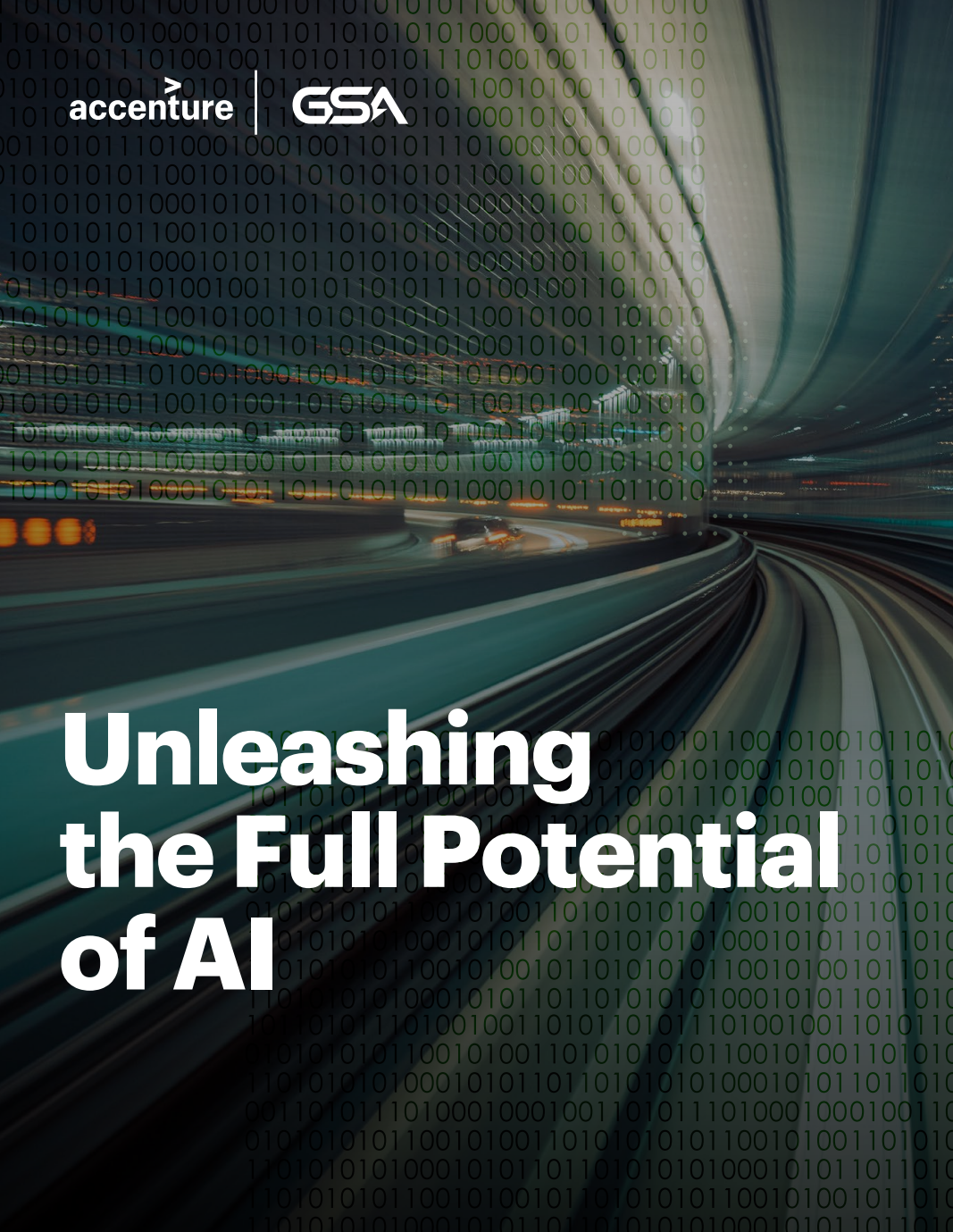


Unleashing the Full Potential of AI

Abstract

Artificial Intelligence (AI) is fundamentally transforming the way the world lives, works and plays. From smart manufacturing and predictive healthcare to enhanced gaming experiences and autonomous driving, AI is profoundly affecting nearly every industry.

Up until now, most AI development has been fueled by the availability of data and computing bandwidth to train and develop AI models. While this has been sufficient for most basic AI functions, this technology has far greater potential to evolve every aspect of our daily lives. The semiconductor industry is key to unlocking this full potential, where semiconductor manufacturers have the opportunity to play a leading role in driving this next wave of computing in an economical way.

The earlier days of AI saw flexible and fungible software running on general-purpose hardware across large-scale use cases. Today's business requirements can no longer be addressed by simple brute force of deploying hardware made of the latest advanced nodes every few years. AI implementation has become more use-case specific, and leading AI practitioners have started marrying their software and hardware development to achieve higher compute performance. In addition, different types of AI hardware have emerged to handle either AI training at data centers or inference at the edge.

This proliferation of AI aware architectures emphasizes the focus on hardware from the initial business case development and planning to the final AI deployment. As a result, AI hardware and the semiconductor industry are in a prime position to be a key driver for future AI advancements.

The exponentially increasing complexity of AI algorithms and the way in which that complexity is evolving through both AI training and AI inferencing presents a significant opportunity for semiconductors to contribute in a meaningful way from both a technology and economic perspective. However, despite being the foundational layer for AI, the way the semiconductor industry can best contribute to the growing demands for AI capabilities varies dramatically as a result of AI's growing requirements and complexity.

Collaborative efforts between companies across the value chain will be vital to ensuring these bets pay off.

There are multiple business challenges to overcome, such as rising costs, talent shortage and increasing use-case application diversification. In addition, while the fragmentation in the semiconductor industry, such as the Fabless-Foundry business model, enables continued innovation, it also drives increasing costs for new future technology solutions.

As a result, this industry needs to place huge bets.

This paper presents three key observations on major trends in the semiconductor industry driving the evolution of AI across two critical components of AI application: Training and Inferencing. These shifts will challenge the industry to rethink the way we operate and collaborate, its monetization strategy, energy consumption and the trajectory of AI technology evolution and scalability.

They include:

01.

While software drove the initial burst of advancements in AI, the parallelism of hardware and software development will achieve the next major AI breakthroughs.

02.

Although the growth in AI creates tremendous demand for semiconductors and puts them in a position of power, the industry will continue to experience growing pains.

03.

Demand for compute power far outstrips its supply and the industry's current business model is not well suited to address this imbalance.

Table 1
What is AI Training and AI Inferencing?

	AI Training	AI Inferencing
Characterized By:	Heavy concentration of compute requirements	AI Inferencing Use-case specific requirements; balance between Cloud and Edge opportunities
Challenged By:	Rising cost of R&D, competing silicon volume, scale	Diversity of solutions and industries

Observation #1 While software drove the initial burst of advancements in AI, the parallelism of hardware and software development will achieve the next major AI breakthroughs.

In this section, we will discuss the evolution of AI compute requirement and how the industry has been addressing it.

Since AI's entry into our lives a decade ago, its rapid progress and ability to solve many technical problems have been enabled by software's scalability and flexibility to be tailored to various applications. As AI applications evolve to be even more pervasive, inclusive, complex, and use-case specific, practitioners are demanding more from their hardware and looking to semiconductors for the next phase of innovation. Gartner estimates that more than 50 companies are making chips specifically for AI, and AI-specific chip revenue is expected to reach $76.8 billion by 2025[1]. Realizing the important role of semiconductors in AI development, leaders have been taking matters into their own hands by bringing chip development capability in-house. Gone are the days where AI applications are built by writing software tailored to the end use-case and leveraging standard off-the-shelf chips.

Today's AI leaders know parallelism in the software and hardware development is necessary to turn their impossible visions into reality. In the future, the best computing performance will be achieved by using a combination of chips and software that are designed for each other and for a specific use-case.

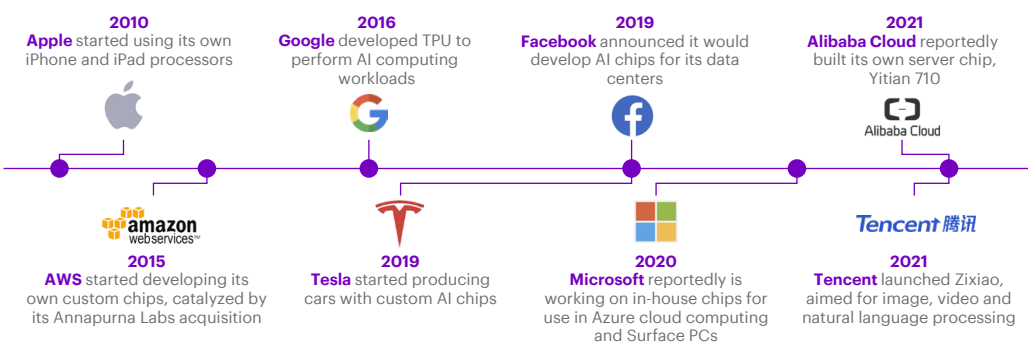


Figure 1
Large tech companies with well-defined AI use cases have been bringing semiconductor expertise to their shores.

In 2020, The Economist reported that the increase in AI compute power requirement would far outpace (more than 6x) the progress that the semiconductor industry currently offers[2] (Figure 2).

With off-the-shelf chips only providing marginal performance boosts, AI hardware companies are exploring new technologies to meet this increasing compute power requirement. While Moore's Law may be enough for non-AI applications such as wireless modems or image processing, it does not provide the compute power for the most advanced AI use-cases where AI models have a larger number of parameters and are becoming more end use-case specific.

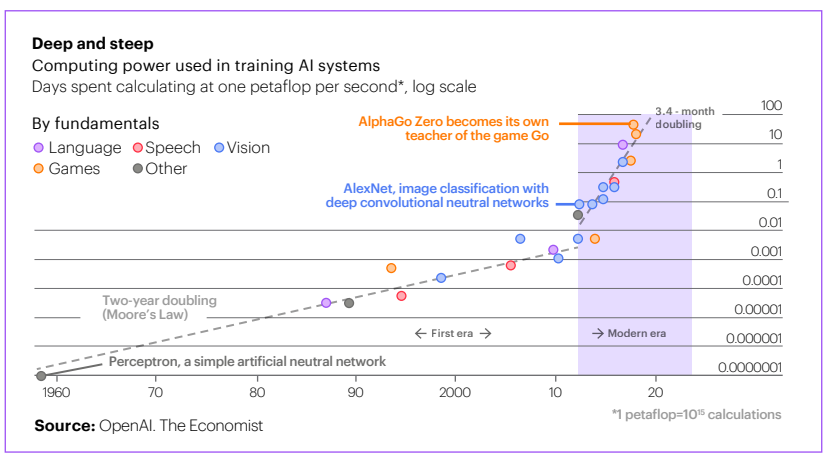


Figure 2
Since 2010, AI computing power requirement has been increasing >6x faster than Moore's Law.

When looking at the increase of AI end use-cases, it is important to analyze the two key phases of AI workflow and their varying compute power requirements. The first phase is AI training, where a set of data is fed into a model, and it "learns" from the dataset and becomes capable of making predictions. The second phase is AI inference, where it uses what AI has learned in training to make a recommendation or solve a particular problem. For AI training components, where compute power requirement is most demanding, end use-case specific accelerators such as TPUs (Tensor Processing Unit) or high-end ASICs designed specifically for the software is the best solution.

However, because AI inference does not require the highest computing firepower, general-purpose CPUs are sufficient. Repurposing the general-purpose processors used in AI inference for different use-cases or software would come with a smaller penalty in performance relative to that for use-case specific accelerators.

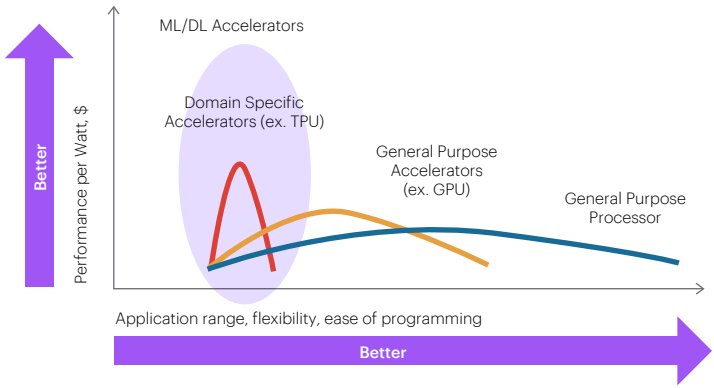


Figure 3
AI accelerators with the best performance are tailored specific to their applications.
Source: Dell

The dichotomy of use-case specific and general-purpose AI hardware is also observed in the worldwide market forecast done by the technology research firm, Allied Market Research: Artificial intelligence chip market; Edge artificial intelligence hardware market.

While CPUs and GPUs will continue to play important roles in the AI space and are projected to grow at a healthy pace through 2030, ASIC implementation of AI is expected to outgrow them both with a CAGR of 39.8% in the forecast period of 2021-2030. In another report, the research firm also points out that the edge AI hardware market is expected to grow rapidly in the forecast period of 2021-2030, driven by their applications in AI inference which has a CAGR of 19.5%.

Observation #2 Although the growth in AI creates tremendous demand for semiconductors and puts them in a position of power, the industry will continue to experience growing pains.

In this section, we will discuss the headwinds that the semiconductor industry is facing as it progresses forward.

On the surface, the semiconductor industry appears to be in an enviable position with high revenue growth driven by increasing demand for AI. Behind the scenes, however, there are painful business operations stemming from the need for continuous strategic investment to develop more complex chips, a larger number of chip types - all while addressing talent scarcity and supply chain issues. According to a survey conducted by GSA (Global Semiconductor Alliance) targeted towards global AI practitioners, semiconductor industry growth will be largely driven by the tremendous worldwide demand coming from autonomous mobility applications, which covers areas broader than autonomous vehicles as shown in **Figure 4**. A wide variety of sensor types, training data and algorithms translate into large data volume, data types and a variety of AI-specific chips.

This trend transcends autonomous mobility and applies to other industries such as manufacturing and finance.

As the survey found, complex and massive data sets born out of AI applications are both a huge demand opportunity and the most pressing business challenge in AI for the semiconductor industry. The survey also noted two important areas that the semiconductor industry could solve to address future AI needs as shown in **Figure 6**: shifting AI requirements and talents.

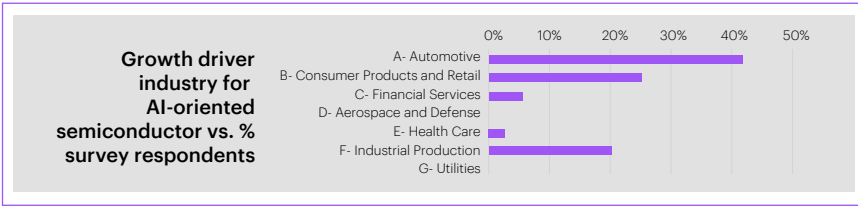


Figure 4

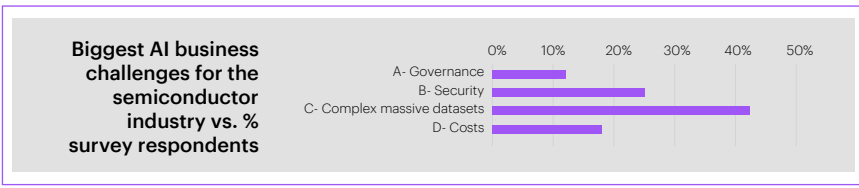


Figure 5

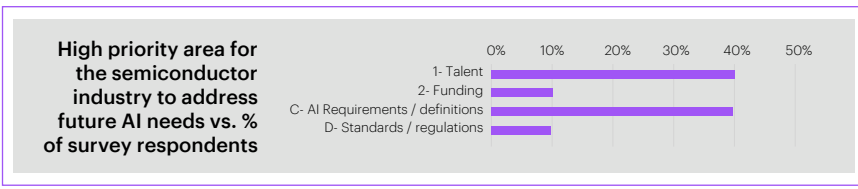
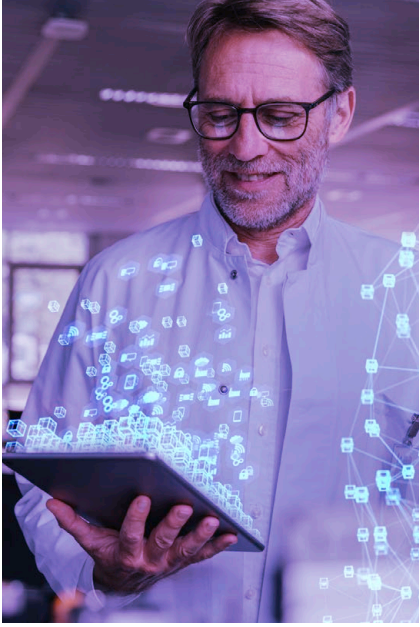


Figure 6

Figure 4, 5, 6 GSA survey results based on input from global audience of semiconductor industry members with expertise and interest in the AI space.

Skilled talent scarcity challenges semiconductor leaders to keep innovating in their recruitment and retention practices as seen in The Competitive Etch: Addressing the talent gap in the semiconductor industry.

The semiconductor industry requires a tremendous level of R&D and capital investments and - equally, if not more importantly - also requires best in class human capital. When the industry was experiencing incredible growth in the 1970s, the best and brightest STEM graduates flocked to leading semiconductor companies such as IBM and Intel. Some started the diaspora of semiconductor giants in Asia such as Samsung Semiconductor, TSMC and SK Hynix. In the following decades, however, STEM talents have been highly favored by other technology industries such as software and social media platforms as well as non-technology industries such as finance and banking.



This global trend continues today, including in regions with historically strong semiconductor industry presence such South Korea, North Europe, and Silicon Valley. Regions such as China, which only recently started having substantial semiconductor activity, are also finding it difficult to find skilled talent. When combined with the shifting boundaries of the semiconductor industry where non-traditional players have begun recruiting semiconductor talent a perfect storm of skilled talent shortage for the industry has emerged **(Figure 7)**[3]. Across the globe, many initiatives have been proposed and executed to improve the talent pipeline. This includes GSA's WLI University Program which brings the semiconductor experience to students and universities around the globe, SEMI Foundation's Global Workforce Development Initiative and SEMI Works as well as regional-level initiatives such as TSMC's Girls in Semiconductor Tour and South Korea's K-Semiconductor Belt Initiative.

However, in addition to recruiting, it is equally important to improve talent retention. This requires an environment where semiconductor talent can retain and grow their skills, become industry leaders, and eventually pass on their knowledge to the next generation.

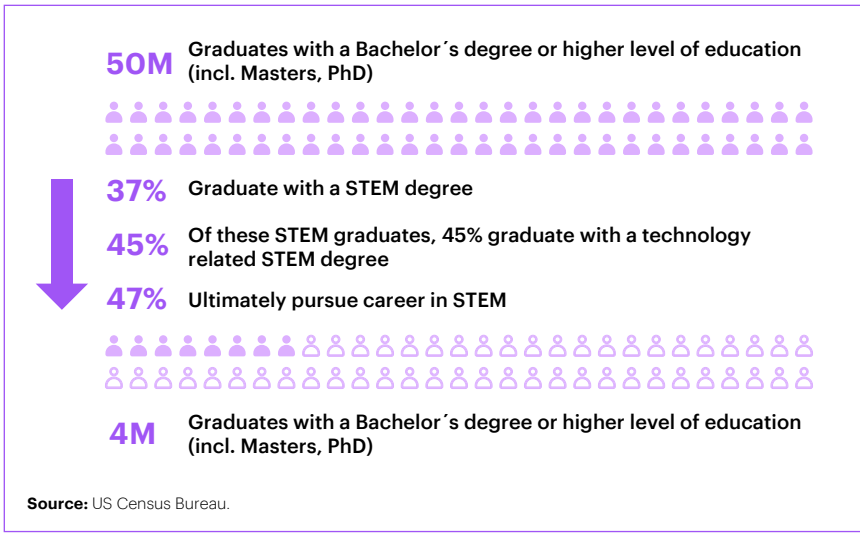


Figure 7
Census data above is from 2019, for US STEM graduates of working age (28-65 years old).
Note: Technology-related STEM degree excludes biological sciences, environmental sciences, agricultural sciences, psychology, social sciences, and multidisciplinary studies.

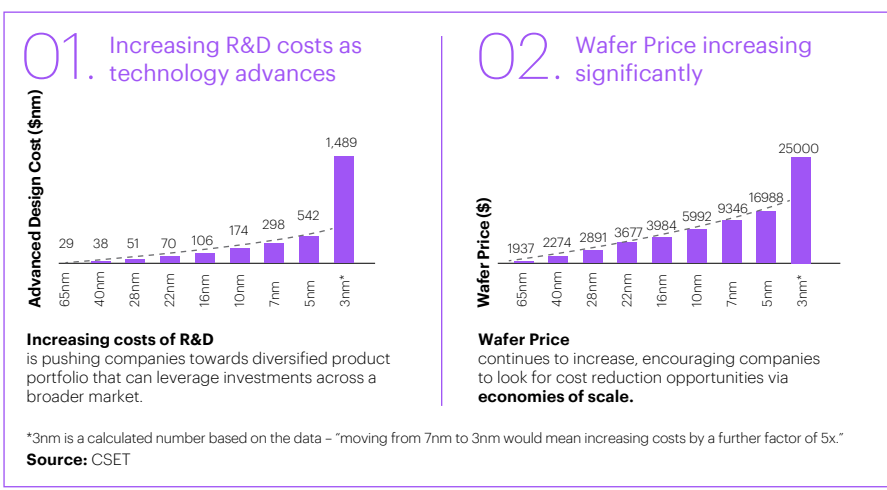
Without a clear retention strategy, the industry is vulnerable to continuous brain drain that has a strong impression on global supply-demand and its ability to innovate and support the rest of the technology industry.

According to the Accenture Tech Vision 2022, the existing talent shortage issue in the technology industry is already significant and will get more severe as the industry, required skills and technology evolve further[4]. Companies that thrive will be those that have a talent pipeline strategy that prioritize early identification of these skills, acquire, and continuously develop them. Without a robust people strategy, the semiconductor industry risks falling further behind in the global competition for top talent. In addition to talent scarcity, another huge hurdle is cost. To be able to play at the frontier of AI, practitioners have to use chips in the trailblazing node (e.g., 5nm or below).

At every node, the physical design and architecture is more complex as are the costs of the associated software, verification, and validation. Weighing the escalating costs against the increasing compute power requirement and larger AI model size raises the question of how far AI advancement can go to meet many needs.

Powerful chips are needed to run compute at scale for GPUs and CPUs, and purposefully designed chips are required for unique use-cases and applications.

According to TSMC, 5nm provides about 20% faster speed and about 40% power reduction than 7nm, while moving from 5nm to 3nm provides a further 15% speed gain and 30% power reduction. However, these performance improvements are dwarfed in comparison to the exponential increase in the design and manufacturing costs **(Figure 8)**.



*3nm is a calculated number based on the data – "moving from 7nm to 3nm would mean increasing costs by a further factor of 5x."
Source: CSET

Figure 8
IBS and CSET research shows an ever-increasing design and manufacturing costs as one moves to more advanced semiconductor nodes. Note that these costs were reported prior to the semiconductor value chain disruptions and increased demand for semiconductor technologies in the last 2 years.

Another challenge the industry has to overcome is how to package AI chips together so that bandwidth is not bottlenecking the overall computing performance and energy consumption is minimized. Today, a large percentage of the energy in AI training is lost in data movement between memory and memory interface. Training complex AI models consumes tremendous amounts of energy and creates major concerns from a carbon emissions perspective.

For example, two options to address demanding AI applications would be either having the compute and data storage within the same chip or using off-chip memory that is specifically designed to handle high bandwidth high-capacity models, as seen in **Figure 9**. Both solutions must comprehend end user and business requirements.

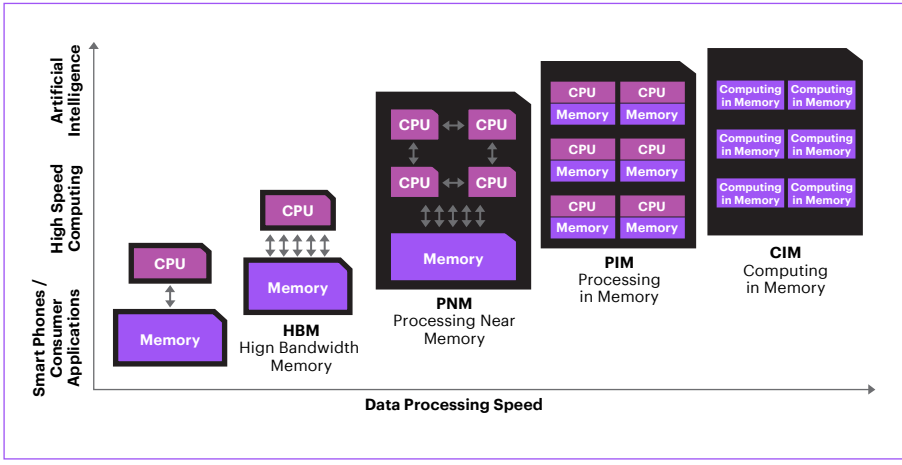


Figure 9
Memory requirements of AI Chips.

The semiconductor industry has been advancing Moore's Law for decades, despite prediction of its death many times in the past. As we look to the future of AI, there are various options for meeting the exponentially increasing compute power that will be required. This can be mapped along two axes as shown in Figure 10:

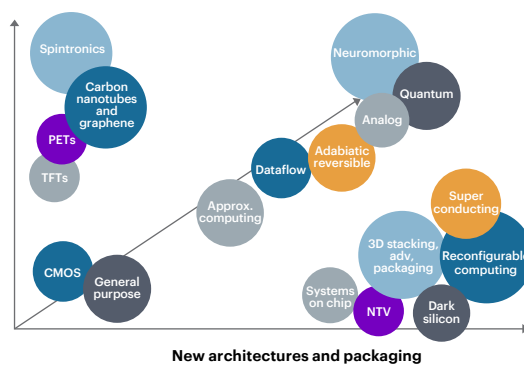
01. New devices and materials such as carbon nanotubes, graphene and spintronics.

02. New architectures and packaging such as 3D packaging, optical interconnects, and reconfigurable computing.

The industry is now at a point where many constraints exist, such as increasing cost and talent shortage. Challenges such as the long-term nature of manufacturing and fungibility constrain the multi-faceted funding options and companies need to decide how to best place bets on forward options. Given that semiconductor technology can take 10-20 years from discovery to volume manufacturing while requiring continued financial investment, this is a difficult question to answer and one that requires collaboration through collective thinking and strategizing by its industry members.

The future of AI compute is uncertain and branching in many different directions

Technology scaling options along three dimensions. All dimensions, which are mutually exclusive, aim to squeeze out more computing performance.



Source: "Computing beyond Moore´s Law, John M. Shakf; Robert Leland

Figure 10
Beating Moore's Law by making chips faster, by making data transfer between chips faster, or by doing both.

Observation #3 Demand for compute power far outstrips its supply and the industry's current business model is not well suited to address this imbalance.

The application of AI is growing at an incredible speed, and the advances in applying AI to solve real-world challenges does not rely on advances at a fundamental conceptual level. Instead, the proliferation of applied AI primarily depends on AI's scalability in terms of technology and economics. However, scalable AI can mean different things depending on whether we view the topic from the training or inference perspective.

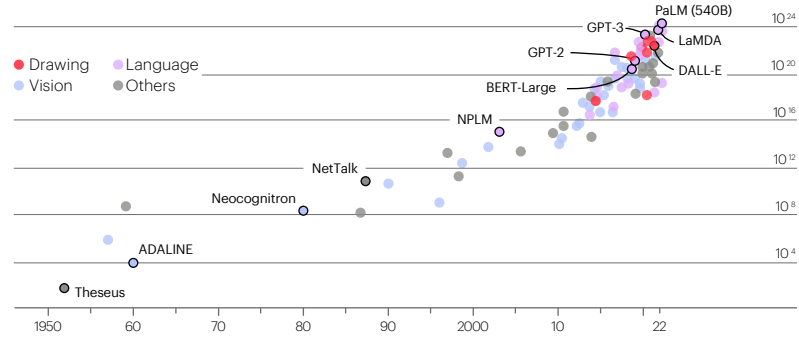
AI scalability at the training level heavily relies on larger data models and neural networks to achieve the objective. In contrast, scalability for inferencing can mean running tasks in compute and power-constrained environments. In this section, we will focus on the relationship between the growing model sizes and their relation to computing demand.

In this section, we will focus on the relationship between the growing model sizes and their relation to computing demand.

As shown in **Figure 2**, the rise in compute demand has a marked inflection point since the emergence of the AlexNet convolutional neural network (CNN) architecture, which consists of multiple layers of neurons, mimicking the way human brain functions[5]. The exponentially increasing diversity of AI use-cases and their complexities require intimate collaboration among semiconductor value chain players to meaningfully propel the AI field forward.

The economics of AI training:
The demand and supply gap in AI compute

As AI model sizes grow, the cost to train a model grows significantly, increasing the demand for AI compute resources. This creates an economic challenge to AI scalability. In recent years, AI practitioners have adopted the approach that foundational AI models work better as they get bigger. As the chart in **Figure 11** highlights, GPT-2 used 40 GB of data and 1.5 billion parameters for its training, while GPT-3 trained on a significantly more extensive data set of 570 GB and 175 billion parameters. GPT-3 has outperformed GPT-2 by a large margin.



Sources: "Compute trends across three eras of machine learning", by J. Sevilla et al., arXiv, 2022; Our World in Data.

Figure 11
The larger the AI model size, the better the AI performance and the prediction it makes. As the cost to train large models across AI applications is growing in an unsustainable fashion, the economics will become more challenging, forcing the AI practitioners to rethink the compute supply-demand equation.

One of the challenges with quantifying the cost of computing needed for training the massive foundational models is the lack of precise data. This requires assumptions and the use of petaflops-days as the measure for the amount of compute power required. Using this measurement, you can conclude the following assumptions:

01. The GPT-3 model has 175 billion parameters and is estimated to require 3640 petaflops-days to train.

02. The Megatron Turing-NLG model is reported to have 530 billion parameters.

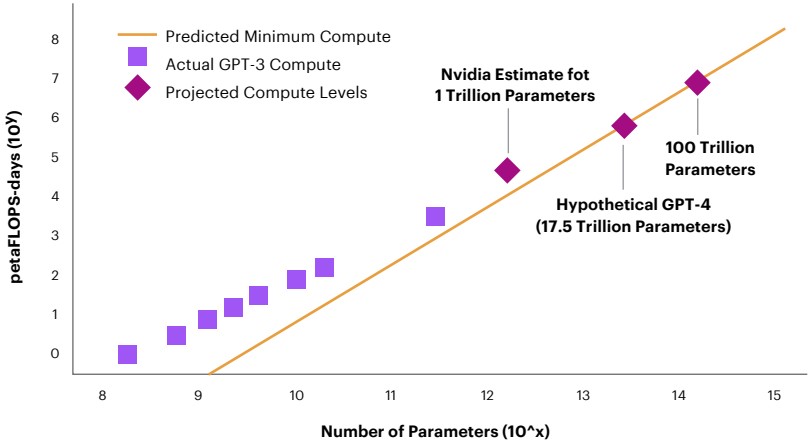
03. The GPT-4 model is projected to be 100 times larger than GPT-3, making it a 17.5 trillion parameter model. This model is estimated to take about 450K petaflops-days to train and cost about $200M.

04. Nvidia estimates that a 1 trillion parameter model might take approximately 42K petaflops-days and cost $19M to train.

05. A 100 trillion parameter model is estimated to cost about $2B to train.

Clearly, the growth of demand for compute is far outstripping the supply. This will likely hinder scalable AI deployment at reasonable cost. This begs the question of where progress will come from in the future. Some approaches being used to answer this include increased algorithmic efficiency, better AI hardware architectures and hardware/software co-design.

There is also a need for industry stakeholders to collaborate up and down the semiconductor value chain to come together to share information and explore new business models to address the scalability challenges to the AI deployment.



Source: OpenAI, Nvidia, and CSET.

Figure 12
Using the assumptions laid out above, a hypothetical GPT-4 AI model is predicted to require 450K petaflops-days and $200M to train.

A Call for Collaboration

AI is at the forefront in changing the world, the way we live and how people perceive what is possible with the technology.

As the implementation of AI becomes more targeted towards the end use case and the bifurcation of AI training and inference continues, the interlock between hardware and software development needs to be more intimate.

Future leaders in the AI implementation will be those that break down barriers between the hardware and software development silos, and instead, place them in the center of the roadmap right from the very beginning.

While this shift places more emphasis on the semiconductor aspect of AI advancement and presents the industry with fantastic upside economic opportunities, it also comes with growing challenges that pull the industry in many different directions; the divergence of AI end use cases, rising costs, talent shortage. For the semiconductor industry to drive the evolution of AI, collaboration will be vital to ensure that the industry as a whole is proceeding in the right direction. We call on all industry professionals to come together and join the GSA AI Interest Group, and engage in collaboratives effort with Accenture, Arm, and GSA.

Together, we can identify, discuss, and address the challenges and benefit from the opportunities this AI paradigm is presenting.

As an industry, we have seen only the tip of the AI iceberg. The next wave of AI will be operating at much higher efficacy and automatically generating original content, not just recognizing patterns. It will democratize prediction with the potential to disrupt and enable new business models, while fueling an explosion of new AI-enabled applications.

Meet the authors

Syed F. Alam
Managing Director
High Tech Global Lead
Accenture Strategy

Timothy Chu
Senior Manager
Businnes Strategy
Accenture

Michael Kurniawan
Manager
Business Strategy
Accenture

Jaya Shukla
Senior Manager
Management Consulting
Accenture

Yanamadala Chowdary
Technology Strategy
Business leader
Arm

Saito Shungo
Director
Program Development
GSA
(Global Semiconductor Association)

References

[1] "AI Chip Startups Pull In Funding as They Navigate Supply Constraints", The Wall Street Journal Online.
https://www.wsj.com/articles/ai-chip-startups-pull-in-funding-as-they-navigate-supply-constraints-11647338402

[2] "The cost of training machines is becoming a problem", The Economist.
https://www.economist.com/technology-quarterly/2020/06/11/the-cost-of-training-machines-is-becoming-a-problem

[3] "United States Census Bureau, Does Majoring In STEM Lead to STEM Jon after graduation: Pathways in STEM".
https://www.census.gov/library/stories/2021/06/does-majoring-in-stem-lead-to-stem-job-after-graduation.html

[4] "Semiconductor Technology Vision 2022", Accenture.
https://www.accenture.com/us-en/insights/technology/technology-trends-2022?c=acn_glb_technologyvisiogoogle_12867730&n=psgs_0322&gclid=CjwKCAjwrZOXBhACEiwA0EoRD-jtATmN46ti_GWg7CTnMvq8olvh230yzRVTzcRO9CK5c0FQcypvUBoCUTkQAvD_BwE&gclsrc=aw.ds

[5] "Convolutional Neural Network", Wikipedia.
https://en.wikipedia.org/wiki/Convolutional_neural_network

About Accenture

Accenture is a global professional services company with leading capabilities in digital, cloud and security. Combining unmatched experience and specialized skills across more than 40 industries, we offer Strategy and Consulting, Technology and Operations services and Accenture Song --all powered by the world's largest network of Advanced Technology and Intelligent Operations centers. Our 710,000 people deliver on the promise of technology and human ingenuity every day, serving clients in more than 120 countries. We embrace the power of change to create value and shared success for our clients, people, shareholders, partners and communities. Visit us at www.accenture.com.

Accenture Semiconductor Practice

Accenture Semiconductor is committed to working with semiconductor manufacturers and companies to help capitalize on the opportunities created by digital disruption and optimize efficiencies across product development, manufacturing, supply chain and business operations. We have deep relationships, experience, and expertise across the semiconductor ecosystem: IDM, IP designers, fabless, foundries and equipment manufacturers. We also have dedicated practice areas and proven results in growth strategy, mergers and acquisitions, engineering operations, silicon design services, supply chain operations, system implementation and manufacturing analytics.
Visit us at www.accenture.com/semiconductors