Training and Deployment Platforms and Packaging

Graphics Processing Units (GPU)

- Mainly developed for graphics rendering and gaming.
- They are found in PCs, game consoles and mobile phones, tablets, embedded systems and cars.

	Q2'20	Q1'21	Q2'21
AMD	20%	19%	17%
Nvidia	80%	81%	83%

PC Discrete GPU shipment market shares

Graphics Processing Units (GPU)

GPUs are fast…

GTX 285 has 240 cores, 1 TFLOPS

GTX 480: 1345 GFLOPS 250W, March 2010

GTX 590: 2488 GFLOPS 244W, March 2011

GTX 680: 3090 GFLOPS 195W, March 2012

GTX 780Ti: 5046 GFLOPS 250W, November 2013 (649$)

GTX 980: 4612 GFLOPS , 165W, September 2014 (549$) (Later: 5632 GLOPS, 250W)

GTX 980 notebook: 4612 GFLOPS, 145 W, September 2015

GTX 1080: 9 TFLOPS, 180W, May 2016 (599$)

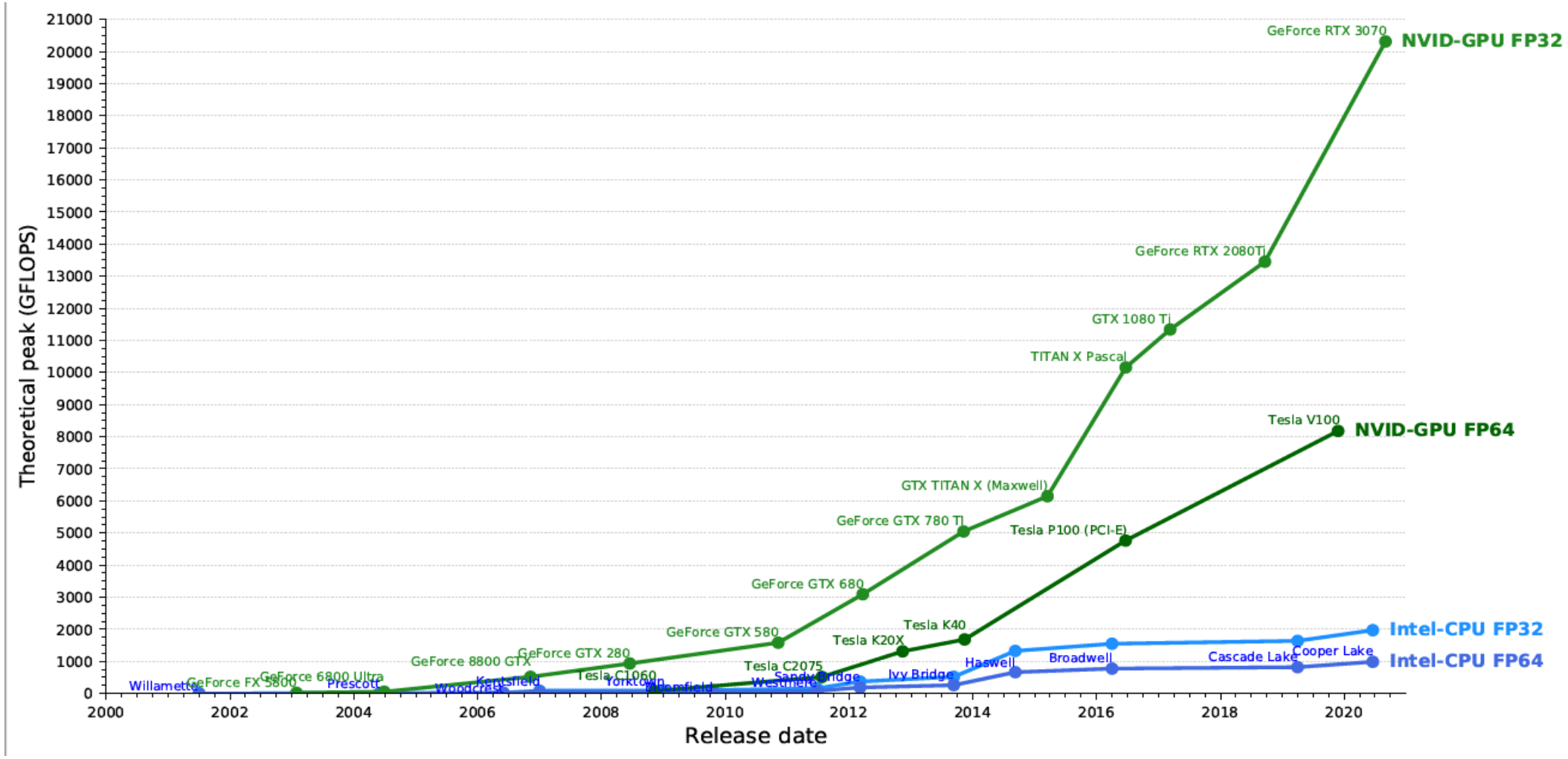
GTX 1080TI: 11.3 TFLOPS, 250W March 2017 (699$)

RTX 2080TI: 13.4 TFLOPS, 250W, Sept 2018 (999$)

RTX 3080: 29.8 TFLOPS, 320W, Sept 2020 (700$)

Note: Intel Core i7-8700K 6-core CPU has a performance of 218 GFLOPS @95W

Graphics Processing Units (GPU)



Training Hardware- NVIDIA

FULLY INTERGRATED DL SUPERCOMPUTER

DGX-1 & DGX Station

DESKTOP

RTX/GTX series

DATA CENTER

Tesla A100
Tesla V100

INFERENCE

DATA CENTER

Tesla A100/V100

Tesla T4

AUTOMOTIVE

Drive PX2

EMBEDDED

Jetson TX2

Jetson Xavier

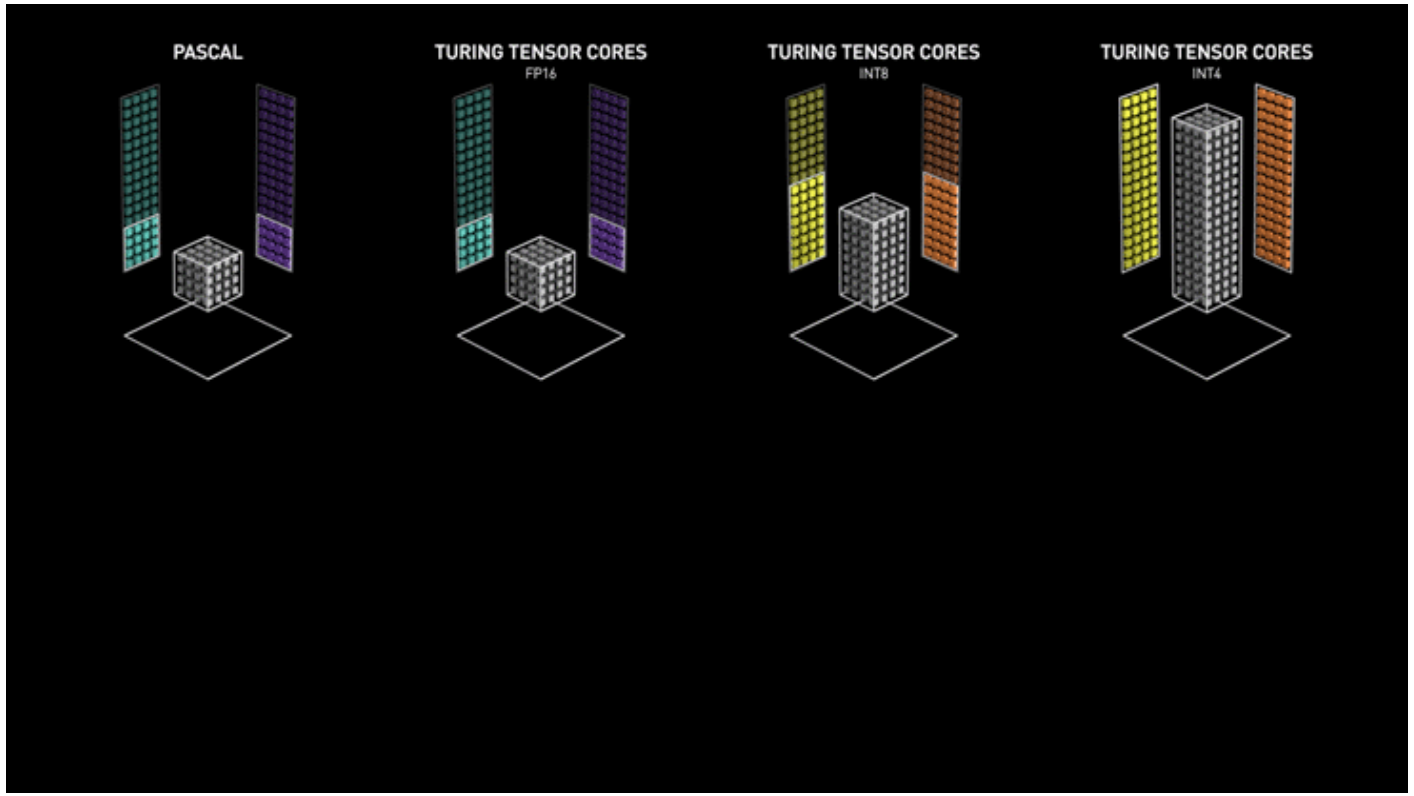
Training Hardware- Goggle

Google Cloud Tensor Processing Units (TPUs)

- TPU's are Application Specific Integrated Circuits (ASIC) designed to accelerate machine learning algorithms
- They basically accelerate linear algebra calculations and can be used to train models using Tensorflow

NVIDIA Tensor Cores

- Tensor Cores are programmable fused matrix-multiply-and-accumulate units that execute concurrently alongside the CUDA cores.
- Tensor Cores implement floating point HMMA (Half Precision Matrix Multiply and Accumulate) and IMMA (Integer Matrix Multiple and Accumulate) instructions for accelerating dense linear algebra computations.



NVIDIA Tensor Cores

- Tensor Cores are already supported for deep learning frameworks: TensorFlow, PyTorch, MXNet, and Caffe2.
- Mixed-Precision Training Guide gives more information about enabling Tensor Cores when using these frameworks
- TensorRT 3.0 release also supports Tensor Cores for deep learning inference.
- cuBLAS uses Tensor Cores to speed up GEMM computations (GEMM is the BLAS term for a matrix-matrix multiplication)
- cuDNN uses Tensor Cores to speed up both convolutions and recurrent neural networks (RNNs).

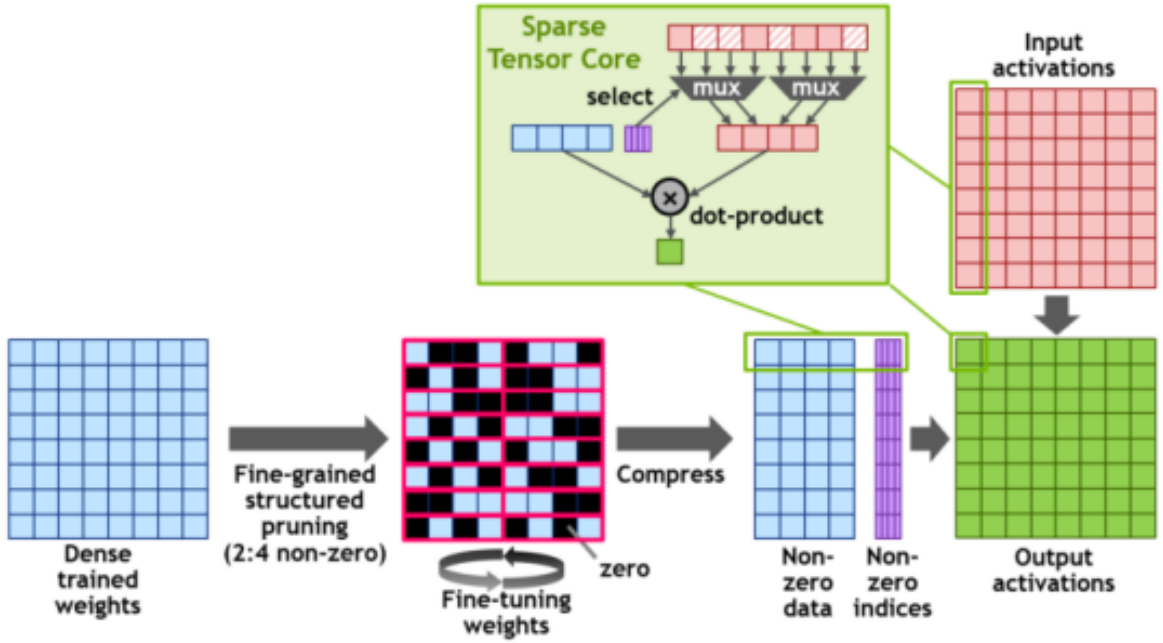
NVIDIA Tensor Cores – 3rd Gen

- The NVIDIA Ampere brings support for new precisions—Tensor Float 32 (TF32) and floating point 64 (FP64).
- TF32 works just like FP32 while delivering speedups of up to 20X for AI without requiring any code change.
- Automatic Mixed Precision can provide an additional 2X performance with automatic mixed precision and FP16 by adding just a couple of lines of code.
- And with support for bfloat16, INT8, and INT4, Tensor Cores in NVIDIA Ampere architecture Tensor Core GPUs create an incredibly versatile accelerator for both AI training and inference.
- HPC grade devices A100 and A30 GPUs also enable matrix operations in full FP64 precision.

NVIDIA Tensor Cores – 3rd Gen

- Ampere architecture brings support sparsity feature.
- Sparsity feature takes advantage of the fine-grained structured sparsity in deep learning networks to double the throughput of the Tensor Core operations.
- Sparsity is constrained to 2 out of every 4 weights being nonzero. It enables the tensor core to skip zero values, doubles the throughput, and reduces the memory storage significantly.
- Networks can be trained first on dense weights, and then pruned, and later fine-tuned on sparse weights.

Figure 5: Ampere GPU 3rd Generation Tensor Core Sparsity



Deep Learning Hardware

TRAINING

FULLY INTERGRATED DL SUPERCOMPUTER

DGX-1 & DGX Station

DESKTOP

RTX/GTX series

DATA CENTER

Tesla A100
Tesla V100

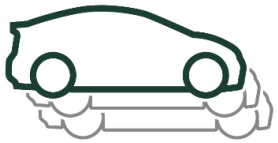
INFERENCE

DATA CENTER

Tesla A100/V100

Tesla T4

AUTOMOTIVE



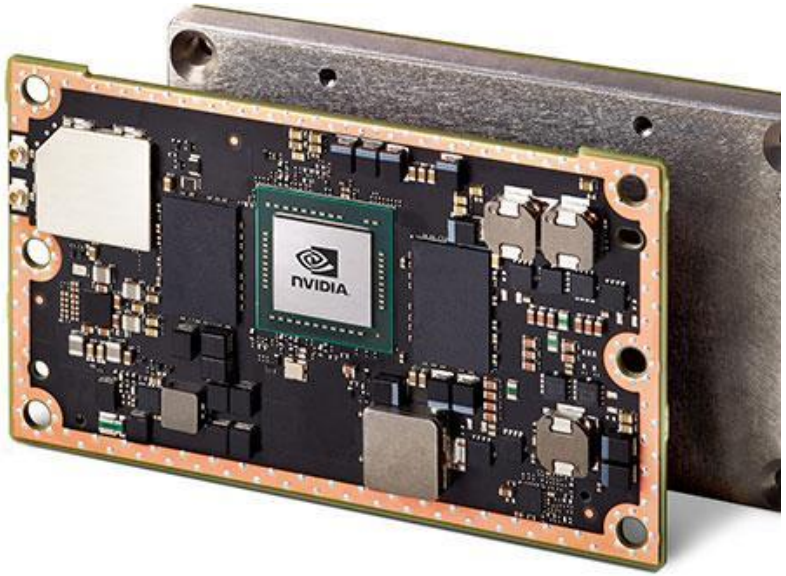
Drive PX2

EMBEDDED

Jetson TX2

Jetson Xavier

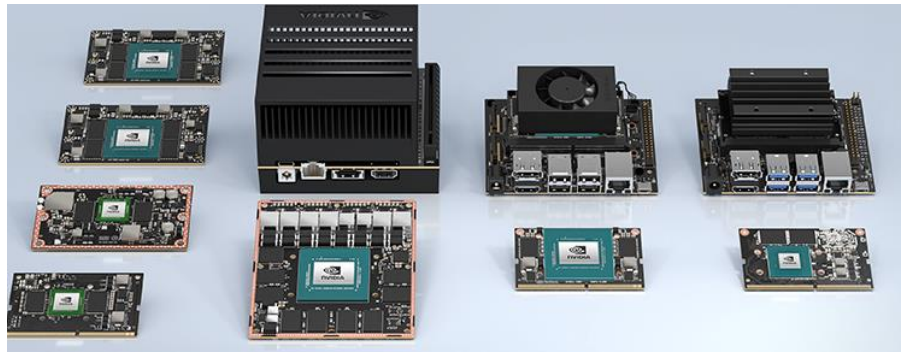
NVIDIA Jetson



	Jetson TX2	Jetson TX1
GPU	NVIDIA Pascal™, 256 CUDA cores	NVIDIA Maxwell ™, 256 CUDA cores
CPU	HMP Dual Denver 2/2 MB L2 + Quad ARM® A57/2 MB L2	Quad ARM® A57/2 MB L2
Video	4K x 2K 60 Hz Encode (HEVC) 4K x 2K 60 Hz Decode (12-Bit Support)	4K x 2K 30 Hz Encode (HEVC) 4K x 2K 60 Hz Decode (10-Bit Support)
Memory	8 GB 128 bit LPDDR4 59.7 GB/s	4 GB 64 bit LPDDR4 25.6 GB/s
Display	2x DSI, 2x DP 1.2 / HDMI 2.0 / eDP 1.4	2x DSI, 1x eDP 1.4 / DP 1.2 / HDMI
CSI	Up to 6 Cameras (2 Lane) CSI2 D-PHY 1.2 (2.5 Gbps/Lane)	Up to 6 Cameras (2 Lane) CSI2 D-PHY 1.1 (1.5 Gbps/Lane)
PCIE	Gen 2 | 1x4 + 1x1 OR 2x1 + 1x2	Gen 2 | 1x4 + 1x1
Data Storage	32 GB eMMC, SDIO, SATA	16 GB eMMC, SDIO, SATA
Other	CAN, UART, SPI, I2C, I2S, GPIOs	UART, SPI, I2C, I2S, GPIOs
USB	USB 3.0 + USB 2.0	
Connectivity	1 Gigabit Ethernet, 802.11ac WLAN, Bluetooth	
Mechanical	50 mm x 87 mm (400-Pin Compatible Board-to-Board Connector)	

Mobile Devices : NVIDIA Jetson

	Jetson Nano	Jetson TX2 Series				Jetson Xavier NX Series		Jetson AGX Xavier Series			Jetson Orin NX	Jetson AGX Orin
		TX2 NX	TX2 4GB	TX2	TX2i	Jetson Xavier NX 16GB	Jetson Xavier NX	Jetson AGX Xavier 64GB	Jetson AGX Xavier	Jetson AGX Xavier Industrial		
AI Performance	472 GFLOPS	1.33 TFLOPS			1.26 TFLOPS	21 TOPS		32 TOPS		30 TOPS	100 TOPS	200 TOPS
GPU	128-core NVIDIA Maxwell™ GPU	256-core NVIDIA Pascal™ GPU				384-core NVIDIA Volta™ GPU with 48 Tensor Cores		512-core NVIDIA Volta GPU with 64 Tensor Cores			1024-core NVIDIA Ampere GPU with 32 Tensor Cores	2048-core NVIDIA Ampere GPU with 64 Tensor Cores
CPU	Quad-core ARM® Cortex®-A57 MPCore processor	Dual-core Denver 2 64-bit CPU and quad-core Arm Cortex-A57 MPCore processor				6-core NVIDIA Carmel Arm®v8.2 64-bit CPU 6MB L2 + 4MB L3		8-core NVIDIA Carmel Arm®v8.2 64-bit CPU 8MB L2 + 4MB L3			8-core NVIDIA Arm® Cortex A78AE v8.2 64-bit CPU 2MB L2 + 6MB L3	12-core NVIDIA Arm® Cortex A78AE v8.2 64-bit CPU 3MB L2 + 6MB L3
DL Accelerator	-	-				2x NVDLA		2x NVDLA			2x NVDLA v2	2x NVDLA v2
Vision Accelerator	-	-				2x PVA		2x PVA			1 x PVA v2	1 x PVA v2
Safety Cluster Engine	-	-				-		-		2x Arm Cortex-R5 in lockstep	-	-
Memory	4GB 64-bit LPDDR4 25.6GB/s	4GB 128-bit LPDDR4 51.2GB/s		8GB 128-bit LPDDR4 59.7GB/s	8GB 128-bit LPDDR4 (ECC Support) 51.2GB/s	16GB 128-bit LPDDR4x 59.7GB/s	8GB 128-bit LPDDR4x 59.7GB/s	64GB 256-bit LPDDR4x 136.5GB/s	32GB 256-bit LPDDR4x 136.5GB/s	32GB 256-bit LPDDR4x (ECC support) 136.5GB/s	12GB 128-bit LPDDR5 102.4 GB/s	32GB 256-bit LPDDR5 204.8 GB/s
Storage	16GB eMMC 5.1	16GB eMMC 5.1		32GB eMMC 5.1	32GB eMMC 5.1	16GB eMMC 5.1		32GB eMMC 5.1		64GB eMMC 5.1	- (Supports external NVMe)	64GB eMMC 5.1



Embedded Devices: Jetson AGX Orin

Figure: 1 Jetson AGX Orin delivers 6x the AI performance of Jetson AGX Xavier

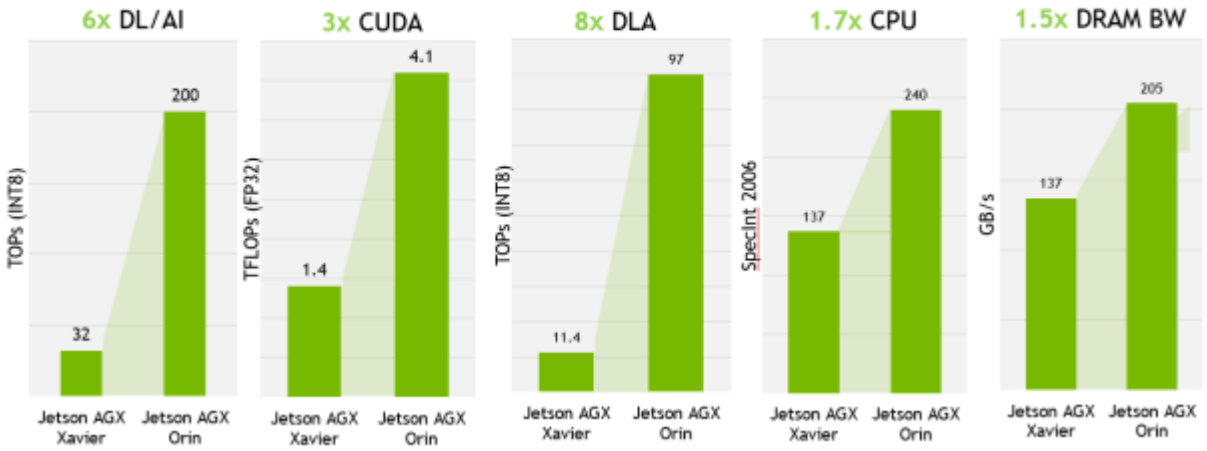
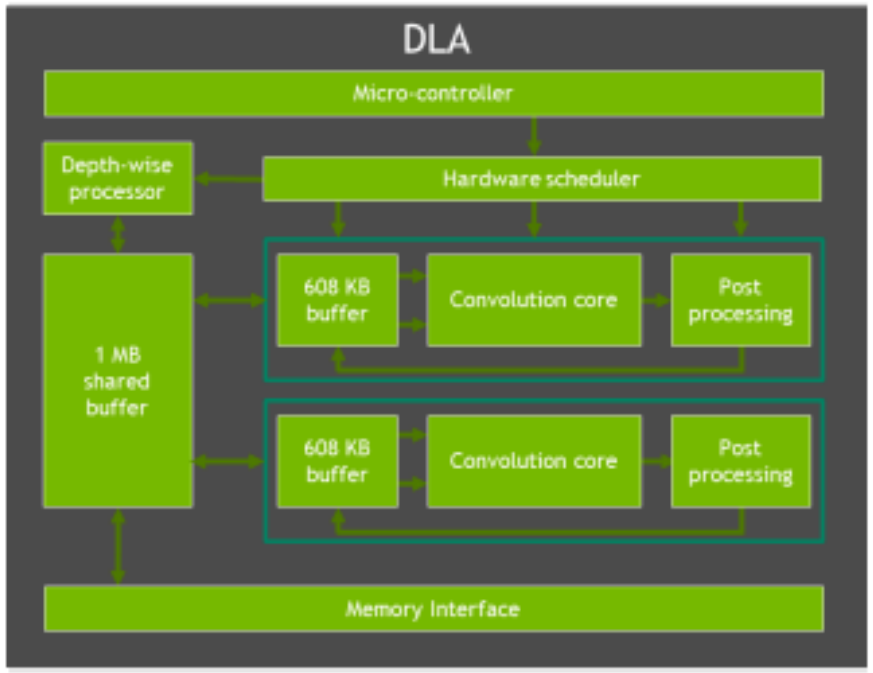


Figure 13: Jetson AGX Orin Developer Kit

Embedded Devices: Jetson AGX Orin

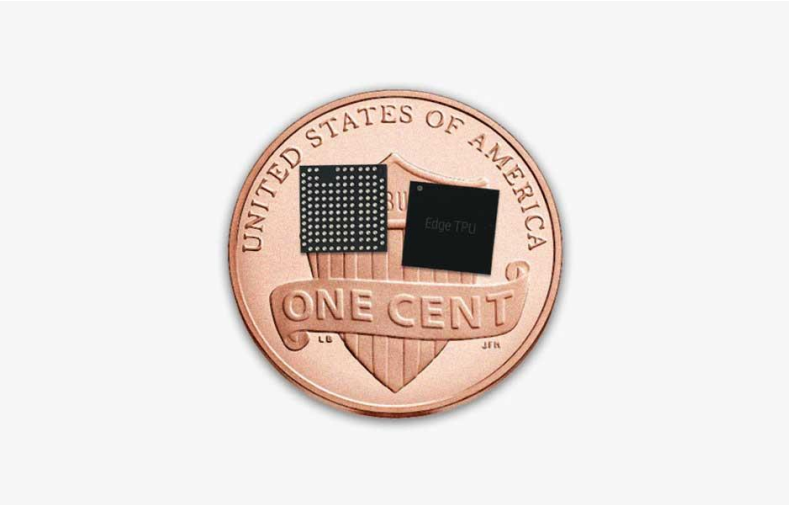
- The deep learning accelerator, or DLA, is a fixed-function accelerator optimized for deep learning operations.
- It is designed to do full hardware acceleration of convolutional neural network inferencing.
- The Orin SoC brings support for the next generation DLA, NVDLA 2.0

Figure 6: Orin Deep Learning Accelerator (DLA) Block Diagram



CNNs on Mobile Devices – Edge TPU

- Edge TPU: a small ASIC designed by Google that provides high performance ML inferencing for low-power devices.
- It can execute state-of-the-art mobile vision models such as MobileNet V2 at 100+ fps, in a power efficient manner.
- Supports Tensorflow Lite



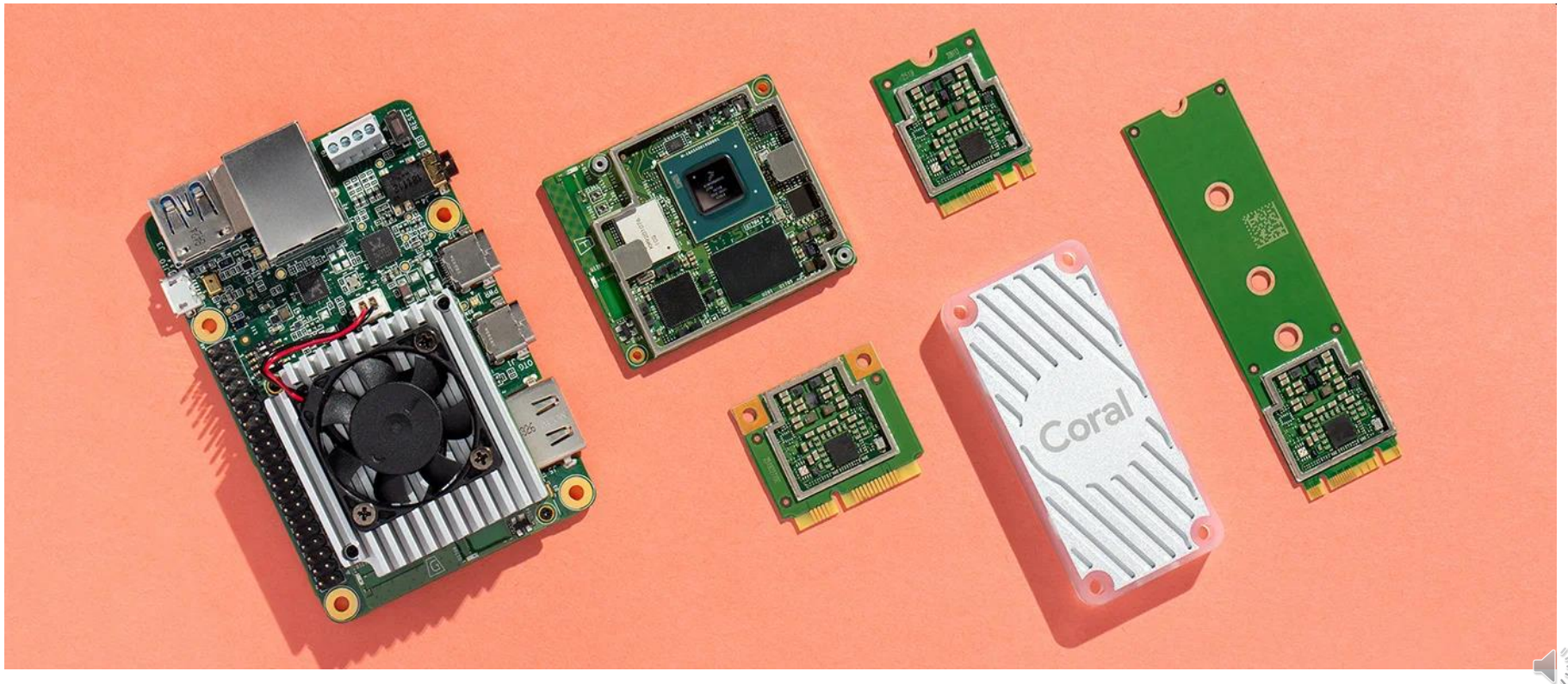
Two Edge TPU chips on the head of a US penny

USB Accelerator with Edge TPU

https://coral.withgoogle.com/tutorials/edgetpu-faq/

Edge TPU – Coral Toolkit

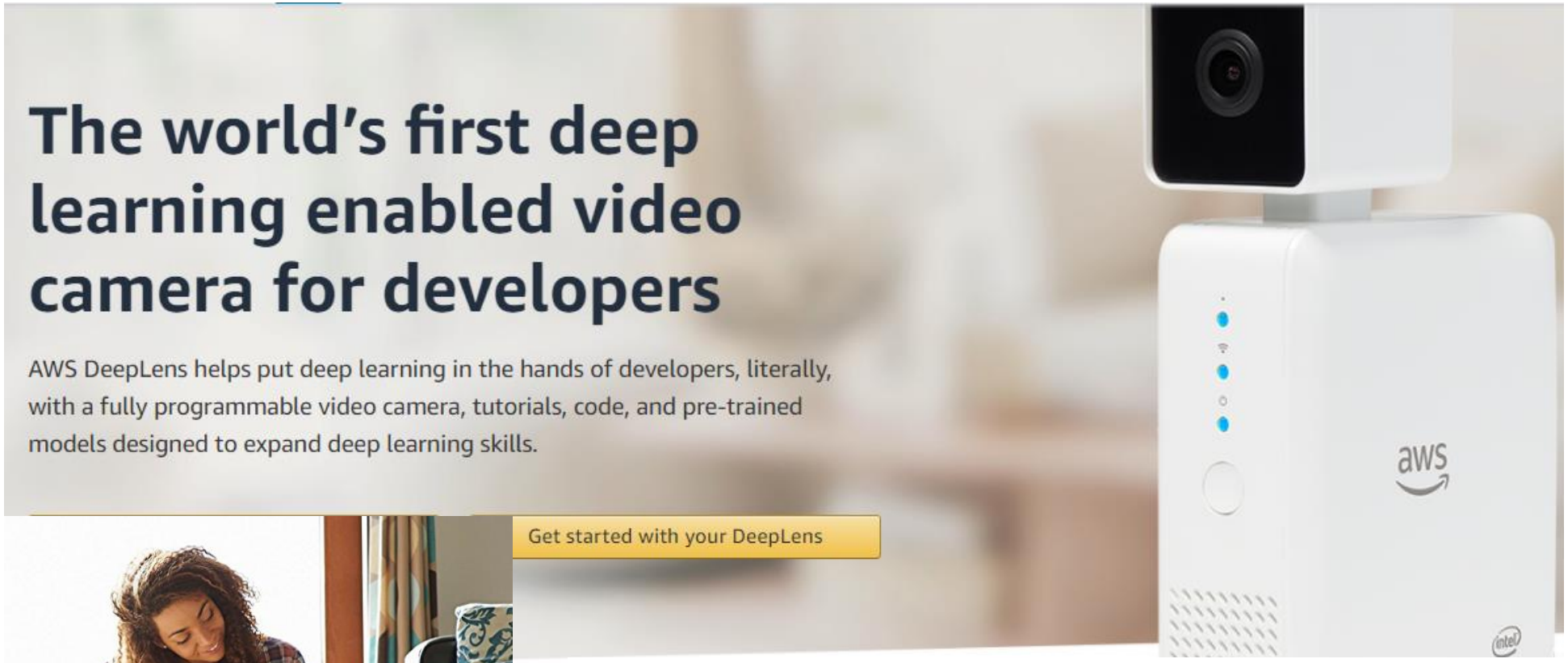
– Coral is a complete toolkit to build products with local AI.
– Prototyping devices include a single-board computer and USB accessory
– Production-ready devices include a system-on-module and PCIe module.



Edge TPU – Intel® Movidius™

- Inference Accelerator

Edge AI– AWS Deeplens



ACTIVITY RECOGNITION

Recognize more than 30 kinds of actions such as brushing teeth, applying lipstick, and playing guitar.

AWS DeepLens - Deep learning enabled video camera for developers

by Amazon Web Services

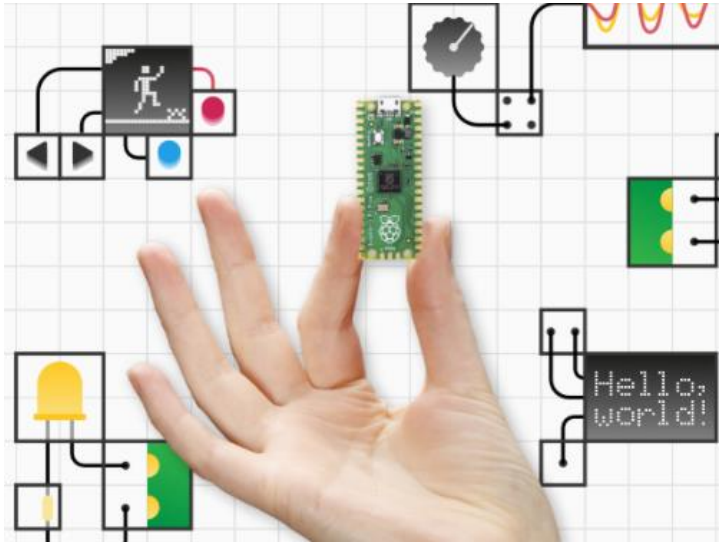
★★☆☆☆ 27 customer reviews | 27 answered questions

Price: $249.00 & **FREE Shipping**. Details

prime | Try Fast, Free Shipping

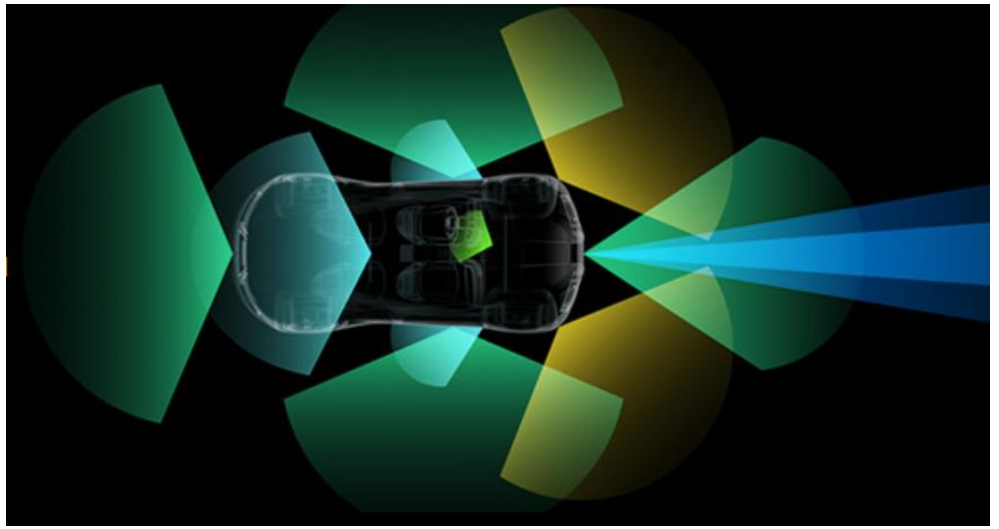
IoT Devices – Raspberry Pi

- Raspberry Pi Pico is a highly cost effective product and supports TensorFlow Lite
- Dual-core Arm Cortex-M0+ CPU
- 264KB RAM
- Up to 16MB off-chip flash memory support.



Deep Learning for Autonomous Driving

NVIDIA DRIVE® embedded supercomputing platforms process data from camera, radar, and lidar sensors to perceive the surrounding environment, localize the car to a map, and plan and execute a safe path forward. This AI platform supports autonomous driving, in-cabin functions and driver monitoring, and other safety features—all in a compact, energy-efficient package.



ODTÜ METU

Deep Learning for Autonomous Driving

NVIDIA DRIVE Orin

The NVIDIA DRIVE Orin™ SoC (system on a chip) delivers 254 TOPS and is the central computer for intelligent vehicles. It's the ideal solution for powering autonomous driving capabilities, confidence views, digital clusters, infotainment, and passenger interaction with AI. The scalable DRIVE Orin product family lets developers build, scale, and leverage one development investment across an entire fleet, from Level 2+ systems all the way to Level 5 fully autonomous vehicles.

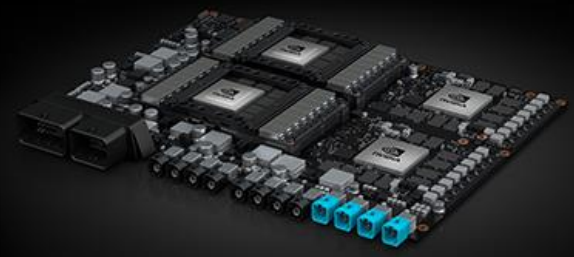
Learn more about the DRIVE Orin SoC ›

Learn more about our DRIVE AGX Orin development platform ›

NVIDIA DRIVE AGX Pegasus

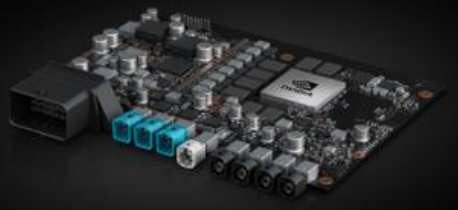
NVIDIA DRIVE AGX Pegasus™ uses the power of two Xavier SoCs and two NVIDIA Turing™ GPUs to achieve an unprecedented 320 TOPS of supercompute capability. The platform is designed and built for Level 4 and Level 5 autonomous systems, including robotaxis.

Learn more about the DRIVE AGX Pegasus Developer Kit ›



ODTÜ METU

Deep Learning for Autonomous Driving



NVIDIA DRIVE AGX Xavier

NVIDIA DRIVE AGX Xavier™ delivers an incredible 30 TOPS for Level 2+ and Level 3 automated driving. At its core is the first-ever production auto-grade Xavier SoC, which incorporates six different types of processors, including a CPU, GPU, Deep Learning Accelerator (DLA), Programmable Vision Accelerator (PVA), Image Signal Processor (ISP), and stereo/optical flow accelerator.

Learn more about the DRIVE AGX Xavier Developer Kit ›

NVIDIA DRIVE Hyperion

NVIDIA DRIVE Hyperion™ is a reference and testing platform for autonomous vehicles. It consists of a complete sensor suite—including 12 exterior cameras, three interior cameras, nine radars, and two lidar sensors—and the Orin-based AI computing platform. Plus, it features the full software stack for autonomous driving, driver monitoring, and visualization. DRIVE Hyperion can be integrated into a test vehicle, letting you evaluate DRIVE AV software and perform data collection for your autonomous vehicle fleet. Software is updated to DRIVE Hyperion using the NVIDIA DRIVE over-the-air update infrastructure and services.

Learn more about Hyperion 8 ›

ODTÜ METU

Power Efficiency on Embedded Devices

- Jetson TX2 was designed for peak processing efficiency at 7.5W of power. This level of performance, referred to as Max-Q, represents the peak of the power/throughput curve.
- Every component on the module including the power supply is optimized to provide highest efficiency at this point.

- The Max-Q frequency for the GPU is 854 MHz, and for the ARM A57 CPUs it’s 1.2 GHz.

- The L4T BSP in JetPack 3.0 includes preset platform configurations for setting Jetson TX2 in Max-Q mode.
- JetPack 3.0 also includes a new command line tool called nvpmodel for switching profiles at run time.

Power Efficiency on Embedded Devices

- While Dynamic Voltage and Frequency Scaling (DVFS) permits Jetson TX2's Tegra "Parker" SoC to adjust clock speeds at run time according to user load and power consumption, the Max-Q configuration sets a cap on the clocks to ensure that the application is operating in the most efficient range only.

- Although most platforms with a limited power budget will benefit most from Max-Q behavior, others may prefer maximum clocks to attain peak throughput, albeit with higher power consumption and reduced efficiency.

Power Efficiency on Embedded Devices

- DVFS can be configured to run at a range of other clock speeds, including underclocking and overclocking. Max-P, the other preset platform configuration, enables maximum system performance in less than 15W.

- The Max-P frequency is 1122 MHz for the GPU and 2 GHz for the CPU when either ARM A57 cluster is enabled or Denver 2 cluster is enabled and 1.4 GHz when both the clusters are enabled.

- You can also create custom platform configurations with intermediate frequency targets to allow balancing between peak efficiency and peak performance for your application. Table 2 below shows how the performance increases going from Max-Q to Max-P and the maximum GPU clock frequency while the efficiency gradually reduces.

Power Efficiency on Embedded Devices

		NVIDIA Jetson TX1	NVIDIA Jetson TX2		
		Max Clock (998 MHz)	Max-Q (854 MHz)	max-P (1122 MHz)	Max Clock (1302 MHz)
GoogLeNet batch=2	Perf	141 FPS	138 FPS	176 FPS	201 FPS
	Power (AP+DRAM)	9.14 W	4.8 W	7.1 W	10.1 W
	Efficiency	15.42	28.6	24.8	19.9
GoogLeNet batch=128	Perf	204 FPS	196 FPS	253 FPS	290 FPS
	Power (AP+DRAM)	11.7 W	5.9 W	8.9 W	12.8 W
	Efficiency	17.44	33.2	28.5	22.7
AlexNet batch=2	Perf	164 FPS	178 FPS	222 FPS	250 FPS
	Power (AP+DRAM)	8.5 W	5.6 W	7.8 W	10.7 W
	Efficiency	19.3	32	28.3	23.3
AlexNet batch=128	Perf	505 FPS	463 FPS	601 FPS	692 FPS
	Power (AP+DRAM)	11.3 W	5.6 W	8.6 W	12.4 W
	Efficiency	44.7	82.7	69.9	55.8

Table 2. Power consumption measurements for GoogLeNet and AlexNet architectures for max-Q and max-P performance levels on NVIDIA Jetson TX1 and Jetson TX2. The table reports energy efficiency for all tests in images per second per Watt consumed.

Packaging: Open Neural Network Exchange (ONNX)

- Training frameworks are not designed for efficient inference.
- Model formats may change in the future.
- Once the training is done, it is desirable that the models are exported in a portable format to use with specialized inference engines.

KEY BENEFITS

Interoperability

Develop in your preferred framework without worrying about downstream inferencing implications. ONNX enables you to use your preferred framework with your chosen inference engine.

SUPPORTED FRAMEWORKS ›

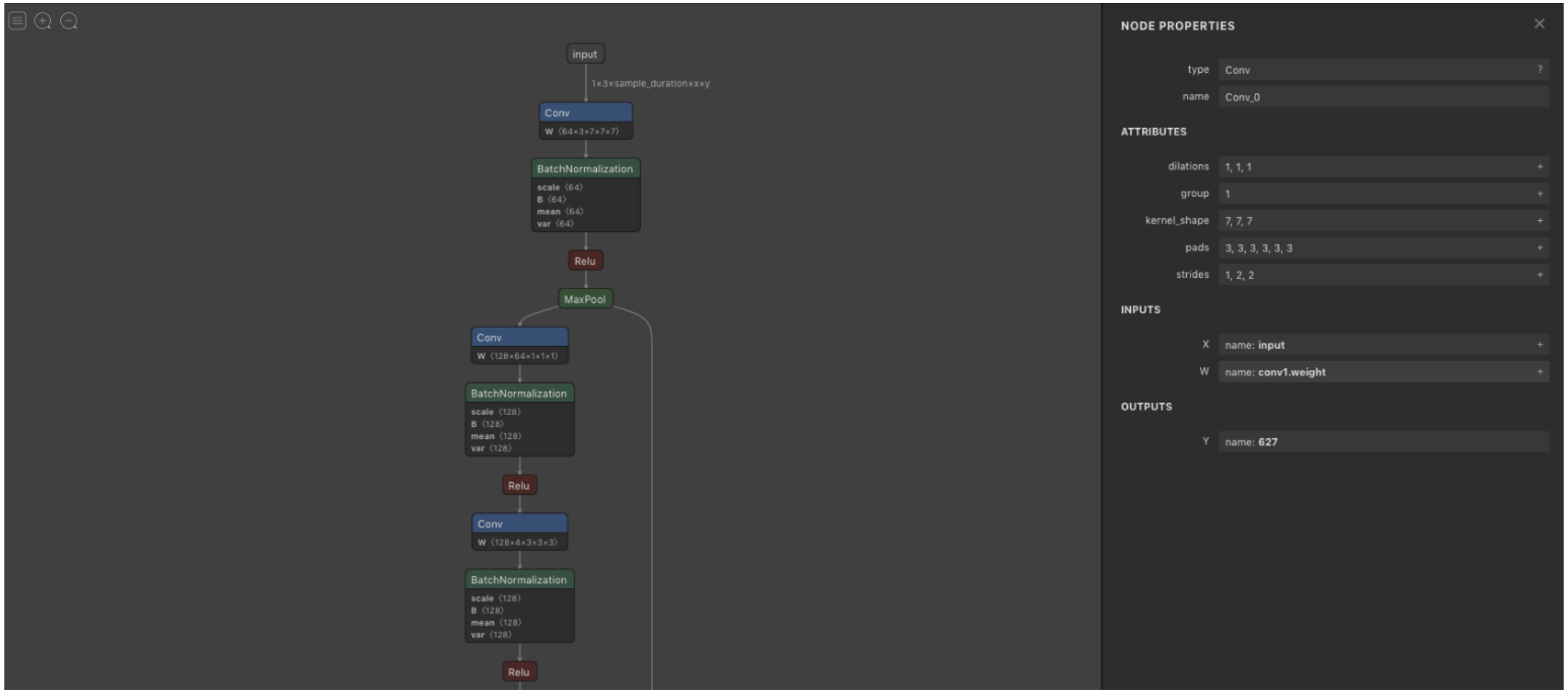
Hardware Access

ONNX makes it easier to access hardware optimizations. Use ONNX-compatible runtimes and libraries designed to maximize performance across hardware.

SUPPORTED ACCELERATORS ›

Packaging: Open Neural Network Exchange (ONNX)

- ONNX provides a definition of an extensible computation graph model.



Packaging: Open Neural Network Exchange (ONNX)

- ONNX stores the data using Protocol Buffer (protobuf); a message gile format developed by Google.
- This format is also used by Tensorflow and Caffe frameworks.
- In protobuf, only the data types such as Float32 and the order of the data are specified, the meaning of each data is left up to the software used.
- ONNX outputs of the frameworks may have redundancies and a simplification step may be used: ONNX Simplifier: https://github.com/daquexian/onnx-simplifier
- ONNX files can be visualized using Netron.

Creating a Demo App - Streamlit

- An app framework built for ML engineers
- Streamlit apps are really scripts that run from top to bottom. There's no hidden state and you can factor your code with function calls.
- There are *no callbacks, e*very interaction simply reruns the script from top to bottom.