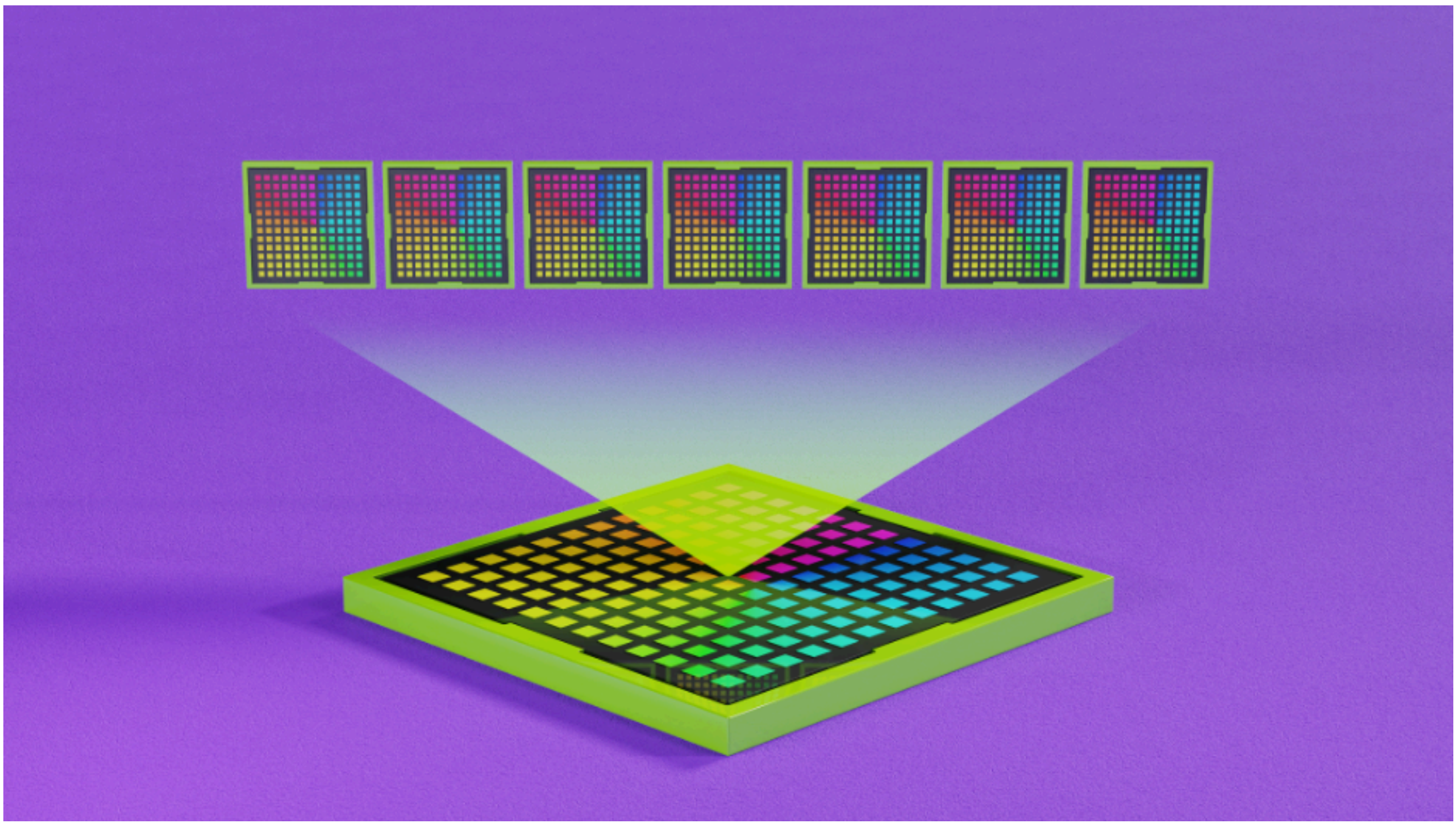
Data Center / Cloud

English

Running Multiple Applications on the Same Edge Devices

Jul 18, 2022

By Troy Estes



Smart spaces are one of the most prolific edge AI use cases. From smart retail stores to autonomous factories, organizations are quick to see the value in this innovative technology. However, building and scaling a smart space requires many different pieces of technology, including multiple applications. Operating multiple applications at edge locations can be tricky.

To do this, organizations may add new hardware to a location so that each application has dedicated compute resources, but the costs associated with purchasing and installing new hardware with every new application can be substantial. Many organizations deploy multiple applications on the same device.

While that is a solution for scale, it can present different challenges.

Many organizations rely on the performance of GPUs to power applications at the edge. Even with high-performance GPU-accelerated systems, having two or more AI applications running concurrently on the same device using time slicing inevitably leads to higher latency with minimal hardware optimization.

With multiple applications running on the same device, the device time-slices the applications in a queue so that applications are run sequentially as opposed to concurrently. There is always a delay in results while the device switches from processing data for one application to another. The amount of delay varies per deployment, but it could be as much as 8ms. That could present serious concerns for applications powering high-speed operations, such as a manufacturing production line.

Because applications are running sequentially, the GPU is only ever used as much as each individual application needs while it is running. For instance, if there are three applications operating sequentially on a GPU and each application requires 60% of the GPU's resources, then at any given time less than 60% of the GPU is used. The GPU utilization would be 0% during each context switch.

workloads continue operating uninterrupted since instances and workloads run in parallel while remaining separate and isolated.

MIG works equally well with containers or virtual machines (VMs). When using VMs, it is easy to virtualize GPUs using NVIDIA vGPU, which can be configured to employ either time slicing or MIG.

MIG for edge AI

When deploying edge AI, optimizing for cost, power, and space are all important considerations, especially if you want to replicate to thousands of edge nodes. By allowing organizations to run multiple applications on the same GPU, MIG eliminates the need for installing a dedicated GPU for each workload, significantly reducing resource requirements.

More than resource optimization, MIG helps ensure predictable application performance. Without MIG, different jobs running on the same GPU, such as different AI inference requests, compete for the same resources such as memory and bandwidth. Due to the competition for resources inherent in time slicing, the performance of one application can be affected by activity in another. For edge AI environments, unpredictable performance can have serious consequences.

For example, a computer vision application monitoring a production line to detect product defects has to be able to react instantaneously to its dynamic environment. It must be able to inspect products quickly, and also to communicate with other machinery to stop the production line in the case of a defective product. For safety and efficiency, organizations must know that the AI applications powering their production lines are running correctly and predictably all of the time.

Jobs running simultaneously with different resources result in predictable performance with quality of service and maximum GPU utilization, making MIG an essential addition to every edge deployment.

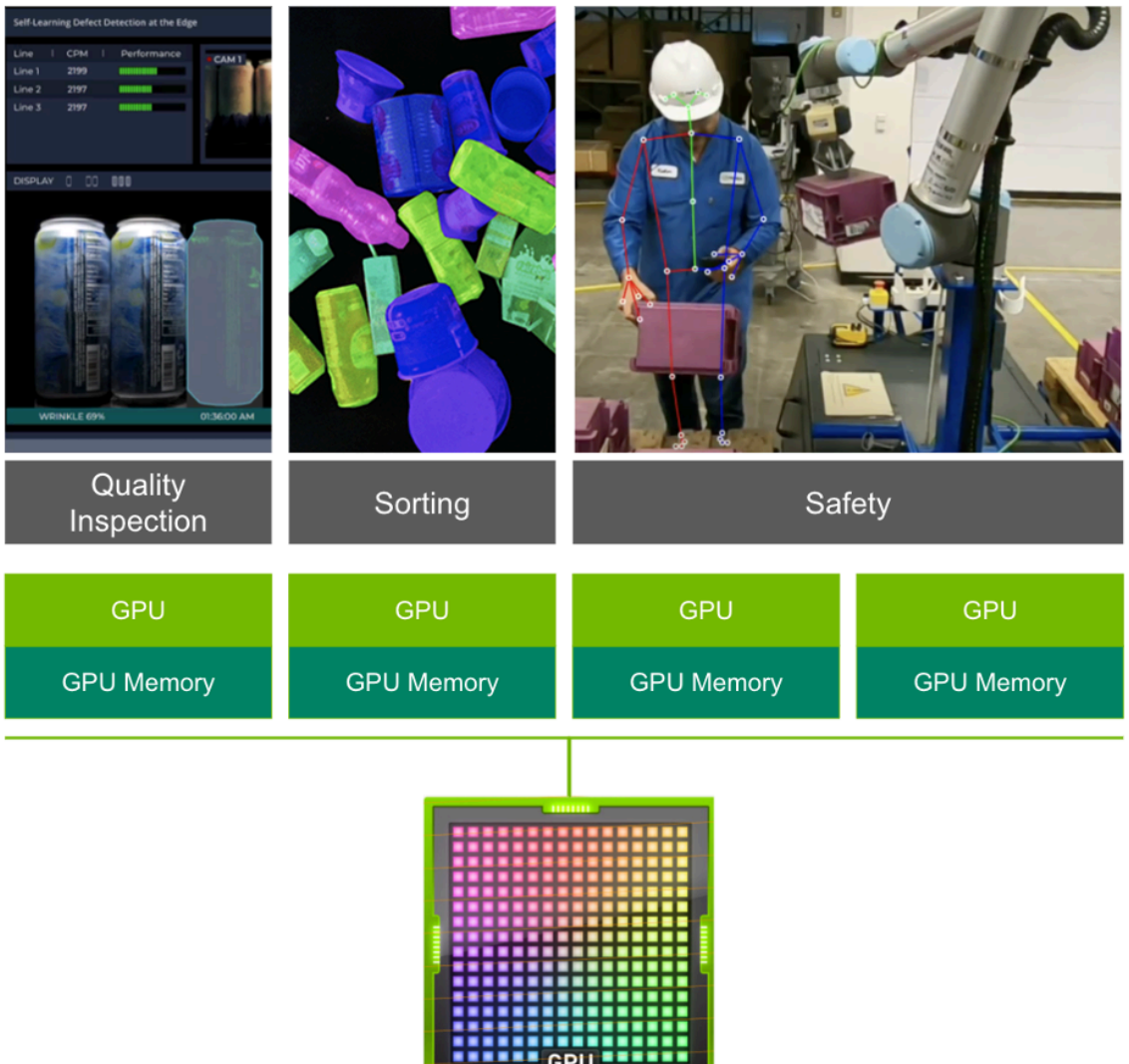


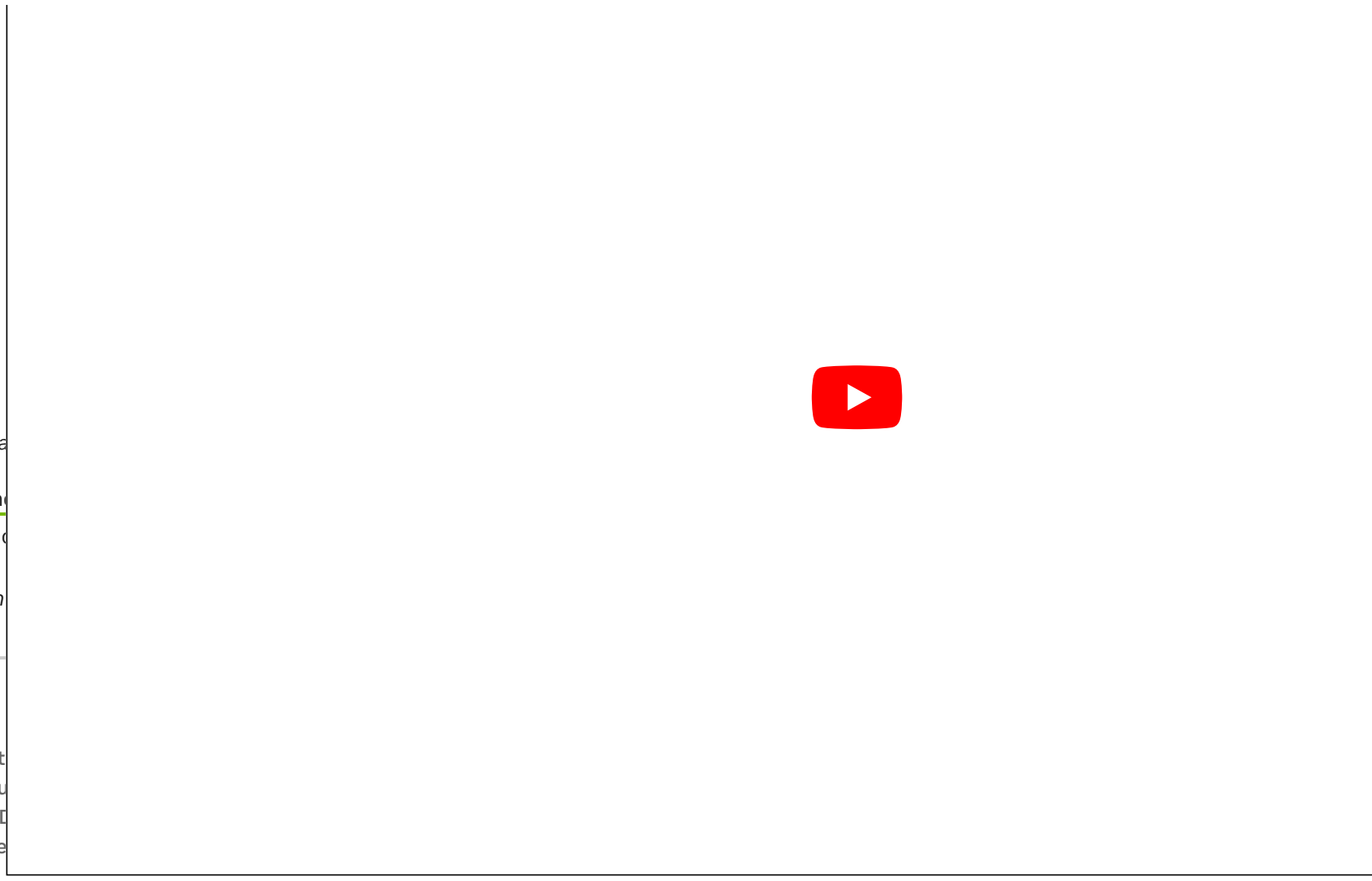
Figure 1. Each MIG instance can handle an independent workload, optimizing environments where multiple use cases need to operate simultaneously

MIG on NVIDIA Fleet Command

Fleet Command is a cloud service that centrally connects systems at edge locations to securely deploy, manage, and scale AI applications from one dashboard. Purpose-built for edge AI, Fleet Command is the best way to orchestrate AI across hundreds or even thousands of devices.

From the Fleet Command cloud platform, administrators have complete control over MIG for edge AI deployments with minimal configuration needed. Using MIG on Fleet Command enables you to make resource utilization decisions across hundreds or even thousands of devices with just a few clicks. You can easily add new MIG partitions, scale down existing partitions, and create custom deployments all from one dashboard.

The combination of MIG and Fleet Command provides organizations with all the functionality needed to have full control over edge AI deployments, leading to better used and more effective workloads. For more information about the entire workflow for using MIG on Fleet Command, see the following video.



Video 1. How to Run Multiple Applications on the Sa

Try Fleet Command yourself with NVIDIA Laun
flow for deploying and managing applications

Sign up for Edge AI News to stay up to date with

Related resources

- **DLI course:** Building Video AI Applications at t
- **GTC session:** Edge Computing 101: An Introdu
- **GTC session:** Getting AI to the Edge with NVID
- **GTC session:** An End-To-End Solution for Ente
- **SDK:** NVIDIA Fleet Command
- **Webinar:** Edge Computing 101: An Introduction to the Edge

Tags

Data Center / Cloud | Edge Computing | Retail / Consumer Packaged Goods | Smart Cities / Spaces | Fleet Command | Tutorial | Featured

About the Authors

About Troy Estes

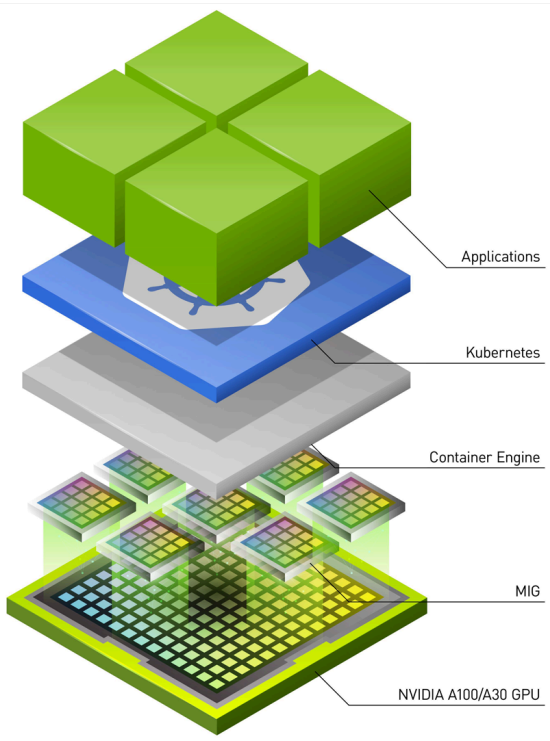
Troy Estes is a product marketing manager in edge and enterprise computing solutions at NVIDIA. Prior to joining the Edge & Enterprise business unit, Troy has worked on marketing campaigns and supporting product GTM in the Autonomous Vehicles business unit and the NVIDIA GRID product group.

View all posts by Troy Estes ›

Comments

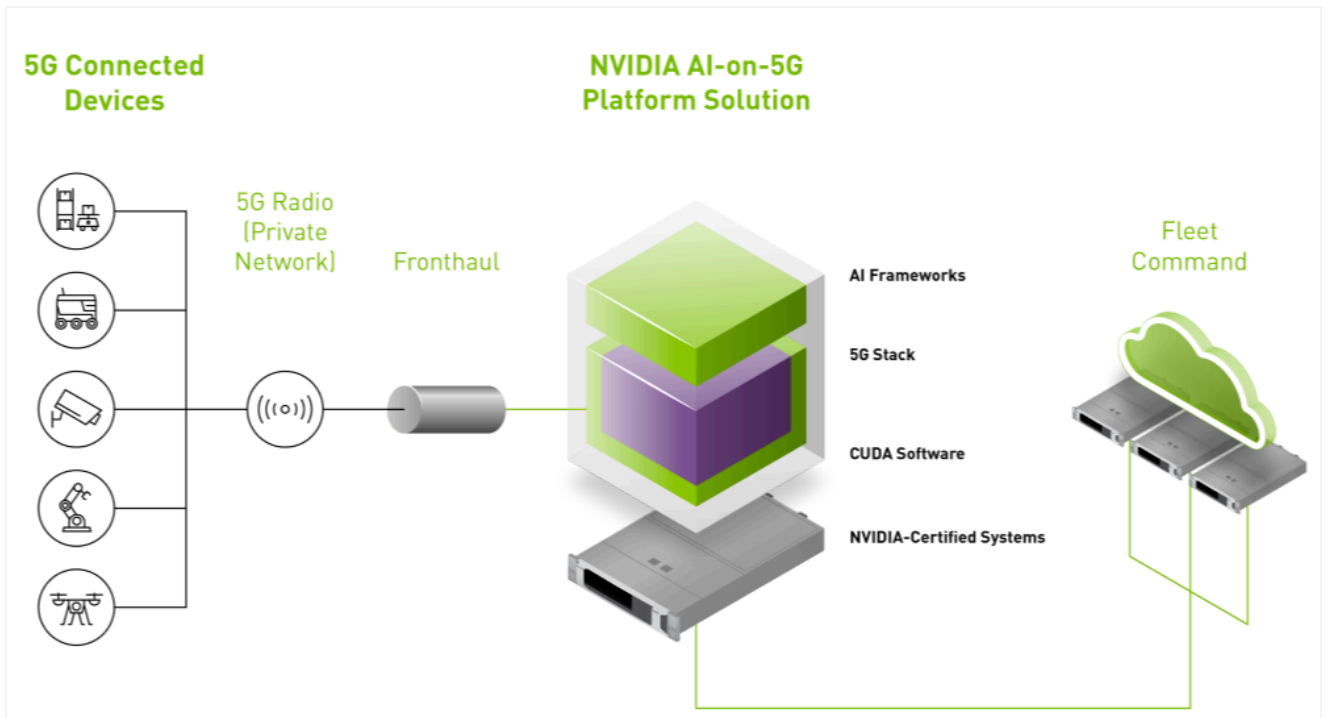
Start the discussion at forums.developer.nvidia.com

Related posts

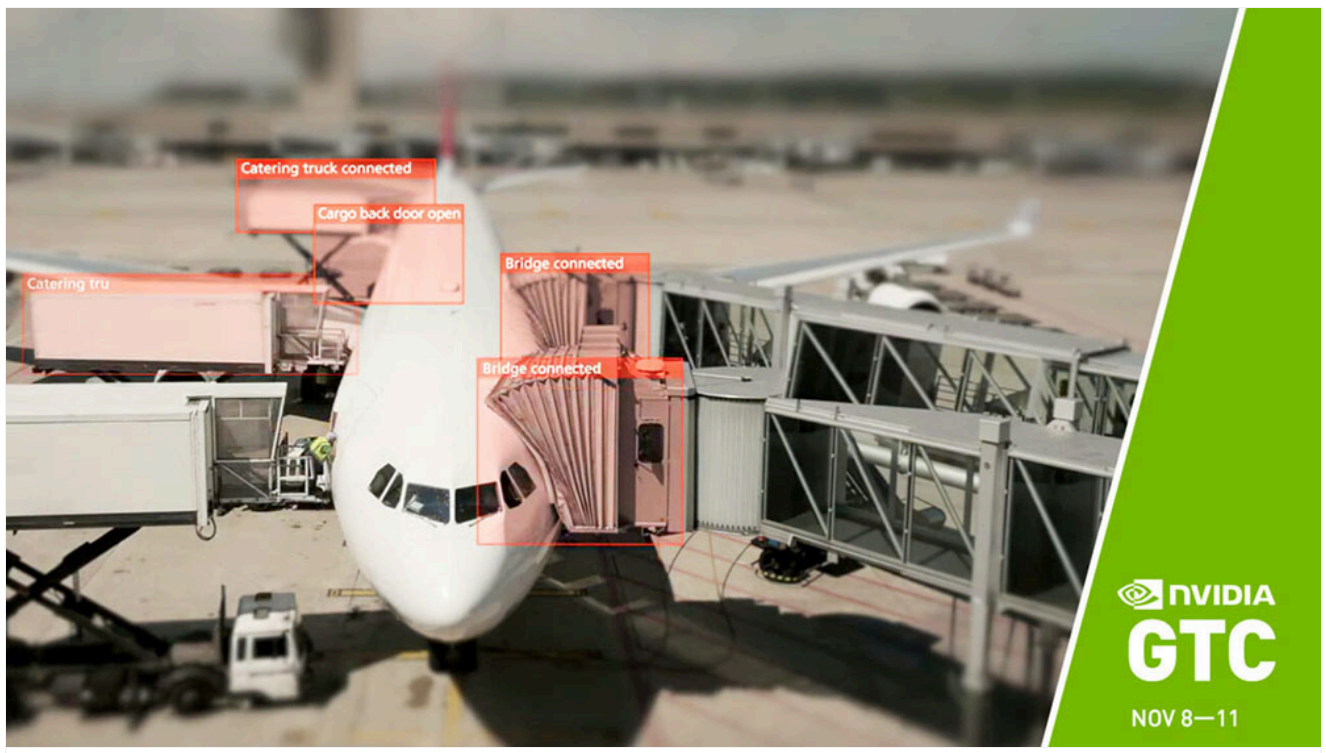


Improving GPU Utilization in Kubernetes

The Need for Speed: Edge AI with NVIDIA GPUs and SmartNICs, Part 1



NVIDIA AI-on-5G for Enterprise: A Converged Platform for AI and 5G at the Edge



NVIDIA GTC: Taking It to the Edge

Deploying and Accelerating AI at the Edge with the NVIDIA EGX Platform