NVIDIA’S XAVIER SOC

Michael Ditty, Ashish Karandikar, David Reed
AUTONOMOUS MACHINES

Xavier — Designed for the next wave of Autonomous Machines

- CARS
- ROBO-TAXIS
- TRUCKS
- DELIVERY ROBOTS
- FLYING CARS
- MEDICAL INSTRUMENTS
- AGRICULTURE
- PICK-AND-PLACE
- LOGISTICS
- MANUFACTURING
# XAVIER INNOVATIONS

World’s First Autonomous Machines Processor

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

- **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

Volta Tensor Core GPU
- 512 CUDA Tensor Cores
- 2.8 CUDA TFLOPS (FP16)
- 22.6 Tensor Core DL TOPS

FP32 / FP16 / INT8 Multi-Precision

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

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

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

16 Lane CSI
- 109 Gbps CPHY 1.1
- 1Gb Ethernet

DLA
- 5.7 TFLOPS FP16
- 11.4 TOPS INT8

256-Bit LPDDR4X
- 137 GB/s

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

Most Complex SOC Ever Made | 9 Billion Transistors, 350mm², 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

Speed up of Xavier over Parker

- SpecInt2K6-Rate (est.): 2.0
- SpecFP2K6-Rate (est.): 2.8
- AnTuTu6: 1.6
- GeekBench4 multicore: 1.8
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 32-bit 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

<table>
<thead>
<tr>
<th>Engine</th>
<th>Function</th>
<th>Description</th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>PVA</td>
<td>Vision Accelerator</td>
<td>Computer Vision Algorithms</td>
<td>1.7 CV TOPS</td>
</tr>
<tr>
<td>DLA</td>
<td>Deep Learning Accelerator</td>
<td>Inference Engine</td>
<td>2x 5.7 TOPS</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphics and Compute</td>
<td>Volta Tensor Core architecture</td>
<td>22.6 DL TOPS 8-bit</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.8 CUDA TFLOPS FP16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.4 CUDA TFLOPS FP32</td>
</tr>
<tr>
<td>SOFE</td>
<td>Stereo &amp; Optical Flow Engine</td>
<td>Dedicated Engines for Stereo &amp; Optical Flow</td>
<td>2x 3.1 TOPS 16 bit</td>
</tr>
<tr>
<td>ISP &amp; VIC</td>
<td>HDR and Lens Correction</td>
<td>High dynamic range support, lens distortion correction, temporal noise reduction</td>
<td>2.4 / 4 GPIX/sec</td>
</tr>
</tbody>
</table>
XAVIER 25X AI PERFORMANCE

<table>
<thead>
<tr>
<th></th>
<th>2X CPU</th>
<th>25X DL / AI (GPU + DLA)</th>
<th>12X COMPUTE (ISP+PVA+CUDA)</th>
<th>11X ACCELERATORS (Stereo, Optical Flow, LDC)</th>
<th>2.3X DRAM BW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parker</td>
<td>63</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
<td>60</td>
</tr>
<tr>
<td>Xavier</td>
<td>125</td>
<td>34</td>
<td>16.1</td>
<td>15.9</td>
<td>137</td>
</tr>
</tbody>
</table>

SpecInt2K6-Rate (est.)

DL TOPS

Equivalent CUDA TFLOPS

Equivalent CUDA TFLOPS

GB/s
## 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
- SPI O
- I2C
- I2S
- UART
- GPIO

©2018 NVIDIA CORPORATION
USE CASE COMPARISON
**XAVIER : AUTOPilot USE CASE**

Example of an Autonomous Machine Mapping on Xavier

<table>
<thead>
<tr>
<th>Component</th>
<th>Capture</th>
<th>Image Processing</th>
<th>Perception</th>
<th>Tracking + Fusion</th>
<th>Localization</th>
<th>Planning</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPX2</td>
<td>Parker ISP</td>
<td>Parker ISP, CUDA-GPU</td>
<td>DL-GPU</td>
<td>CUDA-GPU</td>
<td>CUDA-GPU</td>
<td>CUDA-GPU, CPU</td>
<td></td>
</tr>
<tr>
<td>Xavier</td>
<td>Xavier ISP</td>
<td>Xavier ISP, PVA</td>
<td>DLA, DL-GPU</td>
<td>PVA, SOFE, CUDA-GPU</td>
<td>PVA, CUDA-GPU</td>
<td>PVA, CUDA-GPU, CPU</td>
<td></td>
</tr>
</tbody>
</table>