neuchips

A Purpose-Built

LLM

Accelerator

Llama N3000 Accelerator

Designed for AI Inferencing

The N3000 AI Accelerator leads the industry in performance and efficiency. Purpose-built, the LLM Accelerator improves inferring accuracy, optimizes bandwidth, and increases resource utilization - all while delivering a significantly lower TCO than existing solutions.

Work Group

Host CPU

Offload

N3000 Accelerator

CPU + ASIC Combo

The Perfect Blend of Power & Efficiency

Combining CPU with our ASIC technology provides a more balanced and efficient solution compared to existing solution . Llama N3000 AI Accelerator, supports up to 65B parameters, is the one of the few that has using 7nm ASIC with LPDDR5 and PCIe Gen5 spec at max. 50W TDP.

Optimized for Max. Returns

Maximize your returns with a product that combines low power consumption and high accuracy, offering an attractive ROI. Our ASIC technology is incredibly energy-efficient, slashing operational costs and reducing your carbon footprint.

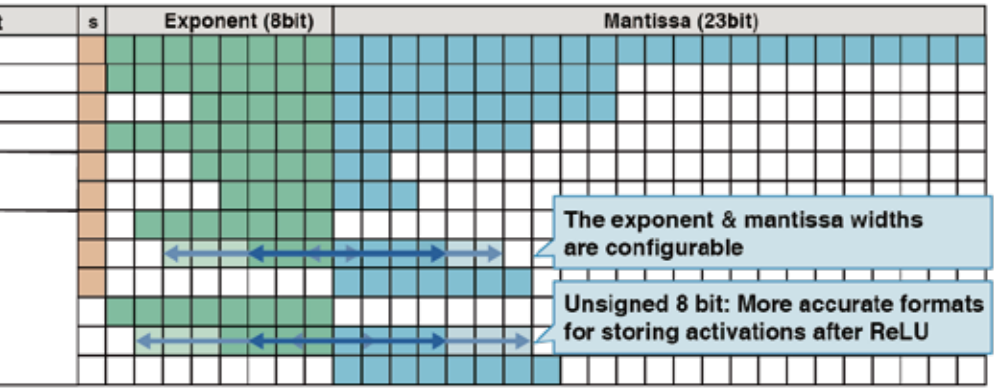
Precision Meets Accuracy

Achieve unparalleled accuracy with minimal perplexity while enjoying rapid inferencing speeds, while keeping power consumption in check, ensuring your business reaps high profits.

Zero Resource Wastage

With ASICs, every bit of processing power is dedicated to your tasks, making your operations more efficient. ASICs are purpose-built for specific tasks, ensuring optimal performance tailored to your needs.

Usage	Format	#bit
Training	FP32	32
Inferencing	TF32	19
	FP16	16
	BF16	16
	NVIDIA FP8	8
	NEUCHIPS FFP8	8

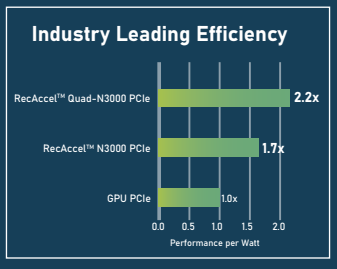


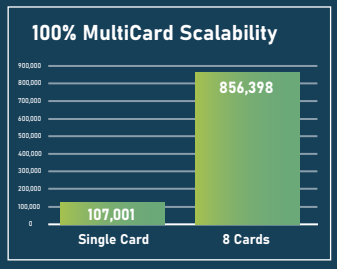
Proprietary Patented FFP8
The Ultimate Memory Fix
Solve LLM Memory Bound Challenge

Feature Highlights

- Easy-to-use Software Stack
  - Frictionless transition from existing development kit
- Patented Flexible Float Point 8
  - Improves inferencing accuracy
  - Quantizes data for memory capacity and bandwidth optimization in LLM workload
- 100% Linear Scaling to Multi-cards
- 4x 64bit LPDDR5 (6.4GHz) with ECC
- Up to 128GB on Card LPDDR5
- PCIe Gen5

Industry Leading Performance





Product Specification

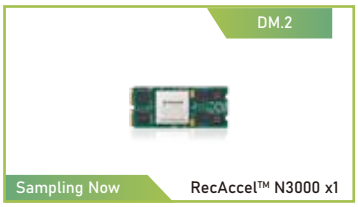
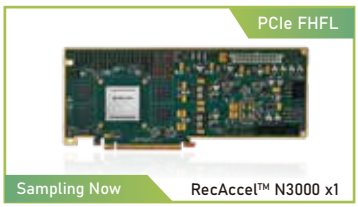
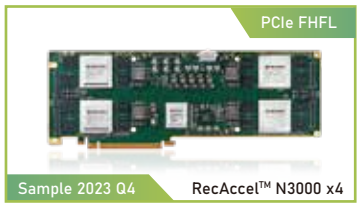
BFLOAT16	32 TFLOPS
INT8	206 TOPS
Memory	32GB LPDDR5
Memory Bandwidth	200 GB/s
Thermal Design Power (TDP)	70W
Form Factor	Full-height, Full-length (FHFL) 10.5" Dual-slot (266mm/10.5 inch)
PCI Express Interface	PCI Express 5.0 x 16 Lane and Polarity Reversal Supported
Interconnect	PCIe Gen5: 64 GB/s
Server Options	Partner and NEUCHIPS-certified Systems with 1-8 Accelerator

Environmental Specification

Ambient Operating Temperature	0°C~50°C
Storage Temperature	-40°C~75°C
Operating Humidity	5%~85% Relative Humidity
Storage Humidity	5%~95% Relative Humidity

MLPerf™ Proven AI Solution

Applications	Recommender, Gen AI...
AI Models	DLRM | LLaMA | GPT...
Frameworks	ONNX PyTorch C++
SDK	RecAccel™ SDK

	RecAccel™ N3000 DM2	RecAccel™ N3000 PCIe	RecAccel™ Quad-N3000 PCIe
			
DLRM Performance	20M inf/sec	30M inf/sec	120M inf/sec
Llama-2 7B[1] (Single Card)	40 tokens/sec	100 tokens/sec	400 tokens/sec
Llama-2 7B[1]	480 tokens/sec (YV 2.5: 12 Cards)	800 tokens/sec (Server: 8 Cards)	3,200 tokens/sec (Server: 8 Cards)
TDP	25 W	50 W	300 W
SRAM	160 MB	160 MB	640 MB
LPDDR 5 Capacity	32 GB	32 GB	256 GB
LPDDR Bandwidth	200 GBps	200 GBps	800 GBps

[1] *Llama-2 7B performance is based on batch size 16.*

Los Altos | Hsinchu | Taipei
neuchips.ai
contact@neuchips.ai