AI-01796 Datacenter Hardware Infrastructure

Owner	Ⓡ Rigel Arcayan
Tags	Investing
Created time	@July 7, 2025 8:50 AM

In AI Datacenters, list the top 5 Server vendors. What are the models they are using? What are components used? What type of hard drives they used and manufacturers name of hard drive.

In the rapidly evolving landscape of AI datacenters, server vendors are crucial in providing the specialized hardware needed for compute-intensive AI workloads. Here's a breakdown of the top 5 server vendors, their models, components, and hard drive types:

Top 5 Server Vendors for AI Datacenters

- NVIDIA (indirectly as a platform provider): While not a traditional server vendor, NVIDIA's HGX platform and GPUs are foundational to almost all high-performance AI servers. They often partner with other server manufacturers to deliver complete solutions.
- Dell Technologies: A major player in enterprise IT, Dell offers a strong portfolio of AI-optimized PowerEdge servers.
- Hewlett Packard Enterprise (HPE): HPE provides a comprehensive range of AI servers, often leveraging their expertise in high-performance computing (HPC) and large-scale deployments.
- Supermicro: Known for its "building block" approach, Supermicro offers a wide variety of highly customizable and GPU-dense server solutions for AI.
- Lenovo: Lenovo has been increasing its presence in the AI server market, offering a range of ThinkSystem servers designed for AI workloads.
- Cisco: Cisco is also emerging as a significant player with their UCS servers tailored for AI, emphasizing integrated, secure, and scalable solutions.

Server Models, Components, and Hard Drives
AI servers are characterized by their focus on accelerated computing, primarily driven by powerful GPUs.

1. NVIDIA (as a platform provider)

- Models/Platforms: NVIDIA HGX platforms (e.g., HGX H100, HGX B200, GB200 NVL72, GB300 NVL72) are the core. These are then integrated into servers by other vendors.
- Key Components:
    - GPUs: NVIDIA A100, H100, H200, B200 (Blackwell), and Grace Hopper Superchips (e.g., GH200). These are the primary compute engines for AI.
    - Interconnect: NVLink and NVSwitch for high-speed, low-latency communication between GPUs within a server or across multiple servers.
    - DPUs (Data Processing Units): NVIDIA BlueField DPUs for offloading networking and security tasks.
    - Networking: NVIDIA Spectrum-X Ethernet platform or InfiniBand for high-bandwidth, low-latency cluster connectivity.
    - Processors (CPUs): Often paired with high-core count Intel Xeon Scalable or AMD EPYC processors.
    - Memory: High-bandwidth memory (HBM) on GPUs, and large capacities of DDR5 RAM for the CPUs.
- Hard Drives: Given the focus on GPU-intensive compute, storage is typically for OS, applications, and frequently accessed datasets.
    - Types: High-performance NVMe SSDs (e.g., PCIe Gen4/Gen5 U.2 or EDSFF) are preferred for their speed and low latency.
    - Manufacturers: Samsung, Kioxia (formerly Toshiba Memory), Micron, Western Digital (via SanDisk/Ultrastar), Seagate.

1. Dell Technologies

- Models: Dell PowerEdge XE9680, PowerEdge R760xa, and other PowerEdge servers optimized for AI and GPU acceleration.
- Key Components:

    - GPUs: Primarily NVIDIA GPUs (H100, A100, etc.), with support for multiple double-wide GPUs.
    - Processors (CPUs): Latest generation Intel Xeon Scalable or AMD EPYC processors.
    - Memory: DDR5 RAM with high capacities.
    - Storage: Support for a mix of hot-swap NVMe, SAS, and SATA drives.
    - Networking: High-speed Ethernet adapters (100GbE, 200GbE, 400GbE) from vendors like Broadcom, Intel, and NVIDIA (Mellanox).
- Hard Drives:
    - Types: Primarily NVMe SSDs for fast data access, with some use of SAS/SATA SSDs and high-capacity HDDs for bulk storage or less performance-critical data.
    - Manufacturers: Samsung, Kioxia, Micron, Western Digital, Seagate.

1. Hewlett Packard Enterprise (HPE)

- Models: HPE Compute XD690, HPE ProLiant Compute DL380a Gen12, HPE ProLiant Compute DL384 Gen12, HPE ProLiant Compute XD685, HPE Cray supercomputing systems.
- Key Components:
    - GPUs: NVIDIA H100, H200, B200, Grace Hopper Superchips. HPE often integrates NVIDIA HGX platforms. Support for AMD Instinct accelerators is also common.
    - Processors (CPUs): Intel Xeon Scalable or AMD EPYC processors.
    - Memory: Large capacities of DDR5 RAM.
    - Storage: EDSFF (Enterprise & Datacenter SSD Form Factor) drives, SFF (Small Form Factor) drives, with a strong emphasis on NVMe.
    - Cooling: Advanced cooling solutions, including direct-liquid cooling, for high-density GPU deployments.
    - Networking: High-speed InfiniBand or Ethernet.
- Hard Drives:

    - Types: NVMe SSDs (E1.S, U.2, 2.5" form factors) for performance, and potentially high-capacity SAS/SATA HDDs for large datasets (e.g., in HPE GreenLake for File Storage solutions).
    - Manufacturers: Samsung, Kioxia, Micron, Western Digital, Seagate.

1. Supermicro

- Models: Supermicro offers a vast array of GPU servers across various form factors (1U, 2U, 4U, 5U, 8U, 10U), including Universal GPU Systems, Liquid-Cooled Universal GPU Systems, and NVIDIA MGX™ Systems.
- Key Components:
    - GPUs: Supports a wide range of GPUs including NVIDIA HGX H100/H200/B200, AMD Instinct MI325X/MI300X/MI250, and Intel Data Center GPU Max Series. They can often accommodate up to 10 double-width PCIe GPUs.
    - Processors (CPUs): Intel Xeon Scalable or AMD EPYC processors.
    - Memory: High DIMM counts (up to 32 DIMMs) for substantial memory capacity (up to 9TB or 12TB).
    - Storage: Up to 24 hot-swap E1.S, U.2, or 2.5" NVMe/SATA drives.
    - Cooling: Offer liquid-cooled systems for high-density deployments.
- Hard Drives:
    - Types: Primarily NVMe SSDs (U.2, E1.S, 2.5"), with options for SAS/SATA SSDs and HDDs.
    - Manufacturers: Supermicro often lists compatible drives from leading manufacturers like Seagate, HGST (a Western Digital brand), and Toshiba.

1. Lenovo

- Models: Lenovo ThinkSystem SR675 V3, SR680a V3, SD650-N V3, and other ThinkSystem servers optimized for AI and HPC.
- Key Components:
    - GPUs: Support for multiple NVIDIA GPUs (e.g., H100, A100) and other accelerators.

    - Processors (CPUs): Latest generation Intel Xeon Scalable or AMD EPYC processors.
    - Memory: High-capacity DDR5 memory.
    - Cooling: Strong focus on liquid cooling solutions, given their experience in supercomputing.
- Hard Drives:
    - Types: NVMe SSDs are standard for performance-critical storage, with SAS/SATA SSDs and HDDs available for other storage tiers.
    - Manufacturers: Common enterprise drive manufacturers like Samsung, Kioxia, Micron, Western Digital, Seagate.

1. Cisco

- Models: Cisco UCS C845A M8 Rack Server, UCS C885A M8 Rack Server.
- Key Components:
    - GPUs: Primarily NVIDIA GPUs, with a focus on high-density configurations (e.g., up to eight GPUs in a 4RU chassis). They leverage NVIDIA MGX modular reference design.
    - Processors (CPUs): Intel Xeon Scalable or AMD EPYC processors.
    - Memory: High-capacity RAM.
    - Networking: Tightly integrated with Cisco's own networking solutions, including high-speed Ethernet.
    - Cooling: Enhanced power delivery and optimized thermal management.
- Hard Drives:
    - Types: NVMe SSDs for performance, often with robust internal storage options.
    - Manufacturers: Industry-standard enterprise drive manufacturers.
      General Hard Drive Types and Manufacturers for AI Servers:
      Types:

- NVMe SSDs (Non-Volatile Memory Express Solid State Drives): These are the predominant choice for AI servers due to their extremely high speed, low

latency, and high IOPS (Input/Output Operations Per Second). They connect directly to the PCIe bus, bypassing traditional SATA/SAS bottlenecks. Form factors include U.2, M.2, and increasingly, EDSFF (Enterprise & Datacenter SSD Form Factor) for higher density and better thermal performance.

- SAS SSDs (Serial Attached SCSI Solid State Drives): Offer better performance and reliability than SATA SSDs, and are often used for enterprise-grade storage where NVMe isn't strictly necessary or for balancing cost and performance.
- SATA SSDs (Serial ATA Solid State Drives): Less common for primary AI data storage due to lower performance compared to NVMe and SAS, but might be used for boot drives or less performance-sensitive data.
- HDDs (Hard Disk Drives): While not ideal for the immediate access of AI training data, high-capacity enterprise HDDs (e.g., 7200 RPM SAS or SATA) are still used for large-scale data lakes, archival storage, and less frequently accessed datasets that feed into AI pipelines.
  Manufacturers:
- SSDs:
  - Samsung: A leading manufacturer of enterprise NVMe SSDs.
  - Kioxia (formerly Toshiba Memory): Strong in enterprise NVMe SSDs.
  - Micron: Produces a wide range of enterprise SSDs, including NVMe and SAS.
  - Western Digital (including SanDisk and Ultrastar brands): Offers a comprehensive portfolio of enterprise SSDs and HDDs.
  - Intel: Known for its Optane SSDs (though their future in this segment is evolving) and other NVMe SSDs.
- HDDs:
  - Seagate: A major global manufacturer of enterprise HDDs (e.g., Exos series).
  - Western Digital (including Ultrastar): Another dominant player in enterprise HDDs.

- Toshiba: Produces enterprise-grade HDDs.

List all peripheral components and manufacturers

In AI datacenters, servers are just one part of a larger ecosystem. "Peripheral components" in this context refer to the essential hardware elements that enable the server to function, communicate, and be managed within the datacenter environment. Here's a comprehensive list of these components and their typical manufacturers:
I. Core Server Components (Internal, but often modular and sourced from specialized manufacturers)
While CPUs, GPUs, and RAM were covered previously as primary components, it's worth reiterating their manufacturers as they are the very heart of the server.

- CPUs (Central Processing Units):
  - Intel: Xeon Scalable Processors (e.g., Emerald Rapids, Granite Rapids)
  - AMD: EPYC Processors (e.g., Genoa, Bergamo, Turin)
  - NVIDIA: Grace CPU (part of Grace Hopper and Grace Blackwell Superchips) - designed for highly integrated AI systems.
- GPUs (Graphics Processing Units / Accelerators):
  - NVIDIA: A100, H100, H200, B200 (Blackwell), Grace Hopper Superchip (GH200), Grace Blackwell Superchip (GB200)
  - AMD: Instinct MI series (e.g., MI300X, MI300A, MI325X)
  - Intel: Gaudi (from Habana Labs, acquired by Intel), Data Center GPU Max Series (e.g., Max 1100, Max 1350, Max 1550)
- RAM (Random Access Memory):
  - DDR5 SDRAM Modules: Crucial (Micron), Samsung, Hynix (SK Hynix), Kingston, G.Skill
  - HBM (High Bandwidth Memory): Integrated directly onto GPUs, manufactured by companies like Samsung, SK Hynix, and Micron.
    II. Storage Components
- SSDs (Solid State Drives):

    - NVMe SSDs (U.2, EDSFF, M.2): Samsung, Kioxia (formerly Toshiba Memory), Micron, Western Digital (including SanDisk and Ultrastar brands), Intel.
    - SAS SSDs: Samsung, Kioxia, Micron, Western Digital, Seagate.
- HDDs (Hard Disk Drives) for Bulk Storage:
    - Enterprise SAS/SATA HDDs: Seagate (Exos series), Western Digital (Ultrastar series), Toshiba.
- RAID Controllers (Redundant Array of Independent Disks):
    - Manufacturers: Broadcom (MegaRAID series, often rebranded by server vendors), Microchip Technology (Adaptec series), Marvell, Intel.
- Storage Backplanes/Caddies: Often proprietary to the server vendor (Dell, HPE, Supermicro, Lenovo, Cisco).
  III. Networking Components
- NICs (Network Interface Cards) / Ethernet Adapters:
    - Ethernet: Intel (Ethernet Network Adapters), Broadcom (NetXtreme-E series, often rebranded), NVIDIA (Mellanox ConnectX series), Marvell (LiquidIO). Speeds commonly 10GbE, 25GbE, 100GbE, 200GbE, 400GbE.
    - InfiniBand HBAs (Host Bus Adapters): NVIDIA (Mellanox InfiniBand series) - crucial for high-performance computing (HPC) and large-scale AI clusters due to extremely low latency and high bandwidth.
- DPUs (Data Processing Units):
    - NVIDIA: BlueField series (BlueField-3, BlueField-4) - offload networking, storage, and security tasks from the CPU, enhancing efficiency for AI workloads.
- Network Switches:
    - Ethernet Switches: Cisco (Nexus series, Catalyst series), Arista Networks, Juniper Networks, Mellanox (now NVIDIA Networking), Dell (PowerSwitch series), HPE (Aruba series).
    - InfiniBand Switches: NVIDIA (Mellanox Quantum series).
- Transceivers (SFP+, QSFP+, OSFP, etc.) and Cables:

- Manufacturers: Finisar (now Coherent), Lumentum, Broadcom (Avago), Sumitomo Electric, Molex, Amphenol. Many generic options available from various suppliers.
  IV. Power and Cooling Components

- PSUs (Power Supply Units):
  - Manufacturers: Delta Electronics, Lite-On Technology, AcBel Polytech, Seasonic, Supermicro (often designs their own). Server vendors typically specify and integrate these. Redundant PSUs are standard.
- Cooling Systems:
  - Fans: Delta Electronics, Nidec, Sunon, Noctua (less common in datacenter, more for specific high-end servers/workstations).
  - Heatsinks: Often custom-designed by server vendors for their specific CPU/GPU configurations.
  - Liquid Cooling Components (Cold Plates, Manifolds, CDUs - Coolant Distribution Units): Asetek, CoolIT Systems, Liebert (Vertiv), Swagelok (fittings), various specialized manufacturers for direct-to-chip or immersion cooling.
- Rack PDUs (Power Distribution Units):
  - Manufacturers: APC (Schneider Electric), Vertiv (Liebert), Eaton, Raritan (Legrand).
    V. Management and Security Components
- BMC (Baseboard Management Controller) / iDRAC / iLO / XClarity Controller:
  - Integrated Circuits: Aspeed Technology (dominant in BMC chips, e.g., AST2500, AST2600), Nuvoton.
  - Firmware/Software: Developed by the server vendors themselves (e.g., Dell iDRAC, HPE iLO, Lenovo XClarity Controller, Supermicro IPMI).
- TPM (Trusted Platform Module):
  - Manufacturers: Infineon, STMicroelectronics, Nuvoton.
- BIOS/UEFI Firmware:

- Developers: AMI (American Megatrends Inc.), Insyde Software, Phoenix Technologies. Server vendors then customize and integrate.
  VI. Chassis and Enclosure Components

- Server Chassis/Cases: Primarily designed and manufactured by the server vendors themselves (Dell, HPE, Supermicro, Lenovo, Cisco) to accommodate specific component layouts and cooling requirements. Supermicro, in particular, is known for its highly customizable "building block" chassis designs.
- Rack Rails: Custom-designed by server vendors for easy installation into standard datacenter racks.
  This comprehensive list highlights that AI datacenters are complex systems relying on a vast ecosystem of specialized components and manufacturers, each contributing to the overall performance, efficiency, and reliability of the AI infrastructure.

Summarize our conversation via a table and list the manufacturers and products they are using. No need to explain the components. I just need a list

Category	Manufacturers	Products
Server Vendors	NVIDIA (as platform provider)	HGX platforms (HGX H100, HGX B200, GB200 NVL72, GB300 NVL72)
	Dell Technologies	PowerEdge XE9680, PowerEdge R760xa, other PowerEdge servers
	Hewlett Packard Enterprise (HPE)	HPE Compute XD690, HPE ProLiant Compute DL380a Gen12, HPE ProLiant Compute DL384 Gen12, HPE ProLiant Compute XD685, HPE Cray supercomputing systems
	Supermicro	Universal GPU Systems, Liquid-Cooled Universal GPU Systems, NVIDIA MGX™ Systems (various 1U, 2U, 4U, 5U, 8U, 10U GPU servers)
	Lenovo	ThinkSystem SR675 V3, SR680a V3, SD650-N V3, other ThinkSystem servers
	Cisco	UCS C845A M8 Rack Server, UCS C885A M8 Rack Server

Category	Manufacturers	Products
GPUs	NVIDIA	A100, H100, H200, B200 (Blackwell), Grace Hopper Superchips (e.g., GH200)
	AMD	Instinct MI325X/MI300X/MI250
	Intel	Data Center GPU Max Series
CPUs	Intel	Xeon Scalable processors
	AMD	EPYC processors
SSDs	Samsung, Kioxia, Micron, Western Digital (SanDisk/Ultrastar)	NVMe SSDs (PCIe Gen4/Gen5 U.2, EDSFF, M.2, E1.S, 2.5"), SAS SSDs, SATA SSDs
HDDs	Seagate, Western Digital (Ultrastar), Toshiba	Enterprise HDDs (e.g., Seagate Exos series, Western Digital Ultrastar series)
Networking	NVIDIA (Mellanox)	Spectrum-X Ethernet platform, InfiniBand
	Broadcom, Intel, Cisco	High-speed Ethernet adapters (100GbE, 200GbE, 400GbE)
DPUs	NVIDIA	BlueField DPUs
Interconnect	NVIDIA	NVLink, NVSwitch