NVIDIA'S XAVIER SOC

Michael Ditty, Ashish Karandikar, David Reed

AUTONOMOUS MACHINES

Xavier — Designed for the next wave of Autonomous Machines

XAVIER INNOVATIONS

World's First Autonomous Machines Processor

Carmel CPU	Volta GPU	DLA	PVA	ISP	MM-Accelerators
8 custom cores, ARM v8.2	512 CUDA Tensor Cores 22.6 int8 DL TOPs	int8/int16/FP16 11.4 DL int8 TOPs	7-slot VLIW 1.7 TOPs	Native HDR Processing TNR 2.4 GPIX/sec	Stereo, Optical Flow, LDC

High Speed I/Os : >40GB/s of IO Bandwidth

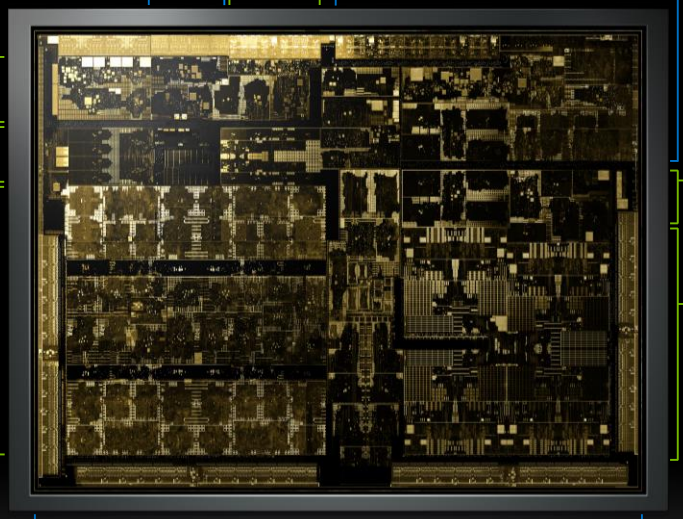
Designed for Safety & Resiliency : ISO26262; ASIL-C

Enhanced Security

Optimized for Energy Efficiency; TSMC 12FFN

XAVIER

World's First Autonomous Machines Processor



16 Lane CSI
109 Gbps CPHY 1.1
1Gb Ethernet

DLA
5.7 TFLOPS FP16
11.4 TOPS INT8

Industry Standard High-Speed IO
PCIe Gen4 Root and Endpoint
USB 3.1 gen2 Host and Device
UFS 2.1 Embedded Storage

Multimedia Engines
1.2 GPIX/s Encode
1.8 GPIX/s Decode
4 GPIX/s Video Image Compositor

Vision Accelerator
1.7 TOPS

Stereo & Optical Flow Engine
2x 3.1 TOPS

Volta Tensor Core GPU
FP32 / FP16 / INT8 Multi-Precision
512 CUDA Tensor Cores
2.8 CUDA TFLOPS (FP16)
22.6 Tensor Core DL TOPS

ISP
2.4 GPIX/s
Native Full-range HDR
Tile-based Processing

Carmel ARM64 CPU
8 Cores
10-wide Superscalar
21 SpecInt2K6 (est.)

256-Bit LPDDR4X
137 GB/s

Most Complex SOC Ever Made | 9 Billion Transistors, 350mm^2, 12FFN | ~8,000 Engineering Years

CARMEL CPU

ARM V8.2 including RAS support

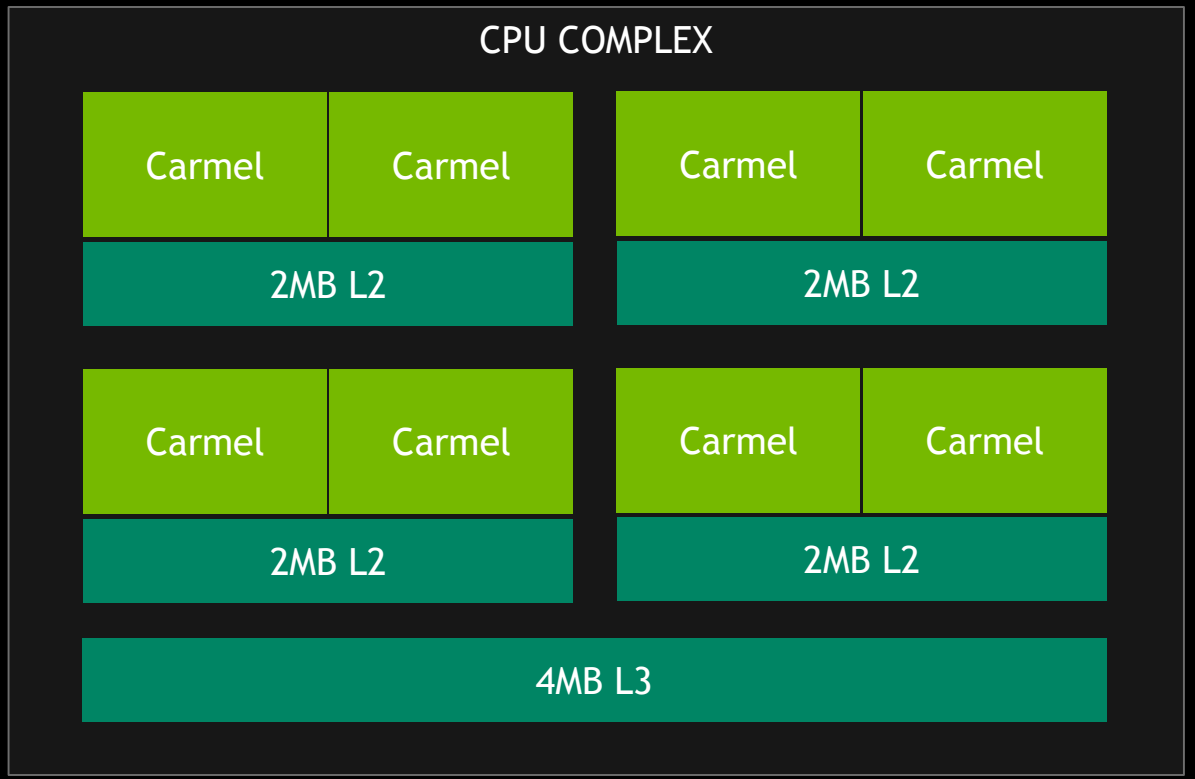
8 NVIDIA Carmel Cores

2 cores + 2MB L2 per cluster

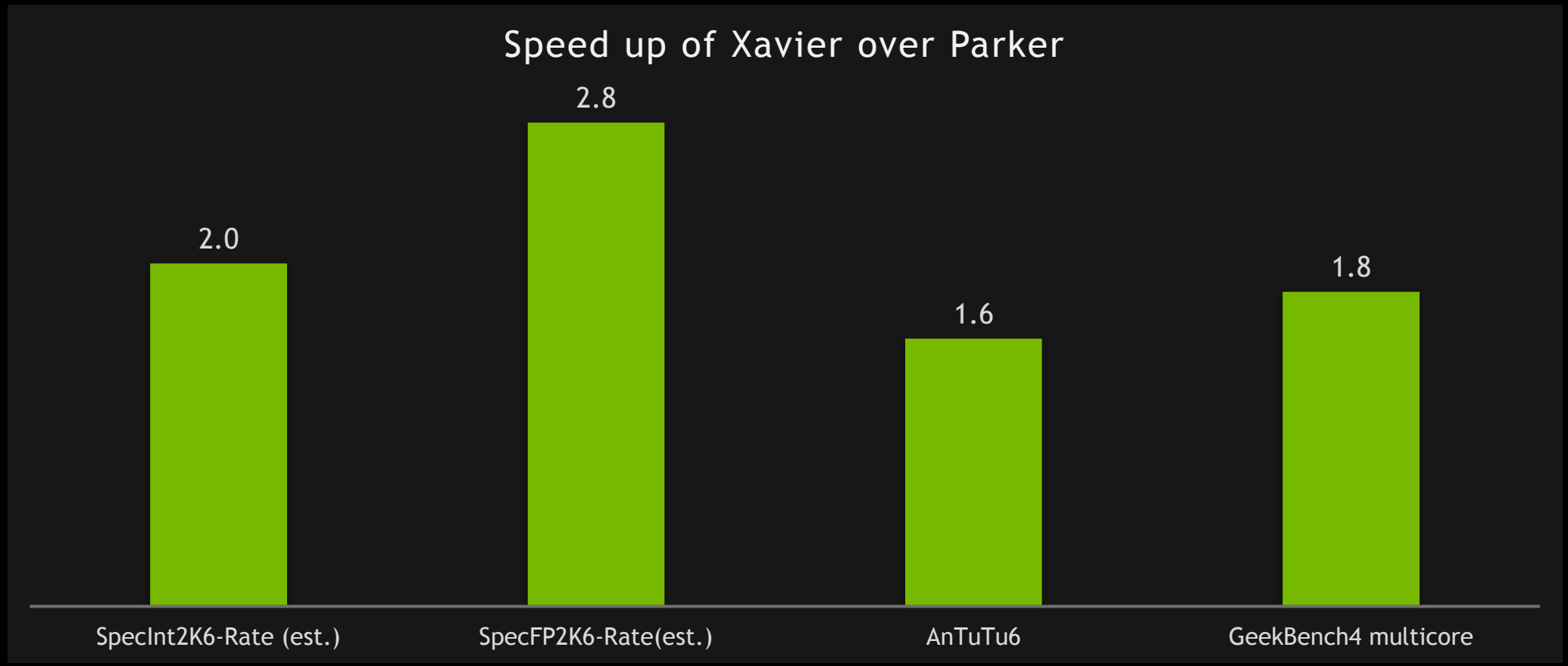
Cache Coherent Across CPU Complex

IO Coherent Memory

4MB Exclusive L3 cache



XAVIER CPU BENCHMARKS



VOLTA GPU

Optimized for Inference

8x Volta SM

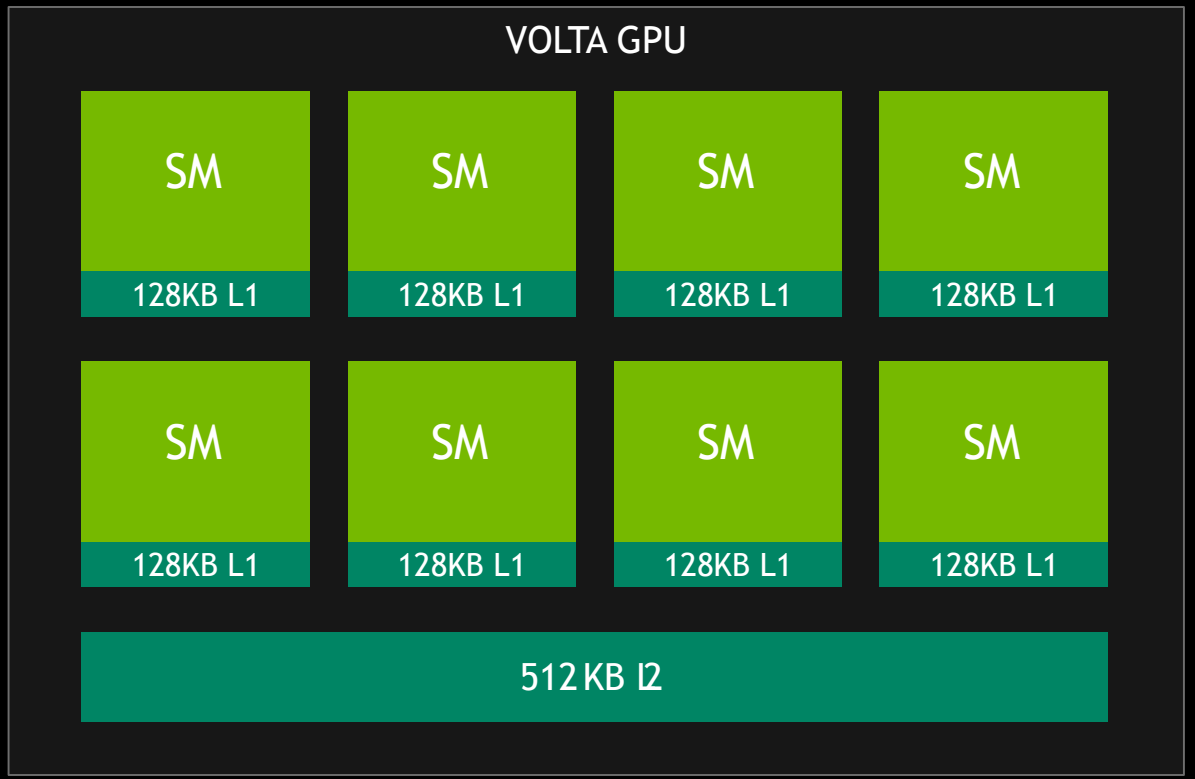
Tensor Cores: fp16, int8

8x Larger L1 cache size

4x faster L2 cache access

22.6 Deep Learning TOPS (int8)

2.1x GFX Performance



DEEP LEARNING ACCELERATOR (DLA)

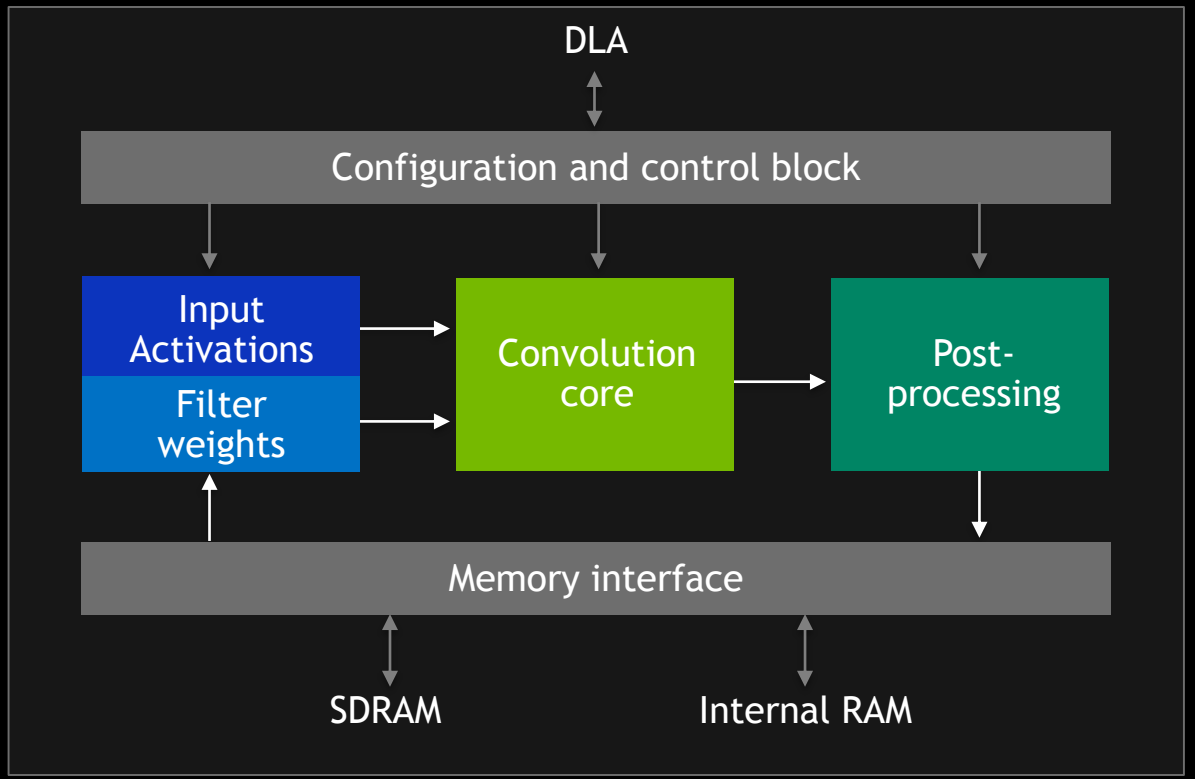
Optimized for perf/mm & power

2x DLA instances

11.4 Deep Learning TOPS (int8)

5.7 Deep Learning TOPS (fp16)

More details in talk tomorrow



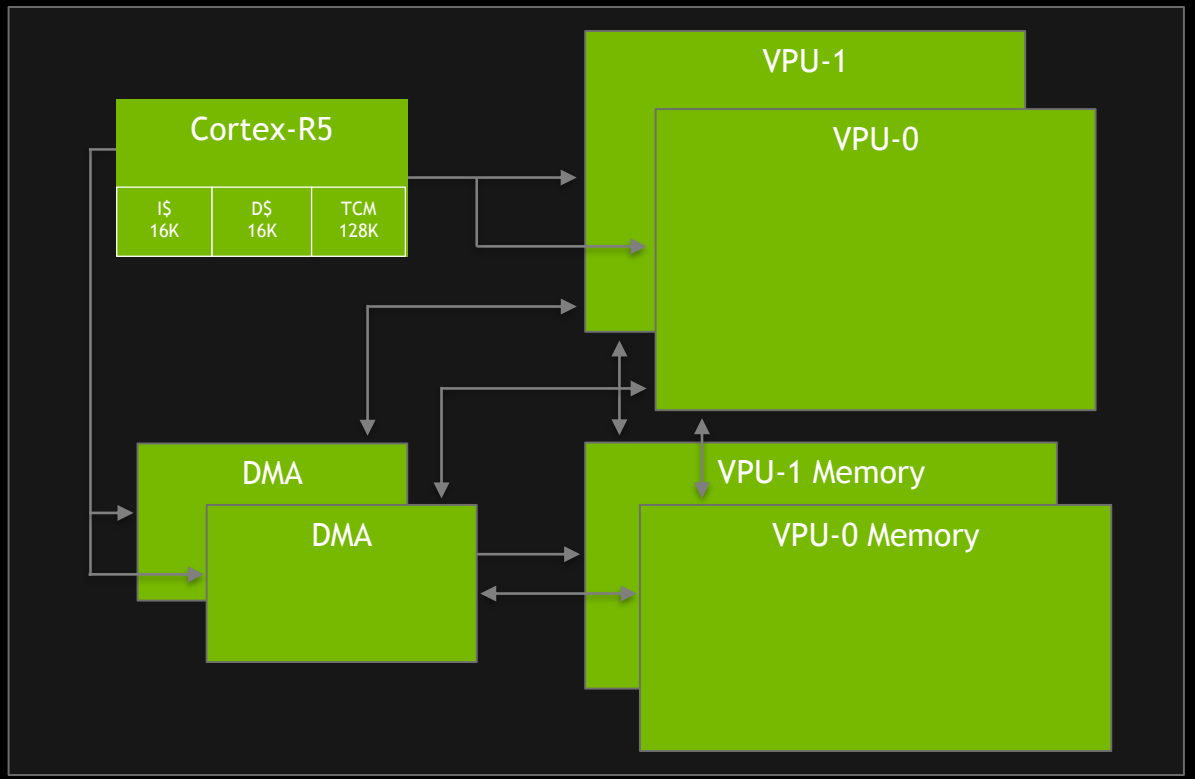
PROGRAMMABLE VISION ACCELERATOR (PVA)

2x PVA

Optimized for imaging & vision algorithms

Each PVA

- Cortex-R5 for config and control
- 2x Vector Processing Units
- 2x DMA for data movement to/from internal/external memories



PVA

Vector Processing Unit (VPU)

7 Slot VLIW architecture

- 2 scalar + 2 vector + 3 memory instructions
- Each vector unit has 32 x 8-bit, 16 x 16-bit, or 8 x 32bit vector math operations
- Additional guard bits for extended precision math
- Table lookup, histogram, vector-addressed store
- Hardware loops and multi-dimensional address generator

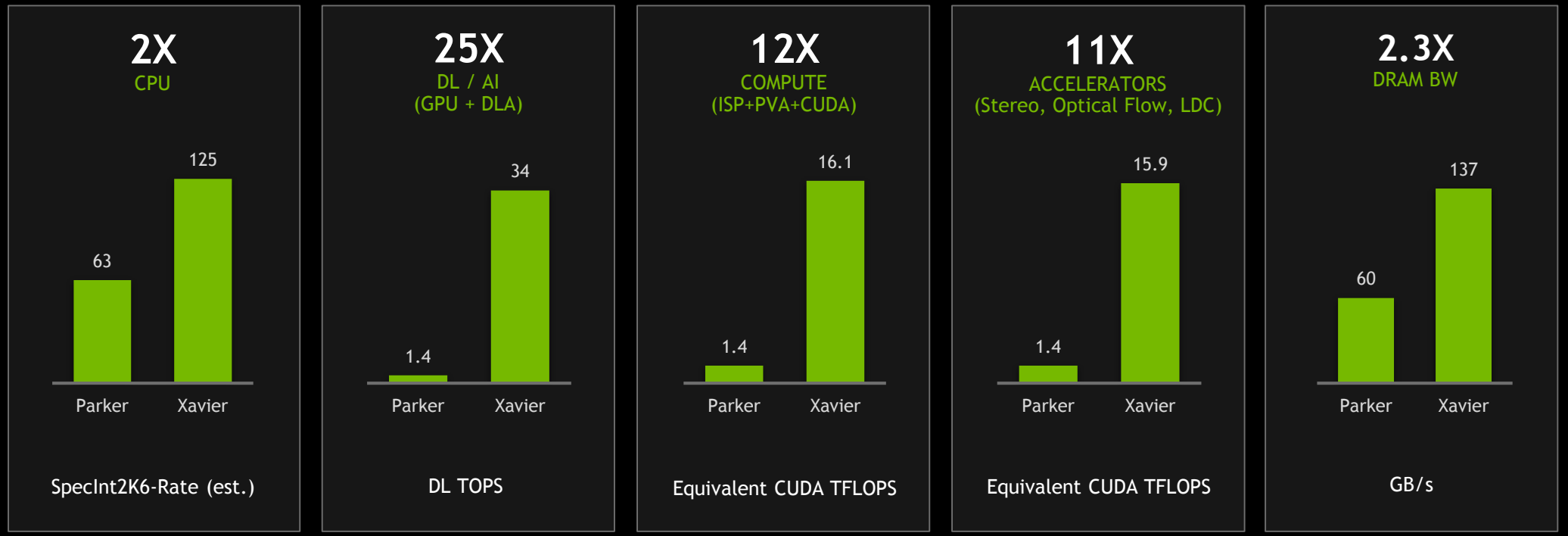
I-cache and local data memory

XAVIER COMPUTER VISION

Multiple Accelerators for Vision Processing

Engine	Function	Description	Throughput
PVA	Vision Accelerator	Computer Vision Algorithms	1.7 CV TOPS
DLA	Deep Learning Accelerator	Inference Engine	2x 5.7 TOPS
GPU	Graphics and Compute	Volta Tensor Core architecture	22.6 DL TOPS 8-bit 2.8 CUDA TFLOPS FP16 1.4 CUDA TFLOPS FP32
SOFE	Stereo & Optical Flow Engine	Dedicated Engines for Stereo & Optical Flow	2x 3.1 TOPS 16 bit
ISP & VIC	HDR and Lens Correction	High dynamic range support, lens distortion correction, temporal noise reduction	2.4 / 4 GPIX/sec

XAVIER 25X AI PERFORMANCE



COMPREHENSIVE HIGH PERFORMANCE I/O SUBSYSTEM

NVLINK

20 GB/s
IO Coherent
Link between Xavier & dGPU

PCIE

Multiple 16GT/s gen4 controllers
x8, x4, x2, x1 configurations
Root port + Endpoint

USB

3x USB3.1 (10 GT/s) ports
4x USB2.0 ports

DISPLAY

4x DP/HDMI/eDP
4K @ 60 Hz
DP HBR3
HDMI 2.0

CAMERA

16 CSI lanes
40 Gbps in DPHY 1.2 Mode
109 Gbps in CPHY 1.1 Mode

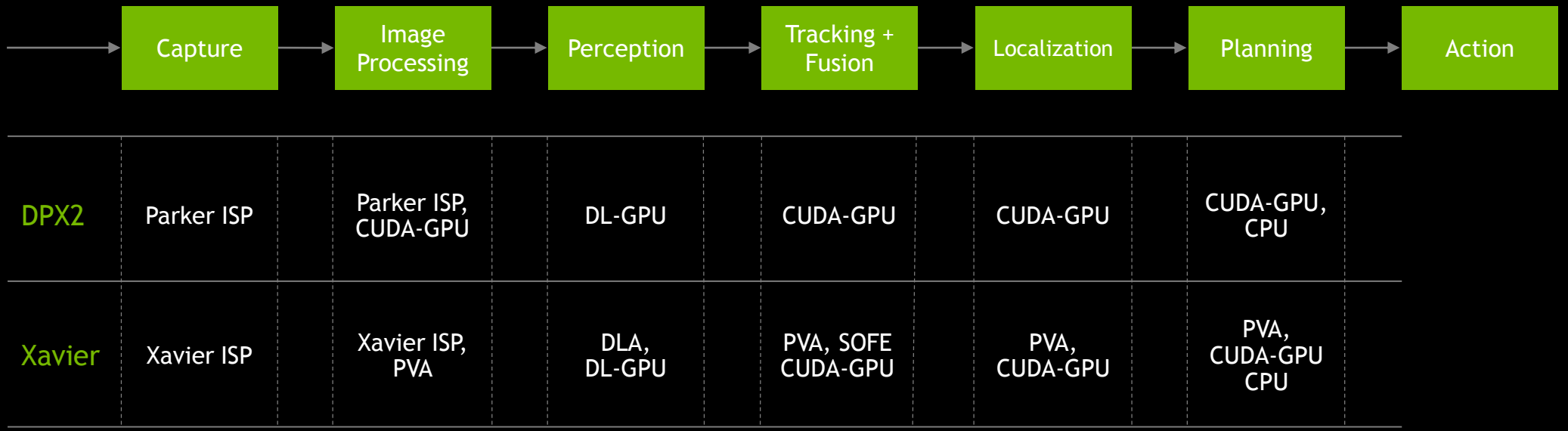
OTHER I/OS

Ethernet UFS SDMMC
CAN SPIO I2C I2S
UART GPIO

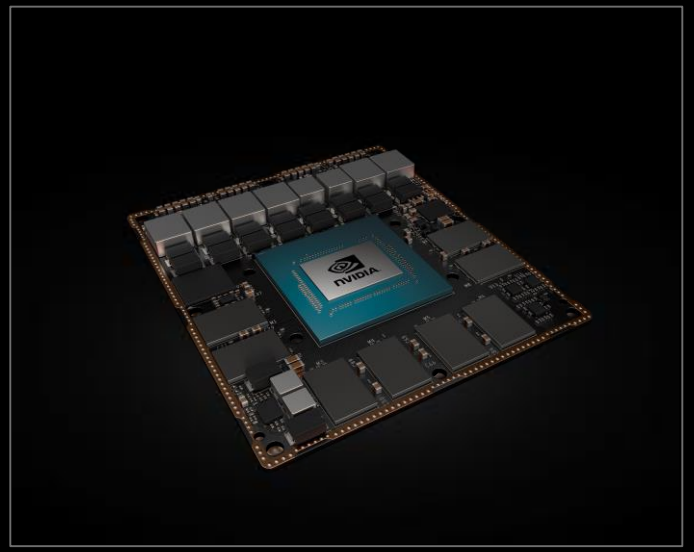
USE CASE COMPARISON

XAVIER : AUTOPILOT USE CASE

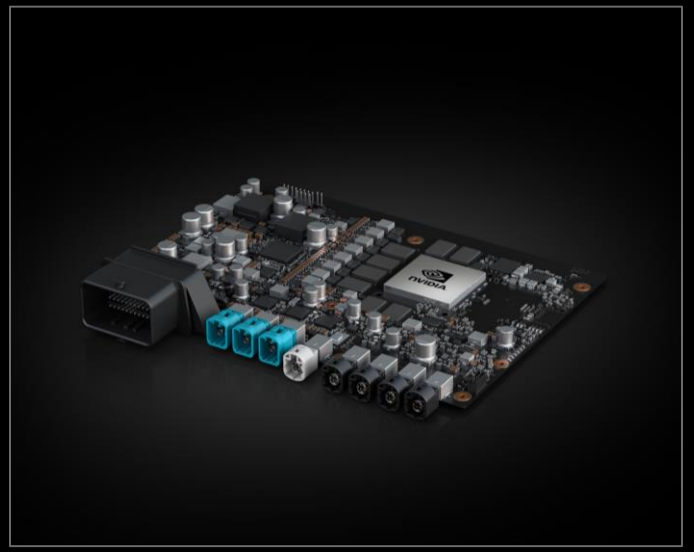
Example of an Autonomous Machine Mapping on Xavier



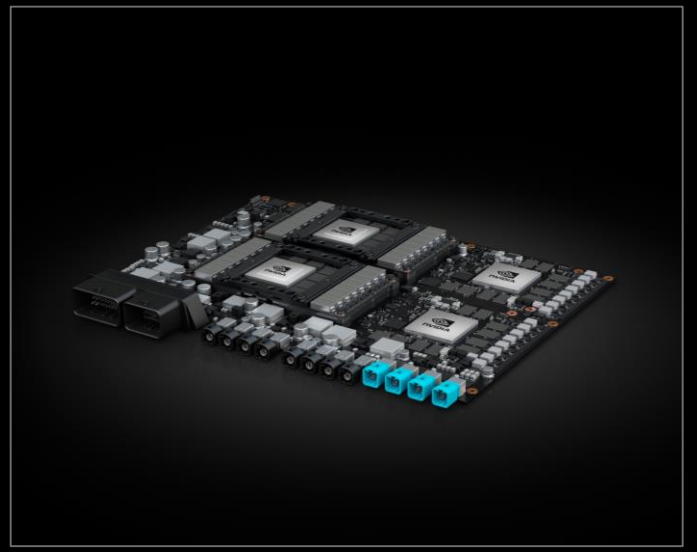
XAVIER



JETSON XAVIER



DRIVE XAVIER



DRIVE PEGASUS


NVIDIA