NVIDIA® A30X

SPECIFICATIONS

Product	NVIDIA A30X Converged Accelerator
Architecture	Ampere
Process Size	7nm | TSMC
Transistors	54.2 Billion
Die Size	826 mm2
Peak FP64	5.2 TFLOPS
Peak FP64 Tensor Core	10.3 TFLOPS | Sparsity
Peak FP32	10.3 TFLOPS
TF32 Tensor Core	82.6 TFLOPS | Sparsity
Peak FP16 Tensor Core	165 TFLOPS | Sparsity
Peak INT8 Tensor Core	330 TOPS | Sparsity
GPU Memory	24 GB HBM2e
Memory Bandwidth	1223 GB/s
NVLink	Third-Generation | 200 GB/s Bidirectional
Multi-Instance GPU Support	4 MIGs at 6 GB Each 2 MIGs at 12 GB Each 1 MIG at 24 GB
Media Engines	1 Optical Flow Accelerator (OFA) 1 JPEG Decoder (NVJPEG) 4 Video Decoders (NVDEC)
Interconnect	PCIe Gen4 (x16 Physical, x8 Electrical | NVLink Bridge)
Networking	2x 100 Gbps ports, Ethernet or InfiniBand
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	230 W