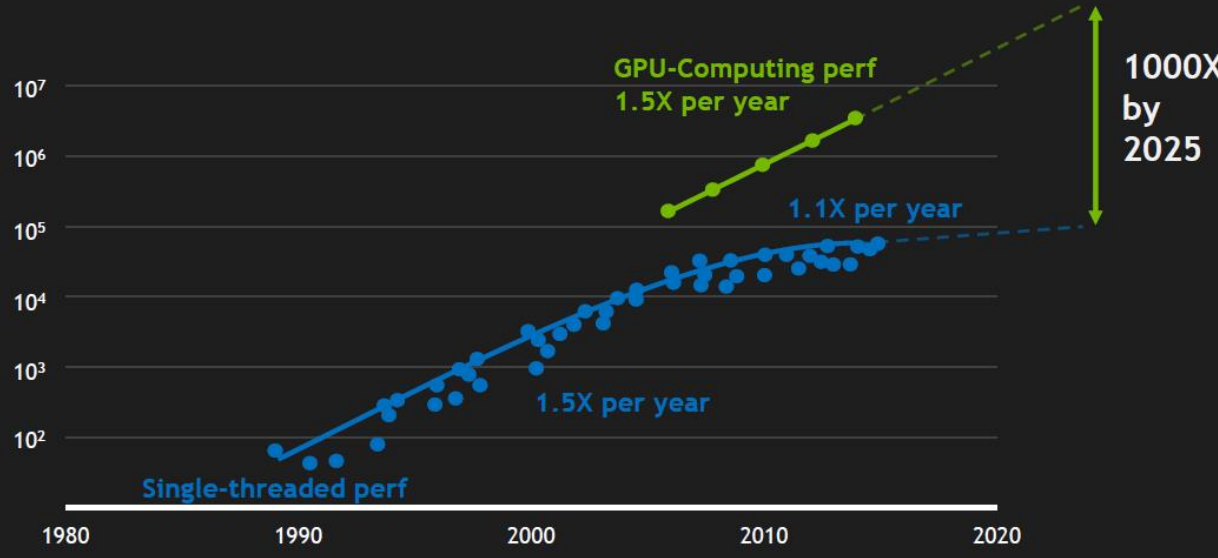
MATLAB EXPO 2017

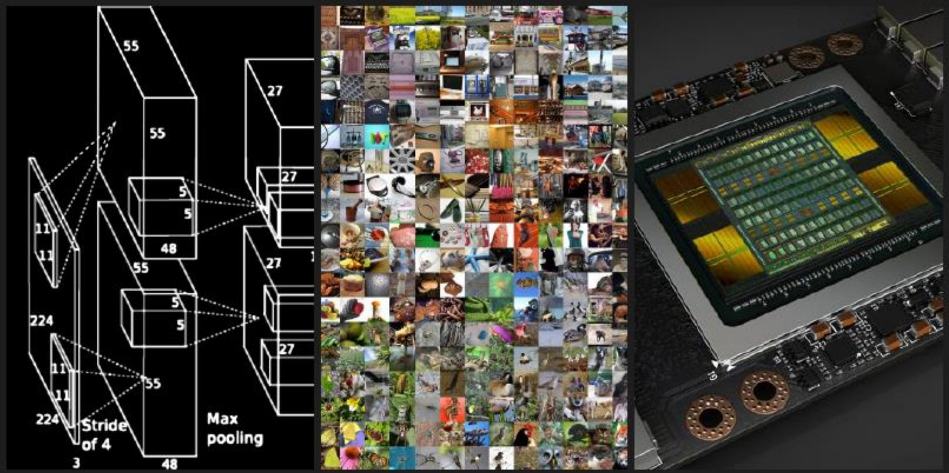
Deep Learning: Transforming Engineering and Science

DEEP LEARNING: TRANSFORMING ENGINEERING AND SCIENCE

THE RISE OF GPU COMPUTING



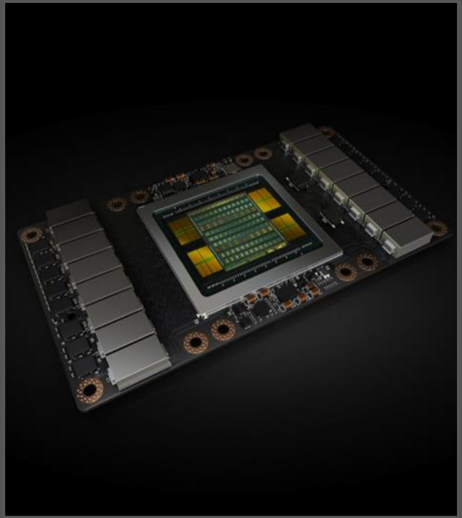
40 Years of Microprocessor Trend Data

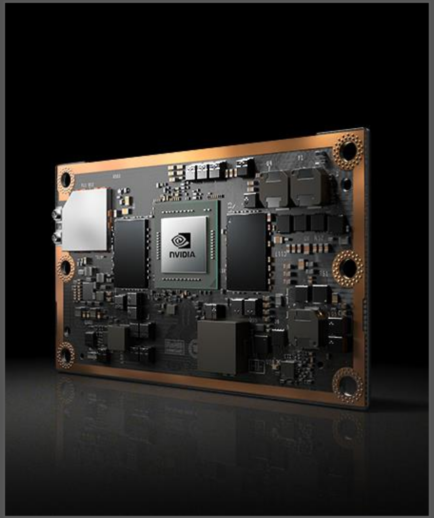


The Big Bang of Deep Learning

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

NVIDIA IS THE WORLD’S LEADING AI PLATFORM



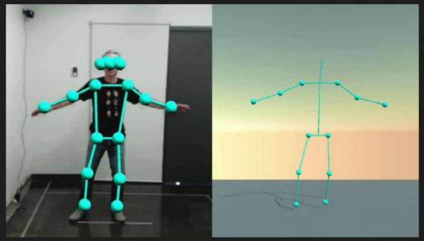



ONE ARCHITECTURE — CUDA

AMAZING ACHIEVEMENTS IN AI

NVIDIA
Interactive Ray Tracing

NVIDIA / Remedy
Audio-driven Facial Animation



WRNCH
Pose Estimation

University of Edinburgh
Character Animation

UC Berkeley / OpenAI
One-shot Imitation Learning

A WORLD OF INTELLIGENT MACHINES

10% of Manufacturing Tasks Are Automated

1M Pizzas Delivered Per Day by Domino’s

100M People 80+ Years Old

Ag Tech: 70% Increase in Farm Yields by 2050

600K Bridges to Inspect in the U.S.

300M Operations per Year WW

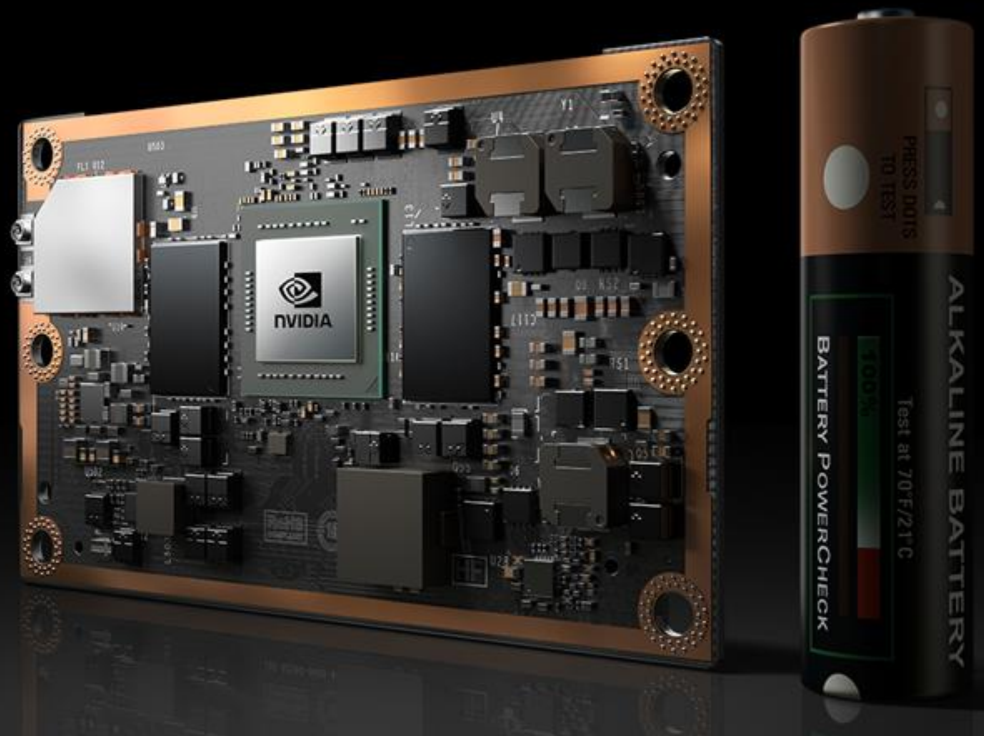
JETSON TX2

SUPERCOMPUTER FOR AI AT THE EDGE

2 Core i7 PCs in <10W

256 CUDA cores

>1 TFLOPS

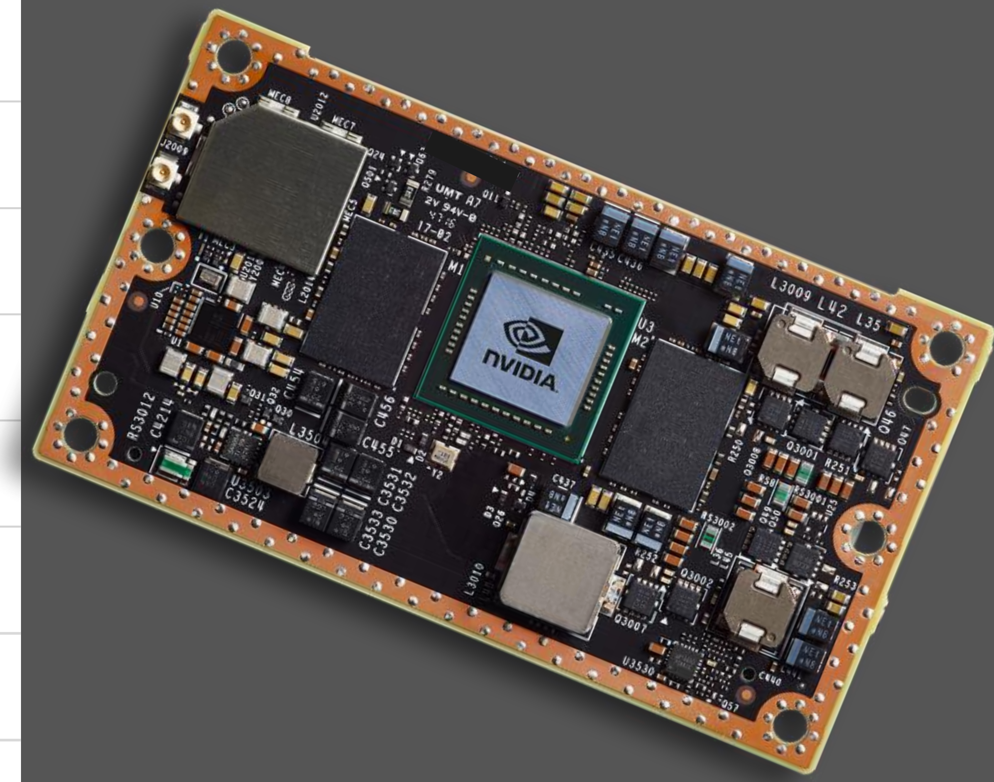


cuDNN, TensorRT

CUDA

Linux, ROS

	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	2160p @ 30	2160p @ 60
Video Decode	2160p @ 60	2160p @ 60 12 bit support for H.265, VP9
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	



INDUSTRY ADOPTION

Manufacturing

Agriculture

Construction

Inventory Management

Logistics/Retail

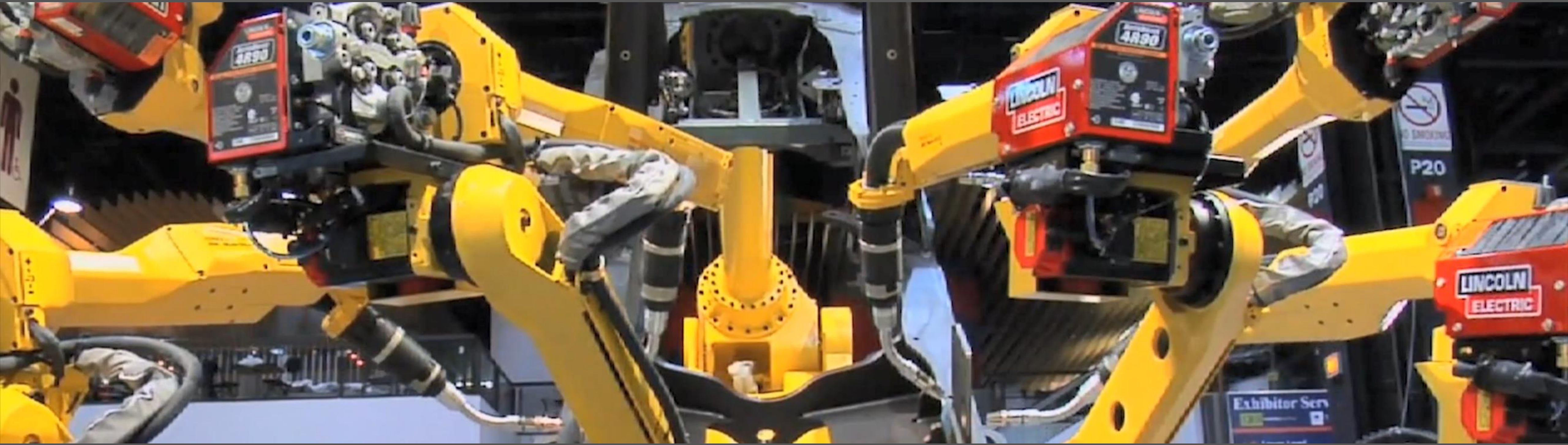
Security

Delivery

Inspection

Autonomous UAV

Social


4R90
LINCOLN ELECTRIC
4R90
P20
LINCOLN ELECTRIC

RESEARCH & EDUCATION ADOPTION

JETPACK SDK FOR AI @ THE EDGE

Sample Code

Nsight Developer Tools

Multimedia API

TensorRT
cuDNN

Deep Learning

VisionWorks
OpenCV

Computer Vision

Vulkan
OpenGL

Graphics

libargus
Video API

Media

CUDA, Linux4Tegra, ROS

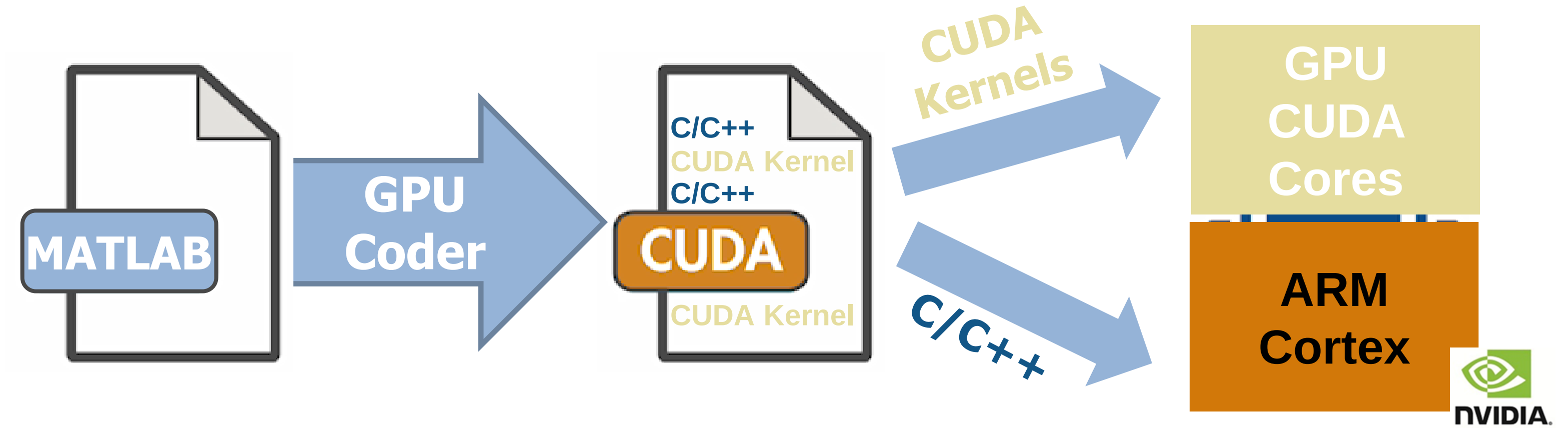
Jetson Embedded Supercomputer: Advanced GPU, 64-bit CPU, Video CODEC, VIC, ISP

How do we target the Jetson TX2 from MATLAB ?

Introducing **GPU Coder**

R2017b

- Generates **CUDA** code, which can be used only on NVIDIA GPUs*
- CUDA extends C/C++ code with constructs for parallel computing

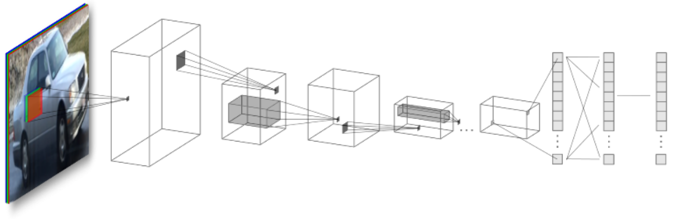


* **Any modern** CUDA-enabled GPU with **compute capability 3.2** or higher

Why Use GPU Coder?

Neural Networks

Deep Learning, machine learning



Up to 7x faster

than state-of-art

Image Processing and Computer Vision

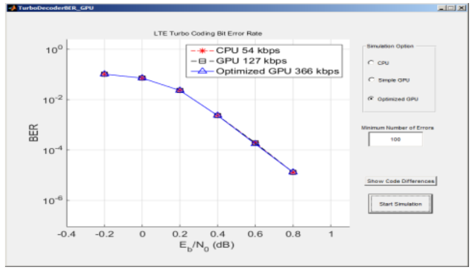
Image filtering, feature detection/extraction

Up to 700x faster

than CPUs for feature extraction

Signal Processing and Communications

FFT, filtering, cross correlation,



Up to 20x faster

than CPUs for FFTs

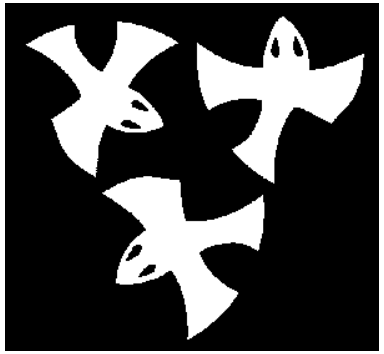
Performance

How fast is GPU Coder?

Fog removal

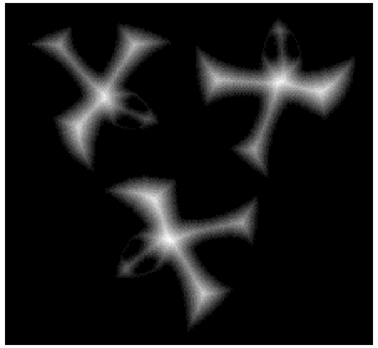
5x speedup

Orders magnitude speedup over optimized C code.



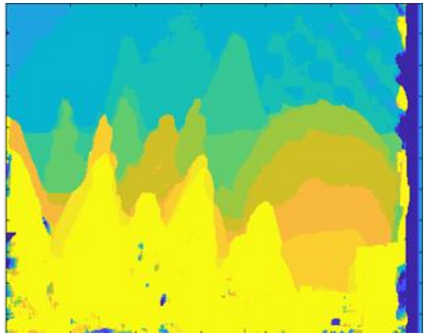
Distance transform

8x speedup



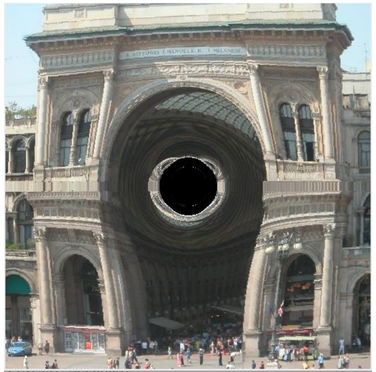
Stereo disparity

50x speedup



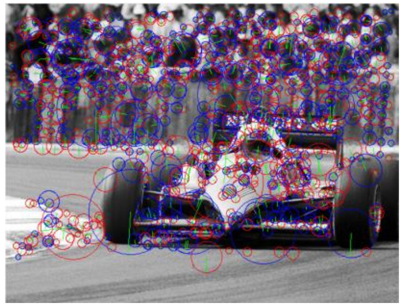
Ray tracing

18x speedup

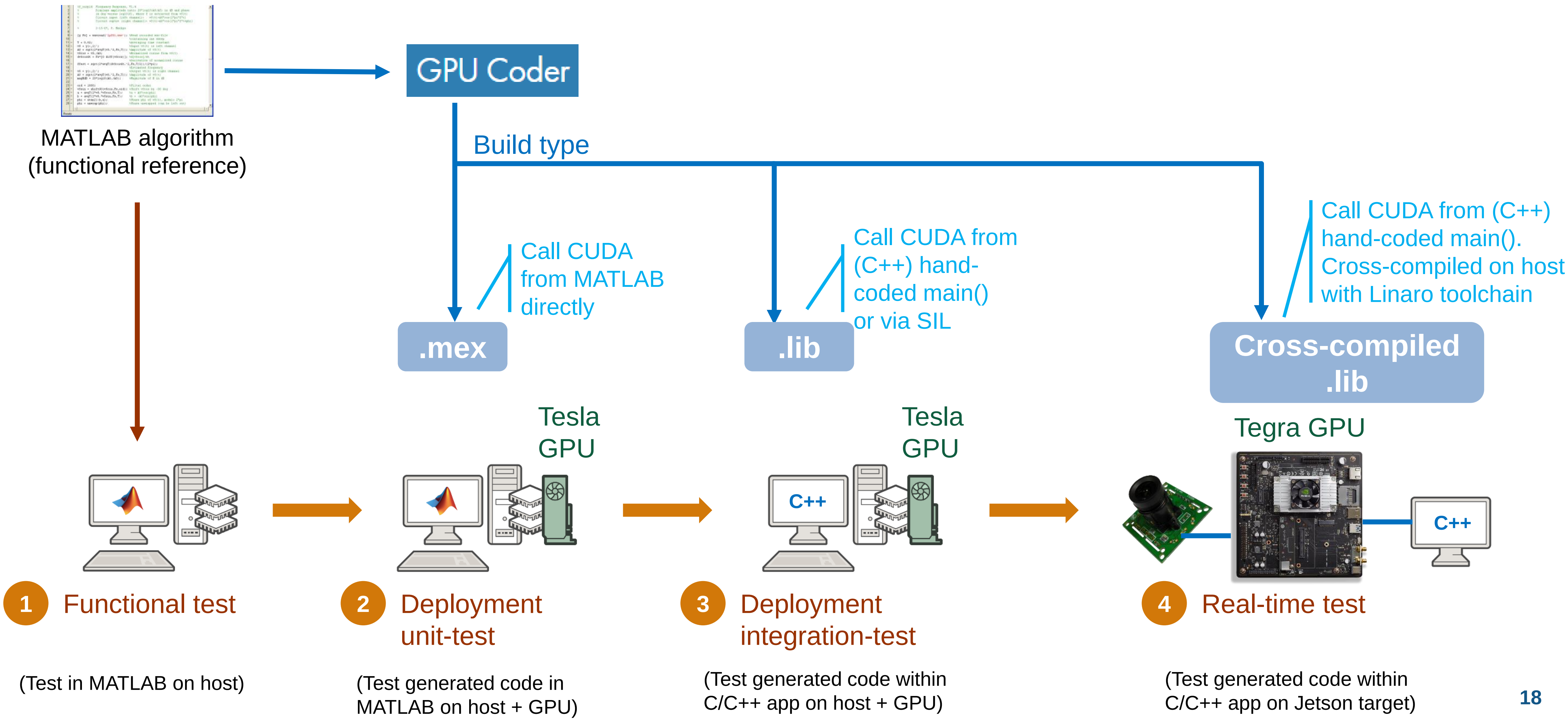


SURF feature extraction

700x speedup

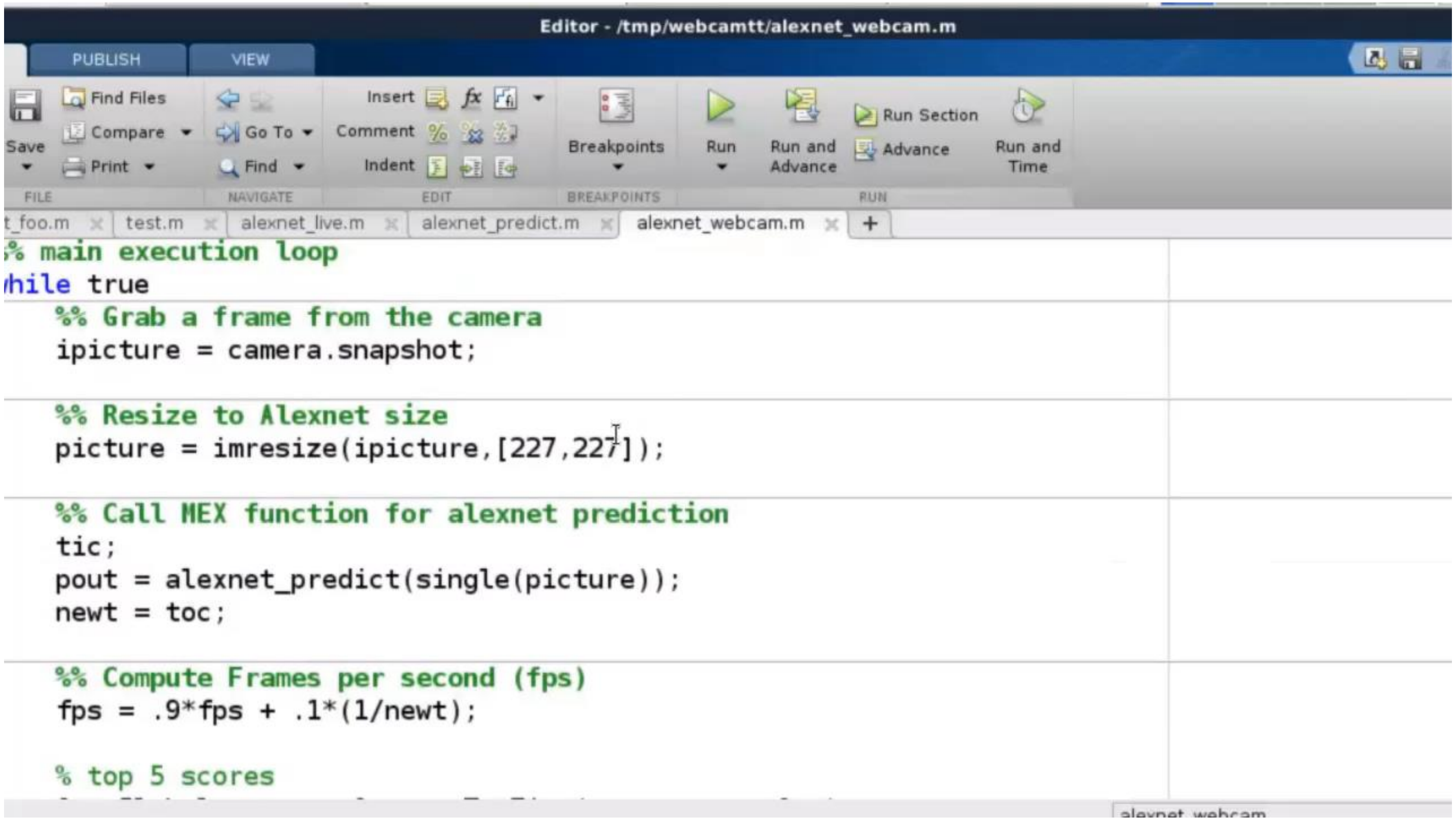


How to Use GPU Coder? Workflow to Embedded Jetson GPU

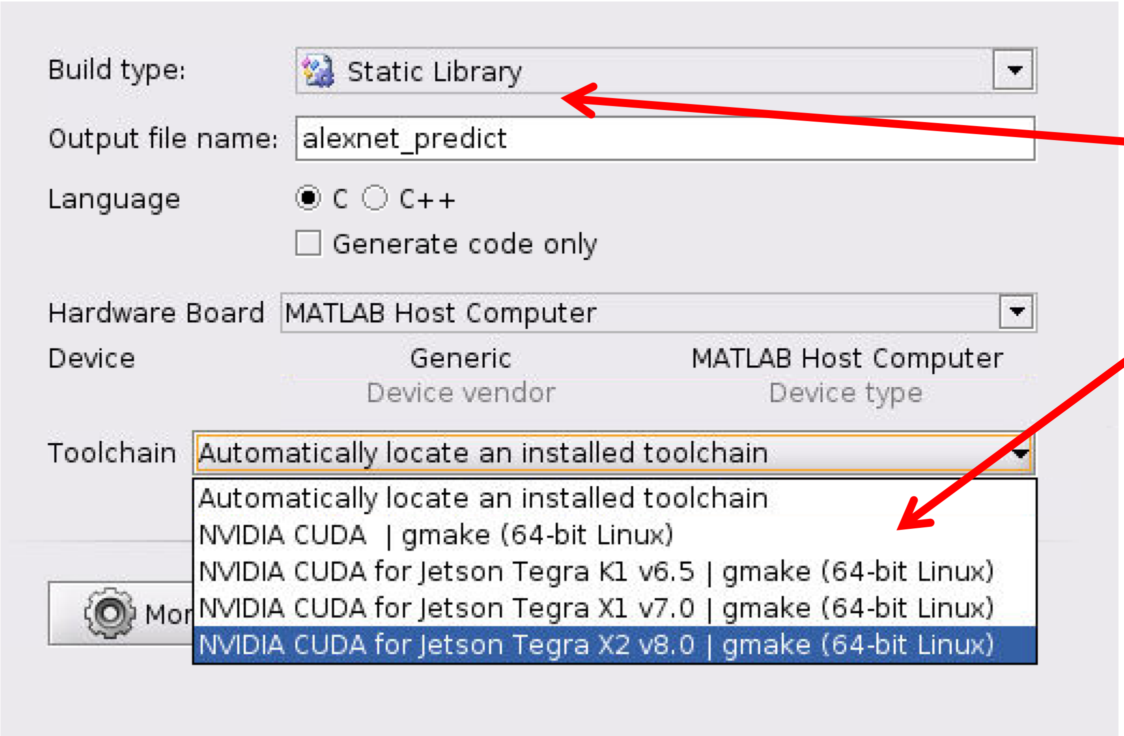


Demo: Generate CUDA Code for AlexNet Prediction

“Hello World” for Deep Learning



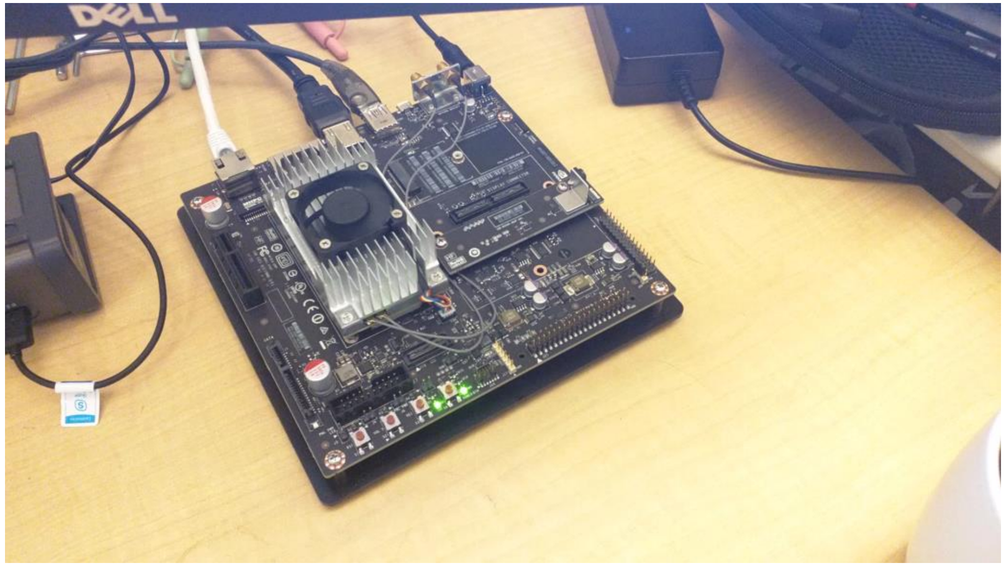
Deployment to NVIDIA Jetson: Cross-Compiled ‘lib’



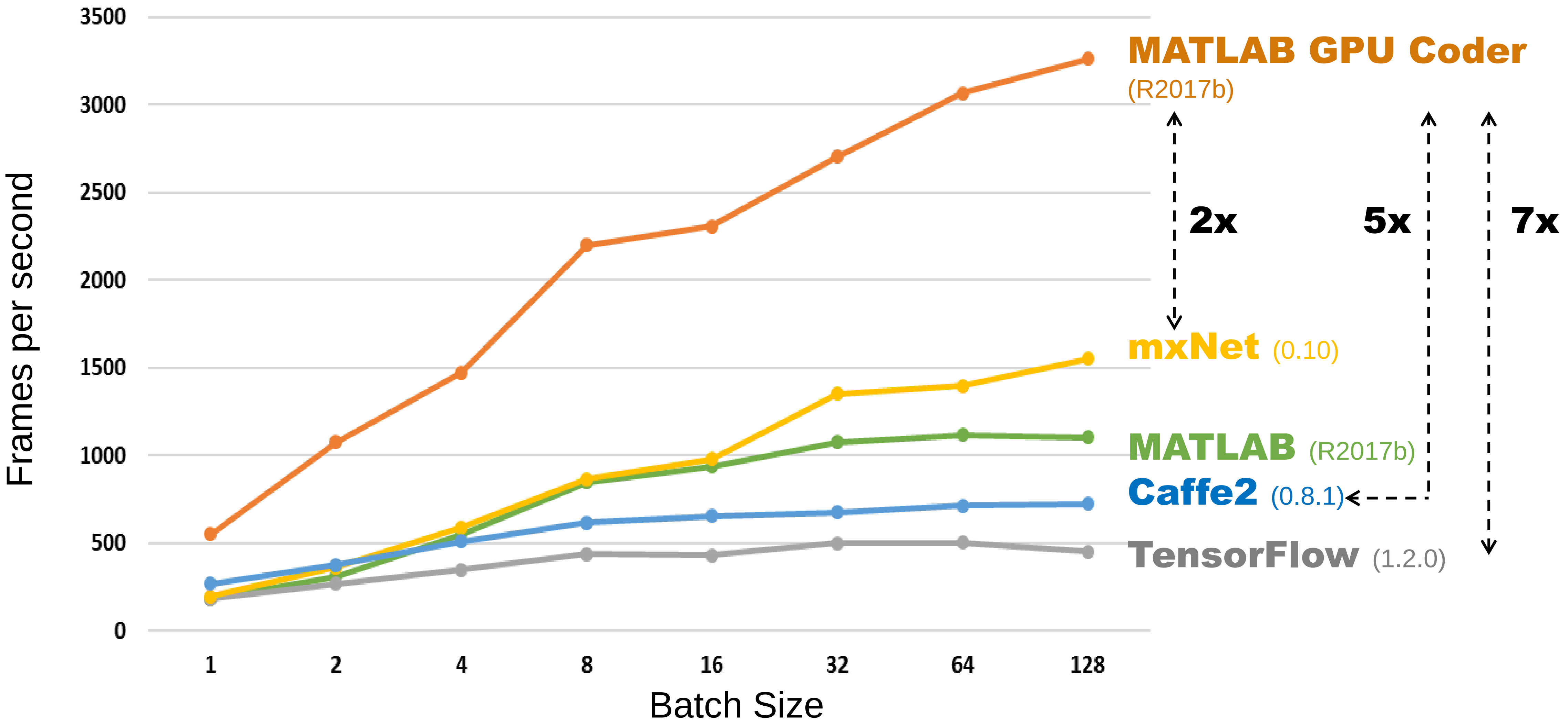
Two small changes

1. Change build-type to ‘lib’

2. Select cross-compile toolchain



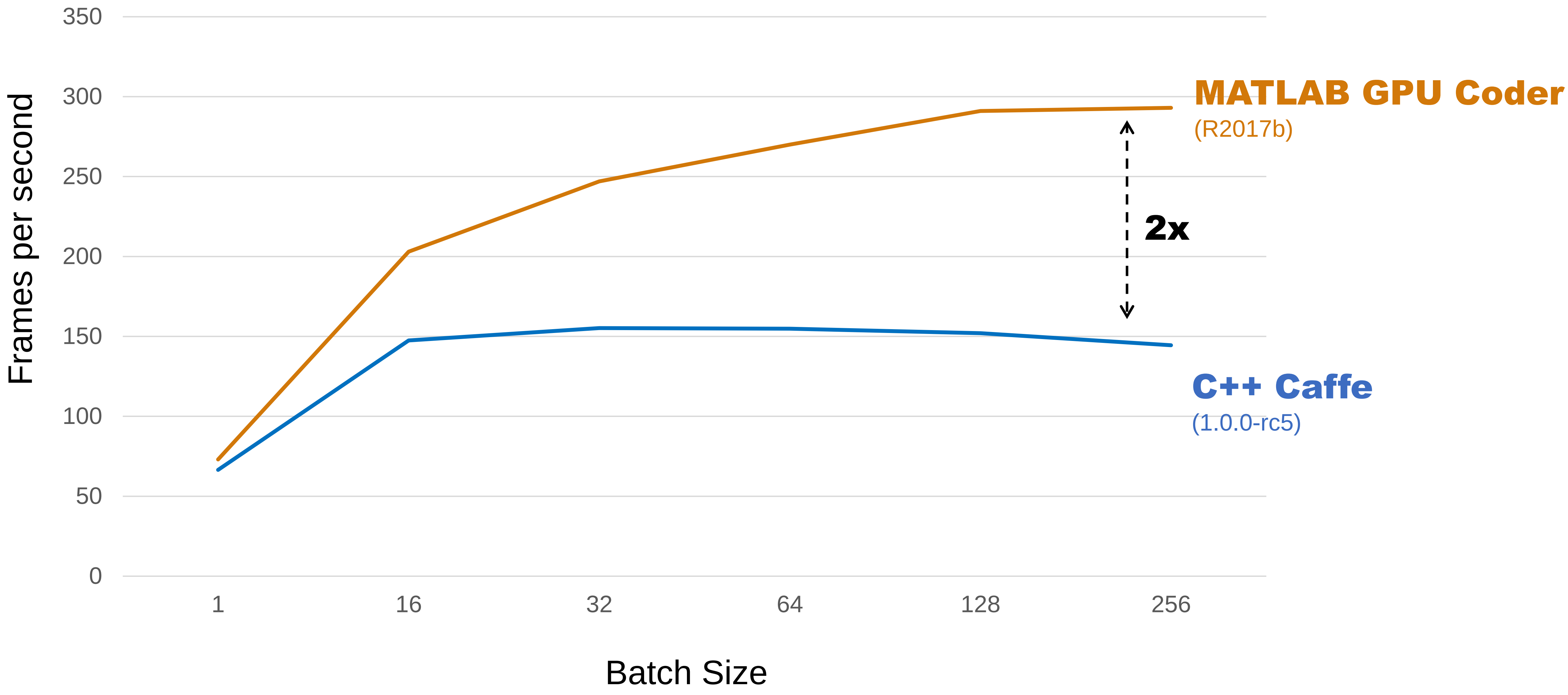
Alexnet Inference on NVIDIA **Titan XP**



Testing platform

CPU	Intel(R) Xeon(R) CPU E5-1650 v3 @ 3.50GHz
GPU	Pascal Titan Xp
cuDNN	v5

Alexnet Inference on **Jetson TX2**: Frame-Rate Performance



Why is GPU Coder Faster than OSS Deep Learning Frameworks?

- OSS frameworks are designed to do many things, including:
  - Training
  - Inference
  - Support various data types (singles, FP16, int8, etc)

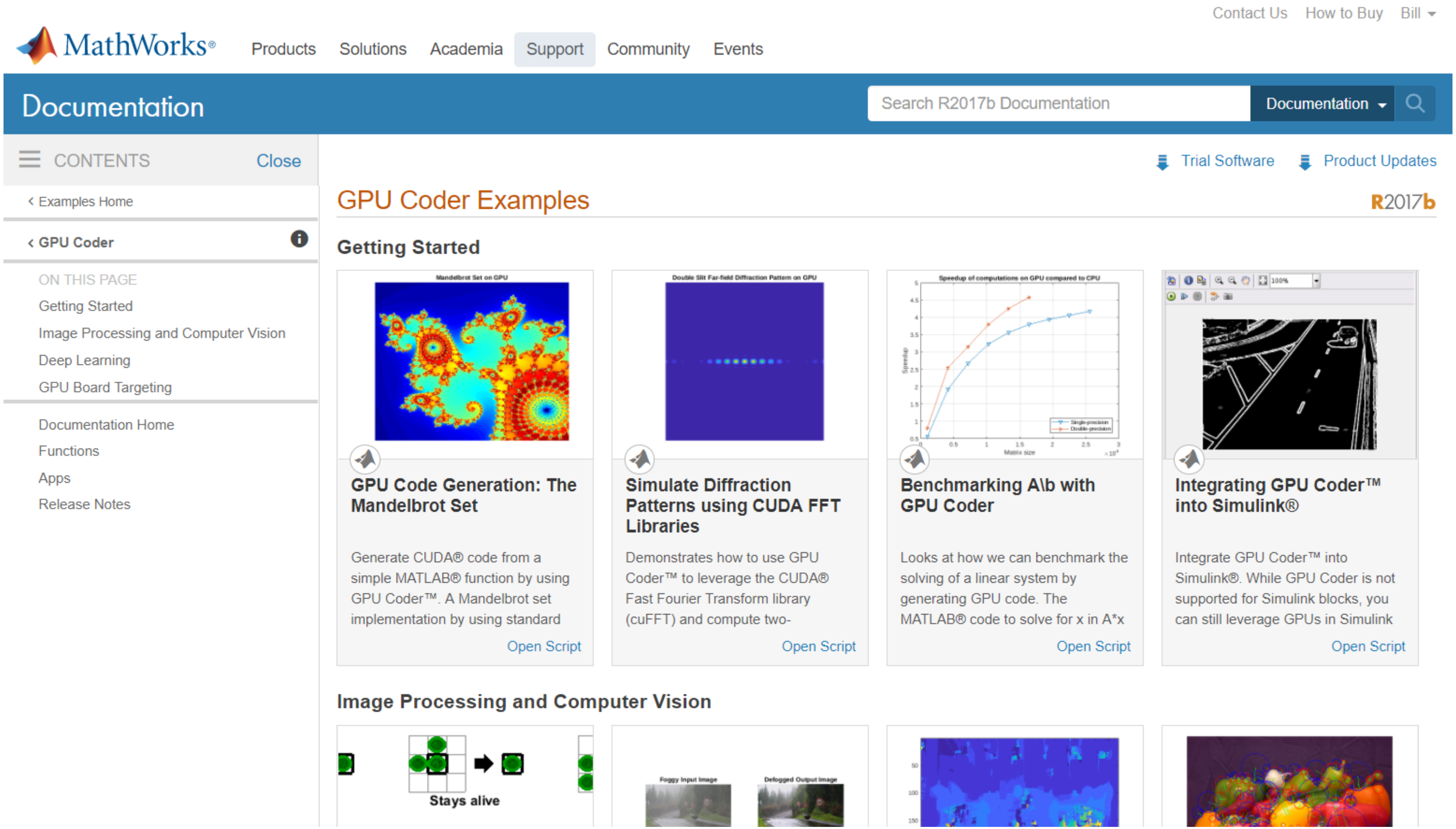
- Tensorflow has the Python overhead

- GPU Coder generates code for the ***specific*** DNN with ***specific*** data types
  - Much less overhead

Additional Features: Optimizations for CUDA Code

- NVIDIA accelerated library support:
  - **cuSolver**: Dense and sparse direct solvers to accelerate computer vision and linear optimization applications
  - **cuFFT**: High-performance computation of FFTs
  - **cuBLAS**: GPU-accelerated implementation of the standard BLAS
  - **cuDNN**: GPU-accelerated library of primitives for deep neural networks

Lots of Examples to Get Started



Contact Us How to Buy Bill

MathWorks® Products Solutions Academia Support Community Events

Documentation

Search R2017b Documentation

Documentation

CONTENTS Close

‹ Examples Home

‹ GPU Coder

ON THIS PAGE

- Getting Started
- Image Processing and Computer Vision
- Deep Learning
- GPU Board Targeting

- Documentation Home
- Functions
- Apps
- Release Notes

Trial Software Product Updates

GPU Coder Examples

R2017b

Getting Started

GPU Code Generation: The Mandelbrot Set

Generate CUDA® code from a simple MATLAB® function by using GPU Coder™. A Mandelbrot set implementation by using standard

Open Script

Simulate Diffraction Patterns using CUDA FFT Libraries

Demonstrates how to use GPU Coder™ to leverage the CUDA® Fast Fourier Transform library (cuFFT) and compute two-

Open Script

Benchmarking A\b with GPU Coder

Looks at how we can benchmark the solving of a linear system by generating GPU code. The MATLAB® code to solve for x in A*x

Open Script

Integrating GPU Coder™ into Simulink®

Integrate GPU Coder™ into Simulink®. While GPU Coder is not supported for Simulink blocks, you can still leverage GPUs in Simulink

Open Script

Image Processing and Computer Vision

- Easily target Jetson TX 2 from MATLAB
- Best in class performance for deep learning

Come See the Demo Live !
Sign Up for 50% Discount on Jetson TX2