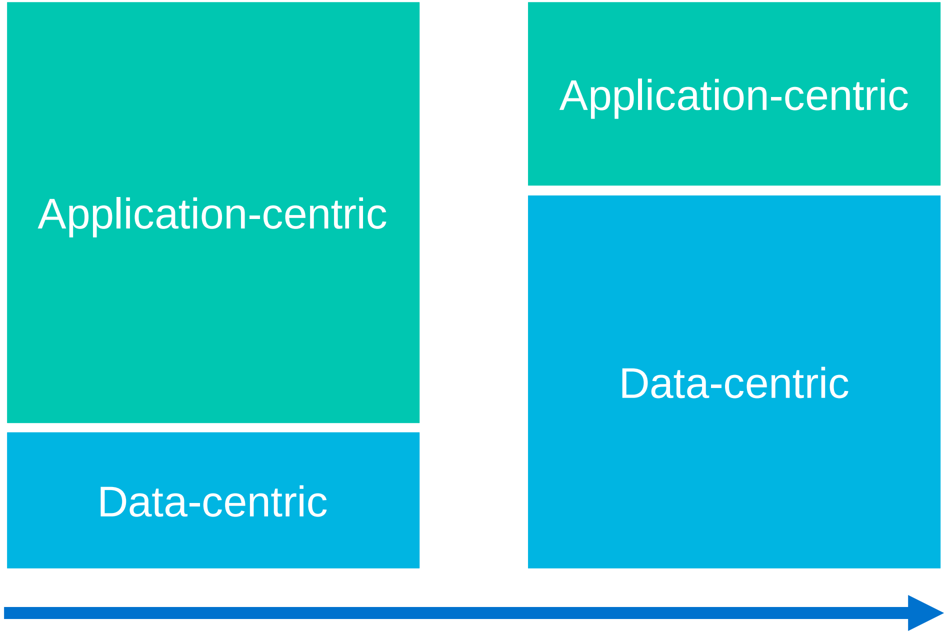
MARVELL™

OCTEON ® 10 and TVM

Dec 2021

Derek Chickles (dchickles@marvell.com)

Workloads are shifting to data-centric compute

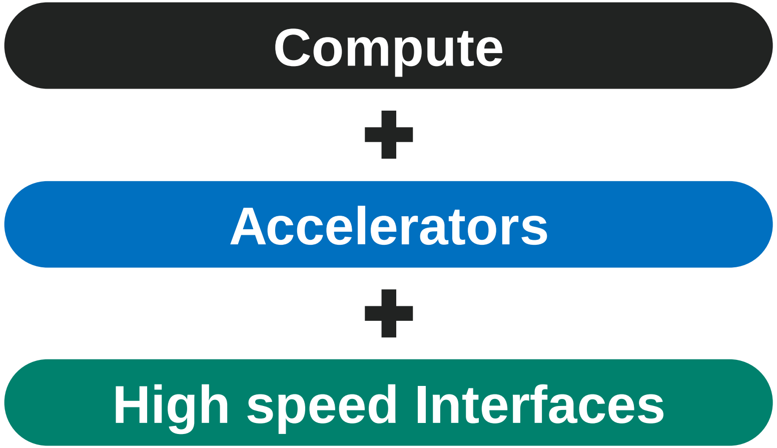


AI, Networking, security, video and storage virtualization

© 2021 Marvell. All rights reserved.

DPU definition

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



Compute

+

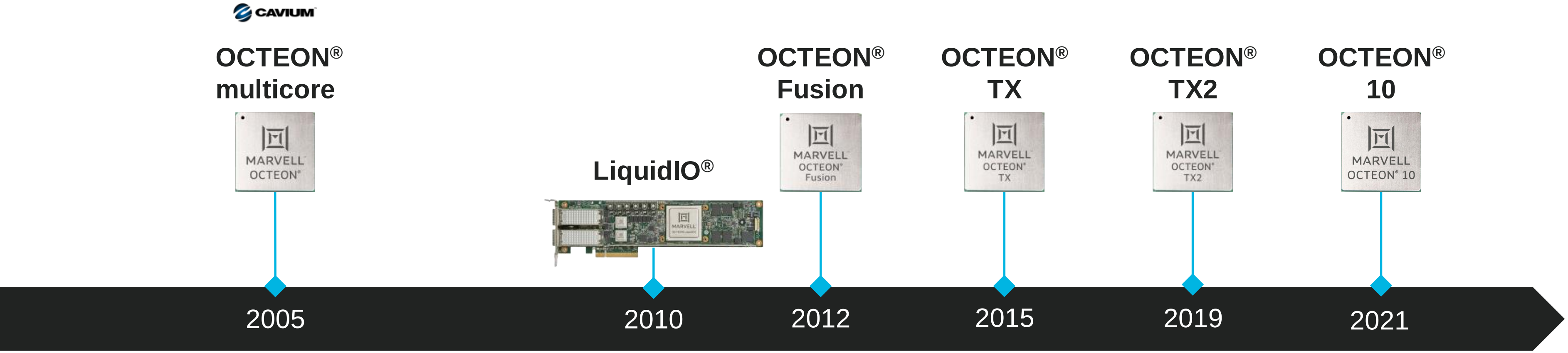
Accelerators

+

High speed Interfaces

© 2021 Marvell. All rights reserved.

OCTEON: The original DPU platform



OCTEON® multicore – 2005 – Industry's 1st DPU

LiquidIO® – 2010

OCTEON® Fusion – 2012

OCTEON® TX – 2015 – First Marvell Arm DPU

OCTEON® TX2 – 2019 – 6th Generation Marvell DPU

OCTEON® 10 – 2021 – 7th Generation Marvell DPU

Applications

Firewall

Cloud

Wireless

Radios

Auto

Edge

© 2021 Marvell. All rights reserved.

OCTEON 10

Industry Firsts

Compute leadership with Arm Neoverse N2 cores

Based on TSMC 5nm process

VPP hardware acceleration

Integrated hardware ML engine

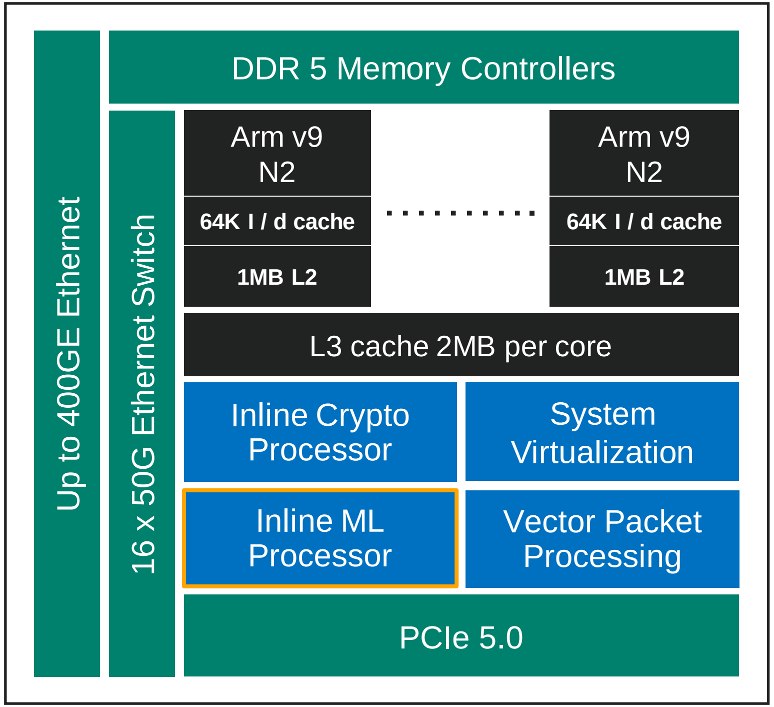
Integrated 1terabit switch

Advanced inline crypto accelerators

Compute leadership with industry-leading performance per Watt

© 2021 Marvell. All rights reserved.

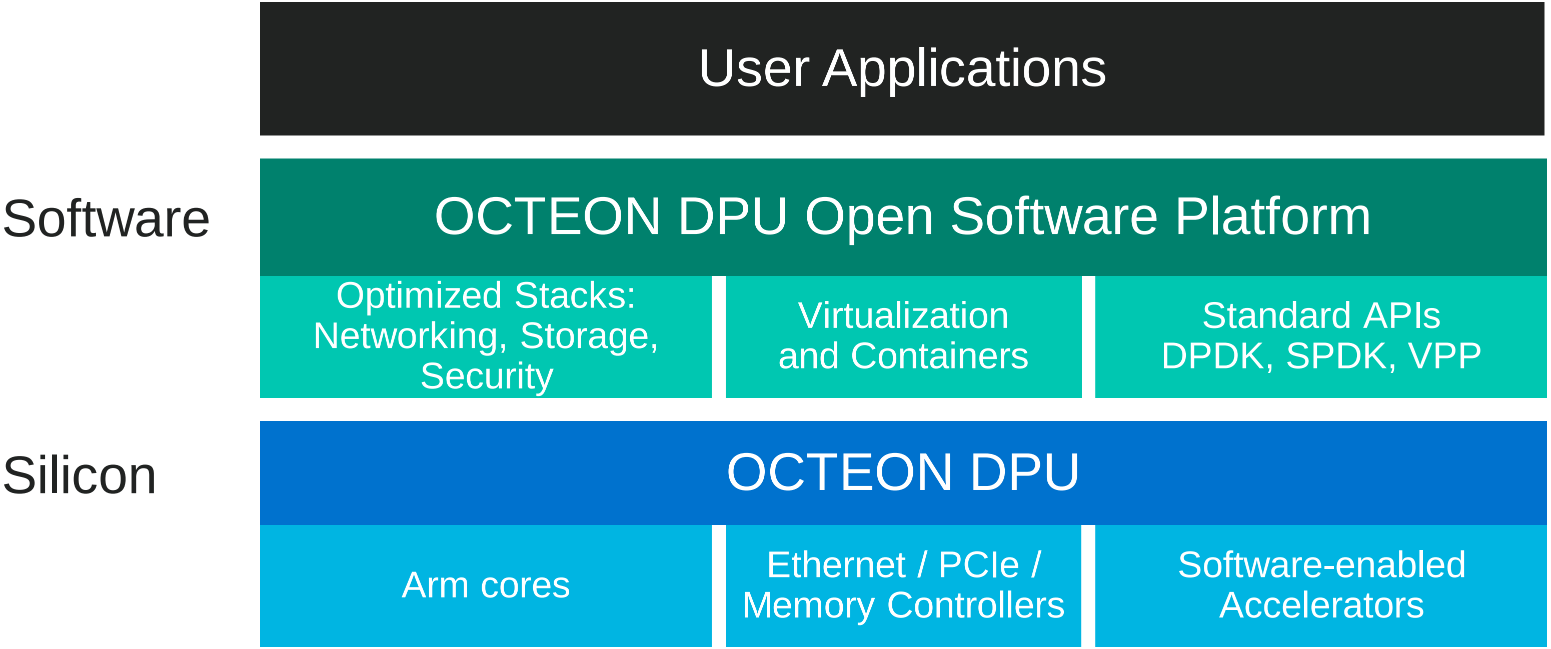
OCTEON 10 innovations



- 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

© 2021 Marvell. All rights reserved.

OCTEON DPU platform



User Applications

Software

OCTEON DPU Open Software Platform

Optimized Stacks: Networking, Storage, Security

Virtualization and Containers

Standard APIs DPDK, SPDK, VPP

Silicon

OCTEON DPU

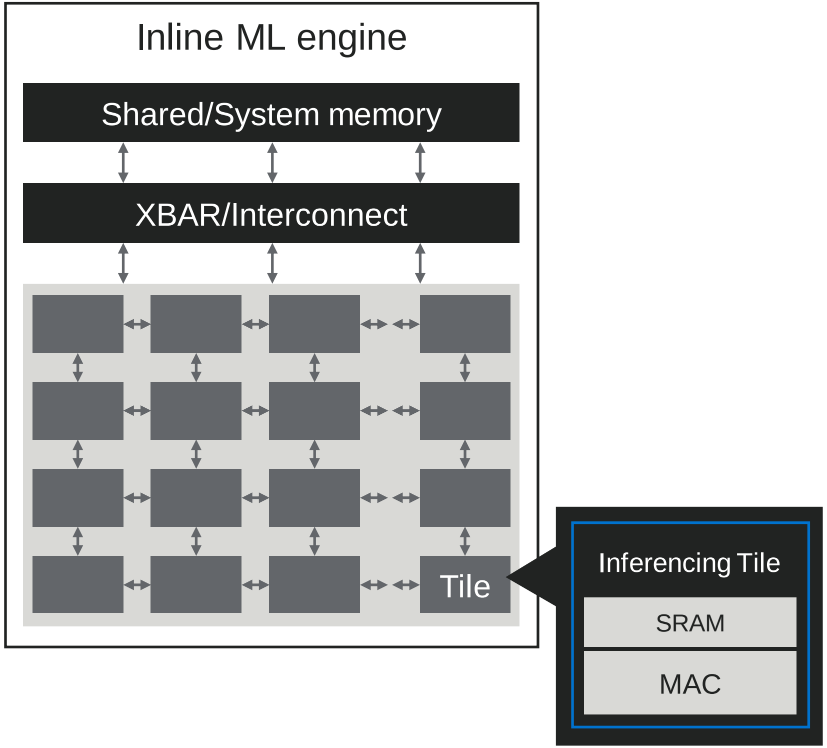
Arm cores

Ethernet / PCIe / Memory Controllers

Software-enabled Accelerators

© 2021 Marvell. All rights reserved.

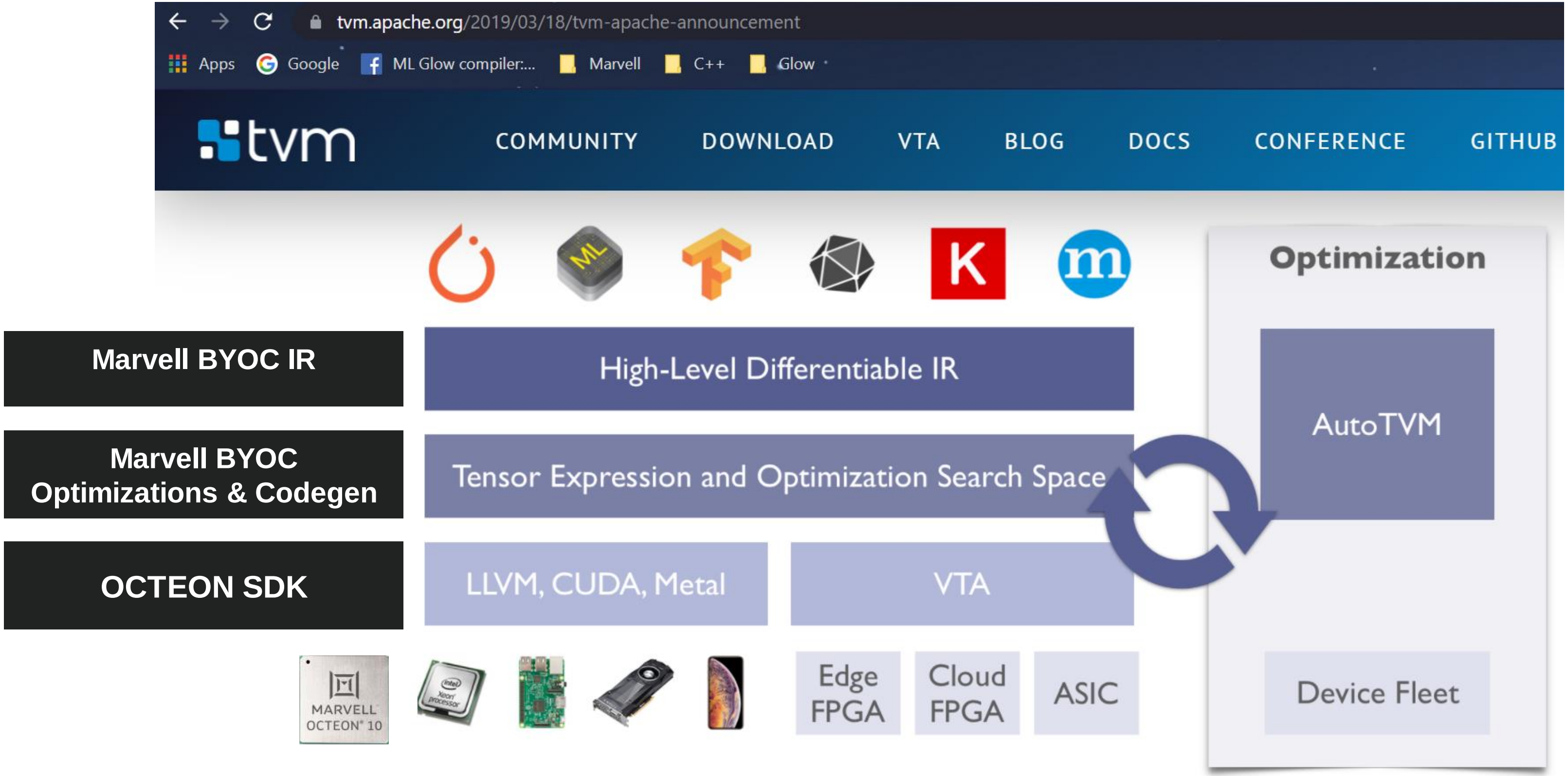
Integrated ML engine



- **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

© 2021 Marvell. All rights reserved.

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



Marvell BYOC IR

Marvell BYOC Optimizations & Codegen

OCTEON SDK

© 2021 Marvell. All rights reserved.

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

© 2021 Marvell. All rights reserved.

Introducing Joe!

Chien-Chun (Joe) Chou

- 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

© 2021 Marvell. All rights reserved.

MARVELL™

Thank You

MARVELL™

Essential technology, done right™