



Turnkey AI inference you can own today

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



8x RNGD

Tensor Contraction Processor (TCP)

4 petaFLOPS

512 TFLOPS (FP8) x 8 RNGDs

384 GB

HBM3 Capacity

12 TB/s

Memory Bandwidth

3 kW

Power Consumption

High throughput, low power, and full control

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. 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)

It’s a scalable, high-performance turnkey solution that you can deploy easily and efficiently in any data centers.

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

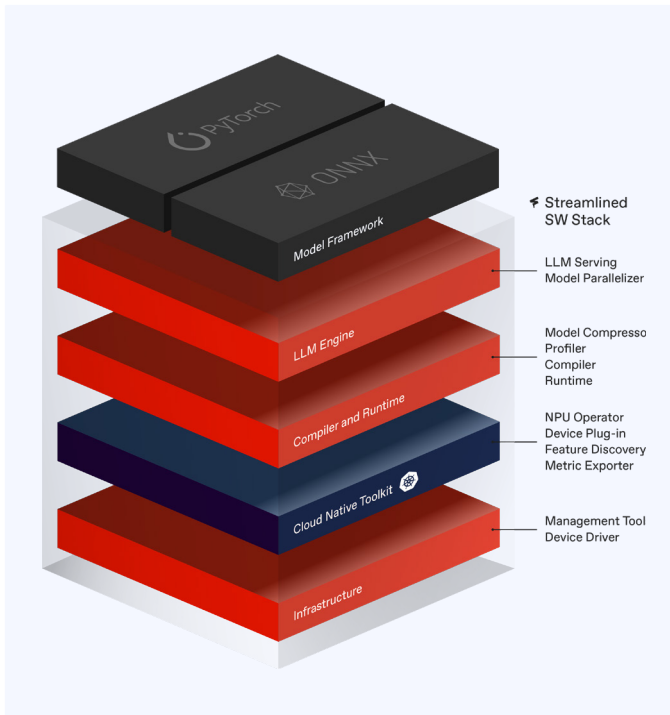
4x more inference capacity

Rack-scale economics: RNGD enable 4x more inference capacity

Max. # of servers per rack	5x	2x
Server power consumption (kW)	3kW	7.5kW
Tokens/s per rack	26,400 tokens/s	6,600 tokens/s
Max. # of users per rack	880	220



Comprehensive software stack for high-performance LLMs



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

Streamlined LLM Stack

From HuggingFace models to Kubernetes workloads, deployment is simplified end-to-end.

Maximum Performance Efficiency

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

Cloud-Native Ready

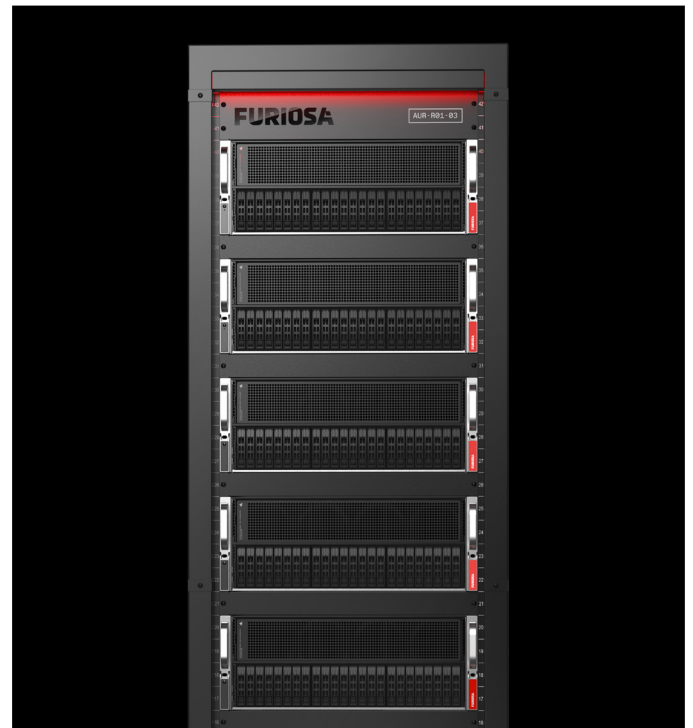
Publicly available container images, SR-IOV slicing, and native Kubernetes Device Plugin/NPU Operator simplify data center management.

OpenAI API & vLLM-Compatible Serving

Furiosa-LLM is a vLLM-compatible serving framework that allows 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	8× RNGD Tensor Contraction Processors (TCP)
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 Core 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



Available to order today. Request a demo at furiosa.ai/signup