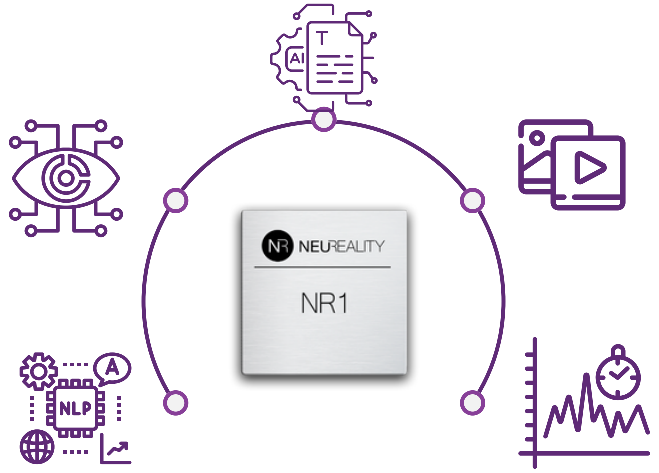
AI Adoption at the Speed and Scale of Your Business

NeuReality delivers the only open, purpose-built AI Inference system architecture, powered by a unique 7nm NR1 Chip to complement any AI Accelerator or GPU.

Unlike traditional CPU-based AI systems, our game-changing NR1 technology unlocks the maximum capability of all GPUs and Accelerators, super boosting them from <50% today to 100% full utilization. That means you get MORE out of your expensive GPU investments, along with scalable AI performance.

Supports Leading AI Technologies



Enables Any Single or Multi-Modal AI Workload

- Finance & Banking
- HealthTech & Insurance
- Biotechnology & Life Sciences
- Government & Smart Cities
- Telecom & Call Centers

Ready to Transform Your AI Capabilities?

Contact us today to learn more about the NR1 AI Inference Appliance and how it can revolutionize your business.

Contact Information

Email: sales@neureality.ai

Website: www.neureality.ai/solutions

bitly

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product.

Experience *Revolutionary* Intelligence with the

NR1® AI Inference Appliance

Fully Compatible with any AI Accelerator

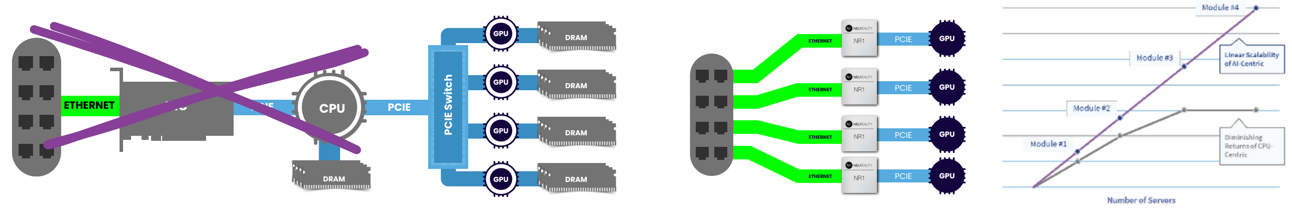
- Perfect companion for any type or number of GPU or AI Accelerators
- 100% linear scalability without performance drop-offs or delays by pairing NR1 with any AI Accelerator or GPU.
- Disruptive technology with 50-90% improved price/performance and the lowest cost per AI query with NR1 versus host CPU and NIC-centric architecture
- Lower environmental footprint with up to 13-15x greater energy efficiency
- Enterprise-ready, out-of-the-box software development and APIs for improved customer experience, faster time-to-market and affordable AI applications

NEUREALITY

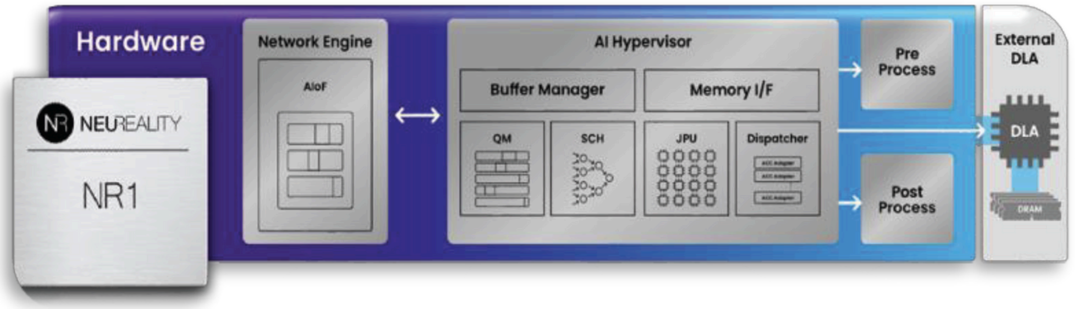
contact us:
sales@neureality.ai or visit www.neureality.ai

From CPU-Centric to Disaggregated AI Inference Architecture

Eliminating system overheads for better performance



Hardware offload of Networking, Data Movement, Processing and Sequencing



High Compute Density, Cost and Power Efficiency

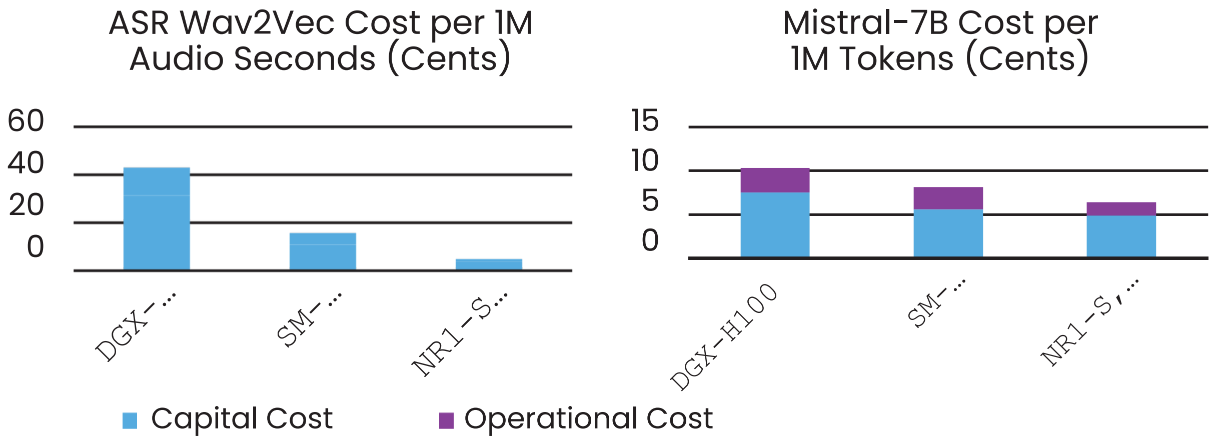
Best-in-class Total Cost of Ownership

Media Processing

Up to 90%

Large Language Models

Up to 50%


ASR Wav2Vec Cost per 1M Audio Seconds (Cents)

Mistral-7B Cost per 1M Tokens (Cents)

NR1® Chip

First True AI CPU Built for Inference at Scale

Linear Scalability	Ultra-Low Latency	Complete offload	Quality of Service	Versatility
Uniform distribution of bandwidth between all clients w/o degradation	Scalable network performance with NR1 AI-over-Fabric™ technology	Hardware-based data movement and processing offloading, accelerated operator libraries	Self-managed in hardware versus software-based SLA enforcement with bottlenecks	Unmatched versatility and adaptation to a wide range of AI applications with ease

Media and Data Compute

4x Video/Image decoders
16x Audio/Speech DSPs
16x vector GP-DSPs
Operator libraries

NR1™ AI-Hypervisor™ technology

Hardware-based sequencing, QOS and data movement

64K Queues, 16K schedulers, 64K Rate Limiter

AI-over-Fabric™ network engine

2x 10/25/50/100 GbE
Efficient AI-over-Fabric (TCP / ROCEv2)
Line rate cryptography
2 tiers of isolated network functions

Performance and Security

PCIe GEN5 x16(RC, EP, SRIOV, multi-device)

20 channels of LPDDR5 16bit 6400 MT/s (up to 160GB, 256GB/sec)

Secured boot support + Root of trust

Get More Out of Your GPUs

Best-in-class Total Cost of Ownership, 50-90% Price/Performance Gains vs CPU-Reliant Systems

- Perfect scalability
- Reduced latency
- High energy efficiency
- High compute density

Mechanical Form Factor	4U, 19" Rack Mount
PCI Express Capability	20 slots of PCIe Gen5 x 16
Compute Capability	Up to 10 NR1 cards carrying up to 16 GPU/AI accelerators
Storage	Up to 10 x 3.84TB E1.S
Power	2+2 Redundance mode, Typical: 3.2KW
Cooling	6 modules, each 2x60x60 dual rotor fans
Software	Server configuration, monitoring, and network security