SOLUTION SHOWCASE

NVIDIA

REAL-TIME AI PERFORMANCE AT THE EDGE

Image courtesy of Kinetic Vision

AI is Now Everywhere

As businesses across industries grapple with vast amounts of data, more complex operations, and more dynamic markets, edge AI is playing a growing role in helping them rapidly respond. Through a combination of computing power, AI technology, data analytics, and advanced connectivity, the edge extends compute capabilities from data centers out to the edge of networks, allowing organizations to act quickly on data where it's captured. Reducing the distance between where data is captured and where it's processed not only alleviates data transit costs, but also improves latency, bandwidth utilization, and infrastructure costs.

Addressing the Requirements of Edge AI

AI at the edge comes with a unique set of requirements. Edge systems, dispersed across vast physical distances, lack the centrality that a data center presents. Software or system updates either need to be performed manually or to be centrally managed to easily deploy, manage, and scale software across vast fleets of devices. Moreover, the security requirements for edge computing infrastructure differ from cloud or data center computing models. Edge locations lack the physical security that data centers have, so an end-to-end security model that protects both the application intellectual property and the sensor data is paramount for a successful deployment.

NVIDIA EGX Platform for Accelerated Edge AI

The NVIDIA EGX™ platform allows enterprise IT to deliver diverse applications on high-performance and cost-effective infrastructure. The platform is a combination of high-performance GPU computing and high-speed, secure networking in NVIDIA-Certified Systems™, built and sold by our partners. The EGX platform allows customers to prepare for the future while driving down costs by standardizing on a single unified architecture for easy management, deployment, operation, and monitoring. The EGX platform supports a vast suite of accelerated applications for edge AI, delivering faster insights where they matter the most.

Generate faster insights from AI deployed across thousands of devices using the NVIDIA EGX platform.

KEY USE CASES

> Inventory optimization and shrinkage reduction in retail stores

> Automation and quality control in manufacturing facilities

> Radiology and patient care in hospitals

> 5G multi-access edge computing (MEC) and virtual radio area network (vRAN) for telecommunications providers

> Freight tracking and route optimization for efficient logistics

PROOF POINTS

> Retailers use AI to reduce shrinkage—the loss of inventory from theft, errors, fraud, waste, and damage—which costs the industry an estimated $100 billion per year globally. AI helps retailers protect their assets with store analytics that monitor points-of-sale and floor merchandise to prevent ticket switching, misscans, and shoplifting.

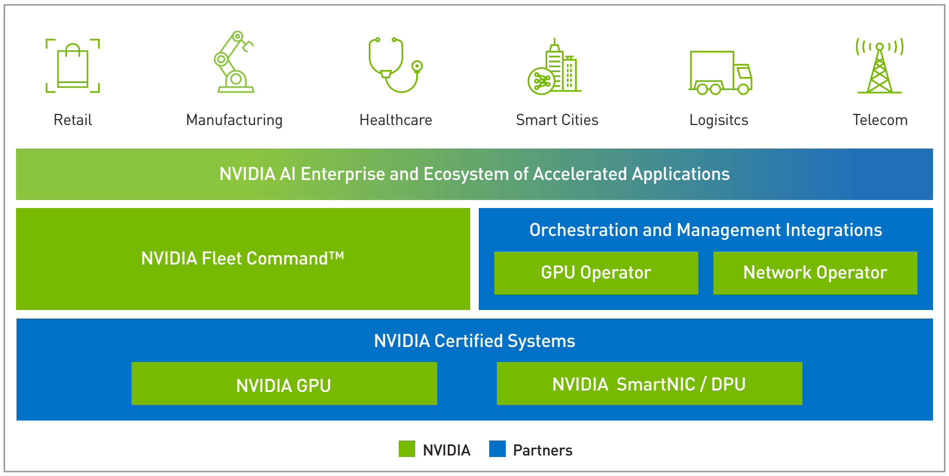

Figure 1. NVIDIA edge computing solutions bring together certified hardware, AI software, and turnkey management services that allow enterprises to harness the power of AI at the edge.

Securely Deploy, Manage, and Scale Applications with NVIDIA Fleet Command

NVIDIA Fleet Command™ is a managed platform for container orchestration that streamlines provisioning and deployment of systems and AI applications at the edge. It simplifies the management of distributed computing environments with the enterprise scale and resiliency of software as a service, turning every site into a secure, intelligent location. Layered security protocols managed by NVIDIA protect intellectual property and application insights as well as automatically identify and defend against threats. Fleet Command leverages a robust set of AI tools and a catalog of ready-to-deploy partner applications to create a secure, end-to-end platform for edge AI that's capable of operating in any environment. Fleet Command provides turnkey AI orchestration that keeps organizations from having to build, maintain, and secure edge AI deployments from the ground up.

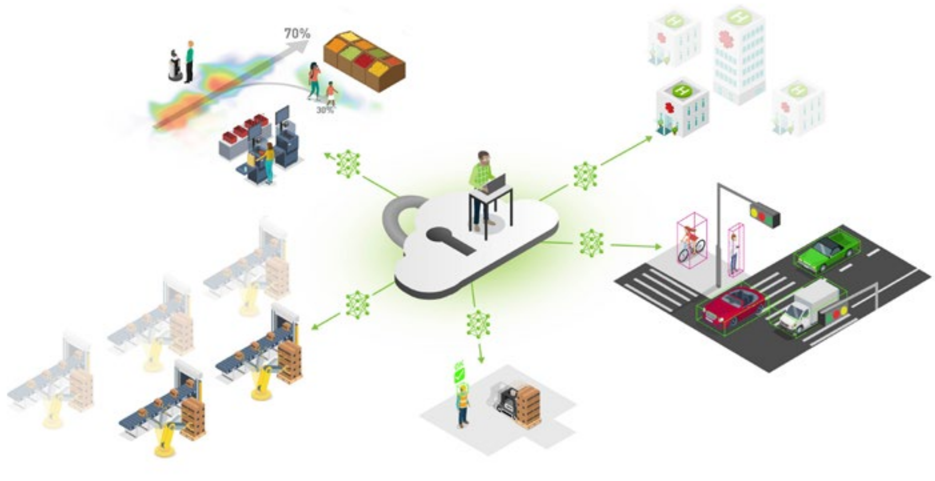

Figure 2. NVIDIA Fleet Command offers a simple, managed platform that makes it easy to provision and deploy AI applications and systems at thousands of distributed environments, all from a single cloud-based console

Figure 3. NVIDIA-Certified Systems bring powerful speedups to AI training and inference.

NVIDIA-CERTIFIED SYSTEMS

> NVIDIA-Certified edge servers are validated to run a range of accelerated workloads with the best performance, and must also meet specific requirements due to their deployment location.

NVIDIA.
CERTIFIED

Securing the Edge

In addition to accelerated computing and simplified deployments, NVIDIA solutions for edge computing offer industry-leading security protocols to ensure data is always protected. All processed data is encrypted in transit and at rest, and the secure and measured boot prevents AI runtime tampering. NVIDIA also provides ongoing managed security, with constant monitoring and automated bug fixes and patches, reducing the need for costly specialized staff to build and maintain these features.. Furthermore, AI applications deployed using Fleet Command are secure by design, built on a zero-trust architecture with layered security designed for edge environments, including a private application registry.

TRY ACCELERATED EDGE AI TODAY

NVIDIA LaunchPad, a program that gives enterprises and organizations immediate, short-term access to NVIDIA AI running on private accelerated compute infrastructure, speeds the development and deployment of modern, data-driven applications. Launchpad enables quick testing and prototyping across the entire AI workflow on the same complete stack you can purchase and deploy.