

NVIDIA® L4

SPECIFICATIONS

FP32	30.3 TFLOPS
TF32 Tensor Core	120 TFLOPS Sparsity
FP16 Tensor Core	242 TFLOPS Sparsity
BFLOAT16 Tensor Core	242 TFLOPS Sparsity
FP8 Tensor Core	485 TFLOPS Sparsity
INT8 Tensor Core	485 TOPS Sparsity
GPU Memory	24GB
GPU Memory Bandwidth	300 GB/s
Display Connectors	None vGPU Only
NVENC NVDEC	2 4 4 JPEG Decoders AV1 Encode and Decode
Form Factor	Single Slot Low-Profile
System Interconnect	PCIe Gen4 x16 64GB/s
Thermal Solution	Passive
Maximum Power Consumption	72 W
Server Options	Partner and NVIDIA-Certified Systems with 1-8 GPUs