Deloitte.

SIA SEMICONDUCTOR INDUSTRY ASSOCIATION



Powering AI: The Semiconductor Ecosystem at the Foundation of Data Centers

June 2026

Table of contents

Executive Summary 3

I. The many chips at the heart of AI 4

Chip innovations driving AI advancements 4

Semiconductor AI chip taxonomy 7

II. Breaking down the AI hardware stack: Chips in an AI data server 9

Compute tray 11

Compute board 12

AI Accelerators 12

High-bandwidth memory (HBM) 13

Central processing unit 15

System memory/random access memory 15

Non-volatile memory express storage 17

Data processing unit 17

Network interface card 18

Power distribution unit 19

Accelerator interconnect tray 21

Switch ASICs 22

Power distribution block 22

Active copper cable chips 22

Backplane connectors 22

Power tray 23

Power supply management module (PSMM) 24

Power Supply Unit (PSU) 24

Network and intelligent platform management interface tray 25

Baseboard management card (BMC) 25

Field replaceable unit 26

Out-of-band (OOB) switch 26

In-band (IB) switch 27

Trusted platform module (TPM) 27

Coolant distribution unit tray 29

III. How chips work together to execute an AI training workflow 30

IV. The AI growth curve: Market view 31

V. The big spend: Semiconductor content value analysis 33

Value concentration vs. cost dynamics 36

VI. From design to data center: Understanding the global supply chain of AI data centers 37

VII. Emerging Horizons in AI Infrastructure 38

VIII. Conclusion 40

Executive Summary

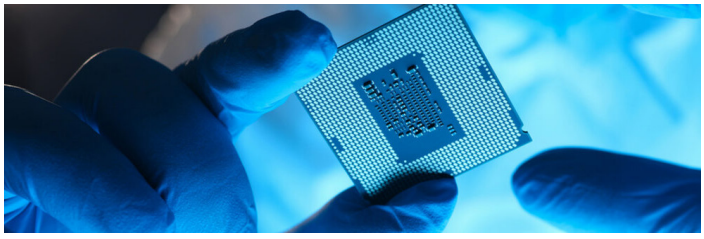
Semiconductors are the foundation of artificial intelligence (AI), a technology that is transforming our economy and society, making entire industries more productive and innovative, and driving major scientific breakthroughs.

Today's AI systems are built on decades of innovation across the entire semiconductor ecosystem. As chip technologies continue to advance, AI will become more capable, energy-efficient, and cost effective. Stronger AI, in turn, will help improve chip design, optimize semiconductor manufacturing, and drive more demand for the broad range of chips that enable AI.

Key Takeaways:

1. **Semiconductors are the fundamental enabling technology of AI.** Chips provide the base hardware layer underpinning modern AI systems and comprise a significant portion of the overall value in a modern AI server:
    - A single AI server rack contains over 4,500 packaged chips, comprised of approximately 20,000 individual dies – i.e., unique integrated circuits.
    - Semiconductors account for more than 95% of a leading AI server rack's content value and more than 50% of the total capital expenditures required for building and operating an AI data center.
2. **AI requires the full range of semiconductor technologies.** To run complex AI training and inference workloads, today's AI data centers need huge amounts of compute, storage and memory bandwidth, power distribution, and networking capabilities – all provided by the full stack of chip technologies. Each of these chip technologies is essential to driving America's AI buildout, while critical dependencies in any of these areas would risk hampering this buildout. Chips in AI data centers include:
    - **Advanced logic chips**, such as AI accelerators, application-specific integrated circuits (ASICs), field-programmable gate arrays (FPGAs), central processing units (CPUs), data processing units (DPUs), and networking chips.
    - **Memory**, such as high-bandwidth memory (HBM), both dynamic and static random-access memory (DRAM and SRAM), and non-volatile flash memory (NAND).
    - **Analog and foundational chips**, such as power chips, transceivers, controllers, and sensors.
3. **AI is a major demand driver for chips across the entire semiconductor industry.** In a positive feedback loop, advances in AI drive demand for improved semiconductor performance and efficiency, while improvements in semiconductor technology enable more powerful and advanced AI systems.
    - To meet global demand for new AI applications, government and industry will invest over **$4 trillion** in new data center infrastructure through 2028, of which up to **$2.8 trillion** will be spent on semiconductors.
    - Annual revenue for semiconductors deployed in AI data centers could reach **over $1.2 trillion** by 2028, a nearly tenfold increase over four years.
    - The AI data center market is experiencing unprecedented growth, projected at an 88.8% Compound Annual Growth Rate (CAGR) from 2022 to 2028. While initial momentum was driven by the rapid adoption of Generative AI, sustained demand remains robust, with a projected 56.3% CAGR from 2025 through 2028.

The entire semiconductor supply chain enables the buildout of AI infrastructure. Without semiconductors, there is no AI. To lead in this transformational technology, government and industry must work together to advance policies to accelerate growth and innovation across the full spectrum of chip technologies and work closely with global partners to build strong and resilient supply chains.



I. The many chips at the heart of AI

Artificial intelligence (AI) has experienced explosive growth in recent years, drawing significant attention to those that train and deploy AI models.[1] A wide variety of semiconductors serve as the backbone and enabling technology of the AI hardware stack, with advancements in chip technology driving increases in processing power, computational efficiency, and overall performance across AI applications. Semiconductors underpin the AI systems embedded in everyday digital experiences.

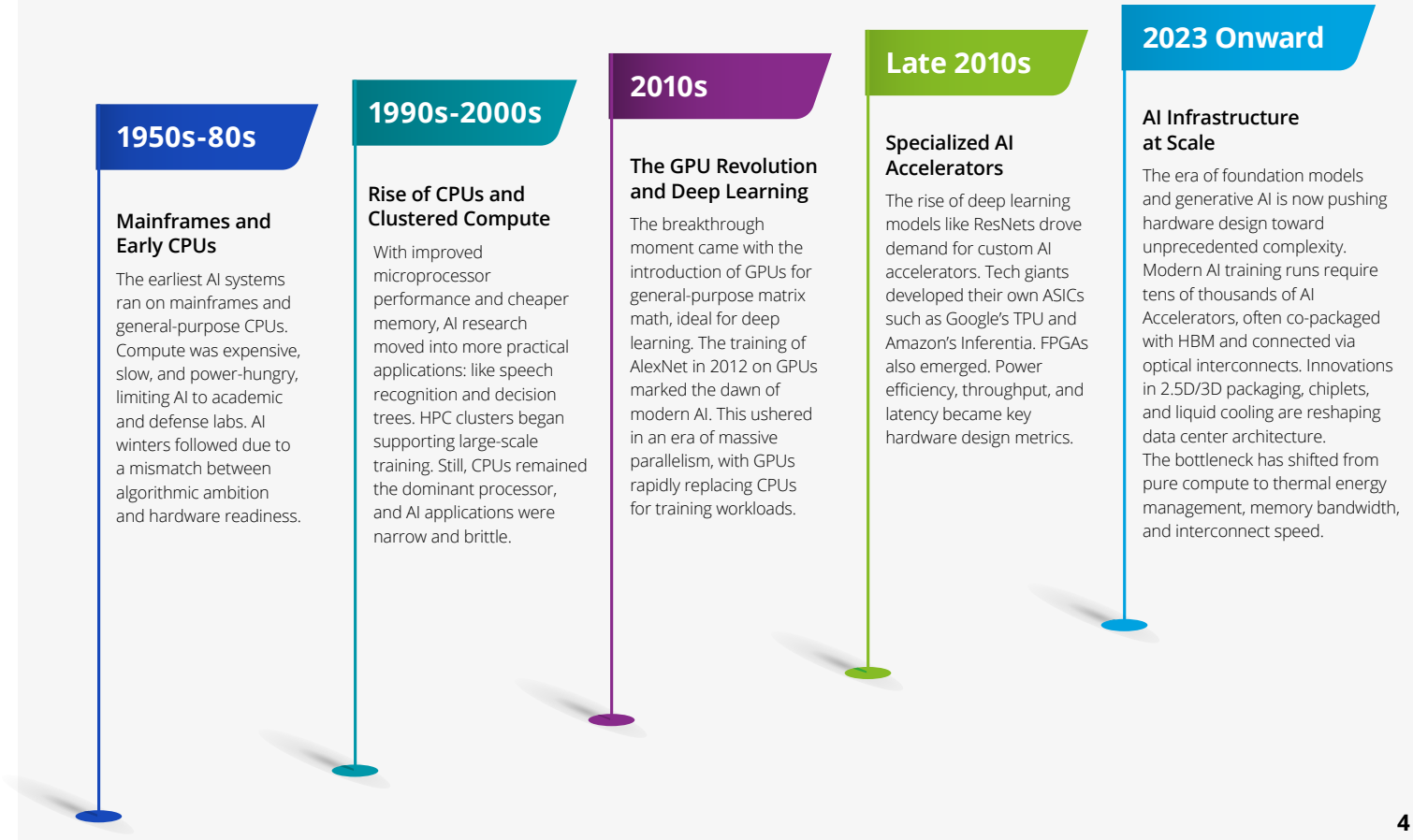
This report offers a unique, inside-out look at the variety of chips that constitute the heart of AI infrastructure by conducting a virtual tear down of a state-of-the-art AI data server, the foundational unit of modern AI infrastructure. Unlike traditional analyses that stop at system-level performance or market sizing, this report dives into the semiconductor content itself within each of the subsystems within a server—mapping the chips, dies, and supporting components that power today's data centers.

We further supplement this analysis by spotlighting where value is concentrated in these server systems and which technologies are critical, from cutting-edge logic built on leading process nodes to mature-node components—like power management integrated circuits (PMICs), electronically erasable programmable read-only memory (EEPROM), compound semiconductors, and microcontrollers—all of which are essential to the functionality of AI systems and infrastructure.

Chip innovations driving AI advancements

Though AI may appear to be a modern development, its foundations stretch through decades of development of state-of-the-art compute power. But functionality was ultimately constrained by limits in hardware available at the time.[2]

Figure 1. Evolution of semiconductor hardware underpinning AI



In the past few years, continued improvements in the scale and sophistication of logic, memory, networking, power, and cooling have paved the way for widespread deployment of high-performance AI systems.

These advancements have catalyzed the rise of AI data centers. While traditional data centers have existed for decades to manage enterprise IT operations, web hosting, and storage, modern AI data centers represent a fundamental capability specialization more than an incremental evolution. Each AI data center server rack integrates an intricate assemblage of advanced semiconductor devices engineered to support parallelization, data proximity, and scalability. An individual AI server rack is composed of approximately 20,000 individual semiconductor dies, consolidated into more than 4,500 packaged chips. These include logic processors delivering high-throughput compute, ultra-low latency memory subsystems, power management units, and networking components.

Semiconductors account for roughly **95%** of a leading AI server rack's content value. Each server rack in an AI data center contains more than **4,500** chips, which are in turn comprised of approximately **20,000** individual semiconductor dies. A leading data center can accommodate over **45,000,000** chips.*

**** Note: Leading AI Data Center assumed to have 10,000 AI Compute racks***

As organizations across industries race to deploy AI-driven solutions, the demand for AI data center capacity and, by extension, advanced semiconductors, has skyrocketed.[3] The impact of this demand surge extends throughout the semiconductor value chain.[4]

Chip designers are under pressure to shorten innovation cycles, releasing new generations of cutting-edge devices more frequently. At the same time, fabricators must deliver major manufacturing technology upgrades to enable the leaps in performance that AI workloads demand. These workloads are pushing the limits of prior hardware generations, exposing critical issues in architecture optimization, heat dissipation, and data movement across massive systems. In response, chipmakers are co-designing hardware and software and pursuing tighter integration of memory and compute, leading to the development of innovative packaging techniques that enable high-density, high-bandwidth configurations.

Figure 2. Demand growth cycle for semiconductors



This is the self-reinforcing cycle of innovation between semiconductors and AI: advances in semiconductor technology and AI systems enable an increasingly sophisticated developer ecosystem which, in turn, demands increasingly powerful AI systems and semiconductor technology. As the developer ecosystem increases in sophistication, AI models scale and require more data, faster processing, tighter coordination across systems, and more computations, pushing the limits of what existing chips and systems can support. This has led to a shift toward highly specialized semiconductor designs, including more efficient and performant processors, more specialized memory stacks, and high-speed interconnects capable of supporting distributed AI workloads across entire data centers. In fact, semiconductor designers and manufacturers are increasingly utilizing AI methods to advance the next generation of their products.

AI Workloads: Training Is the Homework, Inference Is the Test

Throughout this report, we frequently reference two primary AI workloads—i.e., training and inferencing, which represent the distinct phases of AI computation and shape how chips are designed. Let's build a clearer understanding of these two workloads with a simple example:

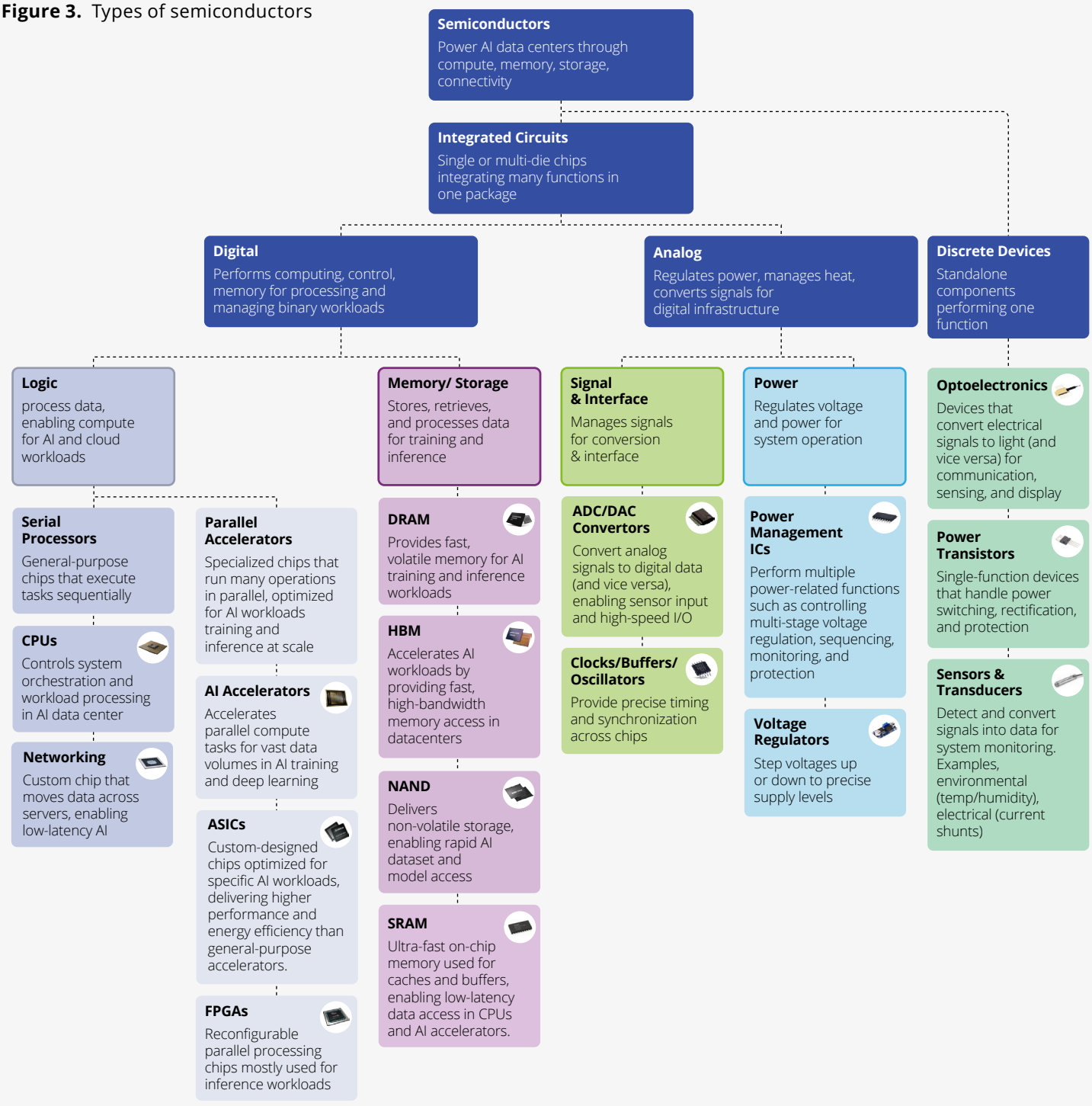
- Training is the process of teaching a model by exposing it to very large sets of data. For example, to build a cat-recognition model, a neural network might be shown thousands of images of cats and non-cats. Through repeated exposure, the model learns to identify patterns, such as ears, whiskers, or body shape, which distinguish cats from other objects.
- Inference is the application of that trained model to new, unseen data. Continuing the cat example, once the model is trained, it can generate a picture of a new cat.

In short, training is how a model learns, while inference is how it applies what it has learned to real-world situations such as generating responses to queries, making predictions, or recognizing patterns.

Semiconductor AI chip taxonomy

AI server racks rely on many types of semiconductor technologies that work together, each tailored to meet the demanding requirements of modern AI workloads (figure 3).[5]

Figure 3. Types of semiconductors



Please note:

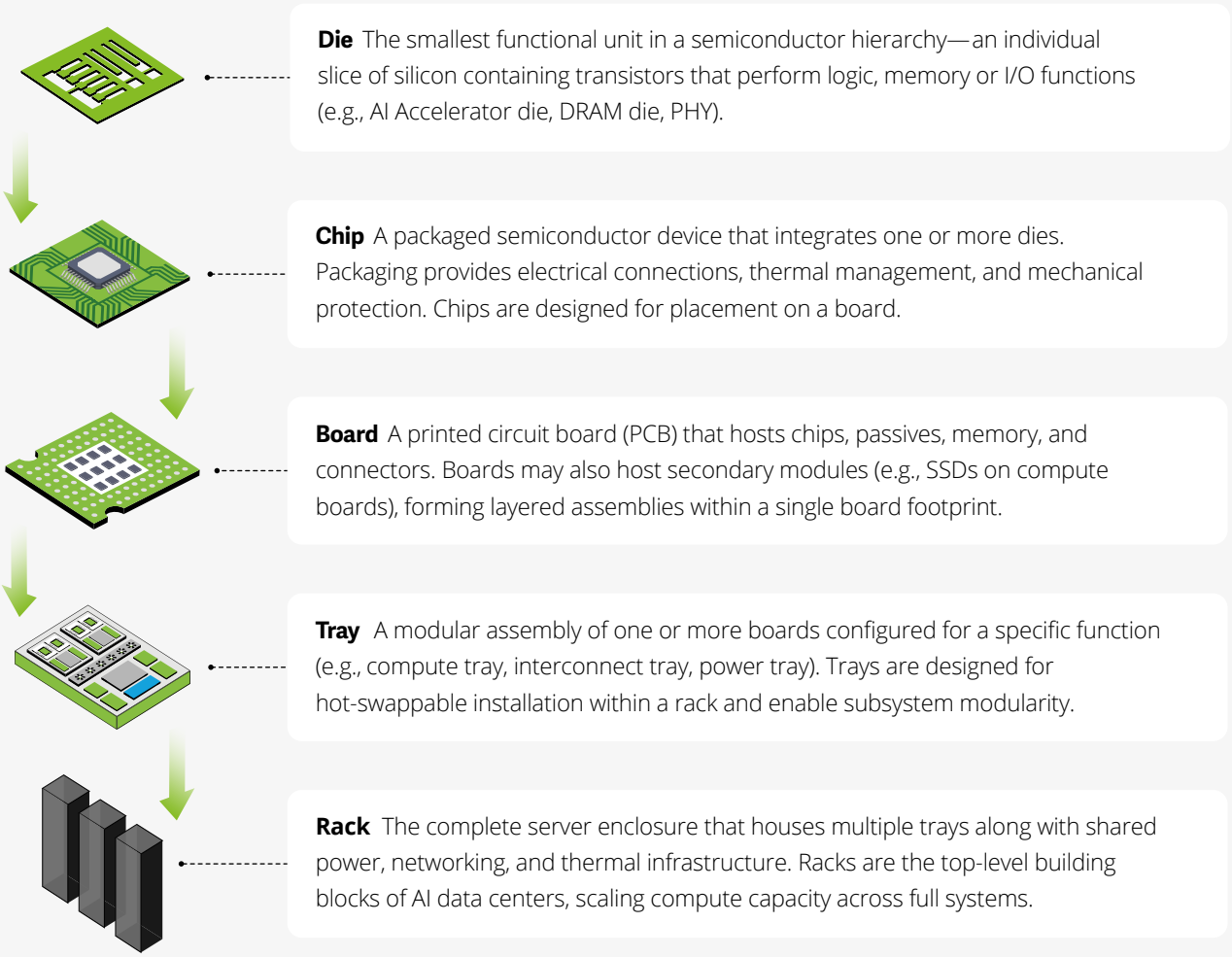
1. Some chips within signal will contain mixed signal i.e. analog and digital attributes
2. The examples are not meant to be exhaustive but indicative, and due to the complex nature of the AI semiconductors, multi-die chip packages contain elements from different classifications.
3. Networking Chips are categorized under logic for the purposes of this report.
4. Discrete devices include both analog and digital semiconductors, but the category is dominated by analog functions. Pure digital discrete devices are outdated and rare, since digital logic is usually integrated into ICs for efficiency.

Within each server are many specialized chips, including the AI Accelerators which perform the parallel processing needed to run complex AI training and inference models. These chips execute billions of operations simultaneously to efficiently process large amounts of information.

Supporting this compute layer, memory semiconductors are critical to performance, enabling fast and reliable data access. As AI models scale, system data volume rises sharply. High-performance memory helps ensure processors are constantly loading data, avoiding bottlenecks, and maintaining system responsiveness. The specialization of memory and logic with AI tasks are increasingly central to next-generation AI system design.[6]

A host of power and networking semiconductors enable efficient energy delivery and seamless interconnection within AI servers and across systems. AI servers use many chips operating at the same time, each tightly coordinated with the rest, making fast and reliable management of power, data, and compute resources essential. In parallel, distributed AI workloads depend on fast, low-latency communication between nodes, making networking semiconductors essential for coordinating compute across groups of interconnected servers.

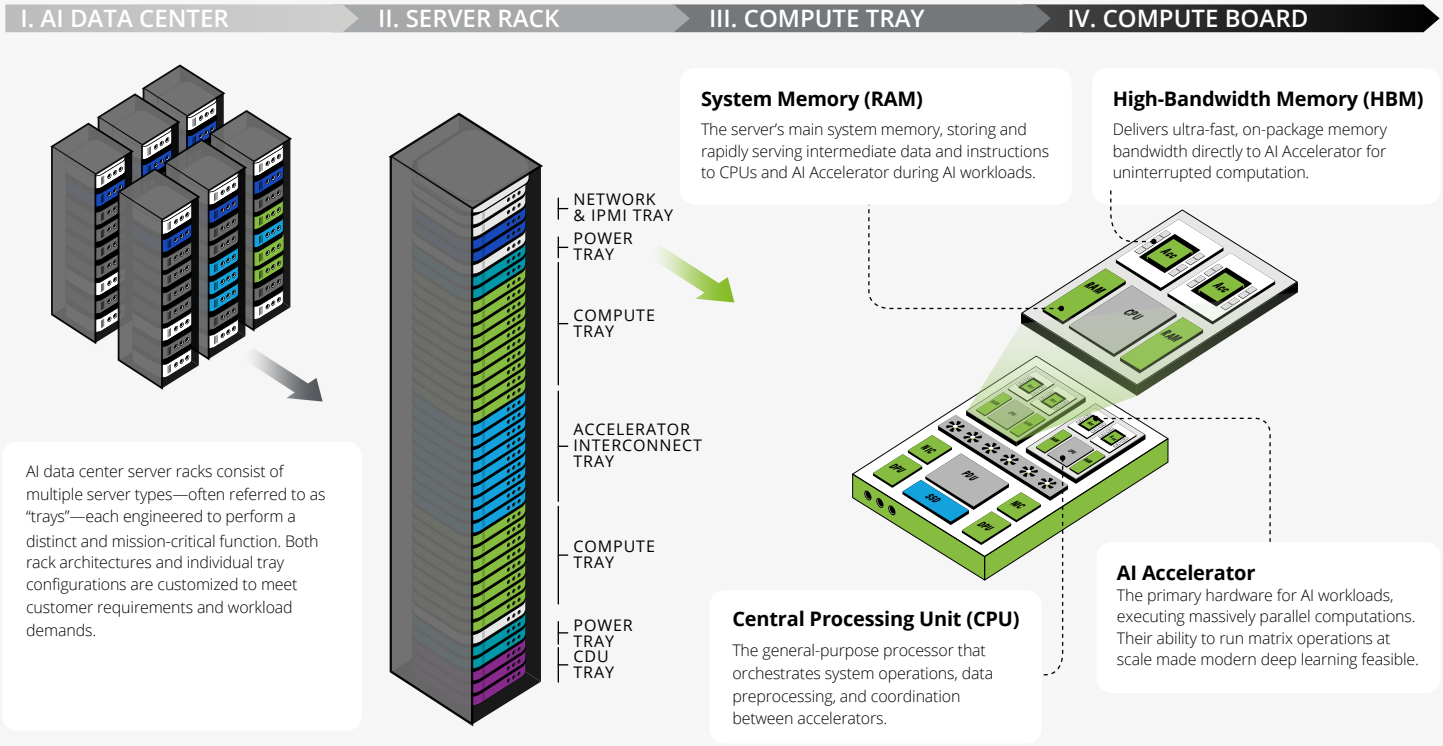
Figure 4. Standard AI hardware stack: From silicon die to rack



II. Breaking down the AI hardware stack: Chips in an AI data server

As the following section will illustrate, breaking down AI server rack hardware reveals a highly modular, vertically integrated system. It is composed of tens of thousands of interdependent semiconductor components across every subsystem within a server, from CPUs and accelerators down to signal conditioning chips, power regulators, memory dies, and control logic.

Figure 5. AI data server teardown



1-2 TRAYS PER SERVER RACK	*2-4 TRAYS PER SERVER RACK*	*1-2 TRAYS PER SERVER RACK*	*8-18 TRAYS PER SERVER RACK*	*4-9 TRAYS PER SERVER RACK*
Network & IPMI Tray	**Power Tray**	**CDU Tray**	**Compute Tray**	**Accelerator Interconnect Tray**
The network and intelligent platform management interface (IPMI) tray provides the control plane infrastructure, combining out-of-band management, high-speed networking, and hardware-rooted security to manage and secure rack-scale operations.	The power tray is the electrical foundation of the rack, housing power supply units (PSUs) and power supply management modules (PSMMs) that convert, condition, and distribute multi-kilowatt loads with high efficiency and real-time monitoring.	The coolant distribution unit (CDU) tray manages rack-level liquid cooling through a closed-loop system connected to cold plates on CPUs, AI Accelerators and memory, ensuring stable thermal performance beyond the limits of air cooling.	The compute tray is the brain of the AI server rack, hosting the primary processing logic. Each tray typically contains one or more compute boards, local Solid-State Drives (SSDs) for storage, Network Interface Cards (NICs) for connectivity, and often Data Processing Units (DPUs) for managing data organization.	The Accelerator Interconnect tray establishes a unified high-performance computing fabric, using switch application-specific integrated circuits (ASICs), power distribution blocks, and active copper cable (ACC) chips to synchronize memory and computation across AI Accelerators.
~100 chips per server rack	***~600 chips per server rack***	***~10 chips per server rack***	***~4,000 chips per server rack***	***~70 chips per server rack***
~$17,000-$25,000 per server rack	***~$50,000-$290,000 per server rack***	***~$15,000-$30,000 per server rack***	***~$1.5M - $3.5M Million per server rack***	***~$10,000-$50,000 per server rack***

Please note: This illustration is for reference only; actual server component configurations can vary significantly depending on specifications, price, vendor, and hyperscaler requirements. On that note, a tray size can also vary between 1U-8U based on the type of tray/function, vendor it is sourced from, data center requirements, etc.

A server rack contains a coordinated set of trays, each designed to deliver a specific function. Each tray, or subsystem, contains a combination of semiconductor and support electronics—loaded with semiconductor content—that operate together to provide the throughput, energy efficiency, and reliability necessary for AI workloads. Understanding this layered architecture is key to appreciating the strategic importance of the semiconductor supply chain that underpins it. As AI continues to scale, the ability to secure, design, and integrate these components will be just as critical as the compute itself.

Generally, there are five types of trays that comprise an AI data server rack found in a modern data center, namely 1) compute trays, 2) power trays, 3) network and intelligent platform management interface (IPMI) trays, 4) AI Accelerator interconnect trays, and 5) coolant distribution unit (CDU) trays. Each individual tray houses differentiated semiconductor components with critical roles. The precise composition of trays in a server rack will vary from data center to data center, with the variation driven by OEM selection, size, layout, power provision, and builder and/or operator of the data center.

This report breaks down the semiconductor content within a generic AI data server, tray by tray.

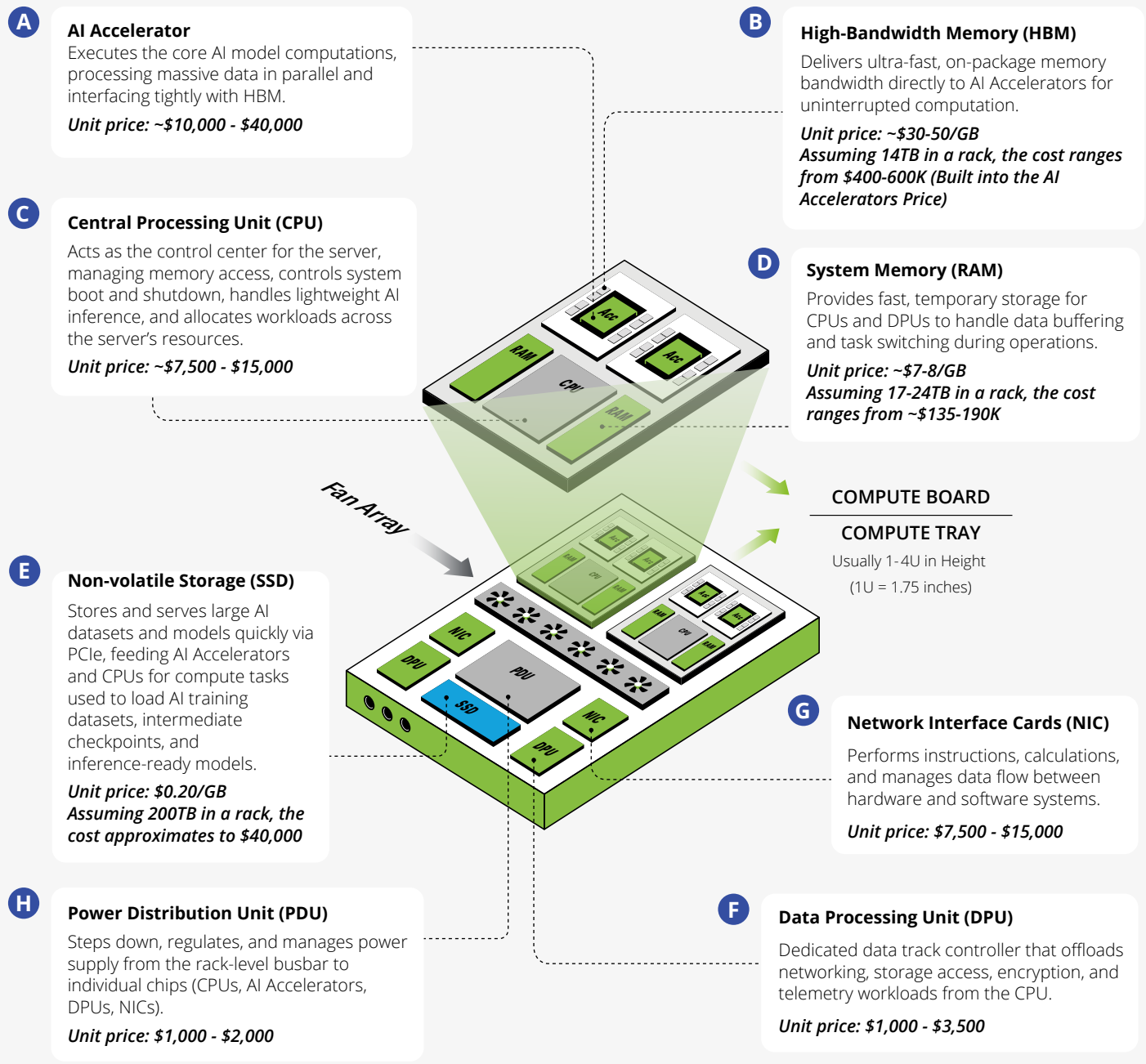


Every component in the system, from the AI Accelerator to individual voltage regulators, plays a vital role in ensuring performance, efficiency, and reliability at scale.

Compute tray

The compute tray is the brain of an AI server rack, providing the primary processing logic required for AI workloads. Each compute tray typically contains one or more compute boards, non-volatile memory express (NVMe) solid-state drives (SSDs) for storage, network interface cards (NICs) for connectivity, data processing units (DPUs) for organizing data into discrete processing units, and a power distribution unit (PDU). This tray bears the most expensive, advanced, and most concentrated semiconductor components within a server rack, as it is directly responsible for performance per watt and AI workload throughput.

Figure 6. Mapping the components of a compute tray



Compute board

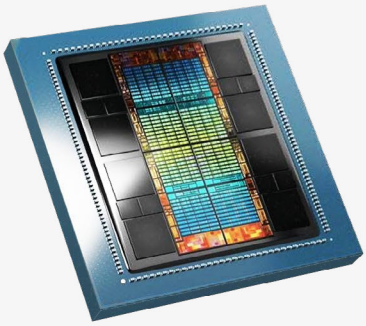
Each compute tray houses one to two compute boards where AI computation is coordinated and executed. Each compute board serves as a high-density processing module within the compute tray. It typically combines one or more high-performance AI Accelerators with a general-purpose CPU, tightly coupled via high-bandwidth interconnects and supported by local random-access memory (RAM). Compute boards are designed for parallelism, thermal efficiency, and signal integrity.

Table 1: Sub-components of a typical AI compute board

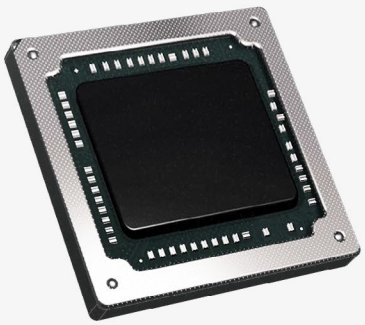
Sub-component	Description	Quantity	Chip type(s)
AI Accelerator	High-performance accelerator with on-package HBM and interconnect die	1–2	Logic and memory—DRAM
Central processing unit (CPU)	Multi-core micro-processor unit for task orchestration, input/output (I/O), and general-purpose workloads	1–2	Logic
System memory (RAM)	High-speed DDR5 or LPDDR5 modules for CPU memory access	8–32 DIMMs	Memory—DRAM
Voltage regulator modules (VRMs)	Power regulators, power management ICs, and discrete metal oxide semiconductor field effect transistor (MOSFET) or insulated-gate bipolar transistor power stages	50–100+	Analog and discrete—power transistors
Control and support ICs	Board management controllers, temperature sensors, oscillators, buffers, clock generators and/or other phase-locked loop	10+	Logic, sensors, and analog
Networking and interconnect	Communication logic ICs (e.g., InfiniBand, Ethernet)	1+	Logic

A. AI Accelerators

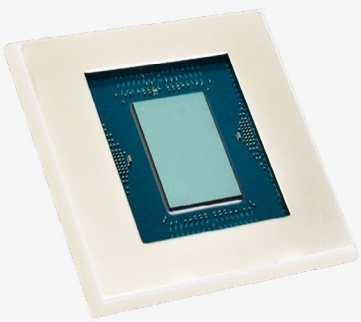
AI accelerators, an umbrella category that includes Graphics Processing Units (GPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs), are specialized semiconductor hardware designed to speed up artificial intelligence workloads by optimizing compute-intensive operations that would otherwise overwhelm general-purpose processors. These accelerators dominate large-scale training and inference in AI data centers, where maximizing throughput and performance is the priority. At the edge, a new class of accelerator has emerged: the Neural Processing Unit (NPU). NPUs are purpose-built to execute neural network workloads with exceptional energy efficiency, optimizing matrix and tensor operations critical for inference.



GPU



FPGA

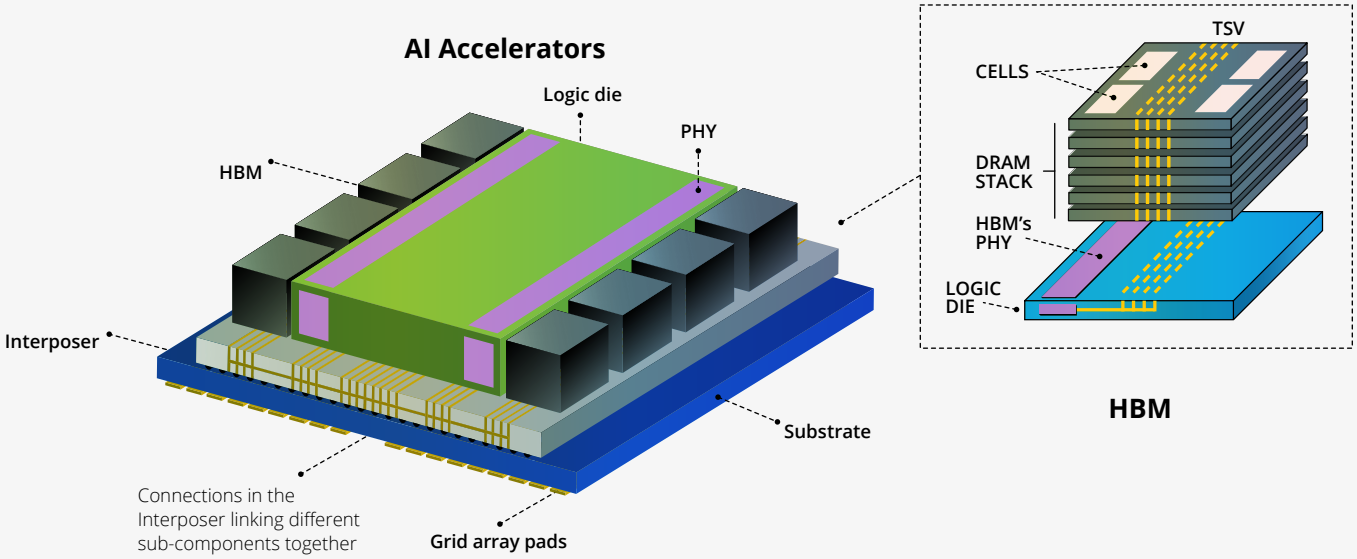


ASIC

While multiple AI Accelerator architectures exist as shown above, GPUs or Graphics Processing units remain the most market-relevant and widely deployed platform for AI training and high-performance inference today. It is the computational powerhouse of modern AI infrastructure. Originally designed to render graphics, GPU architecture allows for extreme parallel processing capabilities to handle billions of computations. Like other modern AI accelerators, they integrate logic dies, which are fabricated on cutting-edge process technology (7 nm), with HBM stacks, enabling the rapid data throughput required for real-time AI applications and large language model processing.

These AI Accelerators rely on advanced packaging technologies, such as 2.5D integration with high-bandwidth memory (HBM) and emerging 3D approaches, which go far beyond conventional packaging, which simply mounts a single die in a package. Unlike conventional packaging methods, advanced packaging tightly integrates multiple dies and memory to maximize bandwidth, reduce latency, and improve power efficiency, making it essential for powering AI data center workloads. It is important to note that these AI accelerators themselves are composed of both active and passive subcomponents. A generic bill of materials (BOM) breakdown for these subcomponents is provided in figure 7.

Figure 7. Schematic breakdown of a modern AI Accelerator and HBM



B. High-bandwidth memory (HBM)

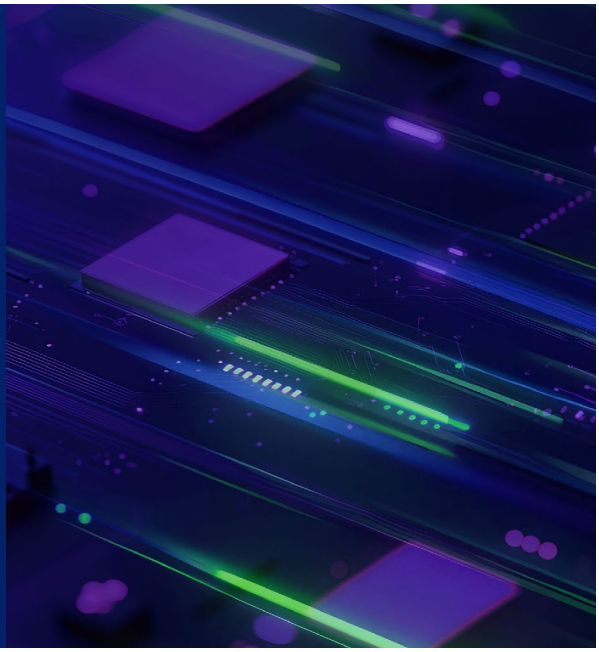
HBM (component B of figure 6) is a type of dynamic random-access memory (DRAM) packaged within an AI accelerator to provide extremely fast and efficient memory performance required by modern AI accelerators. Unlike system memory (described further below) HBM stacks layers of memory vertically using vertical electric connections (i.e., through-silicon vias or "TSVs") to improve the speed, energy efficiency, and scalability of data transfer back and forth to the logic processor.[7] The physical proximity of HBM to the logic dies in an AI accelerator and the massive parallel access capabilities of HBM mitigate memory bottlenecks that would otherwise constrain AI model performance, particularly for large language models that require rapid access to billions of parameters. HBM is often used for both training large-scale AI models and serving high-throughput inference workloads where memory bandwidth directly impacts user experience and operational efficiency. However, HBM is not universally required across all AI systems. Many AI chips, especially those optimized for cost-sensitive, latency-tolerant, or narrower inference workloads, rely on alternative memory architectures such as on-package DDR, LPDDR, or larger on-chip SRAM to balance performance, power, and system cost.

Table 2: Semiconductor components of an integrated AI GPU

Sub-component	Description	Quantity	Chip type(s)
Logic die (ASIC)	These are the AI Accelerators "compute engines." They perform millions of math operations at once, powering AI training and inference. They contain tensor cores, matrix units, and cache.	1–2	Logic
Interconnect PHY (physical link) interface	This is the high-speed communication port that links the AI Accelerator to other AI Accelerators or CPUs using PCI Express (PCIe), Ethernet, or Scale technologies (UAL).	1+	Logic
HBM (high-bandwidth memory)	Stacked DRAM, up to 16 dies per stack, co-packaged with the logic die, connected by through-silicon vias (TSVs). Enables extremely fast data access required for AI workloads.	6–12 stacks (96–192 dies)	Memory—DRAM
Co-packaged interposer or bridge die	A special interconnect layer that connects the AI Accelerator die, Interconnect Phy, HBM, and other chiplets underneath so that they can operate with the efficiency and bandwidth of a unified die. Used in advanced AI Accelerator and CPU heterogenous System-on-Chip architectures	1	Logic

VRMs and Passives

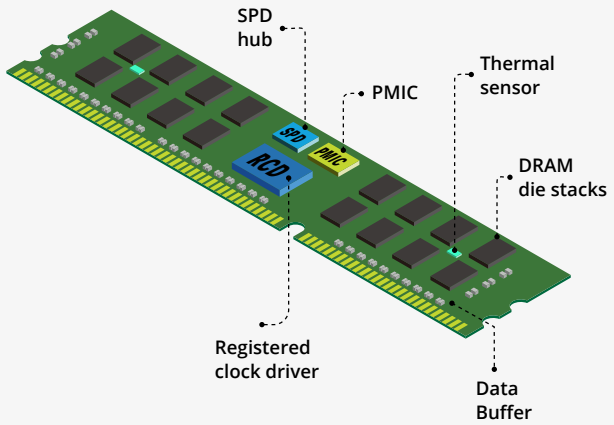
Voltage regulator modules (VRMs) and passive components (capacitors, inductors, chokes, resistors) are the hidden workhorses of AI server hardware. A VRM itself is not a single chip but a small power subsystem typically built from a multiphase controller which drives power stages consisting of MOSFETs which in turn drive driver circuitry, inductors, and capacitors. Together, they step down rack-level voltages to the precise sub-1V rails needed by GPUs, CPUs, and memory. Each accelerator board can host dozens of VRMs and hundreds of supporting passives, and across an entire AI server rack these components number in the thousands. While individually inexpensive and largely commoditized, they are indispensable for maintaining stable, efficient power delivery under the extreme and rapidly changing demands of AI workloads.



C. Central processing unit

The central processing unit (component C of figure 6) serves as the orchestrator and general-purpose compute engine within AI servers, overseeing data preparation (pre and post processing of data), coordinating workload execution, and managing overall system control. The CPU logic die is built on advanced to current generation nodes, depending on performance and cost targets. In AI infrastructure, CPUs are critical for data ingestion, model serving coordination, and running the operating systems and frameworks that support AI workloads. Modern server CPUs uses multi-chip module packaging to combine multiple semiconductor dies with large cache hierarchies, supporting task scheduling and memory management for distributed AI training and inference operations.

Figure 8. Schematic breakdown of a DRAM DIMM



D. System memory/random access memory

System memory/random access memory (component D of figure 6) provides the primary data staging and buffering capability for AI servers, serving as the high-capacity, high-speed storage layer between permanent storage and processing units. System memory/RAM is built from dynamic random access memory die arranged on dual in-line memory modules (DIMMs) or soldered directly to the motherboard. The capacity and bandwidth of system memory directly impact the size of data sets that can be processed simultaneously within an AI data server and the number of concurrent AI workloads a server can support, making it a critical factor in determining overall system throughput and cost-effectiveness.

Table 3: Semiconductor components of DRAM DIMMs

Sub-component	Description	Quantity	Chip type(s)
DRAM die	To store and retrieve data at high-speed using volatile memory cells	8–16 dies per DIMM	Memory—DRAM
Registering clock driver (RCD)	Buffers and synchronizes clock and command signals from the memory controller	1	Analog—signal conversion
Power management IC (PMIC)	Locally manages and distributes power to DRAM chips, reducing motherboard burden	1	Analog—power management
Serial Presence Detect (SPD) EEPROM	Stores configuration data for system BIOS to correctly identify memory	1	Memory—flash
Thermal sensor	Monitors module temperature for thermal throttling or shutdown	1	Sensors

Architectural Trade-offs: Serial vs. Parallel vs. Reconfigurable Compute

CPUs and AI Accelerators differ primarily in their processing architecture. CPUs are optimized for serial processing, executing a few complex tasks sequentially with low latency. CPUs can handle inference on smaller, basic, low-parameter AI models without the need for a dedicated accelerator, making them sufficient for basic AI tasks where latency and scale are less demanding. While AI Accelerators are built for massive parallelism, handling thousands of simpler operations simultaneously. This enables AI accelerators to handle 10-100x more calculations per second than CPUs and achieve far greater data processing throughput, making them especially effective for workloads like AI training and inference, which involve large-scale matrix operations. As a result, AI accelerators have become the compute engine of choice in AI data centers.

As recently as 2015, high-end GPUs were found in gaming machines, not data centers. Most data center tasks were handled by CPUs—serial processors executing one task at a time. In 2016, researchers found GPUs used for gaming could efficiently run machine learning models due to their parallelism. Data centers subsequently began adopting optimized GPUs for AI, though the market remained modest. The 2022 emergence of large language models (LLMs) in generative AI pushed demand for more specialized AI accelerating hardware, often paired with high-bandwidth memory (HBM). These units relied on data center-grade or server-grade CPUs, architecturally similar but separate from standard CPUs, along with other essential components, such as expanded I/O channels, enhanced memory bandwidth, and high-speed interconnect support. By 2025, nearly all the world's top 500 supercomputers use this mix of AI Accelerators, HBM, and CPUs. In effect, today's mega-scale AI data centers are purpose-built supercomputers, designed to train and run AI at scale.

Field programmable gate arrays (FPGAs) are reconfigurable semiconductor devices that can be programmed post-manufacturing to perform a wide variety of logic operations.[8] Unlike CPUs and AI Accelerators, which have fixed instruction sets and execution architectures, FPGAs consist of a flexible matrix of logic blocks and high-speed interconnects that can be rewired through hardware description languages to suit specific workloads. This programmability allows FPGAs to deliver highly customized acceleration for targeted tasks without the upfront cost or lead time of designing custom ASICs.

FPGAs serve as adaptable accelerators in AI infrastructure, occupying a niche where low latency, deterministic execution, and interface flexibility are critical. Unlike other AI Accelerators, which excel at massively parallel matrix operations, or CPUs, which manage general-purpose control tasks, FPGAs are valued for workload-specific acceleration, specializing in inference. They optimize data paths, reduce input–output bottlenecks, and integrate seamlessly with diverse sensors and network protocols, making them well-suited for edge AI, telecom, and specialized data center use cases where configurability and performance-per-watt matter most.[9]

FPGAs are complementary to other AI Accelerators in AI data centers, carving out value in specialized inference niches.

E. Non-volatile memory express storage

Non-volatile memory express solid-state drives (component E of figure 6) store structured and unstructured data to support modern AI workloads. Leveraging the PCIe interface and a streamlined command set optimized for parallelism, NVMe drives ultra-low latency and extremely high throughput, enabling rapid ingestion of large data sets, fast model checkpointing, and seamless inference model loading. These capabilities eliminate I/O bottlenecks that can otherwise stall AI Accelerator utilization, making NVMe storage a core enabler of end-to-end AI pipeline efficiency. NVMe SSDs play a key role in staging large training data sets. In inference-centric systems, these memory drives often act as model caches, serving large models to memory on demand to reduce cold-start latency. Beyond logic components such as the AI Accelerator and CPU, memory components are also composed of multiple sub-components, as demonstrated in the BOM breakdown of a typical NVMe SSD in table 4.

Table 4: Sub-components of a typical data center non-volatile memory express solid-state drive

Sub-component	Description	Quantity	Semiconductor type
NAND flash	Stores persistent data in memory cells using flash technology	256 (8 stacks of 32 dies in a package)	Memory—flash
SSD controller	Manages data flow, temperature control, wear leveling, and flash translation layer	1	Logic
DRAM cache	Temporarily stores metadata and user data for fast access	1	Memory—DRAM
I/O IC	Handles interface protocols (PCIe/NVMe) and signal conversion	1	Analog—interface
PMIC	Regulates power delivery to all active SSD components	1	Analog—power management

F. Data processing unit

The data processing unit (component F of figure 6) is a specialized, programmable infrastructure processor that offloads non-compute tasks from CPUs and AI Accelerators in AI data centers. It operates at the intersection of networking, storage, and security, serving as the controller and data path enabler for efficient resource orchestration. In AI-centric workloads, where massive data sets are streamed and shuffled across AI Accelerators, DPUs optimize network data flow (east-west traffic), handle security tasks like isolating workloads, and collect real-time system data.

Modern DPUs play a key role in making AI infrastructure more flexible, by allowing compute, storage, and networking resources to be dynamically assigned and reconfigured through software. Like AI Accelerators, DPUs are also composed of other semiconductor components (table 5).

Table 5: Semiconductor components of a typical DPU

Sub-component	Description	Quantity	Semiconductor type
Embedded logic die	A small processor inside the DPU that runs its own mini operating system; handles software and management tasks without involving the main server CPU	1	Logic
Packet processing engine	A custom-built engine that moves data quickly between networks, memory, and devices; handles complex tasks like encryption, traffic steering, and telemetry faster than a regular CPU could	1	Logic
On-package DRAM	External memory for buffering, control-plane operations, and software execution	1–2	Memory—DRAM
Physical link (PHY) interface	Transmits and receives Ethernet traffic over fiber or copper cables; converts data to and from the physical link	1	Analog—interface
Flash storage (SPI/EEPROM)	Stores boot firmware, telemetry logs, and operating system images	1	Memory—flash
PMICs/VRMs	Power regulation and distribution for SoC and high-speed I/O domains	6–8	Analog—power management

G. Network interface card

Modern AI workloads require massive bandwidth. Network interface cards (component G of figure 6) enable high-speed interconnectivity essential for distributed AI training and inference serving, supporting terabit-scale data transfers between server racks, compute trays, and storage systems. NICs with advanced features like remote direct memory access (RDMA) and hardware-accelerated networking protocols minimize communication overhead and latency, directly improving training efficiency and inference response times in large-scale AI deployments. The core semiconductor content includes a network controller ASIC alongside analog physical link (PHY) die for signal integrity and clock/data recovery.[10] There are two types of networking ASICs: scale-up, which facilitates low-latency data movement within a system, and scale-out, which enables connectivity and coordination across larger, distributed systems. These chips are often packaged with onboard memory buffers to support congestion control and packet queuing (table 6).

Table 6: Sub-components of a high-performance NIC

Sub-component	Description	Quantity	Chip type(s)
Network controller ASIC	System on chip with embedded DRAM handling packet processing, RDMA, and PCIe/CXL connectivity	1	Logic
Optical transceiver	Module that converts high-speed electrical signals to optical	1	Optoelectronics
SerDes/physical link (PHY) interface	Circuitry that converts digital data to analog signals for high-speed serial transmission	4–8	Analog—interface
DDR memory module	On-board DRAM for buffering, queuing, and control-plane operations	1	Memory—DRAM
PCIe interface	Connector and logic enabling host communication via PCIe Gen4/Gen5	1	Logic
PMICs/VRMs	Power delivery and regulation for core ASIC, I/O blocks, and memory	2–4	Analog—power management
Digital signal processor	Boosts and equalizes signal for long fiber runs, especially in DR4/DR8 optics	1	Logic—digital signal processor
Oscillator	Provides the stable reference clock signal required for NIC timing	1	Discrete – timing
Clock buffers	Distribute the oscillator reference clock to multiple NIC domains while preserving signal integrity	1–2	Analog – timing/buffer
Clock generator	Alternative to oscillator and buffers; derives multiple frequencies from a single reference	1	Analog – timing/ clock generator

H. Power distribution unit

In advanced AI compute trays, the power distribution unit (component H of figure 6) converts and distributes electrical power from high-voltage inputs to precise, low-voltage rails required by CPUs, AI Accelerator, memory, and other subsystems. In advanced AI compute trays, the PDU takes in rack-level 48V or higher and delivers tightly regulated voltages, often below 1V, to high-density components with high transient loads. Internally, they house multiple digital and analog semiconductor components, including power management ICs, voltage regulator modules (VRMs), integrated field-effect transistors (FETs), and microcontrollers. These chips are typically built on current generation to mature process nodes for durability and thermal stability.

As AI data centers demand more power, the chips responsible for managing and converting electricity are evolving to deliver greater efficiency. Compound semiconductors, such as those made from gallium nitride (GaN) and silicon carbide (SiC), are increasingly used in power management systems because they can handle higher voltages with less energy loss and heat.[11] Also, there is a likelihood that AI data centers will move from the current data center mix of higher voltage AC combined with 48V DC power to the racks, toward the proposed 400V DC power supply throughout, which would lead to various efficiency gains. There are even some plans to move to 800V high-voltage DC (HVDC) architectures to support racks that individually require 1 MW (i.e., 1,000 kW) of power in the near future, compared to the 100–120 kW requirements of the most powerful racks in use today.[12]

Compound Semiconductors, Beyond Silicon

Compound semiconductors represent a strategic evolution in materials science, expanding performance envelopes where silicon is reaching physical limits. Materials such as GaN, SiC, gallium arsenide (GaAs), and indium phosphide (InP) enable properties like wide bandgaps, higher electron mobility, and superior thermal conductivity, which are essential for next-generation data center scaling and optical technologies. These advances allow devices to operate at higher voltages, frequencies, and temperatures, reducing the size and cooling requirements of supporting infrastructure.

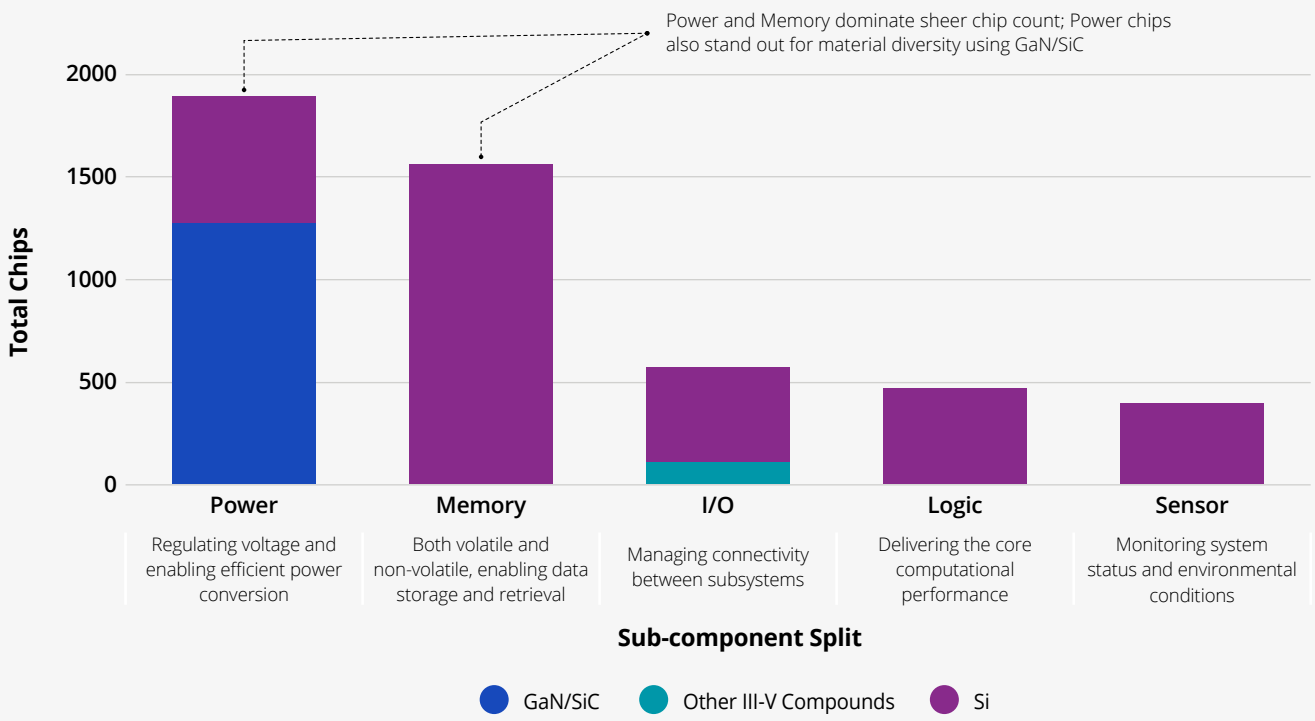


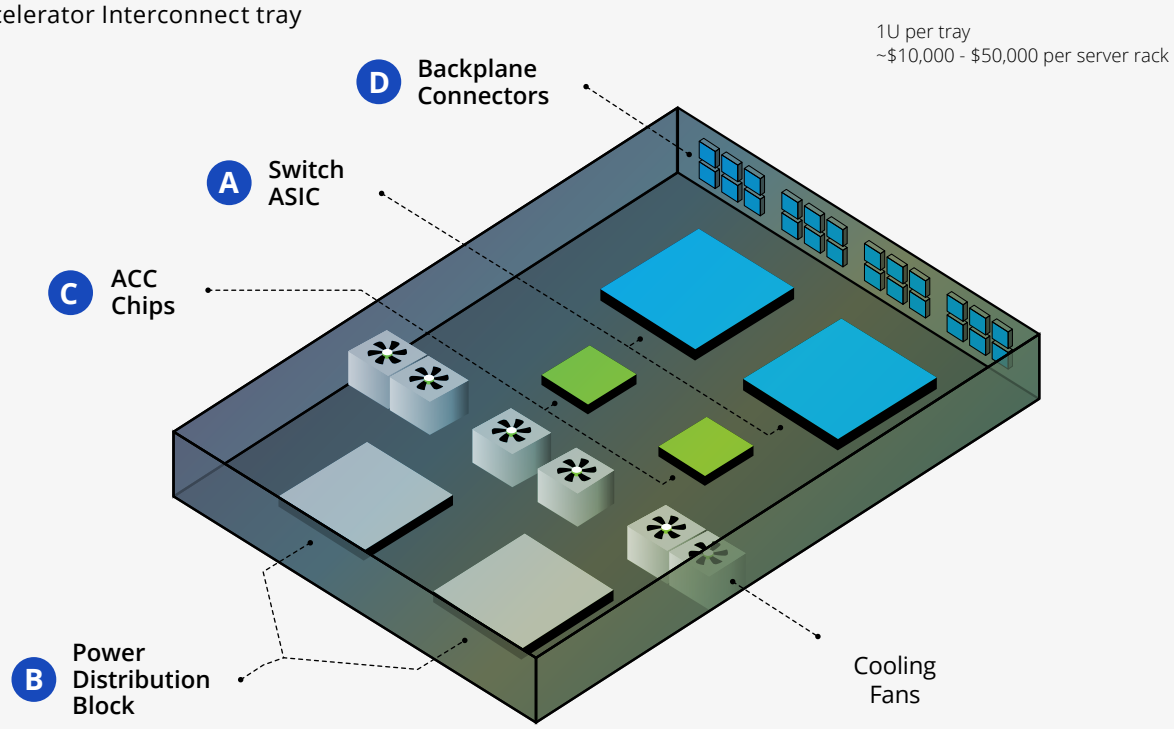
Table 7: Semiconductor components of a typical power distribution unit for AI server compute tray components

Sub-component	Description	Quantity	Chip type(s)
Intermediate Bus Converters (IBCs)	High-power DC-DC converters that step the rack-level 48–51 V bus down to intermediate rails typically at 12 V	2–6	Discrete
Point-of-load (PoL) converters	Local voltage regulators that convert the intermediate rail (12V/6–8V) into low-voltage rails (5 V, 3.3 V, 1.8 V, 1.0 V)	6–12	Discrete
Microcontroller or PMBus SoC	For closed-loop power management, telemetry, and I2C/PMBus control	1–2	Logic—MCUs
Power management ICs (PMICs)	Control logic for sequencing, telemetry, and dynamic voltage scaling	2–6	Analog—power management
Current and voltage sensors	Analog/digital ICs for real-time telemetry and fault detection	5–10	Analog—amplifiers
Drivers & FETs	Performs actual current switching and regulation at IBC level (e.g., 48V->12V)	2–4	Discrete

Accelerator interconnect tray

The Accelerator interconnect tray creates a unified, high-performance computing fabric that enables AI Accelerators within an AI server to function as a single unit. This tray contains high-bandwidth switch application-specific integrated circuits (ASICs), power distribution blocks, and signal-conditioning active copper cable (ACC) chips that orchestrate seamless communication between the dozens of AI Accelerators in a server rack (figure 10). This allows them to share memory, synchronize computations, and collectively tackle large-scale AI workloads that exceed the capabilities of individual accelerators (table 8).

Figure 10. Accelerator Interconnect tray



A. Switch ASICs

Switch ASICs (component A of figure 10) enable high-speed, low-latency communication between multiple AI Accelerators within a server rack or across server racks. Their performance directly impacts the scalability and efficiency of AI workloads, especially in LLM training and other compute-heavy applications that require synchronized processing across multiple accelerators.

B. Power distribution block

The power distribution block (component B of figure 10) within the Accelerator interconnect tray manages localized, high-current delivery to switch ASICs and interconnect components, aggregating and redistributing power throughout a server rack. Internally, this element of the Accelerator interconnect tray incorporates power chips to distribute and regulate power to the trays.

C. Active copper cable chips

The ACC chip (component C of figure 10) is a compact analog signal-conditioning IC embedded at each end of an ACC, enabling high-speed, short-distance communication (up to 5 meters) across trays or racks.

D. Backplane connectors

The backplane connectors (component D of figure 10) serve as the main hub that links multiple AI Accelerators on the compute trays. Backplane connectors are the physical interfaces that attach the Accelerator interconnect tray and other modules to the backplane. Backplane connectors are electromechanical components, not semiconductor devices, but are retained here for completeness given their role in system interconnect.

Table 8: Semiconductor components of a GPU interconnect tray

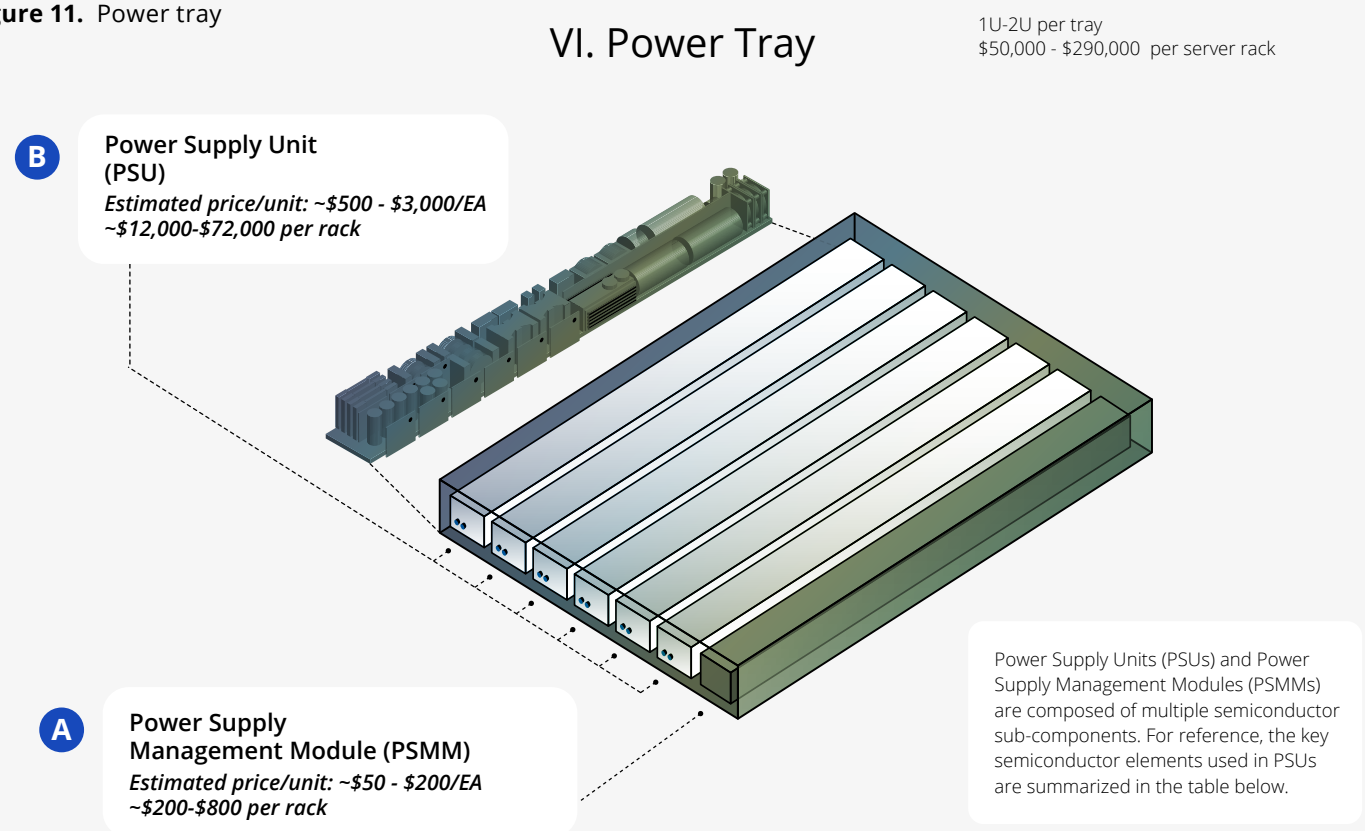
Sub-component	Description	Quantity	Chip type(s)
High-speed PHY interface	Converts digital data into extremely fast electrical signals for sending data between accelerators	2	Logic—digital signal processor
Switch ASIC (fabric switch)	Routes traffic between multiple accelerators so they can all talk to each other efficiently	2 per interconnect tray	Logic
SerDes transceiver	Converts wide data streams into narrow high-speed serial lanes, and back again	16–72+ lanes per Switch ASIC	Optoelectronic
ACC chip	Specialized controller IC that manages low-level control, telemetry, and coordination between multiple AI Accelerator modules	1–2	Analog—Signal conversion
PMICs/VRMs	Incorporated into power distribution blocks to efficiently regulate and distribute power	4–6	Analog—power management
Clock generator/ phase-locked loop (PLL) IC	Provides precise timing signals to synchronize SerDes, switch ASICs, and PHYs	1–2	Logic—digital signal processor
Oscillator	Provides the stable reference clock signal for AI Accelerator interconnect timing	1	Discrete – Timing
Clock Buffers	Distribute the oscillator reference clock to multiple GPU/SerDes domains while maintaining signal integrity	1–2	Analog – Timing/Buffer
Clock Generator	Alternative to oscillator and buffer; derives multiple frequencies from a single source	0–1	Analog – Timing/-Clock Generator

Power tray

The power tray serves as the electrical foundation of an AI server rack, converting, conditioning, and distributing power to all system components within the rack. This tray houses the primary power supply units (PSUs) that must reliably deliver multi-kilowatt power loads while maintaining high efficiency and providing real-time monitoring capabilities. Key semiconductor components within the power tray include bridge rectifiers, which convert the electrical currents; power factor correction (PFC) controller ICs, which optimize the quality and efficiency of the incoming power; and power transistors (often GaN-based), which rapidly switch and regulate voltage conversion for high efficiency.

Unlike the PDU—which is located on or closer to the compute trays and is responsible for localized voltage regulation and fine-grained current delivery to individual components on the compute tray—the power tray operates at the system level handling bulk power conversion and intelligent power orchestration for the entire server rack. Together, these power parts form a hierarchical power delivery network designed for high-density, high-reliability compute environments (figure 11).

Figure 11. Power tray



Within the power tray, the PSU serves as the primary power conversion and conditioning system for AI servers, transforming high-voltage AC power from data center infrastructure into the multiple DC voltage rails required by server components (table 9).

A. Power supply management module (PSMM)

Power supply management module (component A of figure 11) acts as the intelligent orchestrator of power distribution and monitoring within AI servers. PSMM systems help prevent overheating by limiting power when needed, balance electricity across different parts of the server, and support proactive maintenance by constantly monitoring power quality. To do this, the PSMM incorporates a range of chips, including intelligent microcontrollers, voltage regulators for local power, remote access capabilities (Ethernet PHY), and sensors, working together to keep servers running efficiently and minimize downtime.

B. Power Supply Unit (PSU)

Within the power tray, the PSU serves as the primary power conversion and conditioning system for AI servers, transforming high-voltage AC power from data center infrastructure into the multiple DC voltage rails required by server components (table 9).

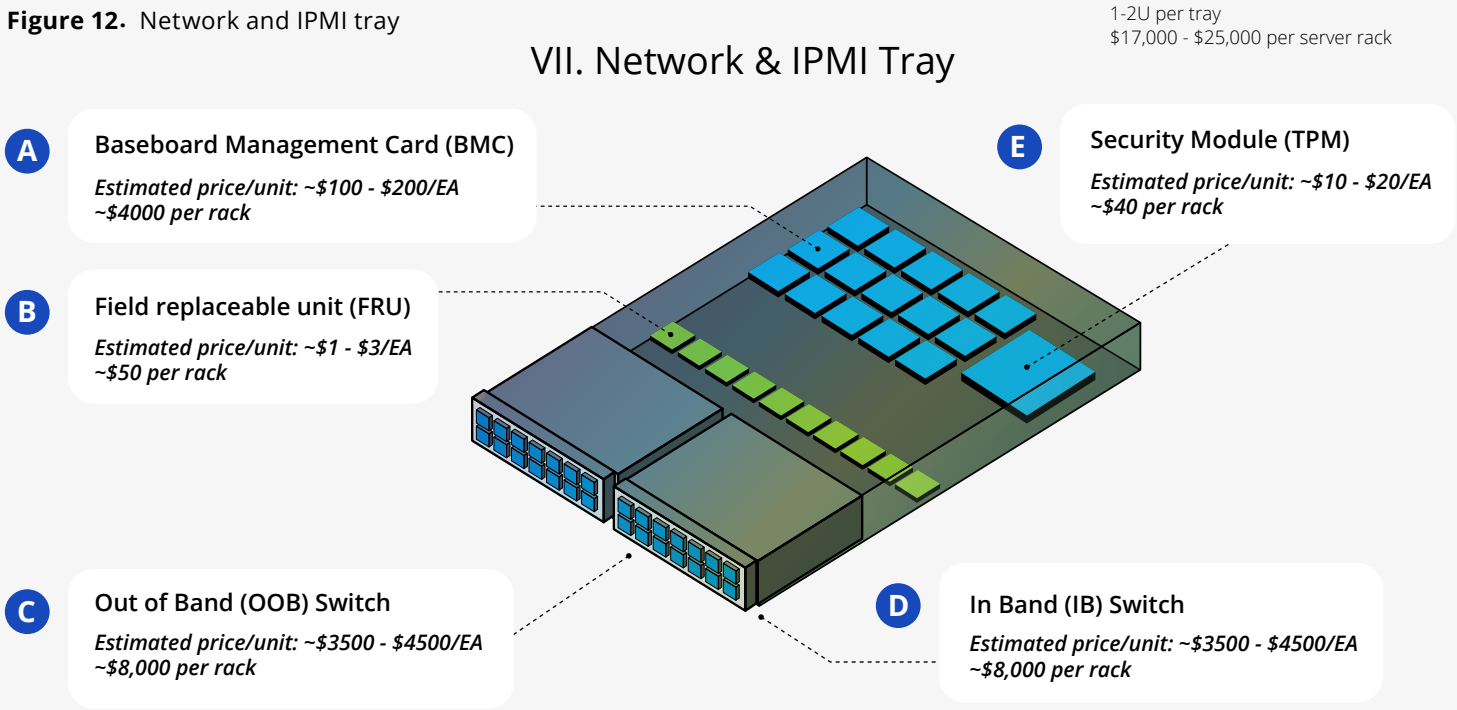
Table 9: Semiconductor components of a server-grade power supply unit

Sub-component	Description	Quantity	Chip type(s)
Bridge rectifier	Converts incoming AC (alternating current) from the mains into unregulated DC (direct current); forms part of the PSU input stage	1	Discrete—rectifiers
Power factor correction (PFC) controller IC	Ensures the input current is drawn in phase with the voltage waveform, improving power efficiency and reducing harmonic distortion	1	Analog—power management
Power MOSFET (e.g., GaN)	High-speed transistors used in switching stages to regulate voltage conversion; enable efficient AC-DC conversion through rapid switching	2–4	Discrete—power transistors
Pulse width modulation (PWM)/resonant controller IC	Regulates switching behavior of the transformer by controlling duty cycle or resonance timing, optimizing power delivery and efficiency	1	Analog—power management
Gate driver IC	Interfaces between controller ICs and power MOSFETs, delivering the correct timing and voltage to switch the transistors	2–4	Analog—power management
Synchronous rectifier FETs	Replaces traditional diodes for rectification using low-resistance transistors, reducing power loss and improving conversion efficiency	2–4	Discrete—power transistors
Microcontroller/power mgt. bus system-on-chip (PMBus SoC)	Embedded control unit for monitoring, telemetry, protection, and communication (e.g., via PMBus or I²C protocols); governs overall PSU logic	1	Logic—MCUs
Voltage and temp sensors	Monitor internal voltage rails, current draw, and thermal conditions to enable system protection and dynamic control	3–5	Sensors

Network and intelligent platform management interface tray

The network and intelligent platform management interface (IPMI) tray is built on a foundation of semiconductor components that enable secure, scalable control and connectivity for AI server operations. Key semiconductors within this tray include baseboard management controllers, hardware security chips, local memory, and high-speed networking switches (figure 12).

Figure 12. Network and IPMI tray



A. Baseboard management card (BMC)

The baseboard management card (component A of figure 12) enables remote power cycling, hardware health monitoring, and system configuration without interrupting running AI workloads (table 10).

Table 10: Sub-components of an AI server rack BMC

Sub-component	Description	Quantity	Chip type(s)
SoC (Microcontroller)	Central processing unit running firmware and IPMI stack for out-of-band management	1	Logic—MCUs
Firmware flash	Non-volatile storage for BMC firmware, logs, and configuration data	1	Memory—flash
Ethernet PHY + magnetics	Provides physical and electrical connectivity for remote management over LAN	1	Analog—interface
Power management IC (PMIC)	Supplies regulated voltages to BMC SoC, memory, and I/O interfaces	1–2	Analog—power management
Watchdog timer IC	Automatically resets the system if a fault or freeze is detected	1	Logic—MCUs
Clock Generator	Provides reference clock signals for BMC and peripheral interfaces, replacing oscillator and buffer in some designs	1	Analog—Timing/-Clock Generator

B. Field replaceable unit

Field replaceable unit electrically erasable programmable read-only memory (FRU EEPROM) (component B of figure 12) acts as digital "name tags" for server components, storing essential details like identity, configuration, and service history. This makes it easy for data centers to automatically track, manage, and maintain hardware.

C. Out-of-band (OOB) switch

The out-of-band (OOB) switch (component C of figure 12) uses specialized semiconductor chips to create a separate network for server management tasks. These switches (powered by network processors and microcontrollers) help maintain reliable oversight and control of the servers, even when the system is under heavy load. This separation is made possible by the advanced networking semiconductors inside the OOB switch, which ensure secure and uninterrupted access for administrators.

Table 11: Sub-components of an out-of-band switch

Sub-component	Description	Quantity	Chip type(s)
Switch ASIC	Core logic chip that routes packets between ports using MAC tables, VLANs, and routing logic in real time.	1	Logic
Ethernet PHY Interface	Converts digital signals to electrical signals for transmission over physical Ethernet media and handles link protocols.	24	Analog—Interface
Management SoC / CPU	Embedded processor that runs the switch OS and handles CLI, API, telemetry, and system management functions.	1	Logic—MCUs
Flash Memory	Non-volatile storage used to hold firmware, configuration files, operating system, and event logs for recovery.	1	Memory—Flash
DRAM	Volatile system memory that provides temporary storage for active data used by the management processor.	1	Memory—DRAM
Power Management ICs	Regulates power delivery to all active SSD components.	1	Analog—Power management

D. In-band (IB) switch

The in-band (IB) switch (component D of figure 12) manages high-speed data traffic between compute nodes and storage systems within the AI data center by supporting advanced networking protocols such as 400G/800G Ethernet or InfiniBand. The IB switch ensures scalable, synchronized communication, essential for training and inference across multi-rack AI clusters.

E. Trusted platform module (TPM)

The trusted platform module (component E of figure 12) provides hardware-based security foundations for AI servers, including secure boot capabilities, cryptographic key storage, and hardware attestation services. TPMs are also known as security modules. In AI environments handling sensitive data or proprietary models, TPM modules ensure system integrity and enable secure multi-tenant operations by providing tamper-resistant security anchors. These capabilities underpin emerging confidential computing architectures, in which data and models are protected not only at rest and in transit but also during execution through hardware-enforced isolation and encryption in CPUs and accelerators. While confidential computing is not yet the default for AI training or inference, adoption is increasing, particularly for inference workloads in regulated or multi-tenant environments, as standards and platform support continue to mature.

Table 12: Sub-components of an In-band switch

Sub-component	Description	Quantity	Chip type(s)
Switch ASIC	Central chip that receives data packets and routes them to the correct port at ultra-high speed	1	Logic
SerDes Transceivers	Converts parallel data into high-speed serial for each port (112G–224G per lane)	100+	Optoelectronic
Ethernet/InfiniBand PHY Interface	Converts data to and from electrical/optical signals, depending on port type (e.g., OSFP)	1	Analog—Interface
Optical Transceivers	Pluggable optics for high-speed fiber links (800G, 1.6T)	72	Optoelectronics
Digital Signal Processor	Boosts and shapes high-speed signals for clean fiber transmission	1	Logic—Digital signal processor
Power Management ICs	Deliver and regulate power across different parts of the switch ASIC and transceivers	10	Analog—Power management
Fan / Thermal Management Controller	Controls temperature with fans, enables local and remote setup via serial and Ethernet ports, and shows system status through indicator lights.	1	Logic—MCUs
Flash Memory	Non-volatile storage used to hold firmware, configuration files, operating system, and event logs for recovery.	1	Memory—Flash
DRAM	Volatile system memory that provides temporary storage for active data used by the management processor.	1	Memory—DRAM



Co-Packaged Optics interconnect

As AI models grow increasingly complex, a large cluster of accelerators must work in unison to develop frontier models with billions of parameters. Similarly, inference solutions based on Agentic AI require progressively larger cluster sizes to deploy multiple agents that can operate concurrently. To meet this insatiable demand for computing and memory, data centers are implementing high-speed scale-up networks that provide high-bandwidth connectivity within a pod or cluster.

Until now, these compute clusters have depended on electrical signals to transfer data. However, as bandwidth requirements increase and cluster sizes expand, it becomes increasingly challenging for electrical connections to meet the reach and bandwidth demands of these networks. High speed electrical signals show significant degradation in signal quality as the network size and distance increase. Optics, and specifically co-packaged optics, offer a pathway to achieving higher bandwidth and energy-efficient scale-up networks. Co-packaged optics technology leverages the advancements made in chiplet technology over the past decade to integrate silicon photonics solutions on accelerator and switch packages. By positioning an optical link close to the accelerator, they provide an energy-efficient solution.

Despite these advancements, several challenges remain in the implementation of co-packaged optics solutions. While CMOS process technology and supply chains are well-established, optical technology is still in the early stages of adoption. Establishing standard protocols for optical communication is crucial for building a supply chain that ecosystem partners can readily access. Standard components, such as laser modules (light sources for optical engines) and pluggable fiber modules, are vital for scaling optics from lab-scale to high-volume manufacturing.



Thermo-mechanical integration

The next phase of AI infrastructure will be constrained by heat flux, mechanical warpage, interface stability, and assembly/manufacturability as much as by compute. Across AI/HPC projections, package areas scale to ~9,000–10,000 mm², die area trends toward ~4,000–5,000 mm², and average accelerator die power is projected to surpass ~5,000 W in the next decade, implying at least a doubling of effective power density even as silicon scales. These conditions increase sensitivity to TIM behavior under real flatness/curvature, clamp-load limits, and cyclic reliability (silicon cracking, delamination, interconnect fatigue) during both package and system-level assembly, making package–retention–cold-plate co-design a strategic requirement.

In parallel, as AI factories and data centers shift from air cooling toward liquid (air declining to ~65% by 2035 while liquid rises toward ~30%, with other such as 2 phase refrigerant cooling ~5%), the critical problem becomes maintaining low and stable thermal resistance at the liquid interface without over-constraining the mechanics. Field realities also matter: reworkability and serviceability increasingly influence uptime and total cost as AI deployments scale.

Looking ahead, thermo-mechanical design will hit an inflection point with emerging system-on-wafer / system-on-panel concepts, where module/tray-level power can push toward the >25 kW class, and with deeper photonic integration to reduce latency, bringing tighter requirements on thermal stability and mechanical alignment. The practical winners will be architectures that treat thermal as a mech-enabled system: engineered for manufacturability at OSAT/EMS scale, validated with correlated test vehicles, hardened to mission-profile cycling, and designed for efficient field service.

Coolant distribution unit tray

The coolant distribution unit tray manages rack-level liquid cooling for high-performance AI servers. As modern AI Accelerator and CPUs in AI workloads generate significantly higher thermal loads than traditional servers, air cooling alone is no longer sufficient. The CDU ensures reliable thermal management by circulating liquid coolant through a closed-loop system connected to cold plates mounted on CPUs, AI Accelerators and memory modules. This not only prevents thermal throttling but also enables higher sustained compute performance in dense configurations.

Semiconductors enable the various pumps, sensors, circuit control valves, and power supply units in these trays. For example, a CDU controller board, equipped with embedded logic semiconductors, manages flow rates, monitors pressure and temperature, and regulates pump speed and valve positions for precise thermal control. An Ethernet interface containing semiconductor chips (PHY and MAC controllers) converts signals and manages data packets, enabling remote monitoring and control, while a local LCD display provides real-time diagnostics and operational status (figure 13).

Figure 13. CDU tray

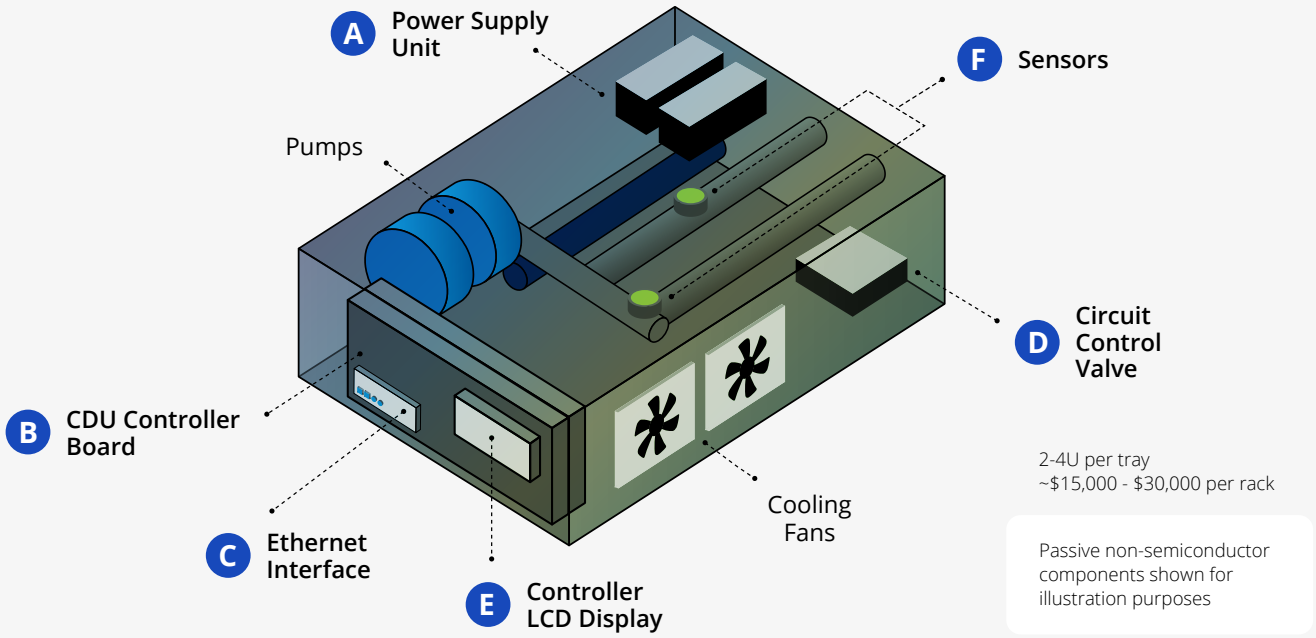


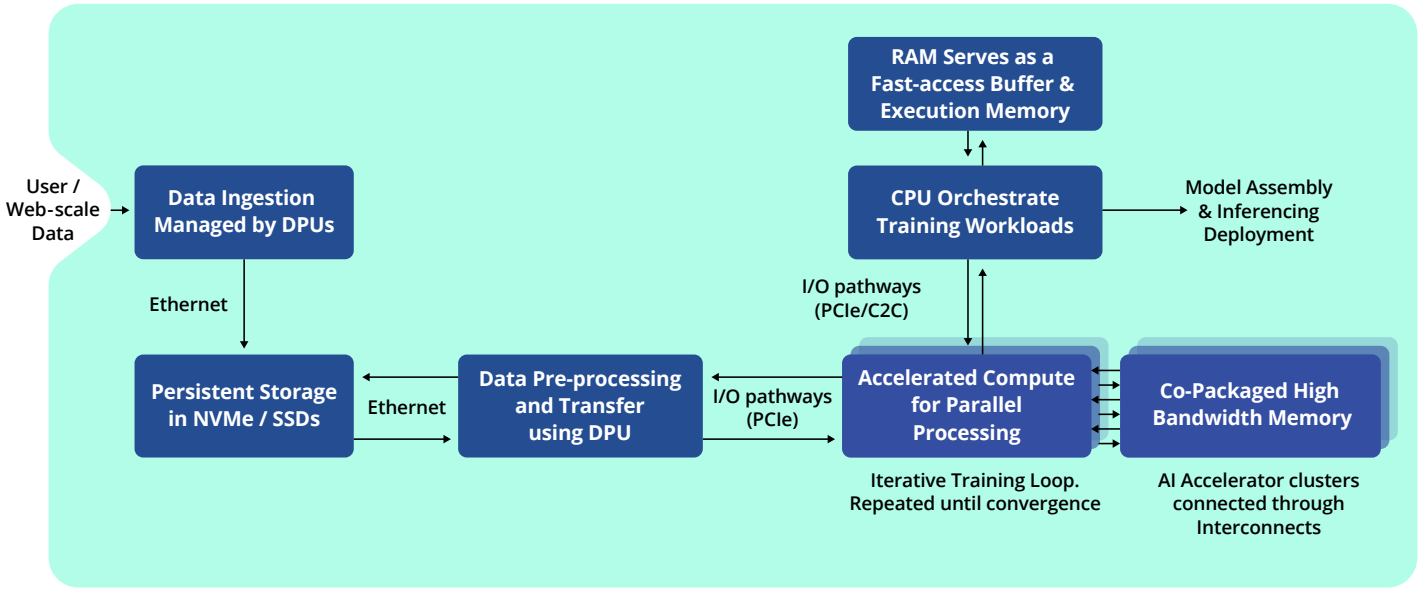
Table 13: Semiconductor components of a coolant distribution unit

Sub-component	Description	Quantity	Chip type(s)
Flow Sensors	Ensures voltage compatibility between MCU and devices	3	Sensor
Pressure Sensors	Includes pumps, flow meters, valves, cold plates, radiators, cooling fans and mechanical connections that circulate coolant across the server rack.	4	Sensor
Temperature Sensors	Measure real-time liquid flow rate (in L/min)	4	Sensor
CDU Controller Board (MCU)	Monitor coolant temperature delta (ΔT)	1	Logic – MCUs
Ethernet/PM Bus Interface	Handles system monitoring, fan/pump control, Ethernet interface	1	Analog – Interface

III. How chips work together to execute an AI training workflow

Training AI models at scale requires a tightly coordinated system of specialized chips, each optimized for a distinct role in the workflow. Figure 14 illustrates how CPUs, AI Accelerator high-bandwidth memory (HBM), non-volatile storage, and optical interconnects operate in unison to manage parallel workloads, synchronize compute across nodes, and ultimately deliver a fully trained model ready for deployment.

Figure 14. Data in a data center: How chips work together in an AI training workflow



AI model training starts with data ingestion, which is handled by data processing units that function as coordinators for web-scalable data entering the system. Once ingested, data is written to NVMe SSDs, which are optimized for checkpointing and feeding multiple accelerators concurrently. This architecture enables rapid throughput and reduces latency during training, especially when paired with technologies like GPUDirect, which allows direct access from storage to AI Accelerator memory, bypassing the CPU altogether.

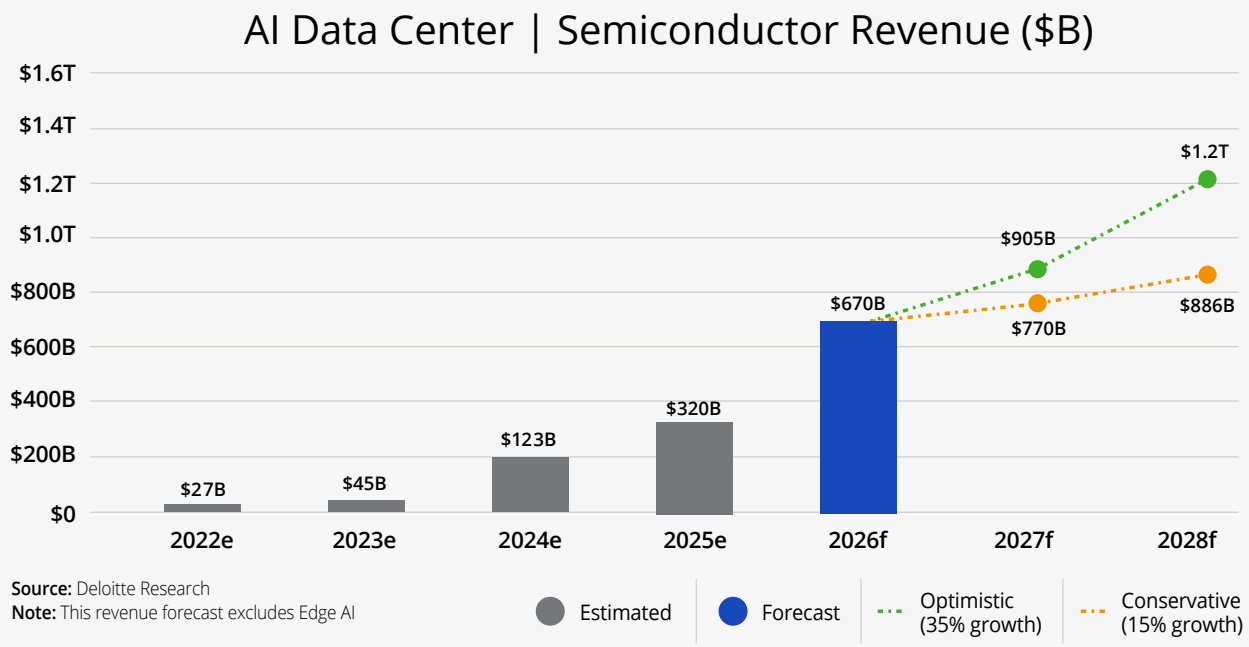
DPUs also perform pre-processing on data at line rate, parsing and formatting it before it's passed directly to the compute units. This minimizes the CPU's workload and ensures that AI Accelerators receive clean, structured data for training. SmartNIC-integrated DPUs can simultaneously support storage offload, network acceleration, and protocol handling, allowing greater scalability with lower power consumption. Once data reaches the AI Accelerator cluster, training is accelerated through highly parallelized compute across thousands of cores, supported by HBM to manage vast model weights and activations across distributed nodes.

Meanwhile, CPUs orchestrate the training workload, managing synchronization, memory coordination, and task distribution. They handle control flow logic and offload activations to balance memory utilization, enabling support for larger models and batch sizes. Leveraging interconnects, AI Accelerators exchange model states and intermediate activations at high speeds, allowing for massive models to be trained iteratively and efficiently across many devices. Once training is complete, CPUs aggregate outputs and finalize model assembly, readying it for downstream deployment.

IV. The AI growth curve: Market view

AI adoption is growing rapidly, which is driving greater demand for the semiconductors that power data centers supporting these workloads. Revenue for semiconductors deployed in AI data centers is projected to reach more than $1.2 trillion by 2028, an almost tenfold increase over five years (figure 15). The market for data center logic chips (primarily AI Accelerators) was approximately $30 billion in 2022, expanded to $70 billion in 2024, and is projected by WSTS to reach $190 billion by 2026.[15].

Figure 15. Total revenue by semiconductors deployed in AI data center servers



Source: Deloitte Research
Note: This revenue forecast excludes Edge AI

Different AI applications require different compute infrastructure, with implications for semiconductor hardware design and deployment. While most data center demand for semiconductors has been driven by training for AI model development, the market for inference (e.g., applying trained models to real world scenarios) is growing.

Training and inference represent distinct computational tasks. AI model training, during which a data center processes massive data sets to learn patterns and make sense of complex real-world variables, requires semiconductors with significant compute power. Chips designed for AI training deliver extreme parallelism and support enormous memory bandwidth, enabling a model to adjust its internal parameters millions, or even billions, of times until it can make accurate predictions or recommendations. This demands hardware with advanced interconnects, dense transistor layouts, and high-speed memory hierarchies.

Training workloads are episodic and increasingly efficient, which will lead to flattening demand for training specific chips. Advances in model architecture and training techniques have reduced the time and compute needed to train large models. Further, the availability of pretrained foundational models further decreases the need for large-scale training runs.

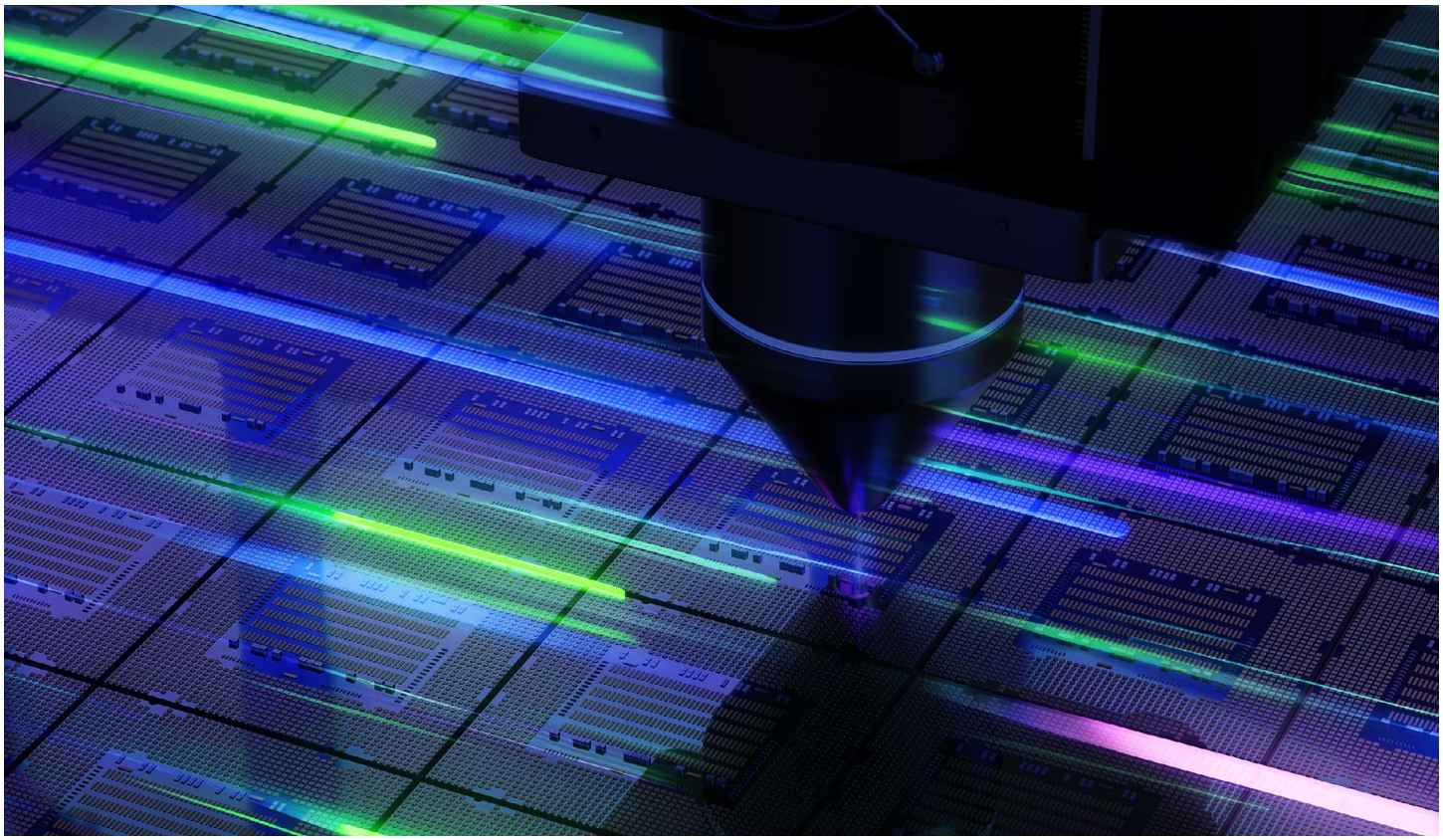
Once trained, models are deployed for inferencing—performing real-world tasks, generating responses to queries, making predictions, or recognizing patterns. Inferencing occurs in a data center, often referred to as "in the cloud," or on a device, referred to as "at the edge" (e.g., inside AI PCs, smartphones, vehicles, or factory systems). In data centers, inference chips are optimized for throughput and latency, leveraging high-efficiency compute to execute operations in milliseconds. At the edge, on-device AI accelerator chips prioritize low latency, small form factors, and energy efficiency, enabling local decision-making without constant server communication. Given the unique characteristics of AI training versus inference workloads, data center operators are increasingly opting for vertically tailored application-specific integrated circuits (ASICs) to optimize performance and efficiency gains.

Inference workloads are predicted to become the primary driver of AI-related chip demand. Demand for inference compute scales with consumer use of generative AI tools, enterprise adoption, and the rise of AI agents embedded into applications.[16] Unlike hyperscale training clusters, edge AI deployments are far less resource-intensive, relying on smaller, energy-efficient chips optimized for inference. Rising inference workloads point toward a growing need for custom, power-efficient silicon: AI Accelerators optimized for throughput, dedicated inference ASICs/FPGAs, CPUs tuned for latency-sensitive tasks, and associated memory and networking chips.[17]

Training and inference workloads currently run on the same hardware components despite their different workload profiles. That said, chipmakers are designing products specifically for inferencing workloads driven by a need for greater efficiency or high throughput.[18, 19, 20, 21] As product differentiation continues to advance, the industry could see a sharper split between training- and inference-specific hardware.[22] Training chips may concentrate on maximizing parallel compute, while inference chips evolve toward ultra-low-latency, energy-efficient designs.

Industry projections indicate that revenue from inference workloads will grow several times over in the coming years, while revenue from training is expected to flatten. For example, Boston Consulting Group estimates training will grow at a 30% compound annual rate from 2023 to 2028.[23] In contrast, inference will grow at 122% over the same period. To contextualize this shift, industry experts forecast inference's share of total demand can grow from 20% in 2024 to 80% by 2032.[24]

AI is poised to drive a substantial portion of the semiconductor industry's growth through the end of the decade, with McKinsey projecting semiconductor industry annual to reach $1.6 trillion revenue by 2030, largely due to AI & data centers.[25] With a semiconductor market of $791.7 billion in 2025[26] Gen AI could therefore represent over 40% of the industry's growth in the coming years.



V. The big spend: Semiconductor content value analysis

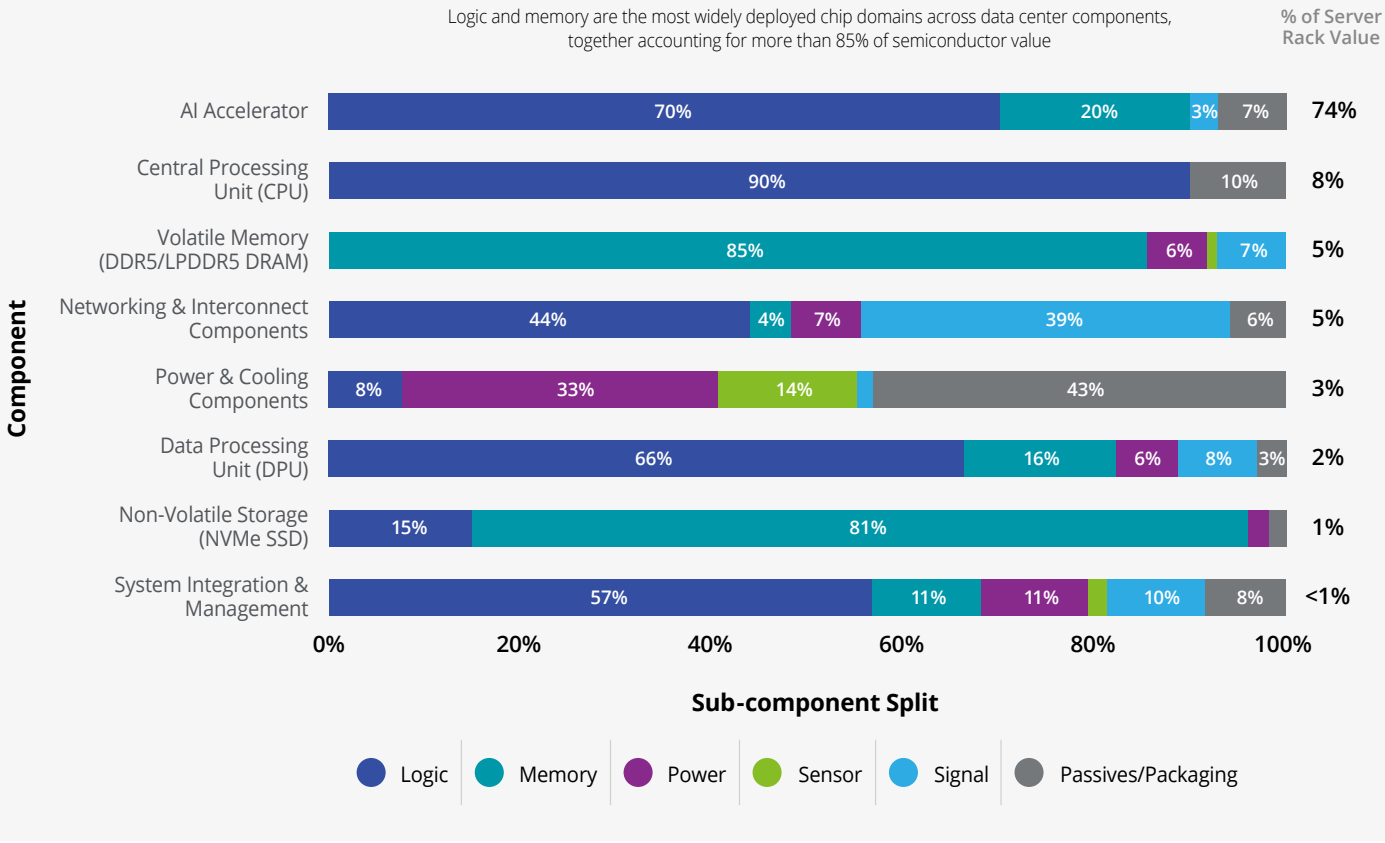
To meet global AI demand, a cumulative $4.0 trillion is expected to be invested in building AI data centers over a period of eight years from 2023 through 2030, of which up to $2.8 trillion is expected to be spent on semiconductors and other hardware for AI server setups.[27] In a typical AI data center, the majority of capital expenditures lies in compute infrastructure, which accounts for more than 50% of an AI data center's total capital expenditures. This investment is concentrated in the AI server rack, a modular, high-value system costing $1.5 million to $4 million each, purpose-built to handle the unique demands of AI workloads. A leading data center could house up to 10,000 racks.

To further understand where this investment would land in the semiconductor value chain, we analyzed the semiconductor content within a standard, industry-grade AI server and how this content translates into economic value across the global semiconductor supply chain. (Note: Figure 16 is based on a single AI data center rack, valued between $1.5 million and $4 million. The percentages shown represent the approximate share of component content within that rack.)

For a modern, leading AI server rack, roughly 95% of content value resides in semiconductors. At the component level, the AI Accelerator captures more than 70% of semiconductor content by value in an AI data server rack. Each GPU module co-packages one or more large, leading-node logic dies, HBM memory stacks, interposers, and specialized interconnects that contribute to this overall value. CPUs, also architecturally critical, account for 8%, and the DPU adds a further 2%.

Overall, logic chips account for 65% of the semiconductor content in a server rack, which also includes AI Accelerators, CPUs, DPUs, NIC controllers, BMCs, and ASICs.

Figure 16. Share of semiconductor content value in an AI data center rack



Memory and storage semiconductors account for more than 20% of a rack's total semiconductor content value, composed of HBM, DRAM, and NAND flash devices integrated in various server trays. These figures reflect the architectural nature of modern AI workloads: vast matrices, distributed models, and large training sets that demand not only raw compute but extreme memory bandwidth and low-latency access to data.

The stability of the entire AI data server also rests on a long tail of lower-cost, current-generation node components and emerging material innovations that account for the remaining approximately 10% of the rack's semiconductor content value. While individually low in unit cost, these chips are critical to voltage stability, signal integrity, and thermal control, particularly in tightly packed racks with megawatt-level power densities. This category includes analog semiconductors such as power management ICs, voltage regulators, and clock generators that ensure reliable power delivery and signal timing. It also includes compound semiconductors (e.g., GaN and SiC power devices) that enable high-efficiency power conversion for liquid cooling pumps and high-voltage distribution. Together, these supporting chips provide the resilience, efficiency, and safety margins that allow the high-value logic and memory components to perform at scale.

As previously noted, the semiconductor value composition within a given server will vary to some extent depending on certain factors (see the appendix for methodology and additional detail). However, this variation is driven primarily by the specific logic chip selection. Server racks that integrate cutting-edge AI Accelerators exhibit a compositional value of 65% to 75% in logic chips, while those using mid-tier AI Accelerators showed only 40% to 50% logic value, with the balance shifting toward memory and storage components. Higher-priced systems are generally characterized by greater compute density and higher peak performance.[28]

The Performance–Pragmatism Trade-Off

System integrators and hyperscalers design AI platforms by first establishing requirements for performance, capacity, power consumption, and total cost of ownership (TCO). These requirements are driven by the intended workload mix, deployment scale, and operational constraints, and they guide a structured component selection process rather than a single "best-chip" outcome. From this perspective, accelerator selection reflects a series of deliberate trade-offs:

- Performance targets (TFLOPs): Not all AI workloads require the maximum compute throughput offered by cutting-edge GPUs. For many training and inference use cases, mid-tier accelerators deliver sufficient performance while improving utilization efficiency and system balance.
- Total cost of ownership (TCO): Accelerator choices are evaluated holistically, incorporating acquisition cost, power consumption, cooling requirements, rack density, and expected operational lifetime. Lower-power or mid-range accelerators can offer superior TCO for workloads that are not performance-bound.
- Capacity and availability considerations: Supply lead times, platform compatibility, and deployment timelines influence accelerator selection, particularly at scale, where predictability and repeatability are critical.

By far, the compute tray accounts for most of the semiconductor footprint within a data server—about 95% of content value, nearly 90% of total number of semiconductor dies, and more than 80% of the number of individual chips (figure 18).

Figure 17. Die count distribution by process node technology for each tray type

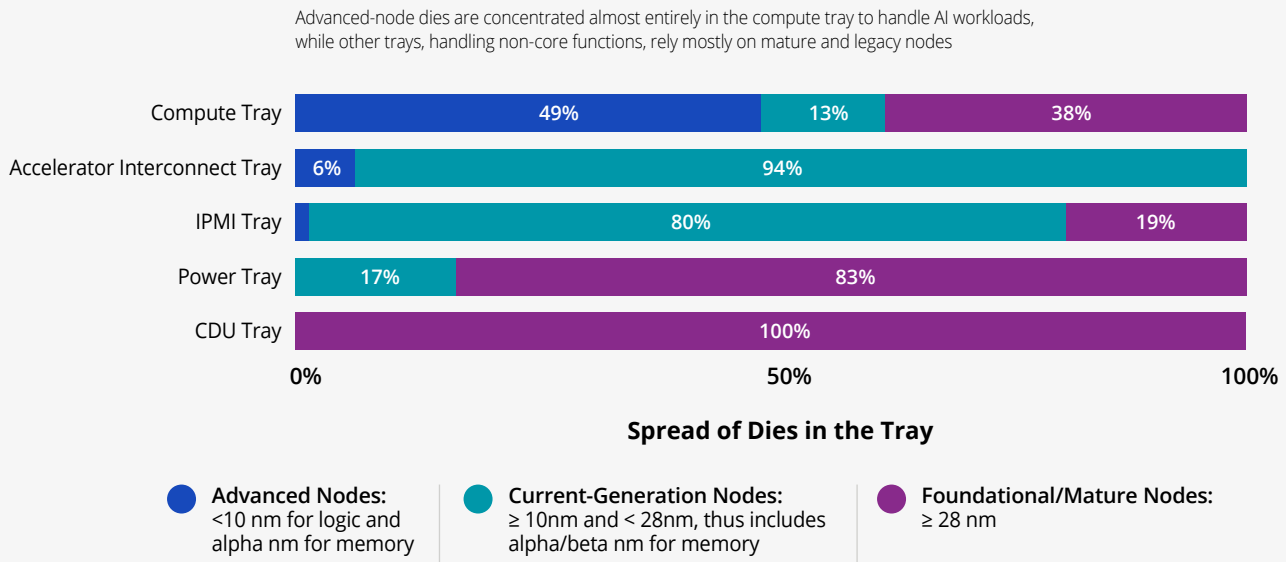
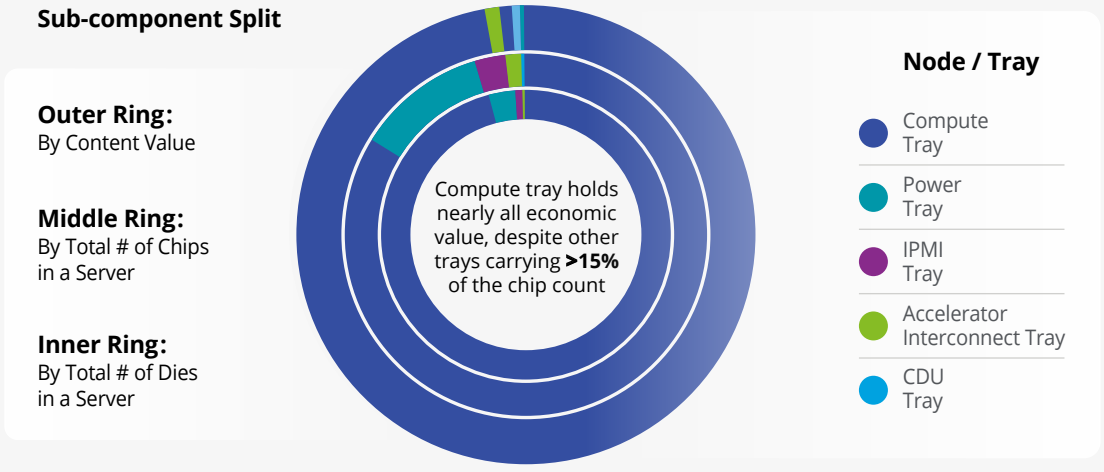


Figure 18. Content value and share of dies/chips by volume for each tray type



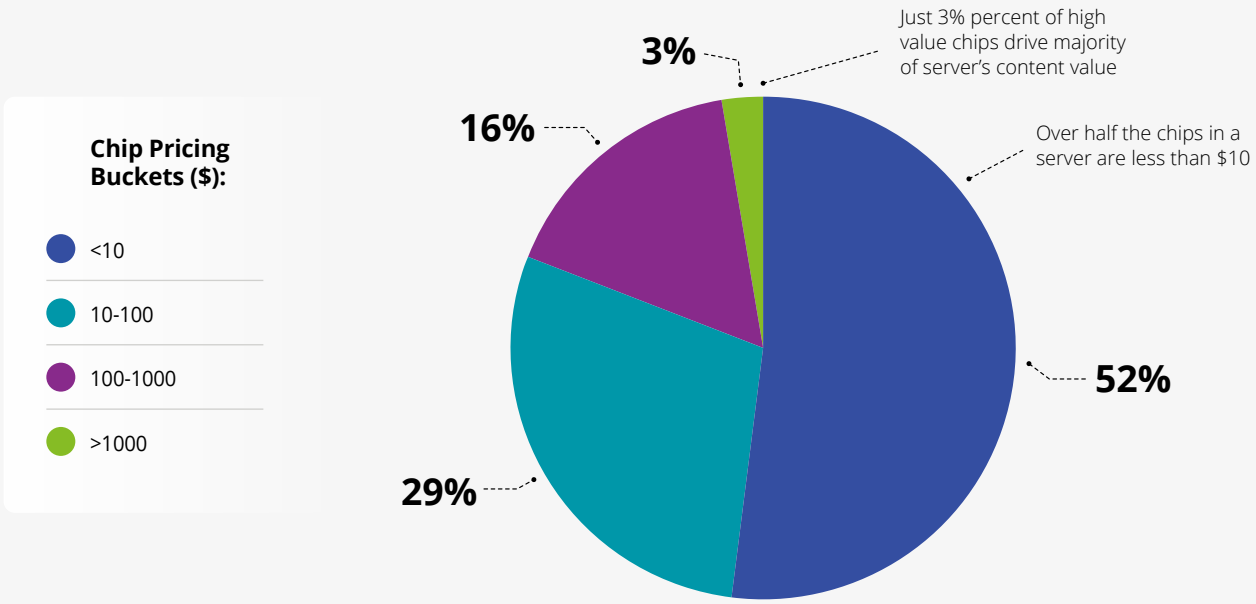
The value distribution of the trays is highly correlated with the level of advancement, or process node, of the semiconductor technology. This report defines advanced nodes as sub-10 nanometer generations, typically fabricated with extreme ultraviolet (EUV) lithography. In addition to the compute tray accounting for the bulk of semiconductor dies overall, most of those dies are manufactured using advanced process technology (figure 18).

The Accelerator interconnect tray and IPMI tray also leverage some advanced dies, but the value of those trays is far lower overall due to the smaller number of dies present and the relative maturity of the technology node used to manufacture those dies.

Value concentration vs. cost dynamics

While a small number of logic chips account for most of the value in an AI data server, the vast majority of components by count sit at the low-cost end of the spectrum but are nevertheless mission-critical for system functionality (figure 19).

Figure 19. Chip distribution across pricing buckets in an AI data center



This duality between a concentrated value stack at the leading edge and critical dependencies across the broader semiconductor spectrum highlights both the opportunities and vulnerabilities shaping the future of AI infrastructure.



VI. From design to data center: Understanding the global supply chain of AI data centers

Up to this moment in this report, we have broken down the chips which go into Data Center AI servers up to the lowest level of detail. To build one of these chips, however, an entire ecosystem needs to be brought together. Data center chip manufacturing requires thousands of steps across a globally distributed supply chain. The process involves two main supply chain segments: one for creating packaged chips from initial designs, and another for integrating these chips into systems ready for data center deployment.

AI servers are not plug-and-play. They must be slotted into pre-provisioned power domains, matched to network topologies that minimize latency, and connected to distributed storage clusters that ensure throughput. Facility operators, cloud hyperscalers, and infrastructure providers all play a role in this deployment choreography and often work closely with semiconductor providers to ensure compute performance requirements are met. As AI infrastructure continues to scale, supply chains are expected to deliver both chips and complete, integrated systems ready to run complex workloads from day one.

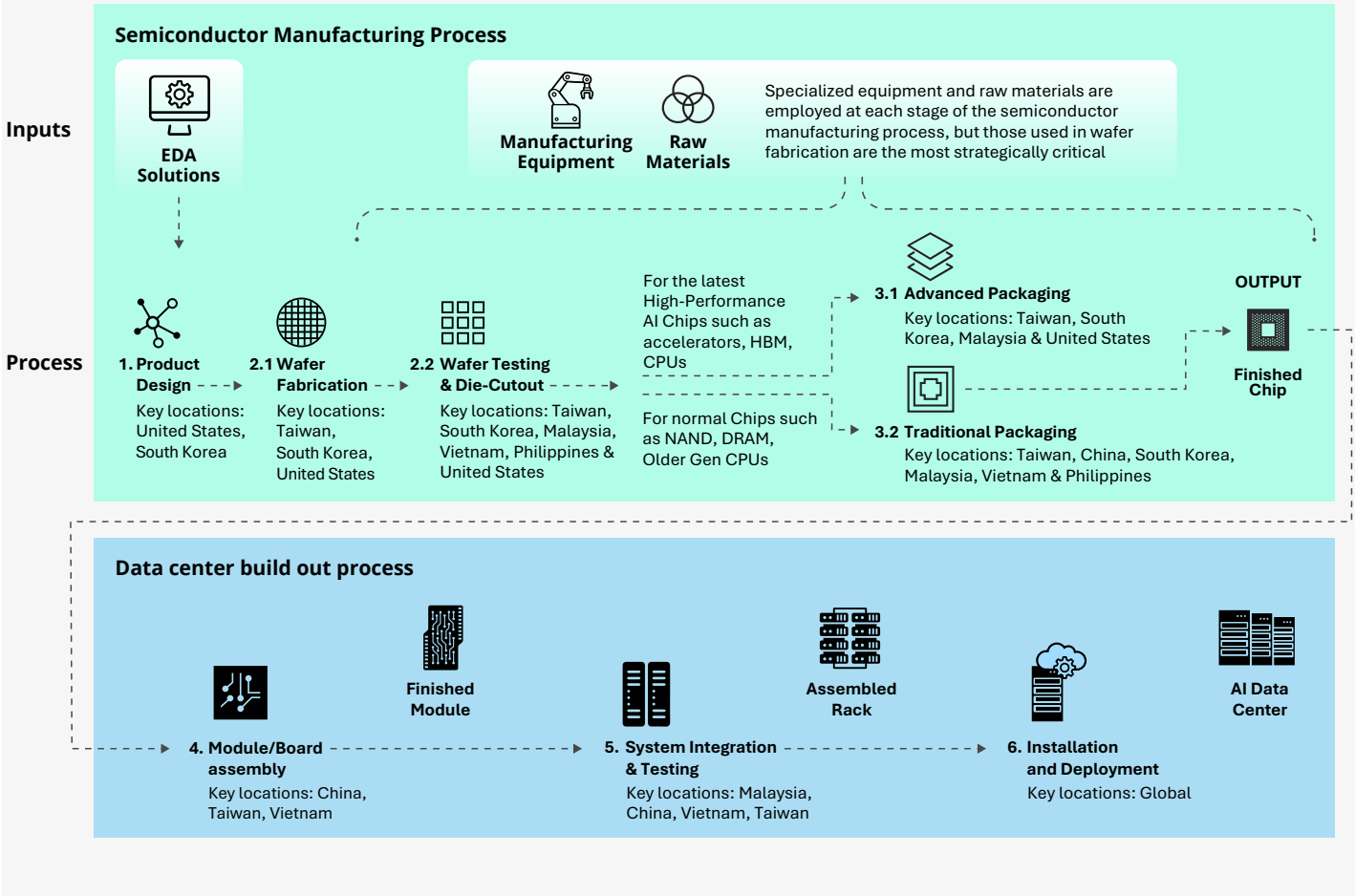
As seen in figure 20, the journey begins with semiconductor design, where engineers use electronic design automation (EDA) tools to architect and simulate integrated circuits and printed circuit boards (PCBs), ensuring functional correctness and manufacturability before fabrication. Next, in semiconductor fabrication, the designs are imprinted on silicon wafers in cleanroom environments using a variety of sophisticated raw materials and advanced equipment by processes such as lithography and etching. In the packaging stage, dies that have passed rigorous testing are interconnected using advanced 2.5D/3D techniques for high-performance chips while standard chips use conventional methods such as wire bonding. The result is a fully packaged and electrically validated chip, ready for integration into systems. Once chips pass electrical validation, they enter module and board assembly, where packaged dies



are mounted on PCBs such as DIMMs (for DRAM), SSD boards (for NAND), or compute boards (for AI Accelerators). System integration and final assembly involve combining compute, memory, storage, networking, and thermal systems into fully functional server units. The exact configuration of the units are adjusted to meet customer requirements for performance, cooling, and power delivery.

2.5D/3D techniques for high-performance chips while standard chips use conventional methods such as wire bonding. The result is a fully packaged and electrically validated chip, ready for integration into systems. Once chips pass electrical validation, they enter module and board assembly, where packaged dies are mounted on PCBs such as DIMMs (for DRAM), SSD boards (for NAND), or compute boards (for AI Accelerators). System integration and final assembly involve combining compute, memory, storage, networking, and thermal systems into fully functional server units. The exact configuration of the units are adjusted to meet customer requirements for performance, cooling, and power delivery.

Figure 20. From design to data center

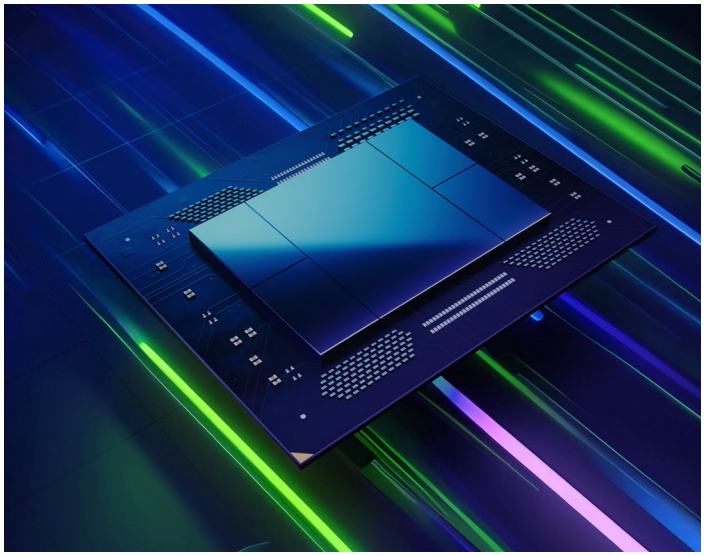


VII. Emerging Horizons in AI Infrastructure

As AI data centers mature, a series of technology shifts and market forces are reshaping how infrastructure is designed, deployed, and sustained. This section highlights the key trends that are starting to drive the next phase of growth:

Impact of edge demand

Inference demand is moving beyond centralized data centers toward hybrid models that combine cloud, edge, and on-device computing. Edge AI is reshaping compute demand by redistributing workloads from centralized data centers to decentralized environments.[29] By 2025, more than 50% of data was expected to be generated by edge devices, and many of these devices are being designed to run AI models independently without relying on cloud infrastructure.[30] This trend is reinforced by the adoption of smaller, task-specific models that can be deployed locally. Hybrid computing is becoming a core element for enterprise AI strategies, driven by the need for faster response times, reduce network dependency, and stronger privacy protections.[31] Expanding edge inference may pull segments of the market toward low-power, cost-optimized semiconductors such as embedded AI SoCs and NPUs, micro-accelerators, and specialized PMICs. This trend could also sustain demand for current-generation and mature-node chips, sensors, and connectivity silicon, as edge deployments prioritize volume, integration, and efficiency over raw performance.



Architecture and packaging transitions

Another significant shift is the move toward domain-specific architecture that tightly integrate high-performance logic, memory, and interconnects. Future AI workloads will increasingly leverage modular designs that allow heterogeneous components (i.e., "chiplets") to be co-packaged with minimal latency penalties. This architectural transition is supported by advances in 2.5D and 3D packaging, which enable tighter integration of compute and memory elements while improving power efficiency and thermal management. Photonics with co-packaged optics (CPO) is also an emerging breakthrough, embedding optical interconnects directly with switch ASICs or accelerators to deliver higher bandwidth and lower energy per bit. Optical interconnects more broadly are positioned as next-generation solutions for latency reduction and energy savings, particularly for large-scale training clusters with multi-node AI Accelerator fabrics.

Hardware–software co-design

Another key trend is the increasing integration and optimization across hardware and software stacks, driving greater efficiency for AI workloads. By controlling silicon architecture, packaging, and deployment strategies end-to-end, these organizations and hyperscalers can tune systems for specific performance, latency, or energy goals. This trend is expected to influence market dynamics by reducing reliance on off-the-shelf solutions and increasing the strategic importance of in-house semiconductor capabilities.

Process technology evolution

Semiconductors continue to advance in density and efficiency, but the pace of raw transistor scaling is slowing. The industry is shifting from FinFETs to nanosheet FET designs, which improve power control and reduce energy loss.[32] At the same time, researchers are exploring new materials such as 2D semiconductors and compound semiconductors to extend performance beyond silicon's limits. With scaling alone no longer sufficient, progress increasingly comes from system-level innovation. While 3 nm technologies are currently pushing the boundaries of transistor scaling, future nodes may incorporate new materials such as nanosheet FETs, gate-all-around (GAA) transistors, and advanced lithographic techniques like EUV with high numerical aperture (High-NA EUV). These approaches allow chips to deliver greater performance, energy efficiency, and cost-effectiveness, even as Moore's Law slows.[33]

Memory and data movement

Continued innovation in HBM and next-generation DRAM is critical. The HBM market alone is forecast to reach $21 billion by 2025, highlighting its critical role in enabling high-bandwidth compute.[34] Future iterations of HBM will feature increased stack heights, improved energy efficiency, and tighter integration with compute cores. Similarly, new memory hierarchies are expected to emerge that combine SRAM, DRAM, and persistent memory for optimized data locality and reduced data movement.

Energy and sustainability imperatives

As energy requirements for AI infrastructure increase, energy efficiency and sustainability will become more pressing concerns, prompting greater investment in power-efficient chip design, dynamic workload allocation, and intelligent cooling systems. Infrastructure providers will need to prioritize computational efficiency per watt of energy, primarily as a cost-saving measure but also as a response to global sustainability targets and energy market volatility. Operators who proactively invest in adaptable, energy-efficient systems will be better positioned to sustain long-term value as AI deployments diversify across sectors and geographies.[35]

With each new generation of data centers, AI computation is becoming significantly more energy efficient, a trend that is reshaping the economics and architecture of modern data

centers. Continued improvements in the energy efficiency of semiconductor hardware that powers AI is key to unlocking AI's potential, from compute-related chips to the chips that manage cooling, steady electricity demand, distribute and transform power, and organize workloads across the entire facility. Gains in performance per watt will allow data center operators to pack servers more densely and boost AI performance without requiring a proportional increase in energy. As chip companies develop solutions that reduce energy requirements per computing unit, a 10 MW data center with more efficient chips could eventually deliver the same processing capacity as 50 MW data centers using earlier generation semiconductor technologies. As electricity costs and grid demand rise, higher energy efficiency is no longer a 'bonus', it is the only way to deploy AI at scale.

Lastly, overall AI energy efficiency is determined not only by the silicon itself, but by the end-to-end data center system. This includes how power is brought into the facility, an area evolving at a pace comparable to chip development, as well as power conversion losses, thermal management efficiency, utilization rates, and workload placement. The architecture of power delivery into AI data centers is now changing on an almost annual basis, and in some cases even faster, collectively governing how much effective compute is ultimately extracted per watt delivered.

VIII. Conclusion

This report illustrates the diversity of inputs within the semiconductor supply chain that underpin AI infrastructure – beyond high-end processors and accelerators to signal conditioning chips, power regulators, memory dies, and control logic. As AI continues to scale, the ability to maintain a competitive semiconductor innovation system will be foundational to the continued advancement of AI. The entire semiconductor supply chain, involving U.S. semiconductor design, manufacturing, and manufacturing equipment companies, enables the buildout of AI infrastructure. In short, without semiconductors there is no AI.

To lead in this transformational technology, government and industry must work together to advance policies to accelerate semiconductor innovation – strengthening capabilities across the full spectrum of chip technologies – and work closely with global partners to build strong and resilient supply chains.



Author Bios



David Isaacs
Vice President, Government Affairs
disaacs@semiconductors.org

David Isaacs is vice president of government affairs at SIA, where he is responsible for all aspects of the association's work related to government policy.



Mary Thornton
Vice President, Global Policy
mthornton@semiconductors.org

Mary Thornton is vice president at SIA, where she is responsible for SIA's global policy and economic security agenda.



Greg LaRocca
Director of Market Research and Economic Policy
glarocca@semiconductors.org

Greg LaRocca is director of market research and economic policy at SIA, where he oversees the analysis of industry data.



Erik Hadland
Director of Technology Policy
ehadland@semiconductors.org

Erik Hadland is the director of technology policy at SIA, where he is responsible for the association's research, development, and technology activities as well as its education and workforce development efforts.



Jeroen Kusters
Principal
jekusters@deloitte.com

Jeroen Kusters is the US semiconductor leader and partner at Deloitte Consulting LLP, where he oversees the semiconductor industry practice, advising semiconductor clients across a variety of domains.



Duncan Stewart
Director
dunstewart@deloitte.ca

Duncan Stewart is the director of TMT research for Deloitte Canada and is the lead researcher on semiconductor topics for the US TMT Center and for DeloitteGlobal.



Amy Scimeca
Manager
aseiler@deloitte.com

Amy Scimeca is a manager at Deloitte Consulting LLP, supporting market strategy, value capture, and supply chain design engagements for semiconductor clients.



Karan Aggarwal
Manager
karaaggarwal@deloitte.com

Karan Aggarwal is a manager at Deloitte Consulting LLP, where he leads large-scale strategic and enterprise transformation projects across US and Asian semiconductor clients.



Pranav Salhan
Consultant
psalhan@deloitte.com

Pranav Salhan is a consultant at Deloitte Consulting LLP, supporting semiconductor clients on operations optimization, supply planning, and enterprise transformation initiatives.

Acknowledgments

We appreciate the invaluable contributions of SIA member companies and subject matter experts who provided feedback that informed this report.

This report would not have been possible without the contributions of our SIA colleagues Aaron Woolf, Alex Gordon, Emma Rafaelof, and Molly O'Leary.

Appendix

Methodology for semiconductor content value analysis (CVA)

CVA quantifies the relative semiconductor content value of an AI data center server setup by evaluating and averaging key hardware components costs across representative system configurations. It is designed to offer insight into how compute, memory, networking, and power infrastructure contribute to the total semiconductor cost within high-performance AI server racks.

Server configuration selection

The CVA is based on three representative AI server rack setups:

- OEM turnkey system: A vertically integrated configuration optimized for performance density, featuring 72 top-tier AI accelerators and custom interconnects.
- Third-party system integrator setup 1: A modular high-performance system with 64 high-bandwidth AI Accelerators (same as OEM set up), emphasizing modularity and power scaling.
- Third-party system integrator setup 2: A mid-tier modular configuration using 32 accelerators across interconnected chassis, representing a performance-efficient, cost-sensitive deployment.

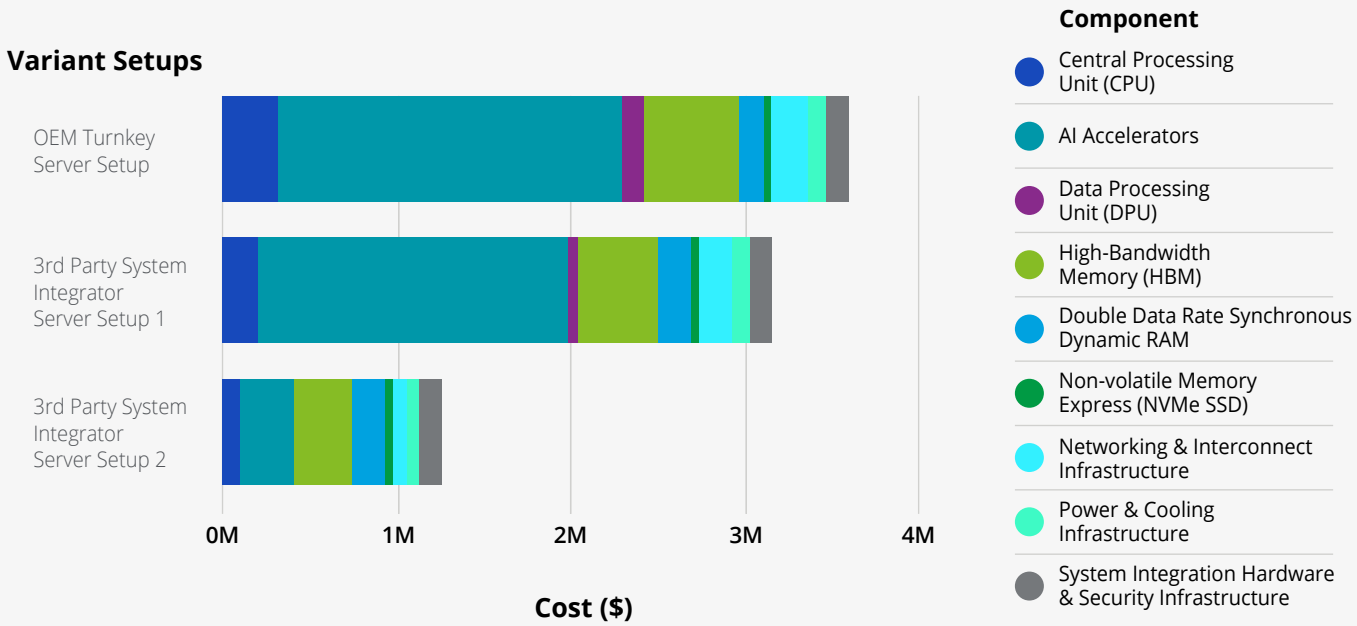
Each setup adheres to a standardized 48U server rack format and is provisioned with 200 TB of storage for comparability. Compute performance is measured using FP16 peak throughput (PFLOPS) to capture theoretical compute capacity for AI workloads. Costs, power draw, and component counts are normalized based on publicly available product specs and internal teardown benchmarks.

Each component category is assigned a value share based on its bill-of-materials (BOM) contribution in the three reference systems. These shares are validated using teardown reports, industry disclosures, and known silicon pricing ranges at volume scale.

Analytical approach and averaging

The value breakdown shown in figure 22 represents an average semiconductor value breakdown across the three server setups discussed above. This approach reflects both OEM turnkey system and third-party system integrator architectures, offering a balanced perspective on how component selection influences total cost and value concentration. We have provided the value breakdown for each of the configurations separately in figure [21].

Figure 22: Value breakdown from reference racks



Glossary

Abbreviation	Definition (AI Data Center Context)
2.5D	2.5 Dimension — advanced chip packaging combining multiple dies on a silicon interposer to enhance AI compute performance.
3D	3 Dimension — refers to stacked chip architectures or 3D packaging used to increase density and efficiency in AI hardware.
Advanced Packaging	Techniques like 2.5D and 3D integration to combine multiple chips for performance.
AI Data Center	A facility designed to support compute-intensive AI tasks like training and inference at scale.
ASIC	Application-Specific Integrated Circuit — custom chips designed to accelerate specific AI tasks, widely used in hyperscale AI data centers.
BMCs	Baseboard Management Controllers — embedded controllers for monitoring server health and performance in AI data centers.
CDUs	Coolant Distribution Units — key elements in liquid cooling systems uased in AI data centers for thermal efficiency.
Chiplet	A modular semiconductor unit integrated with others to form a complete chip system.
CPUs	Central Processing Units — used in AI data centers for orchestrating and managing AI workloads and supporting traditional tasks.
DRAM	Dynamic Random Access Memory — commonly used in AI data centers for temporary data storage during model execution.
EDA	Electronic Design Automation — tools used in designing AI chips for data centers.
Edge AI	Running AI models on local devices to reduce latency and dependency on cloud infrastructure.
EUV	Extreme Ultraviolet — lithography technique enabling high-performance chips used in AI compute nodes.
Fab	Fabrication — refers to semiconductor manufacturing, foundational to producing AI processors used in data centers.
Fabless	A company that designs chips but outsources their manufacturing.
FETs	Field Effect Transistors — fundamental components in chips powering AI accelerators in modern data centers.
Foundry	A semiconductor fabrication facility that manufactures chips based on third-party designs.
FPGAs	Field-Programmable Gate Arrays — reconfigurable chips enabling efficient, task-specific acceleration in AI data centers.
GAA	Gate-All-Around — advanced transistor architecture enhancing performance and energy efficiency in AI hardware.
GPUs	Graphics Processing Units — a subclass of AI Accelerators.
HBM	High Bandwidth Memory — memory type that supports high-speed AI training workloads in modern data center chips.

Abbreviation	Definition (AI Data Center Context)
I/O	Input / Output — essential to AI data processing, impacting memory bandwidth and throughput in data centers.
Inference	Applying a trained AI model to new data for predictions or decisions, typically in real time.
Interposer	A layer connecting multiple chips in advanced semiconductor packaging.
MCUs	MicroController Units — embedded systems that help manage environmental controls and monitoring within data centers.
NAND	NOT AND — a type of non-volatile flash memory widely used in data center storage systems.
NICs	Network Interface Cards — hardware enabling servers in AI data centers to communicate efficiently over high-speed networks.
nm	Nanometer — unit of measurement for semiconductor technology nodes; smaller sizes improve performance and energy efficiency in AI chips.
NVMe	Non-Volatile Memory Express — interface protocol that improves SSD access speed, critical for large data handling in AI workloads.
OEM	Original Equipment Manufacturer — companies producing AI data center equipment under another firm's brand.
PCBs	Printed Circuit Boards — foundational to assembling AI server components in data center infrastructure.
PCIe	Peripheral Component Interconnect Express — a high-speed interface standard for connecting GPUs and accelerators in AI data center servers.
Photonic Computing	Computing using light signals for faster data transfer and lower power consumption.
PHY	PHYsical interface — layer that enables high-speed connectivity between AI chips and memory or interconnects in data centers.
PMICs	Power Management Integrated Circuits — manage voltage and power efficiency of AI chips within data centers.
RAM	Random Access Memory — essential for holding working data and models during training and inference in AI data centers.
SoC	System-on-Chip — an integrated circuit that consolidates multiple functions onto a single die, increasingly used in AI accelerators and edge AI to improve performance, power efficiency, and system integration.
SRAM	Static Random Access Memory — high-speed cache used in AI chips to store intermediate model weights or data.
Training (AI)	The phase where AI models learn patterns from large datasets using computational resources.
TSVs	Through-Silicon Vias — vertical interconnects enabling 3D chip stacking, improving performance and density for AI compute modules.
VRMs	Voltage Regulator Modules regulate power delivery to processors in AI servers for stable operation.

References

1. Krystal Hu, "ChatGPT sets record for fastest-growing user base—analyst note," Reuters, February 2, 2023.
2. Fatima Hameed Khan, Muhammad Adeel Pasha, and Shahid Masud, "Advancements in microprocessor architecture for ubiquitous AI—an overview on history, evolution, and upcoming challenges in AI implementation," Micromachines 12, no. 6 (2021): p. 665.
3. Jesse Noffsinger et al., "The cost of compute: A $7 trillion race to scale data centers," McKinsey Quarterly, April 28, 2025.
4. TrendForce News, "Foundry giants' advanced node expansion efforts beyond 2025: TSMC, Intel, Rapidus and more," January 13, 2025.
5. World Semiconductor Trade Statistics, 2025 Product Classification Guide, December 11, 2024.
6. Ondrej Burkacky et al., "Generative AI: The next S-curve for the semiconductor industry?," McKinsey & Company, March 29, 2024.
7. SK Hynix, "HBM technology: The silent catalyst of the AI revolution," January 30, 2024.
8. Intel, "Field programmable gate arrays (FPGAs) for artificial intelligence (AI)," accessed August 23, 2025.
9. Future Market Insights, FPGA market report, August 23, 2025.
10. Manuel Mota, "Die-to-die connectivity with high-speed SerDes PHY IP," Semiconductor Engineering, March 12, 2020.
11. Matteo Buffolo et al., "Review and outlook on GaN and SiC power devices: Industrial state-of-the-art, applications, and perspectives," IEEE Transactions on Electron Devices 71, no. 3 (March 2024): pp. 1344-55.
12. Nidhi Govil, "Nvidia unveils 800V high-voltage DC power system for next-gen AI data centers," The Outpost, May 28, 2025.
13. Amr Elmeleegy, "NVIDIA contributes NVIDIA GB200 NVL72 designs to Open Compute Project," NVIDIA Developer Blog, October 15, 2024.
14. Duncan Stewart et al., "Silicon photonics: Gen AI communicates at lightspeed," Deloitte, November 19, 2024.
15. World Semiconductor Trade Statistics (WSTS), "H2-2025 Forecast," December 2, 2025
16. Jim Rowan et al., "State of Generative AI in the Enterprise: Quarter four report," Deloitte, January 2025.
17. CloudSyntrix, "The impact of Generative AI on hyperscalers, chipmakers, and inference workload," February 10, 2025.
18. SambaNova Systems, "SN40L RDU AI Chip," accessed October 2025.
19. Qualcomm, "Cloud AI 100 Ultra," Qualcomm Technologies, accessed October 2025
20. David Crawford, Jue Wang, Roy Singh, "AI's Trillion-Dollar Opportunity," Tech Report 2024, Bain & Company, 2024
21. Intel Corporation, "Gaudi AI Accelerator Products," Intel, accessed October 2025,
22. Amazon Web Services (AWS), "AWS Trainium," accessed August 23, 2025.
23. Vivian Lee et al., "Breaking barriers to data center growth," Boston Consulting Group, January 20, 2025.
24. Walters et al., "Rethinking AI demand part 1: AI data centers are experiencing a surge of training demand—what happens when the surge is over?," Alvarez & Marsal," February 25, 2025.
25. McKinsey & Company, "Hiding in plain sight: The underestimated size of the semiconductor industry," January 16, 2026.
26. SIA/WSTS, "Global Annual Semiconductor Sales Increase 25.6% to $791.7 Billion in 2025," Feb. 7, 2025.
27. Jesse Noffsinger et al., "The cost of compute: A $7 trillion race to scale data centers," McKinsey Quarterly, April 28, 2025.
28. Uvation, "Why AI server cost per user is the new metric that matters," July 4, 2025.
29. Josh Schneider and Ian Smalley, "What is a neural processing unit (NPU)?," IBM, accessed August 23, 2025.
30. Adita Agrawal, "The convergence of edge computing and 5G," Control Engineering, August 7, 2023; Baris Sarer et al., "AI and the evolving consumer device ecosystem," CIO Journal by Deloitte for The Wall Street Journal, April 24, 2024.
31. Chris Thomas et al. "Is your organization's infrastructure ready for the new hybrid cloud?," Deloitte Insights, June 30, 2025.
32. Murray Slovick, "From FinFETs to Nanosheets: ICs evolve to keep pace with 'Moore's Law'," The IP&E Specialist, July 6, 2021.
33. Slovick, "From FinFETs to Nanosheets: ICs evolve to keep pace with 'Moore's Law'," July 6, 2021.
34. Vishnu Kumar, "Global semiconductor market to grow 14% in 2025, Gartner report indicates," Circuit Digest, October 29, 2024.
35. Walters et al., "Rethinking AI demand part 1: AI data centers are experiencing a surge of training demand—what happens when the surge is over?" February 25, 2025.
36. Martin Stansbury et al., "Can US infrastructure keep up with the AI economy?," Deloitte Insights, June 24, 2025.



About SIA

The Semiconductor Industry Association (SIA) is the voice of the semiconductor industry, one of America's top export industries and a key driver of America's economic strength, national security, and global competitiveness. SIA represents 99% of the U.S. semiconductor industry by revenue and nearly two-thirds of non-U.S. chip firms. Through this coalition, SIA seeks to strengthen leadership of semiconductor manufacturing, design, and research by working with Congress, the Administration, and key industry stakeholders around the world to encourage policies that fuel innovation, propel business, and drive international competition. Learn more at www.semiconductors.org.

Deloitte.

About Deloitte

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee, and its network of member firms, each of which is a legally separate and independent entity. Please see www.deloitte.com/about for a detailed description of the legal structure of Deloitte Touche Tohmatsu Limited and its member firms. Please see www.deloitte.com/us/about for a detailed description of the legal structure of Deloitte LLP and its subsidiaries. Certain services may not be available to attest clients under the rules and regulations of public accounting.

This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.

Copyright © 2026 Deloitte Development LLC. All rights reserved.