

The GIGABYTE logo is displayed in a bold, white, sans-serif font in the upper right corner. The background of the entire page is a close-up photograph of server hardware, specifically showing two NVIDIA EGX modules installed in a server chassis. The modules are silver with green accents and feature the NVIDIA logo. Various cables and connectors are visible around the modules. A green and white geometric pattern is overlaid on the left side of the image.

GIGABYTE™

EGX Edge AI Computing Platform

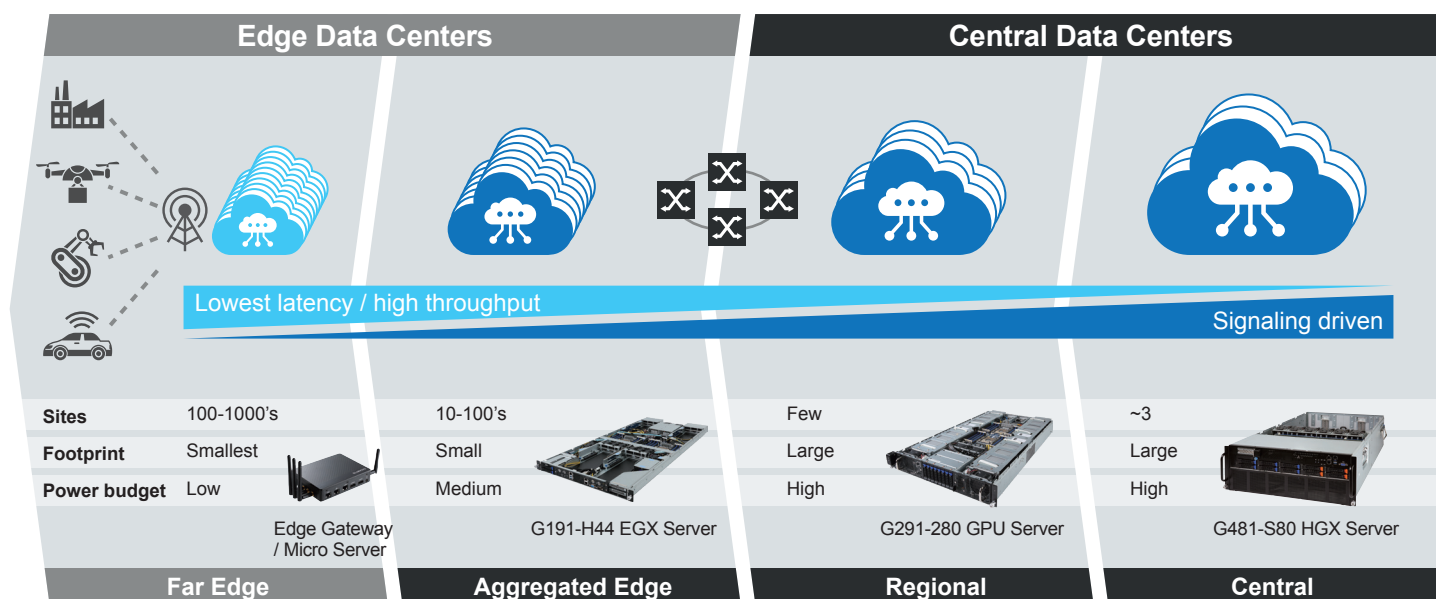


Presenting GIGABYTE's G191-H44 EGX Server, an edge supercomputing platform for IoT, AI, 3D, XR and beyond. EGX is a secure, cloud native platform designed to be deployed as an edge server or in an edge micro data center, and is optimized for processing IoT sensor information (such as video, sonic or LIDAR data) or for multimedia and interactive applications such as cloud gaming or augmented reality. GIGABYTE's G191-H44 EGX server platform features an integrated hardware and software stack, and is ready to use with popular container orchestration platforms such as RedHat OpenShift or deploy ready-to-use containers from NGC (NVIDIA GPU Cloud).

CHALLENGES

There is a big change happening in the computing industry, and it's driven by the massive volume of data that's been generated by billions of IoT sensors, as well as advancements in AI algorithms able to process all of this data. In order to be useful, analysis of this data needs to happen in real time so decisions can be made and acted upon quickly. For example, patients in hospitals can be diagnosed and detected with diseases as soon as an X-ray or MRI scan has been conducted. In smart intersections, vehicles can be identified and the traffic lights can be controlled, so emergency vehicles can be prioritized as well as buses.

The technology now exists to enable these real-time decisions to be made – along with AI algorithms and terabytes of unstructured data from billions of IoT sensors, we now have 5G networking and cloud-native applications driven by Kubernetes. Combining them together now makes it possible to deploy large-scale AI at the edge of the network, away from a centralized data center.



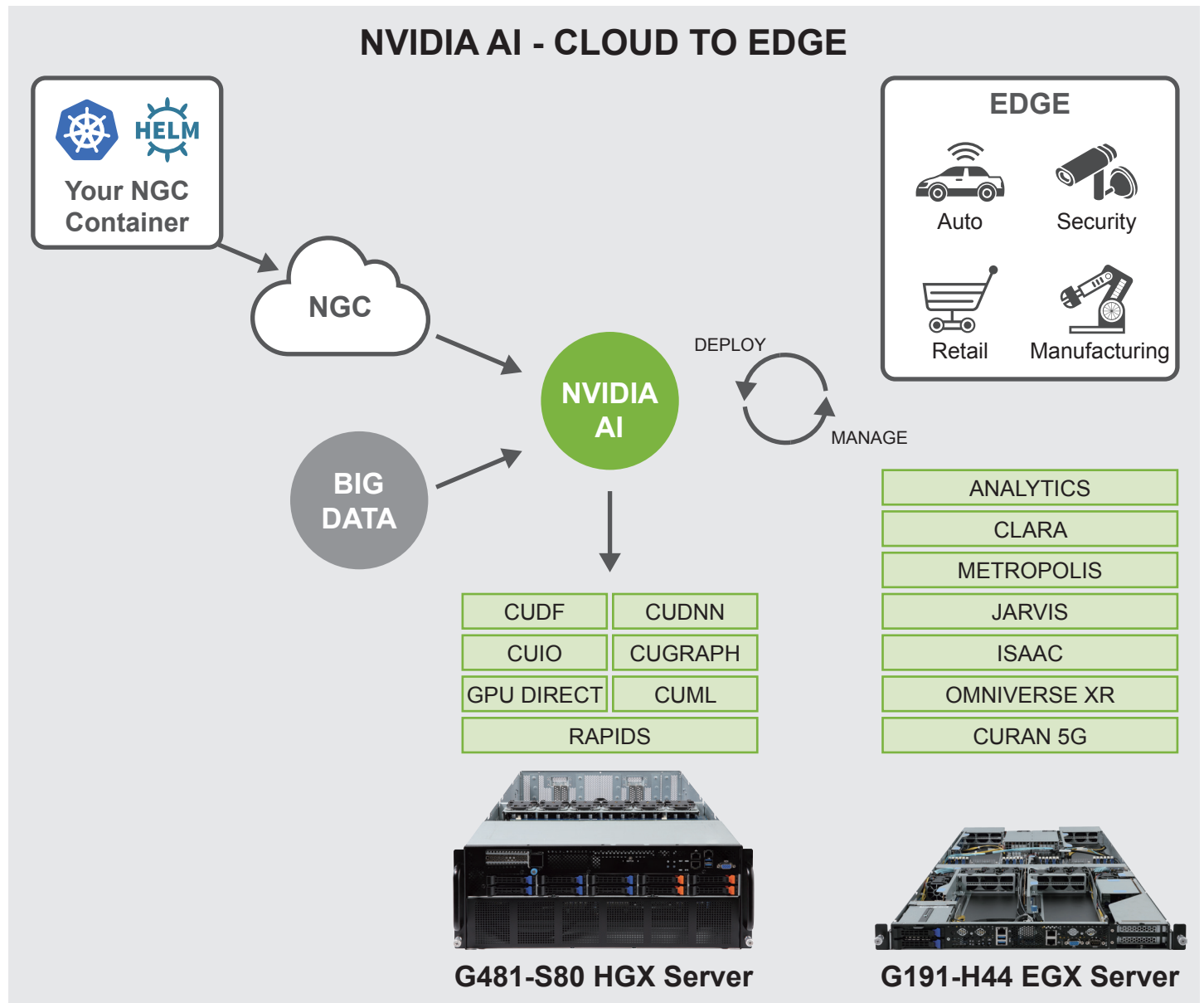
However, because these systems are often distributed across many remote locations, tasks such as operating system installation, Kubernetes deployment, device provisioning and updating, as well as standardizing runtimes, monitoring features, and drivers etc. makes deploying and managing these systems a big challenge.

GIGABYTE and NVIDIA have therefore developed the G191-H44 EGX Edge Computing Platform, which not only features a powerful yet space and energy efficient hardware solution ideal for GPU-accelerated workloads at the edge, but also NVIDIA Edge Stack, an integrated software stack that simplifies and streamlines the entire process of edge AI server deployment and management.

SOLUTION

GIGABYTE's EGX Platform integrates our G191-H44 1U dual 2nd Generation Intel Xeon Scalable Family processor server, 4 x NVIDIA Tesla® T4 accelerators and Mellanox Connect-X 5 smart NICs together with NVIDIA Edge Stack, a hardened software stack built for AI inferencing workloads using containers. NVIDIA Edge Stack provides an enterprise grade container orchestration environment that runs on top of GIGABYTE's G191-H44 server to support all kinds of AI models, and can connect to cloud IoT services such as RedHat OpenShift, AWS GreenGrass, or Azure IoT Edge.

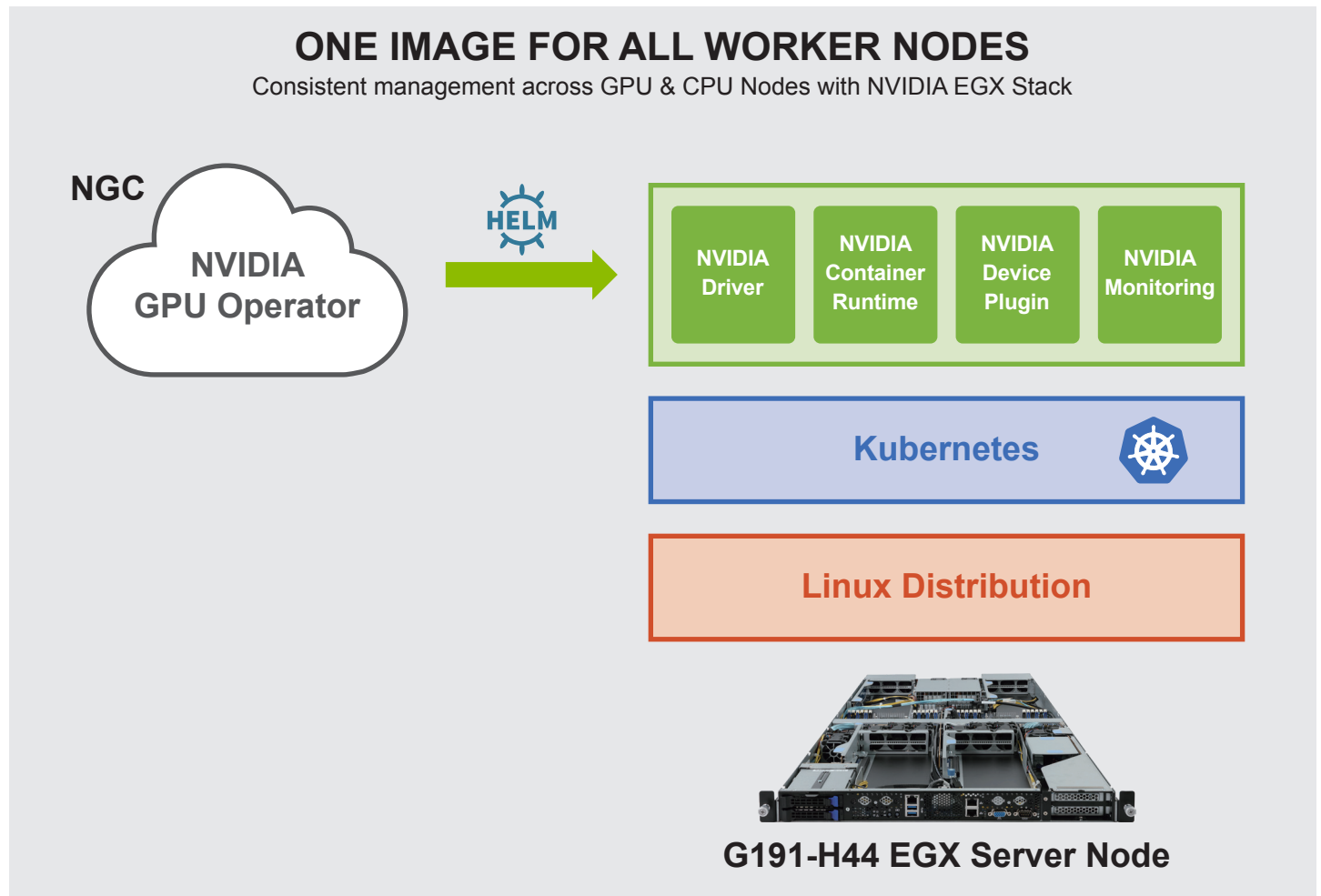
NVIDIA Edge Stack includes NVIDIA drivers, a CUDA Kubernetes plug-in, a CUDA Docker container runtime, CUDA-X libraries and containerized AI frameworks and applications, including TensorRT, TensorRT Inference Server and DeepStream, and can be easily integrated together with leading third party Kubernetes orchestration platforms such as Red Hat OpenShift and be used to deploy a variety of AI-ready containers from NGC (NVIDIA GPU Cloud).



G191-H44 EGX Server is Easy to Deploy with GPU Operator

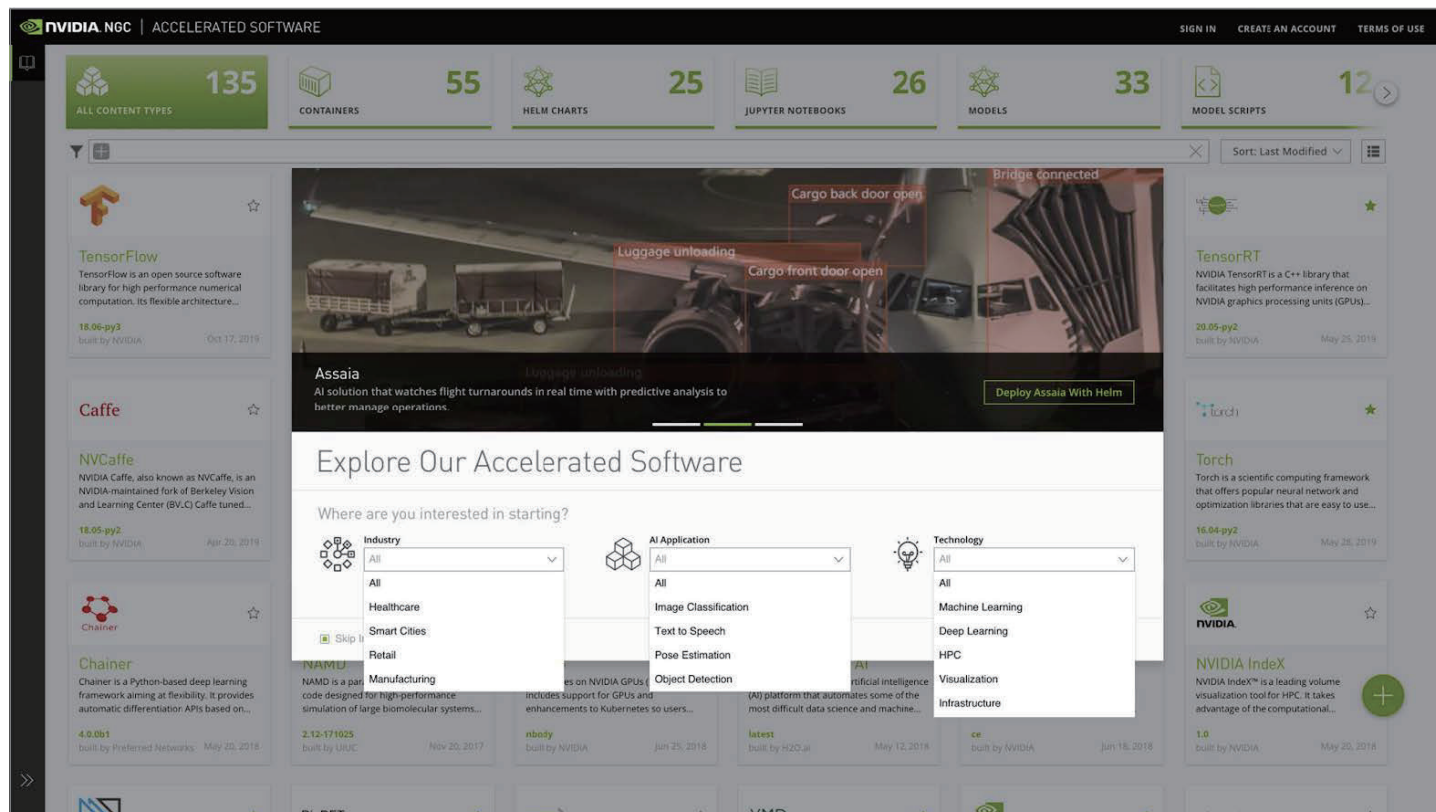
GIGABYTE's EGX server simplifies the deployment and management of GPU-accelerated AI micro-clouds for all industries. The EGX stack is optimized to run container-based applications on cloud-native Kubernetes infrastructure with rapid installation via a Helm chart. This installation gets simpler with NVIDIA GPU Operator, which standardizes and automates the deployment of all the necessary components for provisioning GPU-enabled Kubernetes clusters.

NVIDIA GPU Operator uses containers to automate the setup of a GPU-powered node. The GPU driver, runtime, Kubernetes plug-in, and monitoring plug-in can now be managed by EGX GPU Operator. This infrastructure also manages the complex upgrading of the driver based on cluster policy.



G191-H44 is an NGC Ready Platform

GIGABYTE's G191-H44 EGX Platform is built to be used with NVIDIA NGC (NVIDIA GPU Cloud), an online, free to use platform (and with optional enterprise grade support) which features GPU accelerated containers, a container registry and software for deploying and running these containers. NGC is the go-to place for GPU-optimized and curated containers, deep learning models, and scripts that enable users to get the maximum value from their GPU infrastructure. New NGC containers are published monthly and offer the latest features and superior performance on NVIDIA GPUs, can be used with multi-GPU & multi-node systems for scale-up & scale-out environments, and feature support for Docker & Singularity runtimes as well as VMware vSphere.



GIGABYTE's G191-H44 has been validated as a **NGC Ready for Edge server**, so that NGC containers, models or scripts can be quickly and easily deployed to run inferencing or other GPU-accelerated workloads. NGC has now also expanded to include support for Helm charts, which are essential for fast deployment on any enterprise Kubernetes platform. NGC will include a Helm chart for GPU Operator deployment.

NGC also features NVIDIA Application frameworks that can be easily run on top of NVIDIA EGX-Stack. An example is NVIDIA Metropolis IoT, which can be used for image processing in real time from streams of CCTV cameras for intelligent video analytics and applications that can be used to enable smart cities.

GIGABYTE G191-H44 EGX Platform

NGC



Third-Party ISVs



NVIDIA APPLICATION FRAMEWORKS

METROPOLIS

AI City



CLARA

AI Healthcare



METROPOLIS

AI Retail



ISAAC

AI Manufacturing



NVIDIA EDGE STACK

Kubernetes

Containers

CUDA-X

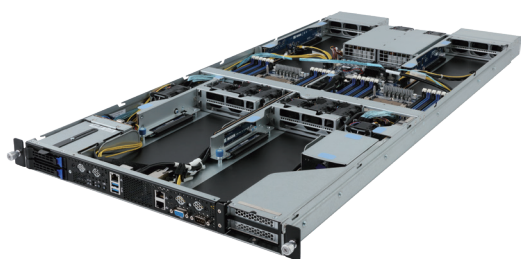
IoT Runtime

G191-H44 EGX EDGE COMPUTING PLATFORM

GPU | AI | Storage | Networking | Security

G191-H44 EGX Server Specifications:

- Dual 2nd Generation Intel® Xeon® Scalable Processors
- 4 x NVIDIA Tesla® T4 PCIe GPGPU cards
- 6-Channel RDIMM/LRDIMM DDR4, 24 x DIMMs
- Supports Intel® Optane™ DC Persistent Memory
- Dual 1Gb/s LAN ports (Intel® I350-AM2)
- 1 x dedicated management port
- 2 x 2.5" hot-swappable + 2 x 2.5" internal fixed HDD/SSD bays
- 2 x low profile PCIe Gen3 expansion slots
- 2 x 80 PLUS Platinum 2000W redundant PSU
- Optimized performance with Mellanox Infiniband and Ethernet 100G products



www.gigabyte.com



[gigabyteserver](https://www.facebook.com/gigabyteserver)



[@GIGABYTESERVER](https://twitter.com/GIGABYTESERVER)



[GIGABYTE](https://www.linkedin.com/company/gigabyte)



GIGABYTE TECHNOLOGY CO., LTD.

- * All intellectual property rights, including without limitation to copyright and trademark of this work and its derivative works are the property of, or are licensed to, GIGA-BYTE TECHNOLOGY CO., LTD. Any unauthorized use is strictly prohibited.
- * The entire materials provided herein are for reference only. GIGABYTE reserves the right to modify or revise the content at anytime without prior notice.
- * All other brands, logos and names are property of their respective owners.