores	850,000	2	2	1,472	24	8	2	2
On-board SRAM	40GB	32MB	288MB	900MB	48MB	128MB	NA	NA
Memory bandwidth	20PB/s	900GB/s HBM	1200GB/s HBM2	47.5TB/s HBM	2.45TB/s HBM2e	204GB/s LPDDR5	NA	13.1TB/s HBM
Bus interface	NA	NA	NA	PCIe 4.0x16	PCIe 4.0x16	PCIe 4.0x8	NA	NA
ASIC provider	TSMC direct	Mediatek	Broadcom	TSMC direct	Alchip	Alchip	Alchip	Alchip
Launch	3Q21	4Q18	4Q21	3Q20	2Q22	2Q22	2022	4Q20
Peak performance (FP32, TFLOPS)	NA	4	4	62	38	NA	2,022	53
TDP (W)	NA	450	175	300	600	75	NA	NA

Source: Company data, Credit Suisse

- **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.

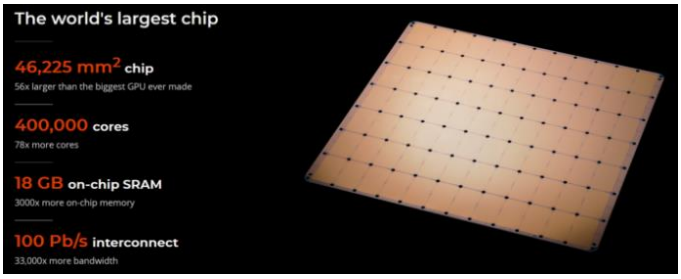
  With its wafer scale engine powering the company's system level solution CS-2, it claims one system can train models with up to 20bn parameters. It is also noteworthy that compared with the traditional GPU cluster where the researchers need to use a special framework to solve the bottleneck on individual 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 combing 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.

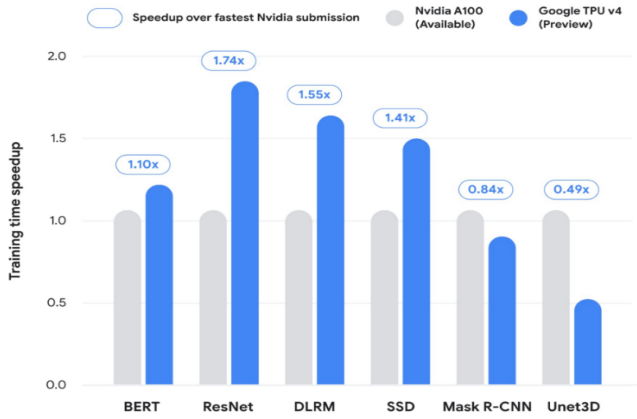
  Following the initial adoption of TPU v1 on TSMC's 28nm in 2015 only for AI inference, the company has been upgrading the chipsets for both AI training and inference through optimizing the chipset architecture and manufacturing technology migration, with the latest version TPU v4 manufactured on TSMC's 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 programing with its autoregressive language model project.

Figure 104: Cerebras Systems wafer size engine for high performance and low latency



Source: Cerebras Systems

Figure 105: Google claims TPUv4 has superior performance on training



Source: Google

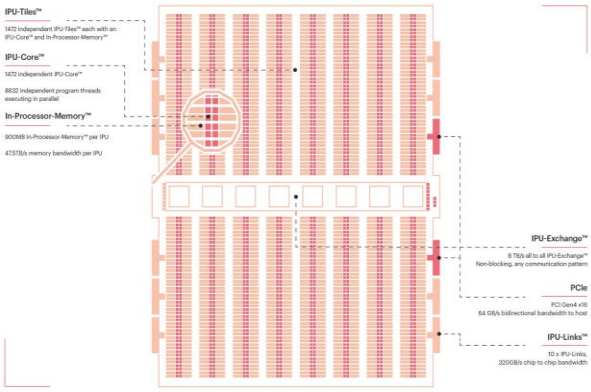
- **Graphcore's Colossus MK2:** Graphcore is a UK semiconductor company founded in 2016 with a focus on developing the accelerators for AI and machine learning. The company introduced its 1st chipset Colossus GC2 on 16nm supporting the tasks in the standard machine learning frameworks. The company subsequently introduced its 2nd generation processor Colossus MK2 - GC200 IPU on TSMC's 7nm, featuring 59bn transistors, 1,472 computing cores and 823mm^2 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.

- **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.

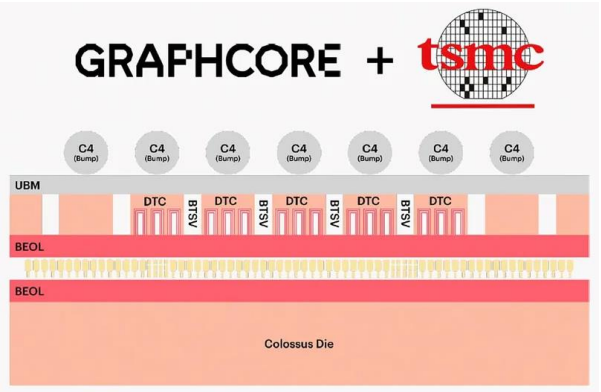
  Habana Labs has migrated its semiconductor roadmap to TSMC's 7nm for its 2nd generation Guadi2 AI training chipset, increasing the number of AI-customized tensor processor cores from 8 to 24, tripling its in-package memory to 96GB of HBM2e at 2.45TB/s bandwidth while adding support for FP8 while integrating 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.

Figure 106: Graphcore Colossus MK2 structure



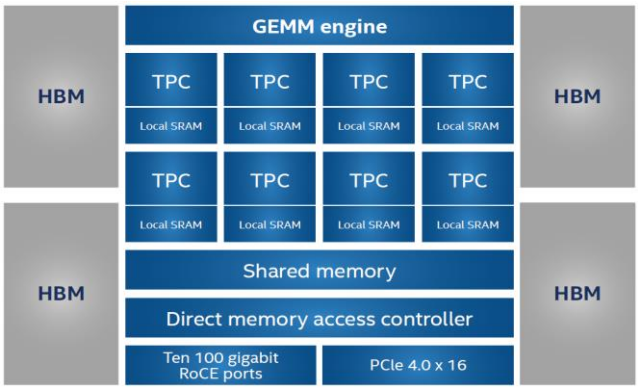
Source: Graphcore

Figure 107: TSMC's WoW packaging enables Graphcore's IPU



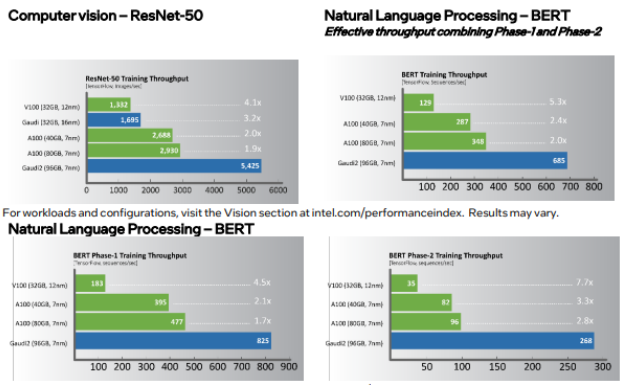
Source: Graphcore

Figure 108: Habana's Gaudi2 features multiple cores and HBM



Source: Intel

Figure 109: Habana's Gaudi2 performance better than Nvidia



Source: Intel

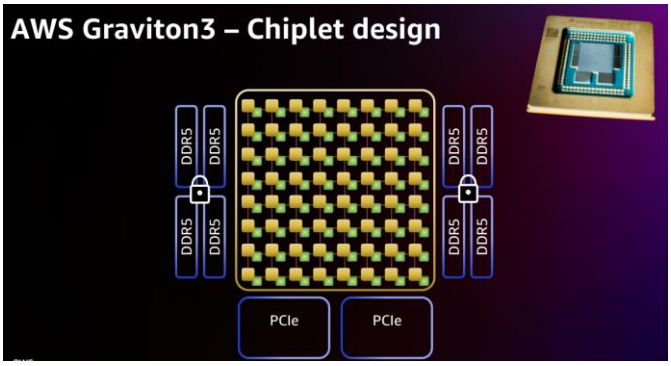
According to Habana, 1,000 Gaudi2 processors have been deployed in its own data center in Israel to support R&D for software optimization while it is also working with multiple customers including Mobileye (training deep learning models for tasks such as object detection and segmentation which enable 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).

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 if 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 110: Amazon's Graviton3 adopts chipset design



Source: Amazon

Figure 111: Graviton3 is widely adopted by aws customers

Source: Amazon

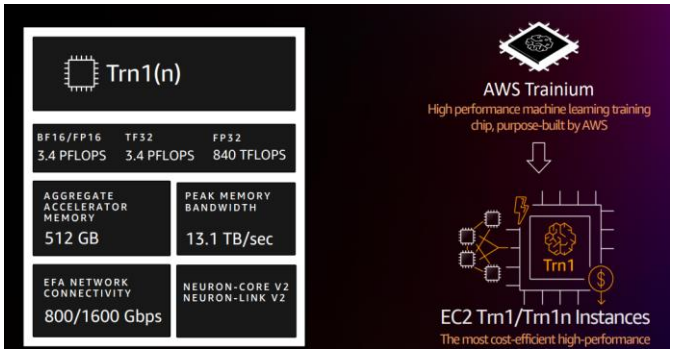
For CPU, Amazon introduced Graviton in 4Q18 featuring ARM's Cortex-A72 cores and subsequently launched Graviton2 in 1Q20 and Graviton3 in 2Q22 to upgrade the performance and power efficiency at much lower cost compared with the off-the-shelf solutions from Intel and AMD. For Graviton2 and Graviton3, 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.

For AI inference, Amazon first introduced Inferentia chipset on TSMC's 16nm in 2018 along with its Graviton CPU, featuring 4 Neuron cores performing up to 128 TOPS while the 2nd 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 1st 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 chipset 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.

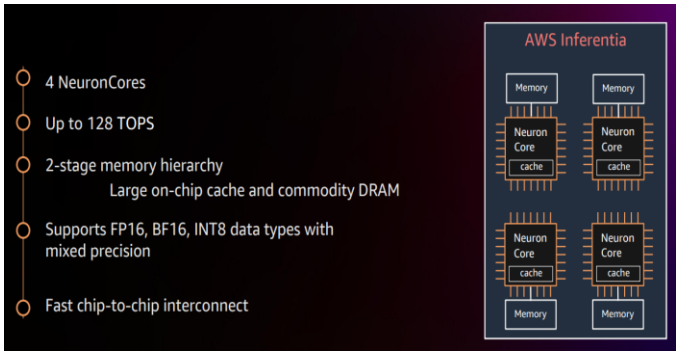
In addition to the diversifying AI chipset project pipeline in the US and EU, China hyperscalers and start up companies were encouraged by local government to develop the high-end chipsets to improve semiconductor self-sufficiency amid geopolitical tension. In the past couple of years, China companies 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 112: Amazon's Trainium chipset strong on AI training



Source: Amazon

Figure 113: Key features for Amazon's Inferentia chipset



Source: Amazon

Figure 114: China AI ASIC solutions

Company	Baidu	Baidu	Tencent/Enflame	Alibaba/T-Head	Cambricon	Corerain	Biren	Huawei
Chipset	Kunlun Gen 1	Kunlun Gen 2	Enflame 2.5	Hanguang 800	Siyuan 370	CAISA3.0	BR100	Ascend 910
Type	ASIC	ASIC	ASIC	ASIC	ASIC	SIC (AI accelerato	GPU	SoC
Function	AI inference	AI inference/training	AI training	AI inference	AI inference/training	AI inference	AI inference/training	AI inference/training
Foundry	Samsung	Samsung	GlobalFoundries	TSMC	TSMC	SMIC	TSMC	TSMC
Manufacturing node	14nm	7nm	12nm FinFET	12nm	7nm	28nm	7nm	7nm+
Die size (mm^2)	NA	NA	NA	709	NA	NA	1,074	456
Number of transistors (billion)	NA	NA	21	17	39	NA	77	NA
Density (mn transistors per mm^2)	NA	NA	NA	24	NA	NA	72	NA
AI-optimized cores	NA	NA	NA	4	4	4	16	32
On-board SRAM	16MB	16MB	32MB	192MB	NA	NA	300MB	32MB
Memory bandwidth	256GB/s	512GB/s GDDR6	819GB/s HBM2e	NA	614GB/s	45GB/s DDR4	2300GB/s HBM2e	1.2TB/s HBM2
Bus interface	PCIe 4.0x8	PCIe 4.0x8	PCIe 4.0x8	PCIe 4.0x16	PCIe 4.0x16	PCIe 3.0x4	PCIe 5.0x16	PCIe 3.0x4
ASIC provider	Samsung direct	Samsung direct	NA	NA	NA	NA	Alchip	NA
Launch	1Q20	3Q21	4Q21	3Q19	4Q21	2Q20	3Q22	3Q19
Peak performance (FP32, TFLOPS)	256 TOPS	640 TOPS	32	NA	256 TOPS	NA	128	512 TOPS
TDP (W)	150	150	NA	NA	NA	NA	550	310

Source: Company data, Credit Suisse

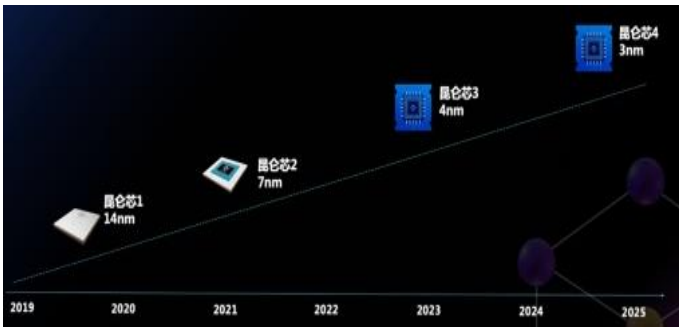
- **Baidu:** Baidu started its AI chipset development from 2011 and spun off its chipset business as an independent company Kulunxin and raised funds in 1Q21 to focus on general purpose AI chipset development for deep learning and machine learning adopting natural language processing, visual recognition, 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 3rd 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.
- **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 first AI training card Yunsui T10 based on its Enflame 2.5 chipset was introduced in 4Q19 and the company launched its 1st 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. first 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 second 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 1st generation chipset. The company targets to introduce its 3rd 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 first 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.

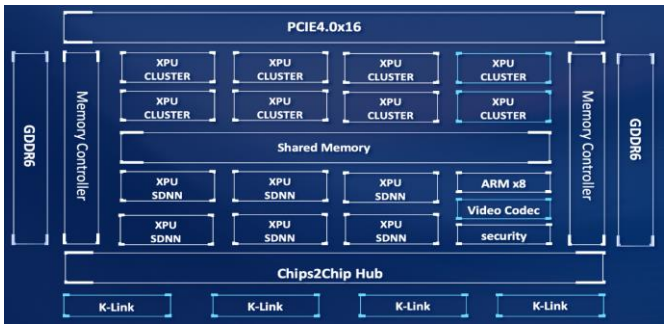
  For CPU, the company introduced ARM based Yitian 710 chipset in 4Q21 based on TSMC's 5nm and packs 128 ARMv9 cores with clock speeds as high as 3.20 GHz. It has eight DDR5-4800 memory channels that can produce up to 307.2 Gbps of transfer speed and 96 PCIe 5.0 lanes which the company claims 30% improvement in cost

Figure 115: Baidu has roadmap through 3nm for Kunlun chipset



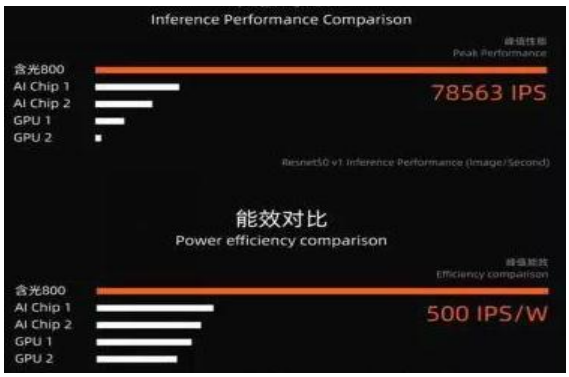
Source: Baidu

Figure 116: Kunlun2 supports GDDR6 and PCIe4.0x16



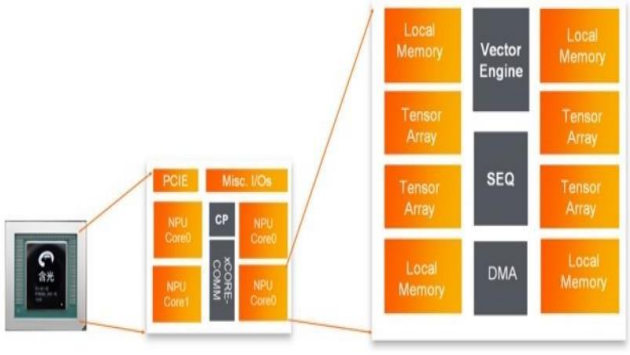
Source: Baidu

Figure 117: Alibaba claims solid performance for Hanguang 800



Source: Alibaba

Figure 118: Alibaba Hanguang 800 structure



Source: Alibaba

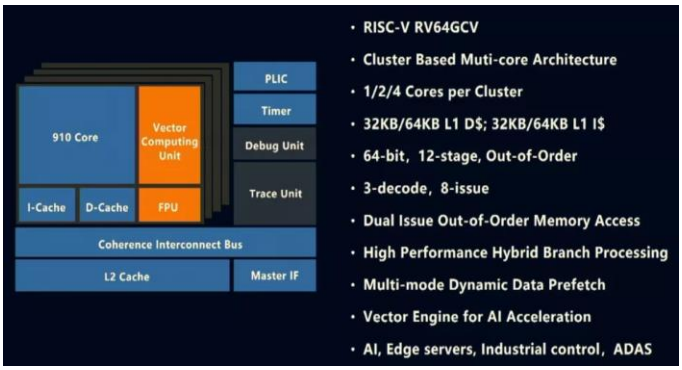
performance and TDP down by 60% compared with the similar products. In addition to ARM CPU, the company also introduced its first RISC-V CPU in 3Q19 Xuentie E902 and upgraded the RSIC-V roadmap through Xuentie C908 as of 2022 to deliver 3.5x better performance on graph classification and neural networking 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.

- **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 first 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 119: Alibaba's Yitian ARM CPU features TSMC's 5nm

Source: Alibaba

Figure 120: Alibaba introduced its RISC-V CPU portfolio

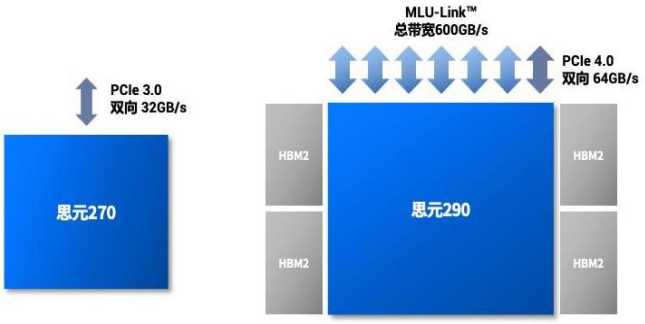


Source: Alibaba

Figure 121: Cambricon's architecture migration supports more features

Source: Cambricon

Figure 122: Cambricon's Siyuan290 chipset supports HBM2

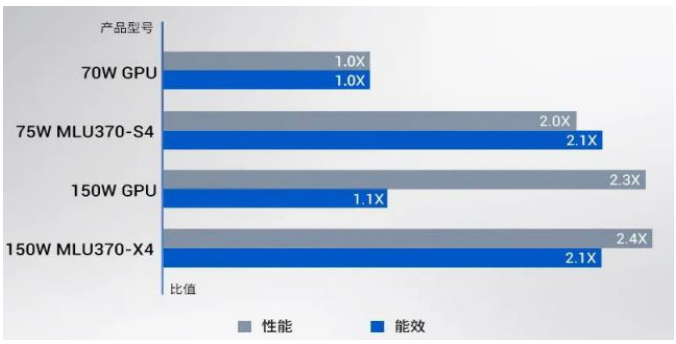


Source: Cambricon

Figure 123: Cambricon shifts its AI chipsets toward chiplet tiles

Source: Cambricon

Figure 124: Cambricon MLU370-X4 improves power/perf.



Source: Cambricon

In 2021, the company launched its 1st 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 1st 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.

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.

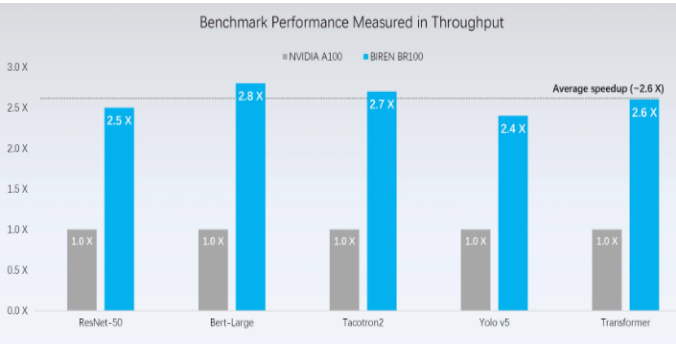
  Biren has been focusing on the general-purpose GPU design and introduced its 1st chipset BR100 on TSMC's 7nm in 3Q22, featuring 77bn transistors in 1,074mm^2 die size, 64GB HBM2e memory supporting 1.64TB/s bandwidth and 819GB/s I/O speed. The company claimed the chipset could deliver up to 256 FP32 TFLOPS 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 125: 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

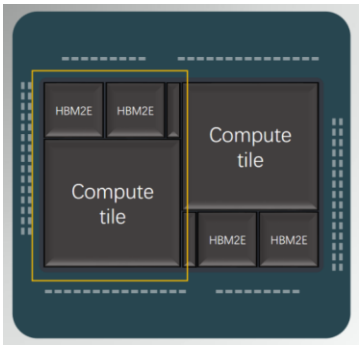
Source: IDC

Figure 126: Biren claims its BR100 has better performance



Source: Biren

Figure 127: Biren BR100 architecture

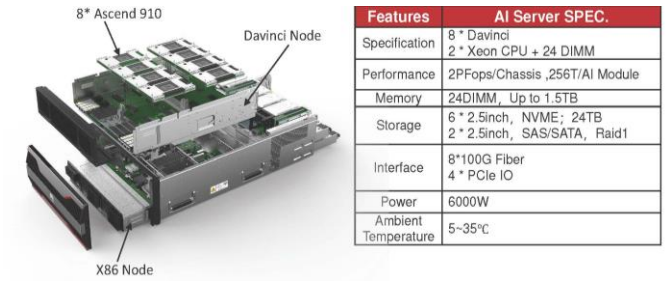


Source: Biren

- **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.
- **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.

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 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.

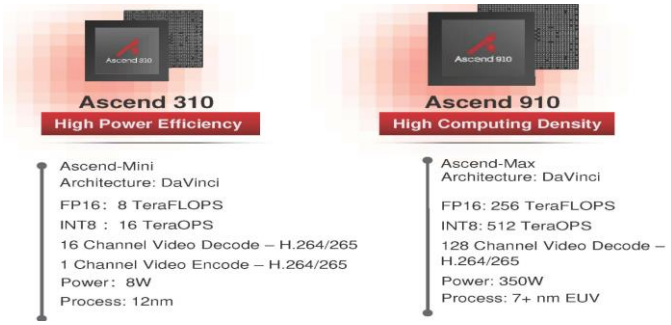
Figure 128: Huawei's AI server based on its Ascend910



Features	AI Server SPEC.
Specification	8 * Davinci 2 * Xeon CPU + 24 DIMM
Performance	2PFops/Chassis ,256T/AI Module
Memory	24DIMM, Up to 1.5TB
Storage	6 * 2.5inch, NVME; 24TB 2 * 2.5inch, SAS/SATA, Raid1
Interface	8*100G Fiber 4 * PCIe IO
Power	6000W
Ambient Temperature	5~35℃

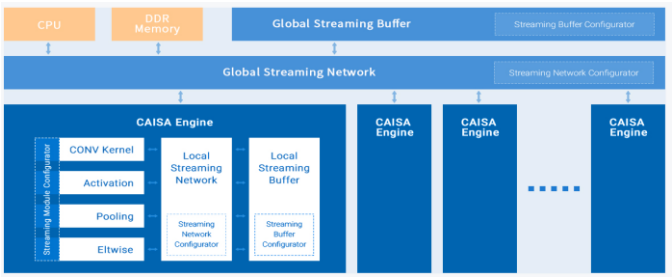
Source: Huawei

Figure 129: Ascend910 vs. Ascend310 spec comparison



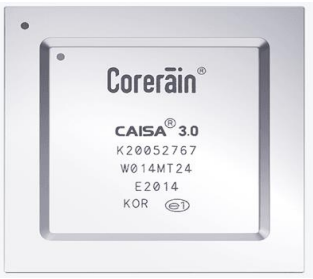
Source: Huawei

Figure 130: Corerain's CAISA chipset architecture



Source: Corerain

Figure 131: Corerain's latest CAISA 3.0 chipset



Source: Corerain

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.

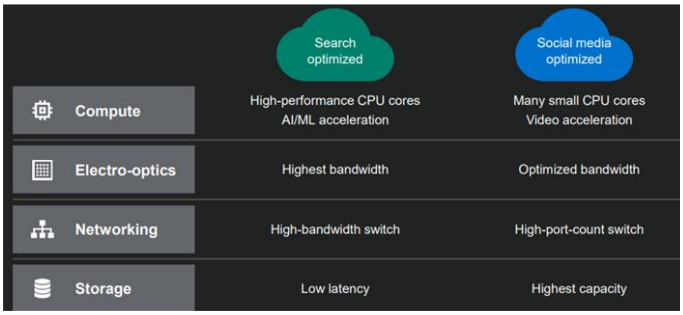
AI could eventually be a catalyst for the Edge compute suppliers, but too early to re-rate Qualcomm and Mediatek

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.

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.

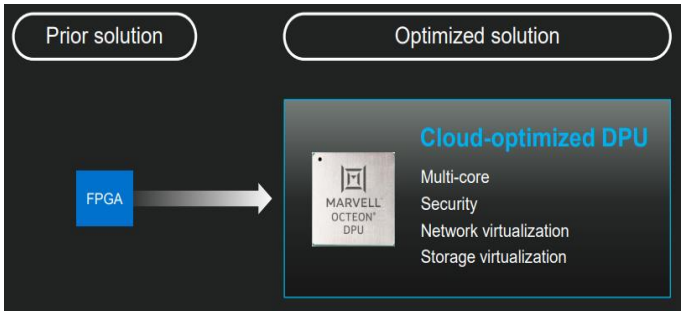
AI would also be integrated in other edge processors to drive overall 14% CAGR as it adds content across embedded processors, MCUs and FPGAs.

Figure 132: Hyperscale customers have unique requirements…



Source: Marvell

Figure 133: Marvell developing custom and optimized solutions for cloud data center



Source: Marvell

Figure 134: 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%
AI AP % of industry units	**19%**	**24%**	**30%**	**36%**	**38%**	**40%**	**41%**		
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	$21.2	13%	1%
AI Processor Premium	**111%**	**94%**	**66%**	**62%**	**61%**	**54%**	**51%**		
AI Application Processor Sales US$mn	$17,773	$26,810	$33,037	$38,229	$43,134	$45,956	$48,640	29%	8%
Industry App. Processor Sales US$mn	$45,009	$57,847	$66,127	$65,428	$70,975	$75,275	$78,352	13%	6%
AI AP % of industry revenue	**39%**	**46%**	**50%**	**58%**	**61%**	**61%**	**62%**		

Source: Gartner April 2022

Figure 135: AI integrated into more embedded processors

AI Integrated sales $mn	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%
Grand Total	**$19,037**	**$29,270**	**$37,374**	**$45,954**	**$54,256**	**$60,672**	**$67,901**	**34%**	**14%**	**100%**
YoY Growth		**54%**	**28%**	**23%**	**18%**	**12%**	**12%**			

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).

Power management: Monolithic Power has strong share in power management on AI boards

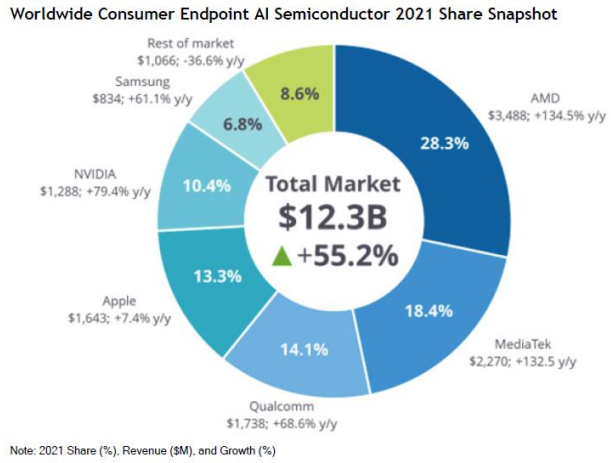
MPWR supplies the power management for both A-100 and H-100 GPUs. For the A-100, we believe MPWR has the majority of the market share, with VICR having a minority share. GPU power is a part of MPWR's enterprise data segment, which represented X of MPWR total revenue in CY22. Looking forward, we expect 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

Scaling still providing some benefits albeit diminishing

TSMC's technology leadership has come from a combination of process technology advances continuing on it is 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 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 of even improve on its innovation pace.

Figure 136: AI consumer edge semiconductors saw strong growth during the pandemic



Source: IDC

Figure 137: AI end points led by AMD in consoles, MTK in smart speakers/TVs

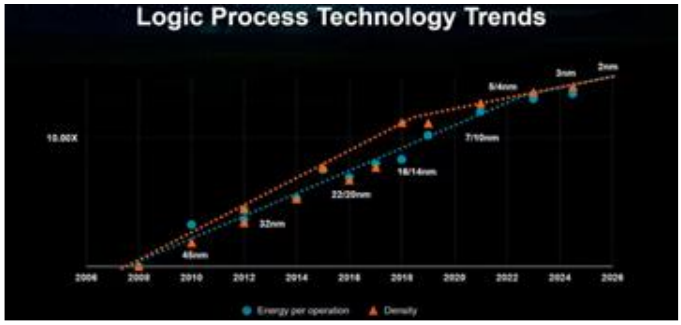
	2020		2021	
Vendor	Revenue ($M)	Share (%)	Revenue ($M)	Share (%)
AMD	1,487.8	18.7	3,488.2	28.3
MediaTek	976.1	12.3	2,269.6	18.4
Qualcomm	1,031.2	13.0	1,738.4	14.1
Apple	1,530.0	19.3	1,642.8	13.3
NVIDIA	718.0	9.0	1,287.9	10.4
Samsung	518.0	6.5	834.3	6.8
Other	1,680.8	21.2	1,065.8	8.6
Total	7,941.9	100.0	12,327.1	100.0

Source: IDC

Figure 138: TSMC maintains scaling to advanced geometries

Source: TSMC

Figure 139: 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.

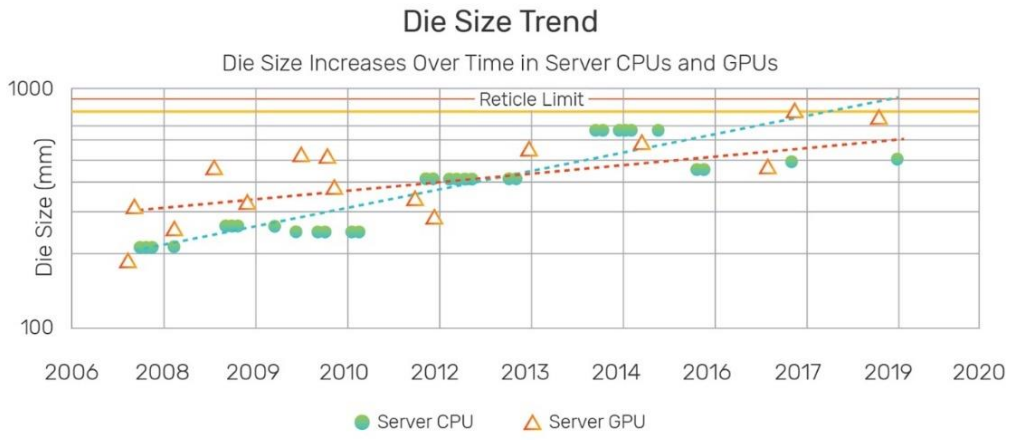
With lower density and cost gains, the drive to offer more performance and functionality at similar power has put processor die sizes on a gradual uptrend. The requirement for more compute in the iPad series and now Mac to run the GPU and AI instructions has led to 20-60% larger die sizes for the compute device, a measure that is also showing with AI through ever larger die sizes for the compute engines.

Figure 140: 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
A8	3Q14	20nm	89.0	2.0	22.5	$8,000	$0.11	198,539,326	43%
A9	3Q15	14/16nm	104.5	2.4	23.0	$8,300	$0.12	195,575,027	-1%
A10	3Q16	16nm	125.0	3.3	26.4	$7,500	$0.11	248,793,600	27%
A11	3Q17	10nm	87.7	4.3	49.0	$10,500	$0.15	330,047,239	33%
A12	3Q18	7nm	83.3	6.9	82.8	$11,750	$0.17	498,267,733	51%
A13	3Q19	7nm	98.5	8.5	86.3	$11,250	$0.16	542,159,052	9%
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%

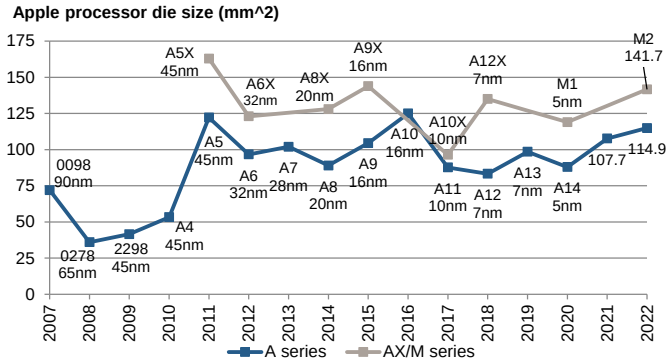
Source: Company Data, Credit Suisse Estimates

Figure 141: Die sizes increasing over time in CPUs and GPUs



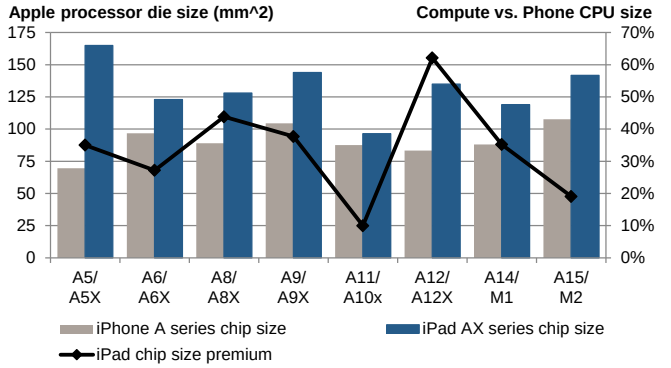
Source: Cadence, Semiconductor Engineering

Figure 142: Apple's die sizes getting larger as density gains fall



Source: Company Data, Credit Suisse Estimates

Figure 143: Apple's compute die size remains larger vs. iPhone



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.

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 well yielding 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.

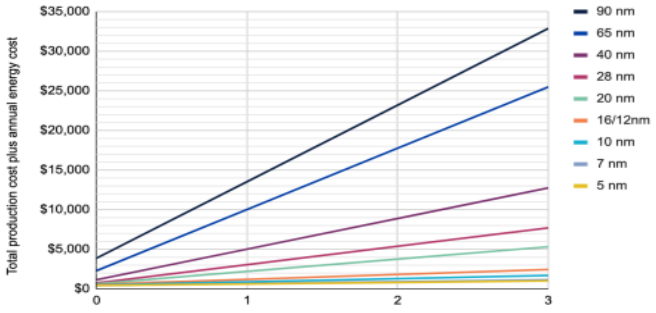
The companies 3DFabric toolkit includes InFO fan-out technology to join multiple chips in a compound material, CoWoS silicon interposer integration of multiple chips on a connector silicon and SoIC which integrates chips in a 3D stack. AMD noted also at its recent ISSCC keynote a piece of achieving higher interconnect density for continued improvements in power/performance requires these gains in chiplet integration in its MI250/300 accelerators.

Figure 144: Production overview of chips by die size

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[90]	$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

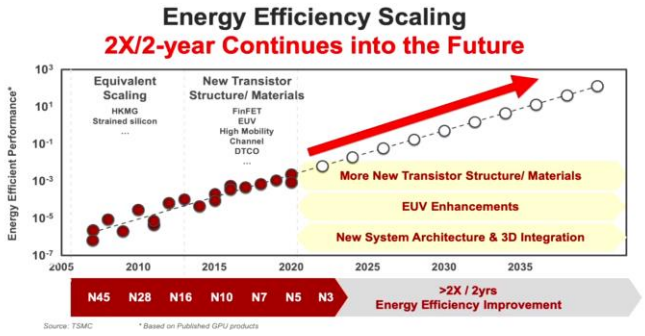
Source: CSET

Figure 145: Production costs underscore importance of energy efficient compute



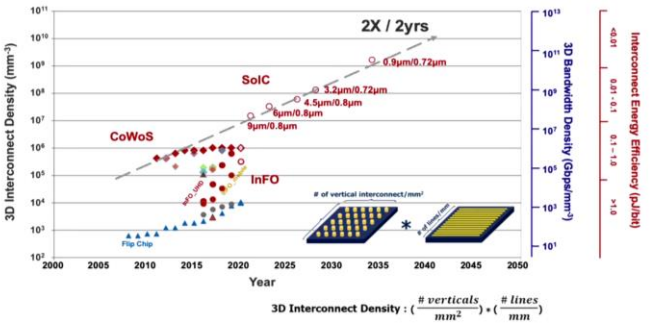
Source: CSET

Figure 146: TSMC's 3DFabric combines FE/BE integration



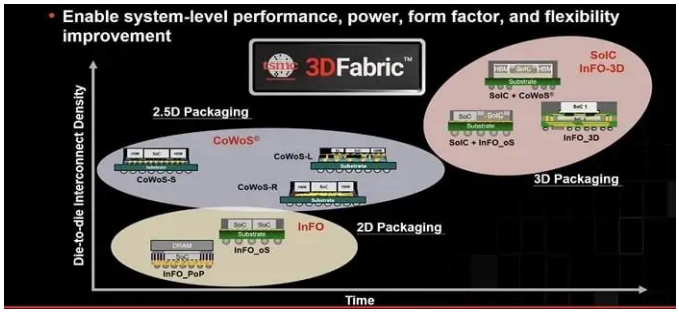
Source: TSMC

Figure 147: TSMC packaging helps maintain 2x scaling



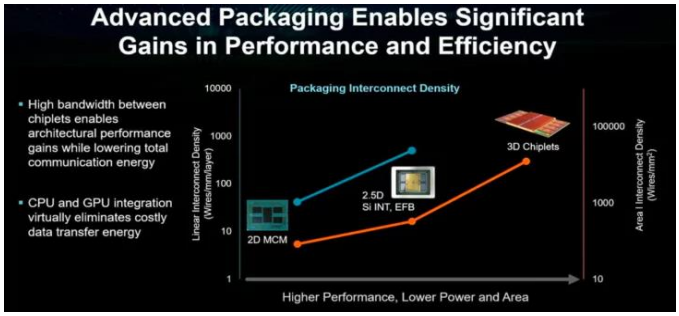
Source: TSMC

Figure 148: Advanced packaging solutions supplement scaling



Source: TSMC

Figure 149: AMD leveraging 2.5D/3D for data center solutions



Source: AMD

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.

For TSMC, the benefits of advancing process have 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.

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.

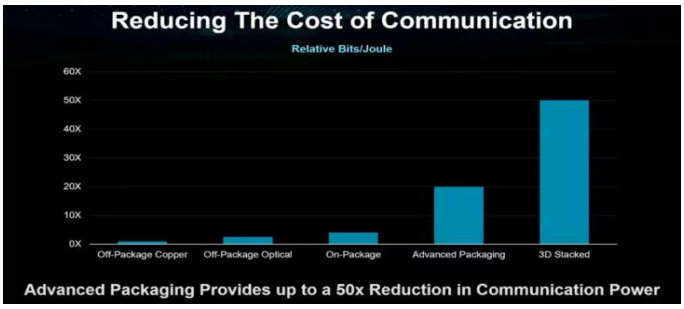
The OSATs though still have a role in final integration, with a number of 2.5D and 3D systems still relying on placement of the integrated system on a high-end ABF substrate for integration on a PCB as well as a final test service. The suppliers have also developed their own fan-out on substrate and 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

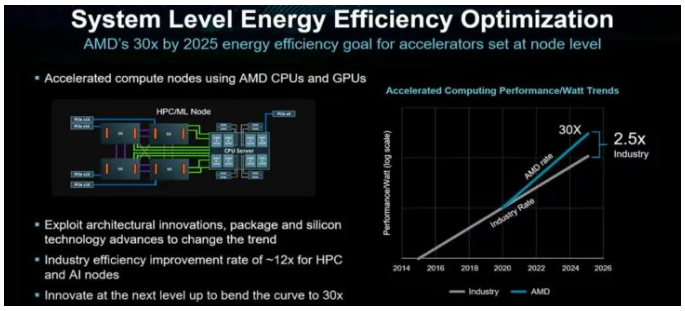
The demand for higher density and performance is also driving more advanced packaging tools including TCB and hybrid bonding. AMD's 3D V-Cache uses TSMC's SoIC with copper-to-copper bonding for its desktop, server CPUs and AI accelerator. The structure provides 3x interconnect energy efficiency and 15x higher interconnect density vs. micro bumps. Intel is also developing its hybrid bonding interconnect (HBI) process for its Foveros Direct technology.

Figure 150: 2.5D/3D stacking reduces system level energy use



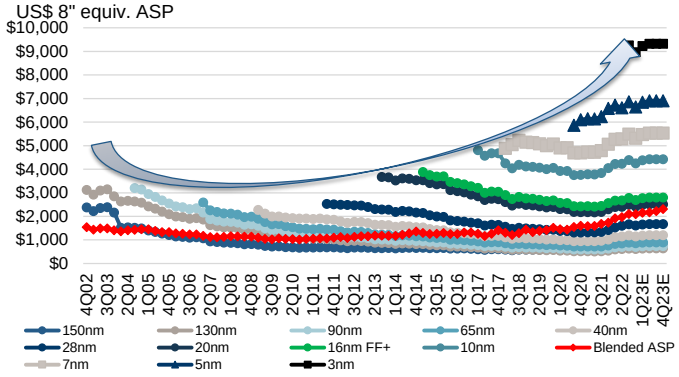
Source: AMD

Figure 151: Compute gains through system level efficiency



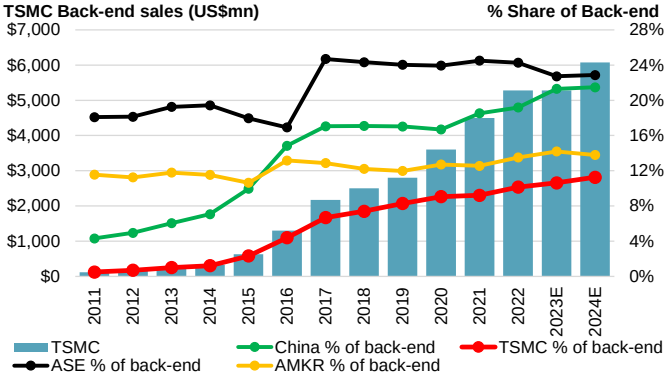
Source: AMD

Figure 152: TSMC pricing leverage migrating to advanced nodes



Source: Company Data, Credit Suisse Estimates

Figure 153: TSMC back-end sales outpace the industry



Source: Company Data, Credit Suisse Estimates

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.

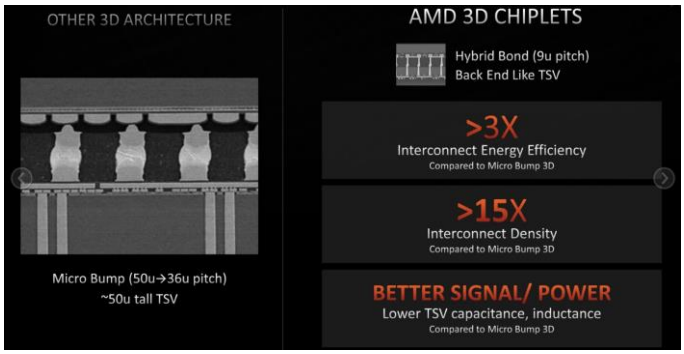
The integration has seen use in putting SRAM on separate chip connected to logic with low latency. For high performance die, advanced substrates and silicon interposers are still used to connect GPU and high bandwidth memory. Our feedback from Semicon Taiwan did note a more gradual ramp for hybrid bonding 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.

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.

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.

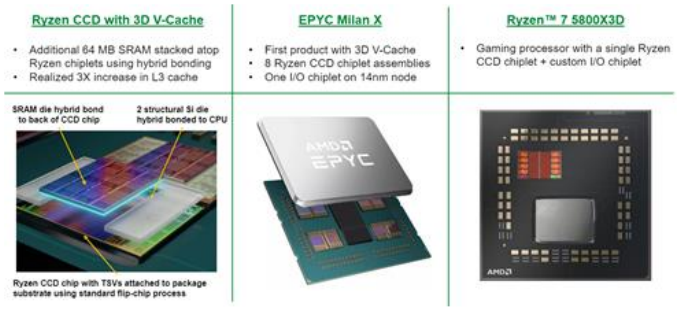
ASM Pacific also presented its solution roadmap for hybrid bonding and ultra fine-pitch TCB bonders for the packaging integration, though noted on its recent earnings call a view the inflection to high volume will take more time to reach the maturity of TCB where it has been shipping for the past 7 years 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 154: Benefits from hybrid bonding for 3D chiplets



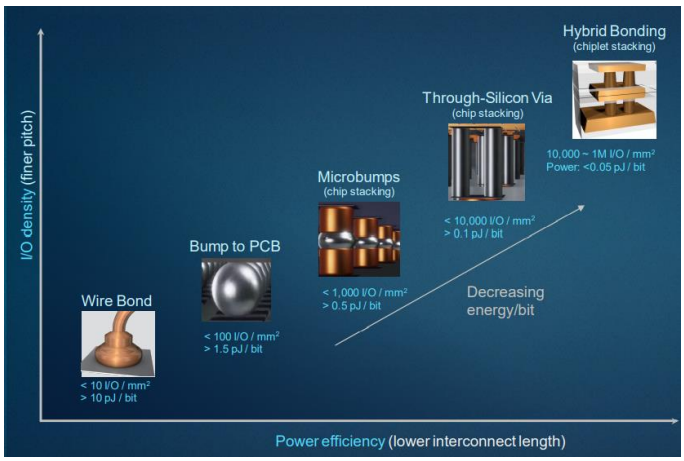
Source: AMD

Figure 155: AMD shipping the first CPUs using hybrid bonding



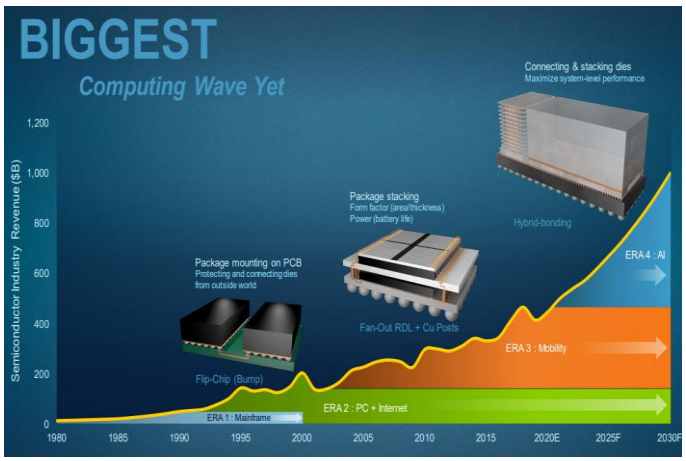
Source: BE Semi

Figure 156: Packaging scaling roadmap toward hybrid bonding



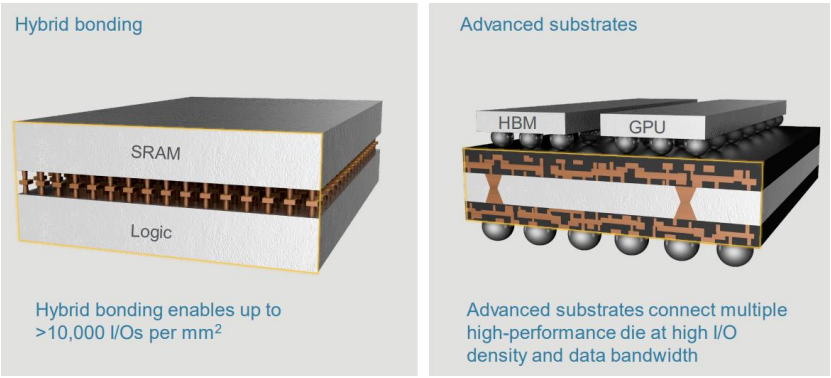
Source: Applied Materials

Figure 157: AI era to be driven by the advanced packaging



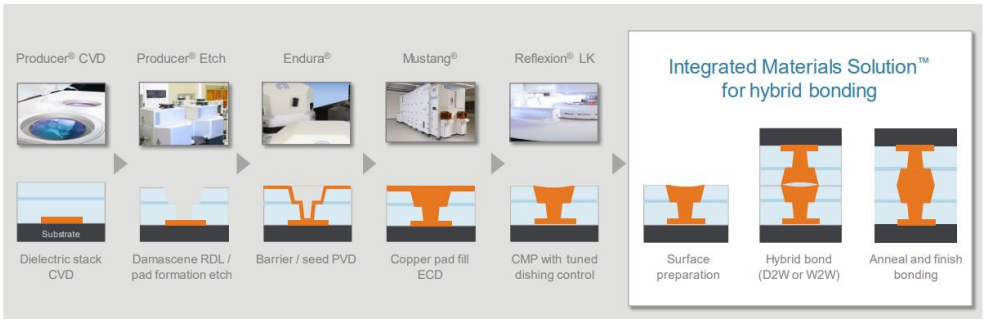
Source: Applied Materials

Figure 158: Hybrid bonding for densely connected 3D stacks



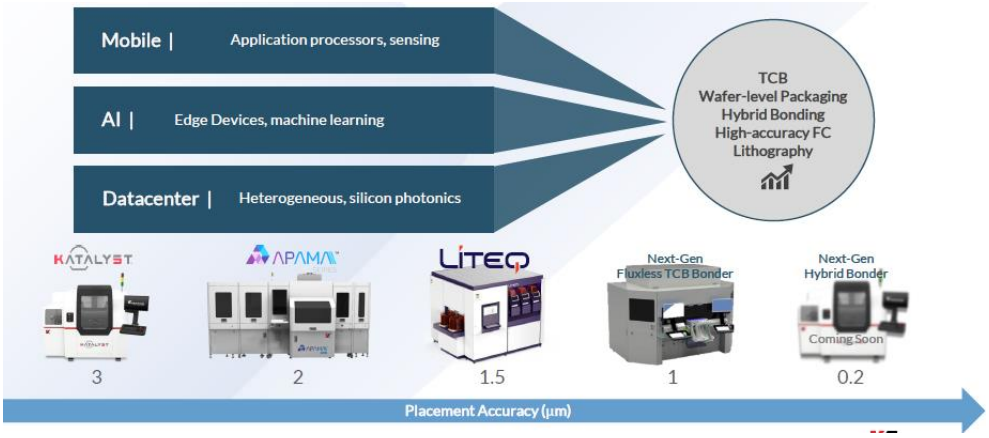
Source: Company Data

Figure 159: AMAT portfolio for hybrid bonding – partnering with Be Semi's bonder



Source: Company Data

Figure 160: K&S tool portfolio for advanced packaging



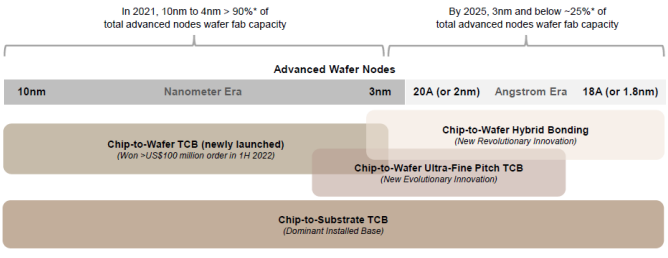
Source: Company data

Figure 161: ASMPT 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 (ie 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 (At Learning Curve Phase)

Source: Company data

Figure 162: ASMP targets chip-to-wafer TCB & Hybrid bonding



Source: Company data

Probe test supplier CHPT targeting AI/HPC projects

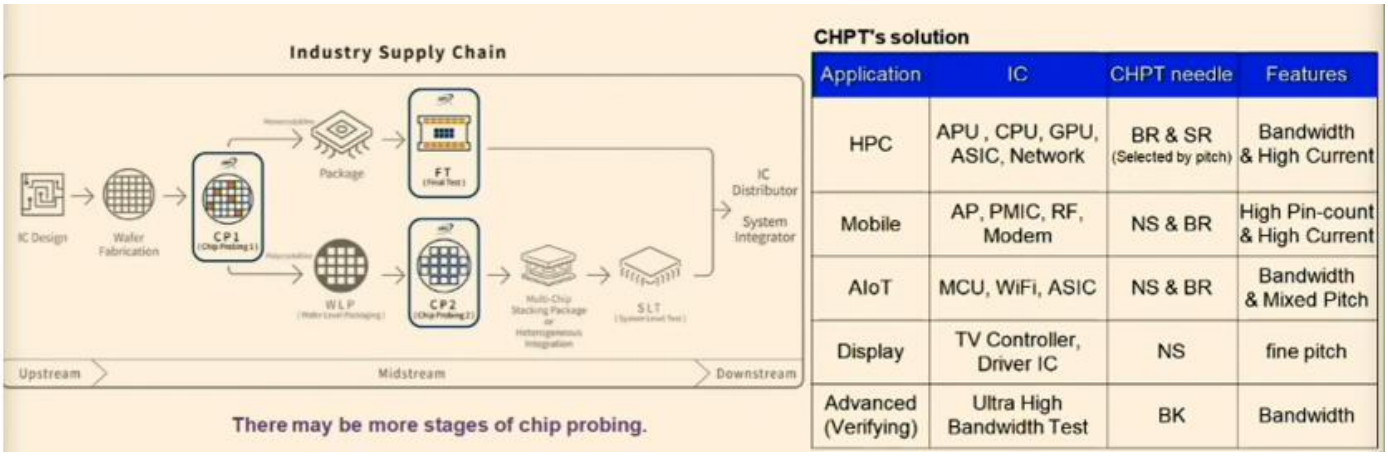
Chunghwa Precision, a Taiwan based advanced probe card supplier is also diversifying from its traditional strength in mobile processors to target high performance applications. We believe the company has about 10% of sales from AI, HPC and GPU with multiple projects in qualification.

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 163: CHPT targeting more HPC projects with its high current probe cards

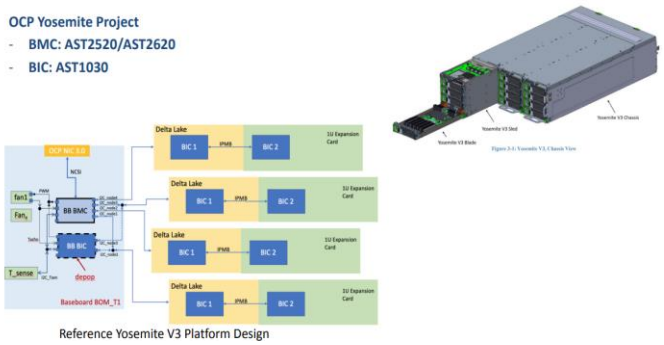


CHPT's solution

Application	IC	CHPT needle	Features
HPC	APU , CPU, GPU, ASIC, Network	BR & SR (Selected by pitch)	Bandwidth & High Current
Mobile	AP, PMIC, RF, Modem	NS & BR	High Pin-count & High Current
AIoT	MCU, WiFi, ASIC	NS & BR	Bandwidth & Mixed Pitch
Display	TV Controller, Driver IC	NS	fine pitch
Advanced (Verifying)	Ultra High Bandwidth Test	BK	Bandwidth

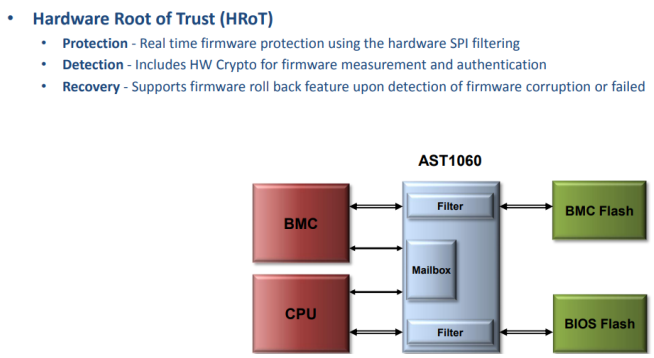
Source: Company data

Figure 164: ASpeed picking up content on CPU add-on cards



Source: Company data

Figure 165: ASpeed root of trust adds to server security



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 166: 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%
Total	**113.1**	**132.7**	**131.5**	**137.5**	**137.8**	**169.0**	**191.9**	**202.3**	**423.1**	**491.3**	**514.3**	**701.0**	**749.2**	**831.8**	**14%**
QoQ / YoY	***-18%***	***17%***	***-1%***	***5%***	***0%***	***23%***	***14%***	***5%***	***-1%***	***16%***	***5%***	***36%***	***7%***	***11%***	
ASpeed BMC units	**2.8**	**3.4**	**3.4**	**3.4**	**3.6**	**4.1**	**3.6**	**3.2**	**8.6**	**11.2**	**12.9**	**14.4**	**13.8**	**15.6**	**13%**
QoQ / YoY	***13%***	***20%***	***0%***	***0%***	***6%***	***14%***	***-11%***	***-12%***	***9%***	***30%***	***15%***	***12%***	***-4%***	***13%***	

Source: Company data, Credit Suisse

Figure 167: 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
Total	**8,107**	**8,876**	**11,496**	**13,290**	**15,584**	**17,120**	**20,573**
YoY Aspeed Growth	**15.3%**	**9.5%**	**29.5%**	**15.6%**	**17.3%**	**9.9%**	**20.2%**
YoY BMC Unit Growth	**15.9%**	**9.1%**	**30.2%**	**14.9%**	**12.0%**	**-4.4%**	**12.6%**
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
Total	**$8.8**	**$9.1**	**$9.0**	**$9.8**	**$11.2**	**$11.4**	**$11.9**
YoY Growth	**-0.3%**	**2.6%**	**-0.2%**	**8.7%**	**14.2%**	**1.6%**	**4.6%**

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	$3,199
Modeled/Actual Sales (US$k)	**$71,581**	**$80,385**	**$103,901**	**$130,503**	**$174,836**	**$195,080**	**$245,172**
Total (US$k)	**$71,581**	**$80,385**	**$103,893**	**$130,539**	**$174,836**	**$195,080**	**$245,172**
YoY Growth	**15.1%**	**12.3%**	**29.2%**	**25.6%**	**33.9%**	**11.6%**	**25.7%**
TWD/US$	30.1	30.9	29.5	27.9	29.8	30.7	30.7
Modeled/Actual Sales (NT$k)	**$2,153,519**	**$2,484,295**	**$3,063,552**	**$3,637,778**	**$5,210,122**	**$5,988,961**	**$7,526,774**
Total (NT$k)	**$2,153,519**	**$2,484,295**	**$3,063,552**	**$3,637,632**	**$5,210,096**	**$5,988,961**	**$7,526,774**
YoY Growth	**13.7%**	**15.4%**	**23.3%**	**18.7%**	**43.2%**	**14.9%**	**25.7%**
QoQ Growth							

Source: Company data

IC Design Service: Alchip, GUC and Socionext set to benefit from growing chipset customization *(Haas Liu)*

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.

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

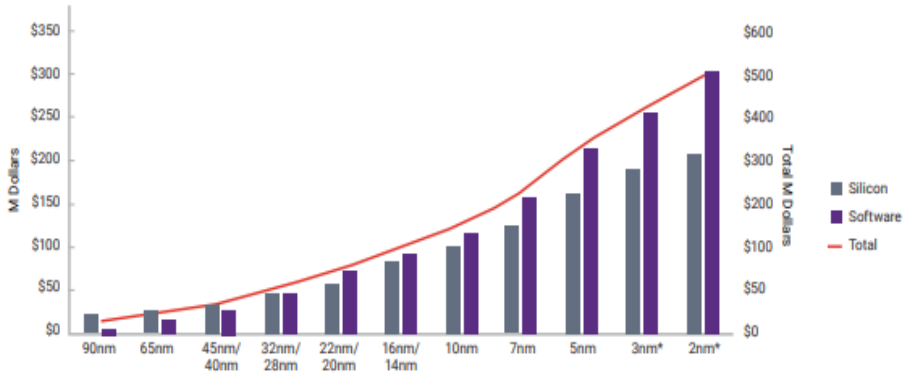
The design service capability and capacity, IP portfolio, success rate, supply chain relationship, target applications and technology/IP support are the important factors when customers choose the service provider. We compare the competitiveness for the major companies including Global Unichip, Alchip and Faraday in Taiwan, Socionext in Japan and VeriSilicon in China as below.

Memory *(Keon Han, Sanguk Kim, DJ Kim)*

Overcoming the memory bottleneck for efficient machine learning (ML) *(Keon Han)*

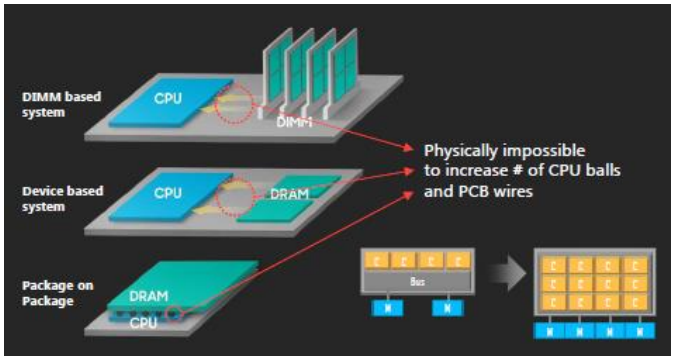
In traditional computer architecture (Von Nuemann model) the data processing for the system and the storage were managed in separate modules. However, historically, the performance gains for the compute performance have exponentially outpaced the improvement in memory speed creating a gap that continues to grow today. The advent of AI shifts compute from serial processing to parallel processing, driving a massive

Figure 168: Chipset design cost surges from US$30-40mn on 28nm to US$500mn+ on 2nm



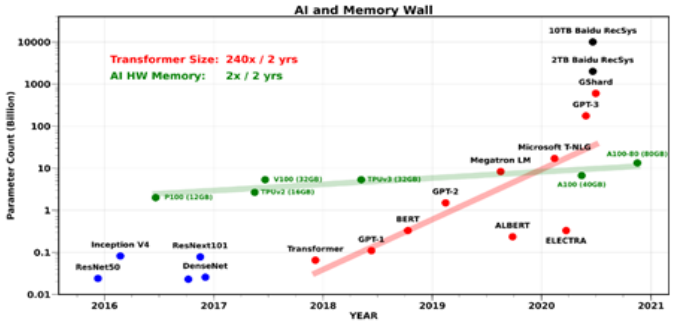
Source: Synopsys

Figure 169: Memory bottleneck feeding the processor data



Source: Samsung

Figure 170: AI Memory Wall – Memory advances cannot keep up with the performance improvement in transformer models



Source: Amir Gholami blog

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.

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 has adopted LPDDR4/5 which combined gives specific choice of subsystem by usage depending on the application and CPU/GPU requirements.

Figure 171: 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).

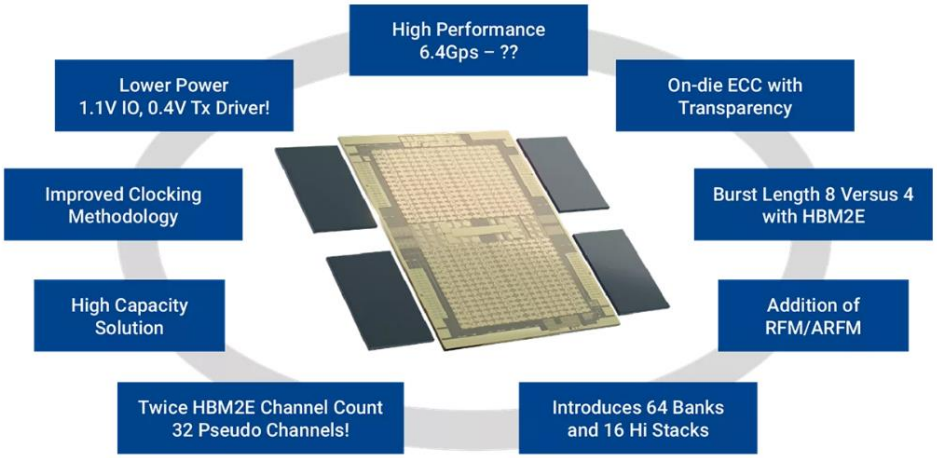
The HBM stack normally is placed adjacent to CPU/GPU connected by a substrate to address the power consumption issues as data movement distance between the memory stack and CPU/GPU resulting in power savings. In addressing the cost issue, HBM3 does not come cheap. Today, it is believed that HBM3 is 2x more expensive than the DDR5 counterpart. It could be too expensive to go mainstream. While traditional DDRs 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.

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.

Figure 172: HBM3 offers power/performance benefits over HBM2E



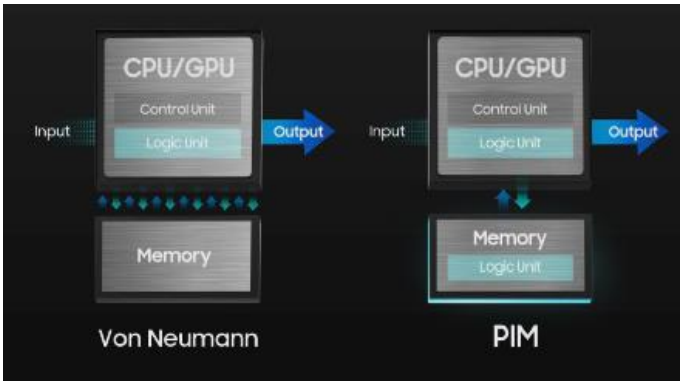
Source: Synopsys

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.

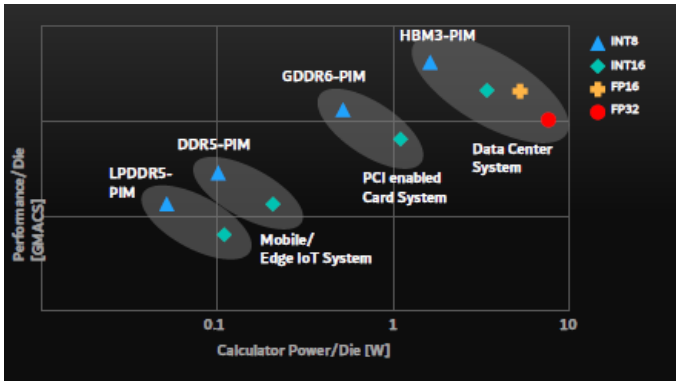
Similarly, SK Hynix has also introduced PIM (they call in AiM – accelerator in memory) in February 2022 with the sample on GDDR6. For now, it is focused more on PIM support on more conventional chips versus on HBM3, believing it is simpler for customers to adopt. SK Hynix researchers did work with Purdue 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 173: Using PIM to overcome the memory bottleneck



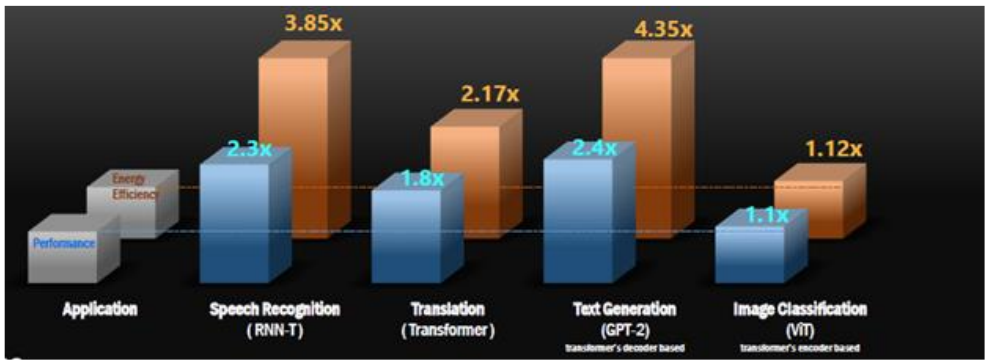
Source: Samsung

Figure 174: Prospective PIM-supported data format



Source: Samsung

Figure 175: Expansion of PIM technology, LPDDR5-PIM



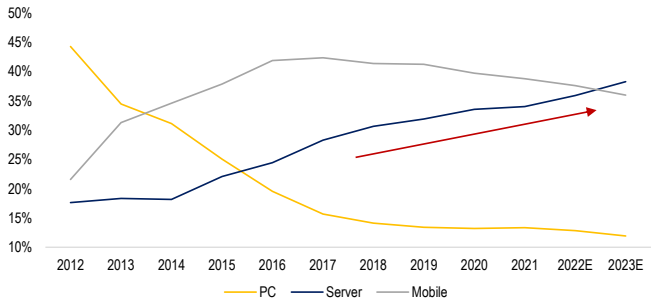
Source: Samsung

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.

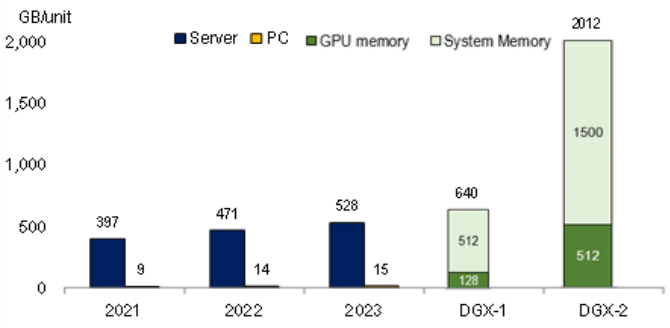
When we compare a simple DRAM requirement for PCs of 15GB per unit in 2023, average server DRAM requirement towers above PC at 528GB per unit. As a sample, for NVIDIA's DGX-1 deep learning system introduced in 2016, it contained total DRAM memory for both accelerator and system at 640GB. It was followed 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 176: Servers the largest segment for DRAM now - AI likely to continue lifting servers' portion of consumption



Source: Credit Suisse Estimates

Figure 177: Average density for a server for DGX-1 and DGX-2 vs. a traditional PC



Source: Samsung

Figure 178: 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 *(Pauline Chen, Akinori Kanemoto, Sanguk Kim)*

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.

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.

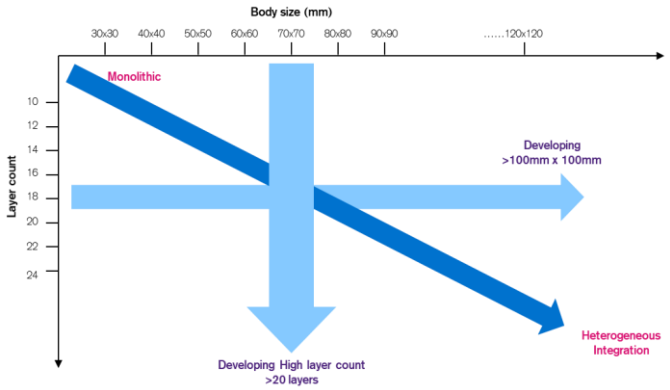
Compared to ASIC, substrates design for GPU typically has a larger form factor and requires more power and cooling, so that it can handle a wider range of tasks. On the other hand, ASICs are typically more optimized for a specific task, which makes it more efficient to achieve higher performance with 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.

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.

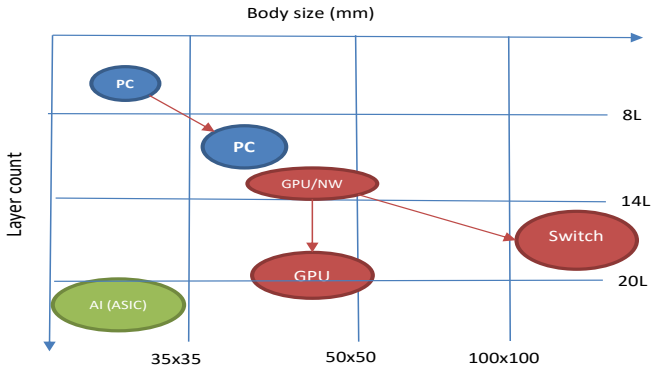
Nevertheless, we acknowledge that silicon substrates' higher production costs (still 3x higher currently) could limit 3D packaging adoption in high-end i.e., HPC, AI and data centers. As a result, organic substrates remain a more cost-effective solution for applications including PC, servers, networking and ADAS in EV, which still accounts for the majority of organic substrate demand.

Figure 179: Trend of Advanced FCBGA substrate



Source: Company data, Credit Suisse

Figure 180: FCBGA substrates by application



Source: Company data, Credit Suisse

Figure 181: 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	SoIC
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

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.

Asia Hardware Stock implications *(Pauline Chen)*

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.

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.

Figure 182: 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	

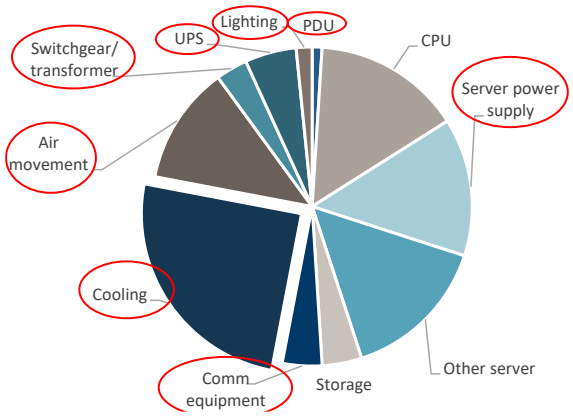
Source: Company data, Credit Suisse
Note: "vvv" stands for key application, "v" stands for minor contribution

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.

Delta well positioned for new data center architecture

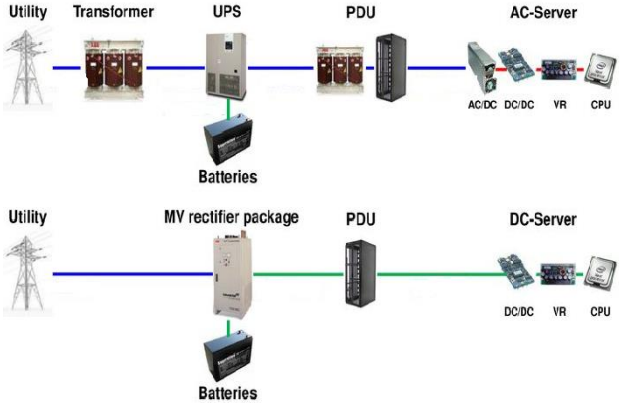
We expect the new data center architecture, including DC power architecture, modularized design, and convergencies of cooling technologies, to drive growth opportunities for the supply chain. **Delta Electronics** is well positioned in this trend, given its full range of product offerings and above-industry-average 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 183: Energy consumption in a data center, and Delta's product offerings



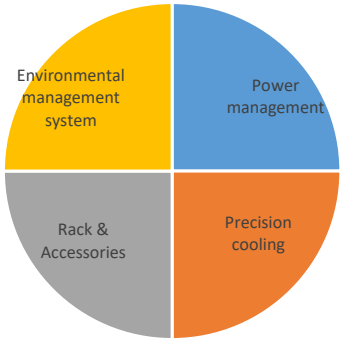
Source: Company data, Credit Suisse

Figure 184: Simplified power supply architecture



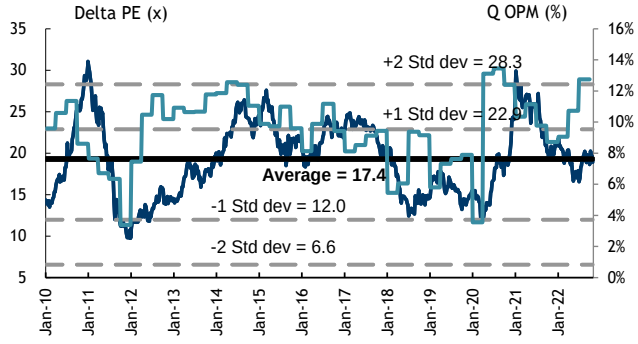
Source: Company data, Credit Suisse

Figure 185: Delta's data center offerings



Source: Company data, Credit Suisse

Figure 186: Delta—OPM back to record high, but P/E is not



Source: Company data, Credit Suisse

Asia Cloud IT Infrastructure Sector *(Jerry Su, Harvie Chou)*

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.

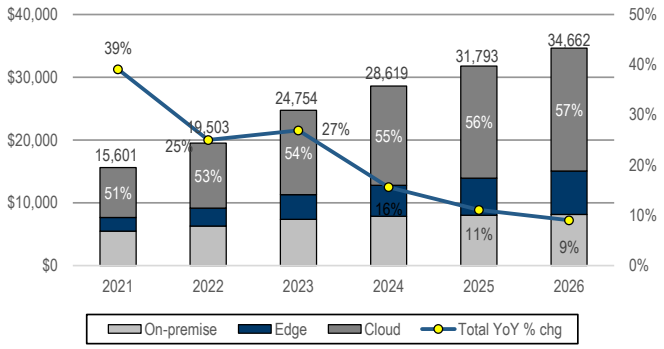
According to IDC, AI servers are servers that run AI platforms (AI applications development including most AI training), AI applications (AI models execution covering AI inferencing), and/or AI-enabled applications (applications with AI functionality). In addition, based on its latest forecasts, it expects 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.

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 versus 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 187: AI server infra for +15% CAGR in 2022-26 with cloud as key driver

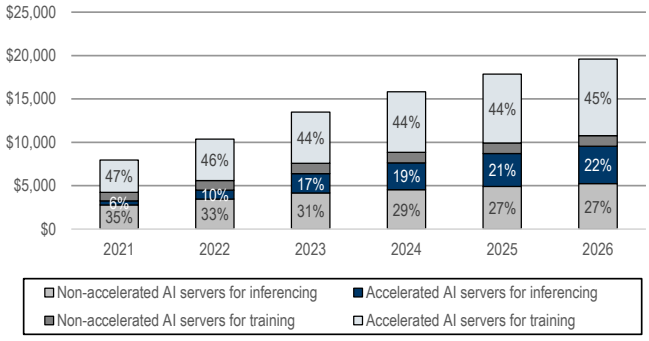
Sales in US$ mn



Source: Company data, IDC, Credit Suisse

Figure 188: Accelerated AI servers with co-compute edging to 70% of mix by 2026

Sales in US$ mn



Source: Company data, IDC, Credit Suisse

Figure 189: 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

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 versus 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 4th Gen Xeon CPU and 8 Nvidia's SXM5 H100 GPU module which could cost around ~US$200K, significantly higher than the blended server 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 preferrable for high-end AI inferencing and T4 could be adopted for entry-level servers.

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 190: Key spec comparison of Nvidia server GPU for parallel compute

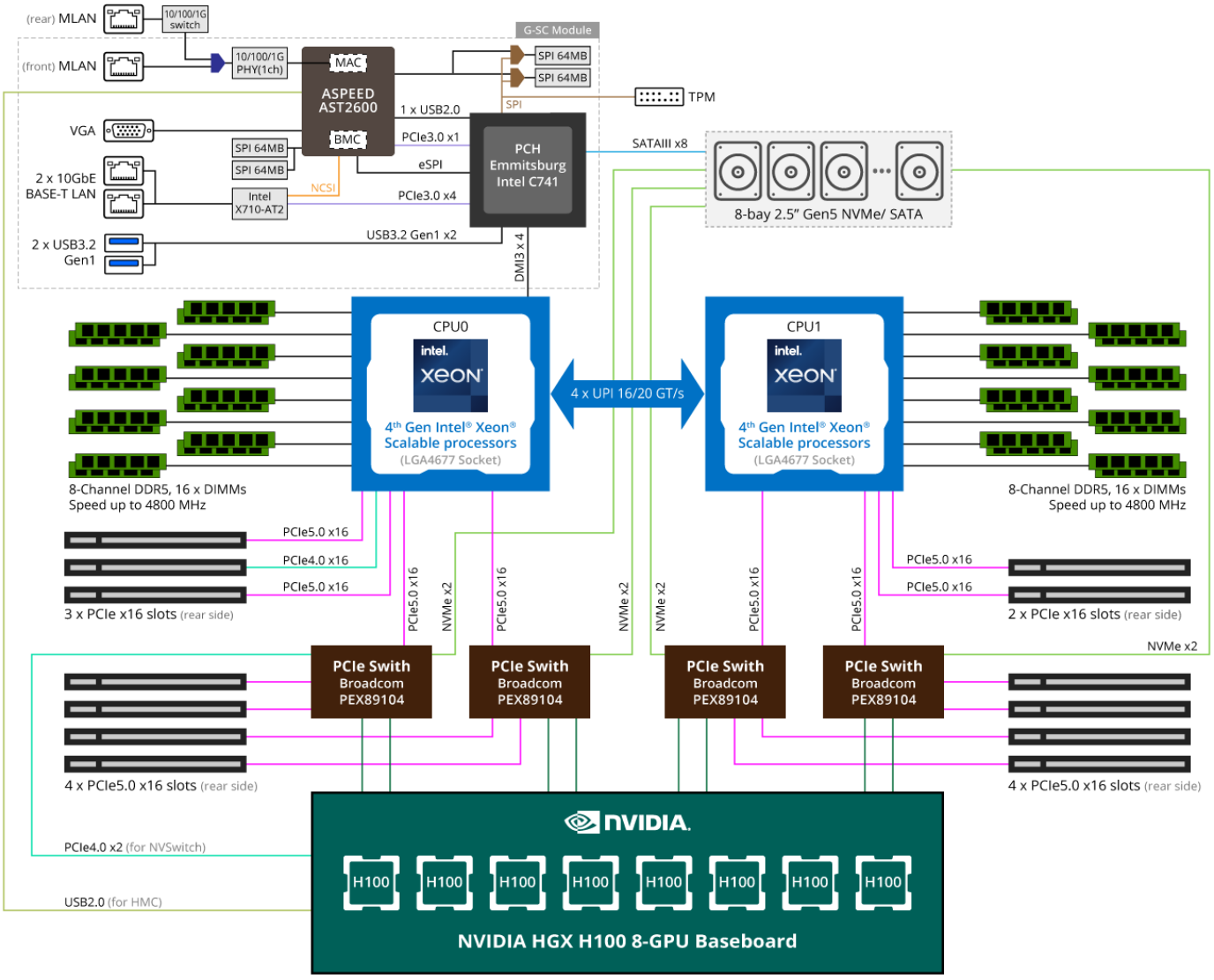
	H100 SXM	H100 PCIe	A100 SXM	A100 PCIe	A30
FP64	34	26	9.7		5.2
FP64 Tensor Core	67	51	19.5		10.3
FP32	67	51	19.5		10.3
TF32 Tensor Core	989	756	156 | 312		82 | 165
BFLOAT16 Tensor Core	1,979	1,513	312 | 624		165 | 330
FP16 Tensor Core	1,979	1,513	312 | 624		165 | 330
GPU memory	80GB	80GB	80GB HBM2e		24GB HBM2
GPU memory bandwidth	3.35TB/s	2TB/s	1,935GB/s	2,039GB/s	933GB/s
Max thermal design power (TDP)	Up to 700W	300-350W	300W	400W	165W

Source: Company data, Credit Suisse

The below figure is the block diagram of Gigabyte's AI server with Nvidia's HGX H100 8 GPU module. Besides adopting the latest Intel 4th 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 191: Block diagram of Gigabyte's HPC server G593-SDO for AI applications

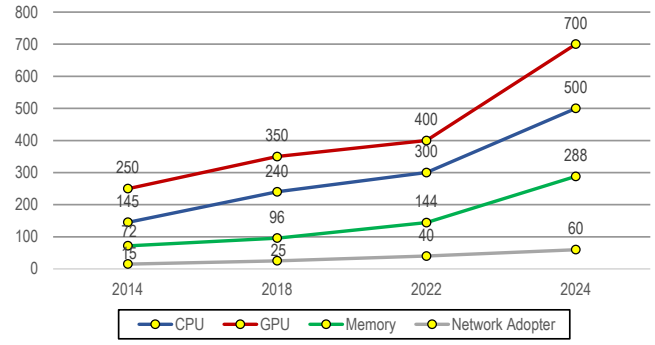


Source: Company data

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 192: Rising TDP across all major cloud components led by GPU and CPU…

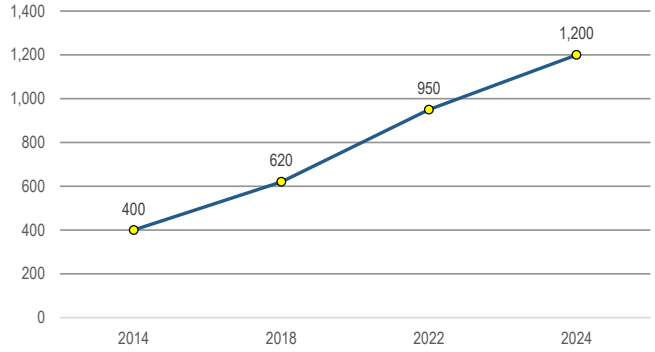
Units in Watts



Source: Company data, Credit Suisse

Figure 193: … and total server system as a result also saw growing energy consumption

Units in Watts for a typical 2 socket server



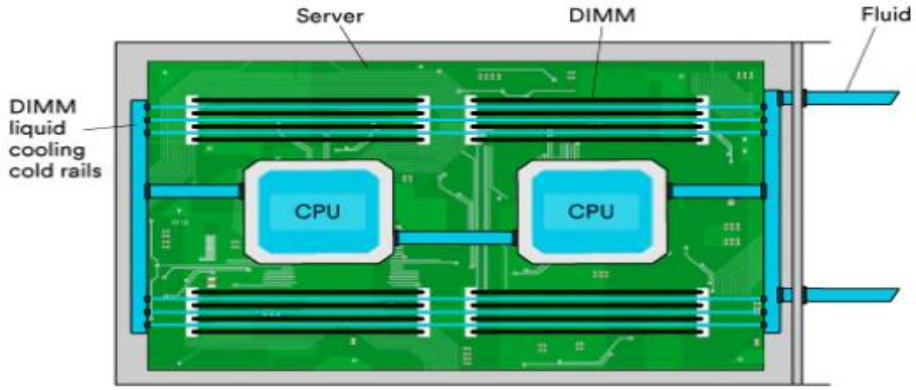
Source: Company data, Credit Suisse

Figure 194: Gigabyte's AI server product portfolio at a glance

	AI platform - G493	Training - G593-SD0	Inferecing - 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

Figure 195: Liquid cooling drives better efficiency with heat transported away by special fluid within cold rails



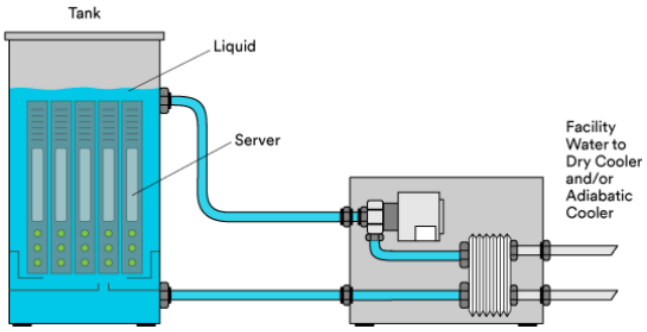
Source: Company data

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.

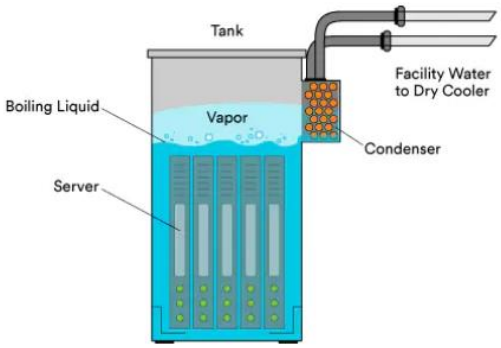
Overall, we see growing value proposition to both cloud ODMs and OEMs from hardware builds perspective in terms of both more complicated board designs and latest cooling solutions adoption along with rise of AI infrastructure. For ODMs, specifically, our checks suggest growing projects engagement by 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.

Figure 196: Single-phase immersion cooling operates under sealed tank environment where fluid remains in liquid state



Source: Company data

Figure 197: Two-phase immersion solution instead cools down hardware with fluid going through evaporation/condensation



Source: Company data

Industrial PC looking for AI opportunities *(Pauline Chen)*

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 AL 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.

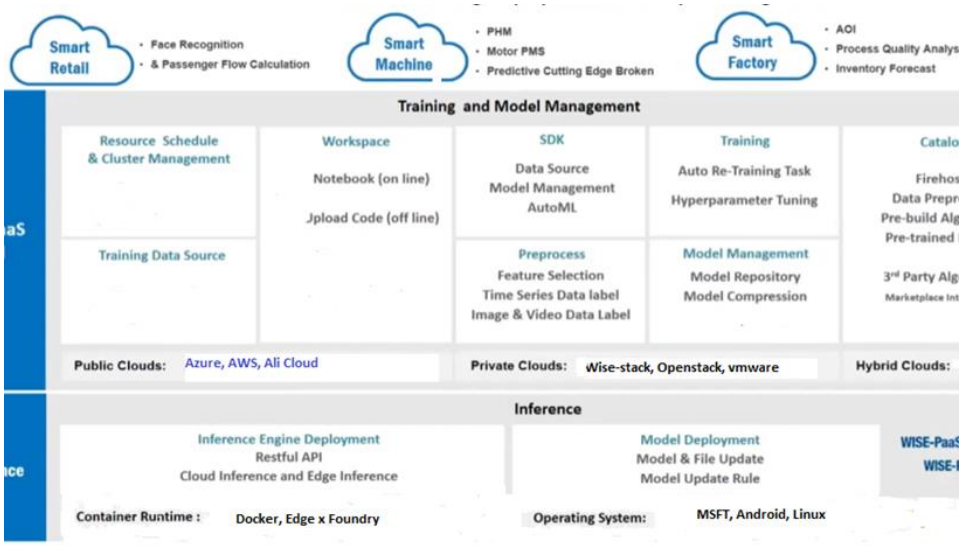
Edge AI opportunities for IPC

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 madding 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.

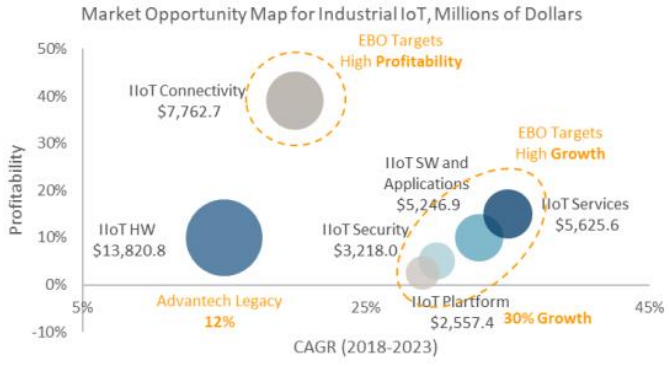
For its software platform, Advantech specifically targets at four verticals, including EMS (Energy Management System), iFactory, iRetail, and machine management. Advantech views itself not just an edge hardware provider, but more a resource integrator for AIoT ecosystem (e.g. to connect and support different domain knowhow suppliers in AIoT).

Figure 198: Advantech's WISE-PaaS/AFS architecture



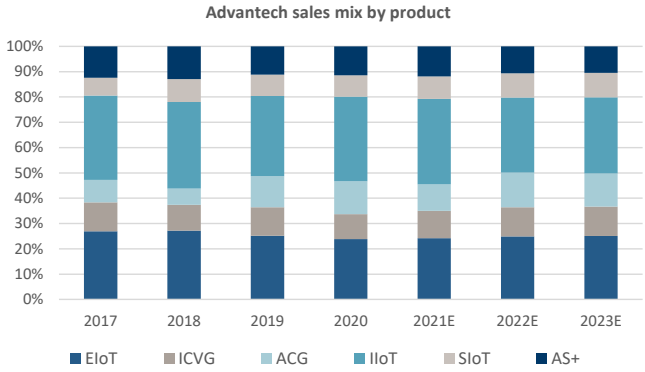
Source: Company data

Figure 199: Gartner market opportunity for IIoT



Source: Company data, Credit Suisse

Figure 200: Advantech's sales mix by product

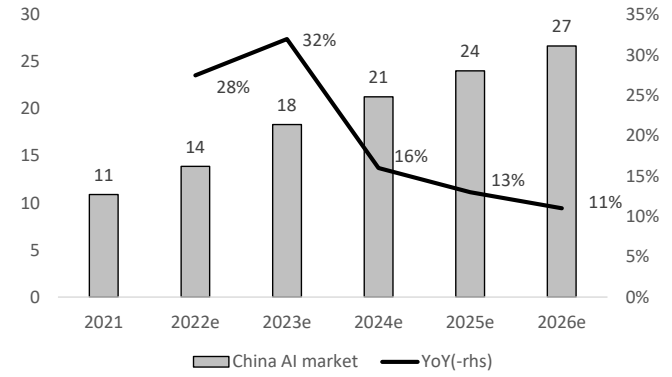


Source: Company data, Credit Suisse

China Technology Sector *(Kyna Wong, Chaolien Tseng, Yufeng Shen, Clive Cheung, Edward Liu)*

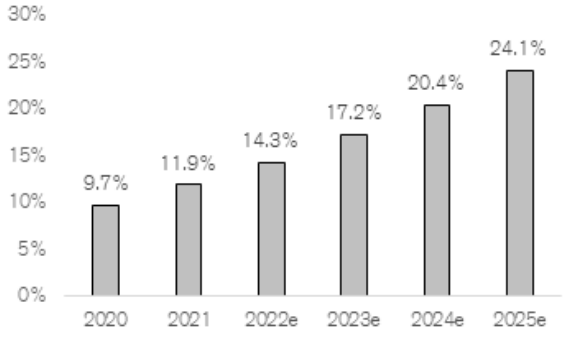
We see many applications and scenarios applicable to ChatGPT and the underlying AI-generated content (AIGC) technology in China along with AIoT, smart city, smart education, smart manufacturing, smart business, Metaverse developments etc. AI has come to play an important role in China's 'Made in China 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.

Figure 201: China AI is expected to grow 41.3%CAGR 20-25E (Rmb Bn)



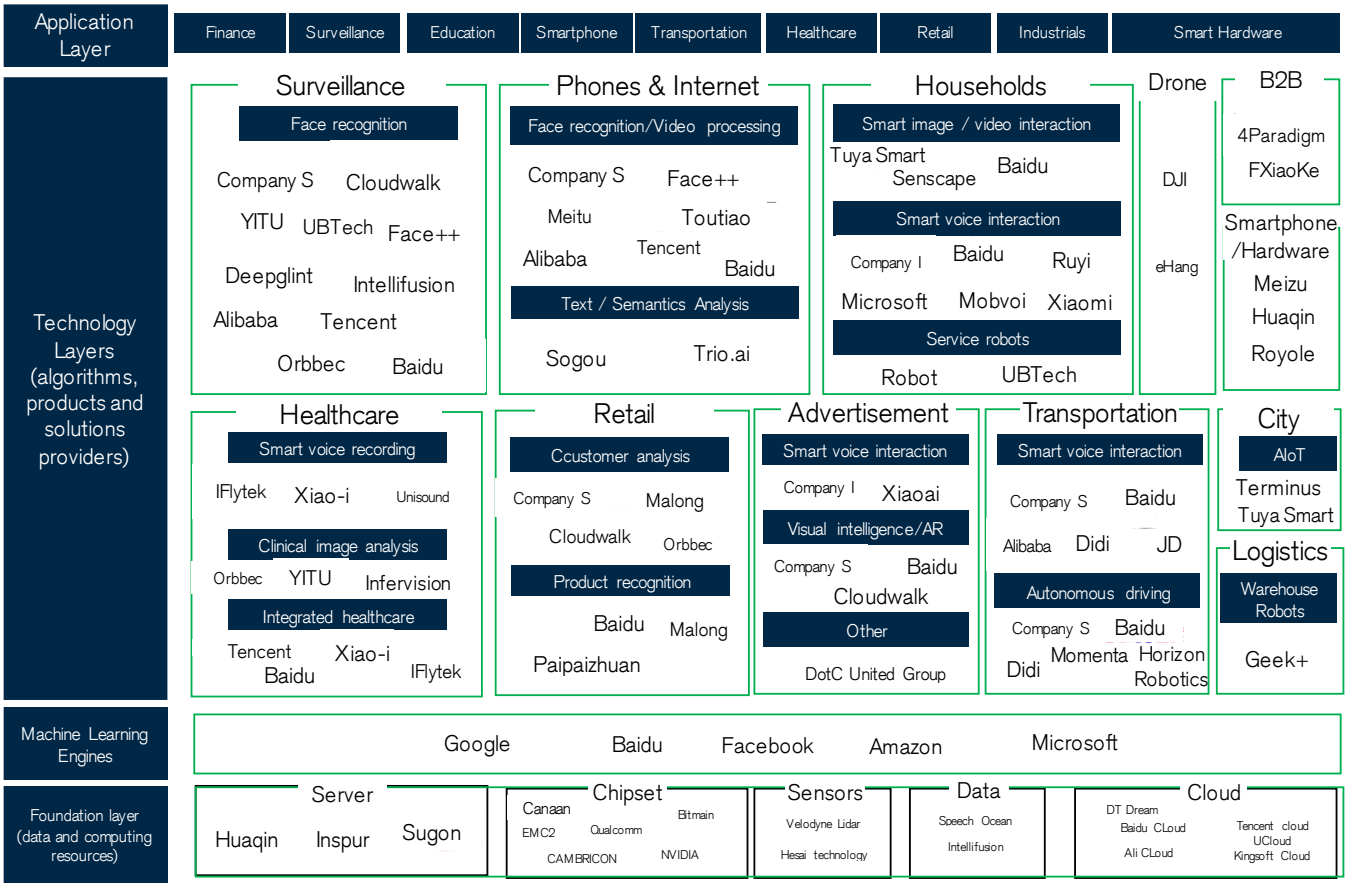
Source: Frost and Sullivan, Credit Suisse

Figure 202: China AI as percentage of Global ramp up



Source: Frost and Sullivan, Credit Suisse

Figure 203: 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.

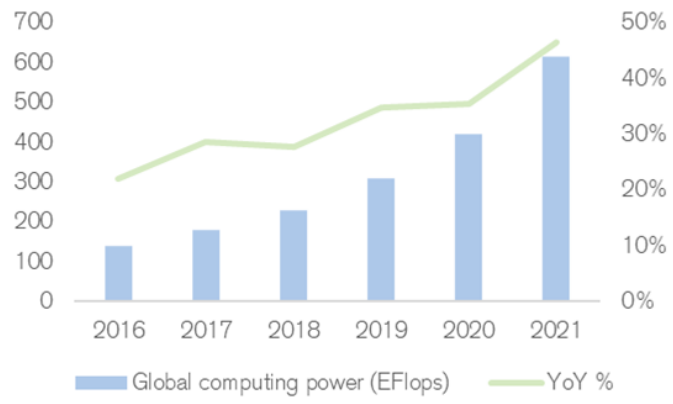
According to Frost & Sullivan, China's AI market has reached approximately RMB186 billion in 2020, accounting for 9.7% of the global AI market, and is expected to reach RMB1,046 billion by 2025, accounting for 24.1% of the global AI market. We expect that China AI industry outgrows in near future thanks 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.

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.

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.

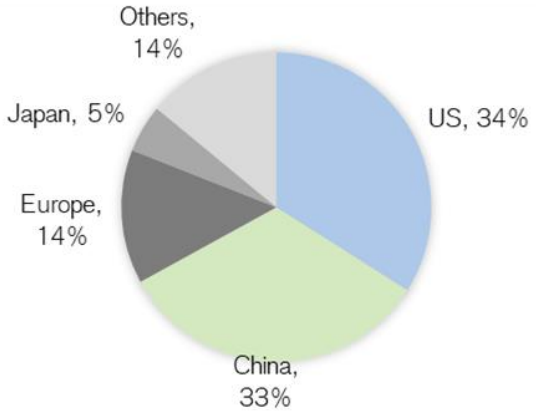
China, which represented 33% of global computing power by 2021, will deliver a similar growth pattern. Its basic computing power reached 95EFlops by 2021, at a 24% CAGR in 2016-21, supercomputing reached 3EFlops by 2021, at a 30% CAGR in 20216-21, while AI computing reached 104EFlops by 2021, at an 85% 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 204: Global computing power grow at a 26% CAGR in 2019-21



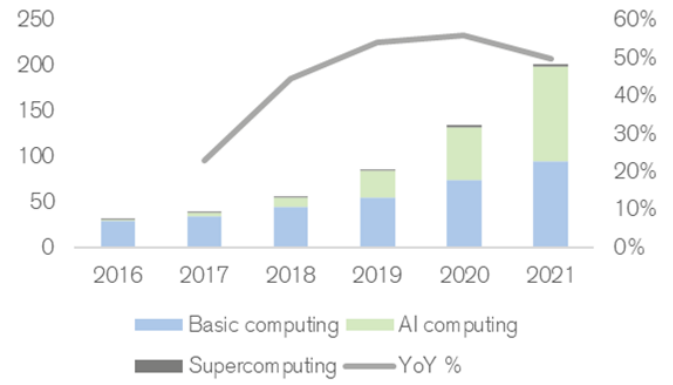
Source: CAICT, IDC, Gartner, HPC TOP100

Figure 205: China represented 33% of global computing power by 2021



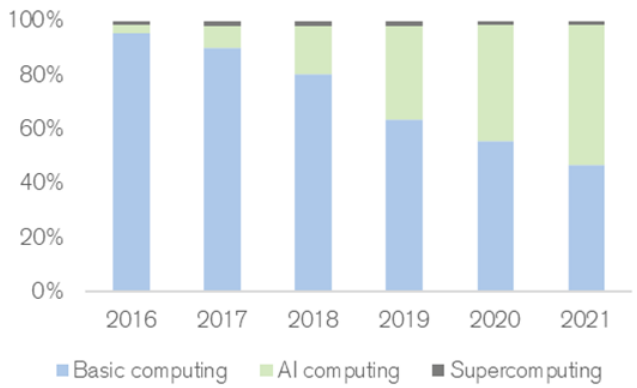
Source: CAICT, IDC, Gartner, HPC TOP100

Figure 206: China computing power outgrows the global



Source: CAICT, IDC, Gartner, HPC TOP100

Figure 207: AI computing has been the largest growth driver



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.

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.

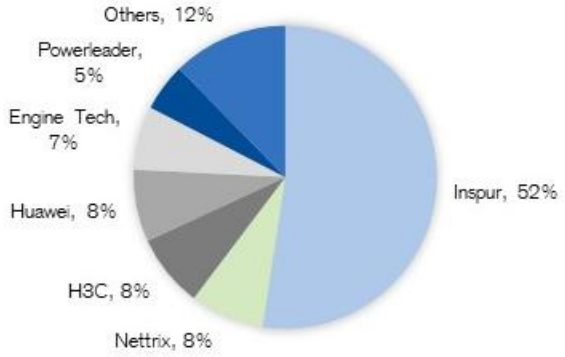
Optical transceivers are also expected to be a key area for investments, to improve the bandwidth to support the enormous data traffic effected to be generated new AI chips as well as providing high data rates support to the data center infrastructures for AI/ML applications. We have already seen upgrades 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.

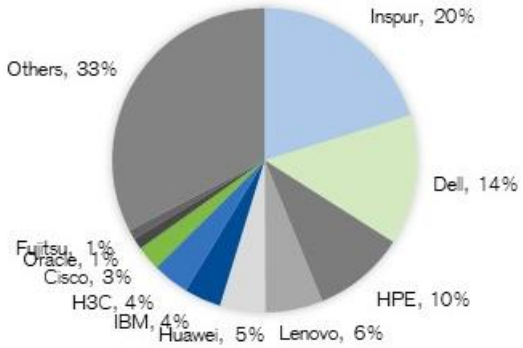
Currently, one of the most representative use cases for AI in China is computer vision. The computer vision software market in China is projected to reach RMB101.7 bn in 2025, representing a CAGR of 43.5% from RMB16.7 bn in 2020.

Figure 208: Inspur represented 52% of China AI server market in 2021



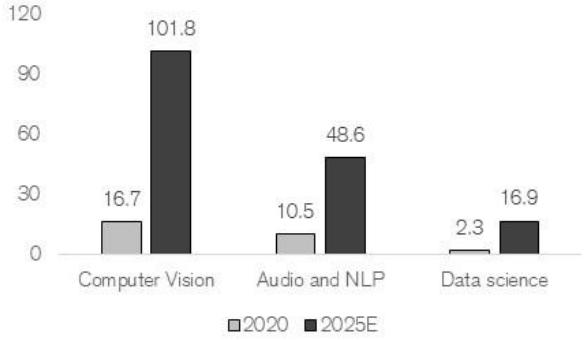
Source: IDC

Figure 209: … and 20% of global AI server market in 1H21



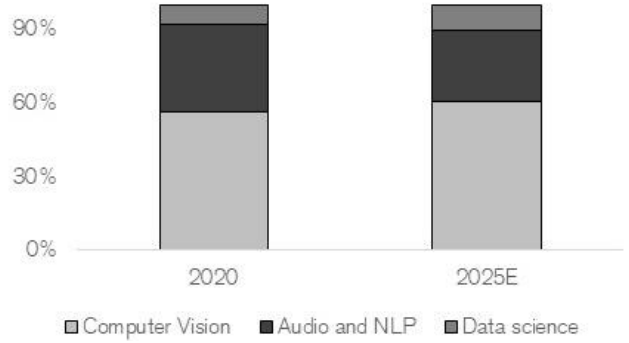
Source: IDC

Figure 210: CV is expected to grow CAGR of 43.5% in 20-25e



Source: Frost and Sullivan, Credit Suisse

Figure 211: Computer vision software is the largest segment of the global AI software market at 46.9% in 2020



Source: Frost and Sullivan, Credit Suisse

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.

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.

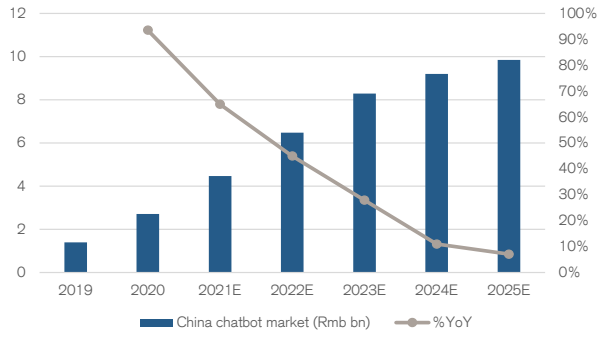
The telcos are also seen to accelerate the integration of new AI elements into their convergence offerings and also expanding their new infrastructure investment to support their overall digital business growth. We see current AI offerings by the telcos are mostly at the PaaS layers, but commercially 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.

European Technology Hardware has strong potential to benefit from generative AI *(Adithya Metuku, Sarah Roberts)*

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.

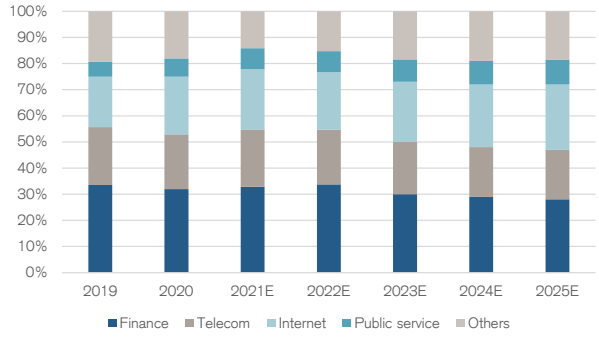
This trend is expected to drive increasing demand for data creation/capture, storage, analytics and transmission. Semiconductors are critical enablers of each of these steps and for driving future trends. Hence, they are key enablers and beneficiaries of generative AI.

Figure 212: China chatbot market can grow at 29% CAGR from 2020 to 2025



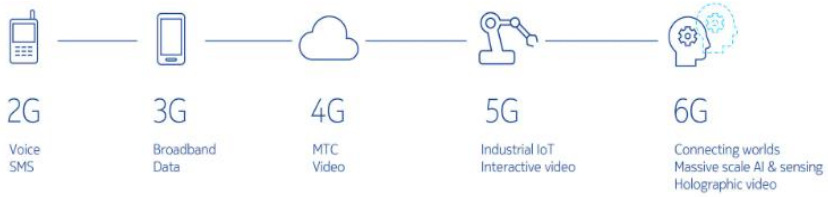
Source: iResearch

Figure 213: Finance and Internet industries accounted for over 50% of usage



Source: iResearch

Figure 214: Next generation networks key to enabling higher data transmission



Source: Nokia

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.

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.

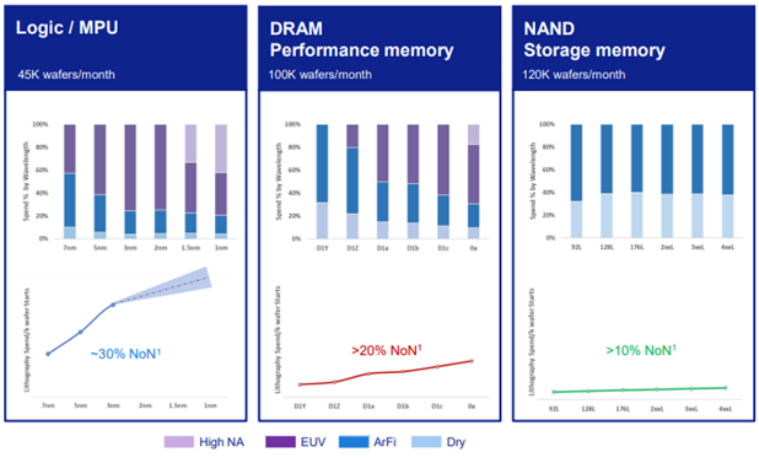
In particular, we highlight ASML and ASM International as our top picks to benefit from increasing demand for generative artificial intelligence and more generally artificial intelligence.

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.

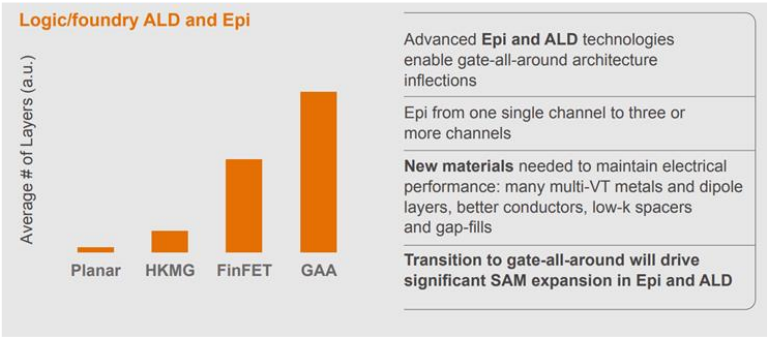
Similarly, leading edge logic and memory devices are increasingly using single wafer atomic layer deposition and epitaxy steps to enable novel device structures. ASM International is the leader in single wafer atomic layer deposition tools and is the #2 vendor of epitaxy tools.

Figure 215: Lithography spending rising materially on future logic and memory nodes



Source: Company data, Credit Suisse

Figure 216: 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.

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.

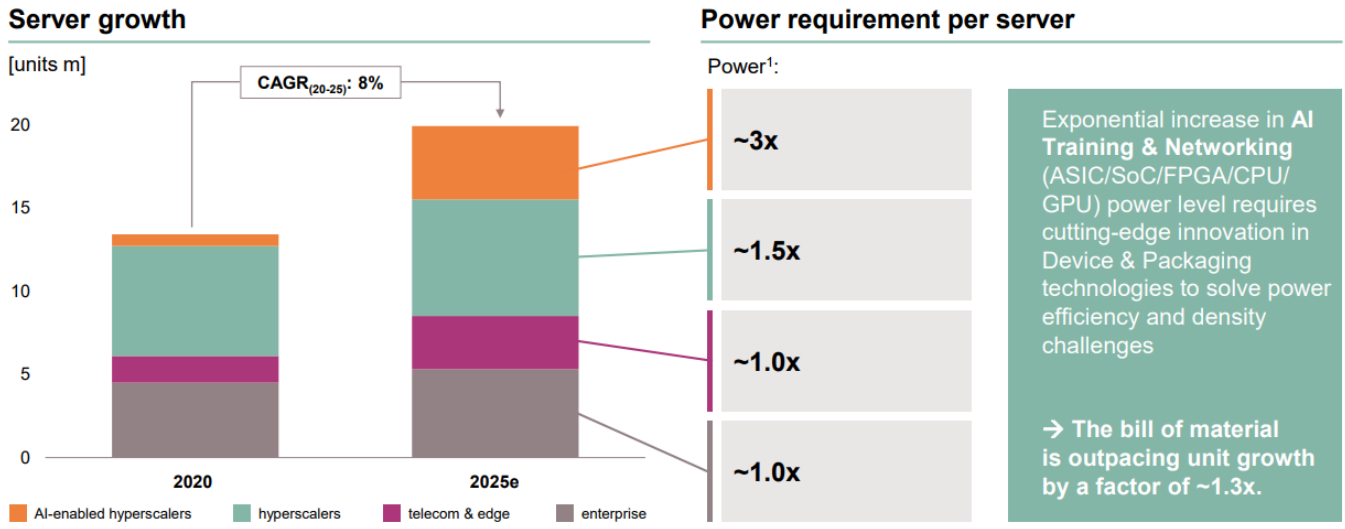
Soitec (SOIT.PA): As explained earlier, generative AI may drive demand for faster and more pervasive connectivity. Soitec is the global leader in RF-SOI substrates which contribute >60% of the company's revenue. RF-SOI content can triple in a device with 5G mmWave connectivity versus a device only with 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 believe there will be content growth for Soitec to benefit from. 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 *(Danny Chan)*

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 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.

Figure 217: Servers for artificial intelligence require materially more power than traditional enterprise servers.



Source: Infineon technologies.

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 downtrend 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. 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.

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's 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.

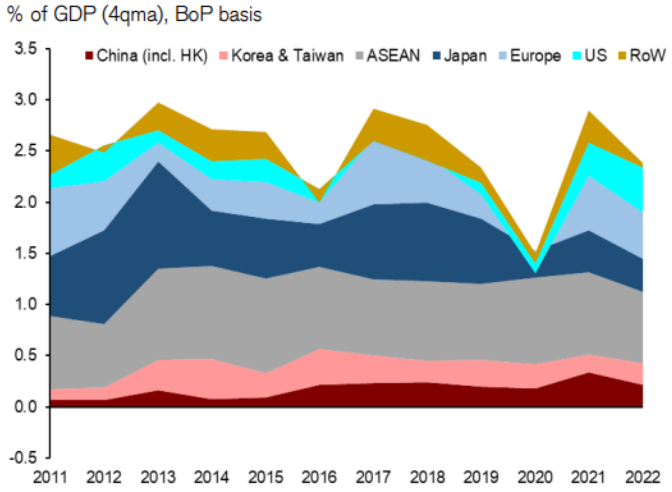
Equipment Makers Weaving in More AI into Their Products

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.

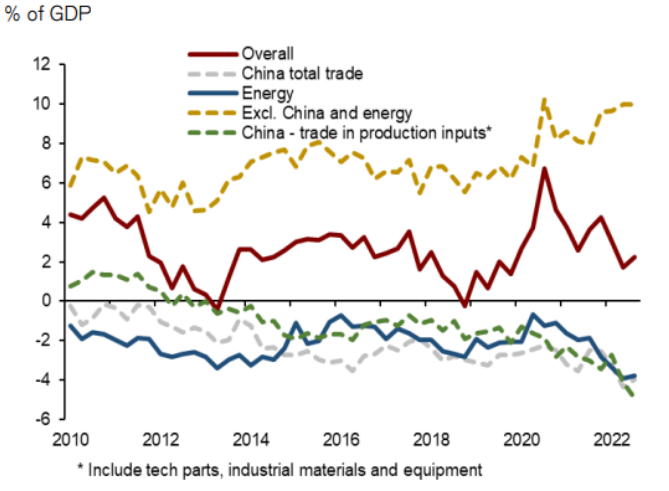
For equipment, we estimate ASEAN players are merely one to two years behind those of global suppliers (based on feedback from the end-customers). We expect the gap to continue in the foreseeable future, as chipmakers will likely continue to depend on the incumbent suppliers for most of their

Figure 218: ASEAN-5 FDI by source



Source: Haver Analytics®, Credit Suisse

Figure 219: ASEAN-6 trade balance



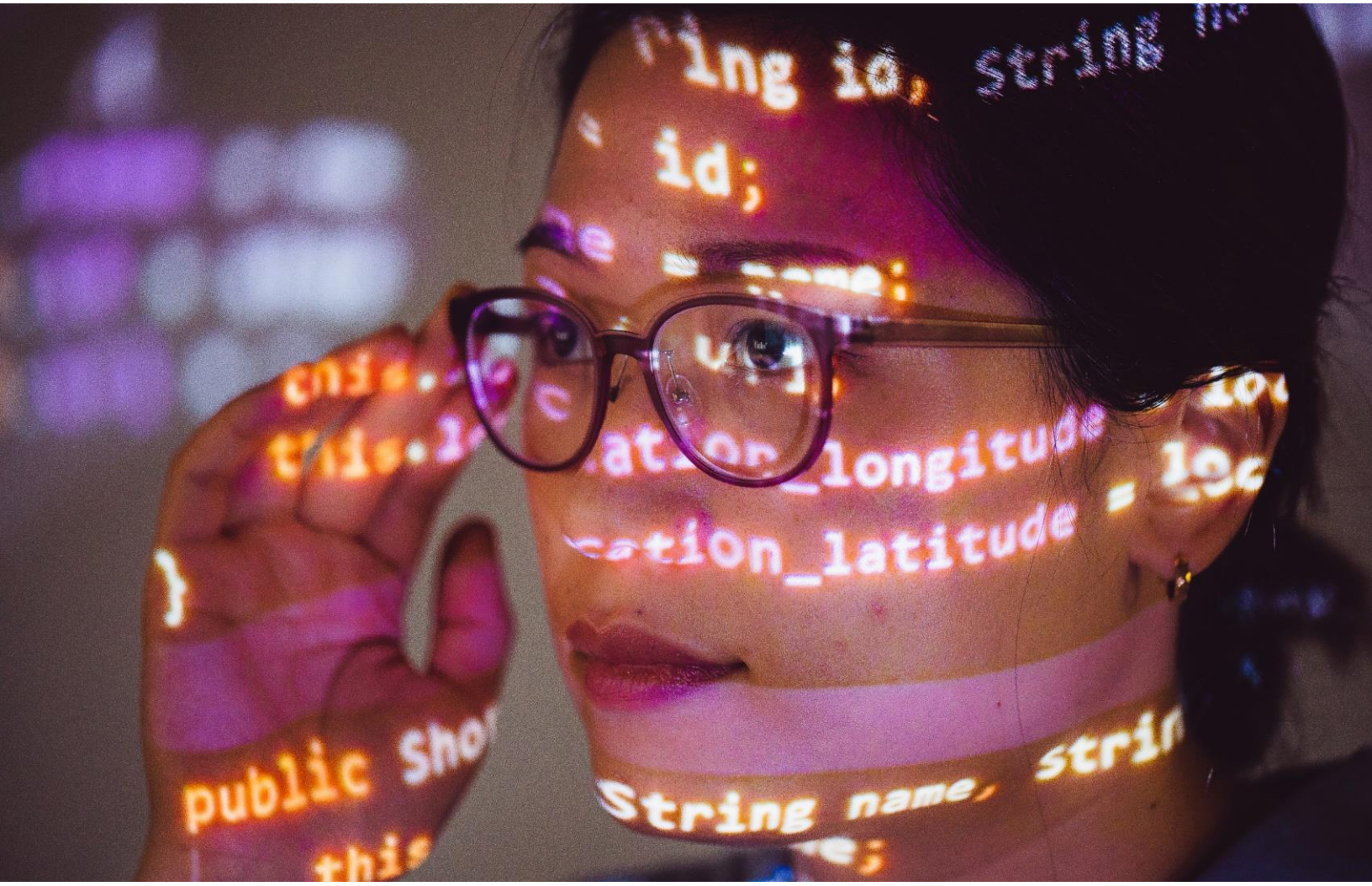
Source: Haver Analytics®, Credit Suisse

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.



“ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness… it’s a preview of progress…” – Sam Altman, CEO of OpenAI

Risks and Regulatory Concerns with ChatGPT and AI Technologies

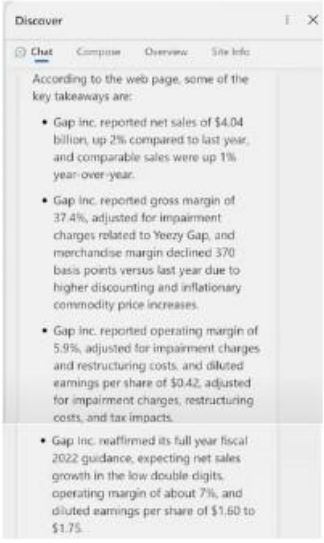
As more businesses and organizations embrace ChatGPT and look to incorporate it (and the OpenAI family of models more broadly) into their daily operations, we note several key risks posed by ChatGPT and similar technologies. We would also highlight there are growing concerns around the state-of-the-art AI capabilities and potential impacts on society altogether.

- **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 owing 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.

Figure 220: OpenAI CEO - Limitations of ChatGPT

Source: Twitter

Figure 221: Bing Chatbot Demo: Key Takeaways GPT 3Q22 EPS



Source: Company data, Credit Suisse

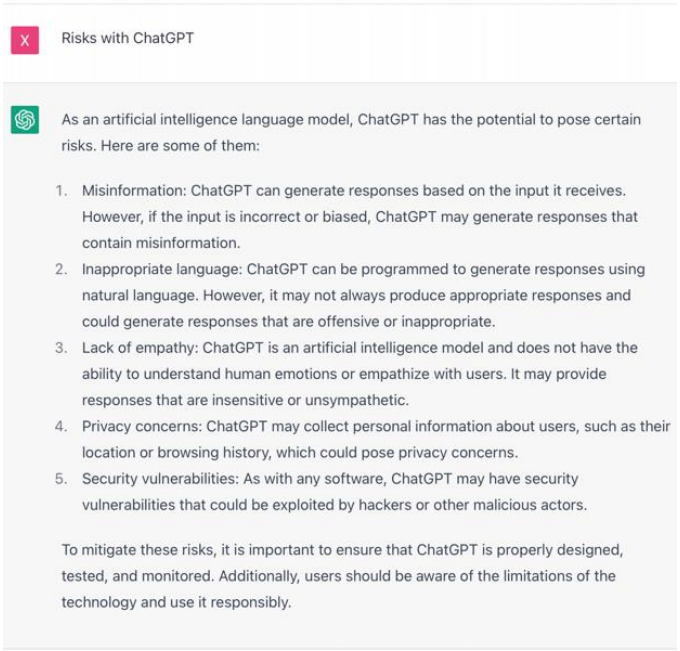
Further, we submitted the following prompt into ChatGPT and below is what it noted as risks when using the tool/service.

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 223, 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.
- **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.

Figure 222: Answers from ChatGPT Regarding the Risks of Using of Its Service



Source: OpenAI.

Figure 223: Open AI CEO, Sam Altman Tweet on the AI Regulation

Source: Twitter

DISCLAIMER

While this post has been prepared by the Securities Research business of Credit Suisse AG, its subsidiary or affiliate ("CS") and may contain references to Securities Research reports and/or Securities Research analysts, it is for information only and does not constitute research by CS. Furthermore, this post is not to be regarded as a sales prospectus or an offer or solicitation of an offer to enter in any investment activity. This post does not take into account your specific investment objectives and needs nor your financial situation. No representation or warranty, either expressed or implied is provided in relation to the accuracy, completeness or reliability of the information contained herein, nor is it intended to be a complete statement or summary of the developments referred to in this post and any liability therefore (including in respect of direct, indirect or consequential loss or damage) is expressly disclaimed. The information and any opinions expressed in this post are subject to change without notice and may differ or be contrary to opinions expressed by other business areas or groups of CS as a result of using different assumptions and criteria. CS is not providing any financial, economic, legal, accounting, or tax advice or recommendations in this post. In addition, the receipt of this post alone by you is not to be taken to constitute you as a client of any CS entity. CS is under no obligation to update or keep current the information contained in this material. Please consult with your client advisor before taking any investment decision. This material is issued and distributed in the United States by CSSU, a member of NYSE, FINRA, SIPC and the NFA, which accepts responsibility for its content. Clients should execute transactions through a Credit Suisse entity in their home jurisdiction unless governing law permits otherwise.