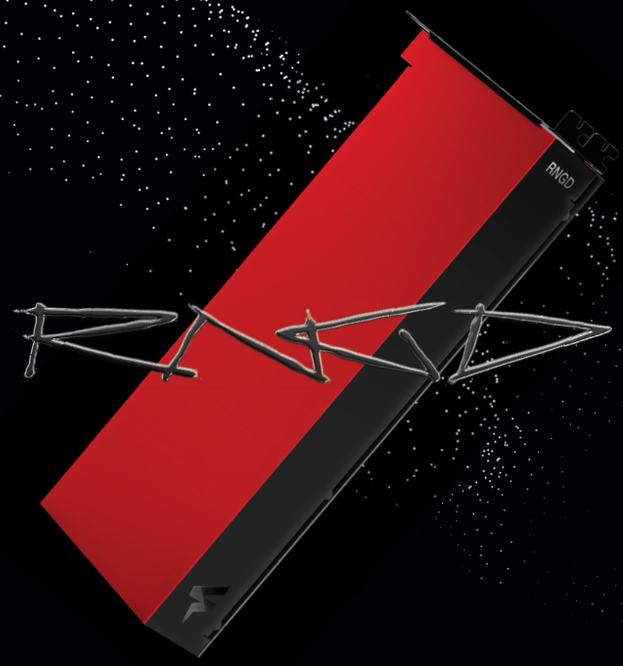
Furiosa RNGD - Gen 2 data center accelerator

Powerfully Efficient AI Inference

Furiosa RNGD (pronounced “Renegade”), powered by Tensor Contraction Processor architecture, delivers high-performance LLM and multimodal model deployment capabilities while maintaining a radically efficient 150W power profile.



150W	256MB	48GB	1.5TB/s	512 TFLOPS (FP8)
TDP	On-chip SRAM	HBM3 Capacity	HBM3 Bandwith	BF16, FP8, INT8, INT4 Support

Llama 3 8B
2048 input tokens / 2048 output tokens

- RNGD FuriosaSDK / FP8 / 3047 tokens/s
- H100 SXM TensorRT-LLM 0.11.0 / FP8 / 8399 tokens/s
- L40S TensorRT-LLM 0.11.0 / FP8 / 1912 tokens/s

GPT-J
MLPerf Data center, closed, offline scenario / 99.9% accuracy

- RNGD FuriosaSDK / FP8 / 15.13 queries/s
- H100 SXM TensorRT-LLM 0.11.0 / FP8 / 30.375 queries/s
- L40S TensorRT-LLM 0.11.0 / FP8 / 12.25 queries/s

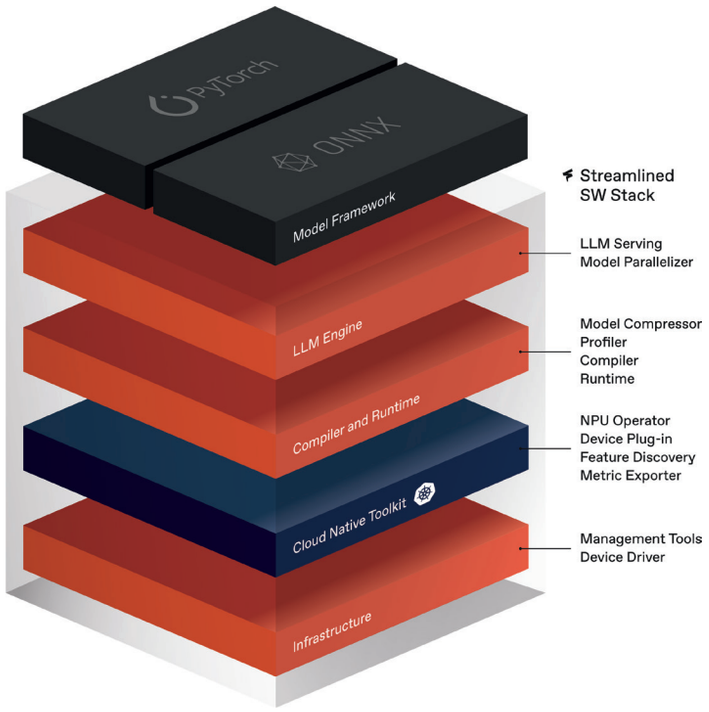
Disclaimer: Measurements by FuriosaAI internally on current specifications and/or internal engineering calculations.
Nvidia results were retrieved from Nvidia website, https://github.com/NVIDIA/TensorRT-LLM/blob/main/docs/source/performance /perf-overview.md, on Aug 25, 2024.

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.

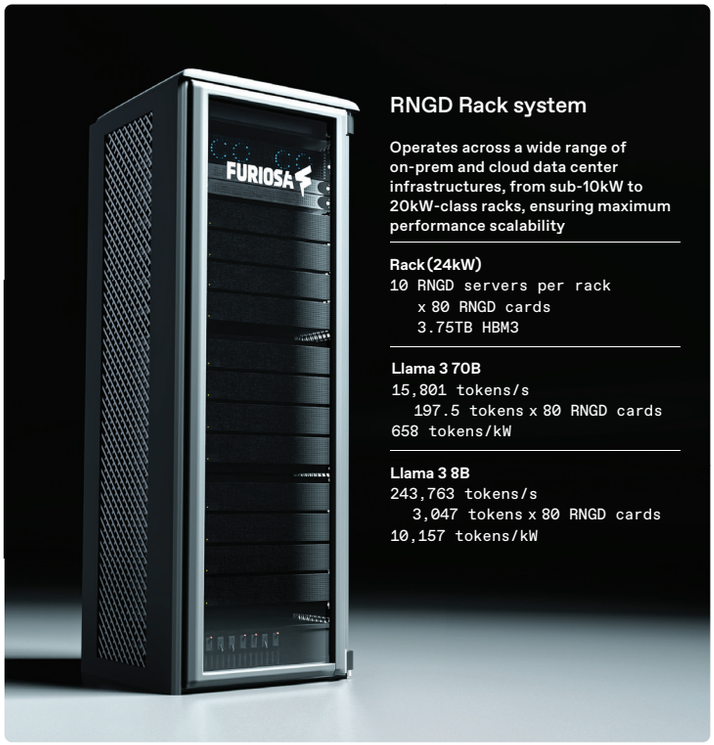
Robust Ecosystem Support
Effortlessly deploy models from library to end-user with PyTorch 2.x integration. Leverage the vast advancements of open-source AI and seamlessly transition models into production.

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	1024 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)	150 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



Be the first to know about RNGD availability and product updates — sign up now on furiosa.ai