# NVIDIA DGX Rubin NVL8

8x Rubin | Rubin | rack | 2026

## OVERVIEW

| 140 | 2.3 TB | 176.0 TB/s | 24.0 kW |
|---|---|---|---|
| FP8 PFLOPS Dense | Aggregate Memory | 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 | 24.0 kW |
| TDP per GPU | — |
| FP8 Efficiency | 5.83 TFLOPS/W |
| Aggregate Bandwidth | 176.0 TB/s |

## SYSTEM DETAILS

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

## INTERCONNECT

NVLink 6 Switch, 28.8 TB/s total

## NOTES

Eight Rubin GPUs with two Intel Xeon 6776P CPUs, liquid cooled. NVIDIA publishes NVFP4 twice and the two are different workloads, not a sparse/dense pair: inference 400 PFLOPS sparse, training 280 PFLOPS dense. The dense training figure is stored as the headline; the sparse inference figure is recorded alongside it. FP8/FP6 training is 140 PFLOPS dense with no sparse counterpart published, and no INT8 figure is given. Networking is 8x OSFP with ConnectX-9 up to 800 Gb/s plus 2x BlueField-4 DPUs. NVIDIA marks the whole specification "preliminary information, all values subject to change".