Workloads are shifting to data-centric compute

AI, Networking, security, video and storage virtualization
A Data Processing Unit (DPU) is a compute entity that is used to move, process, secure and manage data, as it travels or while at rest, to make it available and optimized for application.
OCTEON: The original DPU platform

<table>
<thead>
<tr>
<th>Year</th>
<th>Generation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>Industry’s 1st DPU</td>
<td>OCTEON® multicore</td>
</tr>
<tr>
<td>2010</td>
<td>1st Generation Marvell DPU</td>
<td>OCTEON® Fusion</td>
</tr>
<tr>
<td>2012</td>
<td>2nd Generation Marvell DPU</td>
<td>LiquidIO®</td>
</tr>
<tr>
<td>2015</td>
<td>3rd Generation Marvell DPU</td>
<td>OCTEON® TX</td>
</tr>
<tr>
<td>2019</td>
<td>6th Generation Marvell DPU</td>
<td>OCTEON® TX2</td>
</tr>
<tr>
<td>2021</td>
<td>7th Generation Marvell DPU</td>
<td>OCTEON® 10</td>
</tr>
</tbody>
</table>

Applications:
- Firewall
- Cloud
- Wireless
- Radios
- Auto
- Edge
OCTEON 10
Industry Firsts

- Integrated hardware ML engine
- Integrated 1terabit switch
- Advanced inline crypto accelerators
- Based on TSMC 5nm process
- VPP hardware acceleration
- Compute leadership with Arm Neoverse N2 cores

Compute leadership with industry-leading performance per Watt
OCTEON 10 innovations

- 5nm TSMC process
  - Enables fanless designs
- First inline DPU ML Engine
- Hardware VPP acceleration
- Inline crypto processor
- Arm Neoverse N2 cores
  - Highest SPECint in industry
- PCIe 5.0, DDR5 support
- Integrated with 16x 50GE switch
- 56G SerDes
### OCTEON DPU platform

<table>
<thead>
<tr>
<th>Software</th>
<th>OCTEON DPU Open Software Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimized Stacks: Networking, Storage, Security</td>
<td>Virtualization and Containers</td>
</tr>
<tr>
<td>Standard APIs DPDK, SPDK, VPP</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Silicon</th>
<th>OCTEON DPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arm cores</td>
<td>Ethernet / PCIe / Memory Controllers</td>
</tr>
<tr>
<td></td>
<td>Software-enabled Accelerators</td>
</tr>
</tbody>
</table>
Best-in-class DPU inferencing
- Directly in the data pipeline
- Each ML tile contains private SRAM
- Ultra Low Power

Up to 100x performance vs SW
- Supports INT8, FP16
- Accelerated Tanh and Sigmoid activation functions

Use cases
- Threat detection
- Context-aware service delivery
- QoS
- Beamforming optimization
- Predictive maintenance

Integrated ML engine

Inferencing Tile

Shared/System memory

XBAR/Interconnect

Tile

SRAM
MAC
TVM: Mrvl-BYOC with Marvell Specializations, Software, and Hardware
Marvell TVM Integration

1. Contributions have started (bug fixes)

2. Marvell BYOC up for Review: FP16 compile-time flow

3. Next year: Quantized INT8 flow, run-time flows and more
Introducing Joe!

- Principal ML Software Engineer at Marvell
  - Leading the TVM initiative to deliver end-to-end ML solution for Marvell OCTEON
  - His background:
    - ML compiler frontend/code-gen/backend development and optimization (with GLOW, TVM, ONNX, etc.)
    - ML and SoC hardware architecture
    - Edge and Automotive ML
- Marvell contact: cchou1@marvell.com
- github contact: https://github.com/ccjoechou

Chien-Chun (Joe) Chou
Thank You