FURIOSA

Turnkey AI inference you can own today

Furiosa NXT RNGD (pronounced “renegade”) Server delivers exceptional performance and cost-efficient scalability for inference with advanced LLM and agentic AI applications.

8x RNGD	4 petaFLOPS	384 GB	12 TB/s	3 kW
Tensor Contraction Processor (TCP)	512 TFLOPS (FP8) x 8 RNGDs	HBM3 Capacity	Memory Bandwidth	Power Consumption

High throughput, low power, and full control

Furiosa NXT RNGD Server is powered by Furiosa’s RNGD cards with Tensor Contraction Processor (TCP) architecture, purpose-built for AI inference, balancing performance, programmability, and efficiency.

NXT RNGD Server supports up to 8 RNGD cards, delivering exceptional compute in a single 3 kW server. You can fit up to 5 RNGD servers within a 15 kW rack, generating 3.5x more tokens than a single H100 SXM server (as shown in the chart below).

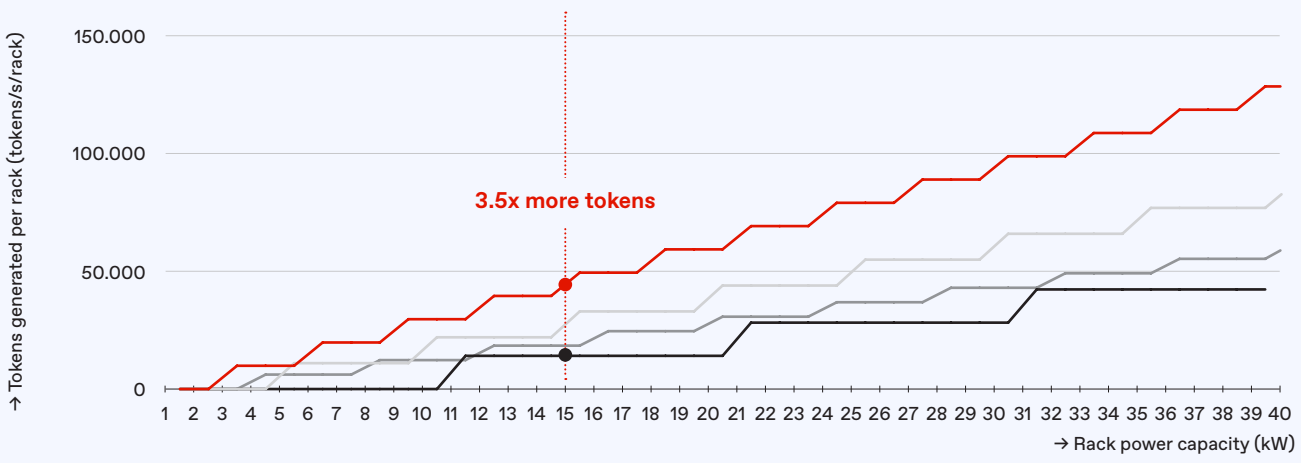
A scalable, high-performance turnkey solution, NXT RNGD Server can be deployed easily and efficiently in any data center.
Key features include:

- High memory capacity for large models and extended sequence KV Cache
- Support for multiple quantization formats, including BF16, FP8, INT8, and INT4
- Scalable efficiency for more tokens per rack than GPUs, significantly reducing TCO

3.5x more compute per rack than GPUs

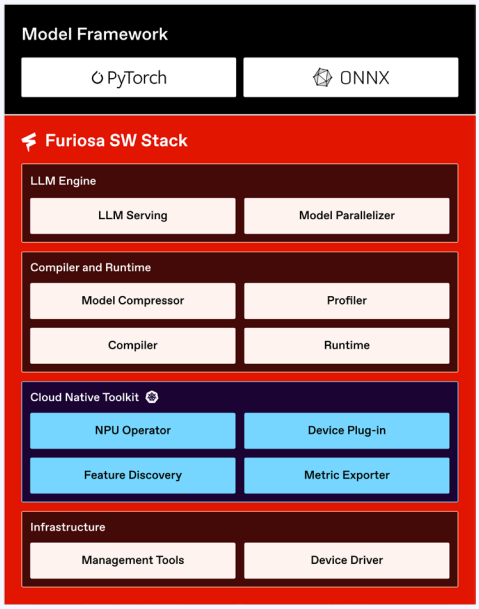
Rack-level performance for RNGD vs. GPUs with Llama 3.1 8B FP8

5x RNGD Server	3x H100 PCIe System	3x L40s System	1x H100 SXM System
49.412 tokens/s	32.957 tokens/s	18.425 tokens/s	14.095 tokens/s



Comparison of rack-level token generation performance under a fixed 15 kW power budget, with all systems running identical workloads

Comprehensive software stack for high-performance LLMs



NXT RNGD Server comes preinstalled with Furiosa's software stack, delivering a turnkey deployment experience that integrates seamlessly into your infrastructure.

Streamlined LLM Stack
From Hugging Face models to Kubernetes workloads, deployment is simplified end-to-end.

Maximum Performance Efficiency
Furiosa Compiler, Furiosa Runtime, and Furiosa-LLM deliver high throughput, low latency, and best-in-class efficiency.

Cloud-Native Ready
Publicly available container images, SR-IOV slicing, and native Kubernetes Device Plugin/NPU Operator simplify data center management.

OpenAI API and vLLM-Compatible Serving
Furiosa-LLM is a vLLM-compatible serving framework that enables you to deploy open-source models quickly and efficiently. With built-in OpenAI API compatibility, enterprises can scale LLM deployments without retooling their stack.

Technical Specifications

Processing Unit	8x RNGD Tensor Contraction Processors (TCPs)
Processing Unit Memory	384 GB total
Performance	4,096 TFLOPS FP8 2,048 TFLOPS FP16 4,096 TOPS INT8 8,192 TOPS INT4
System Power Usage	3,000 W
CPU	Dual AMD EPYC 9354 (Genoa) Processors 64 Cores total, 3.25 GHz (Base), 3.75 GHz (Max Boost)
System Memory	1 TB DDR5
Networking	2 × 25 G dual-port NIC
Storage	OS: 960 GB NVMe M.2 Internal: 2 × 3.84 TB NVMe U.2
OS	Ubuntu 24.04 LTS
Software	Streamlined LLM Stack, including Furiosa Compiler, Furiosa Runtime, and Furiosa LLM
Form Factor	4U Rackmount Server

Sampling today. Request a demo at furiosa.ai/signup