

How chips are shaping data centre requirements, strategies, and markets

This article explores the key chipsets that make up modern data centres and the market dynamics shaping their design, production and supply, as workloads become more specialised.

Jonas Topp-Mugglestone, Consultant

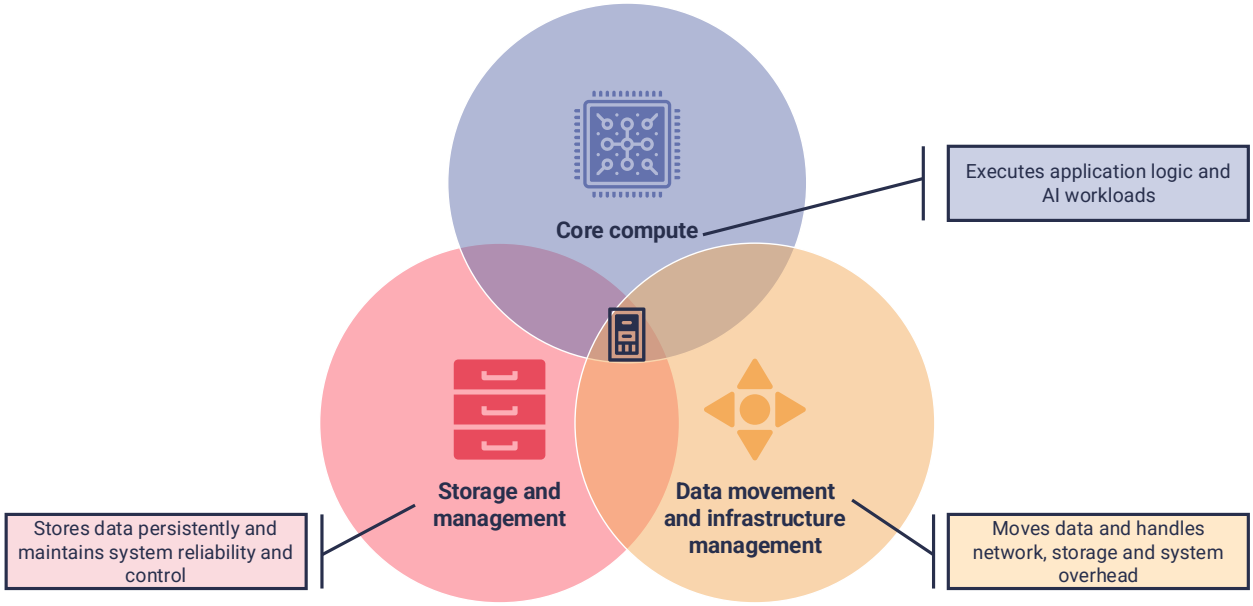
December 2025

© STL Partners

What chips sit inside a data centre today?

Broadly, modern data centres rely on three classes of silicon, each serving a distinct role in the compute stack:

1. Core compute: General-purpose processors, and increasingly specialised accelerators, that execute the majority of application logic.
2. Storage and management: Chips that store and protect data, maintain system integrity and deliver essential control functions, ensuring the persistence, reliability and secure operation of data centres.
3. Data movement and infrastructure offload: Chips that move data and support compute by handling network traffic, security functions and other system overheads, reducing the burden on core processors.



Core compute

CPUs (Central processing units)

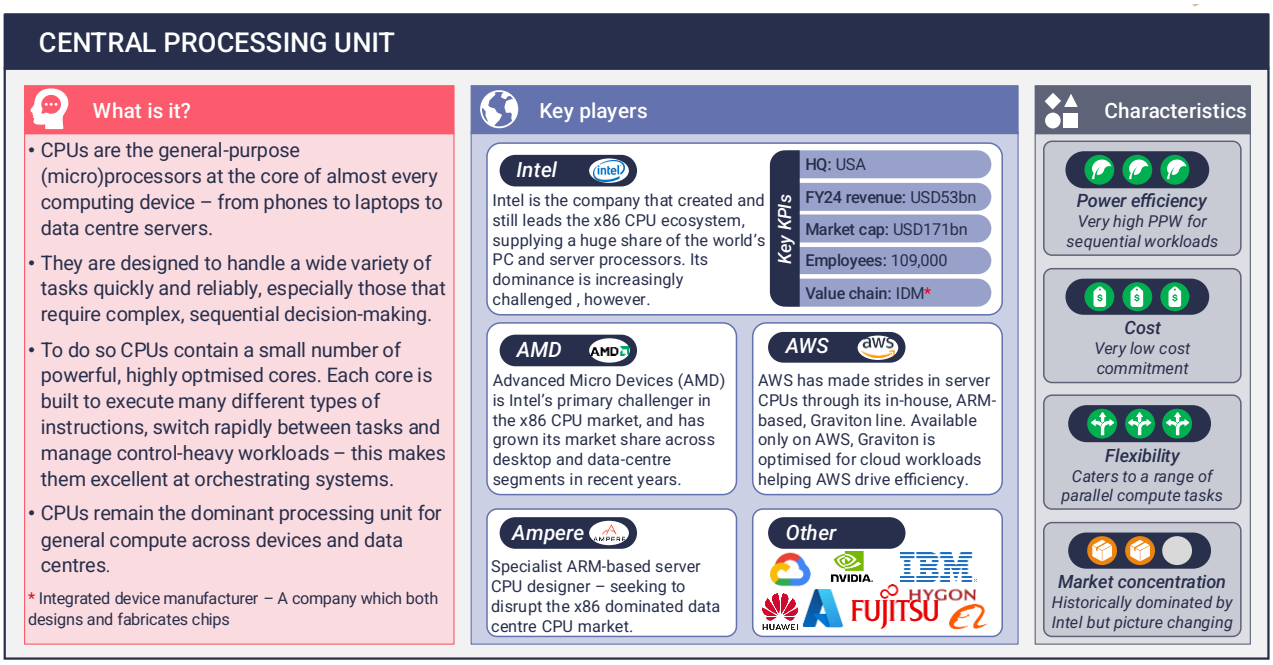
CPUs are the general-purpose processors that provide broad control and execution capability across the data centre. They can run many types of instructions and manage complex logic, making them suitable for tasks that are not purely numerical or require frequent decision-making.

Typical workloads include:

- Operating systems
- Web and application serving
- Virtual machines

© STL Partners

- Pre-and-post processing tasks around AI models



GPUs (Graphics processing units)

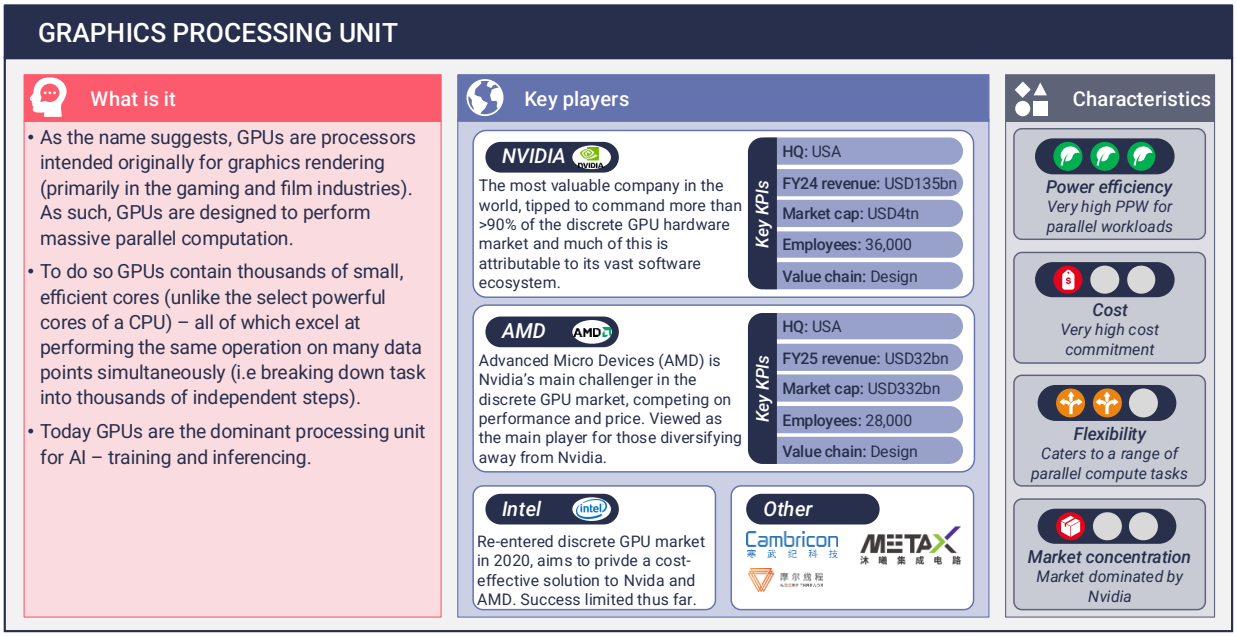
GPUs are general-purpose compute and were originally developed to accelerate graphics rendering, particularly the mathematical operations required to draw and manipulate 3D images in real time for games and visual applications. Their architecture – designed to execute many identical calculations in parallel –made them efficient for transforming and shading millions of pixels per second.

Over time, this parallel design proved extremely effective for matrix and vector computations, which are central to machine learning. As a result, GPUs shifted from being graphics-only devices to becoming the primary engines for AI training and AI inference. Today, discrete GPUs sit at the centre of high-performance computing and AI data centres.

Typical workloads include:

- Training AI models
- AI inference of large models
- Scientific simulations and analytics
- Video processing and rendering

© STL Partners



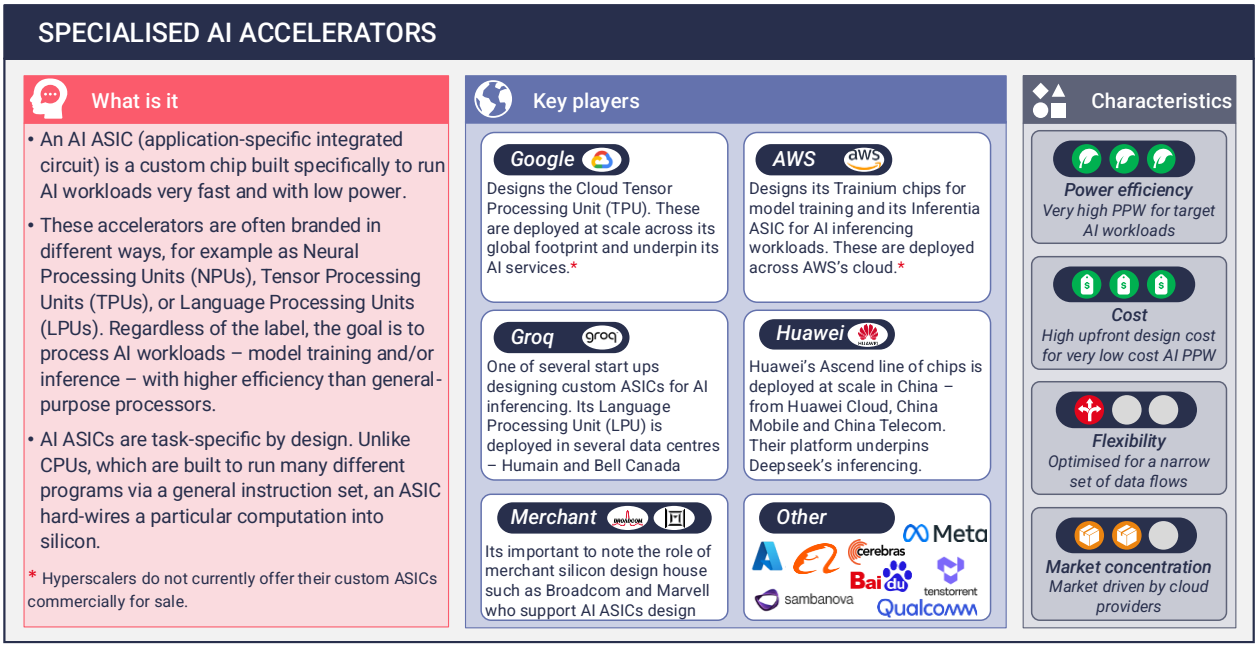
AI ASICS

Custom ASICs (Application-Specific Integrated Circuits) are processors designed for a specific workload rather than general-purpose use. By implementing only the required logic, they offer higher efficiency and lower power consumption for that workload. In modern data centres, ASICs are increasingly important as operators seek cost-efficient and power-efficient alternatives to general-purpose compute (such as CPUs and GPUs).

Typical workloads:

- AI inference acceleration
- Networking and packet forwarding (switch and router ASICs)
- Storage management (SSD controllers for wear-levelling and error correction)

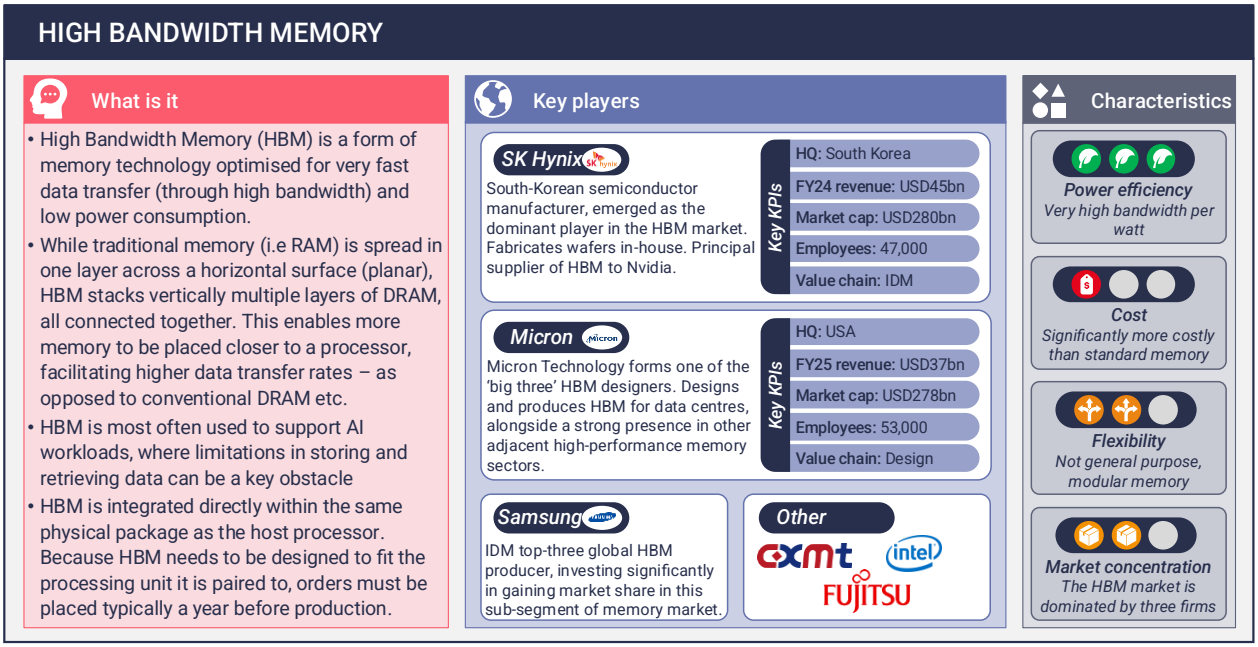
© STL Partners



Memory

High Bandwidth Memory (HBM) is a type of stacked DRAM (Dynamic Random-Access Memory (DRAM) – the main type of working memory used in servers and computing systems) designed to deliver much higher bandwidth than conventional memory, by placing multiple memory dies very close to the processor and connecting them with wide, high-speed interfaces. While conventional DRAM still dominates the overall memory market, AI workloads place far greater emphasis on memory bandwidth than traditional enterprise applications. Rather than serving as general-purpose system RAM, HBM is tightly integrated with GPUs and AI accelerators so they can keep their compute units fed with data, improving performance and power efficiency for bandwidth-hungry workloads. In modern data centres, HBM has become a critical part of "AI factory" designs, enabling dense accelerator clusters to train and run very large models that would otherwise be limited by memory speed.

© STL Partners



Data-movement and infrastructure offload

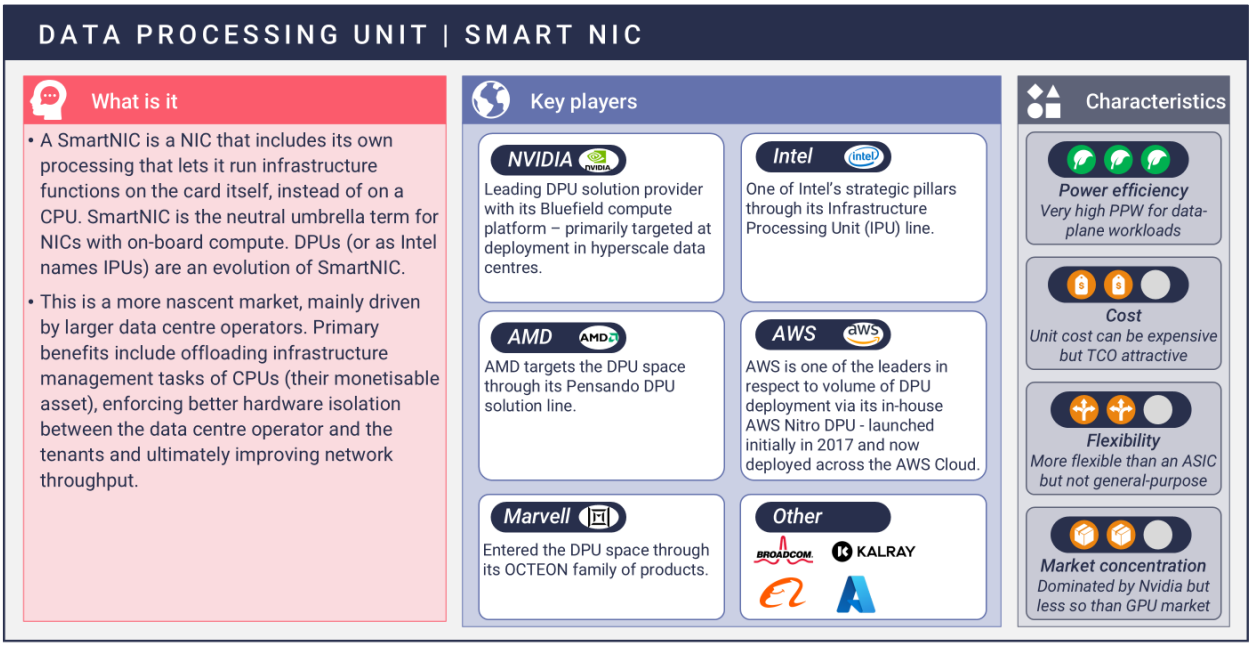
DPUs / SmartNICs (Data Processing Units)

DPUs are network interface cards that include their own processors and hardware accelerators. Their role is to take on infrastructure tasks that would otherwise consume CPU resources. By handling networking, security, and storage operations directly on the NIC, they reduce overhead on the main server and improve system efficiency, in addition to enforcing isolation of the tenant from the data centre operator.

Typical workloads:

- Packet processing and high-speed network I/O
- Virtual switching and routing used in cloud environments
- Encryption, firewalling, and security inspection
- Freeing CPUs from infrastructure management so they can focus on application workloads

© STL Partners



Switch ASICs (Networking Switch Chipsets)

Switch ASICs are the specialised chips inside data centre switches. Their function is to move data between servers at extremely high speed with predictable latency (see our previous article on data centre connectivity). Unlike CPUs or GPUs, they are optimised for fixed-function packet forwarding and routing logic.

Typical workloads:

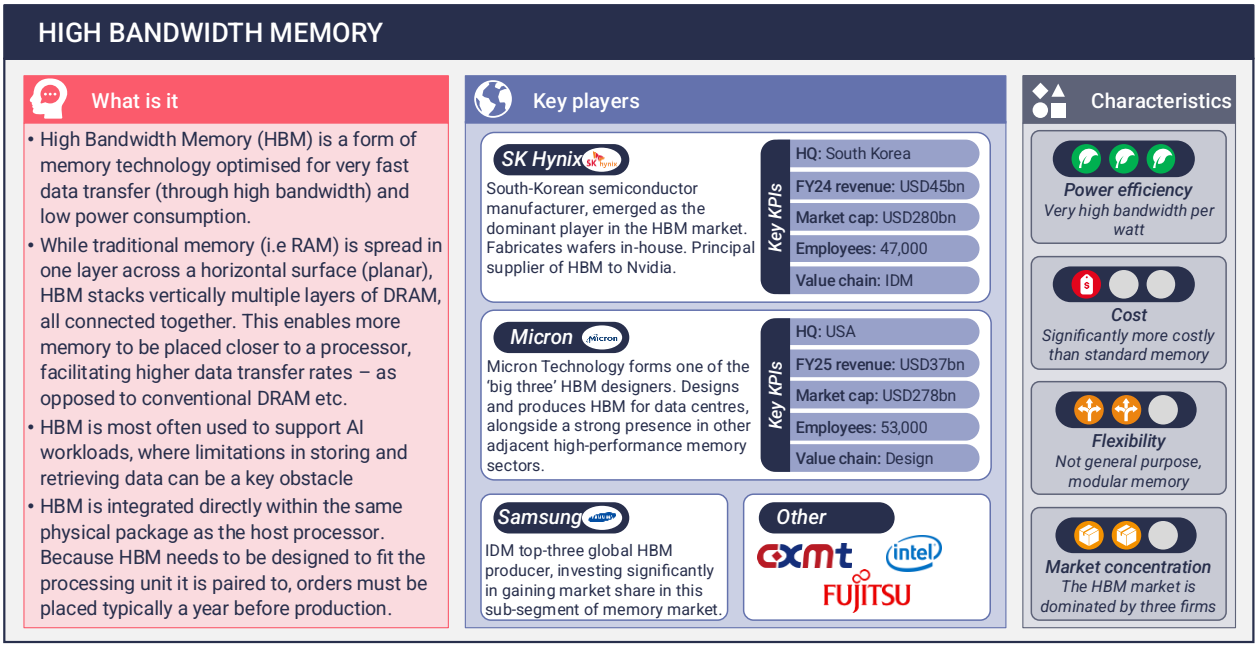
- High-throughput packet forwarding between servers and racks
- Routing, load balancing, and congestion control
- Implementing data centre network policies (e.g., QoS, telemetry, filtering)

Storage components

Memory

High Bandwidth Memory (HBM) is a type of stacked DRAM (Dynamic Random-Access Memory (DRAM) – the main type of working memory used in servers and computing systems) designed to deliver much higher bandwidth than conventional memory, by placing multiple memory dies very close to the processor and connecting them with wide, high-speed interfaces. While conventional DRAM still dominates the overall memory market, AI workloads place far greater emphasis on memory bandwidth than traditional enterprise applications. Rather than serving as general-purpose system RAM, HBM is tightly integrated with GPUs and AI accelerators so they can keep their compute units fed with data, improving performance and power efficiency for bandwidth-hungry workloads. In modern data centres, HBM has become a critical part of "AI factory" designs, enabling dense accelerator clusters to train and run very large models that would otherwise be limited by memory speed.

© STL Partners



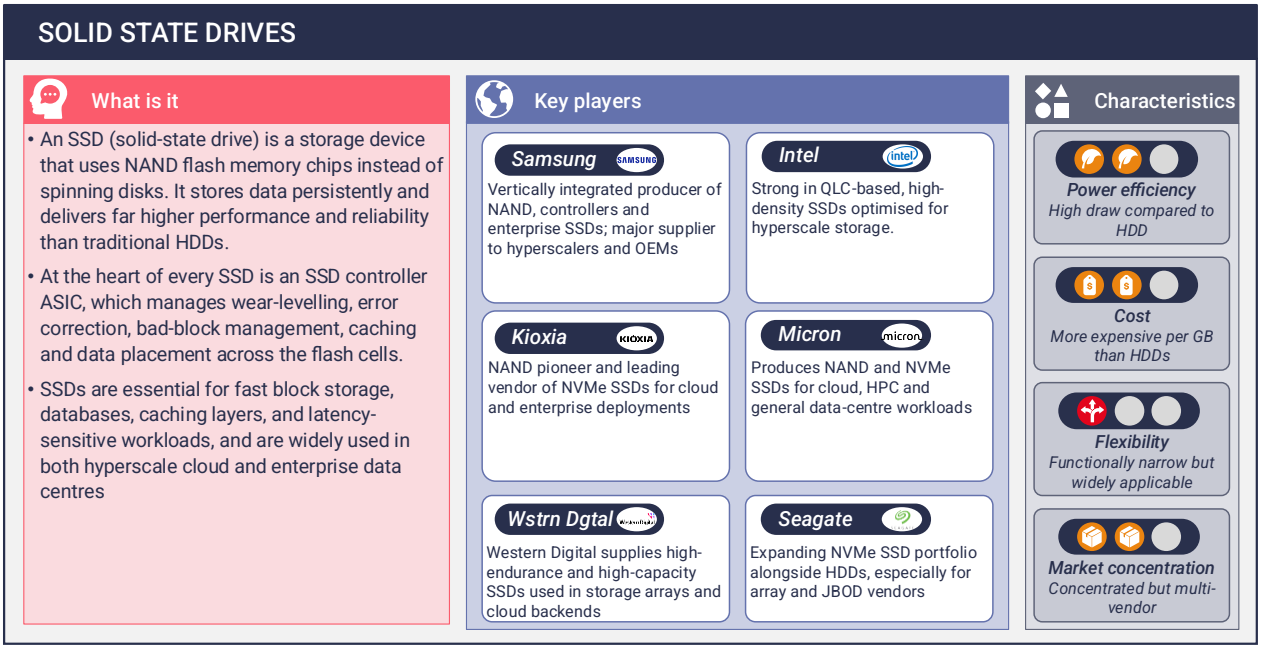
SSDs

SSDs (Solid-state drivers) store data on NAND (non-volatile memory that retains memory without power) flash chips. They provide high performance and low latency compared with hard drives. Each SSD includes a controller that manages how data is written and retrieved, ensuring reliability and consistent performance over time.

Used for:

- Block storage
- Fast databases
- Virtual machines

© STL Partners

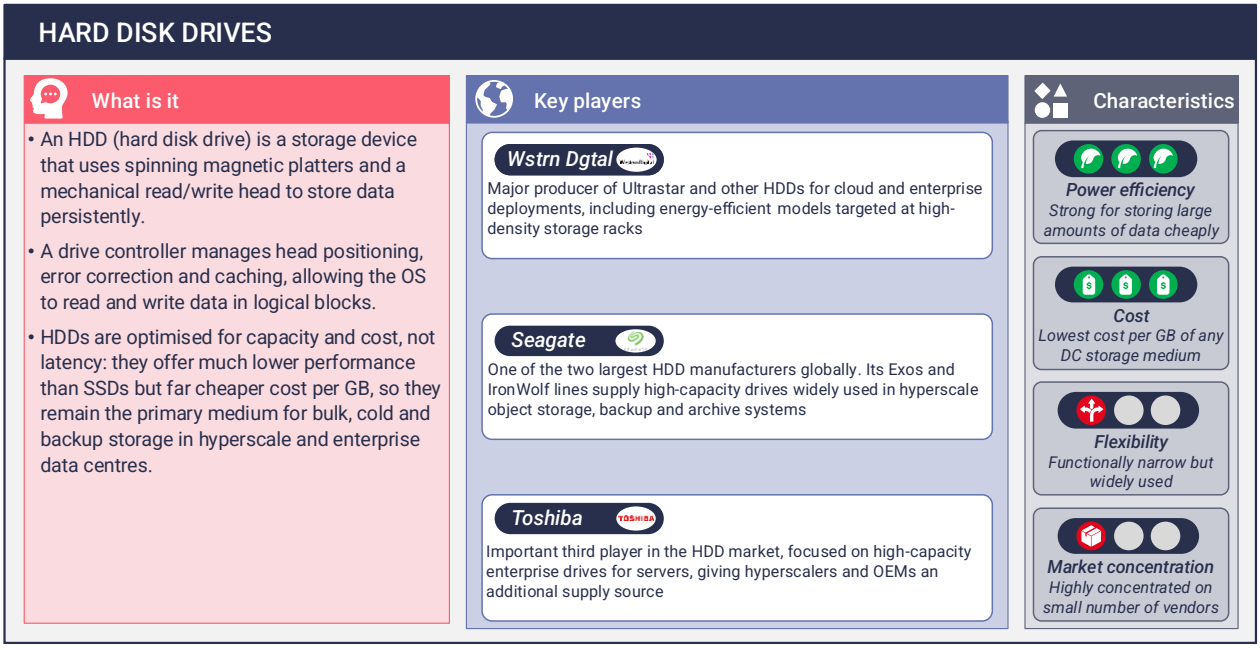


HDDs

HDDs (Hard disk drives) store data on magnetic spinning platters. They offer much higher capacity per dollar than SSDs, though with slower access times. They remain a core part of cloud storage systems where scale and cost efficiency are the priority.

Used for:

- Object storage
- Backup
- Archival data



© STL Partners

Storage controllers

Storage controllers manage the flow of data to and from persistent storage. They handle the complexities of flash media, maintain data integrity, and implement data-protection schemes across devices. In SSDs this logic is embedded; in larger storage systems, it is provided by dedicated controller hardware.

Used for:

- Wear-levelling
- Error correction
- Encryption
- Caching

Summary

This growing reliance on specialised silicon is fundamentally reshaping data centre design. Rather than optimising around general-purpose servers, operators are now designing facilities around accelerator clusters, high-bandwidth memory, dense interconnects and purpose-built networking and offload hardware. Power delivery, cooling, rack density and network topology are increasingly driven by the needs of GPUs, AI ASICs and HBM-rich systems, while DPUs and advanced switch ASICs become essential to scaling performance efficiently. As a result, modern data centres are evolving from uniform compute environments into tightly integrated “AI factories”, where architectural choices are dictated by data movement, memory bandwidth and energy efficiency as much as by raw compute capacity.

© STL Partners