NVIDIA® A100X

SPECIFICATIONS

Product	NVIDIA A100X Converged Accelerator
Architecture	Ampere
Process Size	7nm | TSMC
Transistors	54.2 Billion
Die Size	826 mm2
Peak FP64	9.9 TFLOPS
Peak FP64 Tensor Core	19.1 TFLOPS | Sparsity
Peak FP32	19.9 TFLOPS
TF32 Tensor Core	159 TFLOPS | Sparsity
Peak FP16 Tensor Core	318.5 TFLOPS | Sparsity
Peak INT8 Tensor Core	637 TOPS | Sparsity
Multi-Instance GPU Support	7 MIGs at 10 GB Each 3 MIGS at 20 GB Each 2 MIGs at 40 GB Each 1 MIG at 80 GB
GPU Memory	80 GB HBM2e
Memory Bandwidth	2039 GB/s
Interconnect	PCIe Gen4 (x16 Physical, x8 Electrical) | NVLink Bridge
Networking	2x 100 Gbps ports, Ethernet or InfiniBand
Maximum Power Consumption	300 W
NVLink	Third-Generation | 600 GB/s Bidirectional
Media Engines	1 Optical Flow Accelerator (OFA) 1 JPEG Decoder (NVJPEG) 4 Video Decoders (NVDEC)
Integrated DPU	NVIDIA BlueField-2 Implements NVIDIA ConnectX-6 DX Functionality 8 Arm A72 Cores at 2 GHz Implements PCIe Gen4 Switch
NVIDIA Enterprise Software	NVIDIA vCS (Virtual Compute Server) NVIDIA AI Enterprise
Form Factor	2-Slot, Full Height, Full Length (FHFL)
Thermal Solution	Passive
Maximum Power Consumption	300 W