



FuriosaAI RNGD

Also known as: RNGD · Renegade · Furiosa RNGD · NXT RNGD Server · Furiosa second-generation NPU

RNGD (pronounced "renegade") is FuriosaAI's second-generation, inference-only accelerator: a passively cooled, dual-slot PCIe Gen5 card built on TSMC 5nm, with 48 GB of HBM3 at 1.5 TB/s, 256 MB of on-chip SRAM and a 180 W TDP. It is not a GPU. Its Tensor Contraction Processor (TCP) architecture treats tensor contraction, not 2D matrix multiply, as the primitive, and the Furiosa compiler maps whole models onto it, so it runs only what the Furiosa SDK supports. FuriosaAI sells it as a single card and as the NXT RNGD Server, an 8-card, 4U, roughly 3 kW air-cooled appliance. Where it fits in a build: power-capped, air-cooled racks running mainstream open-weight LLM inference, where tokens per watt matter more than per-card peak throughput or CUDA compatibility.

512 TFLOPS FP8 compute · peak; BF16 256 TFLOPS	1,024 TOPS INT4 compute · INT8 512 TOPS	48 GB HBM3 Memory · 1.5 TB/s, two stacks
256 MB On-chip SRAM · software-managed, not a cache	180 W TDP · earlier docs and Hot Chips 2024 listed 150 W	Dual-slot FH 3/4L, passive Form factor · PCIe Gen5 x16, no proprietary card-to-card link

COMPUTE (TCP ARCHITECTURE)		01/06
Architecture	Tensor Contraction Processor (TCP) · tensor contraction as the primitive operation	
Process / clock	TSMC 5nm, 1.0 GHz	
Processing elements	8 PEs x 64 slices · 8 x 64 TFLOPS FP8 per vendor product page	
PE fusion	Up to 4 PEs fused into one · reported, Hot Chips 2024	
BF16	256 TFLOPS	
FP8	512 TFLOPS	
INT8	512 TOPS	
INT4	1,024 TOPS	

MEMORY		02/06
HBM	48 GB HBM3 · two 12-high stacks, reported	
HBM bandwidth	1.5 TB/s	
On-chip SRAM	256 MB · 512 KB per slice, reported	
ECC	Supported	
Packaging	2.5D, SoC and HBM on one interposer · reported; interposer technology not published	
INTERFACE & FORM FACTOR		03/06
Host interface	PCIe Gen5 x16 · also carries card-to-card traffic	
Peer-to-peer	Card-to-card over a PCIe switch · reported, Hot Chips 2024	
Form factor	Dual-slot, full-height, 3/4 length	
Power connector	12VHPWR	
Multi-instance	Up to 8 instances	
Virtualization	SR-IOV, 8 virtual functions	
Security	Secure boot with root of trust	
POWER & THERMAL		04/06
TDP	180 W · current docs; 150 W in the v2025.1 docs and at Hot Chips 2024	
Cooling	Passive · needs server chassis airflow	
Rack density	5 NXT RNGD Servers per 15 kW air-cooled rack · vendor claim, 40 cards per rack	

NXT RNGD SERVER		05/06
Accelerators	8x RNGD	
Accelerator memory	384 GB HBM3 · 12 TB/s aggregate	
Compute	4,096 TFLOPS FP8, 2,048 TFLOPS BF16 · INT8 4,096 TOPS, INT4 8,192 TOPS	
CPU	2x AMD EPYC 9354 (Genoa) · 64 cores total, 3.25 GHz base / 3.75 GHz boost	
System memory	1 TB DDR5	
Storage	960 GB NVMe M.2 (OS) + 2x 3.84 TB NVMe U.2 (data)	
Networking	2x 25 GbE dual-port NIC · configurable	
Power / chassis	3,000 W, 4U rackmount, air-cooled	
OS / software	Ubuntu 24.04 LTS, Furiosa SDK	
SOFTWARE		06/06
SDK	Furiosa SDK · compiler, quantizer, runtime; PyTorch / torch.compile front end	
Serving	Furiosa-LLM · vLLM-compatible, OpenAI-compatible API server	
Operations	Kubernetes integration, Prometheus metrics	
Parallelism	Tensor parallelism across cards · over PCIe	
CUDA	Not supported · not a GPU; models run through the Furiosa compiler	
PROCUREMENT		
Typical lead time Not published. FuriosaAI announced on January 27, 2026 that its first volume delivery of 4,000 RNGD units had arrived from TSMC and ASUS and that the card and the NXT RNGD Server are available to enterprise customers; order lead times are handled by sales. FuriosaAI also offers evaluation access as bare metal or an OpenAI-compatible API endpoint. Pricing Not published; quote-only for both the card and the NXT RNGD Server. One informal figure of around \$10,000 per card circulated at Hot Chips 2024 (reported, side conversations, not an official price).		
▪ Sold direct by FuriosaAI ("Request a quote" / "Talk to sales"); no public distributor or reseller listing was found. ▪ The NXT RNGD Server is the turnkey route: it ships with the host CPU, memory, storage, OS and SDK already validated for 8 cards. ▪ Single cards go into a host server you qualify yourself: passive cooling, a 12VHPWR feed and a PCIe Gen5 x16 slot per card, and a PCIe switch topology if you want card-to-card traffic off the host. ▪ Volume production only started in early 2026 (first delivery of 4,000 units); confirm allocation and delivery dates in writing before planning a multi-rack deployment.		

FIELD NOTES

- Model coverage is the gating question, not FLOPS. RNGD runs only what the Furiosa compiler and Furiosa-LLM support; check that your exact model, quantization and context length are on the supported list (precompiled artifacts are published up to 32K context) before you buy, because there is no CUDA fallback.
- 48 GB per card is small for large models. Anything past roughly a 30B model at FP8 needs tensor parallelism across cards, and that traffic rides PCIe Gen5 (about 128 GB/s per card), not an NVLink-class fabric. Furiosa says it overlaps inter-chip DMA in the compiler, but large-batch scaling data has not been published.
- The efficiency case is real but vendor- and customer-reported. LG AI Research reported 2.25x better LLM performance per watt than a GPU setup (the GPU is unnamed in Furiosa's post; The Register reports it was A100) and about 50 to 60 tokens/s at batch 1 for EXAONE 3.5 32B in FP16 on 4 cards. Re-run your own workload before you accept the ratio against current GPUs.
- The TDP moved from 150 W (Hot Chips 2024, v2025.1 docs) to 180 W in the current docs. Budget power and airflow at 180 W per card; the NXT RNGD Server is specified at 3 kW for 8 cards plus a dual-EPYC host.
- The fit is a power-capped, air-cooled room: five 3 kW servers per 15 kW rack (vendor claim) without liquid cooling. If the site has liquid cooling and power headroom, per-rack throughput on HBM GPUs is likely higher; RNGD's argument is tokens per watt and per rack under an air-cooling limit.

QUESTIONS WE GET ON THIS PART

Is FuriosaAI RNGD a GPU?

No. It is an inference-only accelerator built on FuriosaAI's Tensor Contraction Processor architecture: 8 processing elements of 64 slices each, 256 MB of on-chip SRAM, and 48 GB of HBM3. It does not run CUDA; models are compiled with the Furiosa SDK and served with Furiosa-LLM, which is vLLM-compatible and exposes an OpenAI-compatible API.

What are the RNGD card's key specs?

512 TFLOPS FP8, 256 TFLOPS BF16, 512 TOPS INT8 and 1,024 TOPS INT4; 48 GB HBM3 at 1.5 TB/s; 256 MB SRAM; PCIe Gen5 x16; 180 W TDP; passively cooled, dual-slot, full-height, 3/4-length card with a 12VHPWR connector; up to 8 instances via SR-IOV.

What is the NXT RNGD Server?

FuriosaAI's turnkey 4U, air-cooled appliance with 8 RNGD cards (384 GB HBM3, 12 TB/s aggregate, 4,096 TFLOPS FP8), two AMD EPYC 9354 CPUs, 1 TB DDR5, NVMe storage and 2x 25 GbE dual-port NICs, rated at 3,000 W. FuriosaAI says five fit in a 15 kW air-cooled rack.

How much does RNGD cost?

FuriosaAI does not publish a price for the card or the server; both are quote-only. An informal figure of around \$10,000 per card was reported from side conversations at Hot Chips 2024, which is not an official price.

REFERENCES

[1] <https://furiosa.ai/rngd>

[2] <https://furiosa.ai/nxt-rngd-server>

[3] <https://developer.furiosa.ai/latest/en/overview/rngd.html>

[4] <https://developer.furiosa.ai/v2025.1.0/en/overview/rngd.html>

[5] <https://furiosa.ai/blog/tensor-contraction-processor-ai-chip-architecture>

[6] <https://furiosa.ai/blog/lg-ai-research-taps-furiosaaai-to-achieve-2-25x-better-llm-inference-in-production-vs-gpus>

[7] <https://www.storagenewsletter.com/2026/02/27/rngd-enters-mass-production-4000-high-performance-ai-accelerators-shipped-by-tsmc/>

[8] https://www.theregister.com/2025/07/22/sk_furiosa_ai_lg/

[9] <https://chipsandcheese.com/p/furiosaaais-rngd-at-hot-chips-2024-accelerating-ai-with-a-more-flexible-primitive>

Latest version of this datasheet: <https://gpusmith.com/hardware/gpus/furiosaai-rngd>. Specifications compiled from the references above; verify against vendor documentation before purchase.