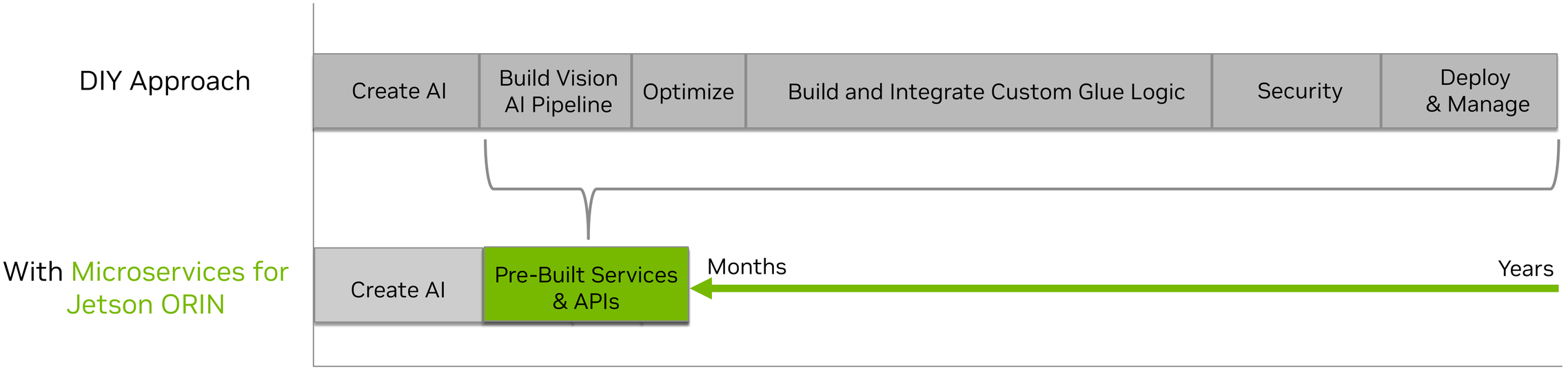
Microservices for Jetson

Chintan Shah, Product Manager

Slashing the Cost of Bringing Edge AI Apps to Production

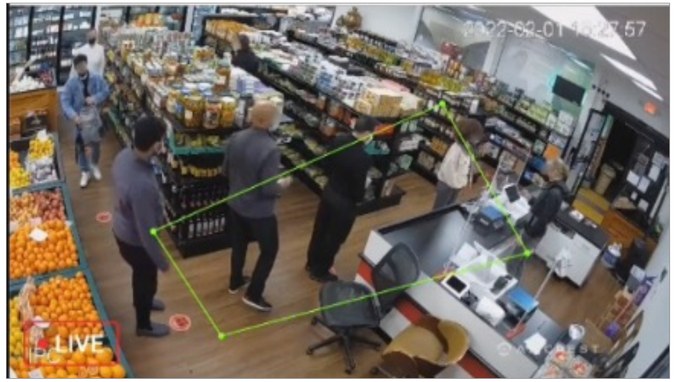


Years to Months = Huge Cost Savings

Microservices for Jetson

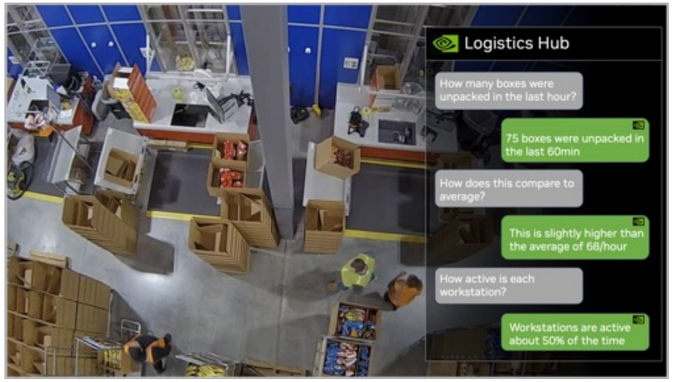
Cloud-Native

- API-driven Microservices
- Fully Containerized
- Modular
- Extensible



Suite of Pre-Built Services

- Sensor Storage & Management
- AI Perception Services
- IoT Gateway
- Monitoring, and more

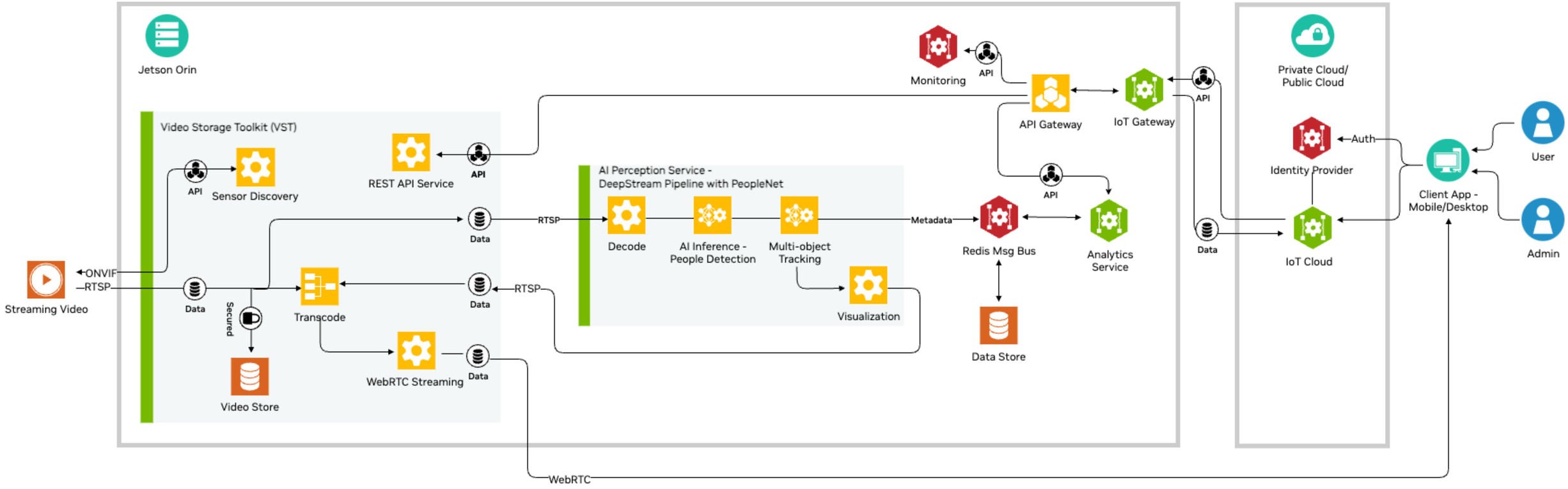


Ready for Generative AI

- Flexible API-driven modules make prompting easy
- Powerful Jetson ORIN Compute capable of running multiple large models

Cloud-Native Vision AI Workflows for the Edge

Use some or all the services



Built for Maximum Performance

Freeing up Compute Resources for other Tasks

Compute Intensive App

		Orin AGX 64	Orin NX 16
Input Streams		16	6
Resource Utilization	CPU	53%	61%
	GPU	42%	28%
	RAM	31GB	8 GB
	DLAx2	78%	33%
	PVA	34%	26%
	VIC	87%	55%

Resources free for other compute-intensive tasks

Maximized utilization of Orin's accelerators

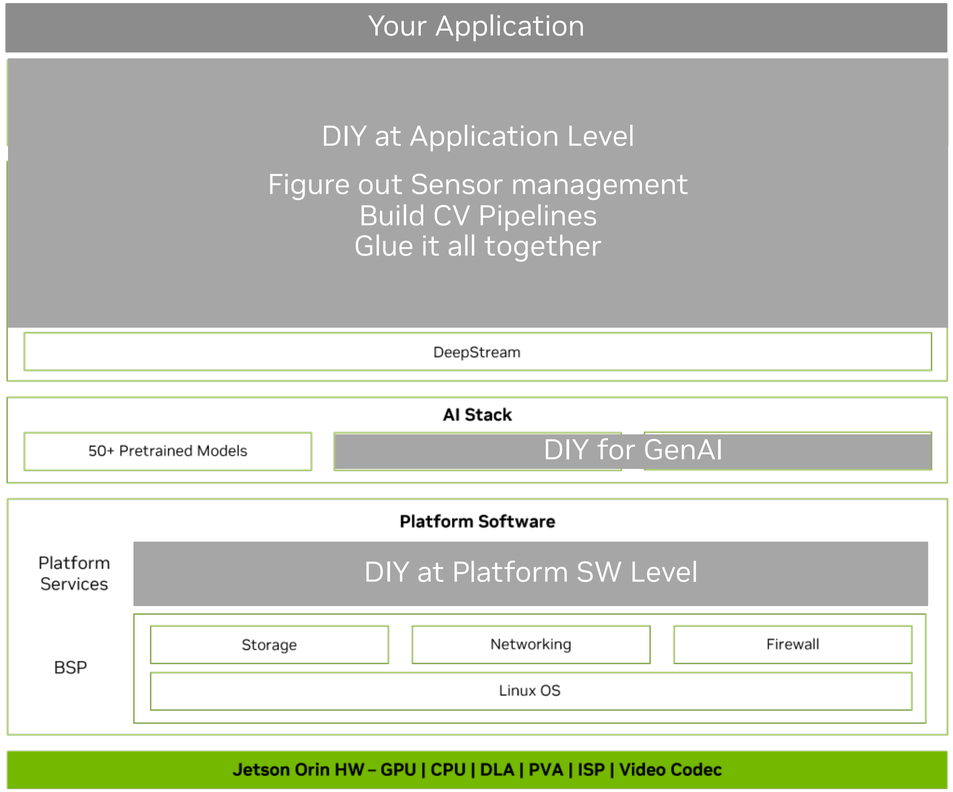
- **Input:** 1080p @30fps
- **AI Perception Service**
  - 16x channels PeopleNet v2.6 (running on DLA)
  - Complex NvDCF Object Tracking (running on PVA)
- **Output:** 4 streams over WebRTC

Collection of APIs & Microservices

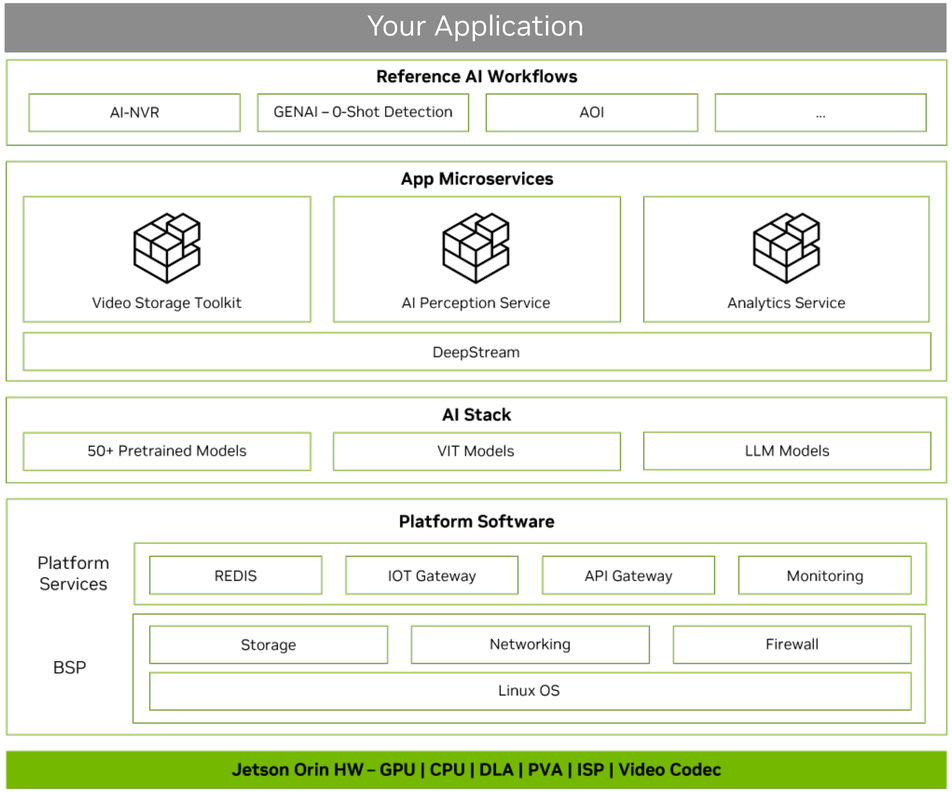
Category	SW Component	Delivery Mechanism (Debian/Container/other)	Hosted Where?
App Services	• Video Storage Toolkit (VST) • AI Perception Service - DeepStream • AI Perception Service - Generative AI • Analytics Service - Line Crossing, ROIs, Counts • Sensor Distribution & Routing (SDR)	Containers	NGC
Cloud Services	• IoT Cloud Service	Containers with deployment scripts	NGC
Platform Services	• Redis • API Gateway • Monitoring • IoT Gateway	Containers	NGC
Reference Workflow	• GenAI Sample App • AI NVR Runtime app	Docker compose pkg	NGC
	Mobile app	APK	NGC

Accelerating Time to Market with Microservices for Jetson

Before
Significant DIY Effort



After
Majority of Processing SW Pre-Built



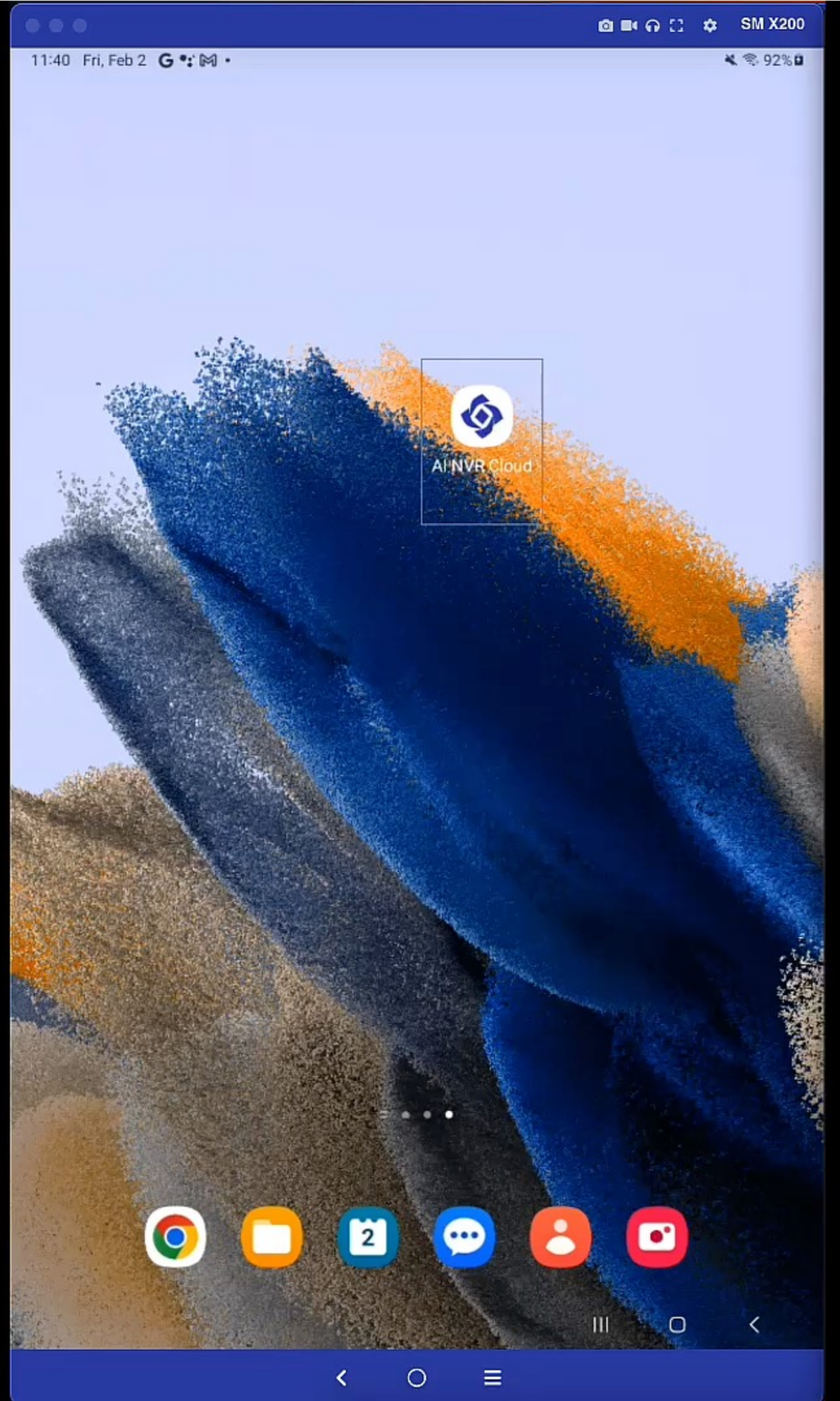
Demo

Terminal Shell Edit View Window Help Tue Jan 23 5:12 PM

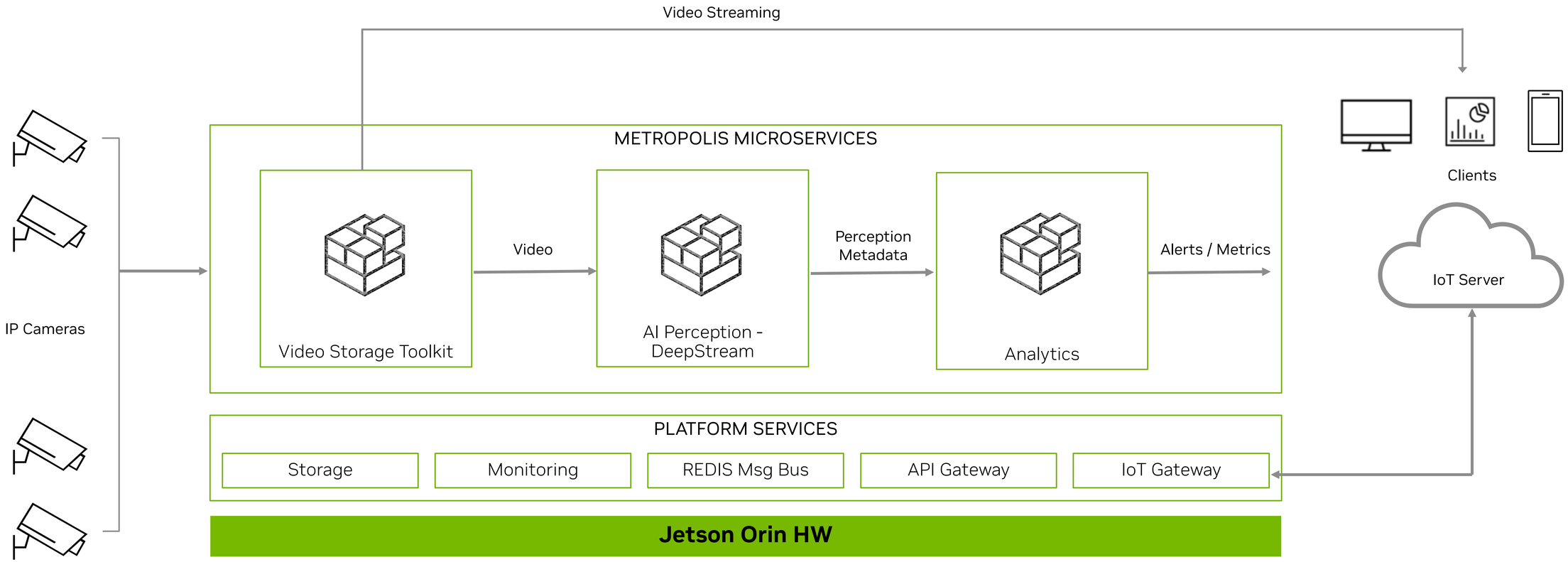
cshah — nvidia@tegra-ubuntu: ~/ai_nvr — ssh nvidia@172.17.170.143 — 245×64

```
base_compose.yaml  compose_agx.yaml  compose_nx.yaml  config  README.md
nvidia@tegra-ubuntu:~/ai_nvr$ sudo docker compose -f compose_agx.yaml down --remove-orphans
[sudo] password for nvidia:
[+] Running 12/12
 ✔ Container deepstream                                   Removed      10.7s
 ✔ Container sdr-emdx                                     Removed      10.2s
 ✔ Container sdr                                          Removed      10.3s
 ✔ Container vst                                          Removed      10.2s
 ✔ Container emdx-webapi                                  Removed      10.3s
 ✔ Container ai_nvr-moj-http-based-init-sdr-emdx-1        Removed       0.0s
 ✔ Container emdx-analytics-02                            Removed      10.2s
 ✔ Container emdx-analytics-01                            Removed      10.2s
 ✔ Container ai_nvr-moj-http-based-init-sdr-1             Removed       0.0s
 ✔ Container ai_nvr-moj-init-vst-1                        Removed       0.0s
 ✔ Container ai_nvr-moj-init-ds-1                         Removed       0.0s
 ✔ Container ai_nvr-moj-http-based-init-emdx-analytics-1  Removed       0.0s
nvidia@tegra-ubuntu:~/ai_nvr$
nvidia@tegra-ubuntu:~/ai_nvr$
nvidia@tegra-ubuntu:~/ai_nvr$
nvidia@tegra-ubuntu:~/ai_nvr$ sudo docker ps
[sudo] password for nvidia:
Sorry, try again.
[sudo] password for nvidia:
CONTAINER ID   IMAGE                                                         COMMAND                  CREATED        STATUS        PORTS     NAMES
fb303a2ddaf0   prom/node-exporter                                            "/bin/node_exporter …"   24 hours ago   Up 24 hours             nodeexporter
0c45f86b9c4c   grafana/grafana                                               "/run.sh"                24 hours ago   Up 24 hours             grafana
290fb841f47d   nvcr.io/e7ep4mig3lne/release/its-monitoring:v1.6.0_arm64      "/root/its_monitorin…"   24 hours ago   Up 24 hours             its-monitoring
6a7f004e3920   prom/prometheus                                               "/bin/prometheus --c…"   24 hours ago   Up 24 hours             prometheus
1fd2bdfde9c2   prom/alertmanager                                             "/bin/alertmanager -…"   24 hours ago   Up 24 hours             alert-manager
5d16583a7227   prom/pushgateway                                              "/bin/pushgateway"       24 hours ago   Up 24 hours             push-gateway
38c669eaa6bc   nvcr.io/e7ep4mig3lne/release/prov-agent:v1.1.0_arm64v8        "/opt/prov-agent/ent…"   24 hours ago   Up 24 hours             prov-agent
fa0c3e9874fe   nvcr.io/e7ep4mig3lne/release/tcpmux-client:v1.2.0_arm64v8     "/opt/tcpmux-client/…"   24 hours ago   Up 24 hours             tcpmux-client
0ca81fcfb1b1   nvcr.io/e7ep4mig3lne/release/ialpha-ingress-arm64v8:0.8       "sh -c '/nginx.sh 2>…"   24 hours ago   Up 24 hours             ingress
2a36814a55f8   redisfab/redistimeseries:master-arm64v8-jammy                 "docker-entrypoint.s…"   24 hours ago   Up 24 hours             redis
3857eaaaadb3   nvcr.io/e7ep4mig3lne/release/vst:nvstreamer_v0.2.24_aarch64   "sh -c '/root/vst_re…"   25 hours ago   Up 24 hours             nvstreamer
nvidia@tegra-ubuntu:~/ai_nvr$ sudo docker compose -f compose_agx.yaml up -d --force-recreate
[+] Running 12/12
 ✔ Container ai_nvr-moj-init-vst-1                        Exited        0.1s
 ✔ Container ai_nvr-moj-http-based-init-sdr-1             Exited        0.1s
 ✔ Container ai_nvr-moj-init-ds-1                         Exited        0.1s
 ✔ Container ai_nvr-moj-http-based-init-sdr-emdx-1        Exited        0.1s
 ✔ Container ai_nvr-moj-http-based-init-emdx-analytics-1  Exited        0.1s
 ✔ Container emdx-webapi                                  Started       0.1s
 ✔ Container emdx-analytics-01                            Started       0.1s
 ✔ Container sdr                                          Started       0.1s
 ✔ Container emdx-analytics-02                            Started       0.1s
 ✔ Container deepstream                                   Started       0.1s
 ✔ Container vst                                          Started       0.1s
 ✔ Container sdr-emdx                                     Started       0.1s
nvidia@tegra-ubuntu:~/ai_nvr$ sudo docker compose -f compose_agx.yaml down --remove-orphans
[+] Running 12/12
 ✔ Container sdr-emdx                                     Removed      10.2s
 ✔ Container emdx-webapi                                  Removed      10.3s
 ✔ Container deepstream                                   Removed      10.7s
 ✔ Container sdr                                          Removed      10.3s
 ✔ Container vst                                          Removed      10.5s
 ✔ Container emdx-analytics-02                            Removed      10.3s
 ✔ Container emdx-analytics-01                            Removed      10.2s
 ✔ Container ai_nvr-moj-http-based-init-sdr-emdx-1        Removed       0.0s
 ✔ Container ai_nvr-moj-http-based-init-sdr-1             Removed       0.0s
 ✔ Container ai_nvr-moj-init-vst-1                        Removed       0.0s
 ✔ Container ai_nvr-moj-init-ds-1                         Removed       0.0s
 ✔ Container ai_nvr-moj-http-based-init-emdx-analytics-1  Removed       0.0s
nvidia@tegra-ubuntu:~/ai_nvr$
```

Chintan Shah


SM X200
11:40 Fri, Feb 2
92%
AI NVR Cloud

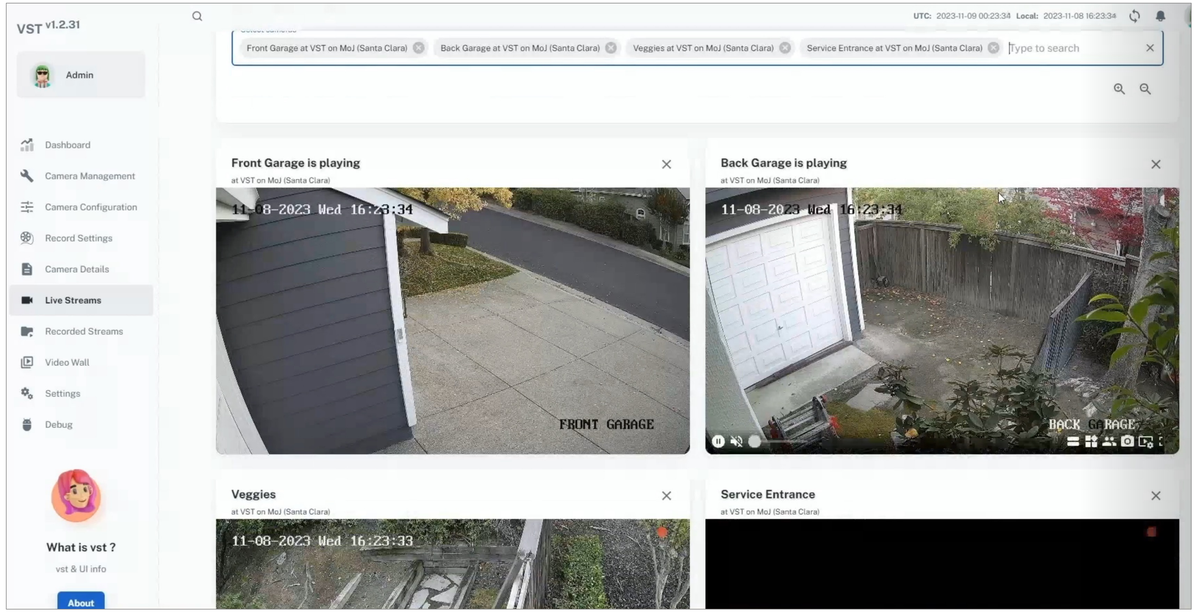
System Architecture for Retail Use Case



Video Storage Toolkit (VST)

WHAT DOES IT HELP WITH

Set of APIs to easily build complex multi-camera ingestion, storage, and streaming.



WHAT'S INCLUDED

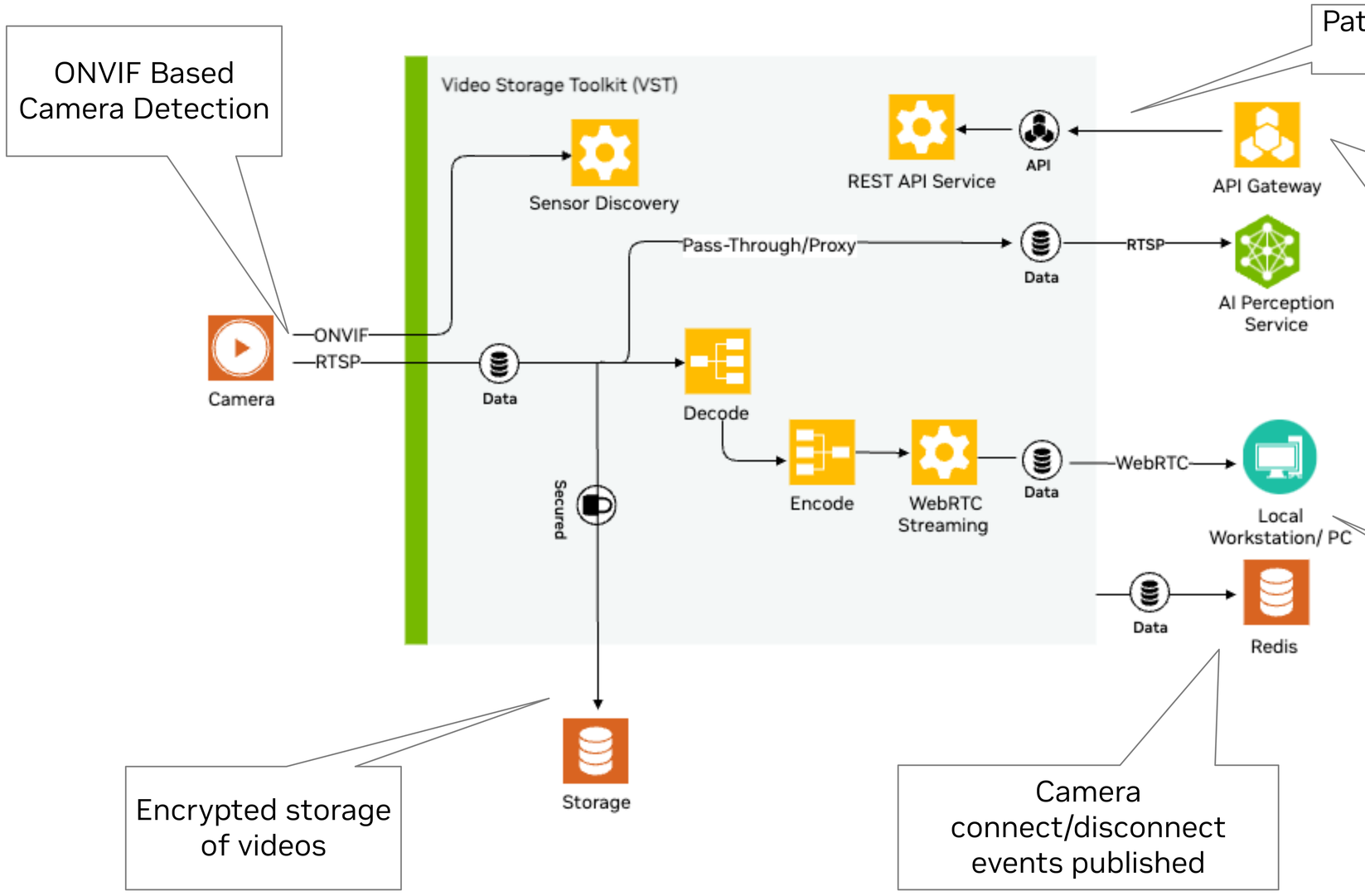
Collection of 20+ APIs including:

- ONVIF camera discovery, monitoring, & management
- Video recording / storage
- Media server, WebRTC streaming
- WebUI for setup and visualization

HOW TO INTEGRATE INTO YOUR APPLICATION

REST APIs

Video Storage Toolkit Functionality



API based interaction

Retrieve Sensor List

`vst/api/v1/sensor/list`

Camera Configuration

`vst/api/v1/sensor/settings`

WebRTC Streaming control plane

`vst/api/v1/live/iceServers`

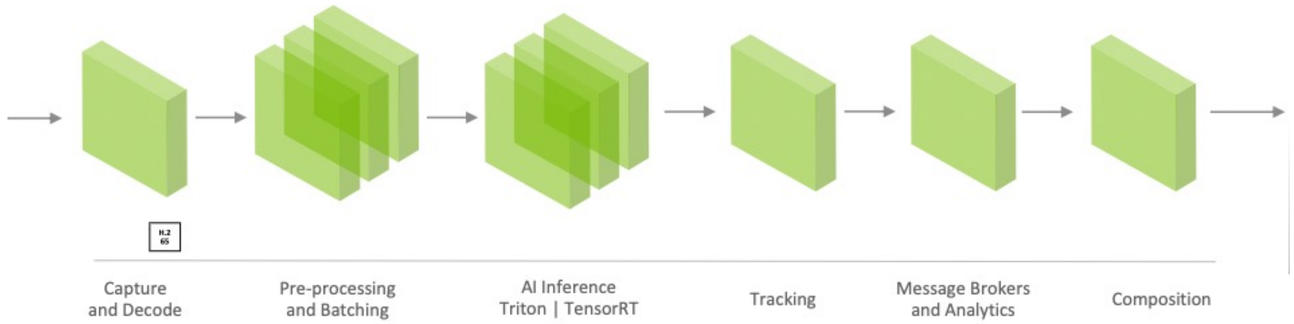
`vst/api/v1/live/stream/start`

`vst/api/v1/live/iceCandidate`

AI Perception Service - Pre-Built DeepStream Pipelines

WHAT DOES IT HELP WITH

Creating an efficient pipeline to process pixels to metadata leveraging DeepStream SDK



WHAT'S INCLUDED

- Up to 16 streams across 2 DLAs (Orin) for PeopleNet
- World-class NvDCF tracker running on PVA
- Dynamic stream discovery, add/remove, reconnection
- Metadata generation and publishing on to Redis
- Integration of app KPIs with Monitoring services

```
"{"version" : "4.0",
   "id" : "2104929",
   “@timestamp" : "2024-02-05T19:56:15.709Z",
   "sensorId" : "Amcrest_2",
    "objects" :
["271893|521.824|50.1403|597.538|142.57|Person",
 "271906|184.405|70.5853|306.508|266.296|Person“]
}
```

HOW TO INTEGRATE INTO YOUR CODE

Deploy using Docker container

Automatic Stream Addition & Removal with SDR

API based interaction

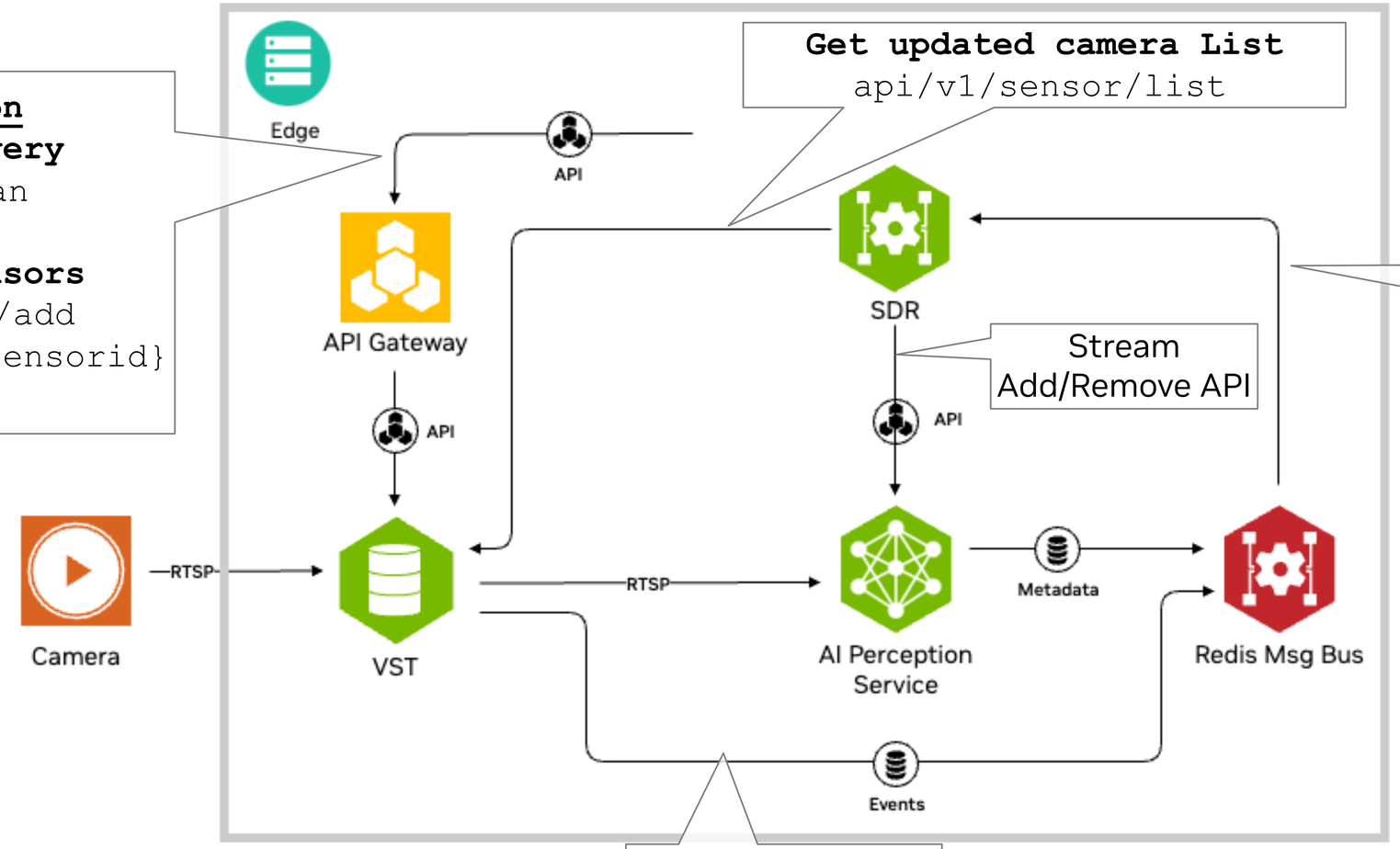
Scan for sensor discovery

`vst/api/v1/sensor/scan`

Manually Add/Remove sensors

`POST vst/api/v1/sensor/add`

`DELETE vst/api/v1/sensor/{sensorid}`



SDR - Sensor Distribution & Routing

Analytics Service - Line Crossing, ROIs, Count

WHAT DOES IT HELP WITH

Using APIs to configure insights and alerts such as line-crossing, ROIs and FOV

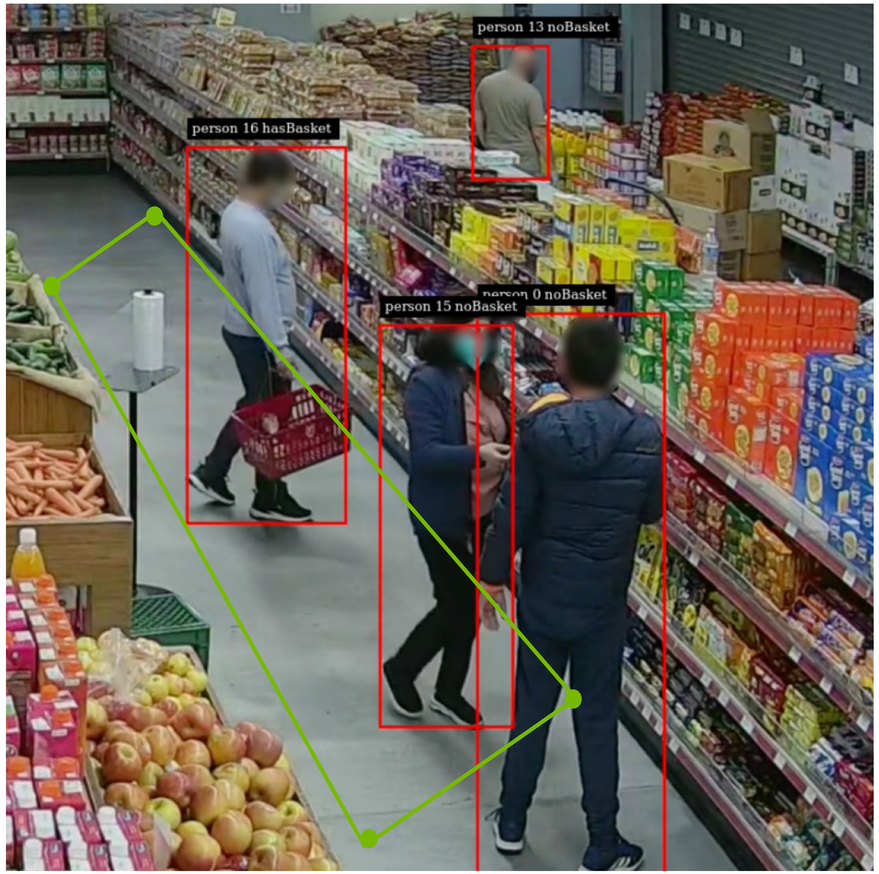
Can help identify traffic flow patterns, entry/exit tracking, encroachment into specific areas

WHAT'S INCLUDED

- Integration into other platform services such as Redis - publishing system event, Monitoring and Storage
- Visualization through overlays
- REST APIs to get insights, set conditions, etc.

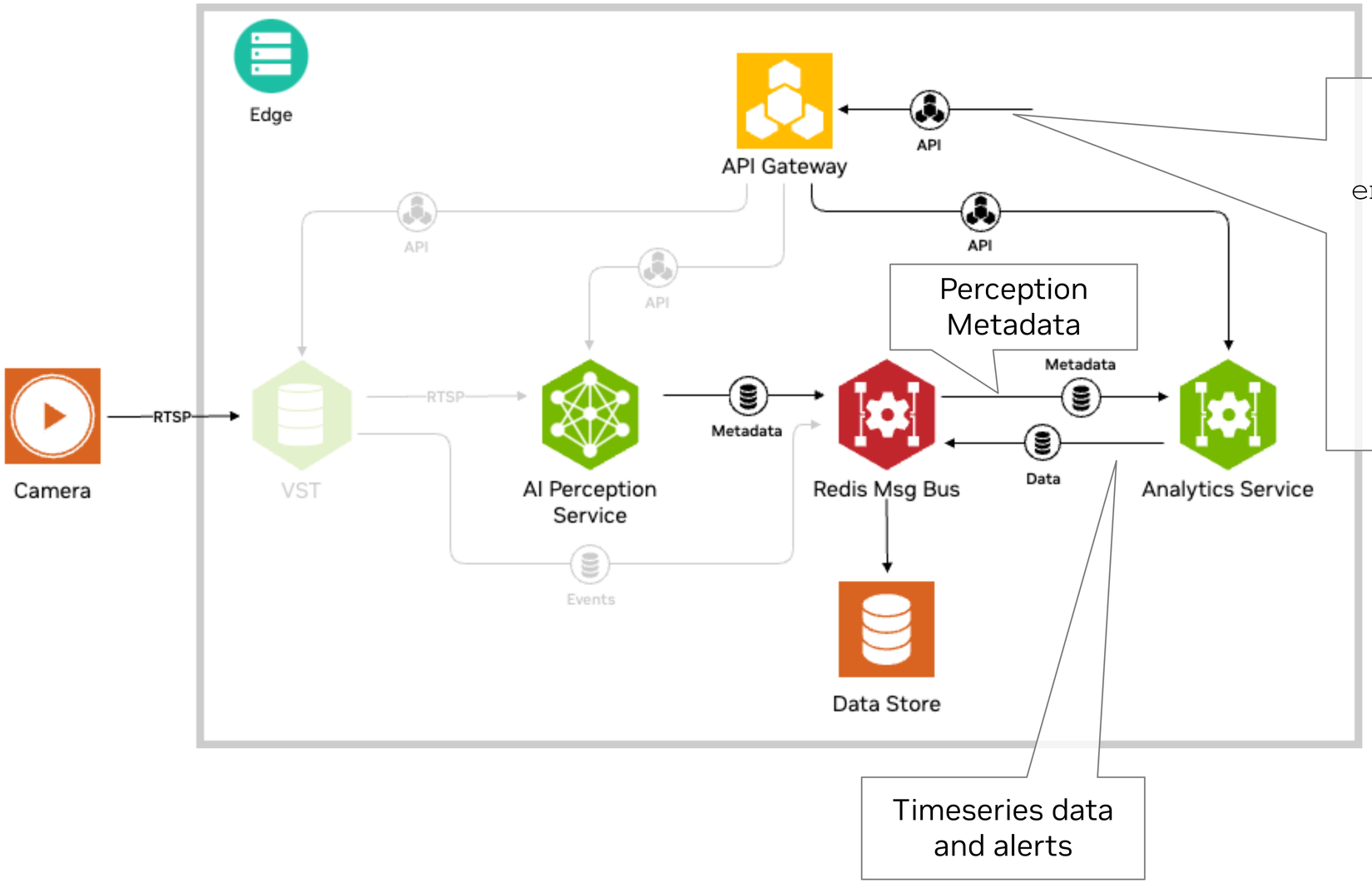
HOW TO INTEGRATE INTO YOUR APPLICATION

REST APIs



ROI - Region of Interest
FOV - Field of View

Streaming Analytics for Spatio - Temporal Understanding



API based interaction

ROI Creation

`emdx/api/config/roi?<sensorId>`

Querying Metrics

`emdx/api/metrics/roi`

Retrieve Alerts

`emdx/api/alerts/roi`

Monitoring / Diagnostics

WHAT DOES IT HELP WITH

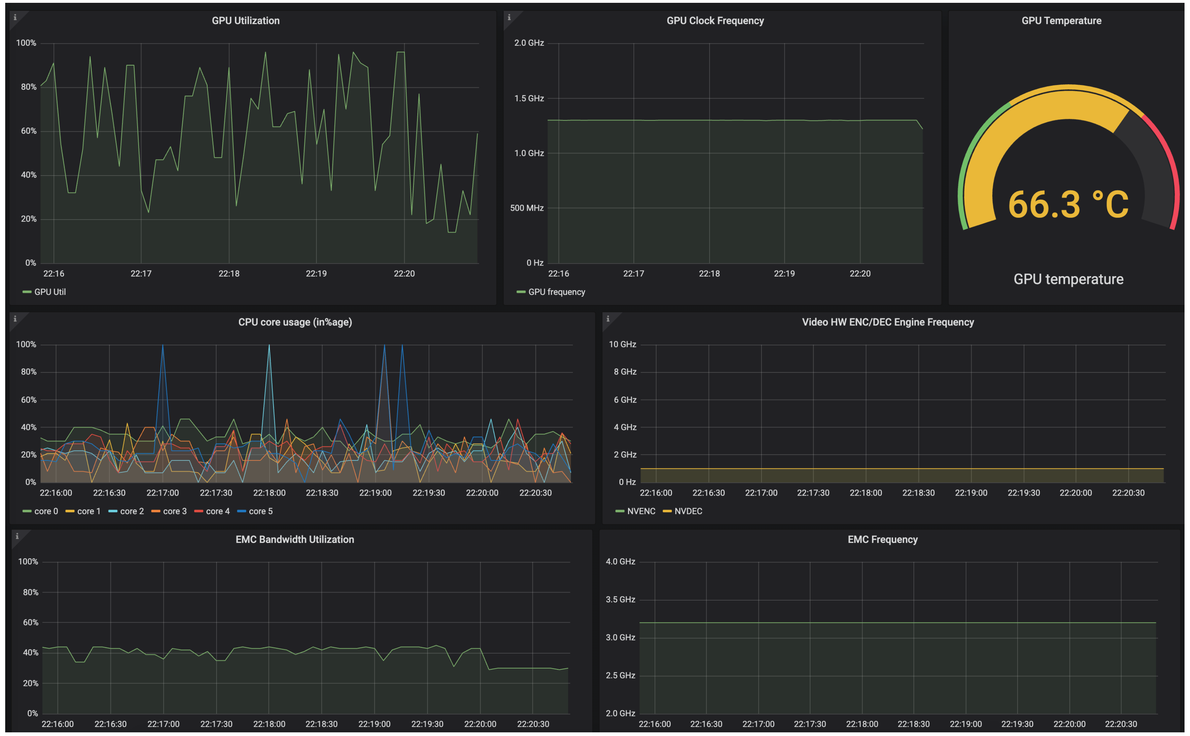
Monitoring of system and application metrics and generating alerts when metrics reach a certain threshold

WHAT'S INCLUDED

- System utilization metrics like CPU/GPU and memory
- QoS metrics like FPS
- Status of microservices
- Connected to 3P services like Prometheus and Grafana

HOW TO INTEGRATE INTO YOUR CODE

Grafana endpoint exposed through API gateway for visualization



IoT and Cloud Service

WHAT DOES IT HELP WITH

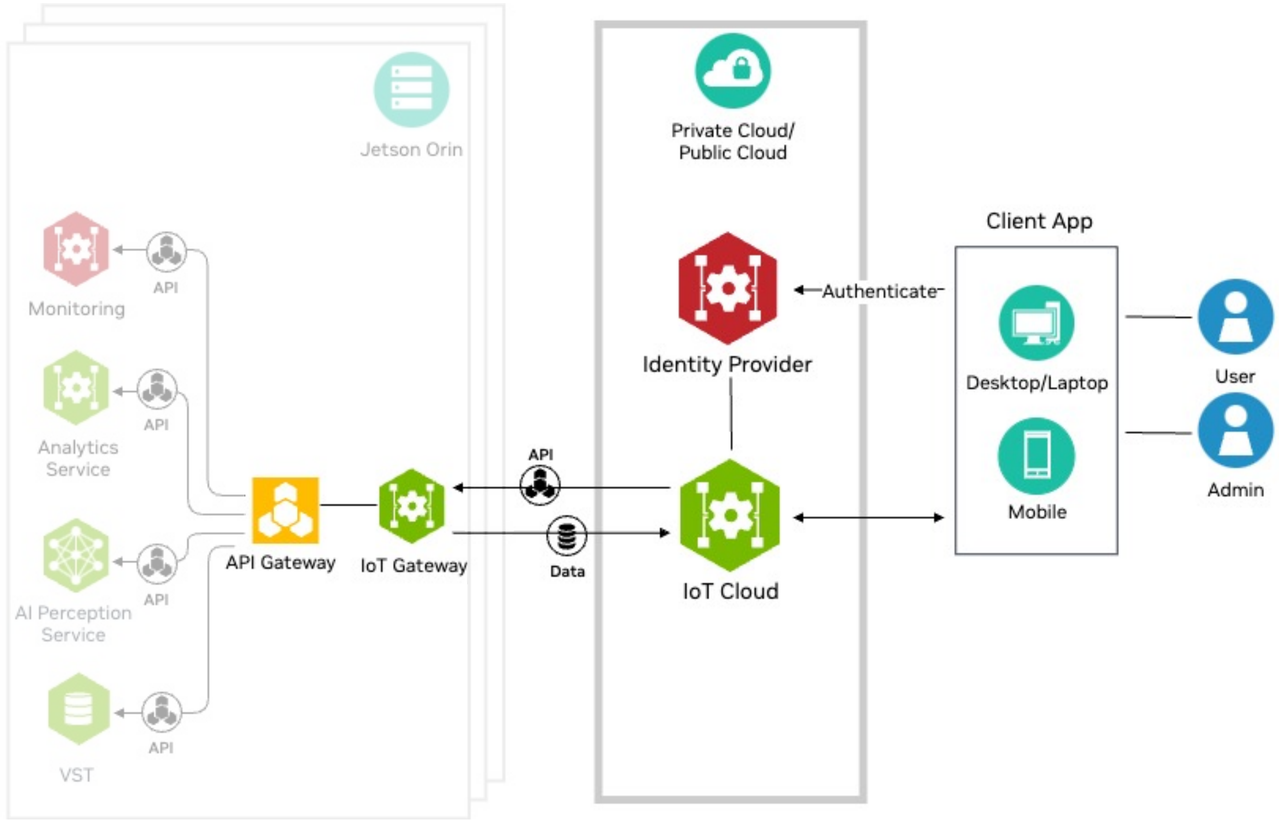
Remote, secure invocation of microservices APIs on devices from client applications

WHAT'S SUPPORTED

- Reference cloud implementation with recipe to deploy on any Cloud
- Setup flows: device securely connecting to cloud through OTP; user getting access authorization through claim code
- "Always on" encrypted, bi-directional communication link between Edge & Cloud
- Public Proxy endpoint in the cloud using which clients can invoke device APIs
- Authentication, authorization, user mgmt., device claim in the cloud

HOW TO INTEGRATE INTO YOUR CODE

REST APIs



Application Deployment Using Docker Compose

List all the microservice containers along with config, storage, startup ordering and deploy!

```
vst:
   image: nvcr.io/e7ep4mig3lne/release/vst:v1.2.37_aarch64
    ……..
    container_name: vst
    entrypoint: sh -c '/root/vst_release/launch_vst --debug-level 3 2>&1
    volumes:
      - ./config/vst/vst_config.json:/root/vst_release/configs/vst_config.json
      - ./config/vst/vst_storage.json:/root/vst_release/configs/vst_storage.json
    depends_on:
       moj-init-vst:
          condition: service_completed_successfully
    deploy:
      resources:
        limits:
          memory: 5600M
      restart_policy:
        condition: always
emdx-analytics-01:
   image: nvcr.io/e7ep4mig3lne/release/emdx-analytics:mmj_v1
   environment:
          CONFIG_LOCATION: "/config"
          PORT: 6001
          INSTANCE_ID: emdx-analytics-01
             ……..
```

Container name

config

Init container for startup ordering

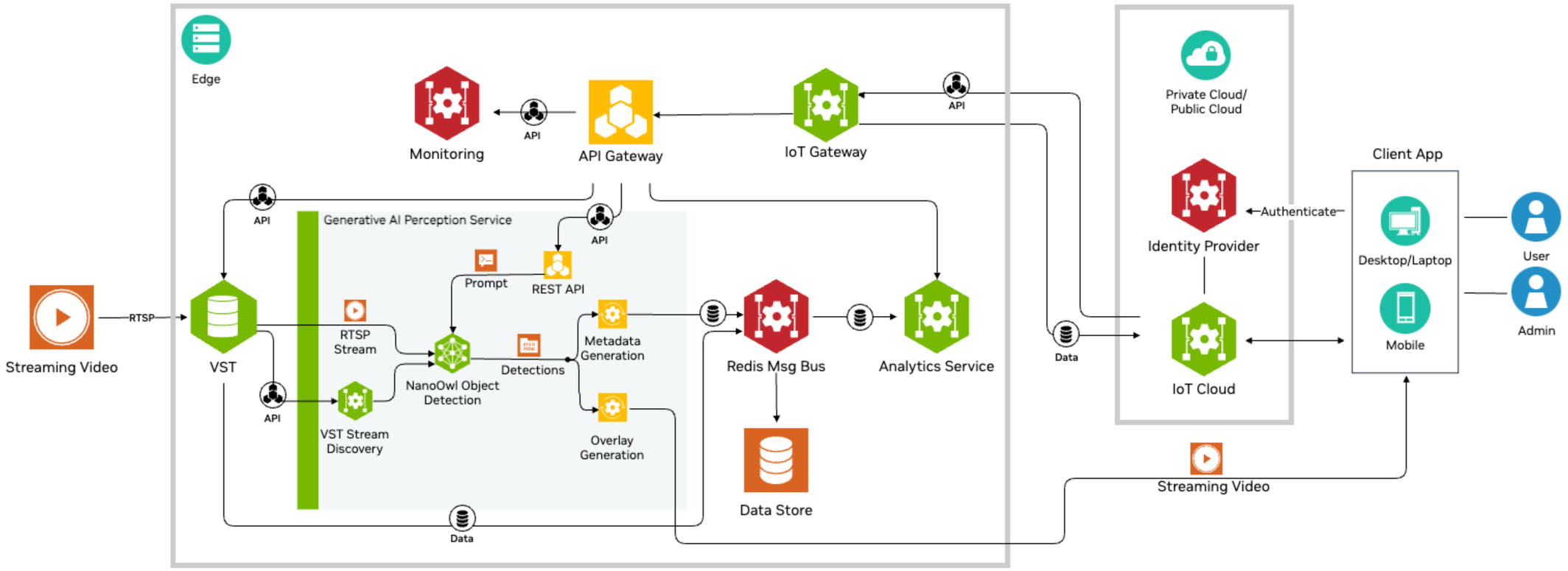
Storage mounts

Failure restart

Environment setting

Generative AI – Reference Application

Integrating Generative AI Services



VST v1.2.29

Admin

Dashboard
Camera Management
Camera Configuration
Record Settings
Camera Details
Live Streams
Recorded Streams
Video Wall
Settings
Debug

What is vst ?

Live streams

Live streams

Stream multiple live streams

Select cameras

owl Type to search

owl is playing

GPU Status

GPU Utilization

GPU Clock Frequency

GPU Temperature

62.8 °C

GPU temperature

CPU core usage (in%age)

Video HW ENC/DEC Engine Frequency

EMC Bandwidth Utilization

EMC Frequency

rosie@tegra-ubuntu:/data/rosie/streaming_owl$

Summary

- Announced availability of Microservices for Jetson ORIN
- Cloud-native Microservice Architecture
- API-driven platform for building AI apps at the Edge
- Accelerate development and reduce time to market

Get Started

https://developer.nvidia.com/metropolis-microservices/jetson-get-started

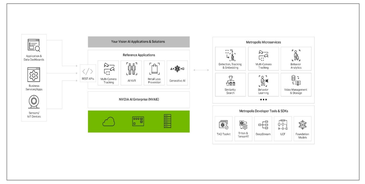
Build complex applications quickly with APIs

Maximize your system resources

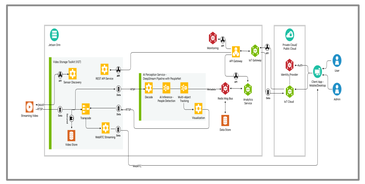
Integrate your own custom microservices

Microservices for Jetson - Resources

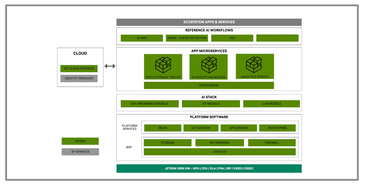
Web



Product Page



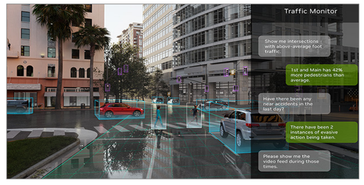
Get Started Page



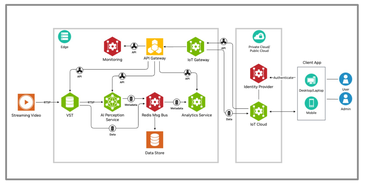
SW Download

Blogs

Dev News

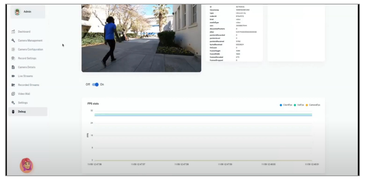


Gen AI

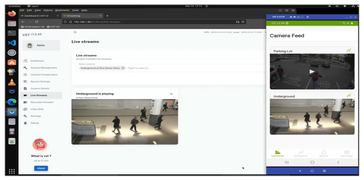


API Workflow

Videos



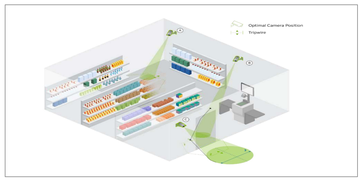
VST Demo



AI NVR Demo

Gen AI Demo

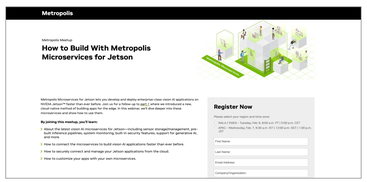
Assets



Whitepaper

Solution One-Pager

Webinar – Part 1

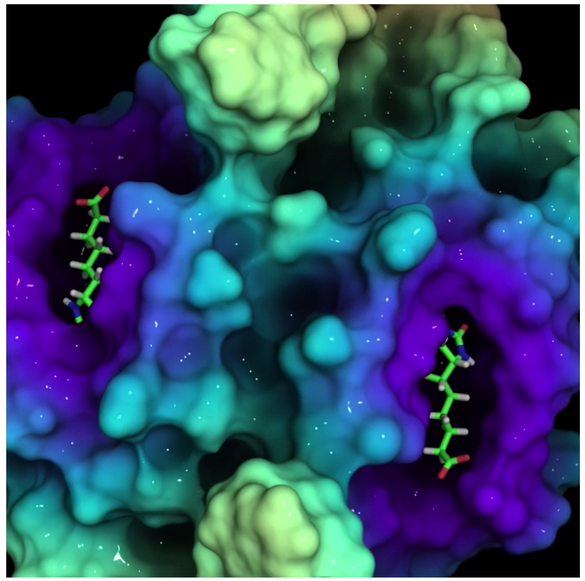


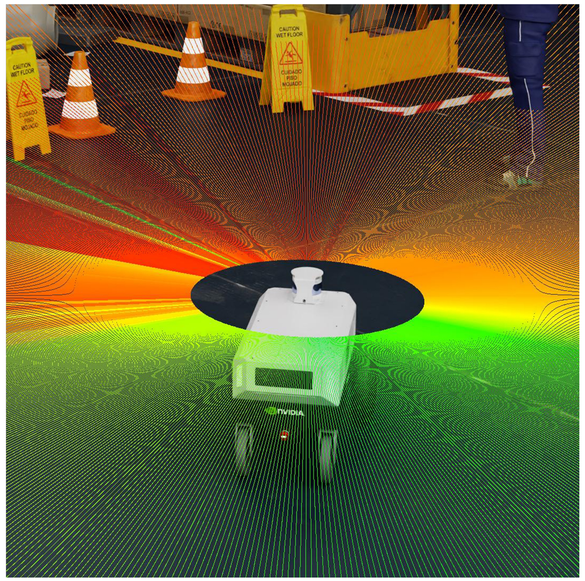
Webinar – Part 2

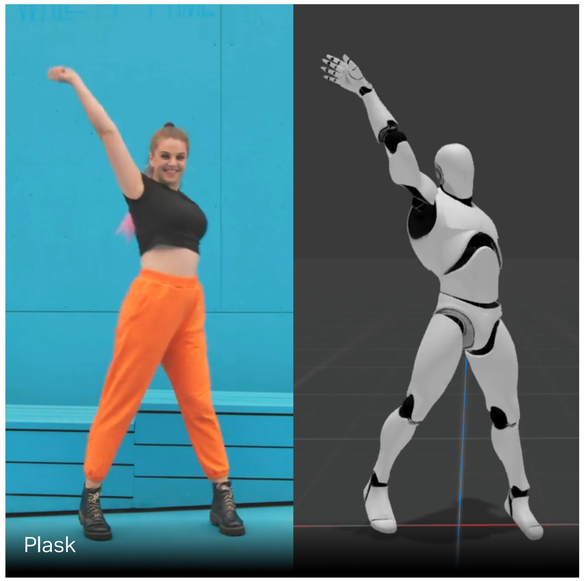
NVIDIA GTC

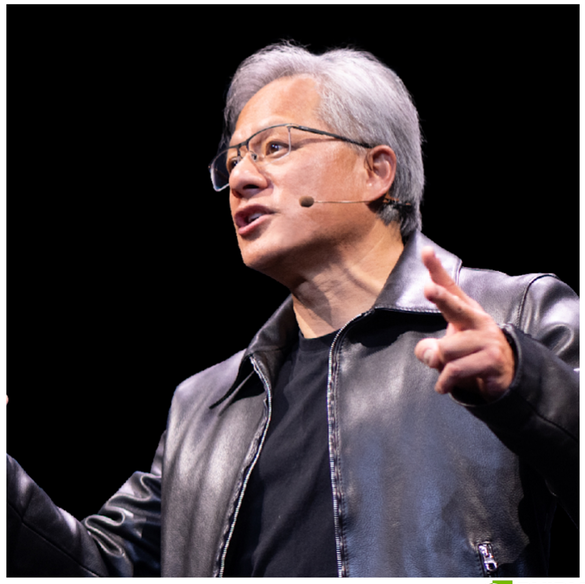
The Conference for the Era of AI

Workshops March 17–21
Conference and Expo March 18–21









NVIDIA

Edge Computing at GTC 2024

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

Register with this link for 25% off your conference pass
www.nvidia.com/gtc/?ncid=GTC-NV481LYM

Empowering Product Creators to Harness Edge AI and Vision

The Edge AI and Vision Alliance (www.edge-ai-vision.com) is a partnership of 100+ leading edge AI and vision technology and services suppliers, and solutions providers

Mission: To inspire and empower engineers to design products that perceive and understand.

The Alliance provides low-cost, high-quality technical educational resources for product developers

Register for updates at www.edge-ai-vision.com

The Alliance enables edge AI and vision technology providers to grow their businesses through leads, partnerships, and insights

For membership, email us: membership@edge-ai-vision.com

Join us at the Embedded Vision Summit
May 21-23, 2024—Santa Clara, California

The only industry event focused on practical techniques and technologies for system and application creators

- *"Awesome! I was very inspired!"*
- *"Fantastic. Learned a lot and met great people."*
- *"Wonderful speakers and informative exhibits!"*

Embedded Vision Summit 2024 highlights:

- **Inspiring keynotes** by leading innovators
- High-quality, practical **technical, business and product talks**
- Exciting **demos**, **tutorials** and **expert bars** of the latest applications and technologies

Visit www.EmbeddedVisionSummit.com to learn more and register