DEEPSTREAM: AN SDK TO IMPROVE VIDEO ANALYTICS

Jason Tichy – NVIDIA Senior Solutions Architect

NVIDIA
A LEARNING MACHINE

NVIDIA has continuously reinvented itself over two decades.

Our invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics, and revolutionized parallel computing. More recently, GPU computing ignited the era of AI.

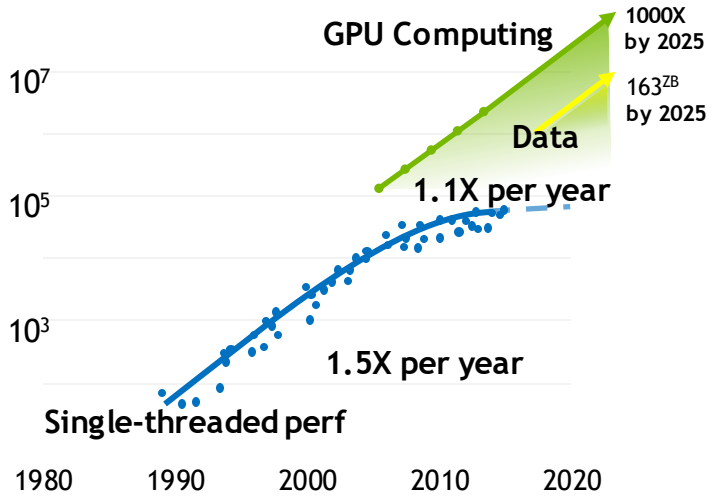
NVIDIA is a “learning machine” that constantly evolves by adapting to new opportunities that are hard to solve, that only we can tackle, and that matter to the world.

TWO FORCES SHAPING COMPUTING

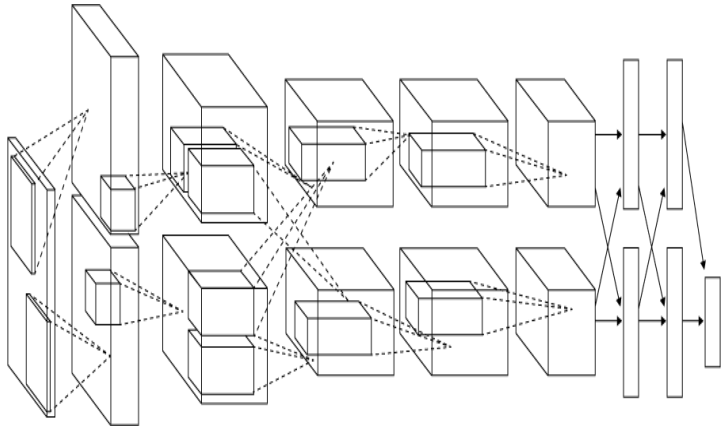
CPU scaling is slowing while the demand for computing power surges ahead.

AI can solve grand challenges that have been beyond human reach, but it must be fueled by massive compute power.

Accelerated computing is the path forward beyond Moore's law, delivering 1000X computing performance every 10 years.



40 YEARS OF CPU TREND DATA



ALEXNET: THE SPARK OF THE MODERN AI ERA

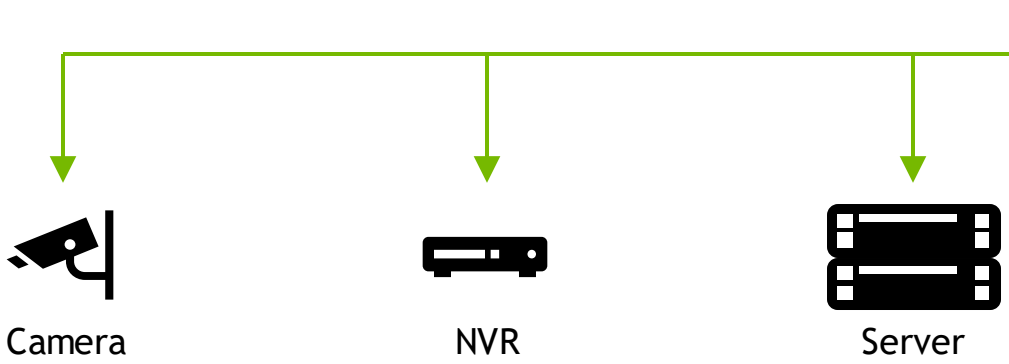
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

Data Growth Source: *Mapping the Future of Silicon for AI* - September 2017

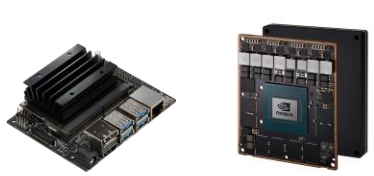
AI EDGE TO CLOUD

EDGE AND ON-PREMISES
Inference

CLOUD
Training and Inference



TENSORRT • DEEPSTREAM • JETPACK

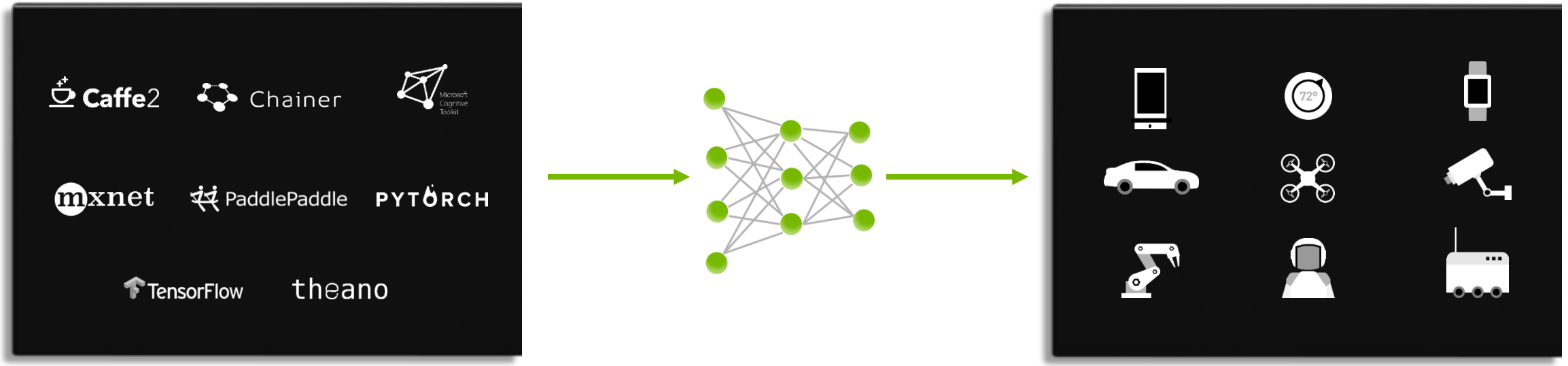


JETSON

TESLA

DGX

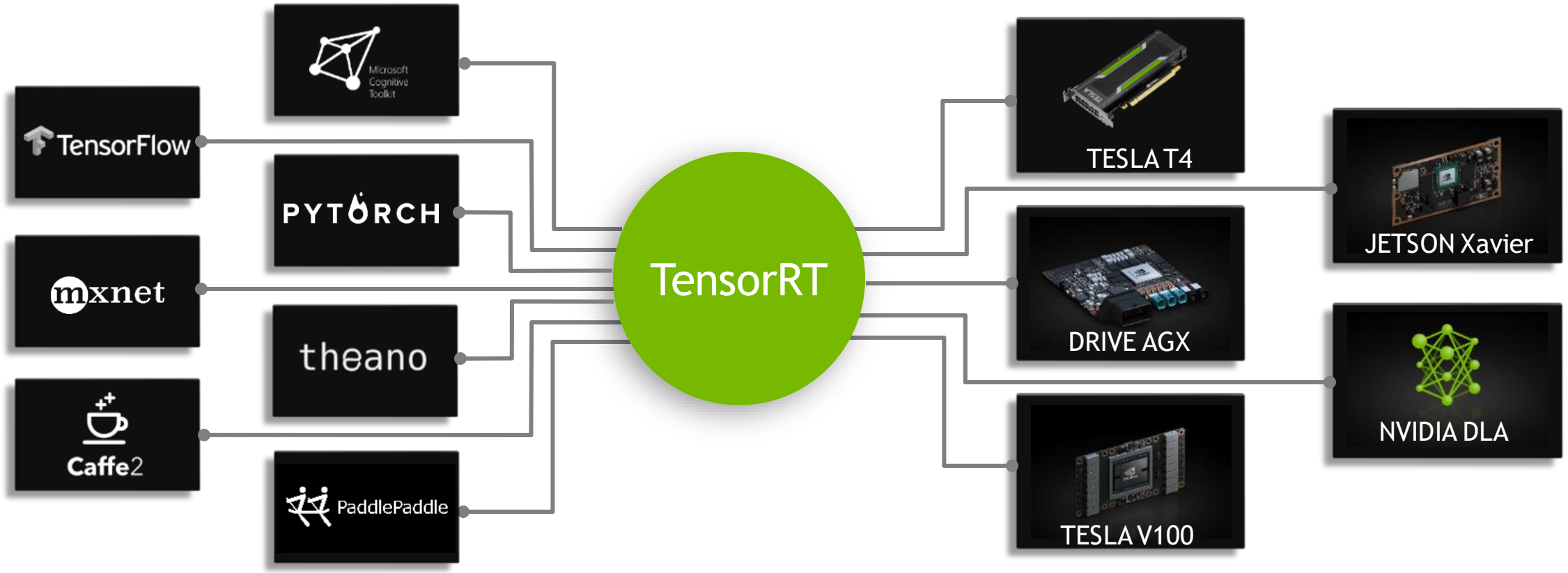
AI INFERENCE NEEDS TO RUN EVERYWHERE



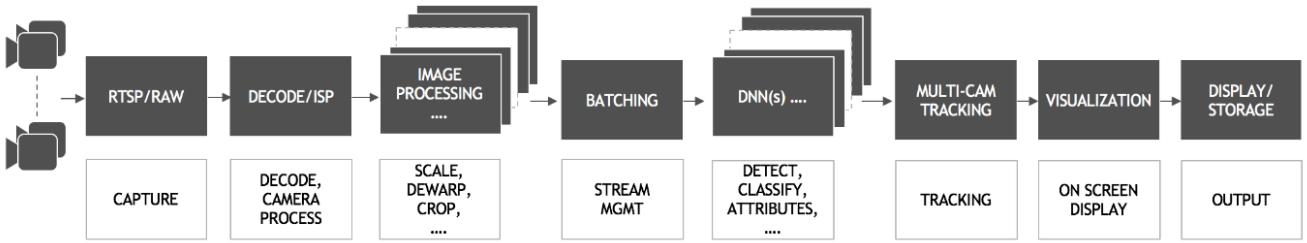
Training DNN Model Inference

NVIDIA TENSORRT

From Every Framework, Optimized For Each Target Platform

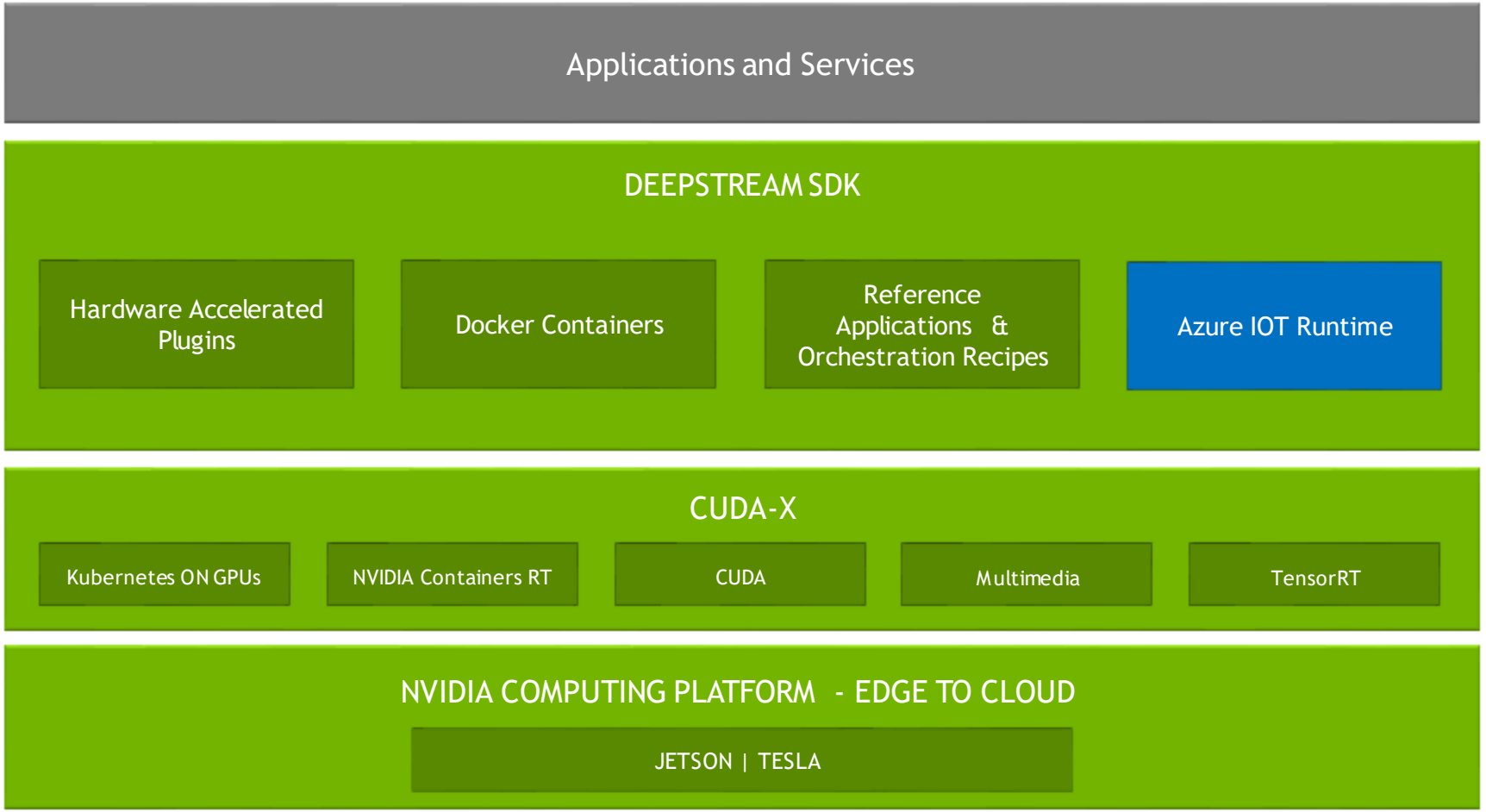


VIDEO ANALYTICS



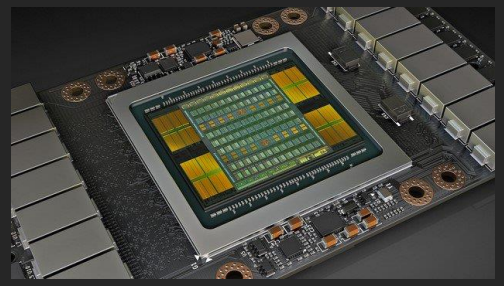
Typical application: 30+ TOPS

WHAT IS DEEPSTREAM?



WHAT'S NEW IN DEEPSTREAM 4.0

UNIFIED SDK , ALL PLATFORMS



Portability from Jetson Nano to T4

TURNKEY IoT INTEGRATION

Microsoft Azure IoT Hub*

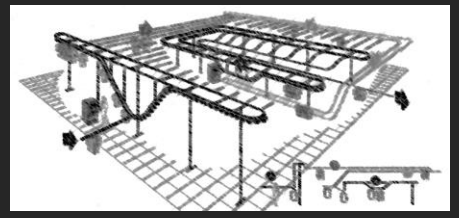
DOCKER CONTAINERS ON NGC

Easy to scale and maintain

MONOCHROME AND JPEG

Enabling More Applications

SUPPORT FOR IMAGE SEGMENTATION



Expanding What is Possible

PLUGIN SOURCES

Inference Decode Messaging

Greater control for your use case

*Containers on Azure Marketplace coming soon. Available on NGC now

NVIDIA DEEPSTREAM

Resources

Downloads, Documentation and Resources:

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

Tesla Docker Container

https://ngc.nvidia.com/catalog/containers/nvidia:deepstream

Jetson Docker Container

https://ngc.nvidia.com/catalog/containers/nvidia:deepstream-l4t

NVIDIA DEEPSTREAM

Performance Driven

Low latency and exceptional performance optimized for NVIDIA GPUs for real-time edge analytics.

Cloud Integration

Pushbutton IoT solution integration to build applications and services with Cloud Service Providers.

Deploy with Ease

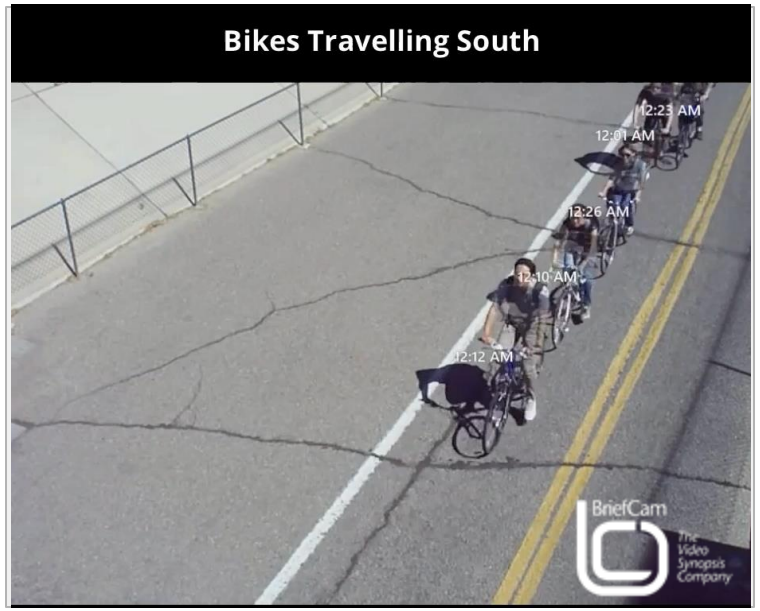
Fast, flexible, and reliable containerized deployment and support for NVIDIA Tesla and Jetson platforms using NGC.

NVIDIA Products	H.264	H.265
Jetson Nano	8	8
Jetson TX1	8	8
Jetson TX2	14	14
Jetson AGX Xavier	32	49
T4	35	68

Data measured using deepstream-app from DeepStream SDK 4.0

FEDERAL USE CASES

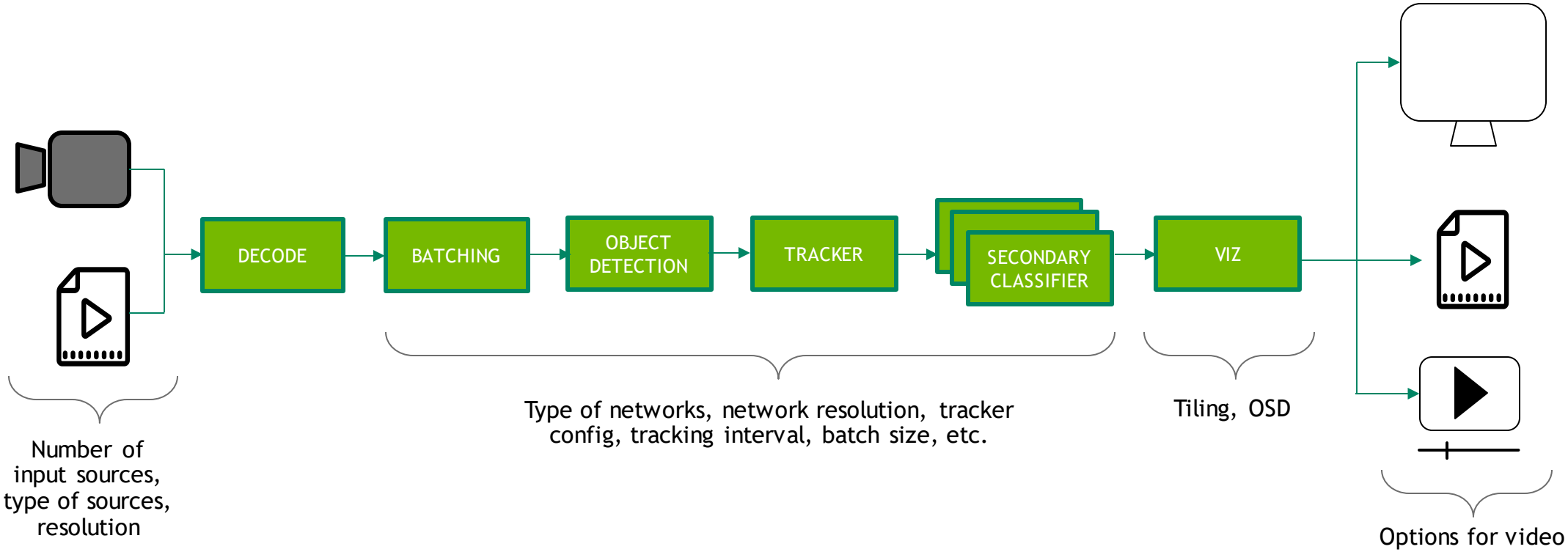
Humanitarian Aid and Disaster Relief



Physical Security and Safety

Logistics and Monitoring

END-TO-END DEEPSTREAM APP



Source code: *sources/apps/sample_apps/deepstream-app*
Config files: *samples/configs/deepstream-app/*

DEEPSTREAM APPLICATION

Application Group

```
[application]
enable-perf-measurement=1
perf-measurement-interval-sec=5
## application group also interfaces with a KITTI metadata format
## folder or output stream, we will not be using this
#gie-kitti-output-dir=streamscl
#kitti-track-output-dir=/home/ubuntu/kitti_data_tracker/
```

DEEPSTREAM APPLICATION

Tiled Display Group (Optional)

```
[tiled-display]
## tiles display controls the interactive gui application if it
#  is needed.
enable=1
rows=2
columns=2
width=1280  # total window value in pixels
height=720  # total window value in pixels
gpu-id=0    # dictate which gpu will be controling this feature
#(0): nvbuf-mem-default - Default memory allocated, specific to particular platform
#(1): nvbuf-mem-cuda-pinned - Allocate Pinned/Host cuda memory, applicable for Tesla
#(2): nvbuf-mem-cuda-device - Allocate Device cuda memory, applicable for Tesla
#(3): nvbuf-mem-cuda-unified - Allocate Unified cuda memory, applicable for Tesla
#(4): nvbuf-mem-surface-array - Allocate Surface Array memory, applicable for Jetson
nvbuf-memory-type=0
```

DEEPSTREAM APPLICATION

Source Group

```
[source0]
enable=1
#Type - 1=CameraV4L2 2=URI 3=MultiURI 4=RTSP 5=CameraCSI (Jetson)
type=3
uri=file://./videos/drones_0%d.mp4
num-sources=1gpu-id=0
# (0): memtype_device   - Memory type Device
# (1): memtype_pinned   - Memory type Host Pinned
# (2): memtype_unified  - Memory type Unified
cudadec-memtype=0
```

DEEPSTREAM APPLICATION

StreamMux Group

```
[streammux]
gpu-id=0
##Boolean property to inform muxer that sources are live
live-source=0
batch-size=4
##time out in usec, to wait after the first buffer is available
##to push the batch even if the complete batch is not formed
batched-push-timeout=40000
## Set muxer output width and height
width=1280
height=720
##Enable to maintain aspect ratio wrt source, and allow black borders, works
##along with width, height properties
enable-padding=0
nvbuf-memory-type=0
```

DEEPSTREAM APPLICATION

Primary-GIE Group (GPU Inference Engine)

```
# config-file property is mandatory for any gie section.
# Other properties are optional and if set will override the properties set in
# the infer config file.
[primary-gie]
enable=1
gpu-id=0
gie-unique-id=1
nvbuf-memory-type=0
config-file=infer_config.txt
```

DEEPSTREAM APPLICATION

Primary-GIE Config File (infer_config.txt linked in top application)

```
[property]
gpu-id=0
net-scale-factor=0.017352074
offsets=123.675;116.28;103.53
model-engine-file=./stanford_resnext_batch4.plan
labelfile-path=./stanford.names
batch-size=4
## 0=FP32, 1=INT8, 2=FP16 mode
network-mode=2
num-detected-classes=80
interval=2
gie-unique-id=1
parse-func=0
is-classifier=0
output-blob-names=boxes;scores;classes
parse-bbox-func-name=NvDsInferParseRetinaNet
custom-lib-path=./libnvdsparsebbox_retinanet.so
```

Tutorial Available at NVIDIA GITHUB
https://github.com/NVIDIA/retinanet-examples

DEEPSTREAM APPLICATION

Primary-GIE Config File (infer_config.txt linked in top application)

```
[class-attrs-all]
threshold=0.3
group-threshold=0
detected-min-w=4
detected-min-h=4
#detected-max-w=0
#detected-max-h=0
```

```
(Optional per-class configuration)
## Per class configuration
#[class-attrs-2]
#threshold=0.6
#eps=0.5
#group-threshold=3
#roi-top-offset=20
#roi-bottom-offset=10
#detected-min-w=40
#detected-min-h=40
#detected-max-w=400
#detected-max-h=800
```

DEEPSTREAM APPLICATION

Tracker Group

```
[tracker]
enable=1
tracker-width=1280
tracker-height=720
ll-lib-file=/opt/nvidia/deepstream/deepstream-4.0/lib/libnvds_mot_iou.so
#ll-config-file required for IOU only
#ll-config-file=iou_config.txt
gpu-id=0
enable-batch-process=1
```

Check the latest DeepStream source for more trackers

DEEPSTREAM APPLICATION

On Screen Display Group

```
[osd]
enable=1
gpu-id=0
border-width=1
text-size=12
text-color=1;1;1;1;
text-bg-color=0.3;0.3;0.3;1
font=Arial
show-clock=0
clock-x-offset=800
clock-y-offset=820
clock-text-size=12
clock-color=1;0;0;0
nvbuf-memory-type=0
```

DEEPSTREAM APPLICATION

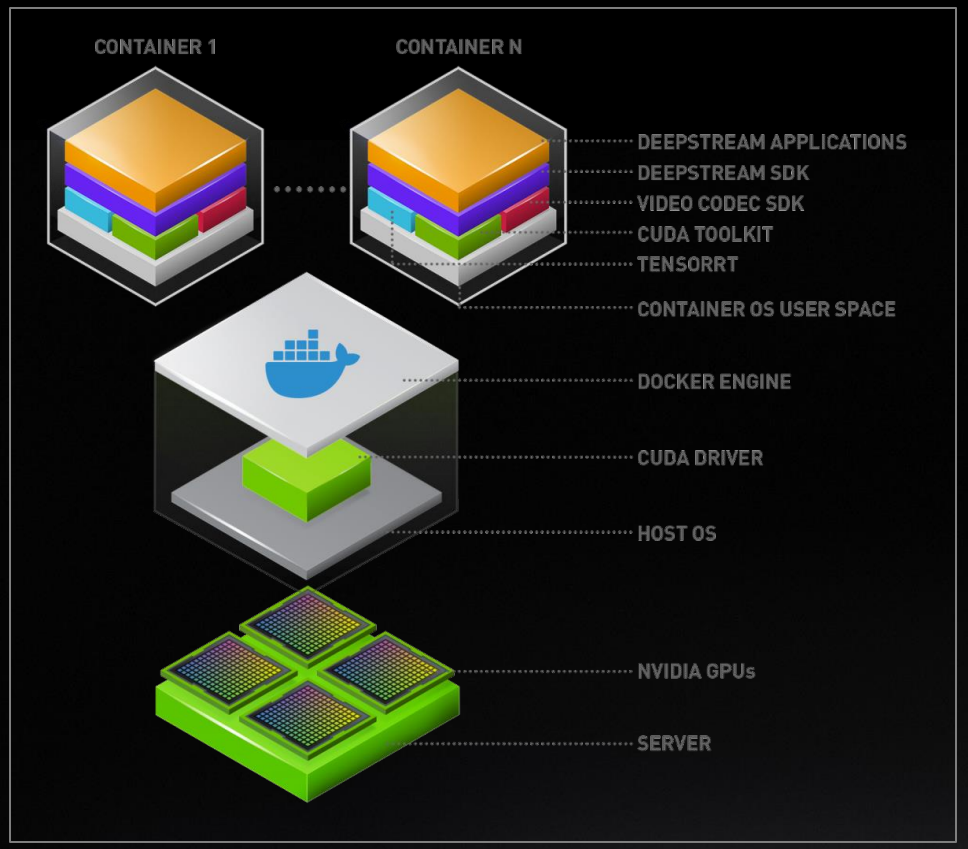
Sink Group

```
[sink0]
enable=1
#Type - 1=FakeSink 2=EglSink 3=File
type=2
sync=1
source-id=0
gpu-id=0
nvbuf-memory-type=0
```

```
[sink1]
enable=0
type=3
#1=mp4 2=mkv
container=1
#1=h264 2=h265
codec=1
sync=0
#iframeinterval=10
bitrate=2000000
output-file=out.mp4
source-id=0
```

```
[sink2]
enable=0
#Type - 4=RTSPStreaming
type=4
#1=h264 2=h265
codec=1
sync=0
bitrate=4000000
# set below properties in case of
RTSPStreaming
rtsp-port=8554
udp-port=5400
```

DEEPSTREAM DOCKER CONTAINER FOR GPU



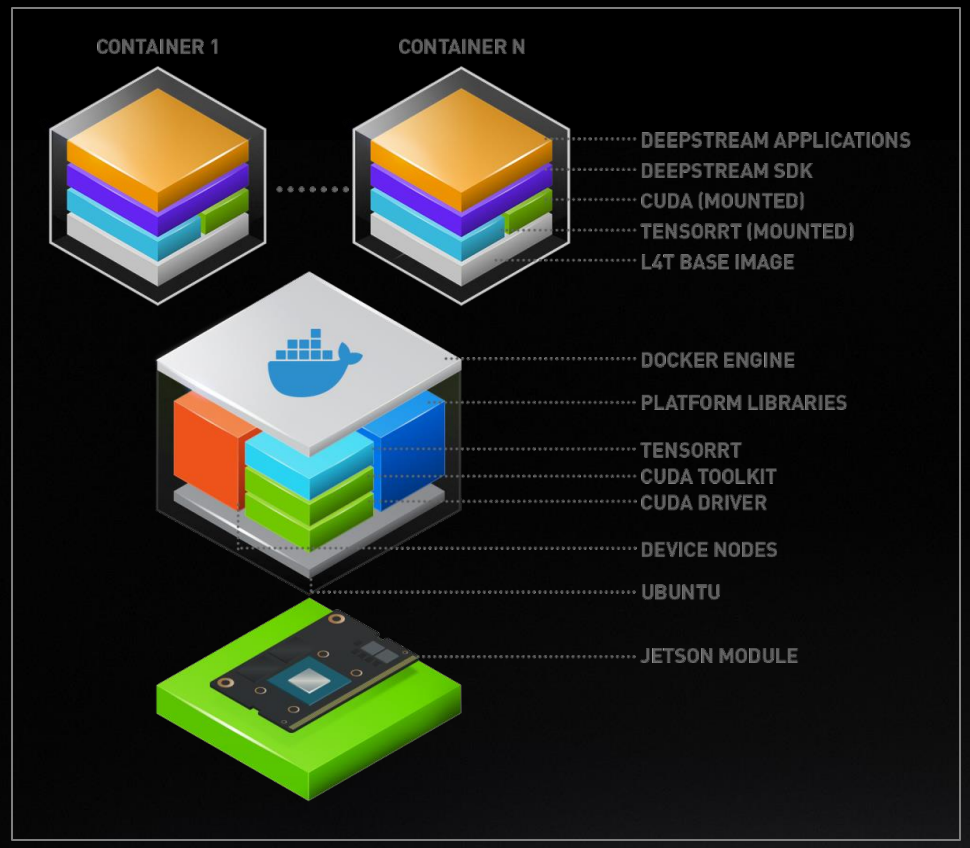
Development and deployment containers

NGC Container:
https://ngc.nvidia.com/catalog/containers/nvidia:deepstream

NVIDIA Container Runtime:
https://github.com/NVIDIA/nvidia-docker

NVIDIA NGC

DEEPSTREAM DOCKER CONTAINER FOR JETSON



Deployment containers

CUDA and TensorRT mounted from Host

NGC Container:
https://ngc.nvidia.com/catalog/containers/nvidia:deepstream-l4t

NVIDIA Container Runtime on Jetson:
https://github.com/NVIDIA/nvidia-docker/wiki/NVIDIA-Container-Runtime-on-Jetson

NVIDIA NGC

NVIDIA

GTC GPU TECHNOLOGY CONFERENCE

November 4 - 6, 2019 | Washington, D.C.

CONNECT

Connect with experts from NVIDIA, GE Healthcare, NSF Carnegie Mellon, Google, and other leading organizations

LEARN

Gain insight and valuable hands-on training through over 100 sessions

DISCOVER

See how GPU technologies are creating amazing breakthroughs in important fields such as deep learning

INNOVATE

Explore disruptive innovations that can transform your work

Join us at GTC DC | Use VIP code NVJTICHY for 25% off | Govt. attends free

Don't miss the premier AI conference.

nvidia.com/en-us/gtc-dc/


NVIDIA