Edge-corE NETWORKS

An Open Infrastructure Solution for Enterprise AI

Nexvec™

May 2025

Nexvec™: An Open Infrastructure Solution for Enterprise AI

Executive Summary:

Nexvec™ is an Edgecore trademarked turnkey solution with purpose built open infrastructure. It integrates disaggregated open Networking, composable Computing, and Storage architectures, along with seamless Day 0 through Day 2 lifecycle management as a total solution for Enterprise AI workloads.

Nexvec™ is designed to be deployable, scalable, manageable, and programmable, meeting the infrastructure demands of next-generation Enterprise AI, particularly for inference, reasoning, and emerging Agentic AI use cases. These technologies aim to retool enterprise workflows, automate cross-functional processes, and enable contextual reasoning by unlocking the value of enterprise data.

This document explores the rapid evolution of AI/ML technologies in the enterprise market, the emerging challenges that organizations face for the adoption of AI solutions effectively, and how **Nexvec™** addresses these needs. It highlights the critical role of open infrastructure in supporting AI workloads and outlines its benefits, particularly as enterprises prepare for the next wave of innovations in the Agentic AI era.

Market movements

The AI/ML landscape is undergoing a rapid transformation, with large language models (LLMs) dominating the spotlight. Innovations in this space are advancing at an unprecedented pace, reshaping industries and redefining the future of digital transformation. Enterprise CXOs can no longer afford to overlook this movement. Most organizations are seeking AI solutions that deliver tangible business outcomes — rather than simply eyeing on LLMs.

Recent breakthroughs in reasoning models, inference, and Agentic AI are paving the way for enterprises to monetize their data, retool business operations, and maximize outcomes. The inference and Agentic AI market is poised to become a key profit center in this next wave of AI evolution. However, enterprises face significant hurdles — including concerns around vendor lock-in, data privacy, data sovereignty, and the growing need for real-time decision-making.

Our rapid business growth over the past two years has been fueled by the increasing demand for AI/ML infrastructure. We have been deeply engaging across a wide range of market segments — from Hyperscalers to Fortune 500 and Global 2000 enterprises — collaborating to build next-generation AI/ML infrastructures. Through joint development and co-innovation with our customers, we have gained invaluable insights and know-how into the operational challenges enterprises face when adopting AI at scale.

Despite growing momentum, most enterprises are still in the early stages of AI adoption — struggling to fully grasp the complexity of AI workloads, let alone establish operational models that effectively deploy AI infrastructure. The intricate interdependencies between computing, storage, and networking add to these challenges, particularly as GPU-to-GPU communication becomes critical to performance across the entire AI stack.

Let's consider some real-world examples:

- **A U.S.-based Fortune 500 retail company** is eager to capitalize on AI's potential but lacks the full-stack in-house engineering expertise needed to integrate the critical infrastructure components required for deployment.

- **A global hospitality company** has embraced disaggregated open networking, open network operating systems, and composable computing architectures to avoid vendor lock-

in and enable more agile innovation. However, without sufficient in-house capabilities and capacities, they rely heavily on external consultants and professional services — diverting focus away from business outcomes and becoming entangled in infrastructure complexities.

- **A Fortune 500 IT company** seeks to implement multi-tenancy across departments, with each group aiming to adopt its own inference models or fine-tune foundation models. Yet concerns over data security have made them hesitant to transition to cloud-based AI services. They also face challenges in enabling collaboration across departments using Agentic AI to retool workflows efficiently.

Challenges

Below are just a few examples of the challenges enterprises face as they strive to harness AI effectively. The journey is complex, but with the right expertise and solutions, businesses can overcome these hurdles and unlock AI's full potential.

Through our collaboration with enterprise customers, we have identified four critical factors that stand out in successful AI adoption:

1. Full-Stack AI Inference Architecture

Enterprises require a robust AI inference architecture to drive business outcomes. AI workloads have fundamentally different infrastructure demands compared to traditional IT — and even within AI, inference introduces unique challenges across computing and networking. The heavy reliance on memory-bound compute creates new optimization requirements that must be addressed to support scalable AI inference operations.

2. Enterprise GenAI Transformation

GenAI, AI inference, reasoning models, and Agentic AI are redefining how enterprises extract value from their data and operations. Unlike traditional big data approaches that primarily "soak" data for analytical insights, these new models enable contextual reasoning across organizations, retooling workflows and accelerating innovation. Many enterprises are restructuring themselves to operate more like technology companies — not just to engineer

new business models from existing assets, but to move beyond proprietary constraints and unlock faster, more open innovation.

3. The Shift Toward Open Infrastructure

To keep pace with rapid AI-driven innovation, enterprises are increasingly embracing open infrastructure**,** aiming to avoid vendor lock-in, maintain the speed required for innovation, and leverage advancements from a broader ecosystem. However, they face significant challenges in balancing vertical innovation through proprietary, customized solutions that offer an integrated user experience with horizontal innovation enabled by the open ecosystem. At the same time, operationalizing disaggregated open networking, composable computing, and storage remains a major hurdle in fully realizing the benefits of openness.

4. Seamless AI Operationalization

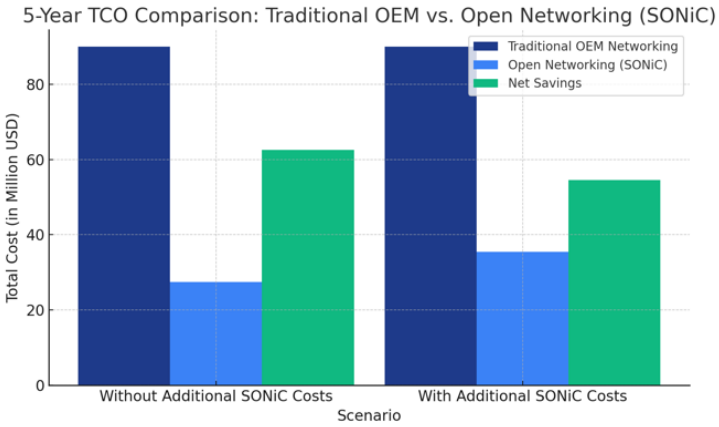
From Day 0 planning to Day 1 deployment and Day 2 operations, enterprises must integrate and operationalize AI building blocks sourced from open ecosystems — all while delivering a seamless, out-of-the-box user experience across the entire lifecycle. The ultimate goal is clear: enterprises must stay focused on driving business value, rather than becoming entangled in the complexities of infrastructure management.

At a high level, building AI infrastructure may not seem like rocket science, but in reality, it involves significant complexity due to the distinct nature of AI workloads. A scale-up architecture is essential to support large inference models, leveraging data parallelization, model parallelization, and Mixture of Experts (MoE) techniques to efficiently distribute data, model partitions, and expert computations across GPU nodes for optimal performance.

For enterprises, multi-tenancy is critical to support diverse inference models across departments. Rather than deploying multiple dedicated clusters, enterprises require dynamic GPU resource allocation with strong isolation — all managed through a single pane of glass using a scale-out architecture.

Finally, it's important to recognize that GPU workloads differ significantly from traditional CPU-based IT workloads. AI infrastructure relies heavily on GPU-to-GPU communication across the underlying fabric to maximize both performance and GPU utilization. This makes a fully

integrated end-to-end management plane — orchestrating both GPU computing and networking — essential for optimizing AI infrastructure at scale.



Openness is not new to us, as it has been a fundamental principle across various layers of the technology stack, from open-source software to open networking and open compute. The industry, including ourselves, has long emphasized the benefits of lower TCO, greater programming flexibility, and freedom from vendor lock-in.

This TCO analysis, generated by ChatGPT, aligns closely with our real-world industry experiences. The first part of the chart highlights the TCO savings 40% ~60% enabled by open networking, while the second part demonstrates that even with the addition of third-party professional services and support, the cost benefits remain substantial.

Yet, despite these clear advantages, why hasn't the adoption of openness in infrastructure accelerated at the pace we expected? Simply put, the answer is rooted in the added complexity associated with using "open".

Open networking enables rich innovations of both hardware and software individually through the process of disaggregation. Each component can be developed independently with the knowledge that they will be reassembled to achieve the intended value. However, reintegrating these disaggregated components into a seamless user experience has remained a challenge. It's like solving a scrambled Rubik's Cube, while all the pieces exist, putting them together efficiently requires significant expertise. Despite considerable progress, there is still no one-stop solution that delivers an out-of-the-box user experience for open infrastructure.

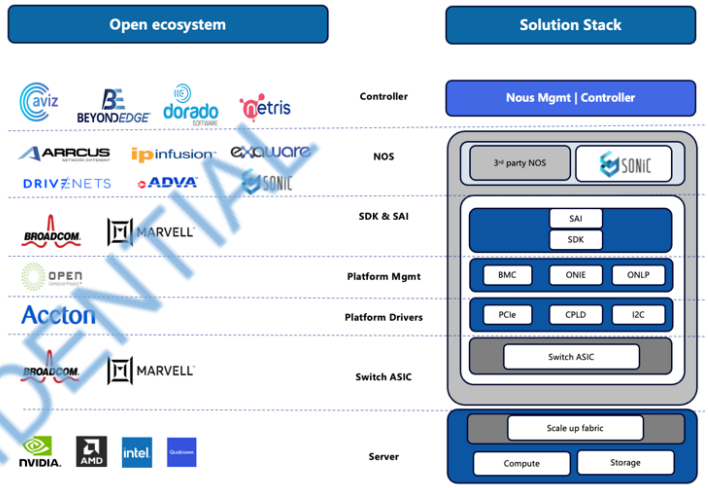
Feature completeness and product quality remain top priorities, and when it comes to AI data centers, the gap is steadily closing. However, brownfield adoption remains an often-cited concern, as new infrastructure must integrate seamlessly with existing environments.

Network automation has matured in enterprise IT, simplifying operations. On the computing side, GPUs for AI workloads lack an equivalent of VMware for CPUs, an abstraction layer that enables

efficient GPU management. GPUs are expensive and depreciate quickly, making optimization, cost efficiency, and rapid deployment critical. Avoiding over-provisioning, right-sizing resources, and accelerating production deployment instead of prolonged lab certifications are essential for maximizing ROI.

When networking and computing are combined to support AI workloads in a Scale-Up and Scale-Out architecture, the complexity doesn't just add up, in fact, it multiplies. Managing both simultaneously demands a holistic approach to ensure seamless integration and efficiency.

If we take a closer look at a typical stack, it spans multiple layers, starting from the compute layer, which includes GPUs, DPUs, memory, and storage, to the networking stack, covering switching ASICs, platform software, SDK/SAI, and network operating systems, all the way up to the management and orchestration layer. The open ecosystem offers a broad range of choices, and this snapshot provides a partial view of its landscape.

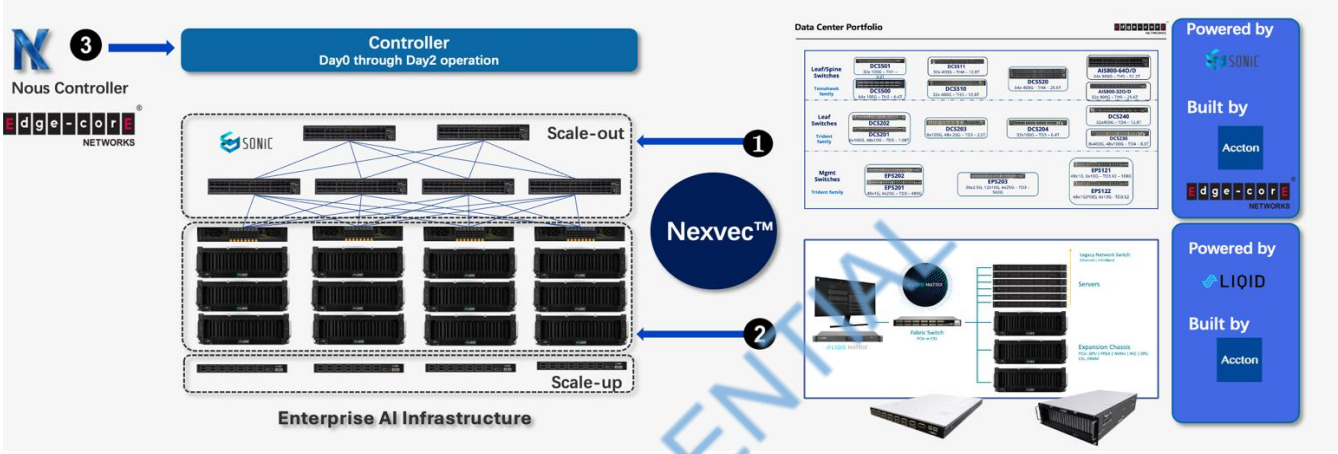


On the one hand, this diversity empowers customers with flexibility, allowing them to cherry-pick the best-fit components. On the other hand, seamless integration remains a challenge. Ensuring interoperability across different components is far from straightforward, especially when vendors operate on misaligned roadmaps driven by their own business priorities.

This is the whole reason we are coming up with **Nexvec™,** the next vector of Open Infrastructure to solve these challenges.

Nexvec™: An Open Infrastructure Solution for Enterprise AI

As a long-standing leader in the open infrastructure industry, we are not satisfied with the current state of adoption, where customers are left to piece everything together on their own to meet their business needs. To address this, we are taking a decisive step to introduce our integrated solution for a seamless user experience. **Nexvec™** is our answer.



Accton Edgecore has long been a leader in open networking hardware and network operating systems, including SONiC. We offer a comprehensive portfolio of data center switches, widely adopted by Fortune 500 and Global 2000 customers. Over the past few years, this portfolio has been successfully deployed in AI scale-out networks for live production environments.

Accton has partnered with Liqid and build a backend Scale-up fabric for composable computing with a PCIe and CXL switch, a 10-slot GPU server shelf and 10-slot memory server shelf, designed to scale-up AI workloads for dynamic resource allocations, particularly inference workloads. Liqid's orchestration platform enables dynamic resource allocation through GPU pooling and memory pooling, ensuring efficient utilization of compute resources. These allocations can be part of standard operational processes or may be programmatic in nature.

While there are various open and proprietary protocols for scaling AI infrastructure, our focus remains firmly on open architectures and solutions. Specifically, our initial **Nexvec™** solution will utilize open standards such as PCIe and CXL for interconnects. Emerging standards like UltraEthernet will also be explored as they mature and become production ready.

Furthermore, Accton Edgecore has introduced the Nous Controller — a centralized management and orchestration platform that unifies both the frontend and backend networks, extending the fabric management all the way to GPU endpoints like SmartNICs. To maximize GPU-to-GPU performance through RDMA Peer-to-Peer communication, extensive optimizations have been built into both the hardware and the Liqid software stack, delivering an enterprise-grade AI infrastructure.

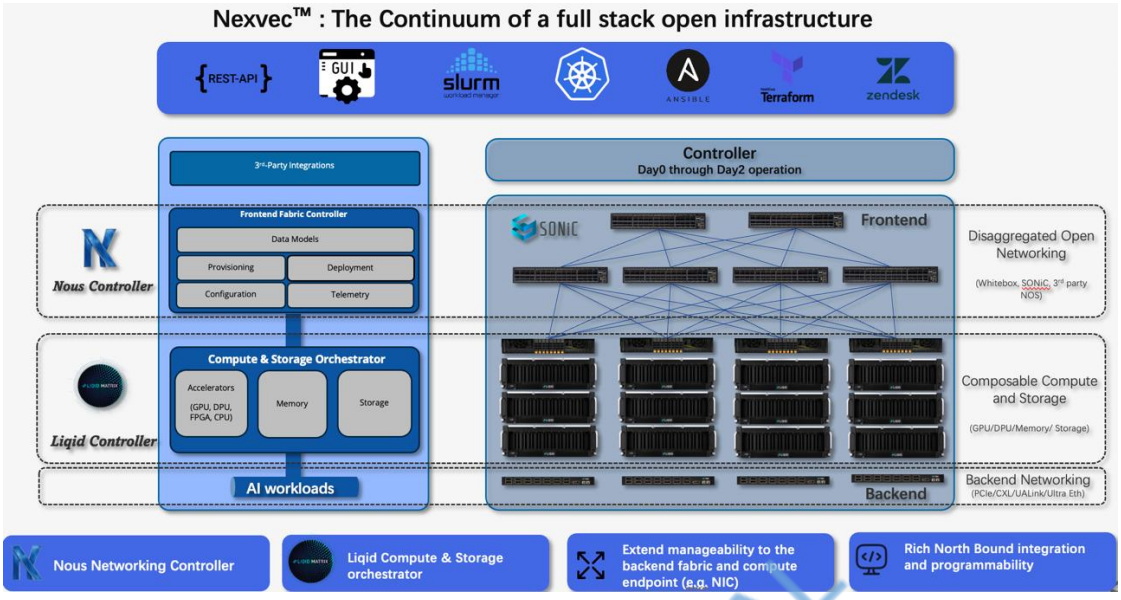
While we remain fully committed to an open partner ecosystem, we believe it is critical to accelerate the adoption of open infrastructure and demonstrate its full potential for AI workloads. This initiative is more than just integration — it's about advancing the competence and maturity of open solutions to meet the demands of modern AI-driven enterprises. The solution will be pre-racked and rolled out for customer deployments, delivering a true out-of-the-box experience.

This is why we call it **Nexvec™** — an innovation that extends the spectrum of openness across Networking, Computing, and Storage. In short, **Nexvec™** is a turnkey open infrastructure that is scalable, deployable, manageable, and programmable for Enterprise AI solutions. It features disaggregated open networking, composable computing and storage architectures, and is designed to optimize performance, utilization, and flexibility across the entire AI stack.

Further, **Nexvec™** enables a seamless continuum of operations across the entire lifecycle - from Day 0 planning to Day 1 deployment to Day 2 operations, ensuring comprehensive infrastructure management.

- The ***Nous*** Controller orchestrates and manages the front-end disaggregated open networking infrastructure.

- The Liqid Controller, ***Liqid Matrix,*** orchestrates and manages backend computing, storage, and networking, bringing dynamic resource allocation to AI workloads.



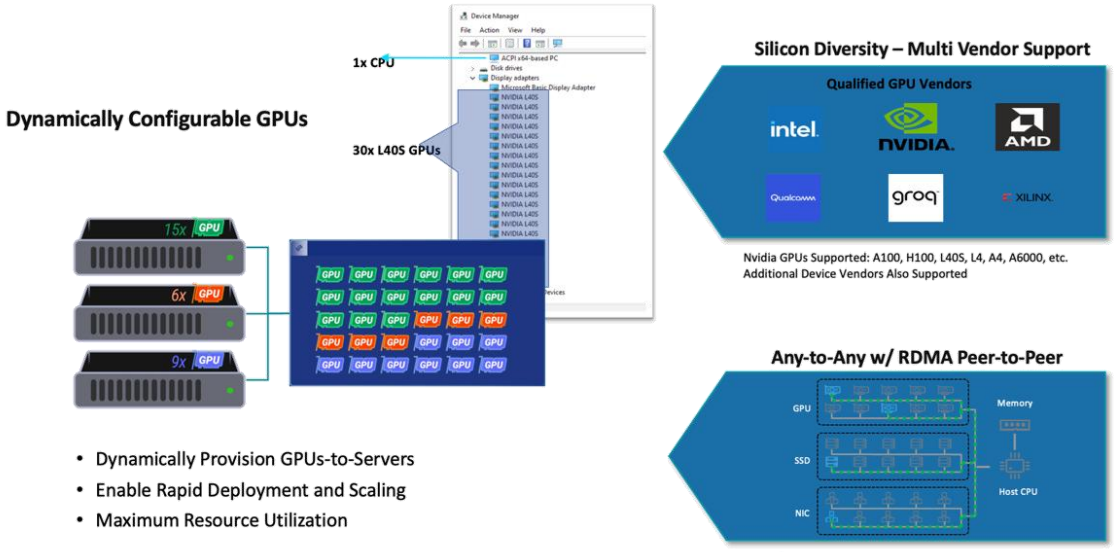
For AI-specific workloads, as previously noted, the network must be workload-aware to support high-throughput GPU-to-GPU peer-to-peer communication. This is critical to ensure low latency while preserving model quality. To achieve this, the Nous Controller will provide visibility into both the backend network and GPU server connectivity endpoints like SmartNICs.

This architecture is designed for seamless integration into existing enterprise brownfield environments. It is API-rich and supports a broad ecosystem of orchestration and automation tools, including:

- Job schedulers such as Slurm
- Kubernetes-based orchestration for AI workloads
- Automation frameworks like Ansible, etc.
- IT service management integration with Zendesk for ticketing and operational support

Built on open technologies, this full-stack infrastructure delivers a scalable, flexible, and automated deployment experience for Enterprise AI.

Now, let's take a closer look at how this architecture enables high-performance, dynamically configurable GPU and memory allocation for enterprise AI workloads, particularly for inference.



Conceptually, this approach is similar to some extend about how VMware virtualizes compute resources, but in this case, it slices and allocates GPU resources across GPU servers, creating virtualized GPU pools. Each pool can support different inference models with multi-tenant isolation, ensuring efficient resource utilization.

Key capabilities of this architecture include:

- Support for heterogeneous GPU environments: The system can host GPUs from different vendors, though only GPUs of the same type can form a GPU virtual cluster.

- PCIe compatibility: Any GPU that is PCIe-compliant and supports standard PCIe protocols can be integrated into the extension shelf.

- Optimized Any-to-Any GPU-to-GPU communication:

    - RDMA Peer-to-Peer allows direct GPU-to-GPU data exchange without CPU or host memory involvement.

    - This is particularly critical for HPC and AI/ML workloads, where minimizing latency and maximizing bandwidth are essential for performance.

By enabling dynamically configurable resources, enterprises gain the agility to optimize GPU utilization, allocate resources on demand, and ensure AI workloads run efficiently without over-provisioning.

It can scale up compute resources from a 10-node POD to 20-node and 30-node configurations and further scale out via the frontend fabric, enabling multi-tenancy support. All operations and management are seamlessly unified under a single pane of glass for efficient and streamlined control.

This solution is well-suited for enterprise adoption. Consider the recently popular DeepSeek model as an example—it requires only few NVIDIA L40S GPUs to support language models ranging from smaller amount of parameters, while thirty L40S GPUs can handle a 671B parameter model for enterprise-grade applications.

Inference workloads are memory-bound, meaning that higher-performance GPUs such as the G200 or H100 require only 9 to 11 GPUs to process the Llama 405B model efficiently.

Composable architecture provides the flexibility to disaggregate and reconfigure compute resources (such as GPUs) with various deployment options. At the same time, from a benchmark performance perspective, it delivers on-par performance compared to traditional fixed server configurations.

In short, composable compute infrastructure optimizes performance, reduces TCO, and enhances agility, making it a compelling solution for modern AI workloads.

Here are two illustrative examples showcasing the value of open infrastructure in AI deployments:

- **Example: AI Inference System (90-GPU Configuration)**
    - Traditional Setup: Requires 22 CPU nodes, 88 GPUs per rack, and consumes 65kW per rack.
    - Nexvec™ solution: Uses 3 fabric units, supports 90 GPUs per rack, and reduces power consumption to 42kW per rack. (a power reduction of 35%)

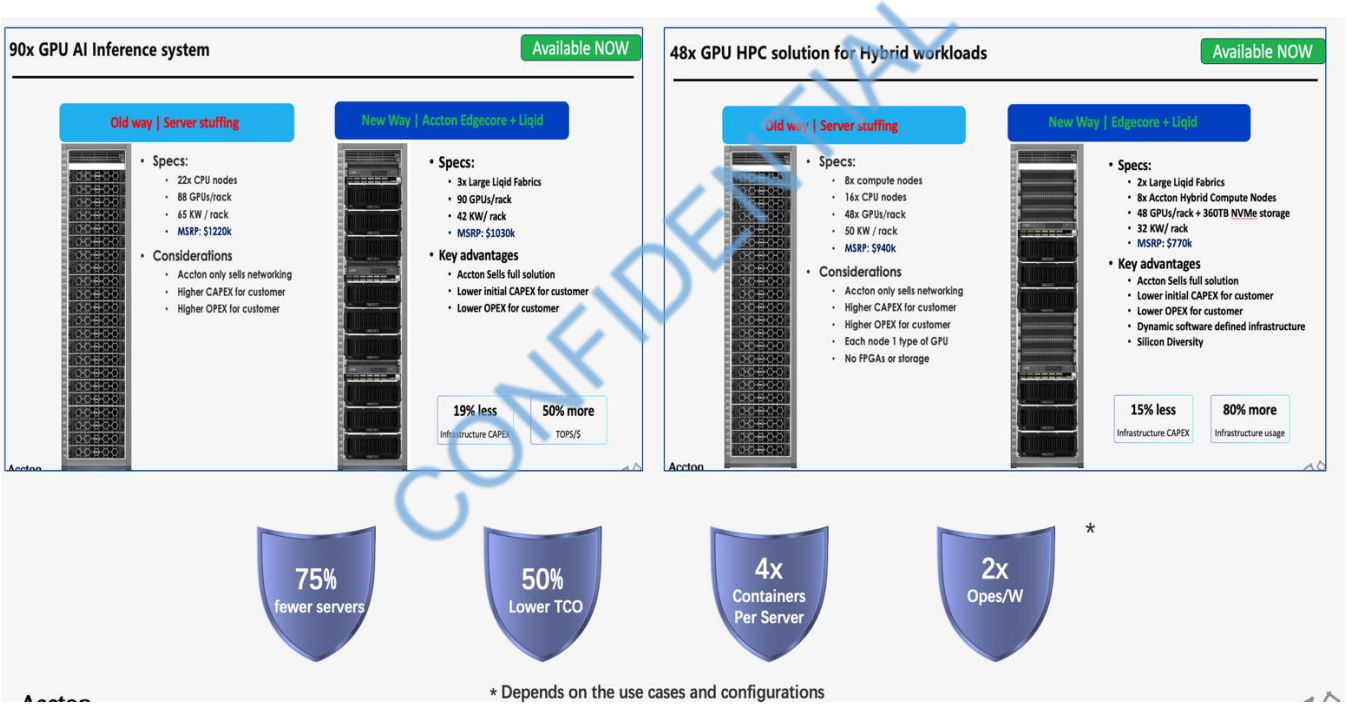
- Key Benefits: Delivers 50% more TOPS while reducing CAPEX by 20%.

- **Example: HPC Solution (48-GPU Configuration)**

  - Traditional Setup: Requires 8 compute nodes, 16 CPU nodes, provides 48 GPUs per rack, and consumes 50kW per rack.

  - Nexvec™ solution: Requires only 2 fabric units, supports the same 48 GPUs per rack, but reduces power consumption to 32kW per rack. (reduced by 35%)

  - Key Benefits: Achieves 80% infrastructure utilization while cutting CAPEX by 15%.



Takeaway:

Nexvec™, is an innovative Open Infrastructure solution for Enterprise AI infrastructure and spans the networking, compute and management domains. As an architected solution, it reduces the operational drag typically associated with modern AI-centric IT infrastructures, and enables higher performance, improved utilization, and significant TCO savings, marking it a compelling choice for modern AI workloads. It is designed and developed to deliver a simpler out-of-the-box experience, enabling enterprises to seamlessly adopt open infrastructures. It offers flexible, dynamically configurable resource allocation through GPU pooling and memory pooling, optimizing AI workloads. **Nexvec™** powers both scalable on-demand resource allocation via Scale-Up and Scale-Out strategies and by simplifying infrastructure lifecycle complexities, enterprises can concentrate on achieving business goals and reduce time to value without dealing with technical difficulties.

Legal Disclaimer:

The information contained in this document must be treated as confidential and proprietary to Accton-Edgecore. It is intended solely for the use of the individuals or entities to whom it is specifically addressed.

This document may contain sensitive business, financial, technical, or other information that is not to be disclosed to unauthorized parties. Any unauthorized review, copying, disclosure, distribution, or use of this information is strictly prohibited and may be unlawful.

You are further reminded that:

- This information is provided for discussion purposes only and does not constitute a binding offer, agreement, or commitment unless expressly stated otherwise in a separate written agreement.
- No representation or warranty, express or implied, is made as to the accuracy or completeness of the information contained herein.
- Any opinions or views expressed in this presentation are those of the presenter and may not necessarily reflect the official position of Accton-Edgecore.
- You are responsible for ensuring that your access to and use of this information complies with all applicable laws and regulations.

By accepting or viewing this document, you agree to maintain the confidentiality of the information contained herein and to not disclose it to any third party without the prior written consent of Accton-Edgecore.

About Accton-Edgecore Networks

Edgecore Networks Corporation, a wholly owned subsidiary of Accton Technology Corporation, is a leading provider of open infrastructure solutions. Edgecore Networks delivers comprehensive wired and wireless products and solutions through channel partners and system integrators worldwide, serving AI/ML, Cloud Data Center, Service Provider, Enterprise, and SMB customers.

Edgecore Networks is committed to advancing open infrastructure beyond networking. Edgecore is expanding its portfolio to include open compute solutions, enhancing its ability to deliver integrated infrastructure that meets the evolving needs of modern data centers and service providers

For more information about Nexvec™, visit us at:

www.Edge-core.com/Nexvec