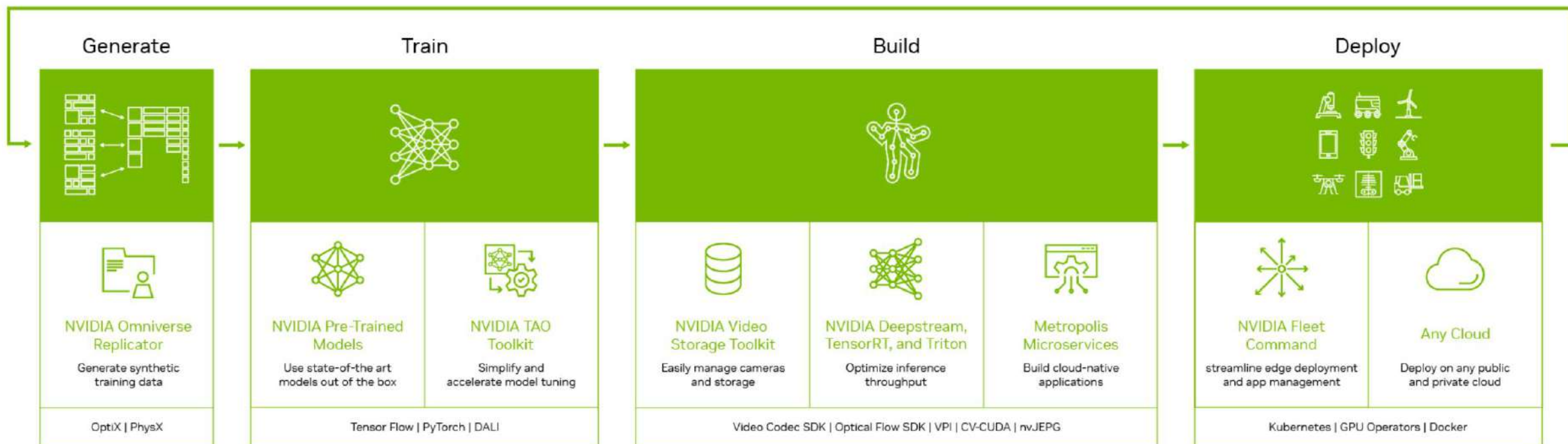# NVIDIA Metropolis

NVIDIA Metropolis features GPU-accelerated SDKs and developer tools that help developers optimally build, deploy, and scale AI-enabled video analytics and IoT applications–from the edge to the cloud.

End-to-End Vision AI Development

Fast-Track Data Generation, AI Model Creation, App Development, Inference, and Scalability

Generate	Train		Build			Deploy	
NVIDIA Omniverse Replicator	NVIDIA Pre-Trained Models	NVIDIA TAO Toolkit	NVIDIA Video Storage Toolkit	NVIDIA Deepstream, TensorRT, and Triton	Metropolis Microservices	NVIDIA Fleet Command	Any Cloud
Generate synthetic training data	Use state-of-the art models out of the box	Simplify and accelerate model tuning	Easily manage cameras and storage	Optimize inference throughput	Build cloud-native applications	streamline edge deployment and app management	Deploy on any public and private cloud
OptiX | PhysX	Tensor Flow | PyTorch | DALI		Video Codec SDK | Optical Flow SDK | VPI | CV-CUDA | nvJEPG			Kubernetes | GPU Operators | Docker	

Explore the benefits

Faster builds

Use and customize high-performance, pretrained models, or your own models, to streamline deploying AI applications across a range of industries. Jump-start application development by building off modular microservices and reference applications.

Lower cost

Powerful SDKs including NVIDIA TensorRT™, DeepStream, and TAO Toolkit reduce overall solution cost by maximizing inference throughput and optimizing hardware usage on NVIDIA platforms and infrastructure.

Flexible deployments

Manage and scale AI deployments securely with Fleet Command™ and deploy with flexibility using cloud-native Metropolis Microservices and containerized apps with options for on-premise, cloud, or hybrid deployments.

Powerful tools for
AI-enabled video analytics

The Metropolis suite of SDKs provides a variety of starting points to accelerate and optimize any aspect of AI application development and deployment.

NVIDIA Omniverse™ Replicator

Generate physically accurate 3D synthetic data at scale, or build your own synthetic data tools and frameworks. Bootstrap perception AI model training and achieve accurate Sim2Real performance without having to manually curate and label real-world data.

100+ Model Architectures

Image Classification	Object Detection	Segmentation
EffcientNet ResNet	RetinaNet YOLOV3/V4	UNET MaskRCNN

25+ Task-Based Models

People Detection	Vehicle Classification	2D/3D Pose Estimation

ALPR	ASR	TTS

Pretrained models

Eliminate the time-consuming process of building models from scratch. Choose from over 100+ permutations of highly accurate models and generic neural network architectures or start with our task-based models to recognize human actions and poses, detect people in crowded spaces, classify vehicles and license plates, and much more.

TAO Toolkit

The Train, Adapt and Optimize (TAO) Toolkit is a low-code AI model development solution that lets you use the power of transfer learning to fine-tune NVIDIA pretrained models with your own data and optimize for inference–without AI expertise or a large training dataset.
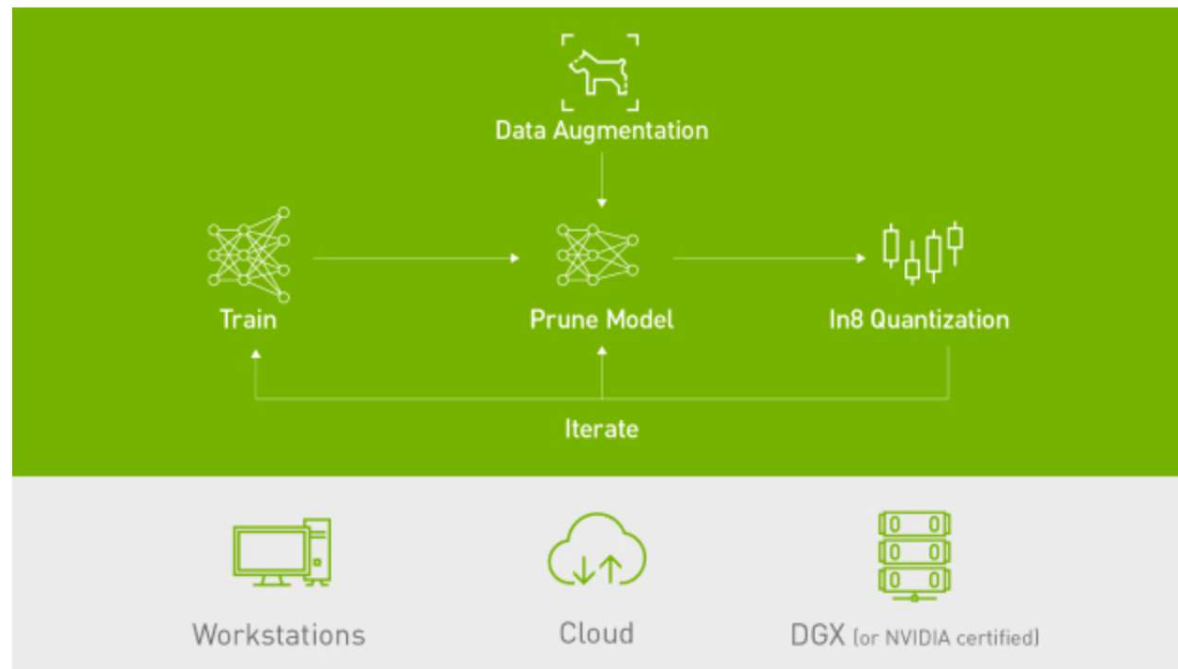
Many AI model frameworks

Create your AI models and applications on these popular NVIDIA-supported AI frameworks. Integrate any existing AI model into the Metropolis workflow and easily customize existing models in TensorFlow, PyTorch, and more by easily converting to TAO.

- > TensorFlow
- > Onnx
- > PyTorch
- > Mxnet
- > PaddlePaddle

TensorRT

This SDK for high-performance deep learning inference includes an inference optimizer and runtime that delivers low latency and high throughput, both on edge devices and within the cloud. TensorRT is supported on all popular frameworks, including TensorFlow and PyTorch. Powering NVIDIA solutions such as JetPack™ and DeepStream, TensorRT is a gateway to accelerated inferencing.
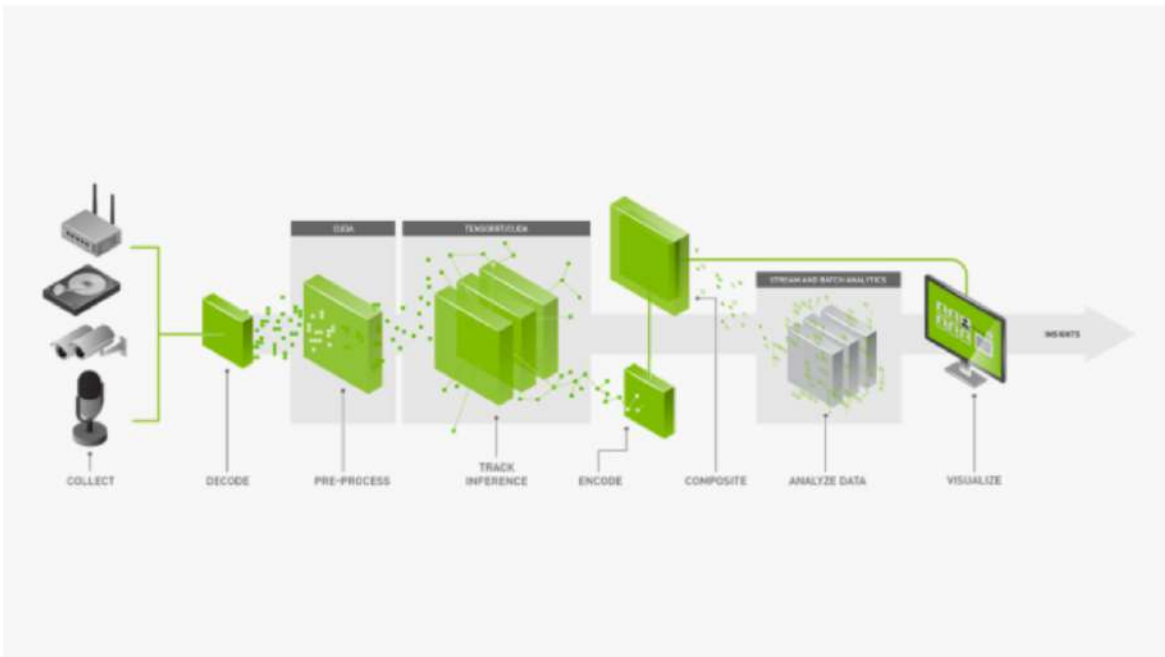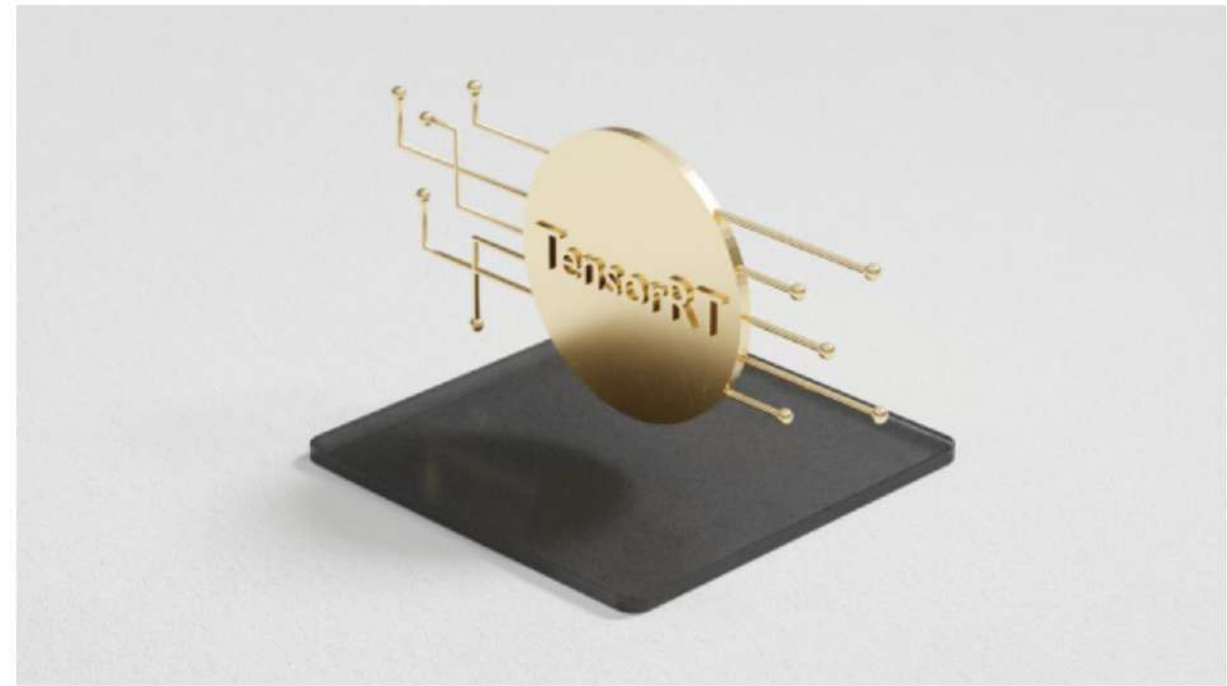

DeepStream SDK

NVIDIA DeepStream SDK is a complete streaming analytics toolkit based on GStreamer for AI-based multi-sensor processing, video, audio, and image understanding. It's ideal for vision AI developers, software partners, startups, and OEMs building IVA apps and services.

Triton Inference Server

The NVIDIA Triton™ open-source, multi-framework inference serves software to deploy, run, and scale AI models in production on both GPUs and CPUs. It supports all major frameworks, including TensorFlow and Pytorch, and maximizes inference throughput on any platform.
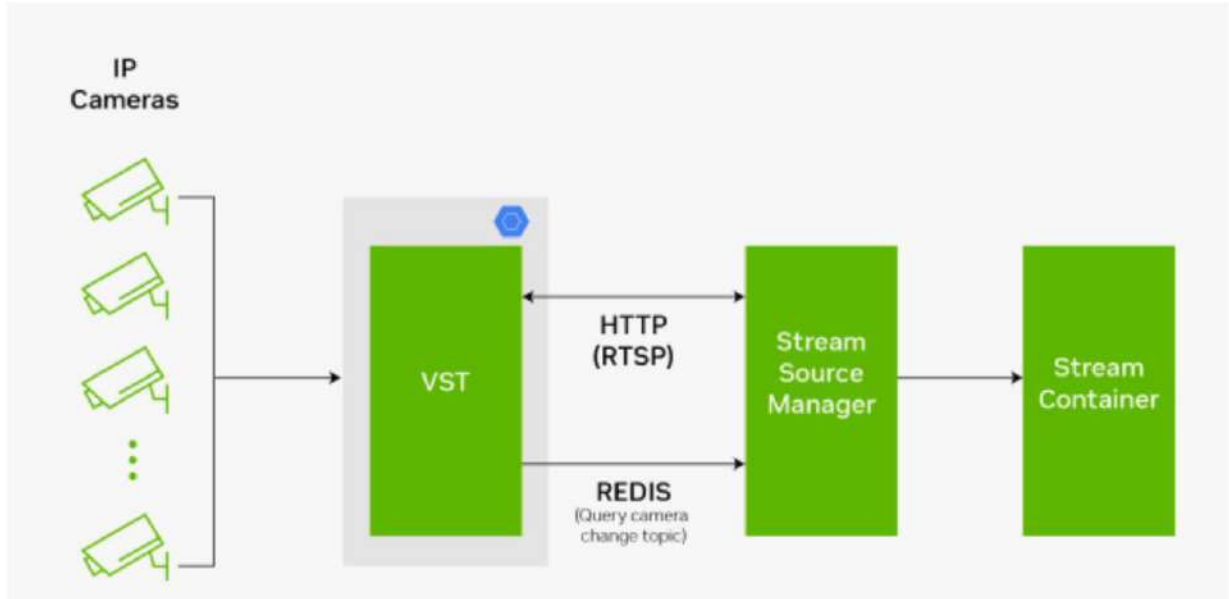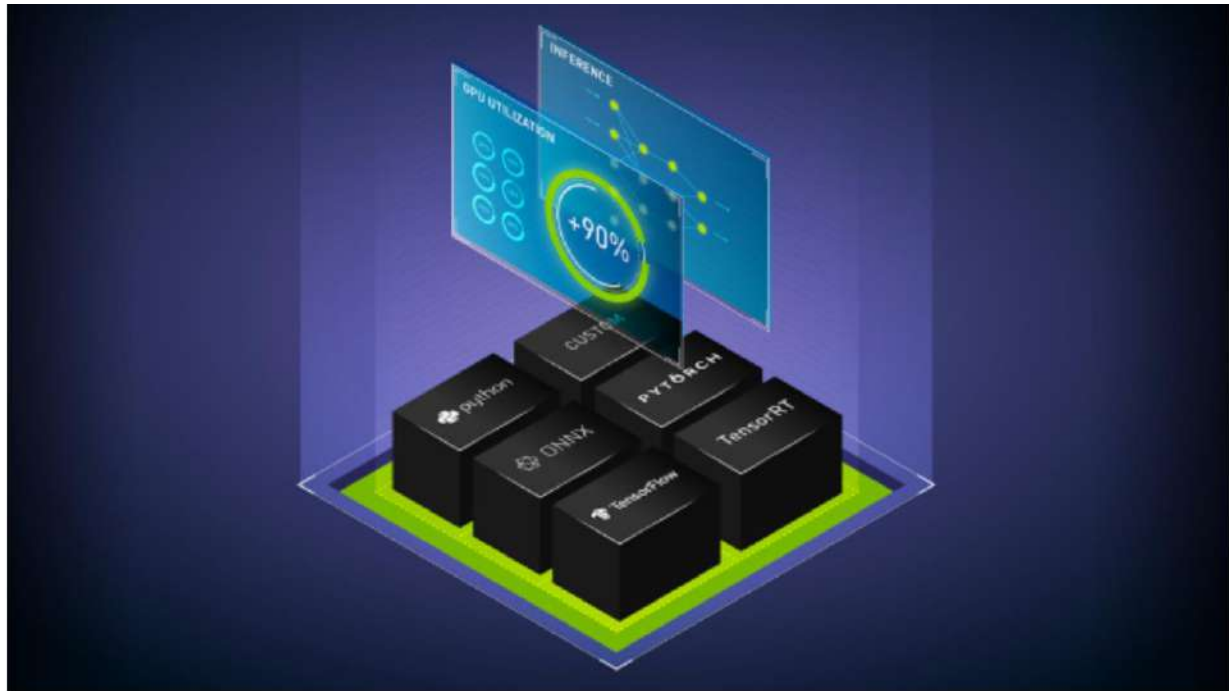
Video Storage Toolkit (VST)

Easily manage and store footage for large volumes of video cameras with hardware-accelerated video decoding, streaming, and storage. Get started quickly with the included web-based user interface and take advantage of VST flexibility through intuitive REST APIs. It's available for NVIDIA Jetson Xavier™ and Orin™ devices.

Metropolis Microservices

This suite of cloud-native microservices and reference applications fast-tracks development and deployment of vision AI applications. Unlock business insights for a wide range of spaces, ranging from roadways to airports to retail stores, in significantly shortened development cycles.


Metropolis Microservices
Customizable vision AI apps
and building blocks

CUDA-X libraries

Take advantage of low-level libraries and primitives for computer vision and more that can help with pre-processing and model performance. NVIDIA® CUDA-X™, built on top of NVIDIA CUDA®, is a collection of libraries, tools, and technologies that deliver dramatically higher performance in compute-intensive algorithms spanning complex math, deep learning, and image processing.

> Video Codec SDK   > cuTensor

> DALI   > nvJPEG

> cuDNN   > VPI

> Optical Flow SDK   > CV-CUDA

> Performance Primitives
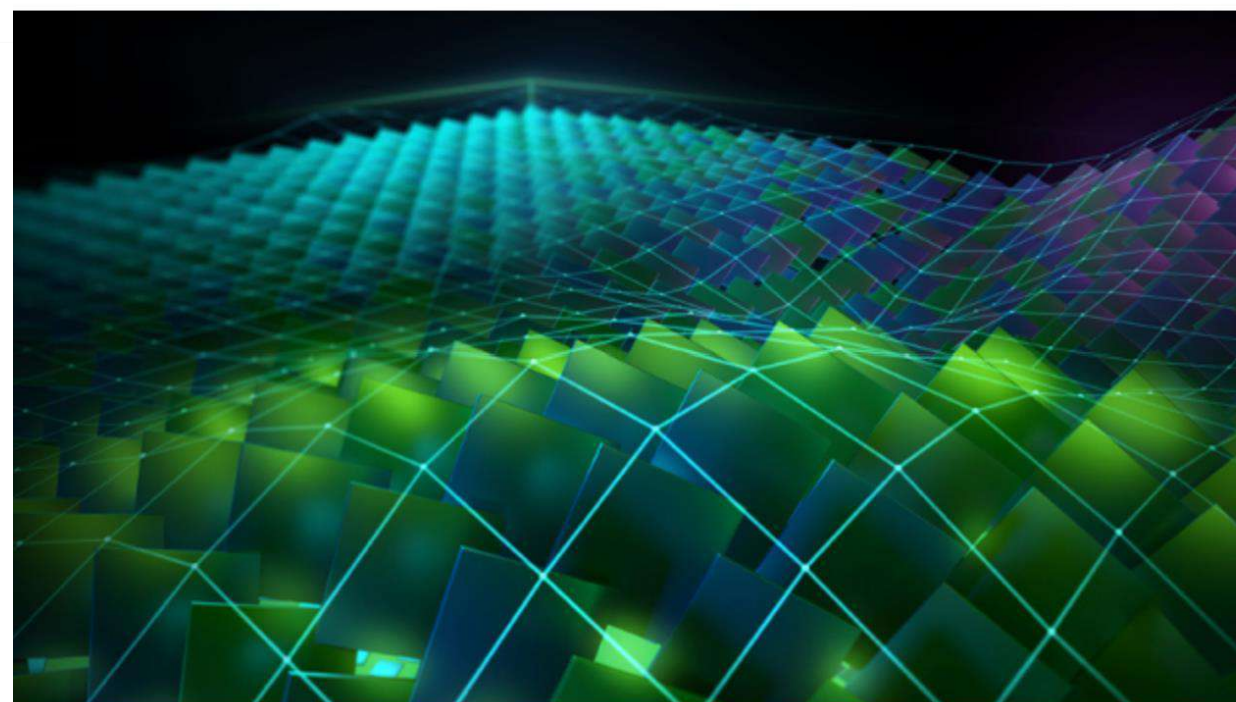
Fleet Command

Streamline the provisioning and deployment of systems and AI applications at the edge with NVIDIA Fleet Command. A managed platform for container orchestration, it simplifies the management of distributed computing environments with the scale and resiliency of the cloud, turning every site into a secure, intelligent location.

Cloud containers

Combine NVIDIA SDKs to create containerized applications easily with Docker, Kubernetes, and GPU Operators to deploy cloud-native solutions on Jetson, x86, and dGPU.