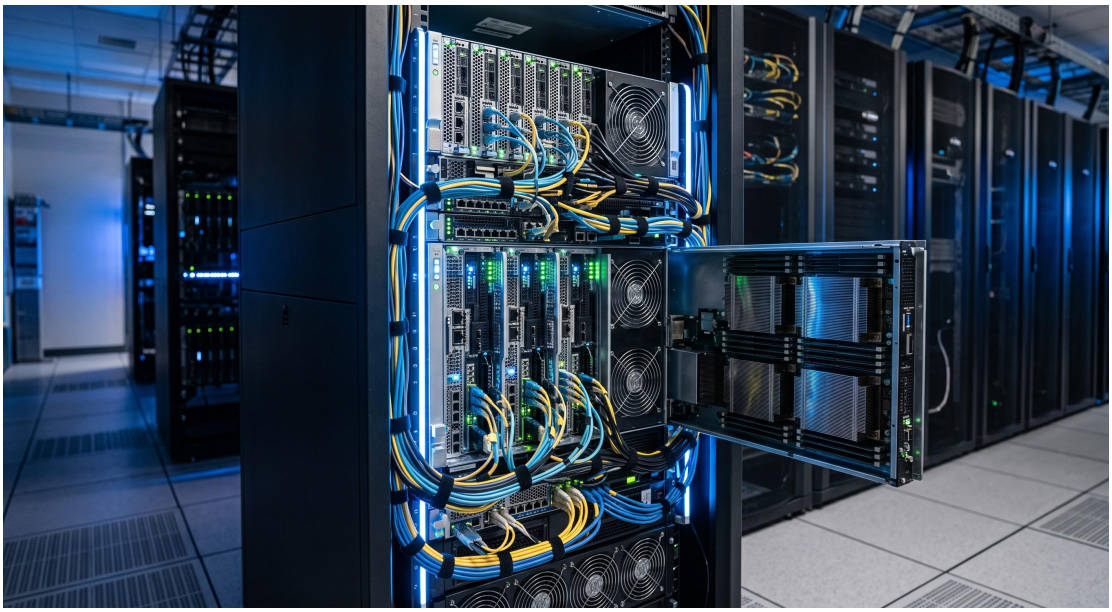
GPU Server Compute Module Selection and Form Factor Reference - Official Technical Overview & Hardware Datasheet

EXECUTIVE SUMMARY

This document provides a comprehensive technical reference for the selection and deployment of GPU server compute modules within modern datacenter and telecom infrastructure environments. As artificial intelligence, machine learning, and high-performance computing workloads continue to proliferate across carrier-grade networks, the strategic selection of GPU compute modules and their associated form factors has become a critical architectural decision. This datasheet outlines the key selection criteria, form factor considerations, thermal design parameters, power delivery requirements, and interoperability standards necessary to ensure optimal performance, scalability, and lifecycle sustainability. The information presented herein is intended for network architects, systems engineers, and procurement specialists tasked with designing and deploying GPU-accelerated compute infrastructure.



ARCHITECTURE & CHASSIS DESIGN

GPU server compute modules are engineered to deliver massive parallel processing capability within standardized server chassis architectures. The architectural foundation of these modules is built upon a modular design philosophy that decouples the GPU compute element from the host server platform, enabling flexible configuration and simplified maintenance. Key architectural considerations include the selection of the appropriate GPU form factor, such as the standard PCI Express add-in card, the SXM module, or the OAM (Open Accelerator Module) specification, each offering distinct advantages in terms of power density, interconnect bandwidth, and thermal management.

The chassis design must accommodate the physical dimensions, power delivery,

and cooling requirements of the selected GPU modules. For high-density deployments, a 2RU or 4RU chassis is typically employed to provide sufficient airflow and power distribution. Backplane interconnect topologies, including PCIe Gen4 and Gen5, as well as proprietary high-speed interconnects such as NVLink, must be carefully matched to the GPU module interface to prevent bottlenecks. Redundant power supply units with N+1 or 2N configurations are recommended to ensure continuous operation in mission-critical environments.

HARDWARE FEATURES

GPU server compute modules support a wide range of hardware features designed to maximize computational throughput and system reliability. These include support for multiple GPU modules per chassis, high-bandwidth memory (HBM) integration, advanced error correction code (ECC) memory, and hardware-assisted virtualization. The modules are typically equipped with dedicated management interfaces, such as BMC (Baseboard Management Controller), which provide out-of-band monitoring and control capabilities. Additional features include PCIe hot-plug support, secure boot functionality, and cryptographic acceleration engines for data-at-rest and data-in-transit protection.

COMPLIANCE & STANDARDS

The GPU server compute modules and associated chassis are designed to comply with a broad spectrum of international standards and regulatory requirements. These include but are not limited to IEC 60950-1 and IEC 62368-1 for electrical safety, FCC Part 15 and EN 55032 for electromagnetic compatibility, and RoHS and WEEE directives for environmental sustainability. Telecom-specific compliance may include NEBS Level 3 certification for deployment in central office environments and ETSI standards for European network infrastructure. All modules are manufactured in ISO 9001 and ISO 14001 certified facilities.

TECHNICAL SPECIFICATIONS

The following table summarizes the key technical specifications for the reference GPU server compute module platform. Actual specifications may vary depending on the specific GPU module selected and the host chassis configuration. All parameters are subject to change without notice.

Parameter	**Specification**
Form Factor	1RU / 2RU / 4RU Chassis (module dependent)

GPU Module Interface	PCIe Gen4 x16 / SXM4 / OAM
Maximum GPU Modules per Chassis	Up to 8 (varies by form factor)
Interconnect Bandwidth	Up to 600 GB/s (NVLink) or 128 GB/s (PCIe Gen5)
Memory Type	HBM2e / HBM3 with ECC
Power Supply	1+1 Redundant AC/DC, 2000W to 3000W per PSU
Cooling	Air-cooled (front-to-back) or liquid-assisted
Management	BMC with IPMI 2.0, Redfish API
Operating Temperature	10°C to 35°C (standard), -5°C to 55°C (ruggedized option)
Compliance	IEC 62368-1, FCC Part 15, CE, RoHS, NEBS Level 3 (optional)

ORDERING OPTIONS

A variety of ordering options are available to meet diverse deployment requirements. Customers may select from different GPU module configurations, including varying memory capacities, interconnect types, and cooling solutions. Chassis options range from 1RU to 4RU form factors, with either AC or DC

power input. Redundant power supply configurations, rail kits, cable management arms, and advanced management software licenses are available as separate line items. Please consult your authorized sales representative for detailed SKU information and configuration guidance.