

NVIDIA HGX Rubin NVL8

8x Rubin

Rubin

baseboard

2026

OVERVIEW

140 FP8 PFLOPS Dense	2.3 TB Aggregate Memory	176.0 TB/s Aggregate Bandwidth	— Aggregate Power
--	---	--	---

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP8	140.0	17500.0	—
NVFP4	280.0	35000.0	—
FP6	140.0	17500.0	—

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

CONFIGURATION

Accelerator	Rubin SXM
Count	8
Memory per GPU	288 GB
Bandwidth per GPU	22.0 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
FP8 Efficiency	—
Aggregate Bandwidth	176.0 TB/s

SYSTEM DETAILS

Vendor: NVIDIA	Family: HGX	Architecture: Rubin
Form Factor: baseboard	Release Year: 2026	Process Node: —

INTERCONNECT

NVLink 6 Switch, 28.8 TB/s total

NOTES

Eight Rubin SXM on an HGX baseboard, listed by NVIDIA alongside HGX B300 and HGX B200. NVIDIA publishes NVFP4 twice and the two are different workloads, not a sparse/dense pair: inference 400 PFLOPS sparse, training 280 PFLOPS dense. FP8/FP6 training is 140 PFLOPS dense. System power is not published for the baseboard; the 24 kW figure on DGX Rubin NVL8 is the full chassis including CPUs. NVIDIA marks the specification preliminary.