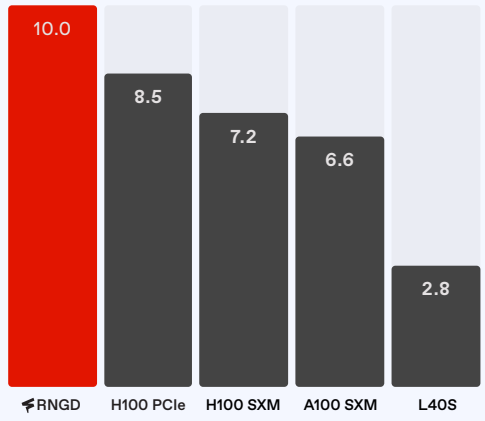
Powerfully Efficient AI Inference for Data Centers

Furiosa RNGD (pronounced “Renegade”) delivers world-class performance and unparalleled deployment flexibility for AI inference. Breakthrough power-efficiency provides a scalable, secure, and cost-effective solution for any use case—from on-prem to enterprise data centers to sovereign AI clouds.

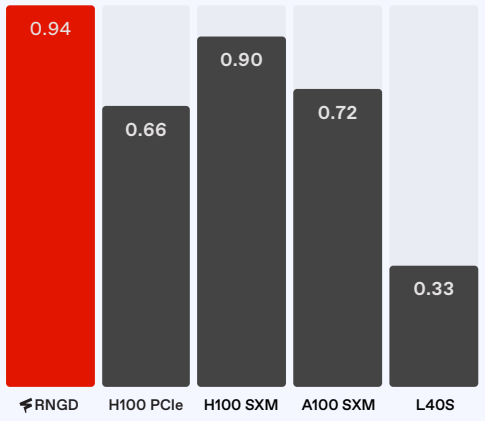
180W	256MB	48GB	1.5TB/s	512 TFLOPS (FP8)
TDP	On-chip SRAM	HBM3 Capacity	HBM3 Bandwidth	BF16, FP8, INT8, INT4 Support

Superior performance per watt

1 card
(token/s/W)



8 cards
(token/s/W)



Llama 3.1 8B

8B BF16 | Concurrency 64 | 1024 input / 1024 output tokens

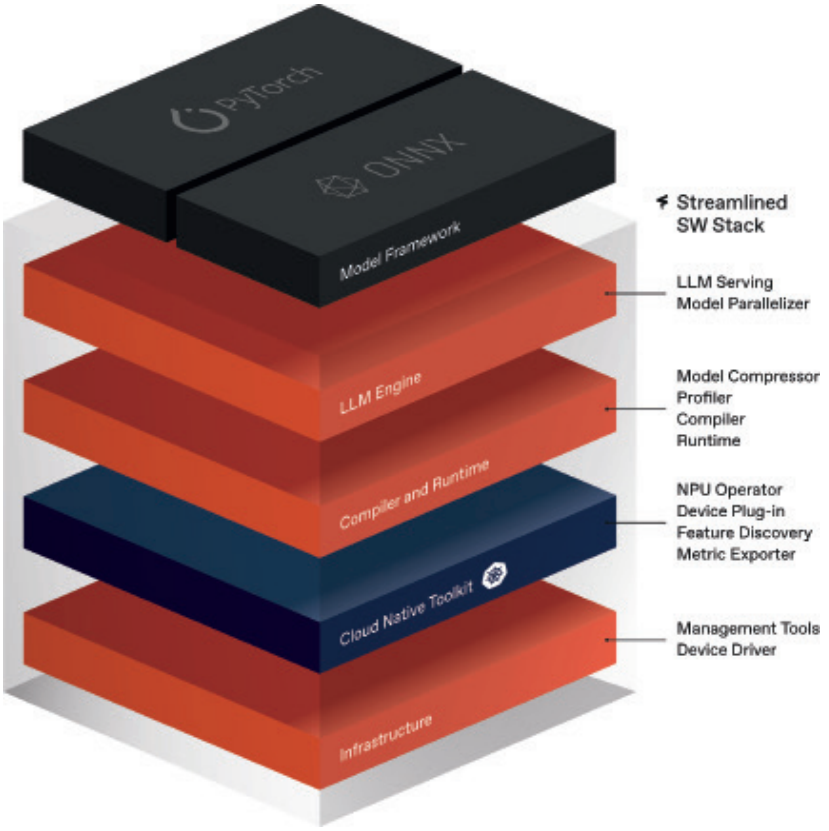
- RNGD FuriosaSDK 25.3.0 | 1725.3 tokens/s | 173.3 W
- H100 PCIe vLLM 0.9.1 | 2530.0 tokens/s | 297.1 W
- H100 SXM vLLM 0.9.1 | 3755.7 tokens/s | 519.1 W
- A100 SXM VLLM 0.9.1 | 2227.8 tokens/s / 337.6 W
- L40S vLLM 0.9.1 | 1165.4 tokens/s / 294.1 W

Llama 3.3 70B

BF16 | Concurrency 64 | 1024 input / 1024 output tokens

- RNGD FuriosaSDK 25.3.0 | 1057.3 tokens/s | 1057.3 W
- H100 PCIe 1167.1 tokens/s | 1167.1 W
- H100 SXM vLLM 0.9.1 | 3235.9 tokens/s | 3235.9 W
- A100 SXM VLLM 0.9.1 | 1789.1 tokens/s / 1789.1 W
- L40S vLLM 0.9.1 | 671.2 tokens/s / 671.2 W

Disclaimer: RNGD results are based on internal measurements by FuriosaAI using SDK 2025.3.0.
GPU results were obtained using vLLM 0.9.1 on RunPod under comparable test conditions.



Furiosa SW Stack

Built for LLM Inference
Comprehensive software toolkit for optimizing large language models on RNGD. User-friendly APIs facilitate seamless state-of-the-art LLM deployment.

Robust Ecosystem Support
Effortlessly deploy models from library to end-user with PyTorch 2.x integration. Leverage the vast advancements of open-source AI and seamlessly transition models into production.

Maximizing Data Center Utilization
Ensure higher utilization and flexibility for small and large deployments with containerization, SR-IOV, Kubernetes, as well as other cloud native components.

Technical Specifications

Architecture	Tensor Contraction Processor
Process Node	TSMC 5nm
Frequency	1.0 GHz
BF16	256 TFLOPS
FP8	512 TFLOPS
INT8	512 TOPS
INT4	1,024 TOPS
Memory Bandwidth	HBM3 1.5 TB/s
Memory Capacity	HBM3 48 GB
On-Chip SRAM	256 MB
Interconnect Interface	PCIe Gen5 x16
Thermal Solution	Passive
Thermal Design Power (TDP)	180 W
Power Connector	12 VHPWR
Form Factor	PCIe dual-slot full-height 3/4 length
Multi-Instance Support	8
Virtualization Support	Yes
SR-IOV	Yes
ECC Memory Support	Yes
Secure Boot with Root of Trust	Yes

More tokens per rack, unlimited flexibility

- Deploy on-prem or in large-scale data centers
- Up to eight RNGD accelerators into a single air-cooled 4U server
- Up to five RNGD Servers in a single, standard 15kW air-cooled rack
- One RNGD rack generate 3.75x more tokens vs. GPU rack
- Compatible with all standard server manufactureres

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