SXM6 is the name used in vendor and secondary sources for NVIDIA's sixth-generation **SXM** socket and module form factor, designed for the company's Blackwell-architecture data center GPUs (such as the B100 and B200). It is a mezzanine design: rather than plugging a GPU into the system as a PCIe expansion card, the GPU mounts directly onto the system baseboard. This direct-mount approach is what lets SXM6 deliver high power and, critically, host the NVLink 5 GPU-to-GPU interconnect fabric used in NVIDIA's HGX and DGX systems.

It is worth flagging up front that "SXM6" appears predominantly in secondary, OEM, and encyclopedia material rather than in surfaced primary NVIDIA documentation. The functional description below is well-corroborated; the exact label and some of the specifications attributed to it should be read with that caveat in mind.

Explainer · June 3, 2026

What SXM is, and why it exists

SXM is a high-bandwidth socket/module form factor for mounting NVIDIA data center GPUs directly to a system, used instead of plug-in PCIe cards. The alternative is a standard PCIe add-in card, which slots into a motherboard through a PCIe edge connector. SXM replaces that edge-connector relationship with a direct mezzanine mount: the GPU module lands on a dedicated baseboard (the HGX or DGX baseboard), and power, signaling, and interconnect all flow through that baseboard rather than through a PCIe slot or external cables.

The core reason it exists comes down to interconnect. The baseboard hosts the NVSwitch fabric that connects multiple GPUs to each other via NVLink, so GPU-to-GPU traffic never has to traverse PCIe. This allows up to 8 GPUs to be fully interconnected with one another **without using PCIe for GPU-to-GPU traffic** — a major advantage for the dense, multi-GPU workloads typical of large-scale AI training and inference.

In other words, PCIe still has its place, but for the high-volume data exchange between GPUs that defines these systems, the SXM design routes that traffic over a dedicated, far higher-bandwidth fabric carried through the baseboard.

The interconnect: NVLink 5

In Blackwell systems, the GPU-to-GPU interconnect carried is **NVLink 5**, delivering up to roughly **1.8 TB/s bidirectional bandwidth per GPU** (18 links at approximately 100 GB/s each).

An important distinction: this 1.8 TB/s figure is a property of NVLink 5 and the GPU itself, not a rating of the SXM6 socket. The socket form factor provides the physical and electrical platform that the interconnect runs through, but the bandwidth number belongs to the protocol and the GPU package.

What the "6" means: the generational lineage

The generation number tracks alongside both the GPU architecture and the NVLink generation it carries. Each GPU generation has its own SXM revision tied to an NVLink generation:

- **SXM3** — V100 (NVLink 2)
- **SXM4** — A100 (NVLink 3, ~600 GB/s)
- **SXM5** — H100 (NVLink 4, ~900 GB/s)
- **SXM6** — Blackwell, B100/B200 (NVLink 5, ~1.8 TB/s)

This lineage is the clearest way to understand what "6" signifies: it is the sixth iteration in a sequence where each step pairs a new GPU generation with a faster GPU-to-GPU fabric. Note one numbering subtlety: the sixth SXM revision (SXM6) pairs with the fifth NVLink generation (NVLink 5), so the two numbers do not line up by coincidence of count. The bandwidth figures shown are protocol/GPU properties across these generations, not socket ratings.

Power delivery

SXM6 modules are paired with high-power Blackwell GPUs — the B100 at roughly 700 W and the B200 at roughly 1000 W TDP — with power delivered through the baseboard socket rather than via external cables. This is the central engineering rationale for the direct-baseboard design. A PCIe x16 slot delivers only about 75 W from the slot itself, with supplemental connectors adding more, but it still falls far short of the 700–1,000 W range that Blackwell-class GPUs require. The baseboard socket can deliver that power directly, absorbing the power and thermal challenge that would otherwise demand external cabling and connectors rated for those loads.

These TDP numbers are GPU-SKU figures rather than a documented socket-level rating. The takeaway is the engineering logic, namely that GPUs drawing this much power need power delivered directly through the baseboard, expressed as GPU power figures rather than as a spec of the SXM6 socket itself.

Memory

Blackwell SXM6 GPUs use **HBM3e** memory. This is best understood as a property of the GPU package rather than a specification of the SXM6 socket. Both B100 and B200 are marketed at roughly 192 GB of HBM3e and around 8 TB/s of memory bandwidth, with the two SKUs differentiated primarily on compute throughput and TDP rather than on memory capacity or bandwidth.

What remains contested or unconfirmed

The term "SXM6" appears consistently across vendor product pages, OEM listings, and encyclopedia entries, but has not been surfaced in primary NVIDIA architecture documentation or hardware specifications in the available evidence. NVIDIA's own Blackwell technical materials describe NVLink 5, NVLink-C2C (chip-to-chip), and PCIe Gen 6 interfaces without using the label "SXM6." This does not mean the designation is wrong — it is the consistent term in partner and aggregator sources — but it means the label is best understood as vendor/OEM nomenclature rather than a confirmed primary NVIDIA product name.

Similarly, the expansion of "SXM" to "Server PCI Express Module" appears in encyclopedia sources but has not been confirmed in any primary NVIDIA document in the available evidence, so it should be treated as unverified.

Finally, whether SXM6 applies broadly to all Blackwell die variants (including the base GB100) or only to specific board-level SKUs is also unresolved in the available evidence.

Summary

SXM6 is, in short, the sixth-generation form factor in which NVIDIA's Blackwell GPUs are mounted directly to HGX/DGX baseboards. Its purpose is to deliver the high power these GPUs require and to host the NVLink 5 fabric that interconnects them at up to ~1.8 TB/s per GPU, sidestepping PCIe for GPU-to-GPU communication. The key mental model is to separate the socket (the physical platform) from the things often attributed to it — the bandwidth belongs to NVLink 5, and the power and memory figures belong to the GPU SKUs that fill the socket. And bear in mind that "SXM6" itself is a label drawn from vendor and secondary sources rather than confirmed primary NVIDIA documentation.

WHERE SOURCES DISAGREE

Does "SXM6" appear in primary NVIDIA documentation, or only in vendor/OEM and encyclopedia material?

One reading takes "SXM6" as an established NVIDIA designation, since it appears consistently across vendor product pages and encyclopedia entries. Another account notes that NVIDIA's own surfaced architecture and Blackwell materials describe NVLink, NVLink-C2C, and PCIe Gen 6 interfaces without using the term "SXM6" at all, suggesting the label originates predominantly with partners, OEMs, and aggregators

Unresolved: no primary NVIDIA hardware specification using the term was surfaced in the available evidence

Does "SXM" officially expand to "Server PCI Express Module"?

One reading presents this expansion as fact, quoted directly from an encyclopedia entry. Another account notes the expansion is not confirmed by any primary NVIDIA source in the evidence

Unresolved: only secondary sources state the expansion and no NVIDIA document confirming it was found

Is "B300" a canonical NVIDIA GPU name, or a platform/marketing label for Blackwell Ultra?

One reading treats "B300 SXM6" as a defined product, per vendor and cloud-provider product pages specifying 288 GB HBM3e and NVLink 5. Another account notes NVIDIA's own technical material uses "Blackwell Ultra" rather than "B300," and that OEM references like "HGX B300" appear to be platform-level nomenclature

Unresolved: the "B300" naming and its module designation come only from partner/marketing sources, not primary NVIDIA branding

Whether the SXM6 form factor extends to the base Blackwell die (GB100) or specific board-level SKUs only

One reading lists SXM6 broadly for Blackwell. The available evidence does not specify whether GB100 uses SXM6 or whether the form factor is tied only to particular products like B100/B200

Unresolved: the corpus raises this as an open question without resolving it

Exact per-GPU memory bandwidth and TDP across B100 vs B200

One reading distinguishes the GPUs by these figures (e.g., ~8 TB/s as a B200-class number, 700 W vs 1000 W). Another account, drawing on independent comparison tables, lists both B100 and B200 at up to ~8 TB/s and 192 GB HBM3e, differentiating mainly on compute throughput and TDP

Unresolved: the figures derive from secondary vendor/provider tables of varying reliability rather than NVIDIA datasheets