Edge Computing 101
An Introduction to the Smart Edge

March, 24

Chen Su, Sr Technical Product Marketing Manager

Polling Question 1

Polling question: How much do you know about edge computing?

- **Learning** the concept
- **Researching** use cases for edge computing
- **Evaluating and Prototyping** solutions
- **Implementing** solutions in production

AGENDA

What is Edge Computing and its benefit?

Use Cases and Technology Trends

Challenges of building solutions and NVIDIA’s Edge AI Offering

Resources to Advance your Knowledge at GTC

What is Edge Computing?

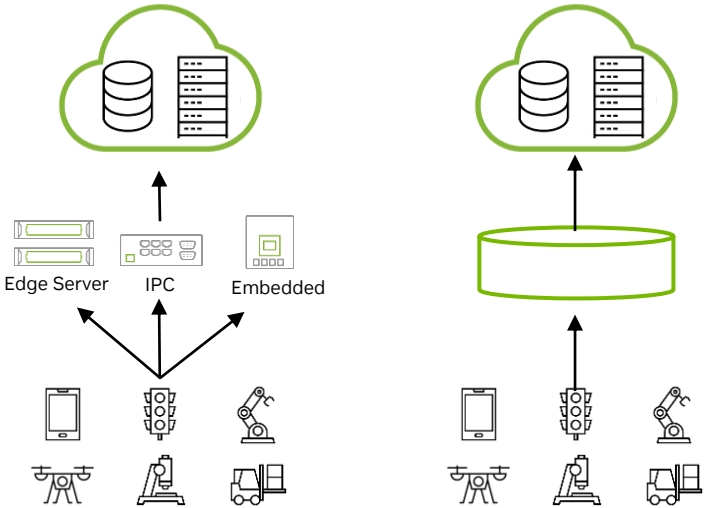
Low Latency, reduced bandwidth requirement, data privacy and improved efficiency

EDGE COMPUTING

Brings computation closer to the network edge where the data is gathered at source.

Average response time is milliseconds

Significantly reduced bandwidth



CLOUD COMPUTING

Location coverage is global because data centers are located around the world.

Average response time can still be milliseconds, but also minutes or days.

Requires a larger amount of bandwidth.

https://blogs.nvidia.com/blog/what-is-edge-computing/

Benefits of Edge AI

Building Intelligent Locations

Edge Computing

LOWER LATENCY

Instantaneous results from low latency environments drive operational safety and better customer experience

REDUCED BANDWIDTH

Reduces data transit and storage costs. Additional bandwidth allows organizations to add more sensors and AI applications

DATA SOVEREIGNTY

Ensure data sovereignty as well as protect privacy and intellectual property

AI Inference

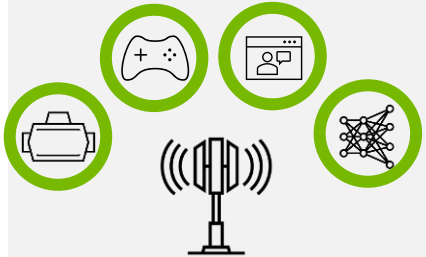
Intellligence

Infer information from data such as text, image, videos, etc to provide insights, make predictions, and take actions

Deliver actionable real-time insights and intelligence with continuous improvement

What Are The Different Types Of Edge AI?

PROVIDER EDGE



Content Delivery
Gaming
AR/VR
AI-as-a-Service

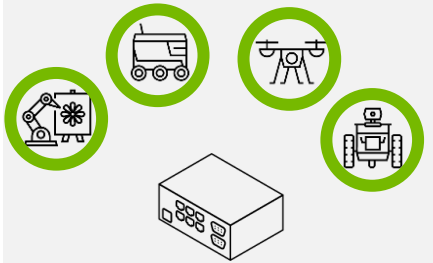
ENTERPRISE EDGE

Intelligent Warehouses
Micro Data Centers
Remote/Branch Offices
Smart Retail Stores

INDUSTRIAL EDGE

Private 5G
Factory Inspection
Medical Device
Autonomous Checkout

Embedded Edge



Robotics
Drone
Intelligent Traffic Systems
Autonomous Checkout

Edge AI Use Cases Accelerates Digital Transformation

Intelligence is Instantaneous

1 Transportation and Logistics

- Digital Signage
- Suspicious Activity Monitoring
- Warehouse Autonomous Mobile Robot
- Traffic flow management

2 Industrial and Manufacturing

- Industrial Inspection
- Perceptive Robotics
- Materials Handling
- Factory Floor Video Analytics
- Digital Twin and Sensor Fusion
- Preventive Maintenance
- Additive Manufacturing

3 Smart Retail

- Automated Checkout
- Store Traffic Analytics
- Inventory Management
- Shopper Analytics
- Digital Signage
- Social Distancing Detection

4 Smart City

- Traffic Analytics
- Vehicle Counting
- Number Plate Detection
- Surveillance and Public Safety
- Smart Parking System

5 Healthcare and Life Science

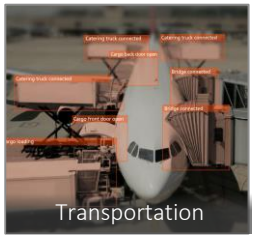
- Surgical Robot
- Medical Image Assistant
- Telepathology
- Patient Health Monitoring
- Digital Health System

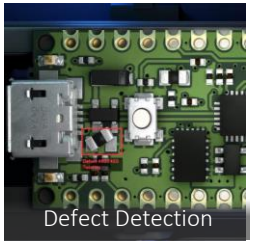
6 Agriculture

- Intelligent Robot Assistant for Harvesting
- Autonomous Tractor
- Selective Spraying system
- Smart Farm Machines
- BeeHome with Robots and AI
- Livestock Health Management

Technology Trends Pushing Us To The Edge

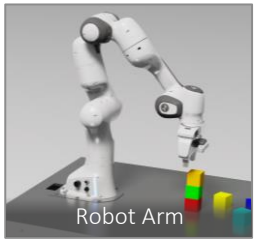
The convergence of AI and IoT forces new infrastructure

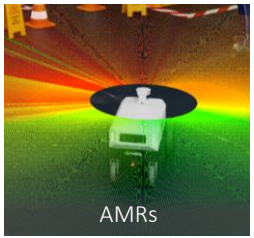




Intelligent Video Analytics

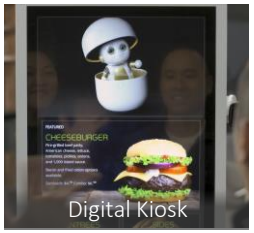
Create actional insights

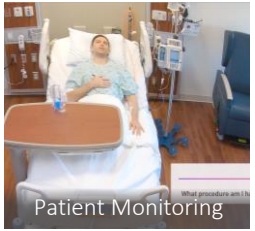


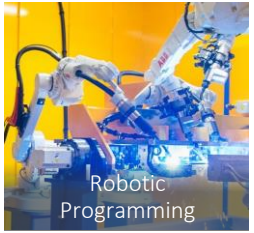


AI-powered Robotics

Deploy Autonomous Robots



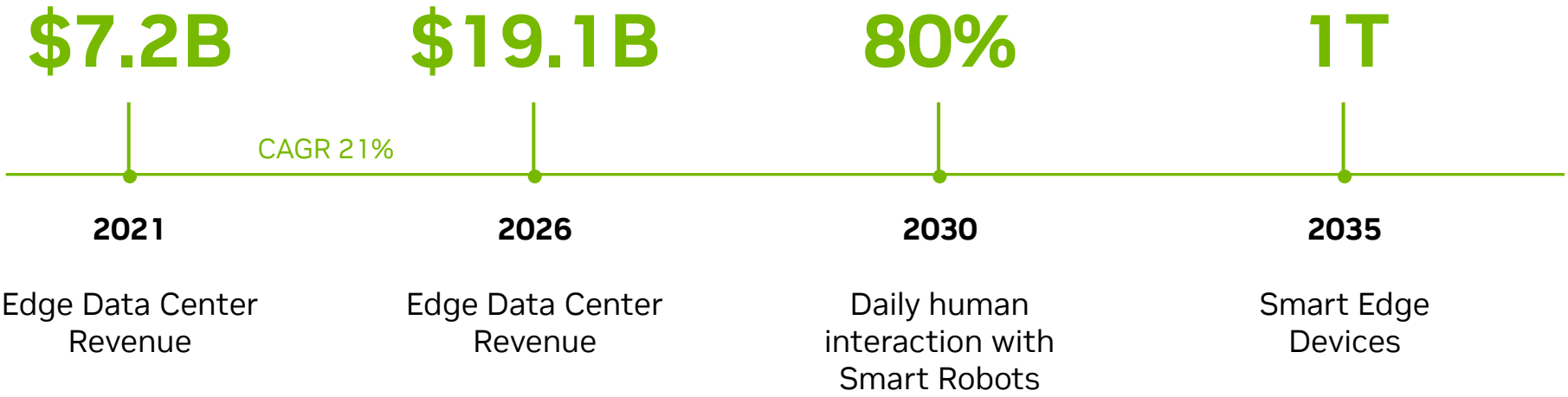




Generative AI

Build your AI-assist Agent

Edge AI Demand is Exploding



Markets and Markets: Marketing Research
Gartner: Emerging Technologies: Smart Robot Adoption Generates Diverse Business Value
arm: Secure Compute for a Cloud-Native Edge

Challenges To Develop and Deploy Edge AI Application

Meeting the various unique requirements and taking years from prototyping to productions

Designed for the Edge
Durable and small enough for industrial environments with low power consumption

Edge Specific Software
SDKs and frameworks to ensure applications are developed to be optimized for the edge

Security
Edge devices are often in public places and more vulnerable to physical and cyber attacks

Low Latency
High speed connectivity is necessary to connect devices with the cloud and / or data center

Data Storage & Processing
Must be able to compute and store large amounts of data to reduce latency and improve performance

Distributed Computing
Need to coordinate widely dispersed edge devices in order to perform complex AI tasks

Orchestration
Tools to deploy, configure, and monitor edge devices and applications

Functional Safety
Assurance that systems function safely and reliably, especially as human and machines interaction increases

NVIDIA-Certified Systems

Certification for Enterprise and Industrial Edge

Benefits

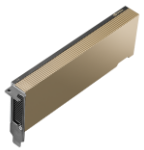
NVIDIA Certified

Validates the Best Baseline Configuration

- HIGH PERFORMANCE
- MANAGEABILITY
- SECURITY
- SCALABILITY

Products

Enterprise



Data Center Edge GPU (T4, A2, L4)
40-72W
Up to 192 GB
Server Rack

Industrial

IGX Orin
10W-70W
64GB
198mm X 243mm

Applications

Recommender System
Computer Vision
Generative AI
Route Planning

Healthcare
Manufacturing Inspection
Proactive Safety
Generative AI
Sensor Fusion

NVIDIA AI Enterprise

End to end generative AI software

MLOps

AI Applications

NVIDIA AI Enterprise

Infrastructure Management

Cloud Native Management and Orchestration
GPU Operator, Network Operator

Cluster Management
Base Command Manager Essentials

Infra Acceleration Libraries
Magnum IO, vGPU, CUDA

Application Frameworks

LLM
NeMo

Speech AI
Riva

Cybersecurity
Morpheus

Medical Imaging
Clara

More

AI Development

Data Science / Prep
RAPIDS, RAPIDS Accelerator for Apache Spark

Model Training and Customization
NeMo, TAO, PyTorch, TensorFlow

Deploy at Scale
Triton Inference Server

Optimize for Inference
TensorRT, TensorRT-LLM

Cloud | Data Center | Workstations | Edge

NVIDIA Jetson For Embedded Edge

Designed for Scalable, Flexible Hardware Systems Deployed at Far Edge

Benefits

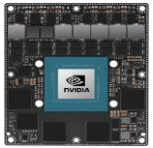
Energy Efficient

Small Footprint

Flexible IO

Customizable

Products



Jetson AGX Orin Series
15- 60W
32GB/64GB
100mm x 87mm

Jetson Orin NX Series
10 - 25W
8GB/16GB
45mm x 70mm

Jetson Orin Nano Series
7 - 15W
4GB/8GB
45mm x 70mm

Applications

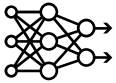
Computer Vision
Robotics & Drone
Healthcare
Manufacturing Inspection
Smart Kiosk
Sensor Fusion
Generative AI

Jetson Software

Accelerates AI Applications and Time-to-Market

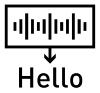
APPLICATION FRAMEWORK

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



Vision AI
Metropolis

Robotics
Isaac



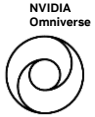
Speech AI
Riva

AI DEVELOPMENT

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

Pre-trained models

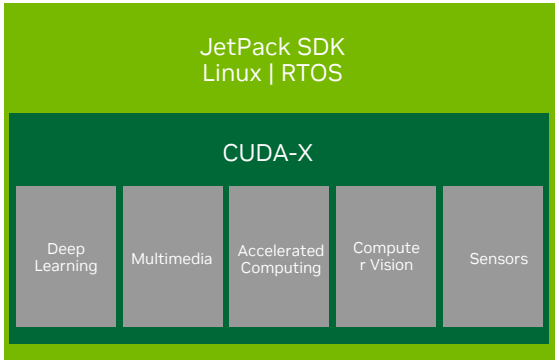
TAO Toolkits



Synthetic Data Generator

JETPACK SDK

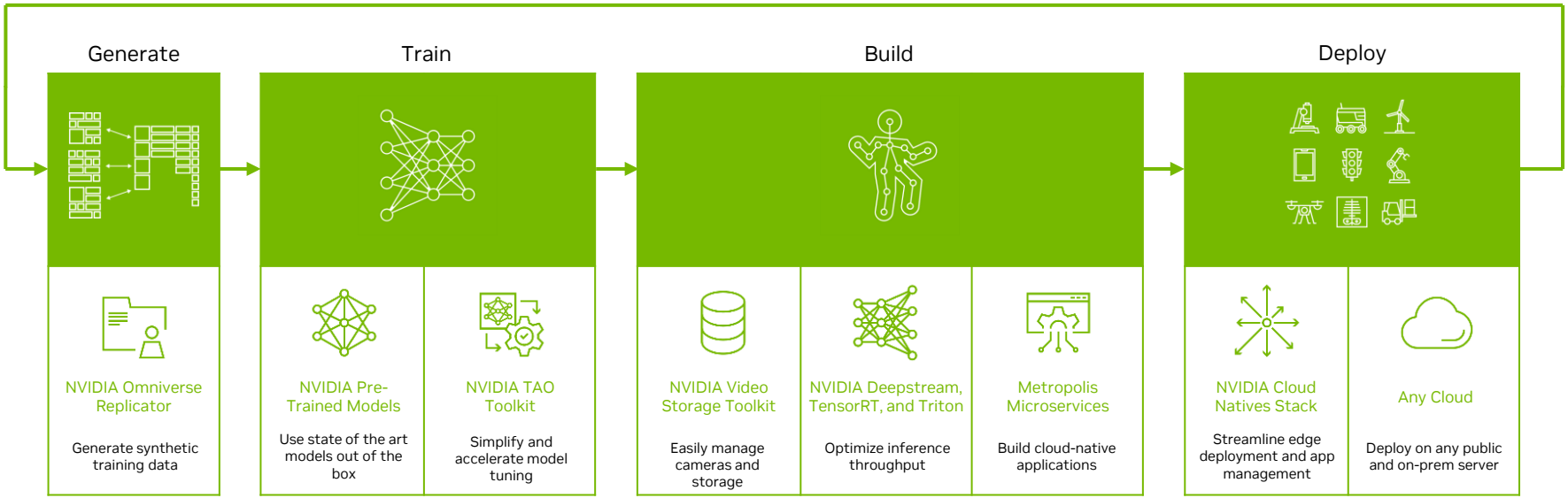
Comprehensive SDK for building end-to-end accelerated AI pipeline




Honey Barbecue
Classic
Classic
Barbecue
ANSWER THE GROWL
Doritos
Doritos
ANSWER THE GROWL
Cheetos
Flamin' Hot
PEPCOIN
Cheetos
PUFFS
Cheetos
PUFFS
Cheetos
Crunchy
Cheetos
Crunchy

End-to-end Vision AI Development With Metropolis

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



NVIDIA Industry Safety Technology

Deploying AI to assist safety of robotics and manufacturing

Simulation predictive safety:

- Isaac Sim add-on for representation of safety assets (safety hazards and safety fields representation & relationships) for safety verification and validation, safety confidence view and safety risk assessment visualization

ISO 13482, ISO/DIS 10218-2, ISO 3691-4, ISO/TR 23482, ISO 18464

Safety foundations:

- SW: Safety Extension Package, to enable safety diagnostics and safety communication
- HW: SoC safety diagnostics, including Functional Safety Island (FSI) and Safety MCU

IEC 61508 & ISO 13849

Outside-in proactive AI safety:

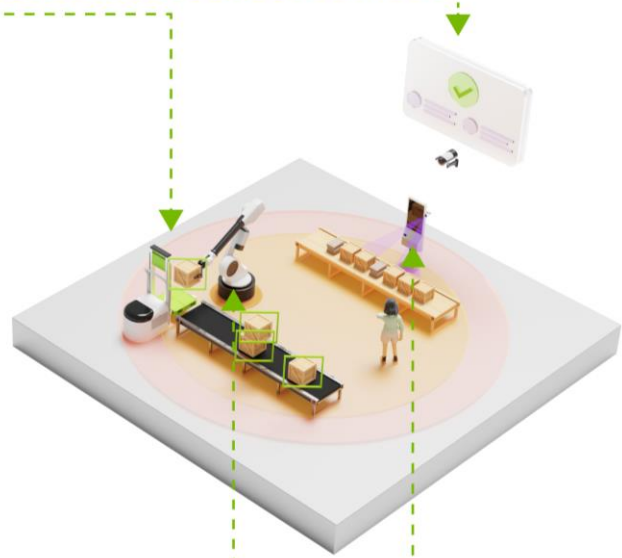
- Metropolis add-on for stationary camera-based monitoring and supervision with AI of industrial machines, AMRs, inspection of safety-related equipment, workers safety and safety hazards detection

ISO/IEC TR 5469, ISO/IEC TS 22440, ISO 13857, IEC 61508

Inside-out reactive AI safety:

- ISAAC add-on for perception-based AMR and manipulators safety

IEC 61508 & ISO 13849, ISO 10218, RIA 15.08

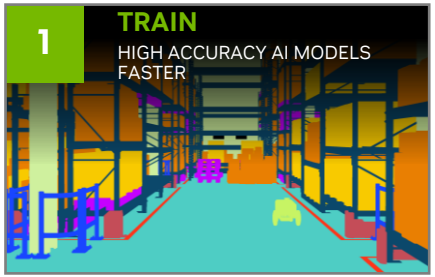



FRAGILE

End-to-end Robotics With NVIDIA Isaac

Smarter Robots Developed Faster Leveraging NVIDIA AI and Omniverse

Leverage **Isaac Sim** - available local and in the cloud - to create physically accurate virtual robots in photorealistic environments to develop and test every aspect of their operation.



Start with **NVIDIA pre-trained models** and augment with synthetic data generated in **Omniverse Replicator** and use **NVIDIA TAO** to train and achieve target performance

Save time by leveraging **Nova Orin** and add AI to your NVIDIA **Jetson** based robots with **Isaac ROS** for Robot Perception,and **DeepStream** for Vision AI

Optimize productivity of managing fleets of robots with **Isaac Mission Dispatch** for task assignment, **CuOpt** for route optimization and **Isaac Mapping** for 3D Mapping services.

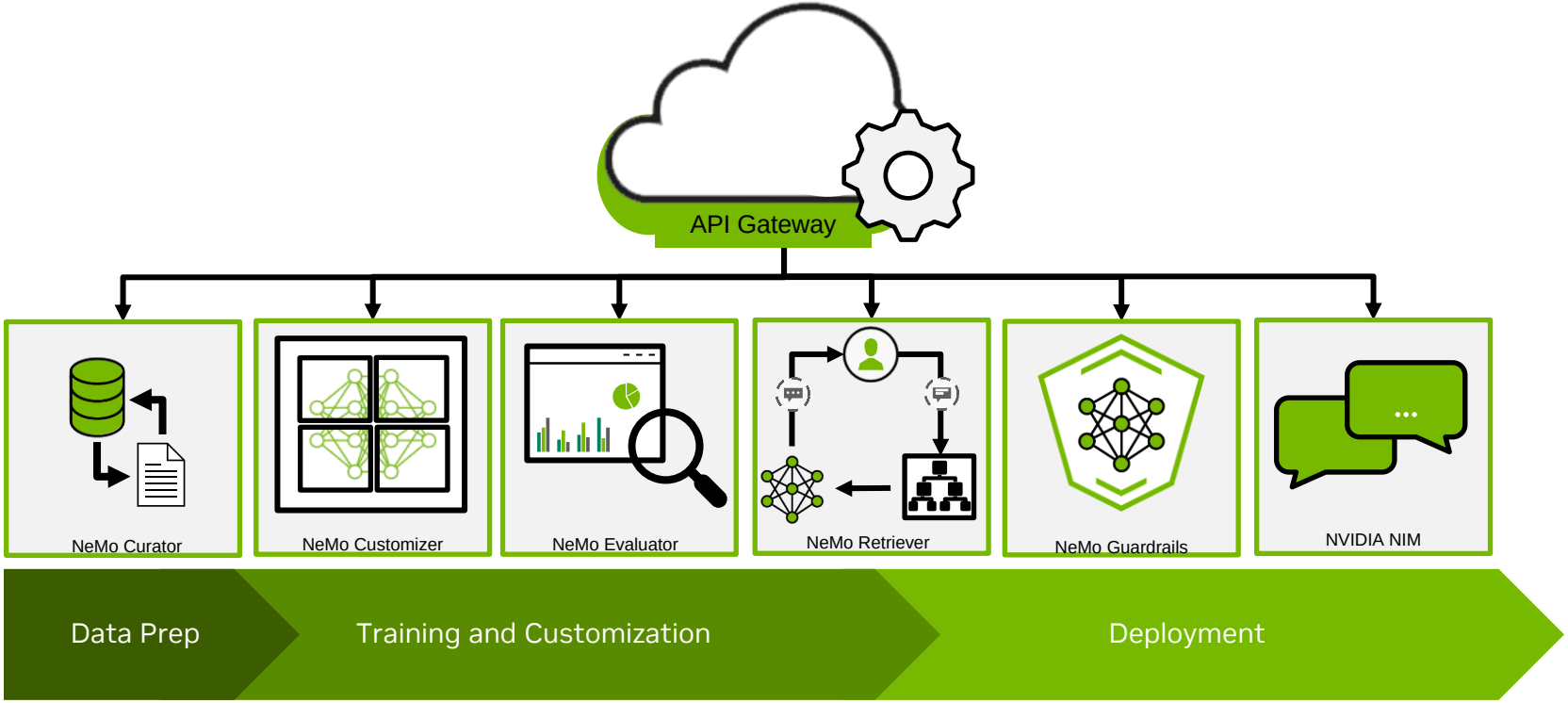
NVIDIA

```
    input_right_camera_frame_(declare_parameter<std::string>("input_right_camera_frame", "")),
    input_imu_frame_(declare_parameter<std::string>("input_imu_frame", "imu")),
    // This is the gravity vector in ROS coordinate frame.
    // Default value assumes that camera is horizontal to the floor.
    gravitational_force_(declare_parameter<std::vector<double>>(
        "gravitational_force",
        {0.0, 0, -9.8})),

    // frames hierarchy:
    // map - odom - base_link - ... - camera
    publish_odom_to_base_tf_(declare_parameter<bool>("publish_odom_to_base_tf", true)),
    publish_map_to_odom_tf_(declare_parameter<bool>("publish_map_to_odom_tf", true)),
    invert_odom_to_base_tf_(declare_parameter<bool>("invert_odom_to_base_tf", false)),
    invert_map_to_odom_tf_(declare_parameter<bool>("invert_map_to_odom_tf", false)),
    map_frame_(declare_parameter<std::string>("map_frame", "map")),
    odom_frame_(declare_parameter<std::string>("odom_frame", "odom")),
    base_frame_(declare_parameter<std::string>("base_frame", "base_link")),

    override_publishing_stamp_(declare_parameter<bool>("override_publishing_stamp", false)),

    path_max_size_(declare_parameter<int>("path_max_size", 1024)),

    msg_filter_queue_size_(declare_parameter<int>("msg_filter_queue_size", 100)),
    image_qos_(parseQosString(declare_parameter<std::string>("image_qos", "SENSOR_DATA"))),
    // Subscribers
    left_image_sub_(message_filters::Subscriber<sensor_msgs::msg::Image>(
        this, "stereo_camera/left/image", image_qos_)),
    left_camera_info_sub_(message_filters::Subscriber<sensor_msgs::msg::CameraInfo>(
        this, "stereo_camera/left/camera_info")),
    right_image_sub_(message_filters::Subscriber<sensor_msgs::msg::Image>(
        this, "stereo_camera/right/image", image_qos_)),
    right_camera_info_sub_(message_filters::Subscriber<sensor_msgs::msg::CameraInfo>(
        this, "stereo_camera/right/camera_info")),
    imu_sub_(create_subscription<sensor_msgs::msg::Imu>(
        "visual_slam/imu", rclcpp::QoS(100),
        std::bind(&VisualSlamNode::ReadImuData, this, std::placeholders::_1))),
    // Publishers:
```

Tabnine

Building Generative AI Applications for the Enterprise

Build, customize, and deploy generative AI models with NVIDIA NeMo.

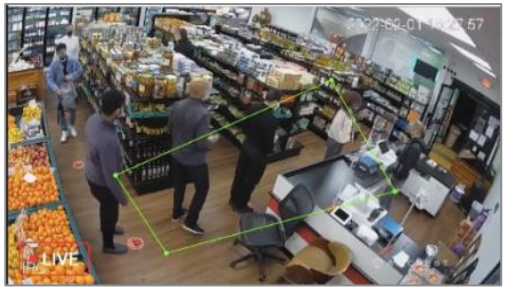


Microsoft Azure aws Google Cloud ORACLE NVIDIA DGX Cloud DELL Technologies Hewlett Packard Enterprise Lenovo SUPERMICRO NVIDIA

Productize Gen AI with Jetson Microservices At the Embedded Edge

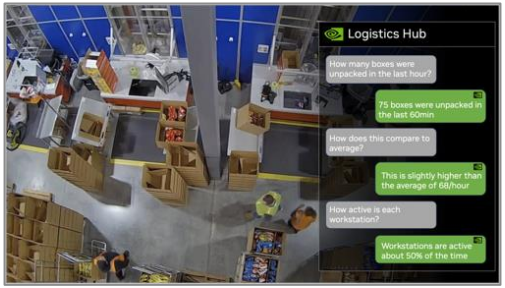
Cloud-Native

- API-driven microservices
- Fully containerized
- Modular
- Extensible



Suite of Pre-built Services

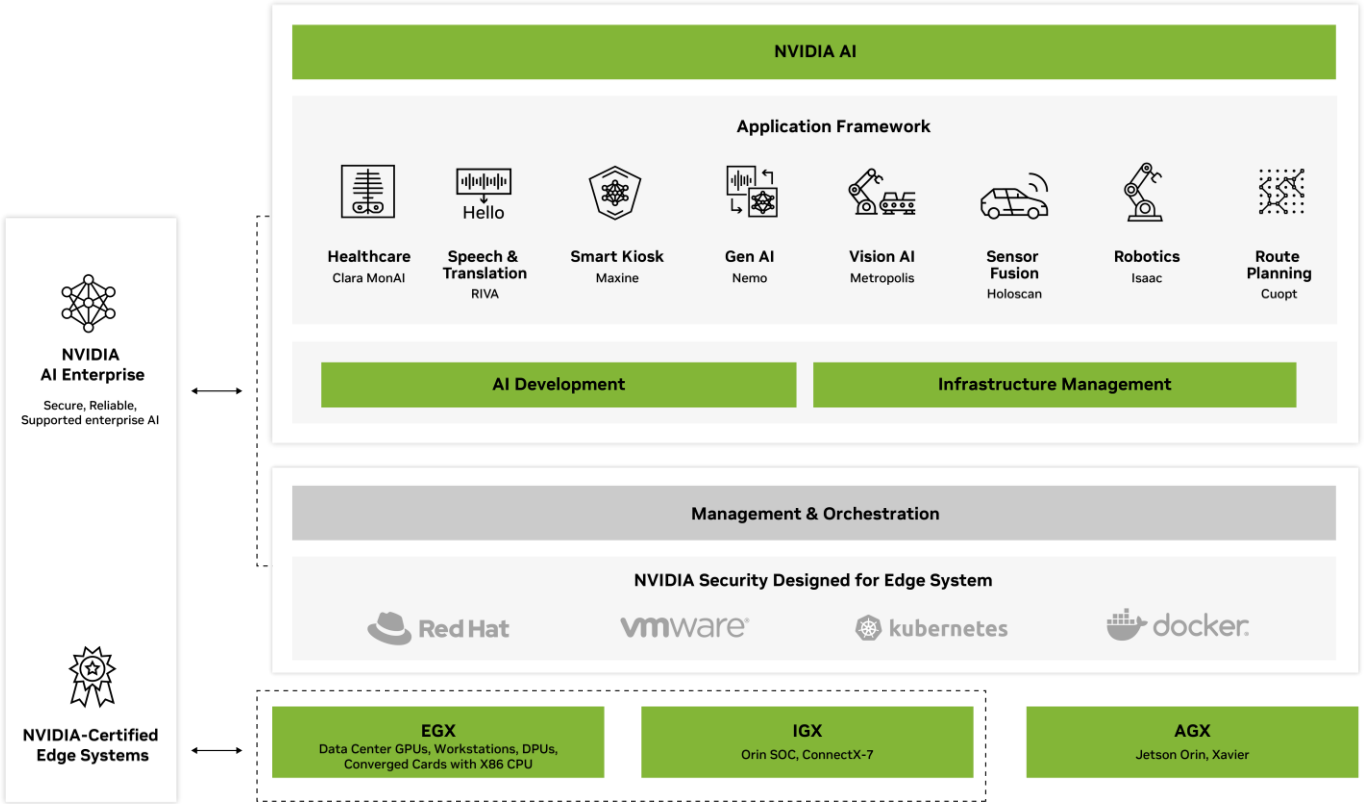
- Sensor storage & management
- AI perception services
- IoT gateway
- Monitoring, and more



Ready for Generative AI

- Flexible API-driven modules make prompting easy
- Plug and Play Open-Source Community Models optimized for Jetson from NVIDIA Jetson AI Lab

NVIDIA Edge AI Stack



Edge Computing Conference Sessions

Discover how AI is transforming edge computing solutions in retail, manufacturing, healthcare, smart cities, and more.

Featured Talks

Edge Computing 101: An Introduction to the Smart Edge
NVIDIA
Monday, March 18, 10:00 AM PDT

A New Class of Cloud-Native Applications at the Far Edge with Generative AI
NVIDIA
Tuesday, March 19, 9:00 AM PDT

Transforming Agriculture with AI and Computer Vision
Blue River Technology (John Deere)
Tuesday, March 19, 9:30 AM PDT

Connect With the Experts: Connect With Jetson Embedded Platform Experts
NVIDIA Panel
Tuesday, March 19, 2:00 PM PDT

Functional Safety for Industry 4.0: Keeping Supply Chains Safe, Secure, and Efficient using AI at the Edge
Amazon, Rockwell Automation, SICK
Wednesday, March 20, 8:00 AM PDT

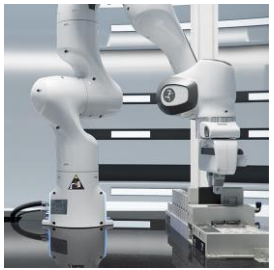
Democratizing AI for Agriculture: Bridging the Digital Divide
Monarch Tractor
Wednesday, March 20, 9:00 AM PDT

AI-Based 6D Object Pose Estimation on Jetson: End-to-End Training and Deployment within the NVIDIA Ecosystem
D3
Wednesday, March 20, 11:00 AM PDT

Special Events and Show Floor Exhibits

AI at The Edge Pavilion

Jetson and Robotics Developer Day
Thursday, March 21, 8:00 AM PDT



Edge Computing Conference Sessions
March 18–21 | www.nvidia.com/gtc

Robotics at GTC 2024

Learn about how NVIDIA and the latest developments in AI are transforming the robotics industry.

Featured Talks

Transforming Agriculture with AI and Computer Vision
Blue River Technology (John Deere)
Tuesday, March 19, 9:30 AM PDT

Next Phase of Industrial Robot Skills with AI
Yaskawa Electric Corp.
Tuesday, March 19, 10:00 AM PDT

Robotics and the Role of AI: Past, Present, and Future
Fireside Chat with Marc Raibert
Tuesday, March 19, 3:00 PM PDT

Robotics in the Age of Generative AI
Google DeepMind
Wednesday, March 20, 10:00 AM PDT

Breathing Life into Disney's Robotic Characters with Deep Reinforcement Learning
Disney Research
Wednesday, March 20, 11:00 AM PDT

Panel Discussion on the Impact of Generative AI on Robotics
Ambi Robotics, Covariant, Scaled Foundations, Vayu Robotics
Wednesday, March 20, 2:00 PM PDT

Robotics Conference Sessions

March 18–21 | www.nvidia.com/gtc | #GTC24

Featured Partners

Boston Dynamics, Disney Research, Google DeepMind, John Deere, Techman Robot, Yaskawa Electric Corp.

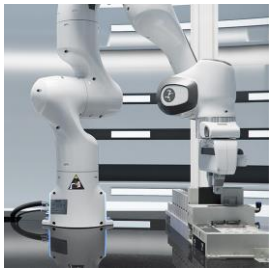
Special Events and Show Floor Exhibits

NVIDIA Robotics Pavilion

AI at The Edge Pavilion

Metropolis Pavilion

Jetson and Robotics Developer Day
Thursday, March 21, 8:00 AM PDT




NVIDIA
Thank You