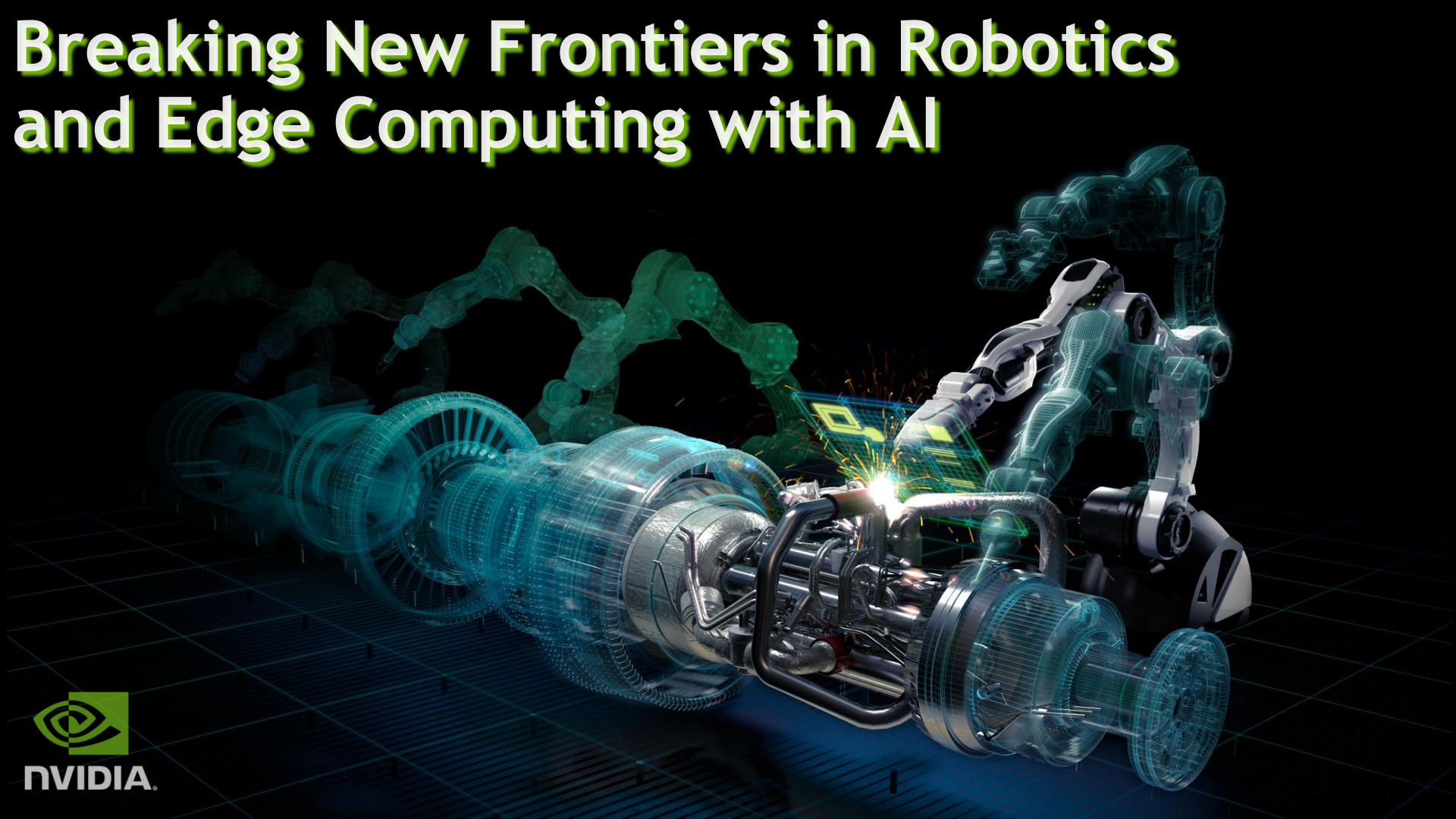
Breaking New Frontiers in Robotics and Edge Computing with AI



NVIDIA

Webinar Agenda

Topic:

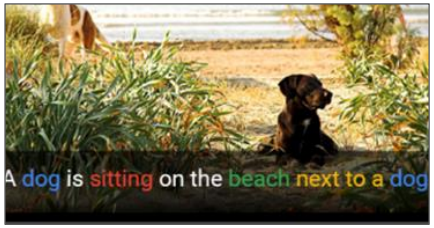
- AI at the Edge
- Jetson TX2
- JetPack 3.1
- 2 Days To A Demo
- Case Study

- Isaac Initiative
- Reinforcement Learning
- Conclusion / Q&A

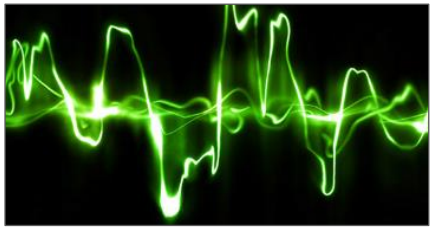
AMAZING ACHIEVEMENTS IN AI

Play Go

Play Games

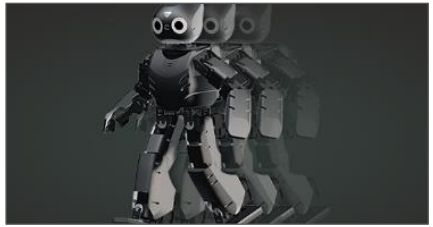


Write Captions



Speech Synthesis

Learn Motor Skills



Learn to Walk

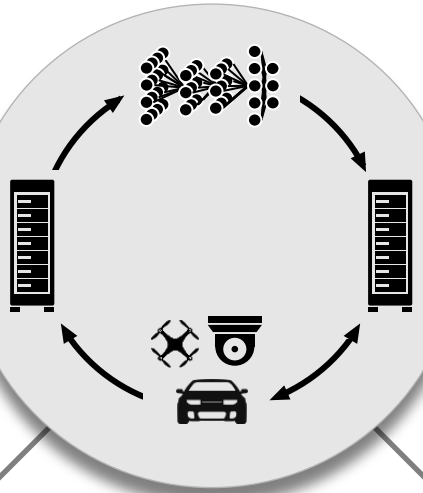
Drive

Fly

GPU DEEP LEARNING IS A NEW COMPUTING MODEL

DGX-1

Training



Tesla

Inference in the Datacenter

Jetson

Inference at the Edge

WHY AI AT THE EDGE MATTERS

BANDWIDTH

1 billion cameras WW (2020)
10’s of petabytes per day

LATENCY

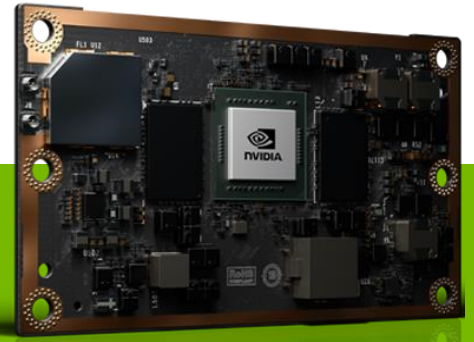
Safety-critical services
Realtime decisions

PRIVACY

Confidentiality
Private cloud or on-premise storage

CONNECTIVITY

50% of populated world < 8mbps
Bulk of uninhabited world no 3G+



NVIDIA Jetson TX2

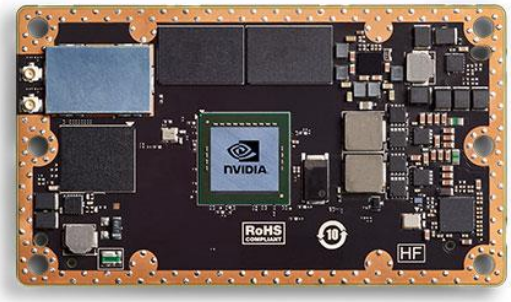
64-bit ARM Cortex-A57 + NVIDIA Denver 2 CPU

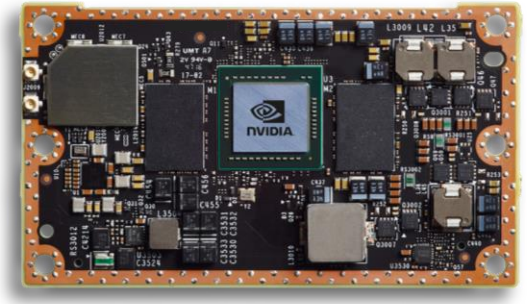
256-core NVIDIA Pascal GPU

8GB LPDDR4, 32GB eMMC

4Kp60 encode/decode

NVIDIA





	JETSON TX1	JETSON TX2
GPU	Maxwell	Pascal
CPU	64-bit A57 CPUs	64-bit Denver 2 and A57 CPUs
Memory	4 GB 64 bit LPDDR4 25.6 GB/s	8 GB 128 bit LPDDR4 58.4 GB/s
Storage	16 GB eMMC	32 GB eMMC
Wi-Fi/BT	802.11 2x2 ac/BT Ready	802.11 2x2 ac/BT Ready
Video Encode	4Kp30 | (2x) 1080p60	4Kp60 | (3x) 4Kp30 | (8x) 1080p60
Video Decode	4Kp60 | (4x) 1080p60	(2x) 4Kp60
Camera	1.4Gpix/s Up to 1.5Gbps per lane	1.4Gpix/s Up to 2.5Gbps per lane
Mechanical	50mm x 87mm 400-pin Compatible Board to Board Connector	

DUAL OPERATING MODES

MAX-Q: Maximum Efficiency

Maximum energy efficiency

Up to 2x the energy efficiency of Jetson TX1

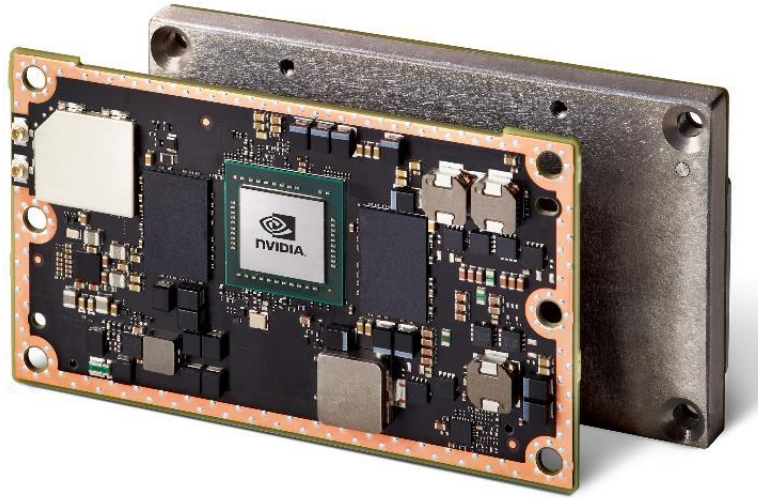
Less than 7.5 W

MAX-P: Maximum Performance

Maximum performance

Up to 2x the performance of Jetson TX1

Less than 15 W



JETSON TX2 DEVELOPER KIT

Open-source reference design

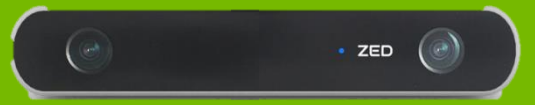
MIPI CSI-2 camera module

EDU discount available

NVIDIA

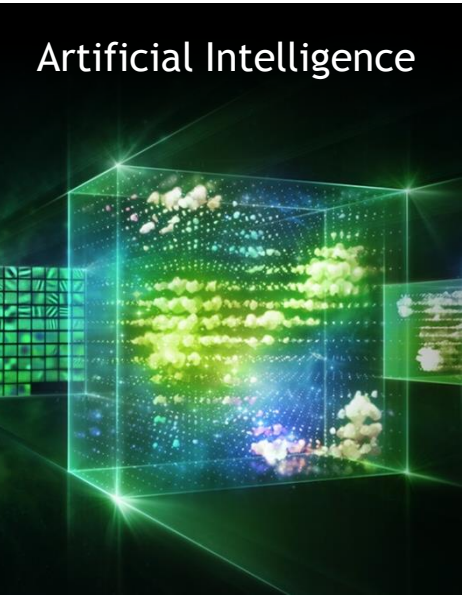
JETSON Ecosystem

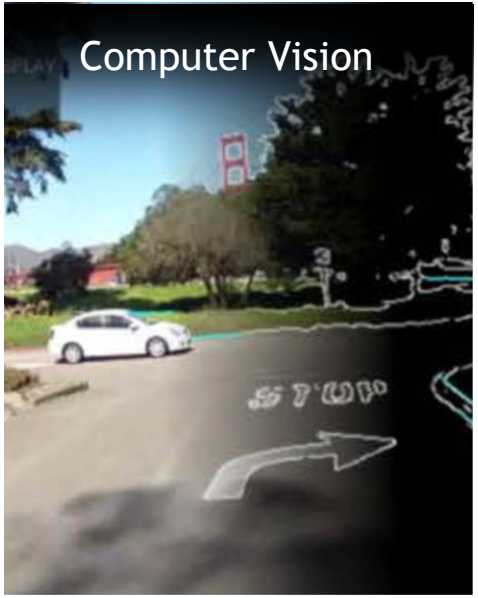
- Miniature carriers
- Enclosures
- Cameras
- Custom Solutions

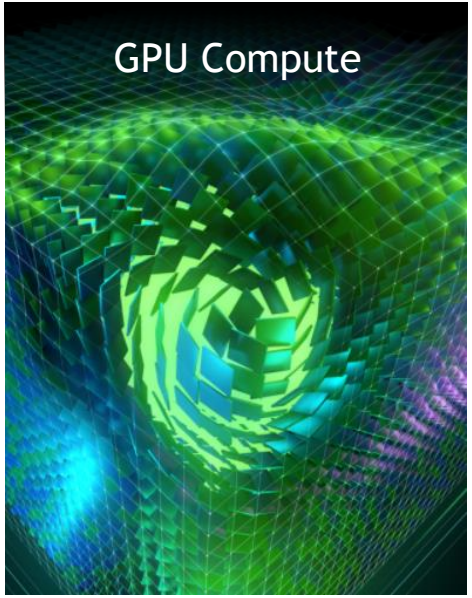


NVIDIA JETPACK

SDK for Intelligent Devices







TensorRT

cuDNN

NVIDIA VISIONWORKS

OpenCV

NVIDIA CUDA

nvGRAPH

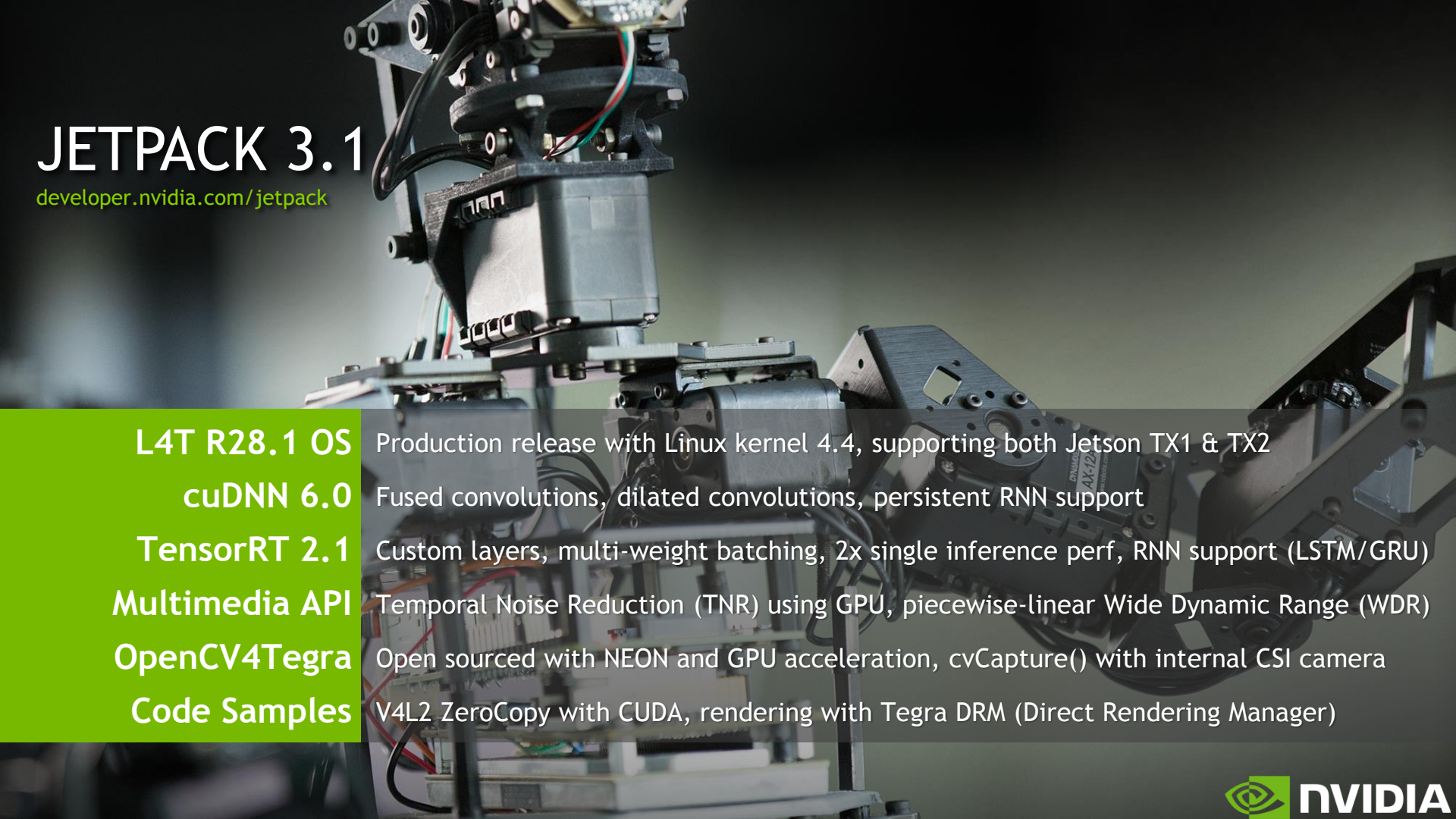
cuBLAS

cuSPARSE

libargus
Low level
camera API

gstreamer

V4L2



JETPACK 3.1

developer.nvidia.com/jetpack

L4T R28.1 OS	Production release with Linux kernel 4.4, supporting both Jetson TX1 & TX2
cuDNN 6.0	Fused convolutions, dilated convolutions, persistent RNN support
TensorRT 2.1	Custom layers, multi-weight batching, 2x single inference perf, RNN support (LSTM/GRU)
Multimedia API	Temporal Noise Reduction (TNR) using GPU, piecewise-linear Wide Dynamic Range (WDR)
OpenCV4Tegra	Open sourced with NEON and GPU acceleration, cvCapture() with internal CSI camera
Code Samples	V4L2 ZeroCopy with CUDA, rendering with Tegra DRM (Direct Rendering Manager)

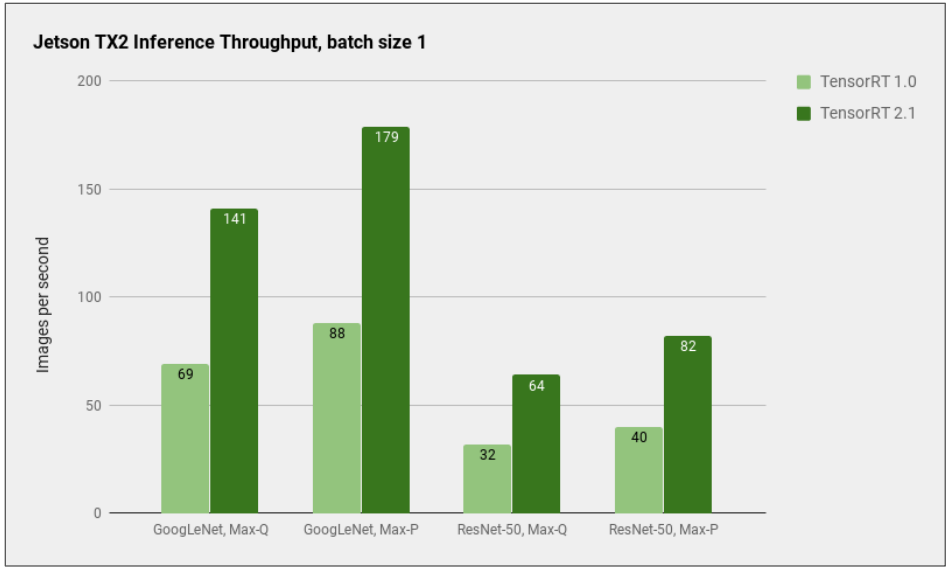
NVIDIA

JetPack 3.1

2x Low-Latency Inference Performance
for Jetson TX1 and TX2

Software Components	
Linux4Tegra R28.1	Linux kernel 4.4
Reference Root OS	Ubuntu 16.04 LTS aarch64
Inference Runtime	TensorRT 2.1
CUDA Toolkit 8	cuDNN v6.0
VisionWorks 1.6	OpenCV4Tegra 2.4.13
OpenGL 4.5	EGL 1.4 | OpenGL ES 3.1
Multimedia API SDK	Argus | V4L2 | GStreamer
Tegra System Profiler 3.8	Tegra Graphics Debugger 2.4

developer.nvidia.com/jetpack



NETWORK	LATENCY		Speedup
	TensorRT 1.0	TensorRT 2.1	
GoogLeNet, Max-Q	14.5ms	7.1ms	2.04x
GoogLeNet, Max-P	11.4ms	5.6ms	2.04x
ResNet-50, Max-Q	31.4ms	15.6ms	2.01x
ResNet-50, Max-P	24.7ms	12.2ms	2.03x

JetPack 3.1 Doubles Jetson's Low-Latency Inference Performance
https://devblogs.nvidia.com/parallelforall/jetpack-doubles-jetson-inference-perf

Realtime AI

Low-Latency Inferencing

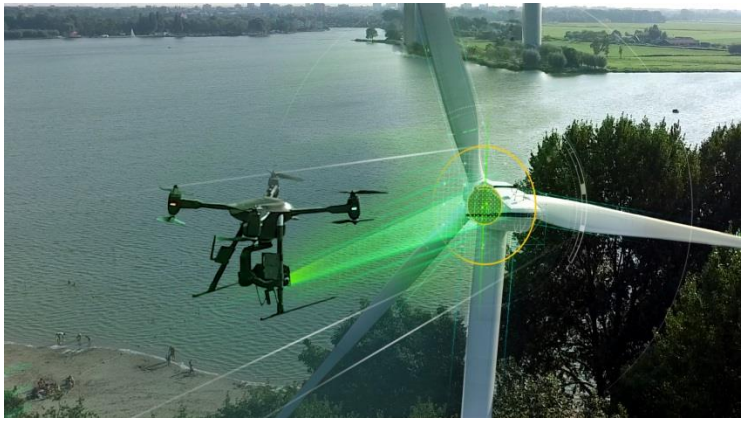
Higher batch sizes increase throughput, but add a frame of latency for each additional instance.

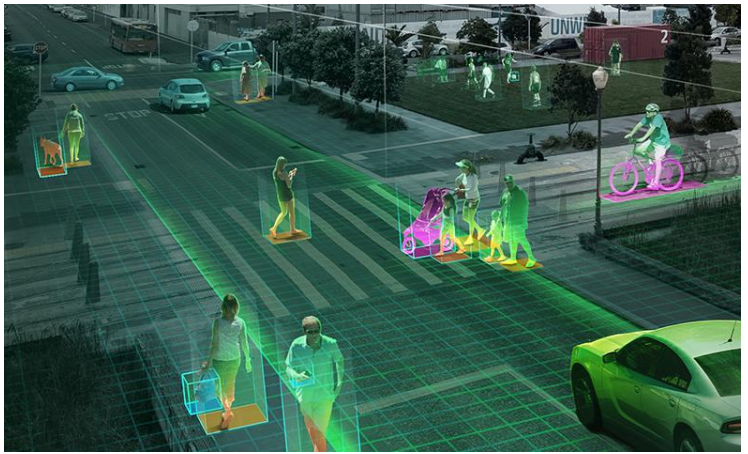
A batch of 1 single frame results in the lowest latency, useful for edge systems with realtime constraints like:

- Tracking
- Motion control
- Obstacle detection
- Collision avoidance
- Path following
- Autonomous navigation

With TensorRT 2.1, single-batch performance is doubled with latencies down to 5.5ms for GoogleNet recognition.

developer.nvidia.com/tensorRT





NVIDIA TensorRT 2

Deep Learning Inference Optimizer and Runtime

High-performance neural network inference optimizer and runtime engine for production deployment

Maximize inference throughput for latency-critical services for production in the cloud and embedded

Optimize pretrained models to generate runtime engines that maximize inference throughput

New features in TensorRT 2:

- Optimized single batch inference for low-latency services
- Custom layer plugins and support for Reshape, ROIPooling layers, 32-bit RNNs (LSTM + GRU), and Region Proposal Object Detection networks like Faster-RCNN and YOLO

developer.nvidia.com/tensorRT

```
#include "NvInfer.h"

using namespace nvinfer1;

// example plugin definition
class MyPlugin : IPlugin
{
public:
  int getNbOutputs() const;

  Dims getOutputDimensions(int index, const Dims* inputs,
                           int nbInputDims);

  void configure(const Dims* inputDims, int nbInputs,
                 const Dims* outputDims, int nbOutputs,
                 int maxBatchSize);

  int initialize();

  void terminate();

  size_t getWorkspaceSize(int maxBatchSize) const;

  int enqueue(int batchSize, const void* inputs,
              void** outputs, void* workspace,
              cudaStream_t stream);

  size_t getSerializationSize();

  void serialize(void* buffer);

protected:
  virtual ~MyPlugin();
};
```

NVIDIA

NVIDIA TensorRT 2

Deep Learning Inference Optimizer and Runtime

High-performance neural network inference optimizer and runtime engine for production deployment

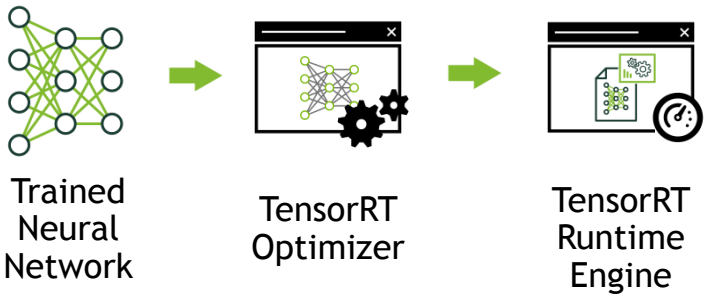
Maximize inference throughput for latency-critical services for production in the cloud and embedded

Optimize pretrained models to generate runtime engines that maximize inference throughput

New features in TensorRT 2:

- Optimized single batch inference for low-latency services
- Custom layer plugins and support for Reshape, ROIPooling layers, 32-bit RNNs (LSTM + GRU), and Region Proposal Object Detection networks like Faster-RCNN and YOLO

developer.nvidia.com/tensorRT



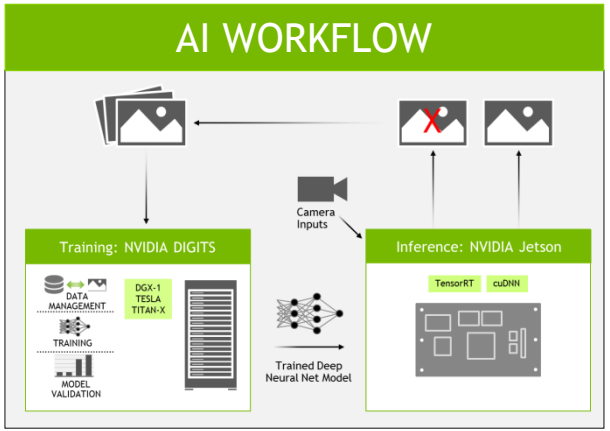
- Fuse network layers
- Eliminate concatenation layers
- Kernel specialization
- Auto-tuning for target platform
- Select optimal tensor layout
- Batch size tuning
- Half-precision FP16 support

NVIDIA

TWO DAYS TO A DEMO

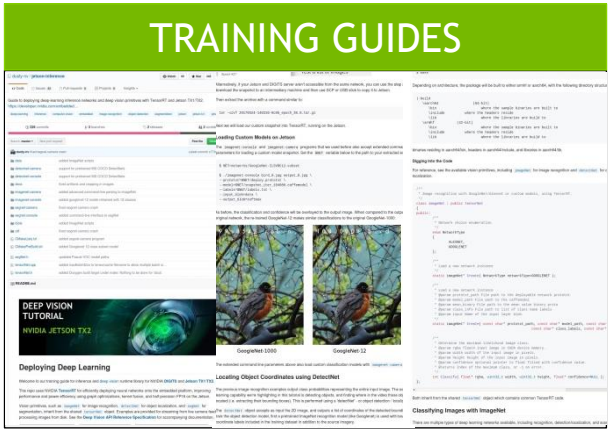
Get Started with Deep Learning

AI WORKFLOW



Train using DIGITS and cloud/PC
Deploy to the field with Jetson

TRAINING GUIDES



All the steps required to follow to train
your own models, including the datasets.

DEEP VISION PRIMITIVES

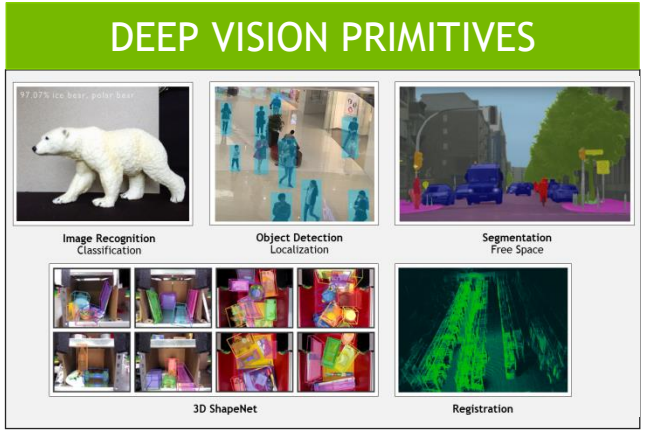


Image Recognition, Object Detection
and Segmentation

github.com/dusty-nv/jetson-inference

NVIDIA

Two Days to a Demo

Guide to Deploying Deep Learning

Create runtime primitives from:

- 16 pretrained models of 1000+ objects
- User-customized models
- From the command line

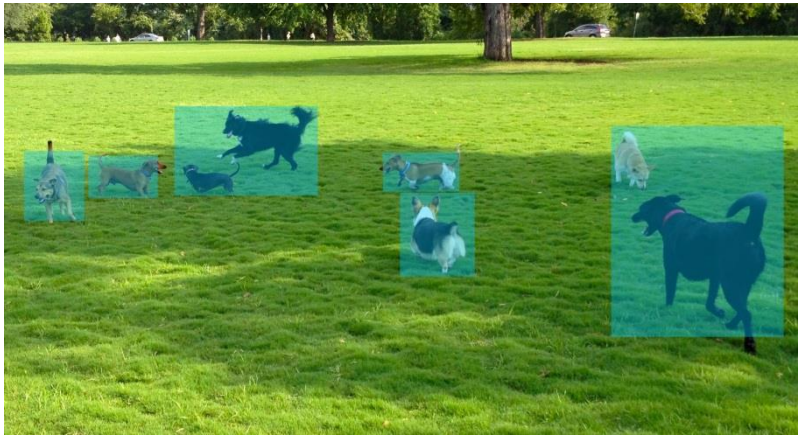
TensorRT API underneath

Live camera streaming

ROS classification nodes

2x faster with TensorRT 2

github.com/dusty-nv/jetson-inference



```
class detectNet : public tensorNet
{
public:
enum NetworkType {
   COCO_AIRPLANE, COCO_BOTTLE, COCO_CHAIR, COCO_DOG,
   FACENET, PEDNET, PEDNET_MULTI
};

static detectNet* Create( NetworkType networkType, float threshold=0.5f );

static detectNet* Create( const char* prototxt_path, const char* model_path,
                          const char* mean_binary, float threshold=0.5f );

static detectNet* Create( int argc, char** argv );

bool Detect( float* rgba, uint32_t width, uint32_t height,
             float* boundingBoxes, int* numBoxes );

};
```

Two Days to a Demo

Guide to Deploying Deep Learning

Create runtime primitives from:

- 16 pretrained models of 1000+ objects
- User-customized models
- From the command line

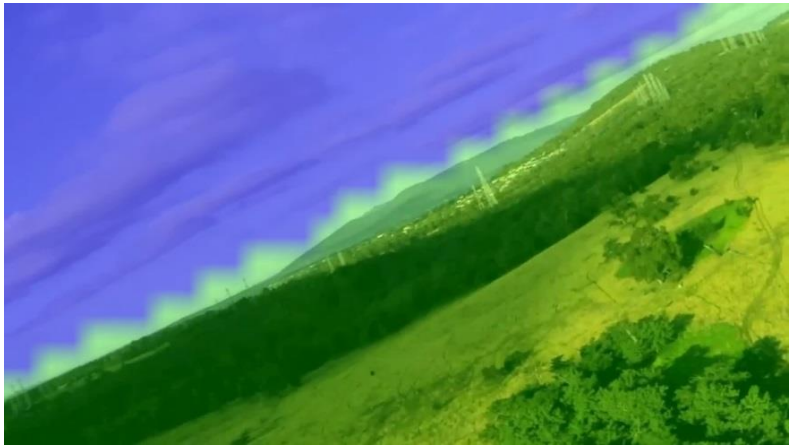
TensorRT API underneath

Live camera streaming

ROS classification nodes

2x faster with TensorRT 2

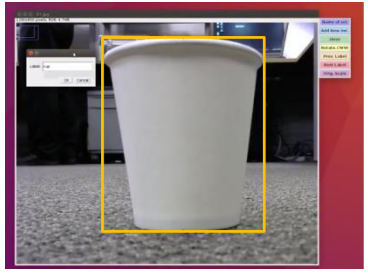
github.com/dusty-nv/jetson-inference



```
#include “segNet.h”

segNet* net = segNet::Create(“/path/to/fcn_prototxt.txt”,
                             “/path/to/model.caffemodel”,
                             “/path/to/my_classes.txt”);

const int size = width * height * sizeof(float) * 4;

void* input = NULL;
void* output = NULL;

cudaMalloc(&input, size);
cudaMalloc(&output, size);

net->Overlay(input, output, width, height);
```

NVIDIA H.S. INTERNS

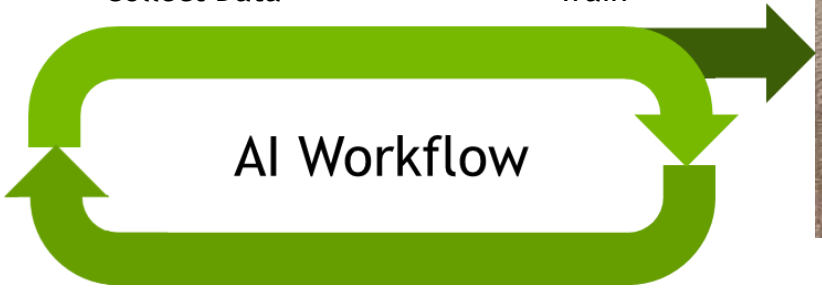
Summer 2017



Collect Data

Train

Test!



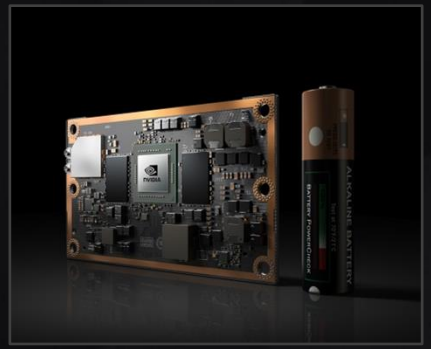
FIRST ROBOTICS COMPETITION

Delivery Bot

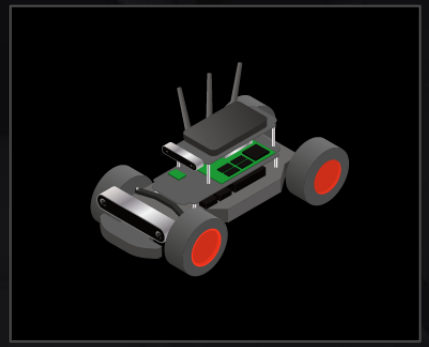
Self-Driving Racecar

NVIDIA

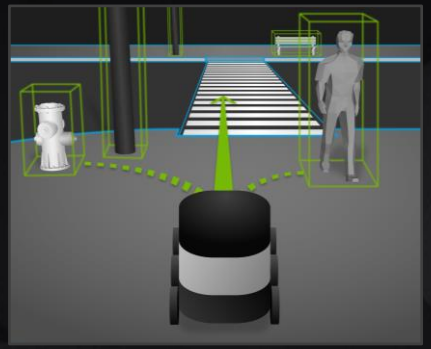
THE ISAAC INITIATIVE



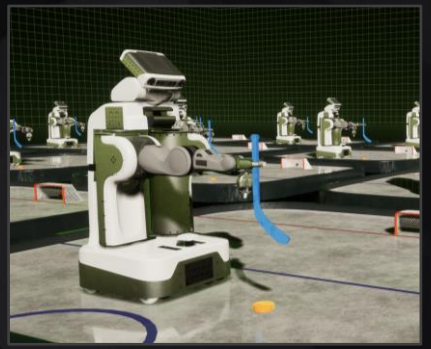
Jetson TX2



AV Reference Platforms

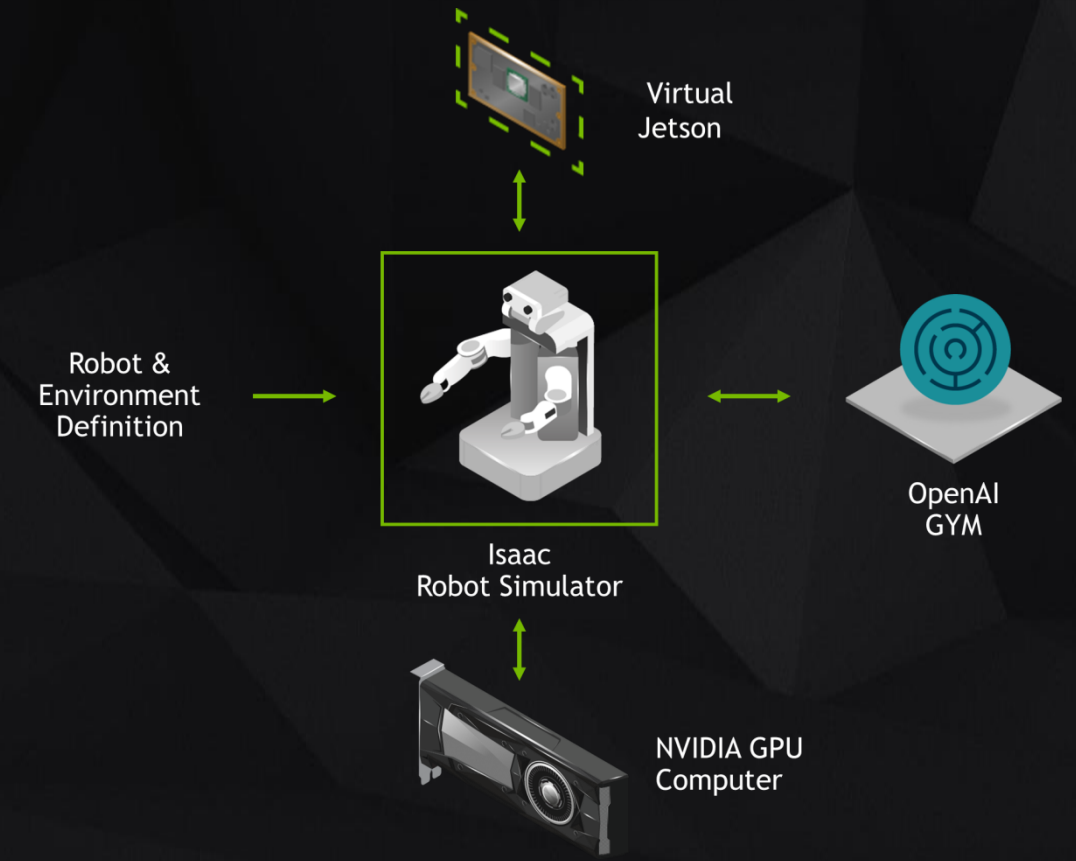


Astro AV Stack



Isaac Lab

ISAAC LAB


Virtual
Jetson
Robot &
Environment
Definition
OpenAI
GYM
Isaac
Robot Simulator
NVIDIA GPU
Computer

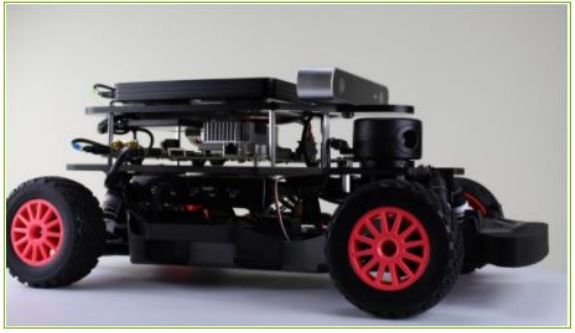
JETSON REFERENCE PLATFORMS

Toyota HSR

Teal Drone

enRoute UGV

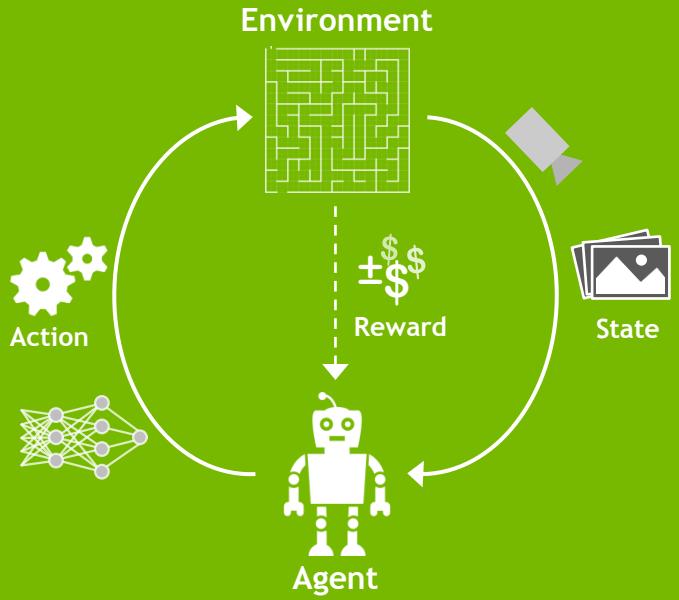
enRoute USV



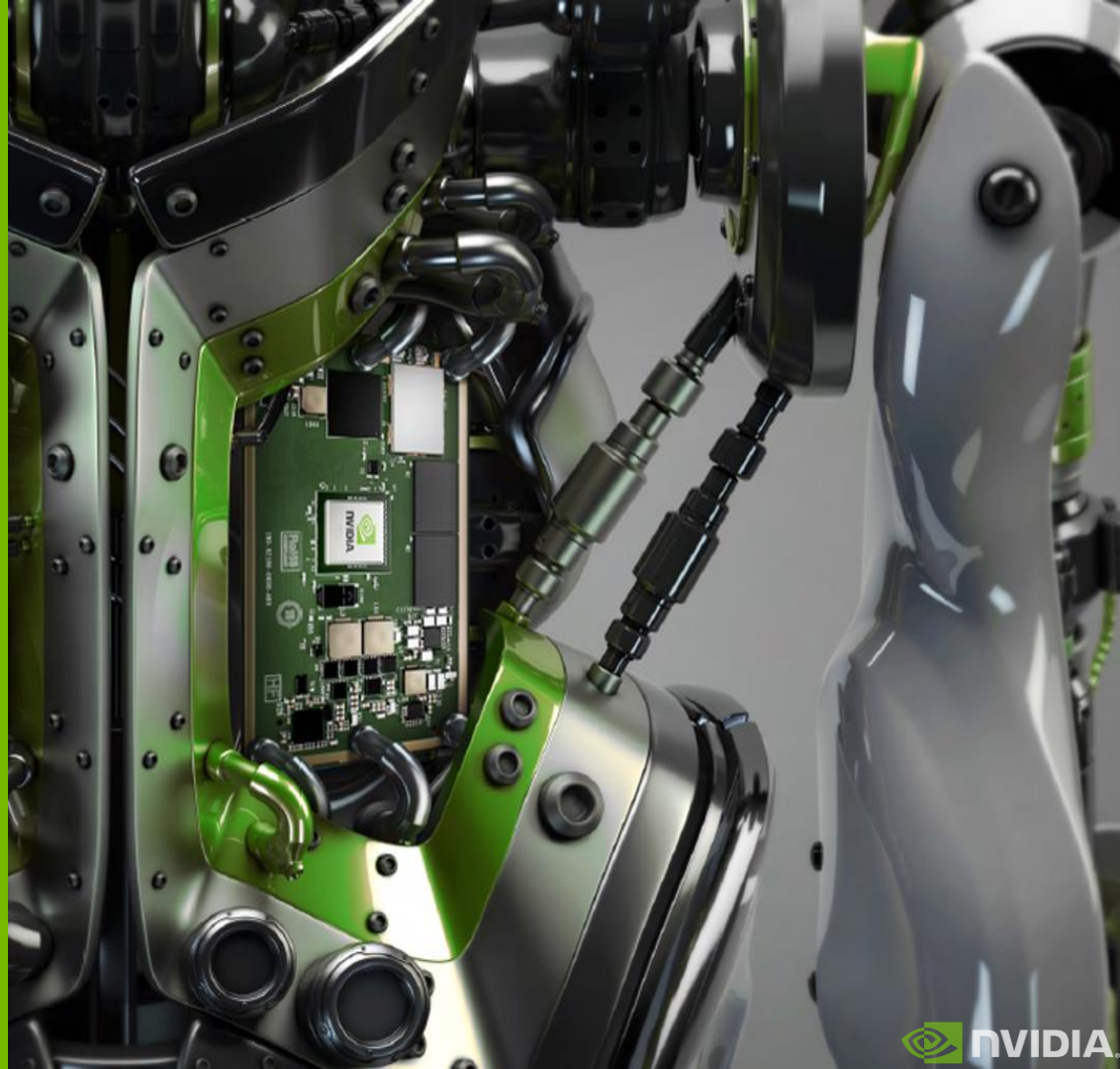
JetsonHacks RACECAR/J

enRoute Industrial UAV

Reinforcement Learning



arXiv:1611.06256 *GA3C: GPU-based A3C for Deep Reinforcement Learning*, Y. Kautz et al., NVIDIA Research, 2016.



TWO DAYS TO A DEMO

Reinforcement Learning Edition

OpenAI Gym

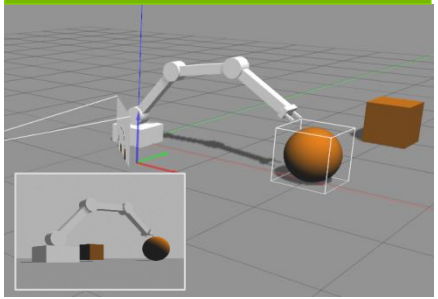
Test environments and games for research and verification

RL Algorithms

```
Do you want to continue [Y/n]? y
Get:1 http://archive.ubuntu.com/ubuntu/ lucid/universe python-keybinde
[12.2kB]
Get:2 http://archive.ubuntu.com/ubuntu/ lucid/universe terminator 0.93
[190kB]
Fetched 202kB in 5s (37.2kB/s)
Selecting previously deselected package python-keybinder.
(Reading database ... 129972 files and directories currently installed
Unpacking python-keybinder (from .../python-keybinder_0.0.4-1_i386.deb
Selecting previously deselected package terminator.
Unpacking terminator (from .../terminator_0.93-0ubuntu1_all.deb) ...
Processing triggers for desktop-file-utils ...
Processing triggers for python-gmenu ...
Rebuilding /usr/share/applications/desktop.en_US.utf8.cache...
Processing triggers for man-db ...
Processing triggers for hicolor-icon-theme ...
Processing triggers for python-support ...
Setting up python-keybinder (0.0.4-1) ...

Setting up terminator (0.93-0ubuntu1) ...
update-alternatives: using /usr/bin/terminator to provide /usr/bin/x-t
ulator (x-terminal-emulator) in auto mode.
```

DQN, DDPG, A3C, Actor Critic
PyTorch and TensorFlow

Robotic Simulation



Observation from vision
Pixels-to-actions

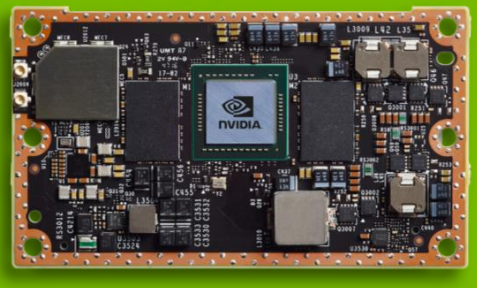
Transfer Learning

Adapt network to real robot
Online learning in the field

github.com/dusty-nv/jetson-reinforcement

NVIDIA

Thank you!



Developer Portal	developer.nvidia.com/embedded
Download JetPack	developer.nvidia.com/jetpack
2 Days To a Demo	github.com/dusty-nv
Jetson Forums	devtalk.nvidia.com
Visit the Wiki	eLinux.org/Jetson
EDU Discount	bit.ly/2veKN1X

Q&A: What can I help you build?

NVIDIA