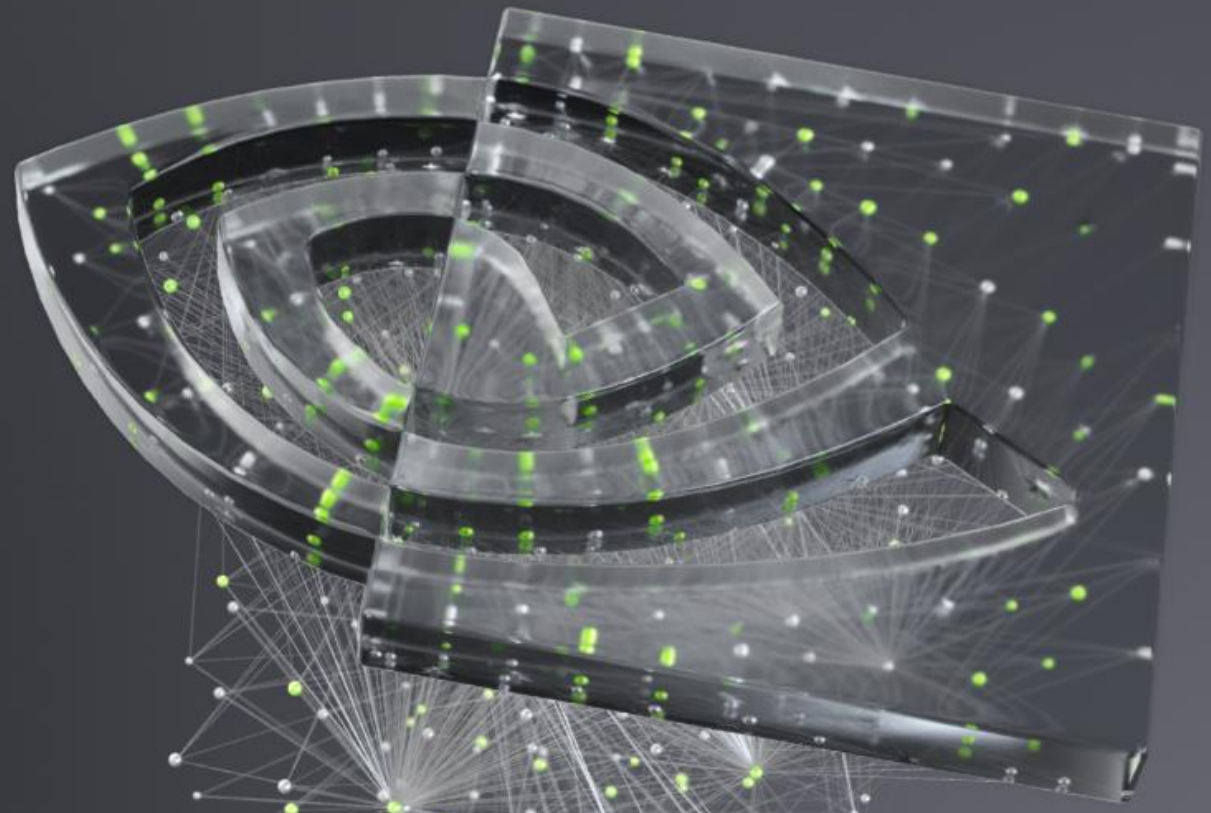


NVIDIA METROPOLIS FOR IVA AS A SERVICE

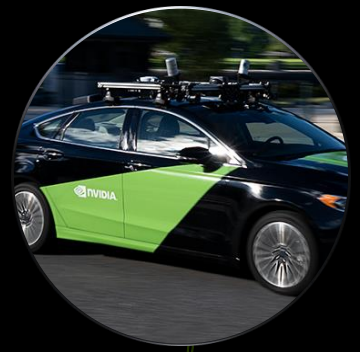
Software Stack | Libraries | SDKs | Pre-Trained Models | Sample Applications | GPU Varieties

Ettikan Kandasamy Karuppiah (Ph.D), Chief Technologist, Asia Pacific South Region.

13rd October 2020

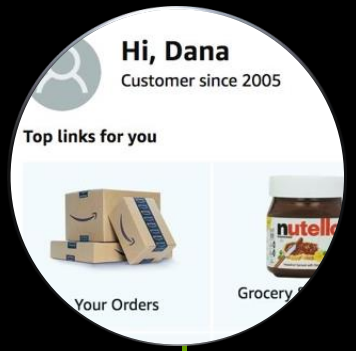
THE MOST POWERFUL TECHNOLOGY FORCE OF OUR TIME

AI ROBOTICS



AI EDGE

AI CLOUD



AI COMPUTING

AI AT THE EDGE IS A LARGE OPPORTUNITY

IOT

IOT devices projected to grow to >150B by 2025, >1T by 2035

5G

5G will deliver 1000X better bandwidth and 10X lower latency than 4G

AI

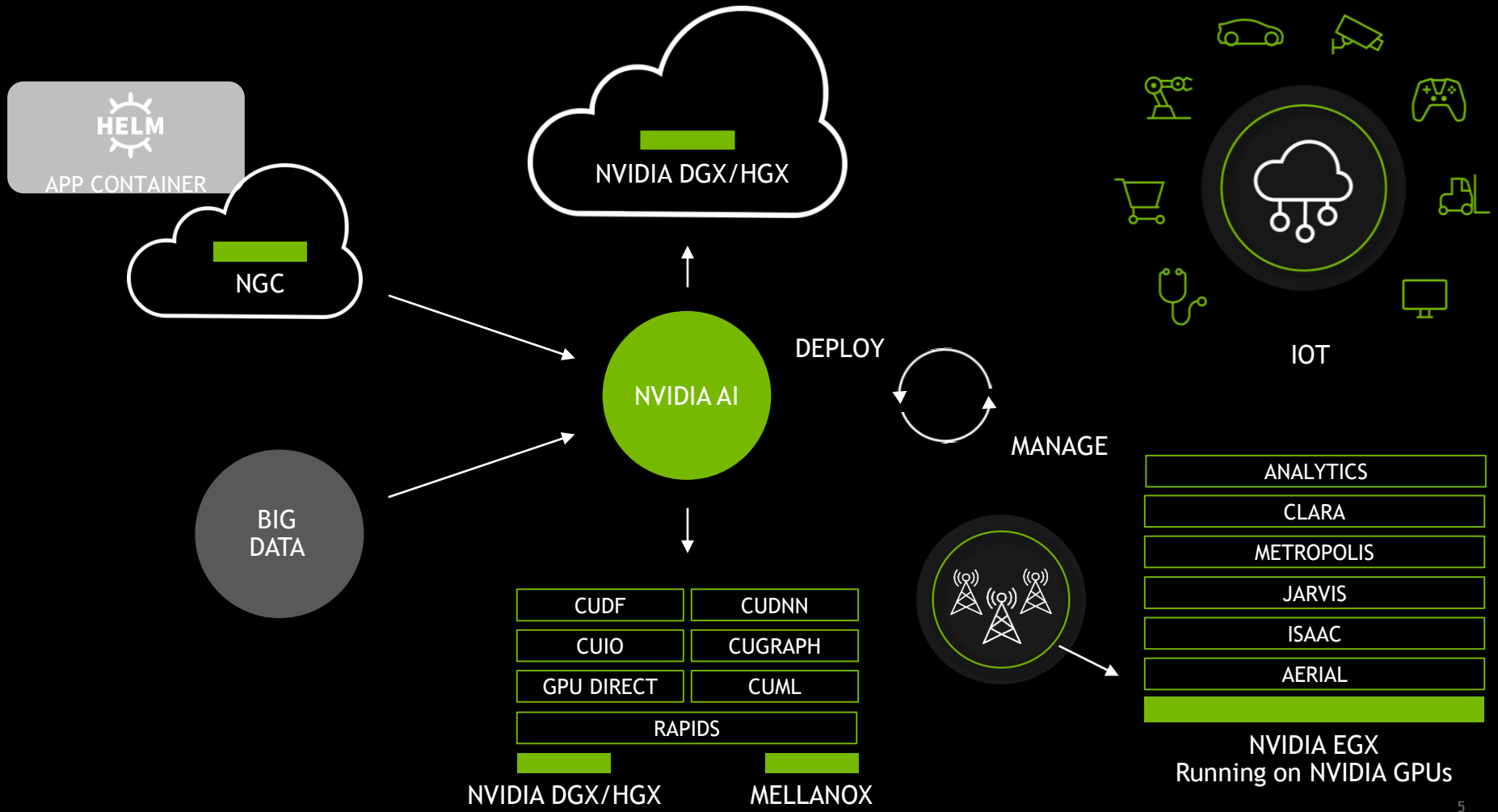
By 2025, AI at the edge has a potential total economic impact of up to $11T/year

INTELLIGENT VIDEO ANALYTICS

- Access Control
- Public Safety
- Critical Infrastructure
- Traffic Engineering
- Retail Analytics
- Parking Management
- Logistics
- Public Transit

$2T INDUSTRY — Increase operational efficiency and safety across many industries using AI

COMPUTING FROM EDGE TO CLOUD



NVIDIA METROPOLIS

NGC

NVIDIA PARTNERS

METROPOLIS SOFTWARE STACK

DEEPSTREAM | TRT, TRITON | TLT | VIDEOCODEC

CUDA-X

EGX SOFTWARE STACK

Kubernetes | Networking | Storage | Security

NVIDIA EGX HARDWARE

T4 | JETSON

https://www.nvidia.com/en-us/autonomous-machines/intelligent-video-analytics-platform/

DEEPSTREAM SOFTWARE STACK

Applications and Services

Python | C/C++

DEEPSTREAM SDK

Hardware Accelerated Plugins | Bi-directional IoT Messaging | OTA model update | Reference Applications & Helm Charts

CUDA-X

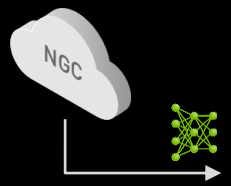
CUDA | TensorRT | Triton Inference Server† | Multimedia

NVIDIA COMPUTING PLATFORM - EDGE TO CLOUD

NVIDIA Containers RT | JETSON | TESLA | EGX | Kubernetes ON GPUs

† - Formerly TensorRT Inference Server

NVIDIA TRANSFER LEARNING TOOLKIT (TLT)



New Classes | Model Pruning | Scene Adaptation

TRANSFER LEARNING TOOLKIT

Data Preparation & Augmentation | Train | Prune

TensorFlow / Keras

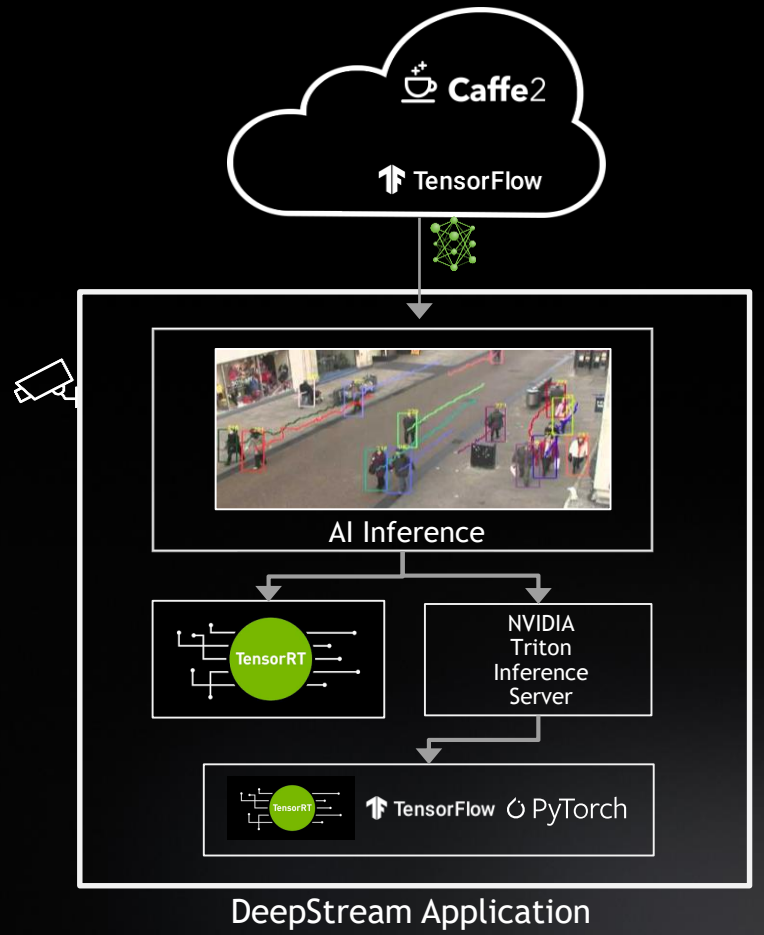
CUDA-X

NVIDIA Containers RT | CUDA | cuDNN | TensorRT

NVIDIA COMPUTING PLATFORM

EGX | DGX | CLOUD | WORKSTATIONS

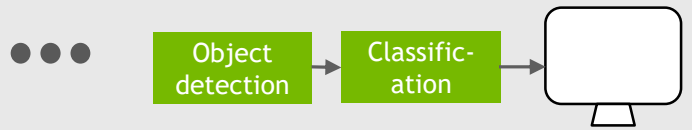
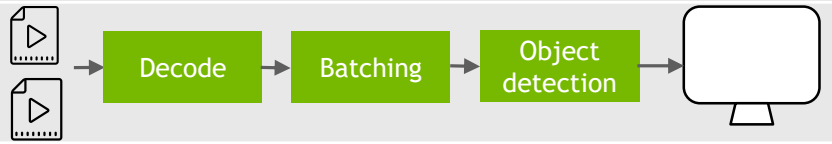
DEEPSTREAM WITH TRITON INFERENCE SERVER



	TensorRT	Triton Inference Server
Pros	Highest Throughput	Highest flexibility
Cons	Custom layers require writing plugins	Less performant than a TensorRT solution

GETTING STARTED APPLICATIONS

Available in C and Python

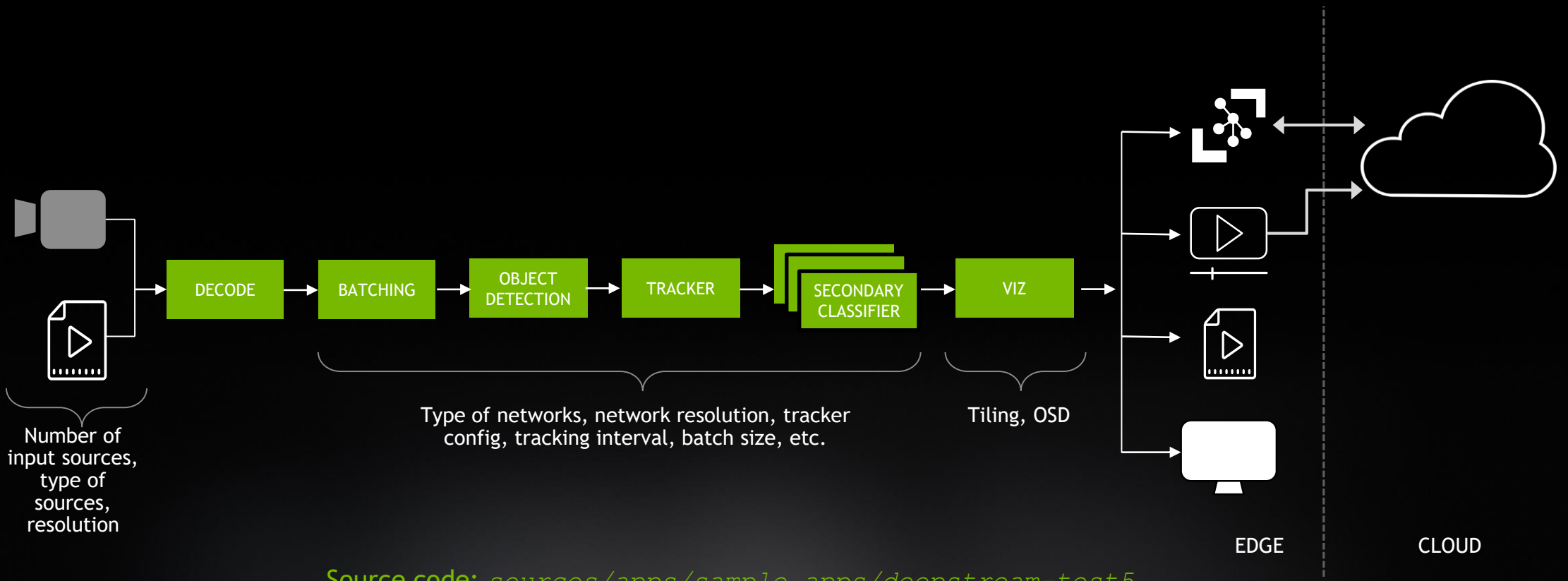
Name	Function	
deepstream-test1	DeepStream Hello world. Single video from file to on screen display with bounding box	
deepstream-test2	Builds on test1 and adds secondary object classification on detected objects	
deepstream-test3	Builds on test1 and adds multiple video inputs	
deepstream-test4	Builds on test1 and adds connections to IoT services thru the nvmsgbroker plugin	

Native C apps: *sources/apps/sample_apps/*

Python apps: https://github.com/NVIDIA-AI-IOT/deepstream_python_apps

END-TO-END DEEPSTREAM APP

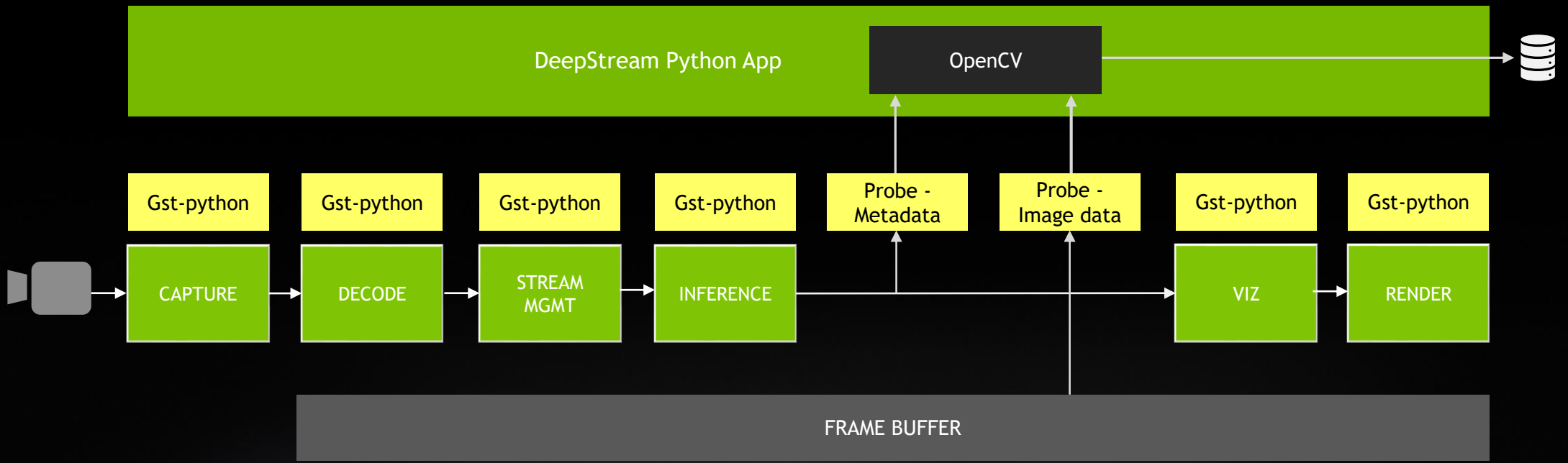
DeepStream-test5



Source code: *sources/apps/sample_apps/deepstream-test5*

Python app coming soon

IMAGE DATA ACCESS IN PYTHON



https://github.com/NVIDIA-AI-IOT/deepstream_python_apps/tree/master/apps/deepstream-imagedata-multistream

NVIDIA GPU

ONE ARCHITECTURE

NVIDIA GPU CLOUD

GAMING | HPC | TRANSPORTATION | HEALTHCARE

PRO VIZ | AI | ROBOTICS | SMART CITY

NVIDIA APPLICATION FRAMEWORKS

CUDA

RTX

DGX

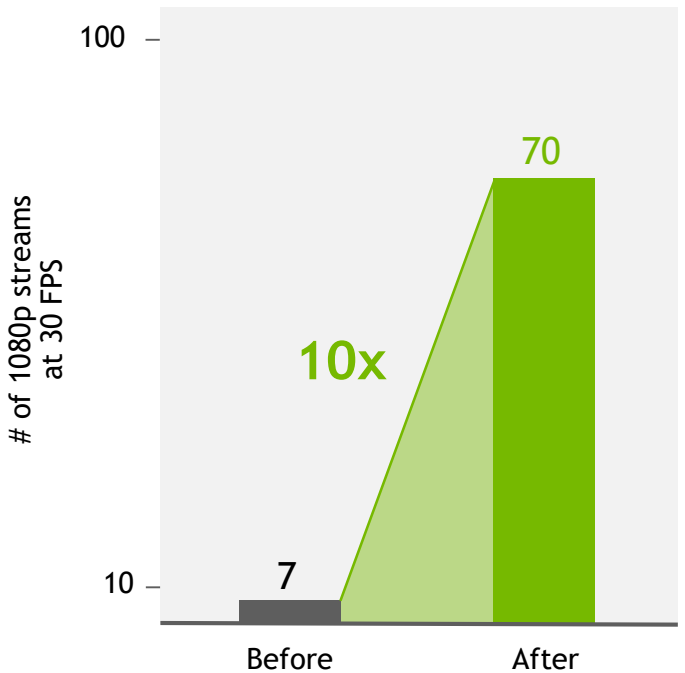
HGX

EGX

AGX

Common Architecture | Common Tools | Common Infrastructure | 1.2m Developers

IMPROVING PERFORMANCE PER DOLLAR



	BEFORE	AFTER
Transfer Learning Toolkit	✗	✓
DeepStream SDK	✗	✓
TensorRT	✗	✓
Video Codec SDK	✗	✓
Number of Channels @30 FPS	7	70

Running 1 detection and 3 classification networks simultaneously on 70 1080p streams @30FPS using NVIDIA T4

Metropolis software and tools to unlock greater performance

LOWERING TOTAL COST OF OWNERSHIP (TCO)

SOFTWARE PARTNER 1 FACIAL RECOGNITION APP	BEFORE CERTIFICATION	AFTER CERTIFICATION
Streams per server	12	47
Number of servers	25	6
Hardware cost	500k	120k

Assume $20,000 per server (four x T4 GPUs)

$380k Savings!

SOFTWARE PARTNER 2 SEARCH IN VIDEO APP	BEFORE CERTIFICATION	AFTER CERTIFICATION
Streams per server	56	90
Number of servers	6	3
Hardware cost	120k	60k

Assume $20,000 per server (four x T4 GPUs)

$60k Savings!

Maximize performance, number of streams and save on TCO

SOFTWARE PARTNER TESTIMONIALS

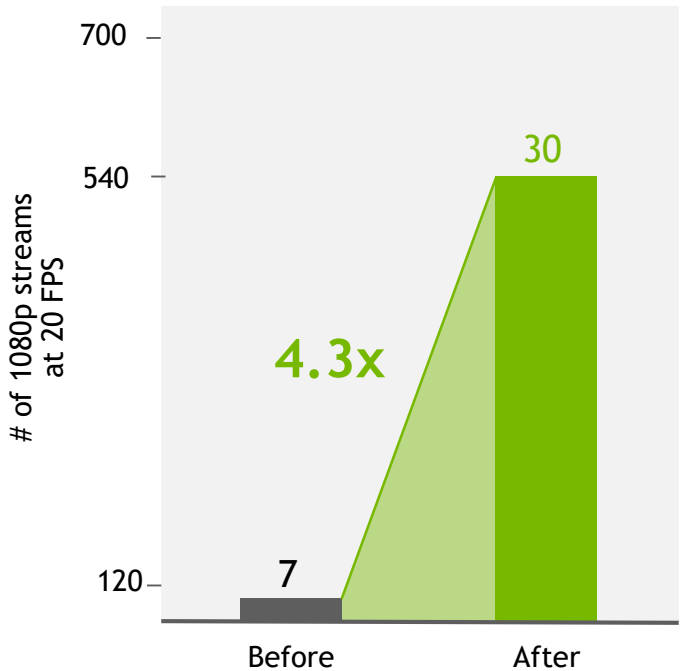
“ Metropolis partner program has been critical for us to make a business impact, reaching and driving new customers and opportunities together. Since completing Metropolis Certification and publishing on NGC, we have noticed a 40% increase in the number of inquiries about our products.”

“ Having IronYun AI NVR available on NGC has allowed our customers and partners greater flexibility in deployment. With the appropriate hardware in the customer's existing infrastructure, we can decrease the delivery time for each PO from 2-3 weeks to merely 30 minutes, even for first-time users, which in turn allows us to focus on product development and customer support with even higher efficiency.”

“ Using NVIDIA’s prowess in accelerated computing we were able to optimize our solution to increase the total number of video streams running on a Dell R740 with 6xT4s by over 250%. Through the Metropolis platform customers across the globe can now easily discover and quickly deploy the SAFR solution to solve critical operational and security challenges.”

TELCO DEVELOPER EXPERIENCE

VNPT-IT/Vietnam Telco Developer Observation



	BEFORE	AFTER
Transfer Learning Toolkit	✗	✓
DeepStream SDK	✗	✓
TensorRT	✗	✓
Video Codec SDK	✗	✓
Number of Channels @20 FPS	7	30

Running 1 detection and 1 recognition networks simultaneously on 30x1080p streams @20FPS using NVIDIA V100.
The plan is to move the inferencing to T4 which has better dollar cost/watt/fps performance

Test System Configuration:
CPU: Intel(R) Xeon(R) Gold 6252 CPU @ 2.10GHz, 8 cores, 8 threads
RAM: 32GB
GPU: V100 16GB

Software Setup:
CUDA: 11
CUDNN: 8.0.1 &
NVIDIA Driver: 440.87
DeepStream NVIDIA Docker: 20.07
DeepStream: 5.0.1
TensorRT: 7.1.2

TELCO DEVELOPER EXPERIENCE

VNPT-IT/Vietnam Telco Developer Observation

- 5x CPU video processing workloads offloaded to GPU, freeing CPU for other application usage
- 2x better GPU/CPU RAM memory utilization
- 2x Yolov3 performance improvement with TLT Model Backbone usage + DeepStream
- 3x Yolov5 performance improvement using TLT + DeepStream
- 2x model performance increase with TensorRT despite similar accuracy maintained
  - ignorable accuracy diff between 0.01- 0.04

INCREASE REACH THROUGH SYSTEM PARTNERS

acer
ADLINK Leading EDGE COMPUTING
ADVANTECH
ASRock Rack
ASUS
Atos
AVerMedia
cisco
DELL EMC
FUJITSU
Hewlett Packard Enterprise
inspur
Inventec
Lenovo
MiTAC
QCT
SUPERMICRO

DELL

“ The Metropolis Certification Program provides DELL with an impressive catalog of AI applications that have been tested on our hardware, enabling faster time-to-revenue and repeatable deployments.”

NVIDIA METROPOLIS ISV PARTNERS

ETHICAL AI

NVIDIA's platforms and application frameworks enable developers to build a wide array of AI applications. Consider potential algorithmic bias when choosing or creating the models being deployed. Work with the model's developer to ensure that it meets the requirements for the relevant industry and use case; that the necessary instruction and documentation are provided to understand error rates, confidence intervals, and results; and that the model is being used under the conditions and in the manner intended.

DEVELOPER RESOURCES

IMPORTANT WEBLINKS:

- Transfer Learning Toolkit | DeepStream SDK | Pre-trained models on NGC
- Documentation | Developer Forum

FREE DLI ONLINE COURSE:

- Getting Started with DeepStream for Video Analytics on Jetson Nano

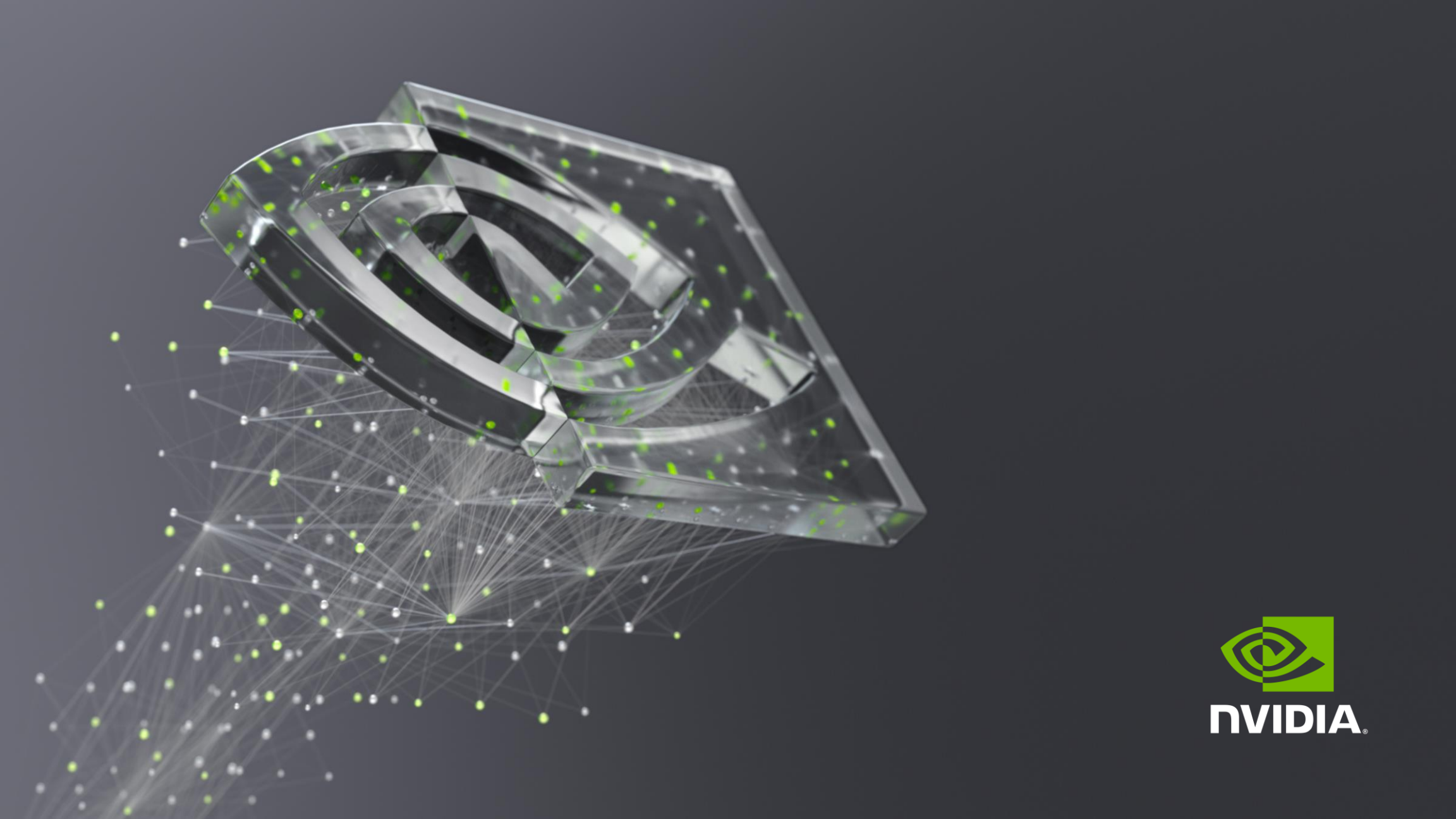
LATEST TUTORIALS:

- Learn about newest features in DeepStream 5.0
- Training with Custom Pretrained Models Using the NVIDIA TLT
- Building a real-time redaction App using DeepStream
- Medium - How to build accurate models using TLT

GITHUB:

- DeepStream reference apps
- Running TLT models on DeepStream example
- DeepStream Python apps

Enroll in the NVIDIA Developer Program to get the latest updates


NVIDIA