



FURIOSA



V-RAPTORTM NEO

High-Efficiency AI NPU Arm Server for LLMs & On-Premises AI

4U Form Factor	64 Cores	4X NPU
--------------------------	--------------------	------------------



Lower Initial Investment

Lower initial investment for NPU inference servers compared to GPU servers.



512TFLOPS^{FP8} at 180W TDP

The FuriosaAI RNGD NPU delivers efficient & stable AI inference through an inference-optimized architecture.



Purpose-Built for AI

Ensure stable service operations through architecture purpose-built for AI & LLM inference.



Provides A Stable Arm Server Platform for AI Inference & LLM Workloads.



The True Cost of AI is Driven by Data, Not Just Compute.

Inference PoC Support



Configuration & Architecture Consulting



Performance Validation & Benchmarking



Technical Support & Operational Guidance

1 ⚡ Furiosa RNGD Delivers Higher LLM Inference Efficiency than H100

Total Design Power

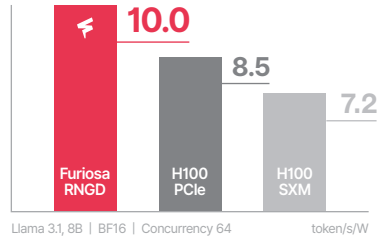
180W

Maximum Performance

512TFLOPS

High-Bandwidth Memory

48GB HBM³



More Efficient than GPU in Power, Cost & Concurrency

Furiosa RNGD-based inference servers offer lower initial deployment costs than GPU-based servers & enable stable AI service through an optimized architecture.

2 Increase Performance & Rack Density

Space Reduction¹

3.0X

Power Reduction¹

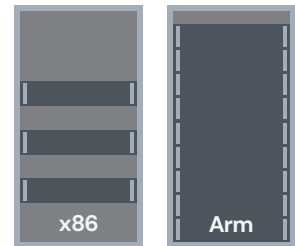
2.8X

Performance¹

2.5X

Fit More Servers in The Same Rack

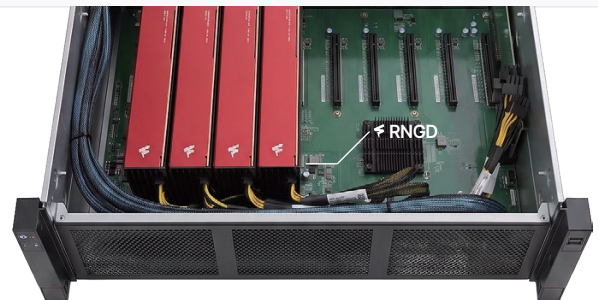
Because V-Raptor™ generates less heat and consumes less power than x86 servers, more systems can be deployed in the same rack while delivering higher performance.



3 Support for A Wide Range of LLMs



4 Up to 4X RNGD NPUs



Specification

1. Ampere Altra Max compared to x86(AMD EPYC 3rd Gen, intel Xeon 3rd Gen) server.

Model	V-Raptor™ NEO	NPU	FuriosaAI RNGD			
Server Type	Rackmount Server	NPU Performance	Up to 512TFLOPS FP8			
Form Factor	4U	NPU TDP	Up to 180W			
Size	790 × 433 × 176.5, mm (D x W x H)	Maximum NPU	Up to 4 NPUs			
Processor	Ampere Altra Q64-22 64-Core, 2.2 GHz	OS	Name		Available	Recommended
CPU TDP	Up to 95W		Ubuntu Server		Up to 24.04 LTS	22.04 LTS
CPU Architecture	Arm v8.2+		SUSE Linux Enterprise		Up to 15 SP7	15 SP7
Memory	DDR4-3200 ECC RDIMM		Rocky Linux		Up to 10.1	9.7
Maximum Memory	Up to 2TB (1 DPC)		Red Hat Enterprise Linux		Up to 10.1	9.7
Storage	2.5″ or 3.5″ SATA or SAS					
Maximum Storage	Up to 4 Bays	OS Installation		Included		
PSU	3+1 Redundant 2,600W PSU	Warranty		From 1 Year Extensible		

XSLAB 엑세스랩

엑세스랩 주식회사 | XSLAB Inc.
sales@xslab.co.kr

Seoul Office

R-1701, 02, 03, Daeryung Post Tower 8th,
43, Digital-ro 26-gil, Guro-gu, Seoul, Korea.

Jeju Office

402, Dongyeon Tower,
4, Sinhyeong-ro, Jeju-si, Jeju, Korea.

