

Furiosa NXT RNGD Server

8x RNGD

Tensor Contraction Processor

rack

OVERVIEW

4.10 FP8 PFLOPS Dense	0.4 TB Aggregate Memory	12.0 TB/s Aggregate Bandwidth	3.0 kW Aggregate Power
---------------------------------	-----------------------------------	---	----------------------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP16	2.0	256.0	—
FP8	4.1	512.0	—
INT8	4.1	512.0	—
INT4	8.2	1024.0	—

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

CONFIGURATION

Accelerator	RNGD
Count	8
Memory per GPU	48 GB
Bandwidth per GPU	1.5 TB/s

POWER & EFFICIENCY

Aggregate Power	3.0 kW
TDP per GPU	180 W
FP8 Efficiency	1.37 TFLOPS/W
Aggregate Bandwidth	12.0 TB/s

SYSTEM DETAILS

Vendor: Furiosa	Family: NXT	Architecture: Tensor Contraction Processor
Form Factor: rack	Release Year: —	Process Node: 5nm

INTERCONNECT

PCIe Gen5 x16 per card; 2 x 25G dual-port NIC for host networking

NOTES

A 4U air-cooled server holding eight Furiosa RNGD cards, and the low-power outlier of this catalogue: the whole machine draws 3 kW, which lets five of them sit in a standard 15 kW air-cooled rack where comparable GPU racks need liquid cooling. Furiosa rates it at 4,096 TFLOPS FP8, 2,048 TFLOPS FP16, 4,096 TOPS INT8 and 8,192 TOPS INT4, with 384 GB of HBM3 across the eight cards at 12 TB/s aggregate. Hosting is a dual AMD EPYC 9354 Genoa pair with 64 cores total and 1 TB of DDR5, with two 25G dual-port NICs and Ubuntu 24.04 LTS. Every figure is Furiosa's own published total and each reconciles exactly against the single RNGD card. One labelling note: this server table calls the 2,048 TFLOPS figure FP16 while Furiosa's card table calls the identical per-card rate BF16 —

the chip runs both at the same speed, and each row here records the label its own document uses. Furiosa publishes rack economics alongside: five servers per rack delivering 26,400 tokens per second and serving 880 users. Furiosa labels none of its figures dense or sparse and publishes no sparse rate.