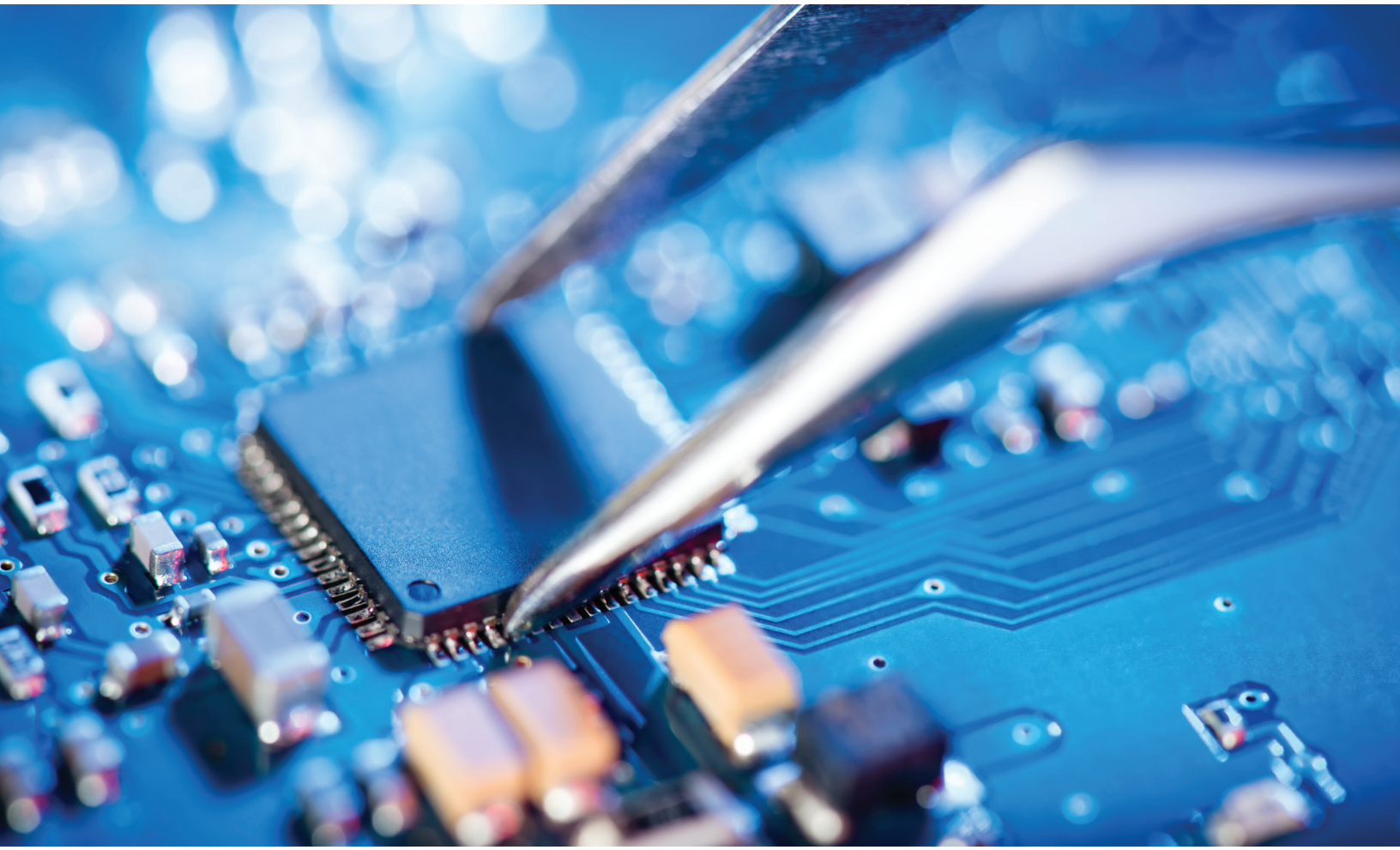
Artificial-intelligence hardware: New opportunities for semiconductor companies

Artificial intelligence is opening the best opportunities for semiconductor companies in decades. How can they capture this value?

Gaurav Batra, Zach Jacobson, Siddarth Madhav, Andrea Queirolo, and Nick Santhanam



© DuKai photographer/Getty Images

Software has been the star of high tech over the past few decades, and it's easy to understand why. With PCs and mobile phones, the game-changing innovations that defined this era, the architecture and software layers of the technology stack enabled several important advances. In this environment, semiconductor companies were in a difficult position. Although their innovations in chip design and fabrication enabled next-generation devices, they received only a small share of the value coming from the technology stack—about 20 to 30 percent with PCs and 10 to 20 percent with mobile.

But the story for semiconductor companies could be different with the growth of artificial intelligence (AI)—typically defined as the ability of a machine to perform cognitive functions associated with human minds, such as perceiving, reasoning, and learning. Many AI applications have already gained a wide following, including virtual assistants that manage our homes and facial-recognition programs that track criminals. These diverse solutions, as well as other emerging AI applications, share one common feature: a reliance on hardware as a core enabler of innovation, especially for logic and memory functions.

What will this development mean for semiconductor sales and revenues? And which chips will be most important to future innovations? To answer these questions, we reviewed current AI solutions and the technology that enables them. We also examined opportunities for semiconductor companies across the entire technology stack. Our analysis revealed three important findings about value creation:

- AI could allow semiconductor companies to capture 40 to 50 percent of total value from the technology stack, representing the best opportunity they've had in decades.
- Storage will experience the highest growth, but semiconductor companies will capture most value in compute, memory, and networking.
- To avoid mistakes that limited value capture in the past, semiconductor companies must undertake a new value-creation strategy that focuses on enabling customized, end-to-end solutions for specific industries, or "microverticals."

By keeping these beliefs in mind, semiconductor leaders can create a new road map for winning in AI. This article begins by reviewing the opportunities that they will find across the technology stack, focusing on the impact of AI on hardware demand at data centers and the edge (computing that occurs with devices, such as self-driving cars). It then examines specific opportunities within compute, memory, storage, and networking. The article also discusses new strategies that can help semiconductor companies gain an advantage in the AI market, as well as issues they should consider as they plan their next steps.

The AI technology stack will open many opportunities for semiconductor companies

AI has made significant advances since its emergence in the 1950s, but some of the most important developments have occurred recently as developers created sophisticated machine-learning (ML) algorithms that can process large data sets, "learn" from experience, and improve over time. The greatest leaps came in the 2010s because of advances in deep learning (DL), a type of ML that can process a wider range of data, requires less data preprocessing by human operators, and often produces more accurate results.

To understand why AI is opening opportunities for semiconductor companies, consider the technology stack (Exhibit 1). It consists of nine discrete layers that enable the two activities that enable AI applications: training and inference (see sidebar "Training and inference"). When developers are trying to improve training and inference, they often encounter roadblocks related to the hardware

Exhibit 1

The technology stack for artificial intelligence (AI) contains nine layers.

Technology	Stack	Definition
Services	**Solution and use case**	Integrated solutions that include training data, models, hardware, and other components (eg, voice-recognition systems)
Training	**Data types**	Data presented to AI systems for analysis
Platform	**Methods**	Techniques for optimizing weights given to model inputs
	Architecture	Structured approach to extract features from data (eg, convolutional or recurrent neural networks)
	Algorithm	A set of rules that gradually modifies the weights given to certain model inputs within the neural network during training to optimize inference
	Framework	Software packages to define architectures and invoke algorithms on the hardware through the interface
Interface	**Interface systems**	Systems within framework that determine and facilitate communication pathways between software and underlying hardware
Hardware	**Head node**	Hardware unit that orchestrates and coordinates computations among accelerators
	Accelerator	Silicon chip designed to perform highly parallel operations required by AI; also enables simultaneous computations

Memory
- Electronic data repository for short-term storage during processing
- Memory typically consists of DRAM[1]

Storage
- Electronic repository for long-term storage of large data sets
- Storage typically consists of NAND[2]

Logic
- Processor optimized to calculate neural network operations, ie, convolution and matrix multiplication
- Logic devices are typically CPU, GPU, FPGA, and/or ASIC[3]

Networking
- Switches, routers, and other equipment used to link servers in the cloud and to connect edge devices

[1] Dynamic random access memory.
[2] Not AND.
[3] CPU= central processing unit, GPU= graphics-processing unit, FPGA = field programmable gate array, ASIC= application-specific integrated circuit.

Source: Expert interviews; literature search

layer, which includes storage, memory, logic, and networking. By providing next-generation accelerator architectures, semiconductor companies could increase computational efficiency or facilitate the transfer of large data sets through memory and storage. For instance, specialized memory for AI has 4.5 times more bandwidth than traditional memory, making it much better suited to handling the vast stores of big data that AI applications require. This performance improvement is so great that many customers would be more willing to pay the higher price that specialized memory requires (about $25 per gigabyte, compared with $8 for standard memory).

AI will drive a large portion of semiconductor revenues for data centers and the edge

With hardware serving as a differentiator in AI, semiconductor companies will find greater demand for their existing chips, but they could also profit by developing novel technologies, such as workload-specific AI accelerators (Exhibit 2). We created a model to estimate how these AI opportunities would affect revenues and to determine whether AI-related chips would constitute a significant portion of future demand (see sidebar "How we estimated value" for more information on our methodology).

Exhibit 2

Companies will find many opportunities in the artificial intelligence (AI) market, with leaders already emerging.

	Opportunities in existing market	Potential new opportunities
Compute	▪ Accelerators for parallel processing, such as GPUs[1] and FPGAs[2]	▪ Workload-specific AI accelerators
Memory	▪ High-bandwidth memory ▪ On-chip memory (SRAM[3])	▪ Emerging non-volatile memory (NVM) (as memory device)
Storage	▪ Potential growth in demand for existing storage systems as more data is retained	▪ AI-optimized storage systems ▪ Emerging NVM (as storage device)
Networking	▪ Infrastructure for data centers	▪ Programmable switches ▪ High-speed interconnect

[1] Graphics-processing units. [2] Field programmable gate arrays. [3] Static random access memory.

Source: McKinsey analysis

Our research revealed that AI-related semiconductors will see growth of about 18 percent annually over the next few years—five times greater than the rate for semiconductors used in non-AI applications (Exhibit 3). By 2025, AI-related semiconductors could account for almost 20 percent of all demand, which would translate into about $67 billion in revenue. Opportunities will emerge at both data centers and the edge. If this growth materializes as expected, semiconductor companies will be positioned to capture more value from the AI technology stack than they have obtained with previous innovations—about 40 to 50 percent of the total.

AI will drive most growth in storage, but the best opportunities for value-creation lie in other segments

We then took our analysis a bit further by looking at specific opportunities for semiconductor players within compute, memory, storage, and networking. For each area, we examined how hardware demand is evolving at both data centers and the edge. We also quantified the growth expected in each category except networking, where AI-related opportunities for value capture will be relatively small for semiconductor companies.

Compute

Compute performance relies on central processing units (CPUs) and accelerators—graphics-processing units (GPUs), field programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs). Since each use case has different compute requirements, the optimal AI hardware architecture will vary. For instance, route-planning applications have different needs for processing speed, hardware interfaces, and other performance features than applications for autonomous driving or financial risk stratification (Exhibit 4).

Overall, demand for compute hardware will increase by about 10 to 15 percent through 2025 (Exhibit 5). After analyzing more than 150 DL use cases, looking at both inference and training requirements, we were able to identify the architectures most likely to gain ground in data centers and the edge (Exhibit 6).

Data-center usage. Most compute growth will stem from higher demand for AI applications at cloud-computing data centers. At these locations, GPUs are now used for almost all training applications. We expect that they will soon begin to lose market share to ASICs, until the compute market is about evenly divided between these solutions by 2025. As ASICs enter the market, GPUs will likely become more customized to meet the demands of DL. In addition to ASICs and GPUs, FPGAs will have a small role in future AI training, mostly for specialized data-center applications that must reach the market quickly or require customization, such as those for prototyping new DL applications.

For inference, CPUs now account for about 75 percent of the market. They'll lose ground to ASICs as DL applications gain traction. Again, we expect to see an almost equal divide in the compute market, with CPUs accounting for 50 percent of demand in 2025 and ASICs for 40 percent.

Edge applications. Most edge training now occurs on laptops and other personal computers, but more devices may begin recording data and playing a role in on-site training. For instance, drills used during oil and gas exploration generate data related to a well's geological characteristics that could be used to train models. For accelerators, the training market is now evenly divided between CPUs and ASICs. In the future, however, we expect that ASICs built into systems on chips will account for 70 percent of demand. FPGAs will represent about 20 percent of demand and will be used for applications that require significant customization.

When it comes to inference, most edge devices now rely on CPUs or ASICs, with a few applications—such as autonomous cars—requiring GPUs. By 2025, we expect that ASICs will account for about 70 percent of the edge inference market and GPUs 20 percent.

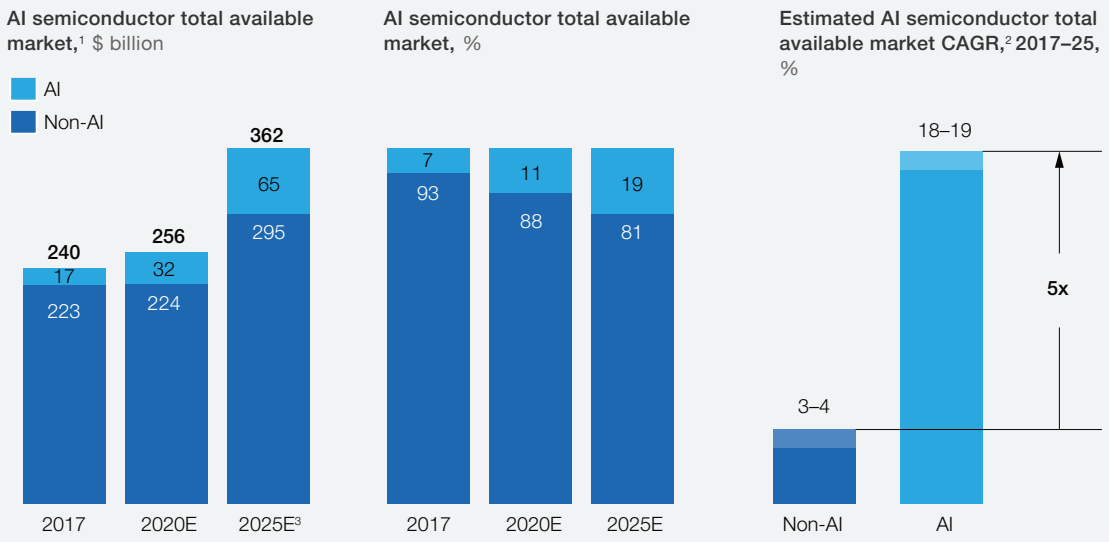
Training and inference

All AI applications must be capable of training and inference. To understand the importance of these tasks, consider their role in helping self-driving cars avoid obstacles. During the training phase, developers present images to the neural net—for instance, those of dogs or pedestrians—and perform recognition tests. They then refine network parameters until the neural net displays high accuracy in visual detection. After the network has viewed millions of images and is fully trained, it enables recognition of dogs and pedestrians during the inference phase.

The cloud is an ideal location for training because it provides access to vast stores of data from multiple servers—and the more information an AI application reviews during training, the better its algorithm will become. Further, the cloud can reduce expenses because it allows graphics-processing units (GPUs) and other expensive hardware to train multiple AI models. Since training occurs intermittently on each model, capacity is not an issue.

With inference, AI algorithms handle less data but must generate responses more rapidly. A self-driving car doesn't have time to send images to the cloud for processing once it detects an object in the road, nor do medical applications that evaluate critically ill patients have leeway when interpreting brain scans after a hemorrhage. And that makes the edge, or in-device computing, the best choice for inference.

Exhibit 3 **Growth for semiconductors related to artificial intelligence (AI) is expected to be five times greater than growth in the remainder of the market.**



[1] Total available market includes processors, memory, and storage; excludes discretes, optical, and micro-electrical-mechanical systems.
[2] Compound annual growth rate.
[3] E = estimated.

Source: Bernstein; Cisco Systems; Gartner; IC Insights; IHS Markit; Machina Research; McKinsey analysis

Memory

AI applications have high memory-bandwidth requirements, since computing layers within deep neural networks must pass input data to thousands of cores as quickly as possible. Memory is required—typically dynamic random access memory (DRAM)—to store input data, weight model parameters, and perform other functions during both inference and training. Consider a model being trained to recognize the image of a cat. All intermediate results in the recognition process—for example, colors, contours, textures—need to reside on memory as the model fine-tunes its algorithms. Given these requirements, AI will create a strong opportunity for the memory market, with value expected to increase from $6.4 billion in 2017 to $12.0 billion in 2025.

That said, memory will see the lowest annual growth of the three accelerator categories—about 5 to 10 percent—because of efficiencies in algorithm design, such as reduced bit precision, as well as capacity constraints in the industry relaxing.

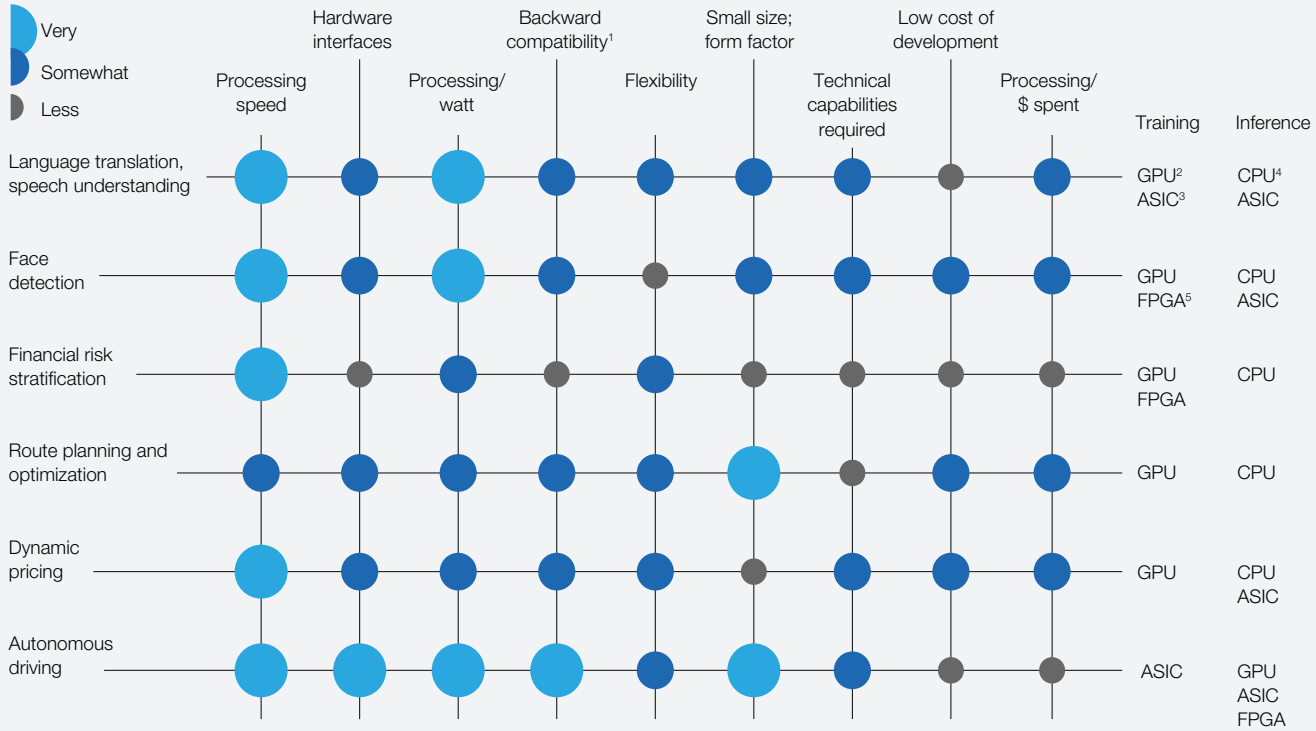
Most short-term memory growth will result from increased demand at data centers for the high-bandwidth DRAM required to run AI, ML, and DL algorithms. But over time, the demand for AI memory at the edge will increase—for instance, connected cars may need more DRAM.

Current memory is typically optimized for CPUs, but developers are now exploring new architectures. Solutions that are attracting more interest include the following:

Exhibit 4

The optimal compute architecture will vary by use case.

Example use-case analysis of importance



[1] Can use interfaces and data from earlier versions of the system.
[2] Graphics-processing unit.
[3] Application-specific integrated circuit.
[4] Central processing unit.
[5] Field-programmable gate array.

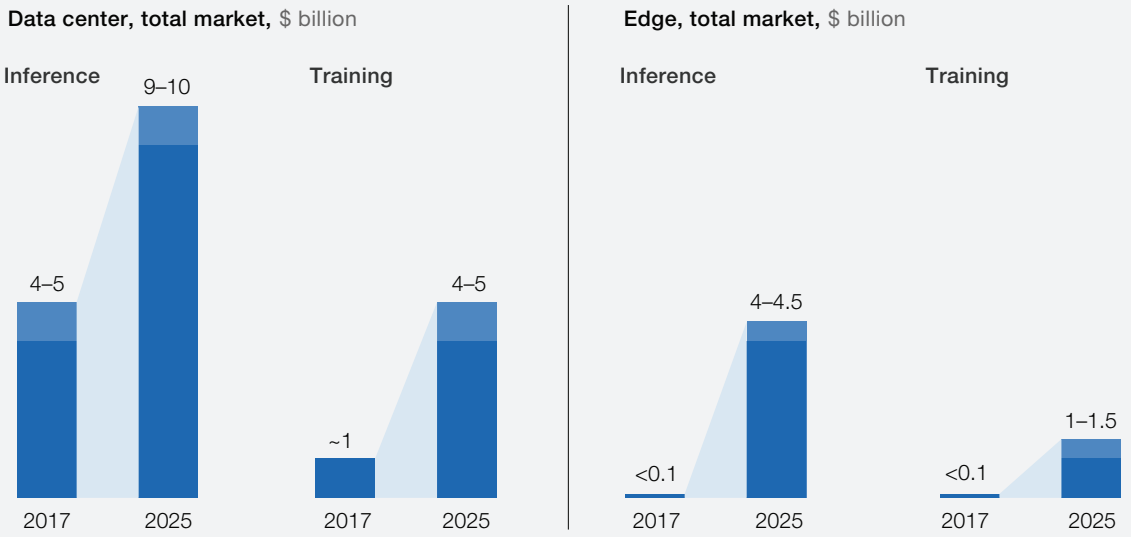
Source: McKinsey analysis

- *High-bandwidth memory (HBM).* This technology allows AI applications to process large data sets at maximum speed while minimizing power requirements. It allows DL compute processors to access a three-dimensional stack of memory through a fast connection called through-silicon via (TSV). AI chip leaders such as Google and Nvidia have adopted HBM as the preferred memory solution, although it costs three times more than traditional DRAM per gigabyte—a move that signals their customers are willing to pay for expensive AI hardware in return for performance gains.[1]

- *On-chip memory.* For a DL compute processor, storing and accessing data in DRAM or other outside memory sources can take 100 times more time than memory on the same chip. When Google designed the tensor-processing unit (TPU), an ASIC specialized for AI, it included enough memory to store an entire model on the chip.[2] Start-ups such as Graphcore are also increasing on-chip memory capacity, taking it to

Exhibit 5

At both data centers and the edge, demand for training and inference hardware is growing.



Source: Expert interviews; McKinsey analysis

a level about 1,000 times more than what is found on a typical GPU, through a novel architecture that maximizes the speed of AI calculations. The cost of on-chip memory is still prohibitive for most applications, and chip designers must address this challenge.

Storage

AI applications generate vast volumes of data—about 80 exabytes per year, which is expected to increase to 845 exabites by 2025. In addition, developers are now using more data in AI and DL training, which also increases storage requirements. These shifts could lead to annual growth of 25 to 30 percent from 2017 to 2025 for storage—the highest rate of all segments we examined.[3] Manufacturers will increase their output of storage accelerators in response, with pricing dependent on supply staying in sync with demand.

Unlike traditional storage solutions that tend to take a one-size-fits-all approach across different use cases, AI solutions must adapt to changing needs—and those depend on whether an application is used for training or inference. For instance, AI training systems must store massive volumes of data as they refine their algorithms, but AI inference systems only store input data that might be useful in future training. Overall, demand for storage will be higher for AI training than inference.

One potential disruption in storage is new forms of non-volatile memory (NVM). New forms of NVM have characteristics that fall between traditional memory, such as DRAM, and traditional storage, such as NAND flash. They can promise higher density than DRAM, better performance than NAND, and better power consumption than both. These characteristics will enable new applications and allow NVM to substitute for DRAM and NAND in others. The market for these forms of NVM are currently small—representing about $1 billion to $2 billion in revenue over the next two years—but it is projected to account for more than $10 billion in revenue by 2025.

The NMV category includes multiple technologies, all of which differ in terms of memory access time and cost, and are all in various stages. Magnetoresistive random-access memory (MRAM) has the lowest latency for read and write, with greater than five-year data retention and excellent endurance. However, its capacity scaling is limited, making it a costly alternative that may be used for frequently accessed caches rather than a long-term data-retention solution. Resistive random-access memory (ReRAM) could potentially scale vertically, giving it an advantage in scaling and cost, but it has slower latency and reduced endurance. Phase-change memory (PCM) fits in between the two, with 3D XPoint being the most well-known example. Endurance and error rate will be key barriers that must be overcome before more widespread adoption.

Networking

AI applications require many servers during training, and the number increases with time. For instance, developers only need one server to build an initial AI model and under 100 to improve its structure. But training with real data—the logical next step—could require several hundred. Autonomous-driving models require over 140 servers to reach 97 percent accuracy in detecting obstacles.

If the speed of the network connecting servers is slow—as is usually the case—it will cause training bottlenecks. Although most strategies for improving network speed now involve data-center hardware, developers are investigating other options, including programmable switches that can route data in different directions. This capability will accelerate one of the most important training tasks: the need to resynchronize input weights among multiple servers whenever model parameters are updated. With programmable switches, resynchronization can occur almost instantly, which could increase training speed from two to ten times. The greatest performance gains would come with large AI models, which use the most servers.

Another option to improve networking involves using high-speed interconnections in servers. This technology can produce a threefold improvement in performance, but it's also about 35 percent more expensive.

Semiconductor companies need new strategies for the AI market

It's clear that opportunities abound, but success isn't guaranteed for semiconductor players. To capture the value they deserve, they'll need to focus on end-to-end solutions for specific industries (also called microvertical solutions), ecosystem development, and innovation that goes far beyond improving compute, memory, and networking technologies.

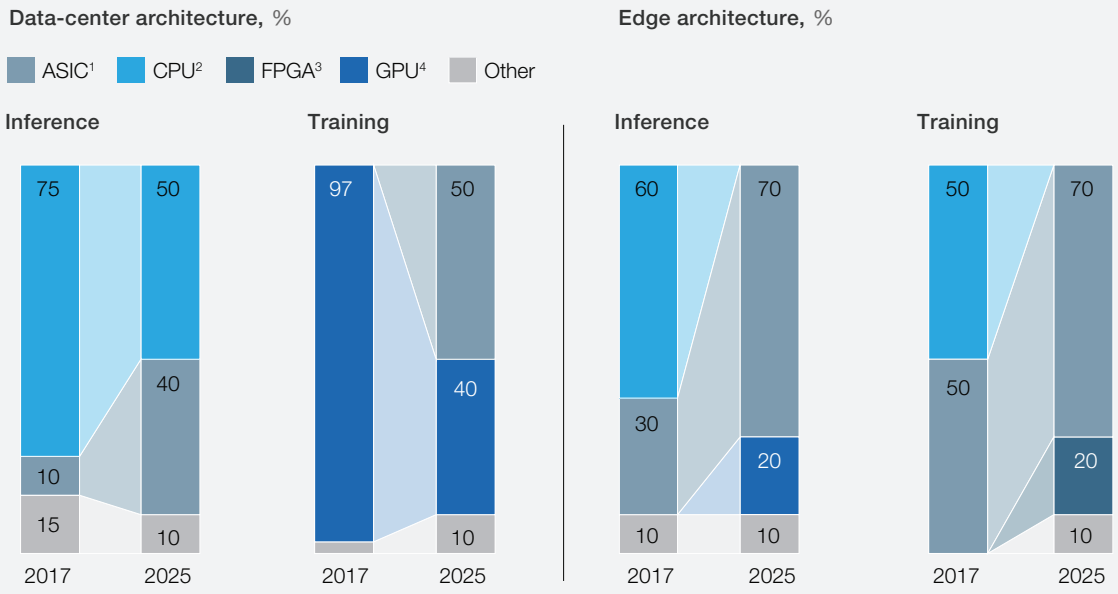
Customers will value end-to-end solutions for microverticals that deliver a strong return on investment

AI hardware solutions are only useful if they're

How we estimated value

We took a bottom-up approach to estimate the value at stake for semiconductor companies. Consider accelerators used for compute functions. First, we determined the percent of servers in data centers that were used for AI. We then identified the type of logic device they commonly used and the average sales price for related accelerators. For edge computing, we conducted a similar review, but we focused on determining the number of devices that were used for AI, rather than servers. By combining our insights for data centers and edge devices, we could estimate the potential value for semiconductor companies related to compute functions.

Exhibit 6 **The preferred architectures for compute are shifting in data centers and the edge.**



1 Application-specific integrated circuit.
2 Central processing unit.
3 Field programmable gate array.
4 Graphics-processing unit.

Source: Expert interviews; McKinsey analysis

compatible with all other layers of the technology stack, including the solutions and use cases in the services layer. Semiconductor companies can take two paths to achieve this goal, and a few have already begun doing so. First, they could work with partners to develop AI hardware for industry-specific use cases, such as oil and gas exploration, to create an end-to-end solution. For example, Mythic has developed an ASIC to support edge inference for image- and voice-recognition applications within the healthcare and military industries. Alternatively, semiconductor companies could focus on developing AI hardware that enables broad, cross-industry solutions, as Nvidia does with GPUs.

The path taken will vary by segment. With memory and storage players, solutions tend to have the same technology requirements across microverticals. In compute, by contrast, AI algorithm requirements may vary significantly. An edge accelerator in an autonomous car must process much different data from a language-translation application that relies on the cloud. Under these circumstances, companies cannot rely on other players to build other layers of the stack that will be compatible with their hardware.

Active participation in ecosystems is vital for success

Semiconductor players will need to create an ecosystem of software developers that prefer their hardware by offering products with wide appeal. In return, they'll have more influence over design choices. For instance, developers who prefer a

certain hardware will use that as a starting point when building their applications. They'll then look for other components that are compatible with it.

To help draw software developers into their ecosystem, semiconductor companies should reduce complexity whenever possible. Since there are now more types of AI hardware than ever, including new accelerators, players should offer simple interfaces and software-platform capabilities. For instance, Nvidia provides developers with Compute Unified Device Architecture, a parallel-computing platform and application programming interface (API) that works with multiple programming languages. It allows software developers to use Compute Unified Device Architecture–enabled GPUs for general-purpose processing. Nvidia also provides software developers with access to a collection of primitives for use in DL applications. The platform has now been deployed across thousands of applications.

Within strategically important industry sectors, Nvidia also offers customized software-development kits. To assist with the development of software for self-driving cars, for instance, Nvidia created DriveWorks, a kit with ready-to-use software tools, including object-detection libraries that can help applications interpret data from cameras and sensors in self-driving cars.

As preference for certain hardware architectures builds throughout the developer community, semiconductor companies will see their visibility soar, resulting in better brand recognition. They'll also see higher adoption rates and greater customer loyalty, resulting in lasting value.

Only platforms that add real value to end users will be able to compete against comprehensive offerings from large high-tech players, such as Google's TensorFlow, an open-source library of ML and DL models and algorithms.[4] TensorFlow supports Google's core products, such as Google Translate, and also helps the company solidify its position within the AI technology stack, since TensorFlow is compatible with multiple compute accelerators.

Innovation is paramount and players must go up the stack

Many hardware players who want to enable AI innovation focus on improving the computation process. Traditionally, this strategy has involved offering optimized compute accelerators or streamlining paths between compute and data through innovations in memory, storage, and networking. But hardware players should go beyond these steps and seek other forms of innovation by going up the stack. For example, AI-based facial-recognition systems for secure authentication on smartphones were enabled by specialized software and a 3-D sensor that projects thousands of invisible dots to capture a geometric map of a user's face. Because these dots are much easier to process than several millions of pixels from cameras, these authentication systems work in a fraction of a second and don't interfere with the user experience. Hardware companies could also think about how sensors or other innovative technologies can enable emerging AI use cases.

Semiconductor companies must define their AI strategy now

Semiconductor companies that are first movers in the AI space will be more likely to attract and retain customers and ecosystem partners—and that could prevent later entrants from attaining a leading position in the market. With both major technology players and start-ups launching independent efforts in the AI hardware space now, the window of opportunity for staking a claim will rapidly shrink over the next few years. To establish a strong strategy now, they should focus on three questions:

- *Where to play?* The first step to creating a focused strategy involves identifying the target industry microverticals and AI use cases. At the most basic level, this involves estimating the size of the opportunity within different verticals,

as well as the particular pain points that AI solutions could eliminate. On the technical side, companies should decide if they want to focus on hardware for data centers or the edge.

- *How to play?* When bringing a new solution to market, semiconductor companies should adopt a partnership mind-set, since they might gain a competitive edge by collaborating with established players within specific industries. They should also determine what organizational structure will work best for their business. In some cases, they might want to create groups that focus on certain functions, such as R&D, for all industries. Alternatively, they could dedicate groups to select microverticals, allowing them to develop specialized expertise.

- *When to play?* Many companies might be tempted to jump into the AI market, since the cost of being a follower is high, particularly with DL applications. Further, barriers to entry will rise as industries adopt specific AI standards and expect all players to adhere to them. While rapid entry might be the best approach for some companies, others might want to take a more measured approach that involves slowly increasing their investment in select microverticals over time.

The AI and DL revolution gives the semiconductor industry the greatest opportunity to generate value that it has had in decades. Hardware can be the differentiator that determines whether leading-edge applications reach the market and grab attention. As AI advances, hardware requirements will shift for compute, memory, storage, and networking—and that will translate into different demand patterns. The best semiconductor companies will understand these trends and pursue innovations that help take AI hardware to a new level. In addition to benefitting their bottom line, they'll also be a driving force behind the AI applications transforming our world.

[1] Liam Tung, "GPU Killer: Google reveals just how powerful its TPU2 chip really is," ZDNet, December 14, 2017, zdnet.com.

[2] Kaz Sato, "What makes TPUs fine-tuned for deep learning?," Google, August 30, 2018, google.com.

[3] When exploring opportunities for semiconductor players in storage, we focused on NAND. Although demand for hard-disk drives will also increase, this growth is not driven by semiconductor advances.

[4] An open-source, machine-learning framework for everyone, available at tensorflow.org.

Gaurav Batra is a partner in McKinsey's Washington, DC, office, **Zach Jacobson** and **Andrea Queirolo** are associate partners in the New York office, **Siddarth Madhav** is a partner in the Chicago office, and **Nick Santhanam** is a senior partner in the Silicon Valley office.

The authors wish to thank Sanchi Gupte, Jo Kakarwada, Teddy Lee, and Ben Byungchol Yoon for their contributions to this article.

Designed by Sydney Design Studio
Copyright © 2018 McKinsey & Company.
All rights reserved.