NVIDIA Vera Rubin NVL72

72x Rubin | Rubin | rack | 2026

OVERVIEW

1,260	20.7 TB	1580.0 TB/s	—
FP8 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP64	2.4	33.3	–
FP32	9.4	130.0	–
TF32	144.0	2000.0	–
FP16	288.0	4000.0	–
BF16	288.0	4000.0	–
FP8	1260.0	17500.0	–
INT8	18.0	250.0	–
NVFP4	3600.0	50000.0	–
FP6	1260.0	17500.0	–
FP32_TC	28.8	400.0	–
FP64_TC	14.4	200.0	–

All figures are dense. Vendors commonly headline the sparse number, which is twice the dense one.

CONFIGURATION

Accelerator	Rubin SXM
Count	72
Memory per GPU	288 GB
Bandwidth per GPU	22.0 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
FP8 Efficiency	—
Aggregate Bandwidth	1580.0 TB/s

SYSTEM DETAILS

Vendor: NVIDIA
Family: NVL
Architecture: Rubin
Form Factor: rack
Release Year: 2026
Process Node: —

INTERCONNECT

NVLink 6 / NVSwitch

NOTES

Rack-scale system pairing 72 Rubin GPUs with 36 Vera CPUs (3,168 Olympus Arm cores total, 54 TB LPDDR5X CPU memory). Headline NVFP4 inference 3,600 PFLOPS dense; NVFP4 training 2,520 PFLOPS dense. Tensor-core-emulated FP32 SGEMM reaches 28.8 PFLOPS, FP64 DGEMM 14.4 PFLOPS. Aggregate memory shown is HBM4 only (CPU LPDDR5X is excluded from GPU memory aggregate). Total chips per rack: 1,296. Specifications preliminary per NVIDIA GTC 2025 footnote.