William Blair

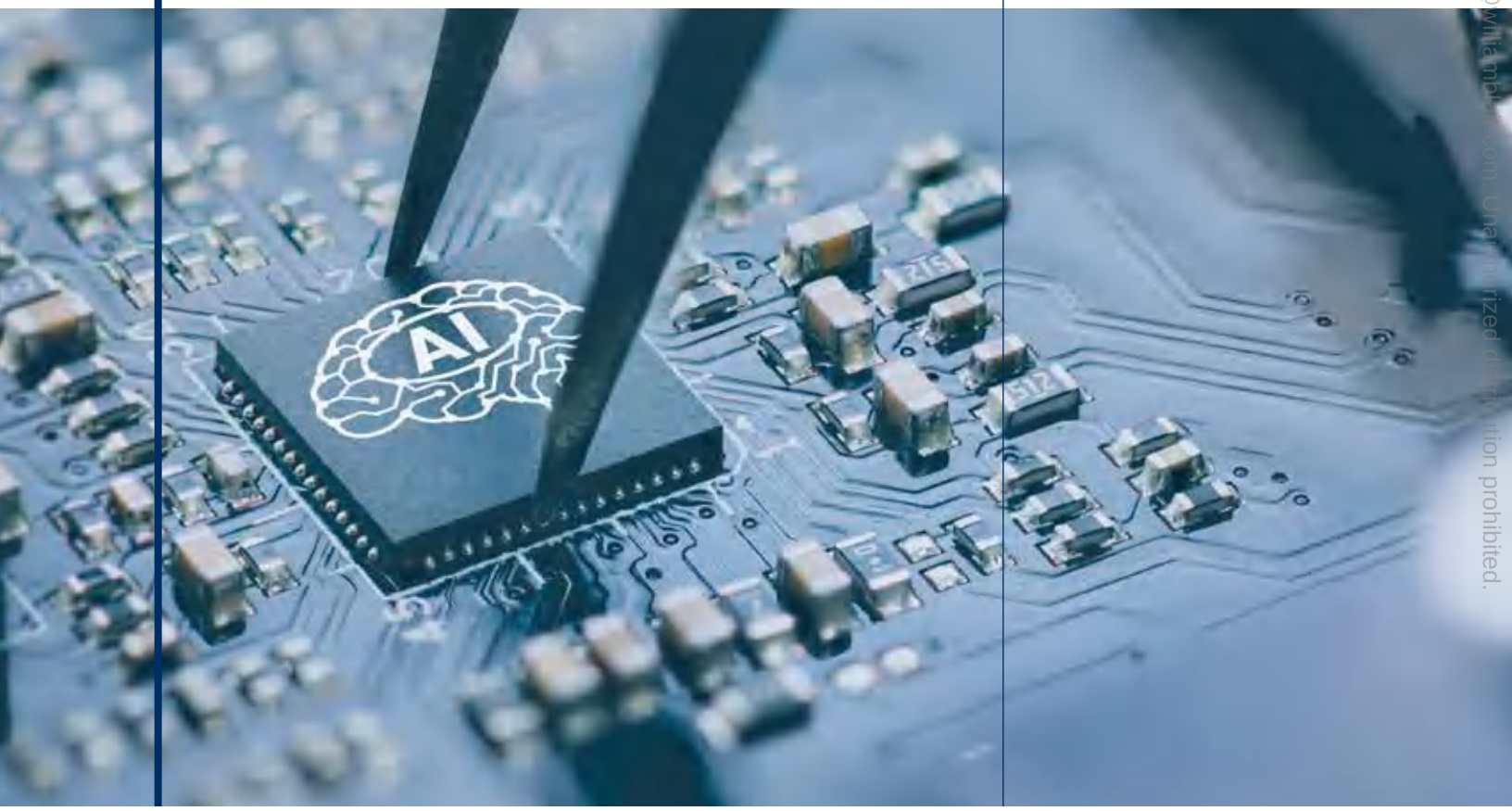
Equity Research
Technology, Media, & Communications | Semiconductors & Infrastructure Systems

September 18, 2024
Industry Report

From Chips to Systems: How AI Is Revolutionizing Compute and Infrastructure

Sebastien Naji +1 212 245 6508
snaji@williamblair.com

Jason Ader, CFA +1 617 235 7519
jader@williamblair.com



Contents

Introduction

It is not controversial to point to AI, and in particular GenAI, as the next generational shift in computing. Similar to prior waves that took us from mainframes to PCs to mobile phones to cloud data centers, each shift has required a rethinking of computing architectures, processor design, and systems engineering. In our view, AI is no different, taking many learnings from the PC and mobile eras and shifting them into the data center. Technical capabilities like parallel computing, system on a chip (SoC), software-defined infrastructure, and systems science are not new, but AI has emerged as a new, large-scale workload that can make use of all these capabilities.

In past reports, the William Blair technology team has discussed the shifts caused by the rise of GenAI since the launch of ChatGPT in November 2022, and how this new technology has reshuffled investments and demand for new infrastructure (network, storage), data services, and security. See our GenAI primer (Generative AI: The New Frontier of Automation) and enterprise GenAI report (The Generative AI Toolchain: How Enterprises Turn Hype Into Reality). In this report, we dig one level deeper to understand fundamentally how the rise in AI impacts the computing layer and with it the entirety of data center infrastructure technology.

Key Takeaways

It's Not Just Better Chips, It's Better Systems

Traditional semiconductor companies have increasingly set their eyes on layers higher up the stack, with increased verticalizations in designs and architectures, shifting the core unit of compute from the chip to the broader computing system. Companies like Nvidia do not view themselves as chip providers but rather as builders of entire computers, where the integration of chips, storage, networking, and software plays a key role in driving performance improvements. In our view, this systems-level approach creates a more substantial and sustainable technical moat for chip vendors like Nvidia than what we're accustomed to seeing in past cycles in the semiconductor industry. For example, the DGX system offered by Nvidia is a complex computer system, comprising 35,000 different components designed to work together to deliver better performance for AI workloads. DGX, like all of Nvidia's solutions, runs the long-established compute unified device architecture (CUDA software) that allows developers to program on top of Nvidia GPUs. The high costs of AI infrastructure—due to the scale and sophistication of systems required—have been a catalyst to vertically integrate more of the semiconductor and IT stack value chain into system-level solutions optimized for a particular use-case. Because of the scale of these systems, incremental improvements in performance or energy efficiency can have a dramatic impact on costs. In the AI era, as the emphasis in computing moves from chips to comprehensive systems, we expect systems companies will accrue the bulk of semiconductor revenue, largely overshadowing discrete chip makers.

AI Represents the Next Generational Shift in Computing

Almost two years following the release of ChatGPT, it is becoming clear that AI represents the key new paradigm for the computing world, requiring a reshuffling of design, architecture, and supply chains to address a burgeoning set of use-cases and workloads. AI follows prior generational shifts in computing, including the shift from mainframe to PCs in the 1980s and from PCs to mobile in the late 2000s. With AI still in its infancy, we expect it will create a massive multitrillion-dollar market opportunity over the next decade as AI technology is integrated across almost all existing processes and solutions (to improve productivity and lower costs) and used to build new tools and applications (e.g., AI-powered applications, omniverse, digital twins, robotics, autonomous vehicles).

Parallel Computing Will Underpin the Majority of Applications

The rise of AI marks a shift from the primacy of serial computing (and the CPU) in the data center to parallel computing (and the GPU/AI accelerator). This shift has been driven by the order of magnitude higher processing power needed to train and operate models that are based on massive troves of data. The vector/tensor mathematics that underpin transformer and deep learning models demand parallelization, where tasks can be broken up into smaller chunks and processed simultaneously. As the benefits of integrating AI models within applications and workflows becomes clearer, the advantages of parallelism have driven an increase in share of GPUs within data center environments—roughly 30% of new chips in data centers were GPUs in 2023, a number we expect to quickly surpass 50% over the next few years, particularly as AI applications represent a larger portion of the overall application estate.

Vertical Integration Keeps Moore's Law Alive

While the traditional performance improvements dictated by Moore's law have slowed—it gets harder and harder to shrink transistors as you hit the limits of physics at the atomic level—semiconductor companies have developed workarounds to continue driving a steady pace of performance improvements for their chips. The semi industry has had to broaden its focus and purview from processor designs to chip systems to superchips, and today, to building vertically integrated computing systems that combine expertise in compute, storage, networking, and software. The rapid rise of GenAI has only made this verticalization shift all the more important, kicking off a new race for accelerating improvements in performance and energy efficiency that expand beyond the chip to the full data center stack. Microsoft CEO Satya Nadella highlighted this point in his March 2024 keynote at the Microsoft Build conference, positing that we are likely entering a "golden age" of systems in IT.

Compute Systems Garner Higher Margins Than Chips Alone

As the core IP of semi companies has expanded from chip design toward systems engineering and software, these companies are able to capture more of the value in the tech stack than ever before. Amdahl's law highlights the limits of simply creating bigger GPUs (i.e., the benefits of higher parallelization are limited by the serial portions of the systems), which has pushed the focus of semi companies up the stack. As more of the IP is driven by valuable software capabilities and system architecture, semi companies have been able to drive better gross margins; Nvidia is the prime example of this with its popular CUDA stack helping it achieve some of the highest gross margins in the industry (in the mid-70% range). While some of this is driven by pricing power, we expect these higher margins will be sustainable for the near term as semi companies develop more proprietary capabilities themselves and are less dependent on improving foundry capabilities as the core driver of performance improvements.

Custom Chip Demand Highlights Verticalization Trend

One key trend over the last several years has been the increasing number of technology companies that are designing their own chips rather than buying off-the-shelf computing from traditional semiconductor companies (like Intel, AMD, Qualcomm). For example, Apple now designs its own chips (A-series for iPhones, M-series for computers) to optimize the performance of its products for its own software ecosystem. Hyperscale technology companies like Meta (with MTIA), AWS (Graviton CPU, Inferentia/Trainium GPUs), Microsoft (Maia GPU, Cobalt CPU), Google (TPUs), and ByteDance/OpenAI more recently are pouring tons of resources into building their own technology stack from the chips all the way up to applications running on top. This highlights the tangible benefits of designing the entire system to maximize performance/cost for a particular use-case. While this highlights the structural shift toward verticalization, it also highlights that for their largest potential customers, chip providers will increasingly be seen as competitors.

Designing Full-Stack Systems Is Hard

While the vertical approach works for the largest tech companies that benefit from massive war chests, it remains out of reach for the majority of organizations that lack the resources or expertise to work directly with the semiconductor supply chain and build their own systems from scratch. Even for the hypescalers like AWS, Meta, and Google, building their own chips and data center systems is a challenging proposition. Our conversations with industry experts highlight the fact that building in-house solutions is challenging. For example, while AWS has had some success with its in-house, Arm-based CPUs (Graviton), its GPU offerings (Trainium, Inferentia) have struggled to gain much traction. This is because AWS not only lags Nvidia's technical lead but also lacks an alternative software ecosystem that can compete effectively with CUDA. The result is that Trainium/Inferentia are useful as low-cost alternatives for certain use-cases, but still need a compatibility layer to run CUDA-based libraries and programs on top. Nvidia's control of the standard software layer for parallel processing gives it a massive lead over competitors and means that hyperscalers today are still playing catch-up with Nvidia rather than establishing their own competitive moat.

Semi Leaders Embrace Verticalization

During the GTC 2024 financial analyst session, Nvidia CEO Jensen Huang made the following statement: "What Nvidia does for a living is not build the chip. We build an entire supercomputer, from the chip to the system to the interconnects, the NVLinks [server interconnects], the networking, but very importantly the software." This assertion not only perfectly encapsulates the idea that Nvidia is incredibly focused on system design and verticalizing the stack to build an entire computing system, but also highlights the importance of software as a valuable layer of this solution. In our view, the semiconductor companies poised to benefit the most from this AI wave are those that have embraced this system approach. Today, we see three vendors at the forefront of this shift: Nvidia, ARM, and Broadcom. Other vendors on the radar include CPU leaders playing catch-up in parallel processing (Intel, AMD), key component vendors of these broader compute systems (like Micron, Monolithic Power, Marvell), and electronic design automation (EDA) firms that enable increased customization of processor/chip system designs (Synopsys, Cadence).

A Brief History of Computing

The semiconductor industry was born in 1947 when a team of researchers at Bell Labs/AT&T successfully demonstrated the first transistor. Their groundbreaking work was published in 1948, eventually earning them the Nobel Prize in physics in 1956. The industry began to take shape in the early 1950s, notably in 1952, when 34 companies licensed AT&T's original semiconductor patents.

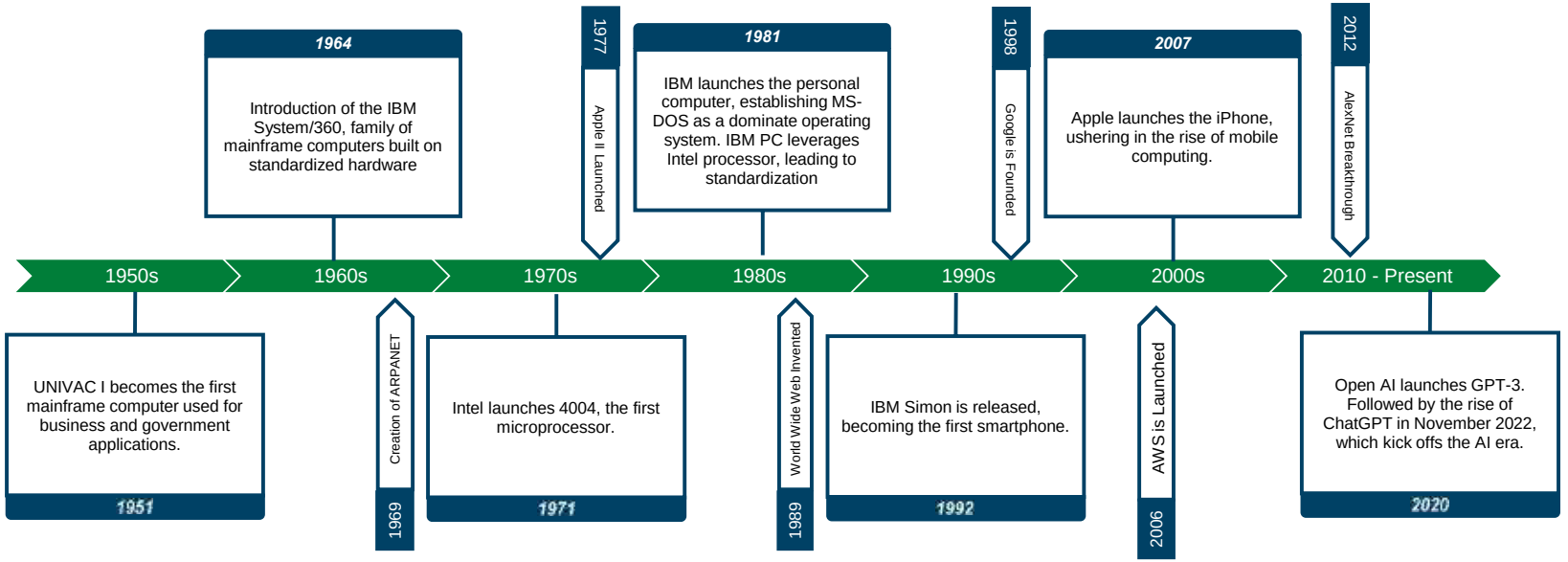
From 1950 to 1970, semiconductor companies increasingly adopted vertical integration. Pioneering firms such as Texas Instruments, Fairchild, and Motorola designed, manufactured, and packaged their semiconductor chips, primarily serving OEMs (e.g., aerospace, telecom, defense). These integrated device manufacturers (IDMs) developed their own process technologies and produced chips in-house, maintaining control over their entire value chain. This approach minimized value loss from double marginalization (e.g., having products marked up multiple times within the supply chain, thereby resulting in higher end-user ASPs) and protected their intellectual property. The integrated business model required substantial investment in cutting-edge manufacturing facilities, which companies justified by producing high volumes of chips at premium margins.

IBM emerged as a successful vertically integrated company during this period, transitioning its mainframe systems business from vacuum tubes to solid-state transistors in the late 1950s. By the early 1960s, IBM had internalized all aspects of its operations, including software development, hardware components, manufacturing equipment, and semiconductor device production.

By 1970, however, the industry began to experience its first wave of deconsolidation. New entrants such as National Semiconductor, Intel, and AMD gained market share by focusing on emerging applications like minicomputers, microcomputers, and eventually personal computers, leveraging new microprocessor technologies to challenge the dominance of established players.

In 1971, Intel released the 4004, the first commercial microprocessor. Boasting more and faster transistors than preceding technology, it combined multiple computational functions on a single integrated chip. The 4004 worked as a reusable module. In 1981 IBM selected Intel's 8088 microprocessor for its first personal computer, establishing a standard that would define the personal computing era. Throughout the 1980s and 1990s, Intel introduced a series of groundbreaking processors, including the 80286, 80386, 80486, and the Pentium, each iteration significantly enhancing computing power and cementing Intel's leadership as the CPU provider of choice for the PC era. These advancements not only drove Intel's revenue and influence, but also allowed the company to set industry standards (from architecture choice to packaging configurations) and drive Moore's law (named after Gordon Moore, Intel's pioneering co-founder and long-time CEO).

Exhibit 1
From Chips to Systems
A Brief History of Computing



Source: William Blair Equity Research

A Computer for Everyone—Rise of the PC and CPU

In the PC era, the CPU quickly became the primary volume driver and largest revenue generator for the semiconductor industry, establishing itself as the foundational platform that propelled innovation in silicon technology. Intel emerged as the primary beneficiary of this transformative shift to CPU-centric computing, leveraging its dominant position to set the pace and direction of Moore's law. Moore's law, which predicted the doubling of transistors on a microchip approximately every two years, was exemplified by Intel's relentless pursuit of higher performance and greater efficiency in its processors.

Through the 1990s and 2000s, Intel's groundbreaking transistor innovations, such as the introduction of strained silicon, high-k metal gate technology, and tri-gate (3D) transistors, underscored its CPU-first strategy. These advancements not only enhanced the performance and capabilities of CPUs, but also established benchmarks for the entire semiconductor industry.

Foundry suppliers, initially catering to diverse applications, adapted these innovations, cascading them into broader semiconductor applications and reinforcing the pivotal role of CPU-centric design in driving technological progress. This era of CPU dominance catalyzed a virtuous cycle of innovation, fueling the exponential growth of personal computing and shaping the trajectory of silicon technology for decades.

As the internet economy expanded in the late 1990s and early 2000s, Intel leveraged its technological expertise from the PC market into the data center space. The introduction of the Pentium Pro in 1995 and the Xeon brand in 1998 marked Intel's dedicated push into servers and workstations.

X86 architecture and Intel's multidecade dominance. One of the key underpinning drivers of Intel's dominance in computing was the popularity of its x86 architecture. It began with the introduction of the Intel 8086 processor in 1978, followed by the 8088 in 1979. The architecture gained significant traction when IBM chose the 8088 for its IBM PC in 1981, establishing x86 as the standard for personal computers. The open architecture of the IBM PC allowed other manufacturers to produce compatible systems, further spreading the use of x86 processors. The "Wintel" partnership, combining Intel hardware with Microsoft software, solidified x86's dominance in the PC market as Microsoft's operating systems became widely used.

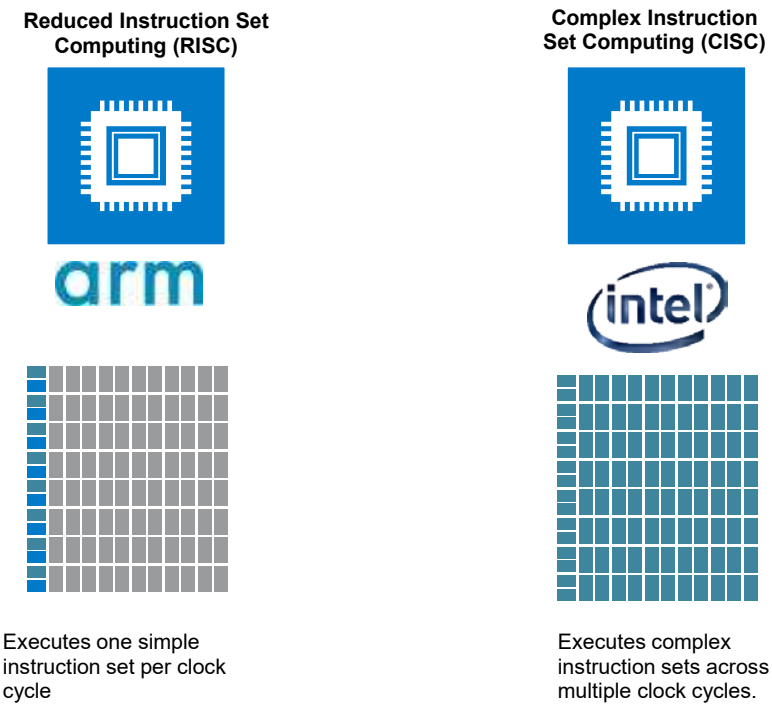
Throughout the 1980s and 1990s, Intel continued to evolve the x86 architecture with the introduction of the 80286, 80386, and 80486 processors, each bringing advancements such as protected mode, 32-bit architecture, and integrated floating-point units. During this time, competitors like AMD and Cyrix emerged, producing x86-compatible processors and driving innovation. The launch of the Pentium brand in 1993 marked a new era of performance with superscalar architecture, further entrenching x86 in the consumer market.

The subsequent improvements in multicore technologies and cost-performance efficiency of x86 processors facilitated the migration from RISC (Reduced Instruction Set Computing) and proprietary architectures to x86 in data centers. For example, IBM had its POWER architecture, used in high-performance servers; Sun Microsystems had SPARC (Scalable Processor Architecture) systems, which were popular for their performance and scalability; and HP used its PA-RISC (Precision Architecture-RISC) processors in its servers.

One of the primary advantages of shifting to x86 architecture was cost efficiency. The x86 processors were mass produced for the PC market, leading to economies of scale that significantly reduced the cost per unit compared to proprietary RISC processors. This made x86-based servers more affordable for a broader range of enterprises.

In the late 1990s and early 2000s, the x86 architecture became a popular choice for servers and data centers, driven by the growth of the internet and enterprise computing. Intel's Xeon processors, specifically designed for servers, and the development of virtualization technologies like VMware and Hyper-V helped establish x86 as the standard architecture in data centers. AMD's introduction of the AMD64 (or x86-64) architecture, which extended the instruction set to 64 bits, also contributed to x86's continued relevance, allowing for greater memory addressing and improved performance.

Exhibit 2
From Chips to Systems
Understanding the X86 Architecture



Sources: William Blair Equity Research

The rise of cloud computing in the 2010s further solidified x86's dominance, with major cloud providers like Amazon, Google, and Microsoft opting for x86-based servers in their hyperscale data centers. Thus, Intel's early success in personal computing directly contributed to the widespread adoption and dominance of the x86 architecture in data centers.

The iPhone Phenomenon and the Rise of Mobile Computing

While RISC was on its way out of the data center, a new wave of computing would put it back into the spotlight. The journey began in the 1980s and 1990s with the development of early mobile phones, which used basic microcontrollers and digital signal processors (DSPs) to handle voice and simple data tasks. As mobile technology advanced, there was growing demand for more powerful and energy-efficient processors. This led to the creation of the ARM (Advanced RISC Machines) architecture, which became a cornerstone of mobile computing. ARM's design, characterized by low power consumption and high performance, was ideal for mobile devices. Companies like Nokia and Ericsson were early adopters, integrating ARM-based chips into their mobile phones. However, undeniably the key inflection in mobile happened with the launch of Apple's iPhone, which led to the rise of the smartphone era.

In the mobile era, the application processor unit (APU), which combines a CPU, a GPU, and additional features into a single system on a chip (SoC), became the key unit of compute. APUs achieved sales volumes that surpassed standalone CPUs substantially (e.g., IDC expects 1.2 billion mobile units shipped in 2024 versus only 250 million PCs), making it the primary driver for the semiconductor foundry sector.

Over the past decade, the preeminence of the mobile SoC platform meant that APUs dictated the evolution and direction of Moore's law. As evidence, the development schedule of Apple's iPhone has exerted considerable influence on the technological advancements and strategic planning at TSMC. Currently, transistor technology is optimized primarily for mobile SoCs, focusing on enhancements that cater specifically to their need for low power consumption, compact size, and high performance. Advancements were made in materials (like FinFET or gate-all-around technologies) and manufacturing processes tailored to the unique requirements of mobile devices (e.g., SoC and wafer-level packaging). Once these technologies are established in the mobile sector, they are then adapted for broader applications, including desktops, servers, automotive electronics, and IoT devices, scaling to meet different performance and environmental needs.

The demand for mobile chips spurred significant investment in semiconductor manufacturing and research, leading to advancements in miniaturization and power efficiency. Companies traditionally focused on desktop and server markets, like Intel, faced new competition from mobile-centric chipmakers. The consumer electronics market expanded rapidly, with mobile devices becoming the primary computing devices for billions of people worldwide. This shift forced semiconductor fabs and suppliers to devote more time and investments to the needs of mobile chip OEMs (e.g., Apple, Samsung) as mobile computing became a larger source of revenue.

Cloud Data Centers and the Emergent AI Wave

The rise of cloud data centers has transformed how businesses and individuals access and use computing resources, driven by the demand for scalable, flexible, and cost-effective IT solutions. Over the last decade, major cloud service providers such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform have significantly expanded their infrastructure to offer a wide range of services, from basic storage and computing power to sophisticated machine learning and AI tools. The evolution of these data centers has been characterized by a shift from traditional, on-premises server rooms to distributed, globally connected facilities that leverage virtualization, containerization, and automation to provide seamless service to users worldwide. This transition has enabled organizations to scale their operations quickly, reduce capital expenditure, and focus on core business activities rather than managing IT infrastructure.

In the past two years, the demand for AI processing capabilities has increased exponentially, driven by growing demand to run larger and more complex workloads on advanced AI applications, such as natural language processing, computer vision, and large-scale data analytics, and the more recent wave of building LLMs and creating new GenAI algorithms, which have all supercharged demand for processing power. The cloud data center has become an early destination for much of this compute (as the CSPs have the capacity, know-how, and money to invest at the forefront of this new technology curve), although we expect this will proliferate across all businesses as the computing systems become more powerful, more efficient, and lower cost.

We Need More Processing!

Serial Processing Versus Parallel Processing

Serial processing, or scalar processing, is the most fundamental form of computation, where a single instruction operates on a single data point. This method is highly sequential, processing one operation at a time. The main advantage of scalar processing is its simplicity and ease of implementation. While it is efficient for simple tasks and straightforward programs, its limitations become apparent with data-intensive and parallelizable workloads.

Because scalar processing is highly sequential, it can serve as a bottleneck for complex computations. For example, consider a simple task like adding two arrays of 1,000 elements each. In scalar processing, this would require 1,000 separate addition instructions, one for each pair of elements. For a more complex operation like matrix multiplication of two 1000×1000 matrices, scalar processing would need 10^9 individual multiplications and summations. These operations are typically executed on single-core CPUs or the basic cores of multicore CPUs, resulting in limited parallelism and efficiency. Performance in scalar processing is usually measured in instructions per second (IPS), highlighting the linear growth tied to clock speed and core count. In AI, early algorithms and smaller-scale models could run on scalar processors, but they quickly outgrew this capability as datasets and model complexity increased.

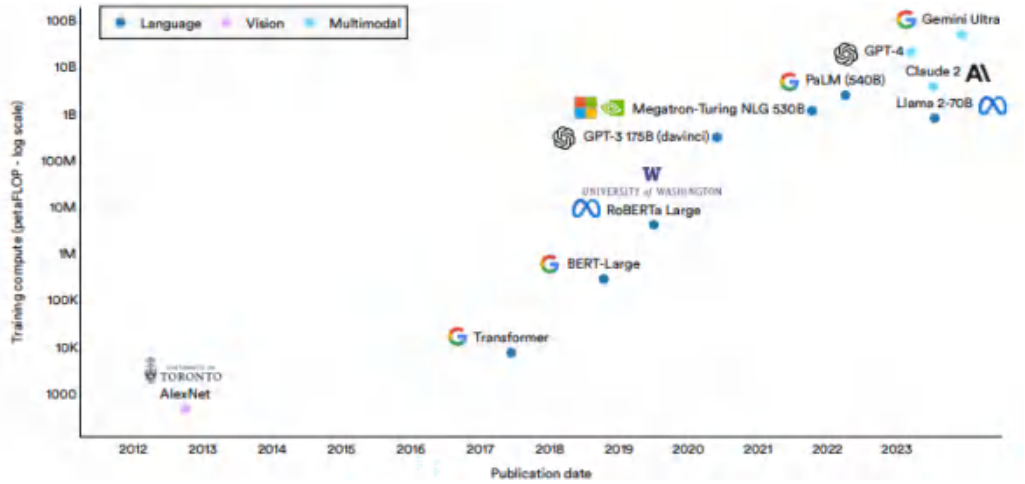
The difference in computation power between traditional CPUs and GPUs is as follows. A CPU will typically have a few cores (in the range of 2 to 16 cores), each of which can run multiple threads. Each core has a high clock speed (2-4 GHz) and a large cache memory (2-16 MB). Meanwhile, a GPU has hundreds of thousands of cores (256 to 4096 cores) that can run thousands of threads simultaneously. Nonetheless, GPU cores themselves have lower clock speeds (between 0.5 and 1.5 GHz) and come with a small cache memory (16 to 64 KB).

Vector Processing. Vector processing enhances performance by allowing a single instruction to operate on multiple data points simultaneously. This technique, often implemented in the form of vector processors or SIMD (single instruction, multiple data) units within CPUs and GPUs, significantly accelerates computations that involve large arrays or matrices. The benefits of vector processing include increased throughput and efficiency for operations like matrix multiplication, which is foundational in many scientific and engineering applications. In the realm of AI, vector processing facilitated more efficient training of machine learning models, particularly those involving linear algebra operations.

Tensor processing represents a further leap, enabling operations on multidimensional arrays (tensors), which are crucial for modern AI applications, especially deep learning. Tensor processing units (TPUs) and other specialized hardware accelerators are designed to handle the vast computational demands of training and inference in deep neural networks. The primary benefit of tensor processing is its ability to perform large-scale, parallel computations with high efficiency, drastically reducing the time required to train complex models. Tensor processing has been instrumental in the advancements of AI, enabling breakthroughs in natural language processing, computer vision, and other fields. While traditional GPUs were generally good at the parallelism for smaller data arrays, the need to perform calculations across massive datasets has pushed the industry to embrace tensor-based computations in their chips. For example, starting in 2017 with its Volta GPU architecture (V100), Nvidia introduced tensor cores that were designed specifically for the matrix multiplications required in machine learning and AI. Then in 2022, the H100 was released with an updated tensor core designed specifically to improve performance and reduce costs when training large LLMs.

How AI Changed the Game. As AI shifts into the mainstream and starts to underpin all types of applications and processes, we have seen an exponential increase in computing demand. The increased need for processing is driven primarily by the need for scale and low latency.

Exhibit 3
From Chips to Systems
LLMs Drive Exponential Demand Growth for Computing



Sources: 2024 AI Index Report, Stanford HAI

First in terms of scale, AI requires handling and processing massive datasets. These datasets are used to train models through complex algorithms that adjust and optimize to millions or billions of parameters. In contrast, traditional computing tasks rarely involve such extensive data manipulation and parameter tuning. Traditional tasks often rely on transaction processing, data retrieval, and software applications that require less computational intensity on a per-operation basis. The scale of AI workloads puts increased demand on core processing capabilities (in terms of FLOPs) and more broadly on more memory/storage, networking, and power demand. For example, ChatGPT queries need nearly 10 times as much electricity as a traditional Google search query—i.e., a single ChatGPT query requires 2.9 watt-hours of electricity compared to 0.3 watt-hours for a Google search, according to the International Energy Agency.

FLOP stands for "floating-point operation." A floating-point operation is a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division. The number of FLOPs a processor or computer can perform per second is an indicator of its computational power. The higher the FLOP rate, the more powerful the computer is. An AI model with a higher FLOP rate reflects its requirement for more computational resources during training.

Second, in terms of speed of responses, AI systems, particularly for inference (model serving), require low latency, particularly as more of these AI systems are required to do real-time processing of data. For example, autonomous vehicle systems and real-time speech recognition/translation require instantaneous processing capabilities.

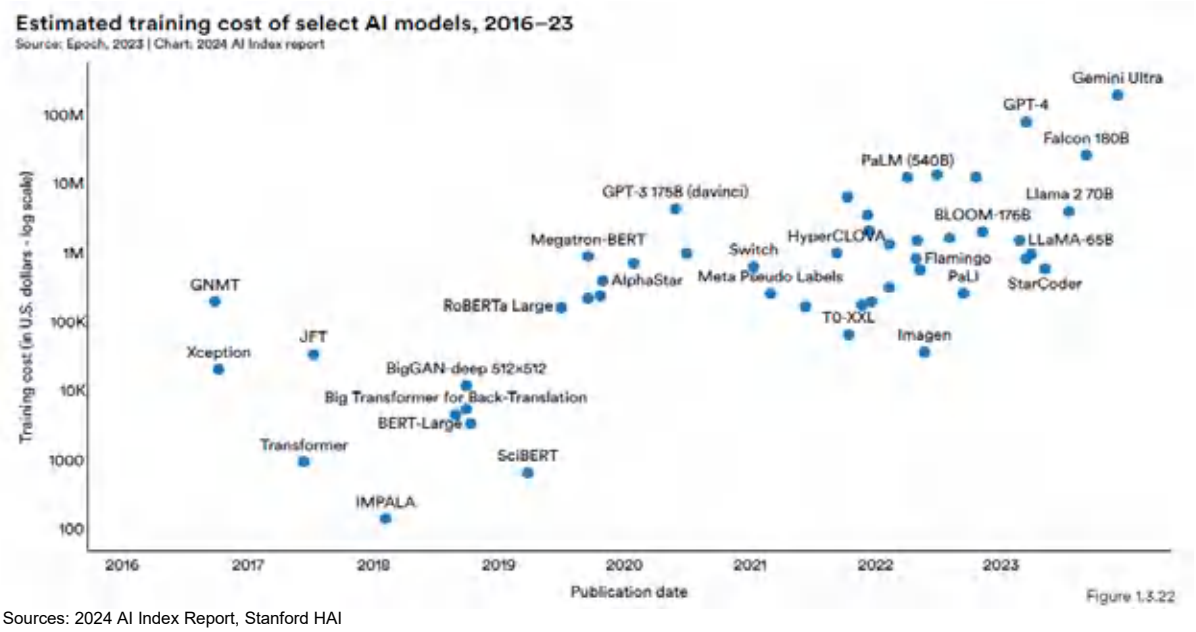
Because the scale and latency of serial processing can be improved only through fundamental improvements in the processing core, the industry has worked around this limitation by shifting its focus to parallel processing. While CPUs can perform a range of operations, including integer, control, floating-point, and I/O operations, GPUs are optimized for floating-point operations and mathematics calculations that can be broken down into smaller parallelized functions. In other words, CPUs are good at general-purpose computing and compute anything we program.

A high-powered CPU can give us one answer quickly, and this characteristic is known as latency: the time it takes to go from cause to effect. GPUs, on the other hand, are often concisely described as trading somewhat lower latency for much higher throughput: a measure of the number of "causes" (i.e., amount of data) being turned into results over time.

On the inference side, latency remains incredibly important, with more limited, diminishing returns than in training—i.e., it probably matters less to the end-customer if training takes a bit longer than if it takes several seconds longer for the customer to get a response from the AI. AI models require considerable computational resources for inference, which still significantly exceeds the demands of traditional computing tasks. AI inference involves applying a trained model to new data to make decisions or predictions. While this process may seem less demanding than training, it still requires high-performance hardware, especially for applications that need real-time processing. Even during inference, AI models benefit from the parallel processing capabilities of GPUs. For example, a model serving multiple users or performing tasks on large datasets simultaneously can leverage thousands of GPU cores to handle many requests in parallel. The general rule of thumb is that the cost of inference come in at roughly the square root of the training costs—though the flip side is that the addressable market for inference is an order of magnitude larger than training, which is dominated by the largest technology companies in the world (Meta, Microsoft/OpenAI, Google).

Bigger Models Require More Compute. Generally, the complexity of the model and the size of the underlying training dataset directly influence the amount of compute needed. The more complex a model is, and the larger the underlying training data, the greater the amount of compute required for training. Over the last five years, the compute usage of notable AI models has increased exponentially. For example, while the original AI transformer model released in 2017 required 7,400 petaFLOPs to train, Google's Gemini Ultra required 50 billion petaFLOPs. Large LLMs, like OpenAI's GPT-4o (trained on 13 trillion tokens) and Meta's LLaMa 3.1, require immense processing capacity to train (see exhibit 3). This exponential growth in computing demand has also driven a similar exponential growth in training costs (see exhibit 4).

Exhibit 4
From Chips to Systems
Costs of Training Expand With Demand for Computing

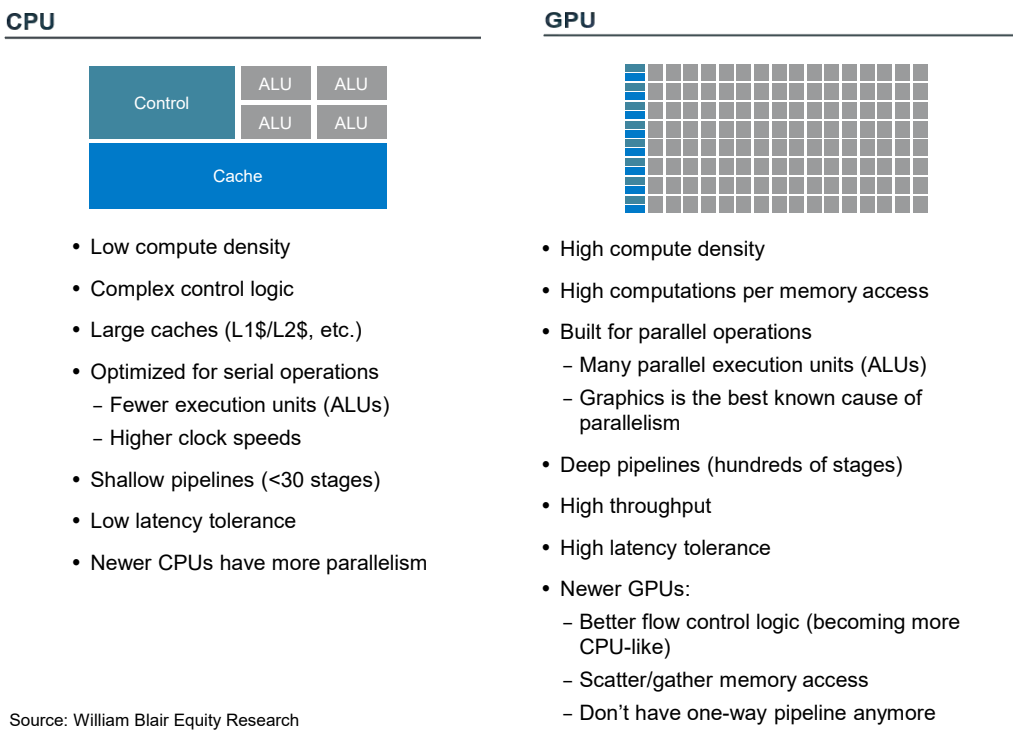


Sources: 2024 AI Index Report, Stanford HAI

In the data center, this means that parallel processing has moved out of its niche in high-performance computing (HPC), gaming, and visualization to take center stage in the AI revolution. This is driving a shift in market share from serial computing, which dominates the industry with CPUs from Intel and AMD, toward parallel computing, where leader Nvidia retains greater than 90% of market share.

While the GPU market will undoubtedly only get more competitive as other companies try to grab market share, we expect that at least over the next 12 months Nvidia will be able to maintain its high market share. This is largely a function of Nvidia having already reserved a substantial majority of TSMC's production capacity through 2025. While TSMC is working to expand its production capacity, with plans to grow the all-important CoWoS capabilities at its fab at a 60% compound annual rate through 2026, the complexity and capital-intensive nature of building new fabs means that supply limitation will likely impede the ability for Nvidia competitors to take meaningful share in parallel computing, at least in the near-term.

Exhibit 5
From Chips to Systems
Comparing the Functions of a CPU With GPU



Source: William Blair Equity Research

The math of parallelization. The vector size in vector processing is determined by the architecture of the processor, specifically the width of the SIMD units or vector registers. This width dictates how many data elements can be processed in parallel with a single instruction.

For example, take a theoretical array of where N=100 and the vector size is V=10. Scalar processing would require NxNxN instructions, or 1,000,000 elements. In vector processing, we would divide each dimension of the array by the vector size. So, the number of instructions needed is (N/V)×(N/V)×(N/V), which in our basic example would result in 1,000,000 instructions divided by the 1,000 vector reduction factor (e.g., 10x10x10), which would result in only 1,000 instructions needed with vector processing. In other words, with vector processing, we can significantly reduce the number of instructions by leveraging the ability to process multiple elements simultaneously. The reduction factor, V^3, highlights the efficiency gains of vector processing over scalar processing, especially as vector sizes increase.

GPUs have a specialized instruction set architecture (ISA) that includes vector operations optimized for the hardware. For example, Nvidia's CUDA and AMD's ROCm provide programming models that allow developers to leverage the full capabilities of the vector units. Nvidia GPUs, for instance, are built around the concept of warps, where each warp consists of 32 threads. Each streaming multiprocessor (SM) can manage multiple warps concurrently, and the vector size (in terms of data elements processed per instruction) is linked to the warp size. With advancements in architecture, such as the move from Volta to Ampere to Hopper to Blackwell, Nvidia has increased the efficiency and throughput of its vector units, allowing for higher performance in parallel tasks.

While clock speeds are not everything, they give us an idea of how quickly instructions tick through a device. The clock speeds of modern CPUs are roughly twice as high as those of the latest and greatest GPUs. As a result, GPU cores are typically slower for serial tasks, but gain a significant advantage when it comes to vector multiplications. While the cores in a GPU are not as fast individually and far more specialized than the cores in a CPU, there are far more of them (16,384 CUDA cores in Nvidia's high-end consumer RTX 4090), so throughput is high for operations that can be parallelized favorably for GPUs.

Training Versus Inference

The compute and processing demands for machine learning inference differ significantly from those of training, reflecting the distinct roles these processes play in deploying AI models. During training, the primary objective is to optimize the model's parameters (weights) by processing large datasets through a series of iterative computations. This involves complex operations like forward and backward propagation, which require substantial computational power and time. Training is highly resource-intensive because it needs to handle massive amounts of data and perform numerous mathematical operations to adjust the model parameters over multiple epochs. High-performance hardware, such as GPUs or TPUs, is typically used to speed up these operations through parallel processing, allowing each data batch to be processed independently. This makes training suitable for environments with abundant computational resources, like data centers.

In contrast, inference is the process of applying a trained model to new data to make predictions or classifications. Unlike training, inference involves only the forward pass through the model, which requires less computation. However, inference often demands real-time or near-real-time processing, especially in applications such as autonomous vehicles, online recommendations, or interactive AI systems. Therefore, the focus during inference is on reducing latency and improving throughput to handle a high volume of predictions per second efficiently. Inference can often be run on more specialized and energy-efficient hardware, including edge devices or specialized inference accelerators, which are optimized for speed and lower power consumption. Techniques

like model compression, quantization, and pruning are frequently used during inference to enhance performance and reduce memory usage, allowing models to be deployed across various platforms with different hardware constraints.

Exhibit 6
From Chips to Systems
Comparing AI Training and Inference Workloads

Metric	Training	Inference
Nature of Workload	High computational demand, iterative optimization	Lower computational demand, forward pass only
Resource Requirements	High memory and processing power, energy-intensive	Optimized for efficiency, lower power consumption
Deployment Environment	Data centers with powerful compute clusters	Cloud and edge devices, requires flexibility
Scalability	Scalable across large clusters of GPUs/TPUs	Scalable across distributed networks
Optimization Focus	Algorithmic efficiency and convergence speed	Model compression and latency reduction
Hardware Used	High-performance GPUs/TPUs	Specialized accelerators, edge devices
Precision	Often full precision (FP32/FP16)	Reduced precision (e.g., INT8) for efficiency
Real-Time Needs	Less focus on real-time processing	Emphasizes real-time or near-real-time processing

Source: William Blair Equity Research

In terms of deployment environments, training is typically conducted in centralized data centers where access to powerful compute clusters allows for distributed processing across multiple nodes. In contrast, inference can be deployed in diverse environments, ranging from cloud servers to edge devices like smartphones and IoT devices. This requires models to be efficient and adaptable to the specific hardware available in each environment. For example, Apple Intelligence, which is being integrated into the latest iPhone 16, will run foundational models locally, on-device.

Furthermore, while training is often scaled horizontally across many GPUs or TPUs to accelerate the process, inference needs to be scalable across distributed networks to handle varying loads effectively. This often involves using cloud-based infrastructure that can dynamically scale according to demand. The optimization focus also differs between the two processes, with training prioritizing algorithmic efficiency and accuracy, while inference emphasizes model compression and speed without significantly impacting accuracy.

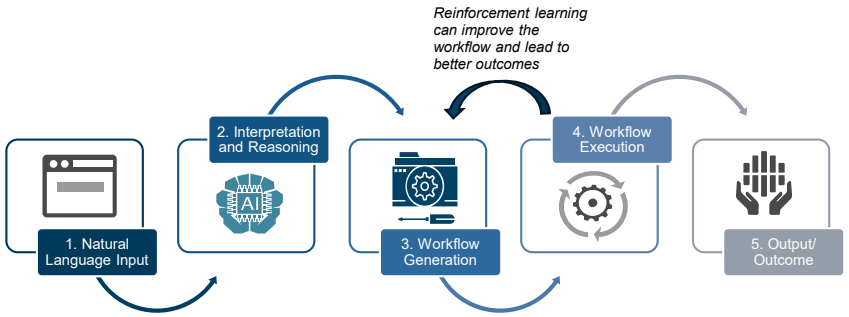
To date, a majority of spending on AI infrastructure has been on the training side. Nvidia estimates roughly 60% of its GPU sales are associated with training clusters. However, over time as more applications leverage AI capabilities, inference use-case and spending should exceed that of training. That is because while the number of companies spending on training LLMs will likely be limited to a handful of large companies and several upstarts, the number of companies using AI in their applications and workflows is substantially larger—beyond the largest technology companies engaged in an "arms race" for building the next-best foundational model, most other organizations will focus on fine-tuning existing models or training smaller, more targeted small language models (SLMs). That means that over time, the inference market size will outgrow the training market substantially and drive sustained demand for AI infrastructure even as investments in training clusters potentially scale back in 2026/2027.

Agentic Systems Require High Compute Capabilities

In AI, an agentic system is characterized by its ability to autonomously interact with its environment and execute tasks, often leveraging large language models for processing and decision-making. In other words, it acts as a layer of intelligent software that can perceive its environment, make decisions, and take actions to achieve specific goals. Agentic AI systems often require many different large language models, each focused on different tasks or parts of a workflow. Current GenAI chatbots and LLMs deliver only about a 15% success rate in complex agentic tasks according to a 2024 Report from Stanford's Human-Centered AI (HAI) Institute. To overcome this and LLMs' inherent gaps in planning, reasoning, and coordination, each of the LLMs in an agentic AI System needs to be manually tuned and optimized.

Agentic AI systems require a dedicated semantic layer to ensure the LLMs not only understand the context of the data, but also retain a shared context across each of the different models within the agentic AI system. Agentic AI systems also need to be optimized at the infrastructure and hardware layer. This requires making a set of highly complex trade-offs across factors such as token(s), cost, batch size, concurrency, and more, depending on the use-case.

Exhibit 7
From Chips to Systems
Understanding Agentic AI Workflows



Sources: William Blair Equity Research

The construction of agentic AI systems typically involves several key components:

1. ***Perception:*** The agent needs sensors or data inputs to gather information about its environment. This could be through cameras, microphones, or data feeds, depending on the application.

2. ***Decision-making:*** A core component is the decision-making mechanism, often implemented using machine learning models such as neural networks, reinforcement learning algorithms, or other AI techniques. This allows the agent to process information and determine appropriate actions.

3. ***Action execution:*** The agent must have a way to interact with its environment, whether through physical actuators in robotics or digital actions in software systems.

4. ***Learning and adaptation:*** Many agentic systems incorporate mechanisms for continuous learning, allowing them to improve their performance over time based on experience and feedback.

5. ***Goal representation:*** The system needs a way to represent and prioritize its objectives, which guide its decision-making process.

Today, agentic AI is starting to be used to improve the performance of AI applications across a range of industries, including powering advanced driver-assistance systems (ADAS) in autonomous cars, assisting medical professionals in analyzing medical data and monitoring patient health, and providing improved customer service through intelligent chatbots that can also take actions to resolve queries.

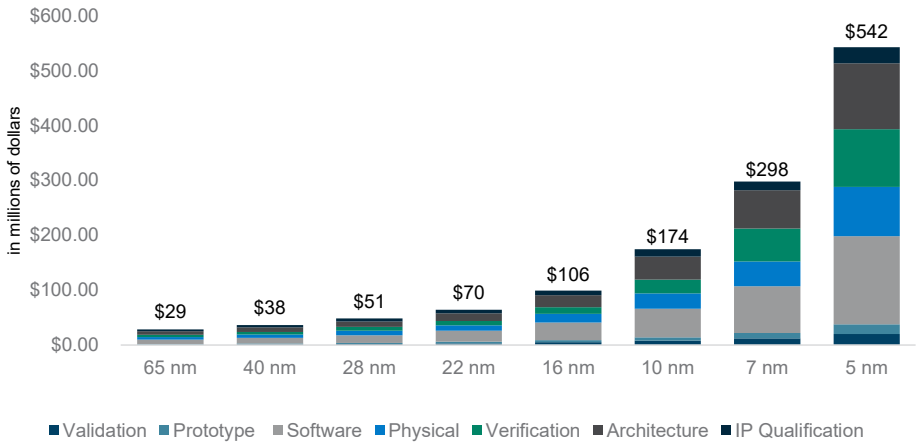
To implement agentic AI, enterprises are leveraging AI libraries like OpenAI Gym, RLlib, and Stable Baselines for training agents in simulated environments. Multi-agent frameworks like JADE (Java Agent Development Framework) and MESA (a similar Python-based framework) allow developers to simulate multiple agents interacting with each other to ensure behavior matches expectations. In robotics, the Robot Operating System (ROS) is a go-to framework for agent-based robotics. It offers a collection of tools, libraries, and conventions that facilitate communication between different components of a robotic system and provides tools for simulation, visualization, and debugging.

Agentic AI is also pulling in demand for contextual data. Graph database vendors like Neo4j are leveraged to help create knowledge graphs that allow agents to improve their reasoning capabilities by better understanding how data is interconnected.

It's Not Just a Chip, It's a System

The AI computing wave comes at a time when performance improvements in processing power are becoming harder and more expensive to come by. The shift from 10 nm to 7 nm to 5 nm to 3 nm transistor node size has taken longer each time, with TSMC, Samsung, and lately Intel all spending substantially larger amounts on fabrication, testing, assembly, and configuration (see exhibit 8). According to TSMC, building a 3 nm factory costs roughly $20 billion. Importantly, Moore's law for many years was driven by a consistent reduction in the size of the transistor, which means that chips could pack more and more of these transistors on the same size chip, driving improvements in processing power and FLOPs. As transistor science starts to run into the limits of physics—measured in atoms—the well-known Moore's law dynamic has had to evolve to keep pace with the continued need for performance improvements in computing.

Exhibit 8
From Chips to Systems
Rapid Rise in Development Costs for Smaller Chips



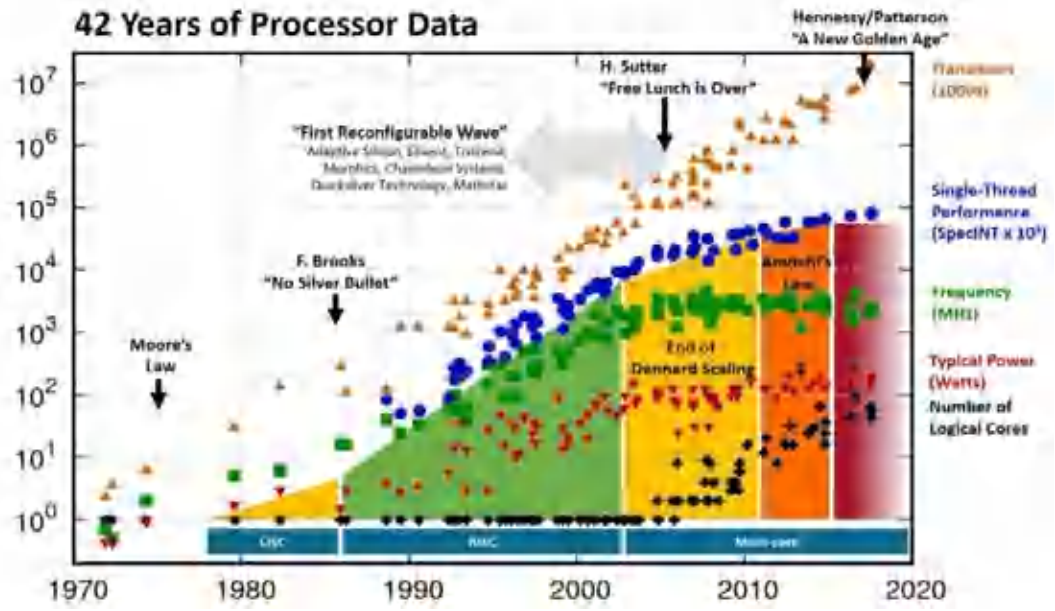
Sources: William Blair Equity Research and IBS

This has shifted the focus of semiconductor vendors to broader systems architectures. The key unit of compute is no longer the processor, but rather a broader system that connects multiple processors together with storage and networking, brought together with a layer of critical software that can also be optimized and improved as underlying hardware specifications change and as new application stacks emerge.

The End of Moore's Law?

Moore's law is a principle formulated by Gordon Moore, the co-founder of Intel, in 1965. Initially, Moore observed that the number of transistors on a microchip doubled approximately every year, leading to a corresponding increase in computational power and decrease in relative cost. In 1975, he revised his observation to predict that this doubling would occur approximately every two years. This prediction, though originally an observation of trends, became a guiding principle for the semiconductor industry, leading to exponential growth in computing capabilities and a rapid pace of technological innovation.

Exhibit 9
From Chips to Systems
Moore's Law and Scaling of Processors



Source: Hennessy and Patterson, Turing Lecture 2018

The "end" of Moore's law is attributed to several technical and economic challenges. As transistors have shrunk to atomic scales, physical limitations such as quantum tunneling, heat dissipation, and power leakage have become significant obstacles, making further miniaturization increasingly difficult and expensive. The cost of developing and manufacturing at the cutting edge of chip technology has skyrocketed, leading to diminishing returns on investment. While innovations such as 3D stacking, chiplet architectures, and alternative materials like graphene are being explored, the straightforward path of shrinking transistor sizes is no longer sustainable at the pace described by Moore's law. As a result, the industry is shifting focus from simply scaling transistors to exploring new architectures and system technologies to continue enhancing processing power.

Pendulum Swings From Horizontal to Vertical Integration

Traditionally, horizontal integration in technology has enabled vendors to focus on a specific layer of the application stack where it has the most differentiation and strongest know-how. In some ways, it allowed companies like Intel to become a quasi-monopoly in the CPU market. Nonetheless, these CPUs had to be built into a broader platform by vendors up and down the stack.

For instance, CPUs, GPUs, memory chips, and other integrated circuits are typically packaged individually. These separate packages are then mounted onto a printed circuit board (PCB), which provides the necessary interconnections through physical traces and soldered connections. This traditional approach often involves complex interconnects between components, which can lead to longer signal paths, increased latency, and potential issues with signal integrity. In addition, managing size and power constraints is handled through the separate packaging and PCB design, which may not always be optimal for high-performance applications.

Many semi players chose to focus on horizontal integration in the early days as the primary growth catalyst to quickly scale core capabilities and divest noncore assets. Standardization of various processes in the value chain, including modular IP blocks for chip design, standard foundry PDKs (process design kits), and packaging standards, helped to simplify intercompany integration challenges enough so that the fabless/foundry model eventually became the de facto and dominant business model. The fabless model really gained popularity in the 1980s and 1990s, where chip design companies could seek to dominate markets by focusing on more innovative designs and shorter delivery times. Companies that specialized and consolidated horizontally earned higher market shares and operating margins by achieving cost synergies and boosting economies of scale.

For example, fabless design companies like AMD and Nvidia were able to drive large improvements in their gross margins by outsourcing fabrication of chips to specialists like TSMC. From its perspective, TSMC was able to benefit from economies of scale by serving a large numbers of fabless chip designers, allowing them to pour even more money into the latest-generation fabrication technologies to maintain its technical leadership (e.g., 3 nm and 5 nm as well as advanced packaging techniques like CoWoS). Perhaps the most cited counterexample of this trend has been Intel, which by continuing to both design and fabricate chips hamstrung its ability to leverage third-party fabs (since its designs were for their specific fab technologies and not easily portable) while its fab business fell behind the innovation curve.

Markets have also recognized the potential for greater return on investment and the improved management of risk in the fabless/foundry model (where there is keen focus on domain specialization). Accordingly, the market has generally rewarded successful fabless companies with higher valuation multiples than their integrated design manufacturer (IDM) counterparts. Even IDMs are realizing the positive effects of specialization in specific end-markets. This is due to a greater focus on microprocessors for key verticals like AI and deep learning accelerators. Intel, for example, has made divestitures that include McAfee in 2016, Wind River in 2018, and its NAND memory business to SK Hynix in 2020.

Mobile revolution sparks early shift to vertical integration. The rise of mobile computing drove a shift toward the system-in-package (SiP) design approach, which replaces the traditional discrete component packaging method. The transition to SiP design began in earnest around the late 2000s and early 2010s. This shift was driven by the increasing demand for miniaturization associated with the rise of smartphones (i.e., every new generation focused on smaller and more compact components to pack into a relatively fixed size product).

SiP design allows for the integration of multiple components into a single package, significantly reducing the overall size of the system. By integrating components into a single package, SiP reduces the distance between them, which minimizes signal delays and improves data transfer speeds—crucial for high-performance computing and mobile devices. Lastly, SiP design offers cost and manufacturing advantages by streamlining the packaging process and reducing supply chain complexity, which contributed to its widespread adoption. As a result of these cost and time-to-market benefits for SiP, semiconductor companies and customers are starting to think less of individual chips and more about the chip system that is used for processing—this has led to the rise of SoC.

We expect the adoption of AI will have a similar structural impact as mobile had on the semi ecosystem and supply chain. While with PCs fabs shifted their supply chains to be optimized for CPU-first technologies, in mobile there was a shift to SoC-first approaches as the unit volume for smartphone chips quickly outpaced the CPU. With the data center GPU becoming more of a system solution rather than a discrete chip solution, we have seen a similar shift within the supply chain to refocus on building accelerated computing systems. For example, TSMC has invested heavily in its CoWoS (chip on wafer on substrate) packaging technology to build the latest generation of Nvidia GPUs (Blackwell), transitioning older fabs to this new technology in anticipation of the changing demand trends from AI. The planned 60% annual growth for these CoWoS facilities and the nearly $70 billion TSMC is making in capex investments for its latest fabs through 2025 is an indicator that we remain in a period of ramping-up supply, driven by visibility and strong demand trends from its end-customers.

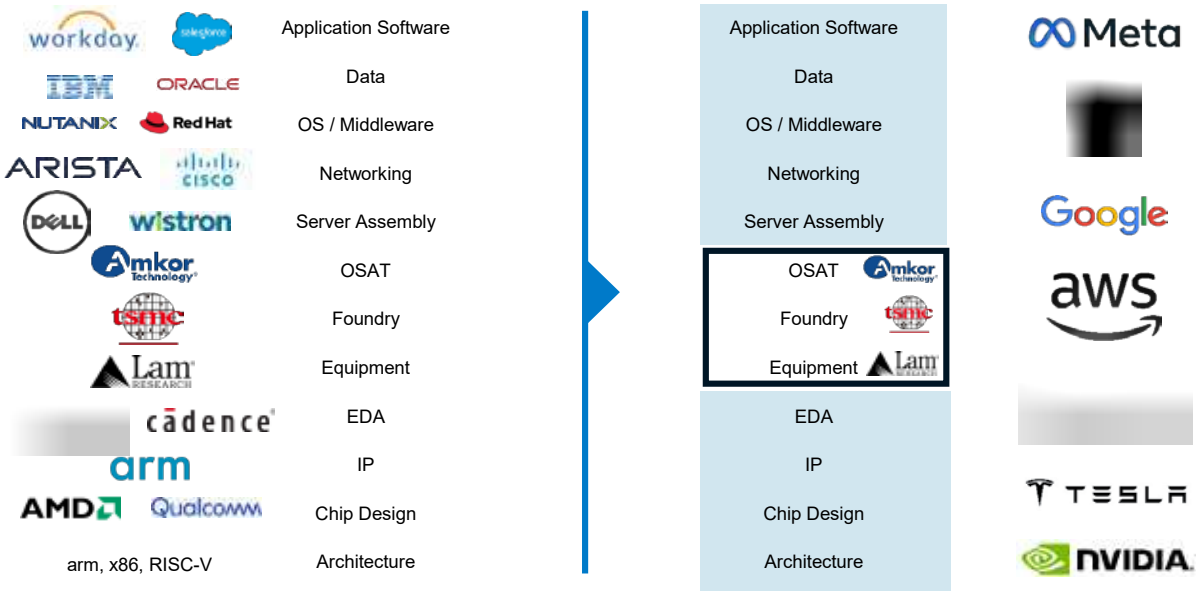
Shift to selling solutions. The pace of innovation in transistor science is no longer the primary driver of semiconductor development. With the breakdown of Moore's law, specific use-cases are now the main drivers of semiconductor technology—fostering a shift from selling devices to selling solutions. Similarly, consumer personas have also changed. Rather than targeting the engineering and procurement departments of immediate downstream customers, semiconductor companies must coordinate their activities with a wider array of ecosystem partners—with the end-customer use-case squarely in mind.

High-growth use-cases include 1) edge/cloud, 2) automotive, and 3) AI, where incremental or even marginal improvements in different parts of the system can have a significant impact at scale on costs, performance, and energy efficiency.

It has become increasingly beneficial to co-design the chip along with the entire system and software layers. This is reminiscent of the early days of the semi industry with AT&T/Bell Labs and IBM in their 1960s to 1970s heyday. Today, bolstered by the fabless model, system integrators and software companies began investing in workload-specific chip designs and system architectures.

This compounded with the commoditization of semiconductor manufacturing has led to the vertical reconsolidation in all parts of the value chain, except for manufacturing (foundries and equipment), as outside of the most advanced nodes, manufacturing technology became commoditized in an industry now dominated by fabless players.

Exhibit 10
From Chips to Systems
AI Leaders Increasingly Embrace Verticalized Approach to Building Data Center Stacks



Note: Logos are meant to be representative and are not an exhaustive list
Source: William Blair Equity Research

Hyperscalers embrace DIY. One customer segment that has been leading this shift back to increased verticalization is the hyperscalers themselves. Hyperscalers are further optimizing their stacks by designing their own chips for consumption by their consumer or enterprise products. Hyperscalers like Apple, Google, AWS, and Meta further benefit from in-house chip design because of their unparalleled access to data, strong ecosystem partnerships, and grounded understanding of end-customer use-cases. Greater control of supply chain and manufacturing processes leads to lower costs and economies of scale. Because optimization of these systems can unlock massive improvements in processing power, the vertical design approach can have immense economic impact—e.g., even a 10% performance improvement on a $1 billion data center (i.e., between 15,000 and 20,000 H100 GPUs) can have a dramatic impact on cost performance.

One high-profile example of this strategy is Apple's shift in 2020 to using in-house central processors (M1) for Mac computers instead of the Intel chips that powered Apple products since 2006. In addition to cost benefits, these new processors serve as a critical focal point of Apple's product differentiation strategy, as the company has touted its custom chip's optimized performance specifications.

Beyond Apple, several leading cloud companies have started to design their own custom chips.

- AWS first unveiled its in-house CPU, Graviton, in 2018, and followed it up with a line of parallel processors Trainium and Inferentia targeted at AI use-cases.
- Google has worked on its TPUs for a long time, leveraging Broadcom to help design and fabricate these chips.
- Starting in 2020, Meta has been developing custom silicon (MTIA) for its content recommendation and ad ranking feeds.
- Microsoft Azure's project Maia is focused on building its own GPU (first announced in November 2023) and it will be paired with the Arm-based Cobalt CPU.

Exhibit 11
From Chips to Systems
Growing Number of Custom AI Chips

AI Accelerator	Manufacturer	Description	Target Use-Case
Tensor Processing Units (TPUs)	Google	Custom-designed ASICs optimized for machine learning workloads, particularly inference. Used in data centers and available on Google Cloud Platform.	High-performance cloud inference
TensorRT and Jetson Platform	NVIDIA	TensorRT is a deep learning inference optimizer for Nvidia GPUs, and the Jetson platform provides AI capabilities for edge computing and robotics.	High-performance inference and edge AI
Movidius Myriad X	Intel	Vision Processing Unit (VPU) designed for efficient AI inference in edge devices, with a dedicated neural compute engine.	Edge devices and AI applications
Neural Engine (ANE)	Apple	A specialized processor in Apple devices to accelerate machine learning tasks on-device, optimizing for low power consumption.	On-device AI processing in Apple devices
Inferentia/Trainium	Amazon	A combination of training/inference chips designed to optimize cost and performance for workloads in AWS cloud environments.	Scalable cloud inference
Versal AI Core Series	AMD (Xilinx)	Combines adaptable hardware and software programmability to deliver efficient AI inference capabilities, optimized for CNNs and other AI models.	Data centers and edge devices
Groq LPU	Groq	A tensor streaming processor designed for high-performance inference, focusing on simplifying AI compute with low latency and high throughput.	High-performance AI inference
Huawei Ascend AI Processors	Huawei	A series of AI processors designed for both training and inference, optimized for high performance and energy efficiency.	Data centers and edge AI applications
Baidu Kunlun	Baidu	AI accelerator designed for deep learning tasks, providing high performance for both cloud and edge deployments.	Cloud and edge AI processing
SambaNova Systems RDU	SambaNova	Reconfigurable Data Unit (RDU) designed for both AI training and inference, emphasizing flexibility and efficiency.	Data centers and AI model training/inference
Cerebras Wafer-Scale Engine	Cerebras	A massive chip designed for AI training and inference with high memory bandwidth and compute power, designed for handling large-scale neural networks.	Large-scale AI workloads
ASICs by Cambricon	Cambricon	A series of AI chips designed for efficient inference across a variety of applications, focusing on flexibility and performance.	General AI inference across applications
Qualcomm AI Engine	Qualcomm	Integrated into Snapdragon processors, optimized for efficient AI inference on mobile and embedded devices, providing power-efficient performance for on-device AI.	Mobile and embedded AI inference
Samsung Exynos NPU	Samsung	Neural Processing Unit integrated into Exynos processors, designed for efficient on-device AI processing in smartphones and other devices.	On-device AI in smartphones
MediaTek APU	MediaTek	AI Processing Unit integrated into MediaTek chipsets, providing dedicated AI capabilities for mobile and IoT devices.	Mobile and IoT AI applications

Source: William Blair Equity Research

Most recently, OpenAI has made the news for its desire to build its own chips. One potential vendor that could help OpenAI in this journey is Broadcom, which was reported to be in talks with OpenAI for custom AI accelerators. Today Broadcom has a roughly $15 billion run-rate business building custom AI chips for Google and Meta, with ByteDance being added as a third customer in March 2024.

Over the longer term, the increased in-house development of chips remains a key risk for semiconductor companies, especially since these OEMs and hyperscalers are today the largest customers of GPUs and CPUs from companies like Intel, AMD, and Nvidia. Of the three companies we are initiating on, Nvidia is probably at the highest risk of being impacted by these in-house competitive offerings. Broadcom, in addition to having a more diversified business with substantial software-based revenue, is also seen as a strategic partner for building custom chips and helping provide the raw silicon used by hyperscalers in building out their networks and data centers. Meanwhile, the ubiquity of ARM architectures across in-house chip designs means that the company is largely unaffected by changing market shares among the different players. ARM's success is more correlated to the overall market for AI infrastructure and the potential share gains versus Intel's x86 architecture.

As for Nvidia's, while in-housing chip designs could weigh on demand for its GPUs, early data points from our conversations with the CSPs have highlighted the difficulty of building a competitive chip to Nvidia's GPUs, but perhaps more importantly, the challenge is recreating a developer ecosystem that can compete with CUDA. Nvidia's software supremacy remains its greatest advantage and should limit the ability of in-house chips to jump in front of Nvidia on the price-performance curve. While Nvidia controls both the chip and software running on top, until a viable CUDA alternative emerges, the OEMs building their own chips will be limited to controlling only half of this hardware-software combination.

Amdahl's law and system-level improvements. Amdahl's law deals with the performance limits of computing systems, particularly in parallel processing. Formulated by Gene Amdahl in 1967 (one of the creators of the original IBM mainframe), Amdahl's law states that the speedup of a task using multiple processors is constrained by the portion of the task that cannot be parallelized. Even if processors become more powerful as predicted by Moore's law, Amdahl's law explains that there are diminishing returns in performance improvements when increasing the number of processors if parts of the task remain sequential. This highlights a critical limitation in system performance, emphasizing that effective parallelization is essential to fully exploit hardware advancements.

Put simply, if a significant part of the computation cannot be parallelized, the performance improvements from adding more processing cores will be limited. Thus, developers must focus on designing algorithms and software that maximize the parallelizable portions of a task, minimize interdependencies, and ensure an even distribution of workloads across available cores. By doing so, they can overcome bottlenecks and make full use of the increased computational power offered by modern hardware, thereby achieving greater performance gains and efficiency.

The consequences of Amdahl's law for semi companies have been manyfold. Amdahl's law has driven the development of heterogeneous computing platforms that combine different types of processing units (e.g., CPUs, GPUs, TPUs) to handle various tasks optimally, leveraging their specific strengths.

The implications of Amdahl's law have also driven semiconductor companies up the stack, designing software that can compel the best price performance out of these systems. They have pushed semiconductor companies to focus on optimizing algorithms to minimize sequential processing. Techniques such as model parallelism and pipeline parallelism are employed to distribute computations more effectively across multiple processing units.

In sum, Amdahl's law illustrates the diminishing returns of adding more processors, guiding AI system architects to balance between increasing computational resources and improving algorithmic efficiency. Semiconductor vendors have developed advanced compilers and toolchains that can optimize code to exploit the full potential of their hardware. These tools help in identifying bottlenecks in parallelism and optimizing sequential code paths. In addition, vendors provide software libraries and frameworks optimized for their hardware to maximize performance by taking advantage of hardware-specific features.

At the forefront of this shift toward parallel computing has been Nvidia. For example, Nvidia no longer builds just chips but entire computing systems. Its DGX server integrates more than 35,000 components and weighs 70 pounds. The complexity of these system-level solutions and associated supply chains creates an incredibly wide moat for any traditional discrete chip company to cross. In the AI era, as the emphasis in computing moves from chips to comprehensive systems, we expect systems companies will receive the bulk of semiconductor revenue, largely overshadowing discrete chip makers.

The importance of software. The software and the developer ecosystems have become a focus of semiconductor companies, understanding that optimizing the ability for apps to run on top of chips would help drive adoption of more GPU resources. Nvidia's CUDA is the best example of the power of building a software ecosystem.

CUDA (compute unified device architecture) is a parallel computing platform and application programming interface (API) that allows developers to use the full potential of Nvidia GPUs for general-purpose processing. CUDA enables developers to write software that can execute thousands of threads (a virtual processing function associated with the physical core) simultaneously, taking advantage of the GPU's massive parallel processing capabilities. This allows complex computations that were traditionally executed serially on CPUs to be performed much faster on GPUs. Today, CUDA mainly competes with AMD's ROCm and Intel's oneAPI platforms.

ROCm (Radeon Open Compute), AMD's open-source alternative, aims to offer flexibility and avoid vendor lock-in by supporting multiple hardware platforms, not just AMD GPUs. ROCm includes tools like HIP (Heterogeneous-compute Interface for Portability), which facilitates the porting of CUDA code to ROCm with minimal changes. This makes it a compelling choice for organizations looking for more hardware versatility. While today CUDA remains by far the largest GPU ecosystem (with more than 5 million developers worldwide), other vendors have focused on the concept of open and interoperable software to differentiate and appeal to a broader set of customers wary of vendor lock-in.

Nvidia has also played a significant role in the advancement of deep learning and AI by optimizing popular frameworks such as TensorFlow, PyTorch, and Caffe for GPU acceleration. An AI framework is a software library or platform that provides developers with the necessary tools and abstractions to design, build, train, and deploy machine learning models efficiently. These frameworks offer prebuilt components for constructing neural networks, optimizing model parameters, and managing data workflows, thus streamlining the development process. These frameworks use CUDA to distribute deep learning computations across the many cores of Nvidia GPUs, significantly speeding up the training and inference processes for neural networks. The development of cuDNN (CUDA Deep Neural Network library) further speeds up AI processing by providing highly optimized routines for standard deep learning operations.

Sitting on top of CUDA, Nvidia has developed a range of GPU-accelerated libraries that allow various scientific and engineering applications to benefit from parallelization. Libraries such as cuBLAS (for linear algebra operations), cuFFT (for fast Fourier transforms), and Thrust (a parallel algorithm library) provide developers with the tools needed to exploit GPU parallelism without writing custom parallel code from scratch. Today, competitive offerings include AMD's accelerated libraries (AMCL) and Intel's Math Kernel Library.

For gaming and graphics use-cases, Nvidia has developed mRTX technology that uses specialized cores in its GPUs, called RT Cores, to perform ray tracing operations in real time. Ray tracing requires tracing the paths of millions of rays of light as they interact with surfaces in a scene. This process is highly parallelizable because each ray can be processed independently, making it ideal for execution on a GPU with thousands of cores. By optimizing both hardware and software for ray tracing, Nvidia has enabled real-time rendering of highly realistic graphics in video games and simulations, previously achievable only in offline rendering systems used in film production. Competitive offerings include AMD's Radeon RX Series with RDNA 2 Architecture and Intel's Arc Graphics.

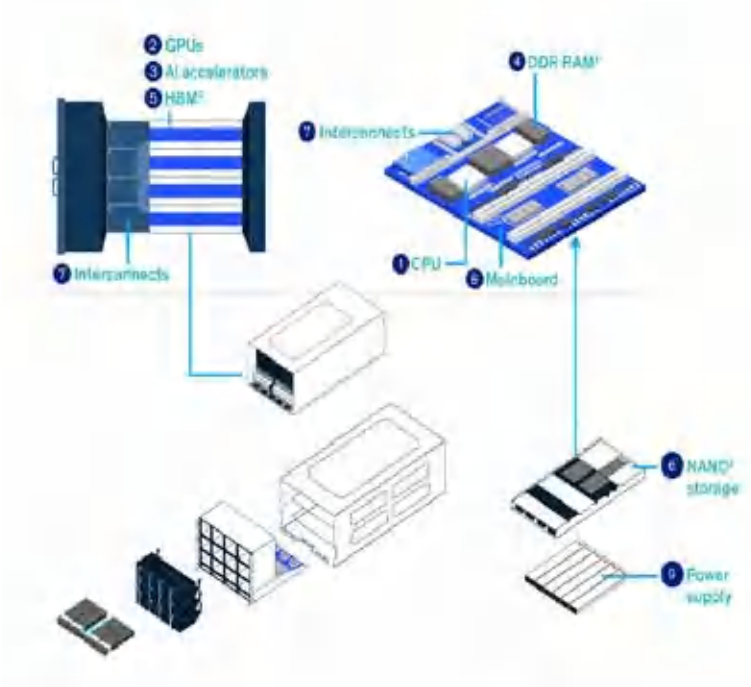
Nvidia OptiX is a ray tracing engine that allows developers to create highly realistic images using GPU acceleration. It takes advantage of the parallel processing power of Nvidia GPUs to perform ray tracing operations, dramatically reducing the time required to render complex scenes. This technology is used in a variety of fields, including design visualization, virtual reality, and visual effects production, enabling artists and engineers to achieve photo-realistic results much faster than with traditional CPU-based rendering. Competitive offerings include AMD Radeon ProRender and Autodesk Arnold and Chaos V-Ray.

In the field of autonomous vehicles, Nvidia's DRIVE platform uses GPUs to process large amounts of sensor data in real time, enabling ADAS and autonomous driving capabilities. The platform's ability to parallelize tasks such as image processing, object detection, and path planning allows vehicles to react quickly and accurately to their environment. Competitive offerings include Intel's Mobileye (still majority controlled) and Qualcomm's Snapdragon Ride. Similarly, in robotics, Nvidia's Jetson platform provides the computational power needed for real-time perception and control tasks, facilitating the development of intelligent robots.

Designing AI Clusters

AI data centers might look similar to traditional data centers, but there are considerable differences in the underlying hardware, software, power, and cooling needs of AI clusters. The term "clusters" references a grouping of multiple servers connected together, able to manage multiple hosts or tenants, and dedicated to a specific workload. These clusters can be relatively small, referring to tens of servers connected together, or large with thousands of servers interconnected together. In this section, we provide a brief overview of the major hardware components that must come together to address an expanding range of AI use-cases.

Exhibit 12
From Chips to Systems
Assembling an AI Server



Source: Dell Technologies

Compute

In response to the increasing demand for computational power in an AI-centric data center, servers will employ high-performance parallel processing chips (either GPUs or specialized ASICs like Google's TPUs), which are commonly referred to as AI accelerators because they offload more repeatable tasks from the CPU to a specialized chip. The types of chips used and the arrangement of them in the server largely depends on the use-case, which in AI can generally be divided into training and inference use-cases—AI training is the process of teaching a model to recognize patterns by adjusting its parameters based on large datasets, while AI inference is the application of the trained model to make predictions or decisions on new data.

Training servers, which make of the bulk of AI chip sales today, require a multitude of GPU servers connected through high-bandwidth, low-latency fabrics. According to research from McKinsey, the prevailing high-performance GenAI server architecture uses two CPUs and eight GPUs for compute. Nonetheless, other form-factors are becoming more common. For example, Nvidia's Grace-Blackwell solution has one CPU for every two GPUs, while Meta leverages boards with a ratio of

one CPU to one GPU to store massive embedding tables and perform pre-/post-processing on the CPU. Over the longer term, we expect that most training workloads will be executed using this type of CPU+GPU combination.

Inference servers are somewhat different. While the same GPUs and AI infrastructure used in training can also be used for inference, inference workloads typically require smaller scale because the amount of data input at prompt for inference is magnitudes smaller than in training.

As GenAI adoption by consumers and businesses increases, the workload is expected to shift predominantly toward inference tasks as more applications are built leveraging trained models and as consumers and workers engage more with AI-based services (e.g., Apple Intelligence, Microsoft Copilot, OpenAI's ChatGPT, Google's AI Search). Every interaction with these solutions will drive inferencing against one model, or more likely many different models (in the same way that accessing an app today depends on hundreds of APIs/integrations). This shift will favor specialized hardware over commodity components due to the advantages in cost, energy efficiency, and optimized performance for specific tasks.

Specialized chips designed for inference are becoming increasingly important. For instance, Google's Tensor Processing Units (TPUs) are highly optimized for inference tasks, providing high performance and energy efficiency for machine learning models. Similarly, Nvidia's Tensor Cores, integrated into its GPUs, accelerate deep learning inference with a focus on efficiency and speed. Meanwhile, Field-Programmable Gate Arrays (FPGAs), such as Intel's Stratix series, offer customizable hardware acceleration tailored to specific inference workloads, providing both flexibility and performance.

Another notable example of a specialized chip designed for AI processing is Groq's LPU (Language Processing Unit). The LPU is designed specifically for AI inference, delivering high throughput and efficiency. Its architecture supports scalable, high-performance processing for large-scale inference tasks. It addresses the key memory bottlenecks of memory access by leveraging more expensive but much-faster static random access memory ([SRAM] versus DRAM that is typical in most other GPU architectures). While this can provide fast inference, it creates its own cost and power limitations because SRAM uses much more power, which leads to its own challenges for adoption at scale—e.g., the memory required to hold LLMs can be an order of magnitude higher than the SRAM that can be fit onto a chip.

Cores. Core counts in CPUs and GPUs are crucial because they directly impact the processing power and efficiency of a computing system. In CPUs, multiple cores allow for simultaneous execution of multiple threads, enhancing multitasking and overall performance. For CPUs, this means faster execution of tasks that are spread across multiple cores, such as data processing, simulations, and software development. Modern AI servers are increasingly using CPUs with higher core counts to handle complex data processing tasks. For instance, server-grade CPUs like the AMD EPYC series and Intel Xeon Scalable processors offer core counts ranging from 16 to 64 cores per chip. CPUs are evolving with more cores to support data-intensive tasks. For example, AMD's EPYC 9004 series offers up to 96 cores, catering to HPC and AI workloads.

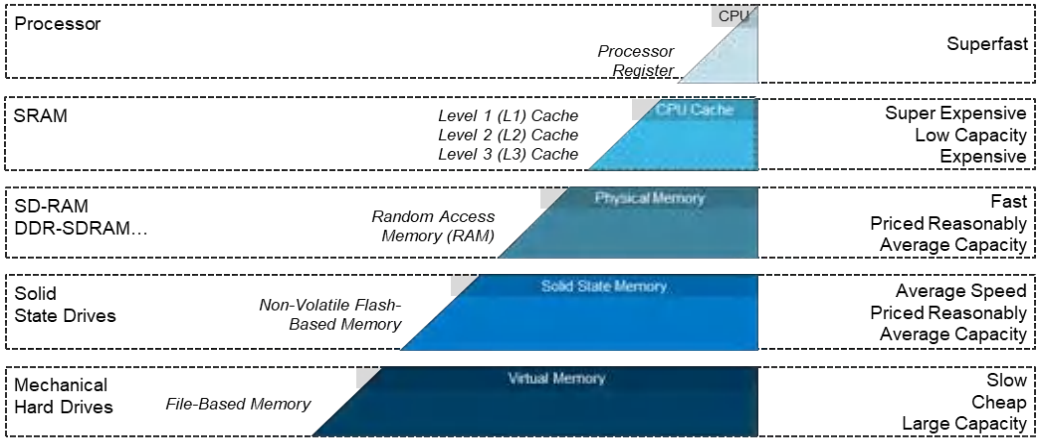
For GPUs, core counts are even more crucial because of their role in accelerating parallel computations. Modern GPUs like Nvidia's A100 and H100 Tensor Core GPUs feature thousands of CUDA cores (e.g., A100 has 6,912 CUDA cores), which are essential for training large-scale AI models and performing high-speed inference.

Unlike traditional processors that rely on multiple cores to handle parallel tasks, Groq's LPUs use a single, highly parallelized core architecture. By focusing on a single core with an architecture tailored for parallelism, this design eliminates some of the overhead associated with managing multiple cores and intercore communication, offering potential advantages in certain AI use-cases (e.g., text-generation and NLP).

Memory/Storage

At a high level, memory is classified into a hierarchy, as shown in exhibit 13, that goes from cheapest and slowest storage capabilities to the fastest but more expensive capabilities. Generally, the fastest memory is located inside the CPU itself (registers), followed by SRAM, which is typically used in caches. SRAM memory leverages a much higher number of transistors to achieve the high speeds required, making it a denser and more power-hungry type of memory than DRAM (dynamic random access memory). Today, GenAI servers use two types of DRAM: 1) high-bandwidth memory (HBM), attached to the GPU or AI accelerators, and 2) DDR (double data rate) RAM, attached to the CPU. HBM has higher bandwidth but requires more silicon for the same amount of data, while DDR RAM is a variant of DRAM memory that provides high-speed volatile memory, facilitating rapid data access.

Exhibit 13
From Chips to Systems
The Memory Hierarchy



Source: William Blair Equity Research

As transformer models grow larger, GenAI servers have been expanding memory capacity. However, the growth in memory capacity is not straightforward, posing challenges to hardware and software design. First, the industry faces a memory wall problem, in which memory capacity and bandwidth are the bottleneck for system-level compute performance. How the industry will tackle the memory wall problem is an open question.

One answer is that future algorithms may require less memory per inference run, slowing total memory demand growth. Second, custom-built AI accelerators can be lighter in memory compared to general CPU+GPU solutions (like Nvidia's) because they are designed for specific use-cases like NLP. Third, SRAM is being tested in various chips to increase the near-compute memory, but its high cost and power consumption limits wide adoption.

NAND. NAND memory is used for data storage (e.g., for the operating system, user data, and input and output). In 2030, NAND demand will likely be driven by dedicated data servers for video and multimodel data. This data will require substantial storage (for example, for training on high-resolution video sequences and retrieving data during inference). McKinsey expects the total NAND demand to be 2 million to 8 million wafers, corresponding to one to five fabs.

Network

GenAI requires high-bandwidth and low-latency connectivity between the servers and between the various components of the servers. A larger amount of network interfaces and switches are required to create all the connections. At its core, AI is a distributed computing problem. AI clusters, which connect thousands of GPUs together, require multiple layers of network connectivity, including a front-end network that connects out to the internet or other data centers; a middle layer that interconnects GPUs together and with NICs (network interface cards), CPUs, and other components (referred to as scale-up network); and a back-end layer that connects the GPU servers together into clusters (referred to as scale-out network).

The AI opportunity for networking is growing as AI clusters become larger, particularly as it relates to back-end networks. As the size of AI clusters scale, the number of GPU servers connected continues to increase, driving greater need for fast, low-latency network fabrics. Broadcom's own data has shown that the number of GPUs per cluster has grown from on average 4,000 in 2022, to 10,000 in 2023, to 30,000 in 2024. By 2027, industry leaders like Broadcom expect we could start seeing clusters with more than 1 million GPUs.

The Ethernet-InfiniBand debate. Over the last two years, the debate around which networking protocol to use in AI clusters has raged between the HPC incumbent InfiniBand and the long-standing standard for traditional data center networks, Ethernet. Nvidia's acquisition of Mellanox (the key provider of InfiniBand technology) has pushed strong growth in InfiniBand over the last few years, as Nvidia promotes, bundles, and pushes its InfiniBand technology as optimal for AI clusters. Nonetheless, we have seen consensus grow that as the Ethernet technology matures and is adjusted to handle some of the challenges of AI networking, it will gain greater and greater share in the AI networking space.

InfiniBand is a high-performance network architecture designed primarily for data centers and HPC environments. It provides very low latency and high throughput, often in the range of microseconds for latency and several hundred gigabits per second (Gbps) for bandwidth. InfiniBand uses a point-to-point switched fabric topology, which helps minimize congestion and optimize data transfer speeds. In addition, it supports remote direct memory access (RDMA), which allows data to be transferred directly between memory locations on different computers without involving the CPU, leading to faster data processing and lower latency. This makes InfiniBand particularly well-suited for applications requiring fast, large-scale data transfers, such as scientific simulations, financial modeling, large-scale data analytics, and of course AI.

Ethernet, on the other hand, is the most widely used networking technology for both local area networks (LANs) and wide area networks (WANs). It is known for its robustness, flexibility, and ease of use. Ethernet has evolved significantly over the years, with current standards like 100/400 GE and soon 800 GE. Despite its higher latency compared to InfiniBand, Ethernet's ubiquity, cost-effectiveness, and extensive support make it suitable for a broad range of applications, including enterprise networking, cloud computing, and internet services. Ethernet also supports various quality-of-service features, which help manage network traffic effectively.

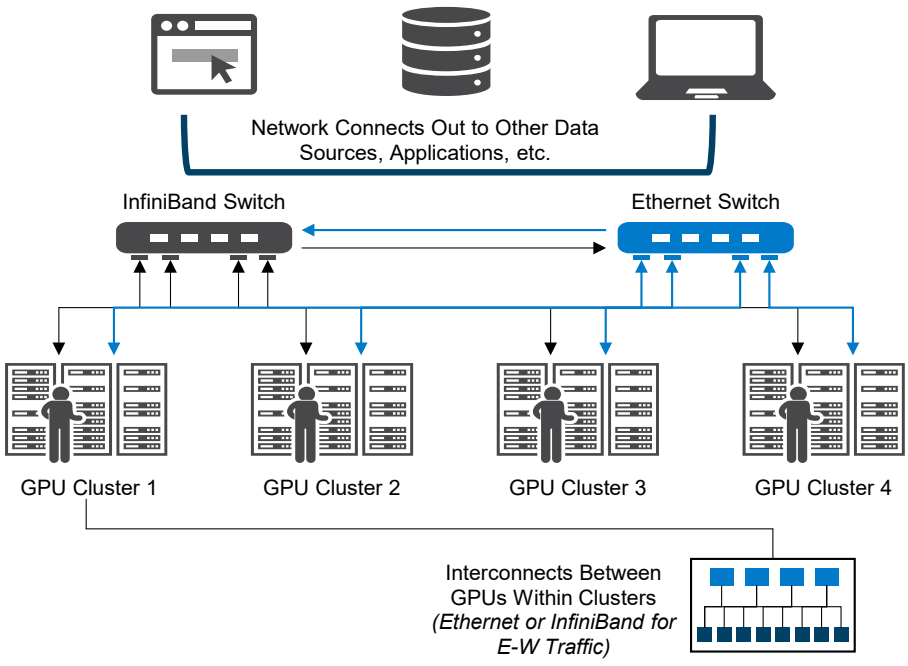
While InfiniBand has benefited from being at the right place at the right time, addressing the high-speed networking needs of GenAI clusters, it suffers from its more restrictive functionality (e.g., expensive, rigid, difficult to scale, sole-sourced from Nvidia). In contracts, Ethernet benefits from

1) a deep multivendor ecosystem for NICs, switches, optics, and tooling; 2) well-understood operations, debugging, and monitoring; 3) robust scalability; 4) a proven record of success across all prior exiting workloads/use-cases; and 5) perhaps most importantly its lower cost to purchase and deploy.

Today InfiniBand has a dominant share in AI clusters (a nascent market), largely because it is bundled with Nvidia GPUs. Nonetheless, Ethernet is expected to catch up and become the dominant approach for AI networking as the industry establishes the right open standards—the Ultra Ethernet Consortium (UEC) expects to release updated modern standards for RoCE (RDMA over Converged Ethernet) in early 2025, with vendor products expected soon after. This next generation of Ethernet standards will leverage dynamic multipath flow, improved packet spraying, and a more scalable control plane to enable the massive bandwidth needs of these GPU clusters. Even Nvidia, which has a built a large InfiniBand business, expects its Ethernet technologies to see immense growth in 2025, becoming a multibillion-dollar business in its own right.

The early nature of the Ethernet network buildouts for AI supports our view that we are not reaching the end of an AI investment cycle, but rather remain several years away from supply potentially outpacing demand. According to commentary from Cisco, Arista, and Nvidia, 2025 will be characterized by the buildout of a first wave of production-grade Ethernet-based AI clusters. Indications are that in 2026/2027, a wave of even larger 100,000-plus GPU clusters will be built. This supports our view that companies like Nvidia and Broadcom are still in the early innings of monetizing this AI investment cycle.

Exhibit 14
From Chips to Systems
Network Topology for Generative AI Clusters



Source: William Blair Equity Research

Power/Energy Consumption
Our William Blair colleague recently published a report examining the impact of AI on energy consumption. For more detailed analysis on the question of energy/power, please see the following report: The Power Behind Artificial Intelligence.

Alongside the growing demand in AI infrastructure has been an increased demand in power/energy. This demand for more energy manifests across two axes: 1) increased demand in overall energy, which has prompted concerns about the ability to generate enough energy with today's production levels to power these energy-hungry net new workloads, and 2) need of higher energy density in data centers, which is requiring a rewiring and redesign of data centers to pack more wafers per square foot to feed the high-density servers.

The rapid growth of AI has sparked significant concerns regarding energy availability and grid stability. As AI applications, particularly in large language models and data processing, continue to expand, their energy demands have escalated dramatically. U.S. data centers alone are projected to see their electricity consumption triple by 2030, potentially reaching up to 390 terawatt-hours, which would account for about 7.5% of the nation's total electricity demand. This surge is placing immense strain on the power grid, with some regions already experiencing delays in connecting new facilities due to insufficient power capacity.

The energy-intensive nature of AI, driven by specialized chips and massive computational needs, has led to concerns that current grid infrastructure may not be able to keep pace with this rapid increase in demand. This is compounded by the slow progress in expanding and upgrading the grid, with some key components like transformers facing significant backlogs. In some cases, utilities have had to delay the closure of coal-fired power plants to meet the sudden spike in demand, which poses a challenge to sustainability goals.

Furthermore, there is a growing recognition that simply improving energy efficiency in AI operations may not be sufficient to curb overall consumption. New strategies, such as the development of edge data centers and the implementation of flexible computing loads that adjust based on grid availability, are being explored to mitigate these challenges. However, substantial investments in grid infrastructure and innovative solutions will be crucial to avoid potential blackouts and ensure that AI can continue to grow without overwhelming the energy supply.

Meanwhile, rack densities, defined as the power consumed by a cabinet of servers, illustrate the significant contrast between traditional and AI data centers. Conventional data centers typically operate at rack power densities ranging from 5 to 15 kW. In contrast, AI training workloads can demand much higher power, reaching up to 150 kW (e.g., Nvidia's latest GB200 NVL72 rack solution requires 120 kW per rack). Experts predict these figures will continue to rise, potentially reaching power densities of 250 kW or even 300 kW in the coming years.

Moreover, as rack power density increases, data centers will transition from air-based cooling to liquid cooling solutions. This shift will necessitate new server and rack designs capable of supporting increased weights, incorporating technologies like direct-to-chip liquid cooling and full-immersion cooling.

Cooling
The evolution of data center cooling is intertwined with the development of computing technology itself. In the early days of computing, mainframe computers and early data centers were primarily cooled using basic ventilation and air conditioning systems, which sufficed for the relatively low-density computational loads of the time.

As computing power increased, data centers began to house more servers and equipment, so the need for more sophisticated cooling solutions became evident. In the 1980s and 1990s, air-based cooling systems were the standard. Raised floor designs, which allowed cold air to be distributed through perforated tiles invented upward to cool equipment, became prevalent. These systems used computer room air conditioners (CRACs) and later computer room air handlers (CRAHs), which were more efficient.

The early 2000s saw the advent of more advanced cooling technologies. Liquid cooling, once deemed impractical for large-scale use, began to make a comeback due to its higher efficiency and heat removal. Innovations like hot aisle/cold aisle containment improved air flow management, significantly enhancing cooling efficiency by separating hot and cold air streams.

As of 2023, the global data center cooling market was valued at about $13 billion and is projected to grow at a compound annual rate of roughly 13% through 2025, according to data from MarketandMarkets. Modern data center cooling technologies have evolved to meet the demands of increased computing power and energy efficiency. Some of the prominent cooling solutions include:

1. ***Air-based cooling:*** still widely used, especially in smaller data centers, but with significant improvements in efficiency. Innovations include more sophisticated air flow management and advanced CRAC/CRAH units.

2. ***Liquid cooling:*** gaining traction for its superior efficiency and heat removal. Two main types are in use:

    a. *Direct-to-chip cooling:* Liquid is circulated directly to the heat-generating components.

    b. *Immersion cooling:* Servers are submerged in a thermally conductive but electrically insulating liquid.

3. ***Free cooling:*** uses external environment conditions, such as cool air or water, to dissipate heat, reducing reliance on mechanical refrigeration. Popular in regions with cooler climates.

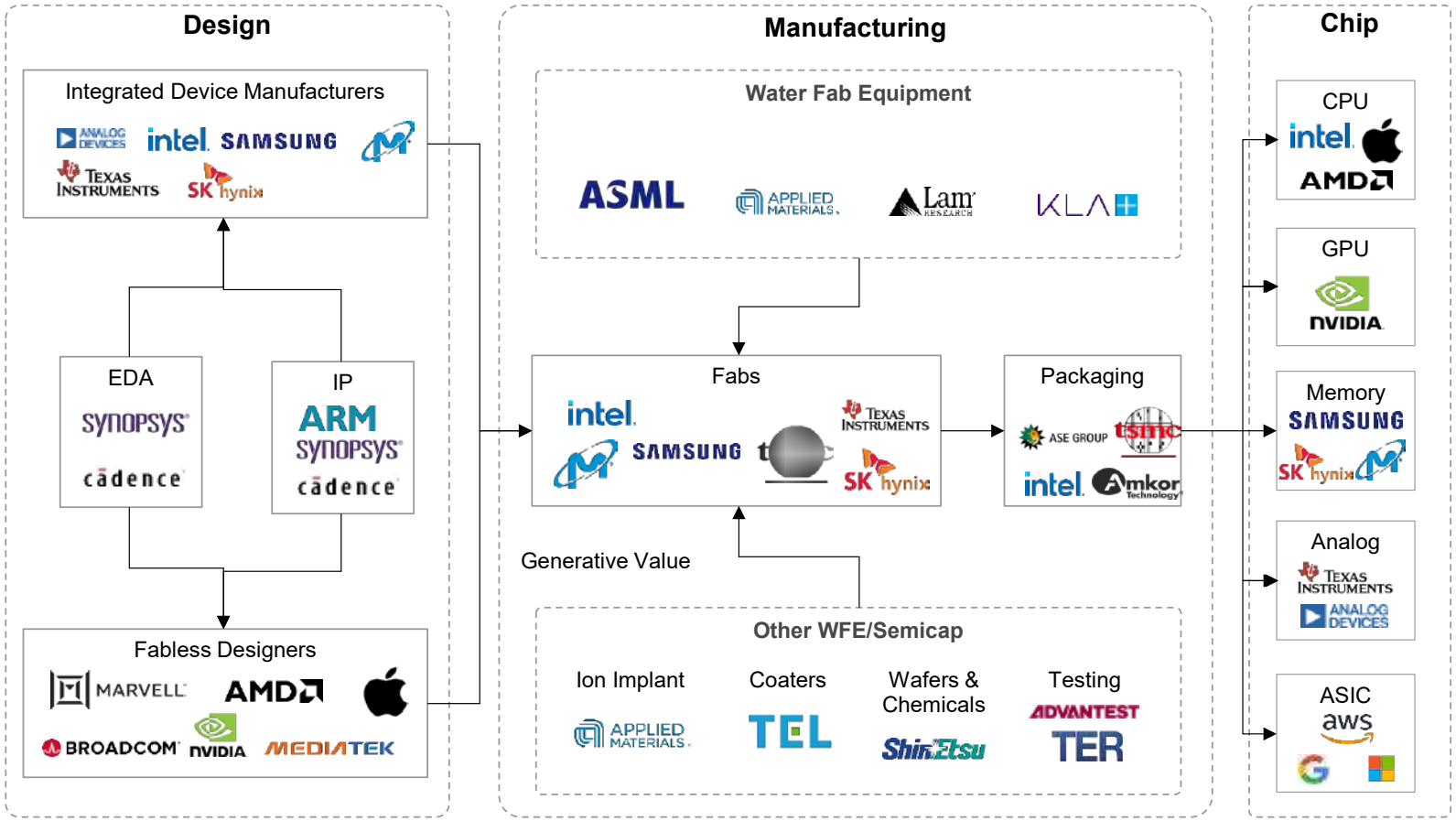
Software

What we have seen in the earliest days of AI is that the software engineers really understand the trade-off between the performance of the system and the precision of the system. If they are somehow limited in bandwidth, energy, or otherwise, they turn it into a software problem. If they do not have enough bandwidth, they can reduce the precision of the numbers, and they train specifically for reduced precision or for sparsity. In the AI field, there is a holistic view of integration between the software side and hardware side. In the same way that programmers have been forced to become more architecturally aware over the last 20 years, considering cache size and processor architecture, in the future programmers will have to be more cognizant about things like power limitations in the system and use tools and APIs that let them trade off power for performance.

Semiconductor Value Chain

While the semiconductor value chain is complex, particularly as the types of chips have become more complex, the basic components are largely the same as they were in the early days. Individual semiconductor chips are designed with the aid of advanced software. Chemicals, gases, and other materials are combined in an intricate series of operations that use complex manufacturing to product wafers containing a large number of dies—each die (assuming it does not suffer from fatal defects) forms the basis for a semiconductor chip. Individual die are cut from fabricated wafers, tested for defects, and assembled into complex packages that combine wire contacts with insulating material to form the finished chip.

Exhibit 15
From Chips to Systems
Semiconductor Value Chain



* Not an exhaustive list of companies/segments

Source: William Blair Equity Research

GPUs, Software, and Structurally Higher Margins for Semi Leaders

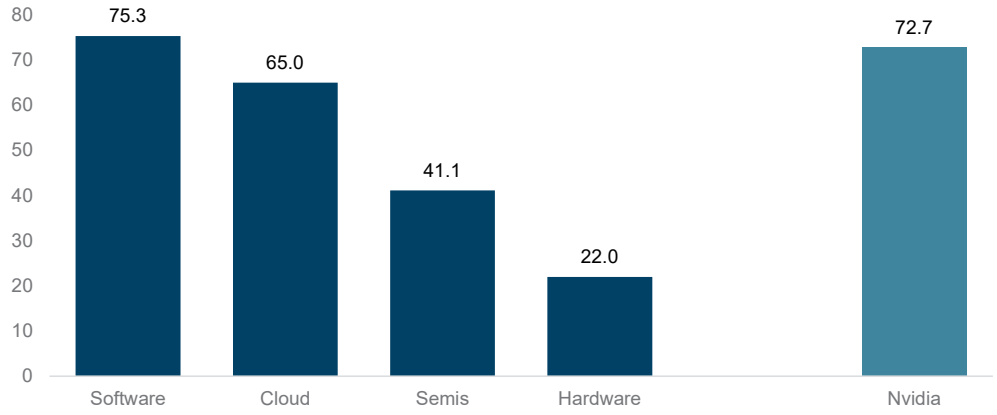
The market has generally rewarded semiconductor companies for specializing in distinct parts of the value chain. This way of working transformed an industry that was initially vertically integrated (semiconductor design, semiconductor manufacturing, and system integration) into an ecosystem focused on specific areas of design, manufacturing, and/or systems.

Today, we are seeing the pendulum swing back toward verticalization as vendors see greater ability to improve chip performance by controlling and optimizing all the data center elements around it. This vertical integration trend is distinctly different from the vertical integration that occurred at the inception of the semiconductor and IDM industry more than 50 years ago. Now, vertical integration is all about optimizing systems for use-cases and capturing more value from the end-customer.

As chip companies move up the tech stack to optimize the performance of their systems and capture more of the value from their end-customers, they have been able to drive better gross margins for their solutions. Another way to think about it is that as more software and systems design IP is integrated into chip solutions, semi vendors are able to drive more differentiation and value for their customers, which allows them to drive better margins. In addition, as companies like Nvidia deliver more of their IP outside traditional chip design, they are able to attract the higher gross margins seen in the software sector. Generally, incremental software improvement costs much less to implement than the incremental chip improvement—i.e., it is much easier to update software then to tape out a new chip.

Nvidia is a prime example of this, with its gross margins having increased from the low- to mid-60% range from 2017 to 2022 to the mid-70% range over the last year. While some of this margin improvement can be associated with Nvidia's pricing power in the market (because of limited supply and still limited competition from other vendors), we expect these gross margins should remain stickier even as core GPU technology becomes increasingly commoditized.

Exhibit 16
From Chips to Systems
Comparing Gross Margins Across the IT Industry



Note: Software margins taken as average of MDB, CRM, ORCL, TEAM, NET, NTNX, DDOG, and BOX gross margins. Semi margins taken as average of INTC, AMD, QCOM, MU, NXPI, TXN, ADI, and Samsung. Hardware margins taken as DELL, HPE, SMCI, and Lenovo. Cloud margins based on William Blair estimates for AWS, Azure, and GCP gross margins and includes IaaS and higher-margin PaaS.

Sources: Nvidia and William Blair Equity Research

Moore's Law Squared?
Moore's law squared refers to the accelerated pace of improvement in AI model performance, driven by advancements in both hardware and software. Traditionally, Moore's law observed that the number of transistors on a microchip doubles approximately every two years, leading to exponential increases in computing power and decreases in cost. Moore's law squared builds on this concept, suggesting that the performance of AI systems improves even more rapidly due to significant algorithmic innovations and software enhancements alongside hardware improvements.

The term captures how specialized AI hardware, such as TPUs and GPUs, boosts computational capabilities beyond what traditional CPU improvements alone would achieve. Concurrently, algorithmic advancements, like the development of transformer architectures and techniques such as model pruning and quantization, contribute to more efficient AI training and inference. Software optimizations in AI frameworks and libraries further enhance performance by maximizing hardware utilization. This synergy leads to exponential growth in AI capabilities that outpaces the original predictions of Moore's law.

Conclusion

Just as every technology wave has had a profound impact on computing (from mainframes to PCs to smartphones, and then cloud), we expect AI will follow a similar trajectory. AI is refocusing the entire semiconductor ecosystem on system-level efficiencies for data centers as companies embark on an arms race to become leaders in AI. As a result, discrete chip manufacturers have had to move up the stack and embrace verticalization to remain competitive. AMD's August 2024 announcement that it was acquiring server building company ZT Systems for nearly $5 billion is one more strong point of evidence that semi companies need to move up the stack or they will fall behind.

We remain steadfastly convinced in the long-term impacts of this shift to AI and the positive benefits it will have on companies that have increasingly adopted a systems approach to their solutions; still, the rapid pace of spending on GPUs and AI infrastructure may create some near-term volatility in results. In this AI arms race, supply and demand may not perfectly match up, which creates heightened risk for oversupply and overcapacity at some point. The reverse is potentially also true, but until we gain a more definitive view on what the end-demand (enterprise/consumer) looks like, and in particular what the ROI is of the hundreds of billions of dollars of spend on AI infrastructure over the last year, it could be challenging for vendors to maintain the pace of outperformance that investors have become accustomed to over the last year.

Nonetheless, commentary from leaders at the largest technology companies keep us bullish on this investment cycle. As Alphabet's CEO noted in his July 2024 earnings call, the risk of underinvesting is dramatically greater than the risk of overinvesting. In September 2024, Oracle CTO Larry Ellison espoused his own belief that AI spending will continue at pace as large technology companies vie to create the best foundational models and as virtually all other companies embed AI into their applications and workflows. The largest companies remain committed to this AI investment cycle, seeing it as existential to remain competitive in this next technology wave.

The prices of the common stock of other public companies mentioned in this report follow:

Alphabet, Inc. (Outperform)	$158.06
AMD	$152.08
Arista Networks (Outperform)	$359.16
Amazon.com (Outperform)	$184.89
Broadcom (Outperform)	$164.02
Cadence Designs Systems	$273.13
Cisco Systems (Market Perform)	$51.03
Intel	$20.91
Marvell Technology (Not Rated)	$73.40
Micron	$87.18
Microsoft (Outperform)	$431.34
Meta (Outperform)	$533.28
Monolithic Power (Not Rated)	$885.23
Qualcomm	$166.61
SuperMicro	$449.10
Synopsys	$498.57

IMPORTANT DISCLOSURES

William Blair or an affiliate is a market maker in the security of Arm Holdings plc, Broadcom Inc., NVIDIA Corporation, Arista Networks, Inc., Cisco Systems, Inc., Oracle Corporation and Microsoft Corporation.

William Blair or an affiliate expects to receive or intends to seek compensation for investment banking services from Arm Holdings plc, Broadcom Inc., NVIDIA Corporation, Arista Networks, Inc., Cisco Systems, Inc., Oracle Corporation and Microsoft Corporation or an affiliate within the next three months.

Officers and employees of William Blair or its affiliates (other than research analysts) may have a financial interest in the securities of Arm Holdings plc, Broadcom Inc., NVIDIA Corporation, Arista Networks, Inc., Cisco Systems, Inc., Oracle Corporation and Microsoft Corporation.

This report is available in electronic form to registered users via R*Docs™ at https://williamblairlibrary.bluematrix.com or www.williamblair.com.

Please contact us at +1 800 621 0687 or consult https://www.williamblair.com/equity-research/coverage for all disclosures.

Jason Ader attests that 1) all of the views expressed in this research report accurately reflect his/her personal views about any and all of the securities and companies covered by this report, and 2) no part of his/her compensation was, is, or will be related, directly or indirectly, to the specific recommendations or views expressed by him/her in this report. We seek to update our research as appropriate. Other than certain periodical industry reports, the majority of reports are published at irregular intervals as deemed appropriate by the research analyst.

DOW JONES: 41606.20
S&P 500: 5634.58
NASDAQ: 17628.10

Additional information is available upon request.

Current Rating Distribution (as of September 18, 2024):

Coverage Universe	Percent	Inv. Banking Relationships *	Percent
Outperform (Buy)	71	Outperform (Buy)	8
Market Perform (Hold)	28	Market Perform (Hold)	1
Underperform (Sell)	1	Underperform (Sell)	0

*Percentage of companies in each rating category that are investment banking clients, defined as companies for which William Blair has received compensation for investment banking services within the past 12 months.

The compensation of the research analyst is based on a variety of factors, including performance of his or her stock recommendations; contributions to all of the firm's departments, including asset management, corporate finance, institutional sales, and retail brokerage; firm profitability; and competitive factors.

OTHER IMPORTANT DISCLOSURES

Stock ratings and valuation methodologies: William Blair & Company, L.L.C. uses a three-point system to rate stocks. Individual ratings reflect the expected performance of the stock relative to the broader market (generally the S&P 500, unless otherwise indicated) over the next 12 months. The assessment of expected performance is a function of near-, intermediate-, and long-term company fundamentals, industry outlook, confidence in earnings estimates, valuation (and our valuation methodology), and other factors. Outperform (O) - stock expected to outperform the broader market over the next 12 months; Market Perform (M) - stock expected to perform approximately in line with the broader market over the next 12 months; Underperform (U) - stock expected to underperform the broader market over the next 12 months; not rated (NR) - the stock is not currently rated. The valuation methodologies include (but are not limited to) price-to-earnings multiple (P/E), relative P/E (compared with the relevant market), P/E-to-growth-rate (PEG) ratio, market capitalization/revenue multiple, enterprise value/EBITDA ratio, discounted cash flow, and others. Stock ratings and valuation methodologies should not be used or relied upon as investment advice. Past performance is not necessarily a guide to future performance.

The ratings and valuation methodologies reflect the opinion of the individual analyst and are subject to change at any time.

Our salespeople, traders, and other professionals may provide oral or written market commentary, short-term trade ideas, or trading strategies-to our clients, prospective clients, and our trading desks-that are contrary to opinions expressed in this research report. Certain outstanding research reports may contain discussions or investment opinions relating to securities, financial instruments and/or issuers that are no longer current. Always refer to the most recent report on a company or issuer. Our asset management and trading desks may make investment decisions that are inconsistent with recommendations or views expressed in this report. We will from time to time have long or short positions in, act as principal in, and buy or sell the securities referred to in this report. Our research is disseminated primarily electronically, and in some instances in printed form. Research is simultaneously available to all clients. This research report is for our clients only. No part of this material may be copied or duplicated in any form by any means or redistributed without the prior written consent of William Blair & Company, L.L.C.

This is not in any sense an offer or solicitation for the purchase or sale of a security or financial instrument. The factual statements herein have been taken from sources we believe to be reliable, but such statements are made without any representation as to accuracy or completeness or otherwise, except with respect to any disclosures relative to William Blair or its research analysts. Opinions expressed are our own unless otherwise stated and are subject to change without notice. Prices shown are approximate. This report or any portion hereof may not be copied, reprinted, sold, or redistributed or disclosed by the recipient to any third party, by content scraping or extraction, automated processing, or any other form or means, without the prior written consent of William Blair. Any unauthorized use is prohibited.

If the recipient received this research report pursuant to terms of service for, or a contract with William Blair for, the provision of research services for a separate fee, and in connection with the delivery of such research services we may be deemed to be acting as an investment adviser, then such investment adviser status relates, if at all, only to the recipient with whom we have contracted directly and does not extend beyond the delivery of this report (unless otherwise agreed specifically in writing). If such recipient uses these research services in connection with the sale or purchase of a security referred to herein, William Blair may act as principal for our own account or as riskless principal or agent for another party. William Blair is and continues to act solely as a broker-dealer in connection with the execution of any transactions, including transactions in any securities referred to herein.

For important disclosures, please visit our website at williamblair.com.

This material is distributed in the United Kingdom and the European Economic Area (EEA) by William Blair International, Ltd., authorised and regulated by the Financial Conduct Authority (FCA). William Blair International, Limited is a limited liability company registered in England and Wales with company number 03619027. This material is only directed and issued to persons regarded as Professional investors or equivalent in their home jurisdiction, or persons falling within articles 19 (5), 38, 47, and 49 of the Financial Services and Markets Act of 2000 (Financial Promotion) Order 2005 (all such persons being referred to as "relevant persons"). This document must not be acted on or relied on by persons who are not "relevant persons."

Any statements in this report that are attributable to IDC Research, Inc. (“IDC”) represent William Blair’s interpretation of data, research opinion or viewpoints published as part of a syndicated subscription service by IDC and have not been reviewed by IDC. IDC’s research is current as of the date IDC published it, not the date that William Blair’s reports are published. Further, IDC’s research contains IDC’s opinion, not representations of fact, and are subject to change without notice.

William Blair & Company, L.L.C. licenses and applies the SASB Materiality Map® and SICSTM in our work.

William Blair

Equity Research Directory

John Kreger, Partner Director of Research +1 312 364 8612
Kyle Harris, CFA, Partner Operations Manager +1 312 364 8230

Scott Hansen Associate Director of Research +1 212 245 6526

CONSUMER

Sharon Zackfia, CFA, Partner +1 312 364 5386
Group Head–Consumer
Lifestyle and Leisure Brands, Restaurants, Automotive/E-commerce

Jon Andersen, CFA, Partner +1 312 364 8697
Consumer Products

Phillip Blee, CPA +1 312 801 7874
Home and Outdoor, Automotive Parts and Services, Discount and Convenience

Dylan Carden +1 312 801 7857
E-commerce, Specialty Retail

ECONOMICS

Richard de Chazal, CFA +44 20 7868 4489

ENERGY AND SUSTAINABILITY

Jed Dorsheimer +1 617 235 7555
Group Head–Energy and Sustainability
Generation, Efficiency, Storage

Tim Mulrooney, Partner +1 312 364 8123
Sustainability Services

FINANCIAL SERVICES AND TECHNOLOGY

Adam Klauber, CFA, Partner +1 312 364 8232
Group Head–Financial Services and Technology
Financial Analytic Service Providers, Insurance Brokers, Property & Casualty Insurance

Andrew W. Jeffrey, CFA +1 415 796 6896
Fintech

Cristopher Kennedy, CFA +1 312 364 8596
Fintech, Specialty Finance

Jeff Schmitt +1 312 364 8106
Wealthtech, Wealth Management, Capital Markets Technology

GLOBAL SERVICES

Tim Mulrooney, Partner +1 312 364 8123
Group Head–Global Services
Commercial and Residential Services

Andrew Nicholas, CPA +1 312 364 8689
Consulting, HR Technology, Information Services

Trevor Romeo, CFA +1 312 801 7854
Staffing

HEALTHCARE

Biotechnology

Matt Phipps, Ph.D., Partner +1 312 364 8602
Group Head–Biotechnology

Sami Corwin, Ph.D. +1 312 801 7783

Lachlan Hanbury-Brown +1 312 364 8125

Andy T. Hsieh, Ph.D., Partner +1 312 364 5051

Myles R. Minter, Ph.D. +1 617 235 7534

Sarah Schram, Ph.D. +1 312 364 5464

Healthcare Technology and Services

Ryan S. Daniels, CFA, Partner +1 312 364 8418
Group Head–Healthcare Technology and Services
Healthcare Technology, Healthcare Services

Margaret Kaczor Andrew, CFA, Partner +1 312 364 8608
Medical Technology

Brandon Vazquez, CFA +1 212 237 2776
Dental, Animal Health

Life Sciences

Matt Larew, Partner +1 312 801 7795
Life Science Tools, Bioprocessing, Healthcare Delivery

Andrew F. Brackmann, CFA +1 312 364 8776
Diagnostics

Max Smock, CFA +1 312 364 8336
Pharmaceutical Outsourcing and Services

INDUSTRIALS

Brian Drab, CFA, Partner +1 312 364 8280
Co-Group Head–Industrials
Advanced Manufacturing, Industrial Technology

Ryan Merkel, CFA , Partner +1 312 364 8603
Co-Group Head–Industrials
Building Products, Specialty Distribution

Louie DiPalma, CFA +1 312 364 5437
Aerospace and Defense, Smart Cities

Ross Sparenblek +1 312 364 8361
Diversified Industrials, Robotics, and Automation

TECHNOLOGY, MEDIA, AND COMMUNICATIONS

Jason Ader, CFA, Partner +1 617 235 7519
Co-Group Head–Technology, Media, and Communications
Infrastructure Software

Arjun Bhatia, Partner +1 312 364 5696
Co-Group Head–Technology, Media, and Communications
Software as a Service

Dylan Becker, CFA +1 312 364 8938
Software, Software as a Service

Louie DiPalma, CFA +1 312 364 5437
Government Technology

Jonathan Ho, Partner +1 312 364 8276
Cybersecurity, Security Technology

Maggie Nolan, CPA, Partner +1 312 364 5090
IT Services

Jake Roberge +1 312 364 8056
Software, Software as a Service

Ralph Schackart III, CFA, Partner +1 312 364 8753
Internet and Digital Media

Stephen Sheldon, CFA, CPA, Partner +1 312 364 5167
Vertical Technology – Real Estate, Education, Restaurant/Hospitality

EDITORIAL AND SUPERVISORY ANALYSTS

Steve Goldsmith, Head Editor and SA +1 312 364 8540

Audrey Majors, Editor and SA +1 312 364 8992

Beth Pekol Porto, Editor and SA +1 312 364 8924

Lisa Zurcher, Editor and SA +44 20 7868 4549