Vision AI at the Edge: From Zero to Deployment Using Low-Code Development NIVIDA Tools, Part 2

Alvin Clark

Technical Product Marketing

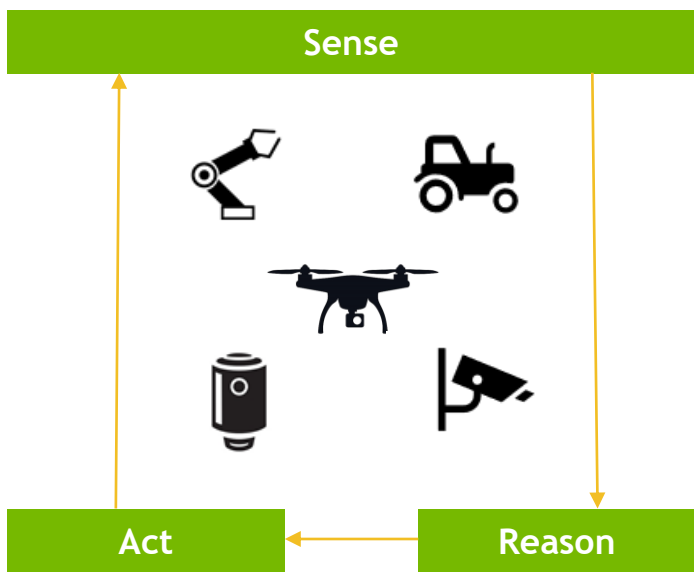NVIDIA

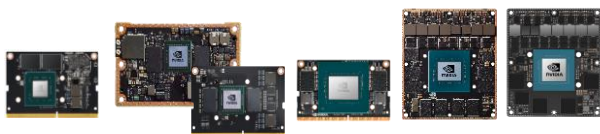# NVIDIA Jetson
Software-Defined AI Platform

AI AT THE EDGE
Sensor Fusion & Compute Performance

Sense

Act ← **Reason**

JETSON COMPUTER

SOFTWARE DEFINED
SDK, Design Tools, Libs, GEMs

Pose est | Speech rec | Gesture rec
Depth est | Path planning | Obj detect

Accelerated Computing | Multimedia
Artificial Intelligence | Computer Vision

JetPack SDK · CUDA · TensorRT · TensorFlow · ONNX · ROS

ECOSYSTEM
Expertise, Time to Market

SENSORS | AI FRAMEWORKS

mxnet
PYTORCH | TensorFlow

ROBOT COMPUTERS | ROBOTICS SOFTWARE SOLUTIONS

aivero | ACCERION | brain | concurrent REAL-TIME
Iris Automation | PX4 Autopilot | ROS | MathWorks

REFERENCE PLATFORMS | DEVICE MGMT & CLOUD INTEGRATION

AWS Greengrass | AWS Robomaker | balena
docker | Microsoft Azure IoT Platform | MPI

NVIDIA.

Jetson Software
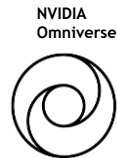## Accelerates AI Applications and Time-to-Market

AI MODEL DEVELOPMENT

Production-ready AI pre-trained models and Toolkits to accelerate development by up to 10X

Pre-trained models

TAO Toolkit

NVIDIA Omniverse
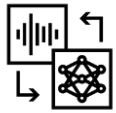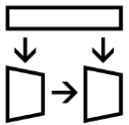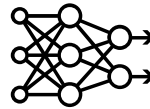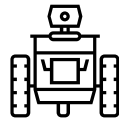
Synthetic Data Generator

APPLICATION FRAMEWORKS

Easy-to-use libraries that support the development of applications

DeepStream

Isaac

Hello

Riva

JETPACK SDK

Comprehensive SDK for building end-to-end accelerated AI pipeline
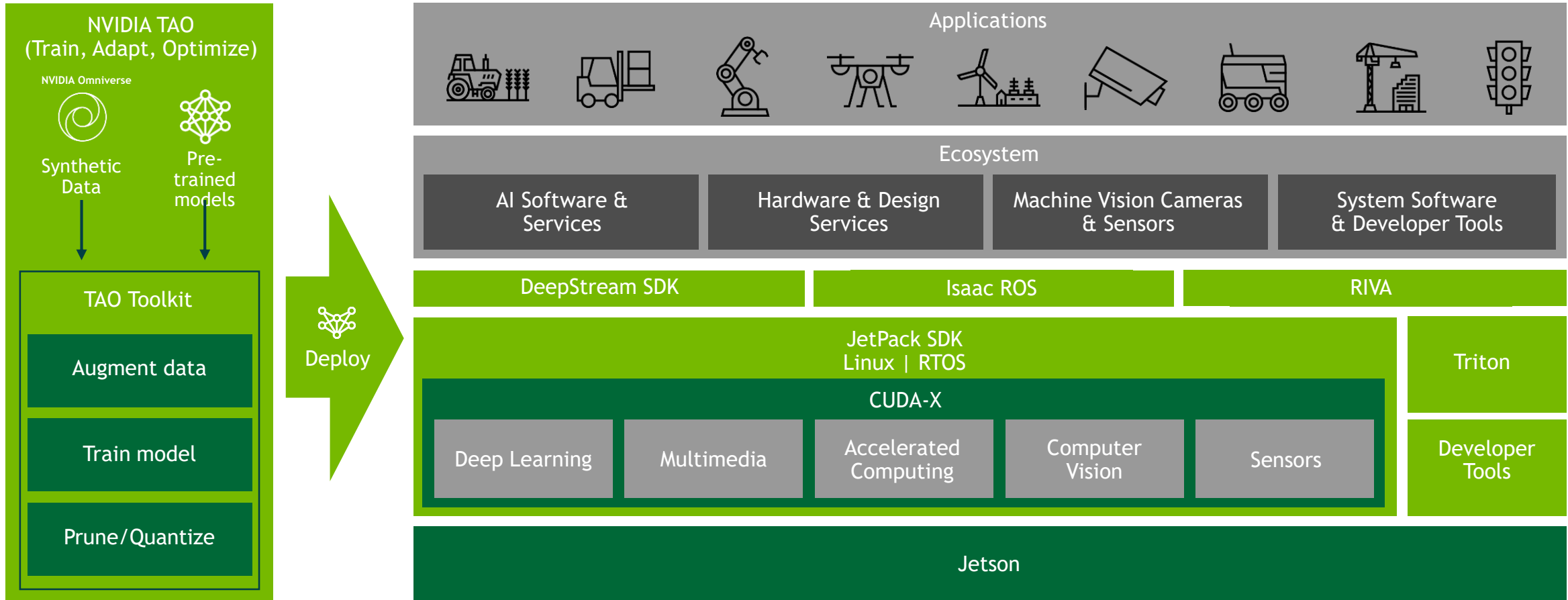
JetPack SDK
Linux | RTOS

CUDA-X

Deep Learning	Multimedia	Accelerated Computing	Computer Vision	Sensors

Jetson Software

NVIDIA TAO
(Train, Adapt, Optimize)

NVIDIA Omniverse

Synthetic Data

Pre-trained models

TAO Toolkit

Augment data

Train model

Prune/Quantize

Deploy

Applications

Ecosystem

AI Software & Services	Hardware & Design Services	Machine Vision Cameras & Sensors	System Software & Developer Tools

DeepStream SDK	Isaac ROS	RIVA

JetPack SDK
Linux | RTOS

CUDA-X

Deep Learning	Multimedia	Accelerated Computing	Computer Vision	Sensors

Triton

Developer Tools

Jetson

CV Is More than AI Inference

| Sensors | Capture and Decode | Pre-processing & Batching | AI Inference | Tracking | Business Rules and Analytics | Composition | Actions |

© 2022 NVIDIA
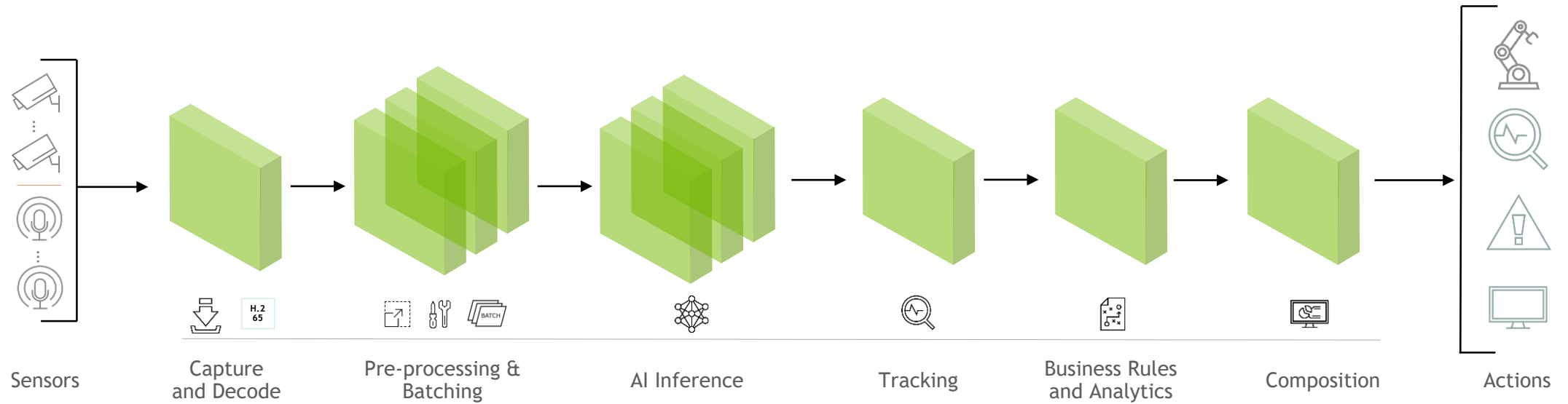
DeepStream SDK
Accelerated and Optimized Applications from Edge to Cloud

Sensors	Capture and Decode	Pre-processing and Batching	AI Inference	Tracking	Business Rules and Analytics	Composition	Actions

Hardware Acceleration

Capture and Decode	Pre-processing and Batching	AI Inference	Tracking	Business Rules and Analytics	Composition
ISP	GPU	GPU	GPU	GPU	GPU
GigE		DLA	CPU	CPU	VIC
NVDEC			VIC		

Edge to Cloud Deployments

Jetson Appliances   EGX Servers   Cloud

© 2022 NVIDIA

6

Pipeline Efficiency with 0 Memory Copies

| Capture | Decode | Batching | Inference | ... | Next Plugin |

CPU to GPU Memory

CUDA HW Buffer Sharing

Returns CUDA HW Buffer for Reuse

Rivermax *(Optional)*

Zero Buffer Copy Path

CPU Memory

Allocates CUDA HW Buffers (GPU Memory)

Uses shared CUDA HW Buffers (GPU Memory)
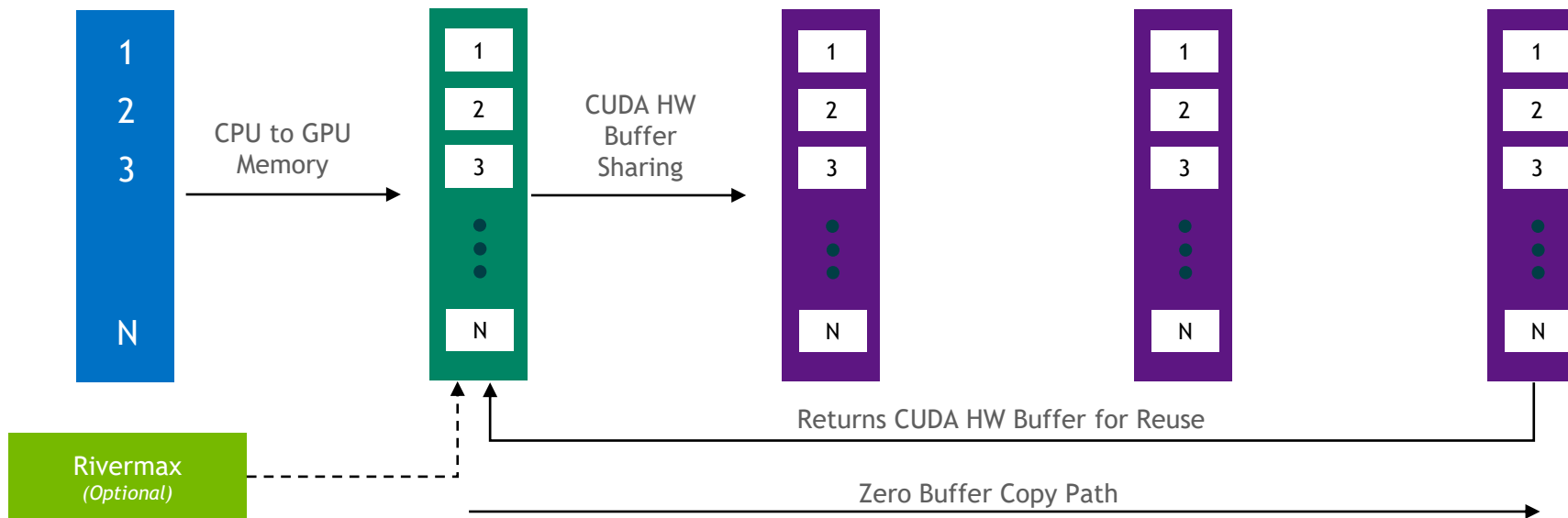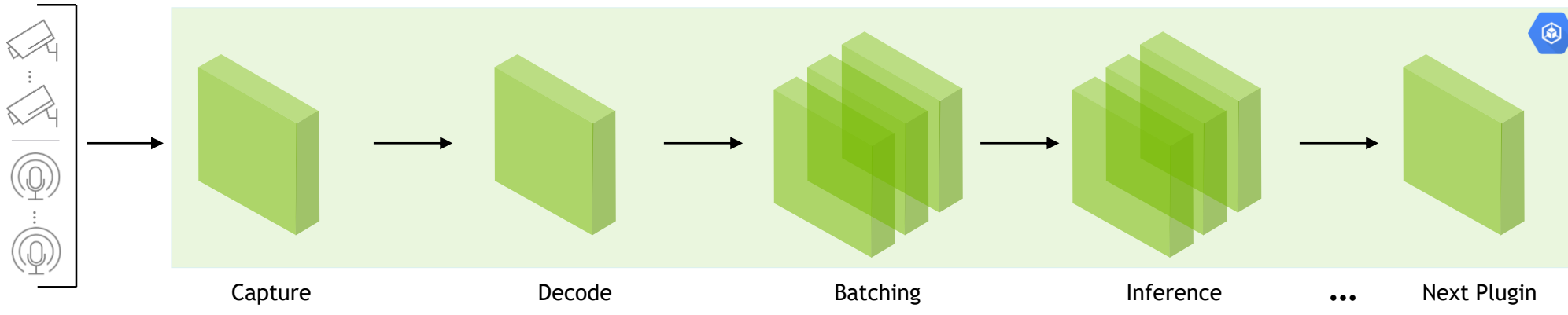
DeepStream Features

Graph Composer
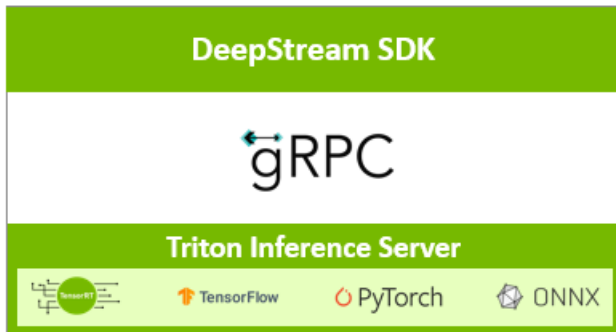Develop production ready apps
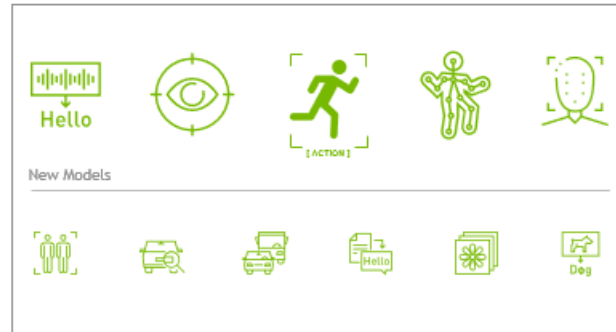
New Models

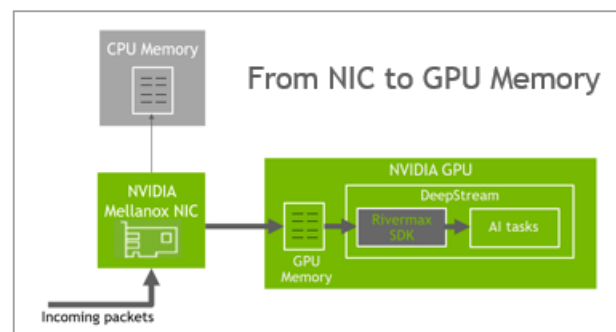**New Models Ready to Use**
Action Recognition, Pose, Gaze, ASR, ...

New Unified Tracker Architecture

Tracker Type	Visual Tracker	State Estimator	Target Mgmt.	Trajectory Mgmt.
NvDCF	●	●	●	●
DeepSORT		●	●	●
IOU		●		●

State-of-the-art Trackers
Bring Your Own Tracker

DeepStream SDK

gRPC

Triton Inference Server

TensorRT   TensorFlow   PyTorch   ONNX

gRPC Support for Triton
Build apps that leverage multiple frameworks

CPU Memory

From NIC to GPU Memory

NVIDIA Mellanox NIC

NVIDIA GPU
DeepStream
Rivermax SDK
AI tasks
GPU Memory

Incoming packets

Rivermax Support
Optimize uncompressed data pipelines

Any Cloud All Your Metadata

Redis   Greengrass   Kafka   Azure IOT   MQTT
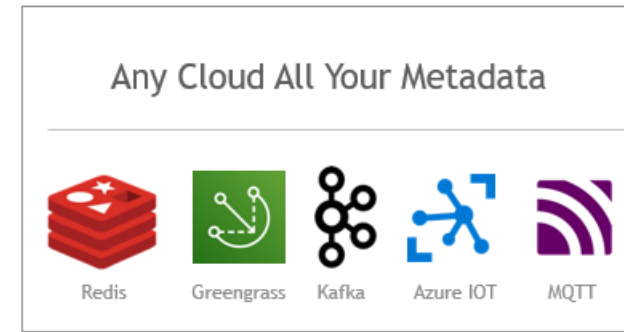
IOT and Cloud Connectivity
Redis, Azure Percept and many others

Developing with DeepStream

Building Blocks

DeepStream Developers

Python Bindings
(source code available now)

Hardware Accelerated
DeepStream Plugins

Plugins to Graph
Composer Extensions

Application Development

Develop in Python

Develop in C/C++

Extensions

Develop in Graph Composer

DeepStream Applications

Application Deployment

Container
Builder

Cloud

EGX Servers

Jetson Appliances

Cloud Native
Deployment

Why Graph Composer?
Benefits of Low-Code Development

Graph Composer for AI Beginners

- Makes NVIDIA Vision AI technology more accessible to developers and organizations

- Makes it easier for beginners to get started with DeepStream on EGX or Jetson

- No prior Gstreamer/DeepStream knowledge required

Graph Composer for Developers

- Allows developers to focus on the application's functionality and optimization

- Provides a platform for quick evaluation and PoC tasks

- Makes it easier to create, integrate, and share custom functionality using extensions

Development Workflow

Graph Composer Workflow Steps

1. Sync extension from NVIDIA Cloud repository

2. Develop custom extension

   - Generate extension from GStreamer plugin (Automatic)

   - Generate extension from codelets (Manual/Advanced)

3. Build and extensions to registry

4. Create and test graph

5. Create deployment container

   - Generate containers for both x86 (dGPU) and Arm64 (Jetson)
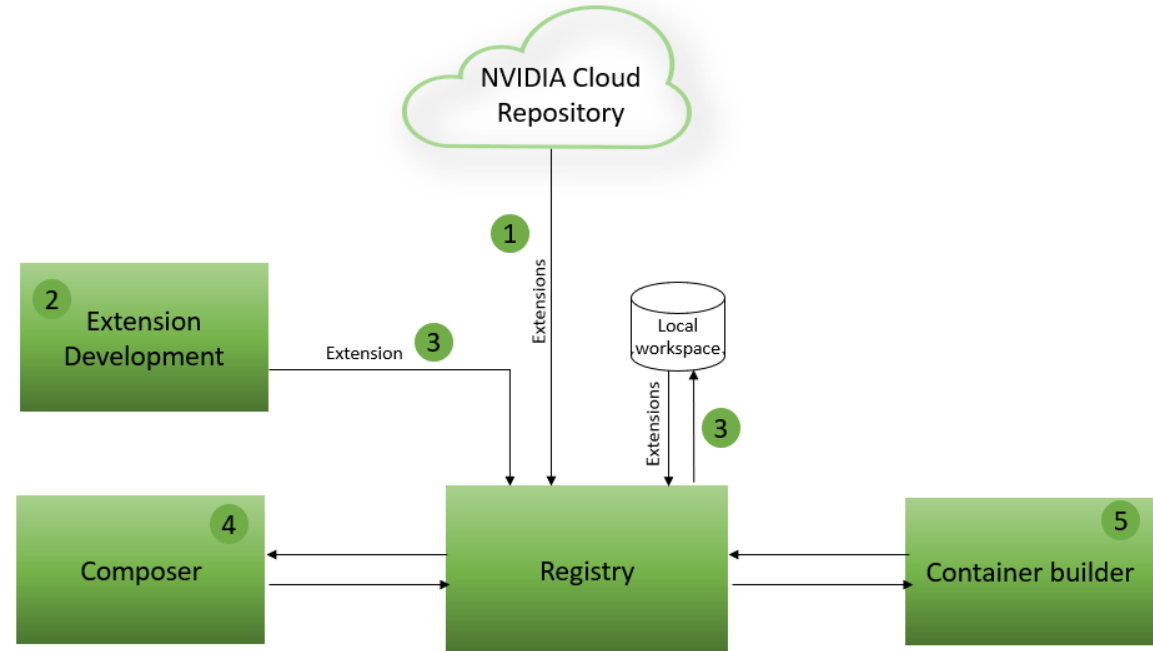
Graph Composer Demo

© 2022 NVIDIA
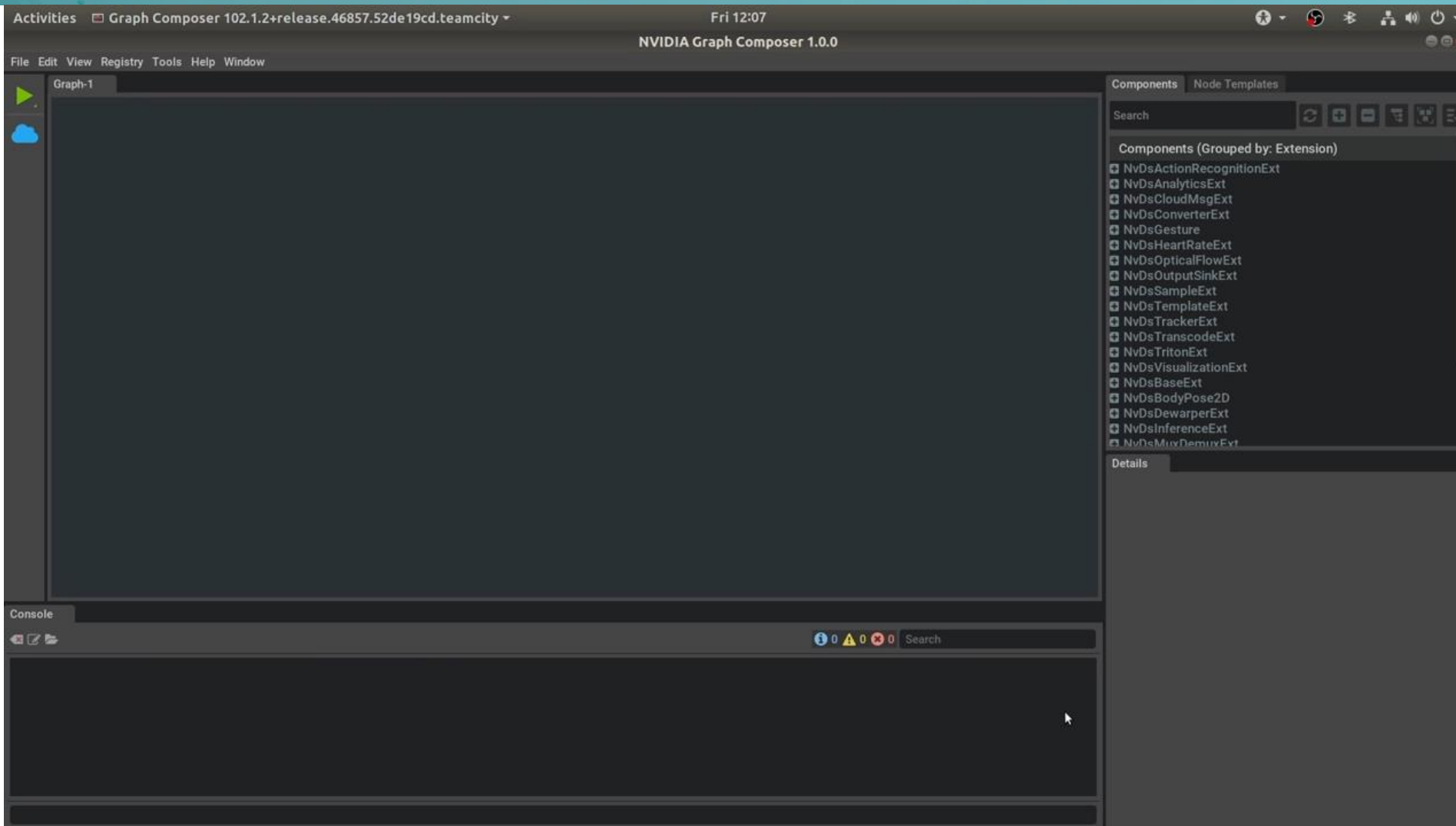
Cloud Native On Jetson
Containers & Micro Services

Easier and Faster Deployment
- Eliminates complex, time-consuming builds and installs

Agile and Easier Development
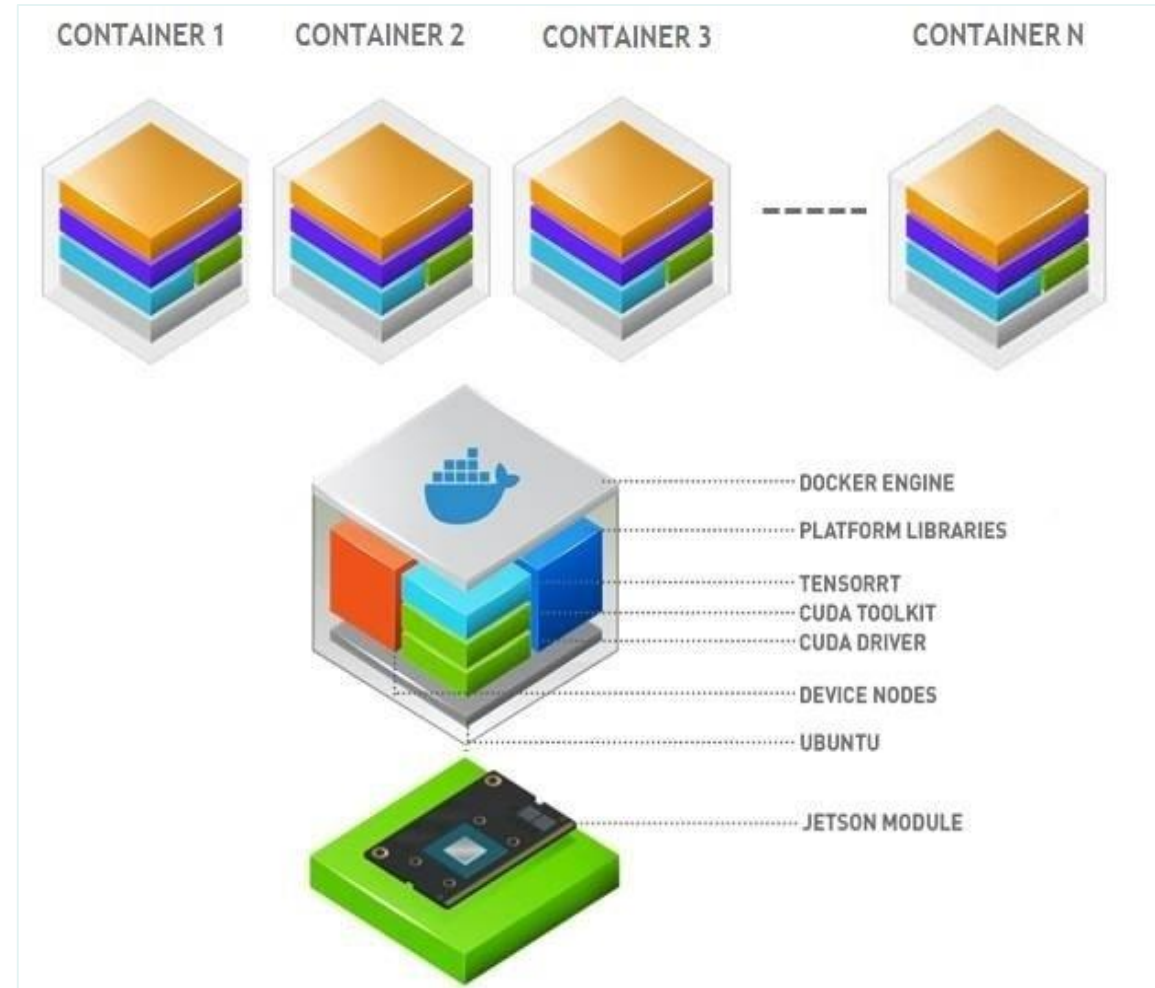- Update specific modules not the whole system

Scalability
- Push the right container to the right platform

Portable
- Deploy across various environments, from test to production with minimal changes

Consistency Across Portfolio
- Update once and push broadly

CONTAINER 1   CONTAINER 2   CONTAINER 3   CONTAINER N

DOCKER ENGINE
PLATFORM LIBRARIES
TENSORRT
CUDA TOOLKIT
CUDA DRIVER
DEVICE NODES
UBUNTU
JETSON MODULE

Call to Action

Developer Resources

DeepStream SDK

https://developer.nvidia.com/deepstream-sdk

NVIDIA Deep Learning Institute

https://www.nvidia.com/en-us/training/

Jetson Module Family

https://developer.nvidia.com/embedded-computing

2022 Embedded Vision Summit

- **NVIDIA - Booth 806**

- **NVIDIA Partner Booths**
  - 324 - deci.
  - 318 - NetworkOptix
  - 417 - VC vision components®
  - 618 - ARROW
  - 523 - SEQUITUR LABS

  - 609 - FRAMOS
  - 608 - D3
  - 709 - Leopard Imaging
  - 604 - BASLER
  - 607 - e-con Systems Your Product Development Partner

Questions and Answer