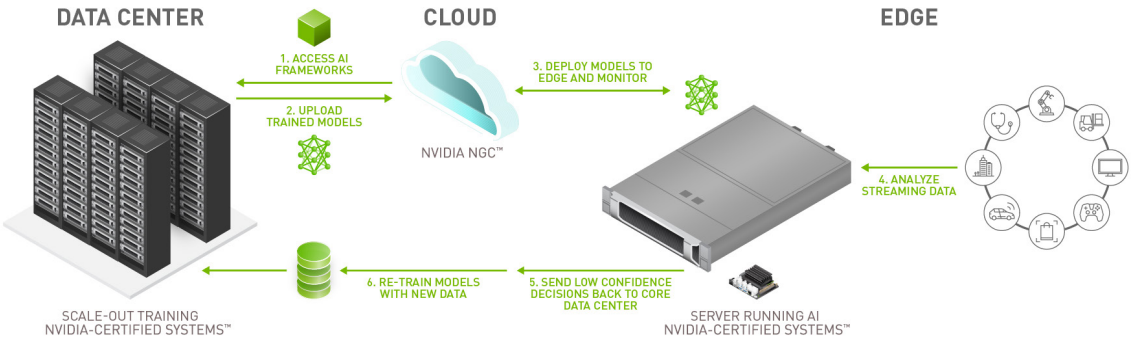
TECHNICAL OVERVIEW

TOP CONSIDERATIONS FOR DEPLOYING AI AT THE EDGE

Billions of IoT sensors—found in retail stores, in hospitals, on factory floors, and more—generate massive amounts of data with the potential to transform business. Because of this, **edge computing**, the process of bringing compute power to where data is collected, is one of the fastest growing trends in enterprise computing. By reducing the distance between where data is collected and where it's processed, organizations can react quickly to real-time insights, unlocking that potential.

Virtually every industry is investing in edge computing to accelerate AI workloads. Over the next four years, enterprise spending on edge hardware, software, and services will increase at an annual compound growth rate of 12.5 percent, amassing to an estimated $250 billion by 2024, according to IDC's 2020 Edge Spending Guide. But before investing in edge computing, organizations need to evaluate if edge computing is right for their needs.



IS EDGE COMPUTING RIGHT FOR YOU?

The first step to deciding whether edge computing is right for your organization is identifying if your use case is an edge use case. The best way to do that is by answering these two questions:

Is your solution "always on"? Always-on solutions are sensors or other pieces of infrastructure that are constantly working or monitoring their environments. Examples of "always-on" solutions include computer vision (as in security cameras for loss prevention), medical imaging (in ERs for surgery support), jobs that train hundreds or thousands of models (hyperparameter optimization), or large multi-week model training or simulation workflows. Examples of systems that aren't always on include small one-off simulations or model training jobs.

1 IDC Press Release, "Worldwide Spending on Edge Computing Will Reach $250 Billion in 2024, According to a New IDC Spending Guide," September 2020.

Does your solution use inference on the data collected from your sensors?
Inference is when data is collected from a sensor, and AI analyzes that data to either make sense of it or make decisions based on it.
Some examples could include obstacle detection for autonomous machines, object classification for smart retail stores, or conversational AI for smart hospital assistance.

If you answered "yes" to both of the questions, there's a good chance you'd benefit from edge computing.

THE TOP 5 BENEFITS—AND CONSIDERATIONS—OF EDGE COMPUTING

Now that you know if edge computing is right for you, take a deeper dive into its specific benefits—and important things to think about before choosing a solution.

Lower Latency: By bringing computing to where data is collected, the distributed infrastructure reduces a major cause of latency. Instead of losing time sending data back to the data center or the cloud, businesses can process data in real time. For example, intelligent sensors embedded in IoT devices can process data from autonomous machines and cameras on a factory floor and instantly alert workers about anomalies, malfunctions, and more. Beyond embedded devices, businesses can place edge servers in close proximity to the sensors—in a server room or closet in a store, hospital, or warehouse—to reduce latency even further. With this on-the-spot insight into their data, enterprises can optimize operations, boost safety in manufacturing, accelerate disease diagnosis, improve customer experiences with faster services, and much more.

Edge deployment is an ideal choice if your organization's current infrastructure isn't able to deliver the necessary real-time, low-latency inference and analytics required for your products and services.

Scalability: Beyond delivering high compute with low latency, integrating edge computing is an ideal way to scale infrastructure.

With a centralized cloud solution, there's limited bandwidth for moving and processing information. The bandwidth cap often leads to higher costs and smaller infrastructure. With edge computing solutions, the data collection

and processing happens on the local network, meaning bandwidth is tied to the local area network (LAN) and has a much broader range of scalability. And since processing happens on the LAN, the cost savings from not needing to move data back and forth to the cloud can be significant.

Remote Management: One daunting aspect of edge computing is the logistics of having the skilled data center technicians to set up and maintain several or even hundreds of systems at remote locations. To avoid this costly and time-consuming process, your edge computing solution must be easy to install and manage. But that still leaves the issue of installing and updating the software deployed on systems at edge locations. The answer to this problem is to adopt a management platform that allows system administrators or IT personnel to easily deploy, manage, and scale applications across the entire edge infrastructure from one location. If you're considering a large edge deployment, you should strongly consider a remote management platform as well. While many require developers to build and maintain them, some are turnkey solutions that can be set up in minutes.

Security: With distributed computing at the edge, security and data privacy quickly come to mind. Ensuring that the localized data, as well as the AI models trained to yield reliable insights, is protected becomes paramount. However, many edge computing solutions put the onus of security on the user. Unless you have security experts on staff, building a security model from the ground up can be prohibiting. That's why it's critical to choose an edge computing platform with secure features already available across the full stack. When evaluating computing platforms, pay close attention to data encryption in transit and at rest. By encrypting both, you greatly increase your security posture and help ensure your data is safe. Also, make sure the edge solution protects the AI runtime from being tampering with.

If you're investigating management platforms in addition to compute infrastructure, look for features that ensure your application is secure before you deploy it. These features can include malware and vulnerability scans or signed containers for more advanced platforms.

If you have a large security team dedicated to creating bespoke solutions, you have much more flexibility in choosing a platform with little or no security features. But in general, you can't have too much security.

Resilience: Since edge solutions are dispersed across many different locations, one pitfall of edge computing comes when a system fails or software has crashed. Theoretically, that means that a skilled IT professional would have to travel to that location and reboot the system or update software, a process that could be both time consuming and expensive. That's why it's important for edge systems to be paired with a remote management platform that takes advantage of resiliency.

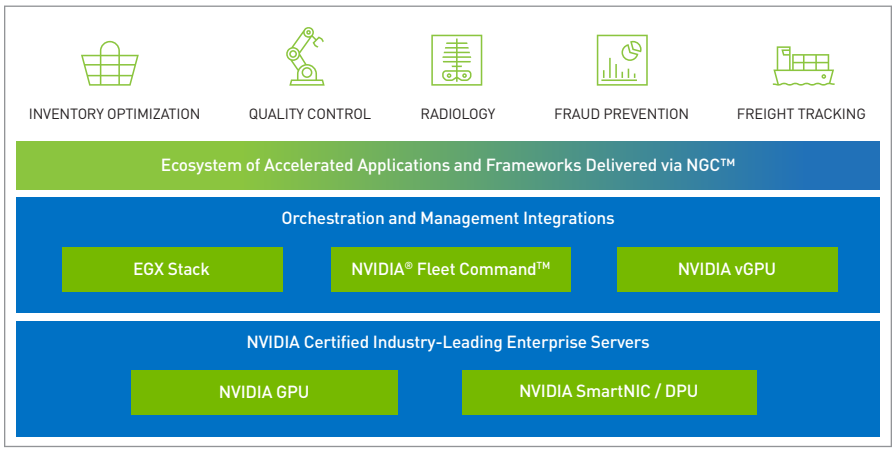
Resilient software is software that can remediate issues on its own without the help of human intervention. But in addition to self-healing applications, resiliency can migrate the workloads of failed systems to other systems on the same network, ensuring zero application downtime and insights that are never lost.

HOW TO DEPLOY AI AT THE EDGE

Getting started with an edge computing platform that offers all these benefits and features is simple. NVIDIA offers an end-to-end solution that brings edge computing intro reach for any organization. It's a two-part solution.

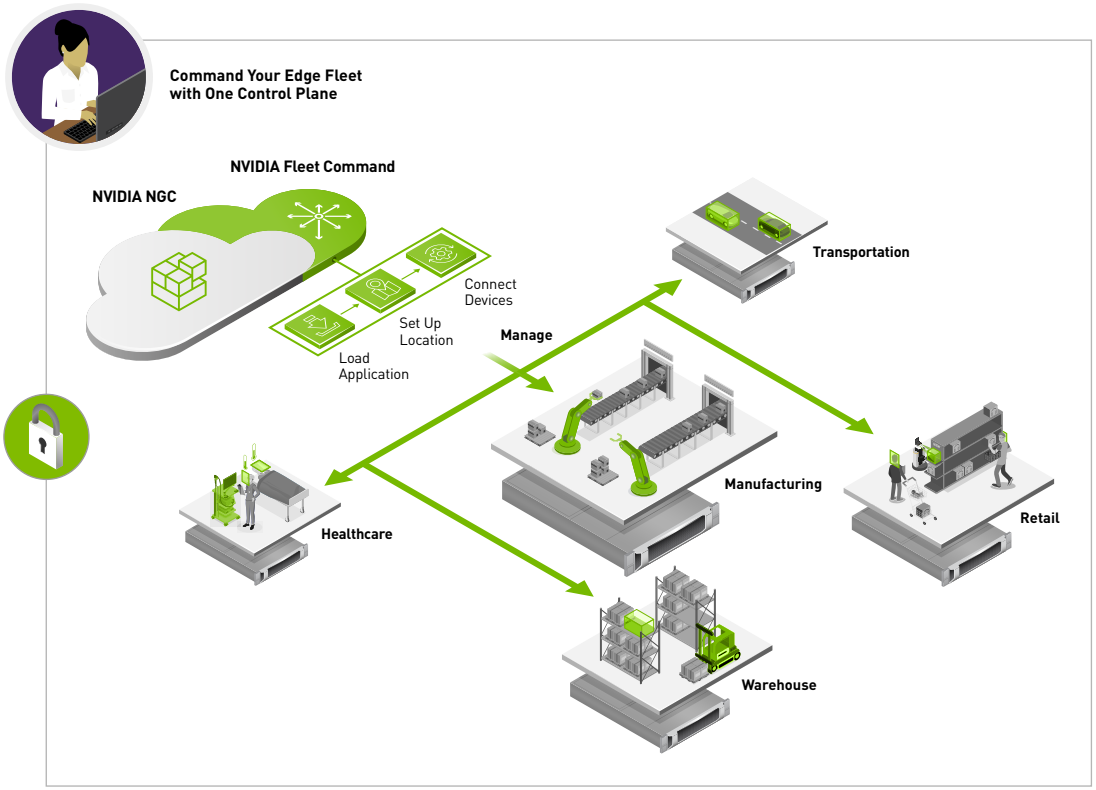
> The Powerful, Secure Platform for AI at the Edge

The NVIDIA EGX™ platform enables 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 NVIDIA 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.



> Easy Management at Your Fingertips

NVIDIA® Fleet Command™ is a hybrid-cloud platform to manage and scale AI deployments across dozens or up to millions of servers or edge devices. Fleet Command allows IT departments to securely and remotely manage a large-scale fleet of deployed systems. Instead of spending weeks planning and executing deployment plans, in minutes, administrators can bring AI to networks of retail stores, warehouses, hospitals, or city streets. Administrators can add or delete applications, update system software over the air, and monitor the health of devices spread across vast distances from a single control plane.



READY TO GET STARTED?

- Learn more about how the **NVIDIA EGX** platform is accelerating AI at the edge.
- Explore the thousands of accelerated AI applications available through the **NVIDIA NGC™ catalog**.
- Learn more about the benefits of an **NVIDIA-Certified System**.
- **Find an NVIDIA-Certified System** available through the world's leading server manufacturers.
- Learn more about deploying applications with the **Fleet Command Solution Brief.**