

Inspur NF5688M6

8x A800

Ampere

rack

2021

OVERVIEW

—	0.6 TB	—	—
Peak Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

CONFIGURATION

Accelerator	A800 SXM4
Count	8
Memory per GPU	—
Bandwidth per GPU	—

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
Efficiency	—
Aggregate Bandwidth	—

SYSTEM DETAILS

Vendor: Inspur	Family: NF	Architecture: Ampere
Form Factor: rack	Release Year: 2021	Process Node: —

INTERCONNECT

One NVIDIA HGX A800 8-GPU baseboard, with NVSwitch giving every pair of GPUs a direct 400 GB/s peer-to-peer path; up to ten 100G or 200G RDMA adapters for scale-out, in a 1:1:1 ratio of GPU to InfiniBand port to NVMe drive

NOTES

Inspur's six-rack-unit AI server, sold by its server arm IEIT SYSTEMS, and one of the few OEM machines whose maker names both the accelerator and the reference baseboard it sits on. It carries a single NVIDIA HGX A800 8-GPU baseboard holding eight SXM4 A800 modules, identified in Inspur's bill of materials as NV_640G_HGX-A800-8GPU, which places 640 GB of GPU memory in the chassis. NVSwitch gives any two GPUs a direct 400 GB/s path, and Inspur balances the machine at one InfiniBand port and one NVMe drive per GPU, with up to ten 100G or 200G RDMA adapters. Hosting is two third-generation Intel Xeon Scalable processors on the C621A chipset. Fully configured the machine weighs 88 kg net and 126 kg shipped, and it draws from six 3,000 W supplies in 3+3 redundancy. Inspur quotes 5 PFLOPS of AI performance for the machine but names no precision anywhere in the 89-page white paper and never uses the words dense or sparse, so no throughput figure is listed here. Inspur also declines to publish a system power figure, stating that consumption varies by configuration.