MARVELL

White Paper

OCTEON® 10 Machine Learning

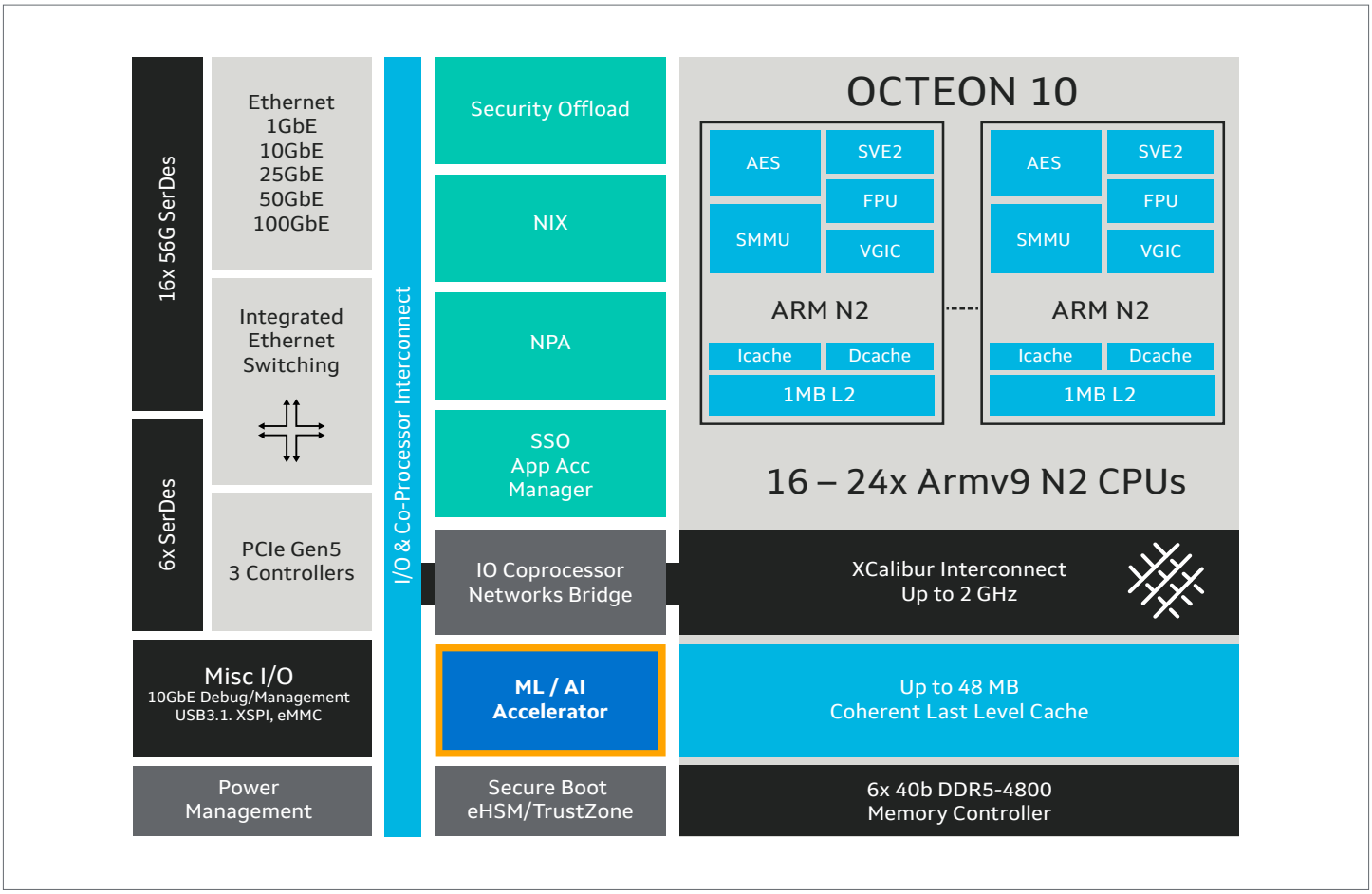
Industry's First Integrated Inferencing Platform

June 2021

OCTEON DPU Background

OCTEON DPUs have traditionally been used extensively as integrated network and security processors in enterprise networking equipment. This equipment was largely built to support a hierarchical network topology of core, distribution and access with firewalls protecting the entry and exit at the perimeters. Most enterprises owned and maintained networking hardware on premise. Over the last decade, enterprises have started moving workloads and access outside of this controlled environment and into public, private and hybrid networks accessed through both wired and, increasingly, wireless communications networks.

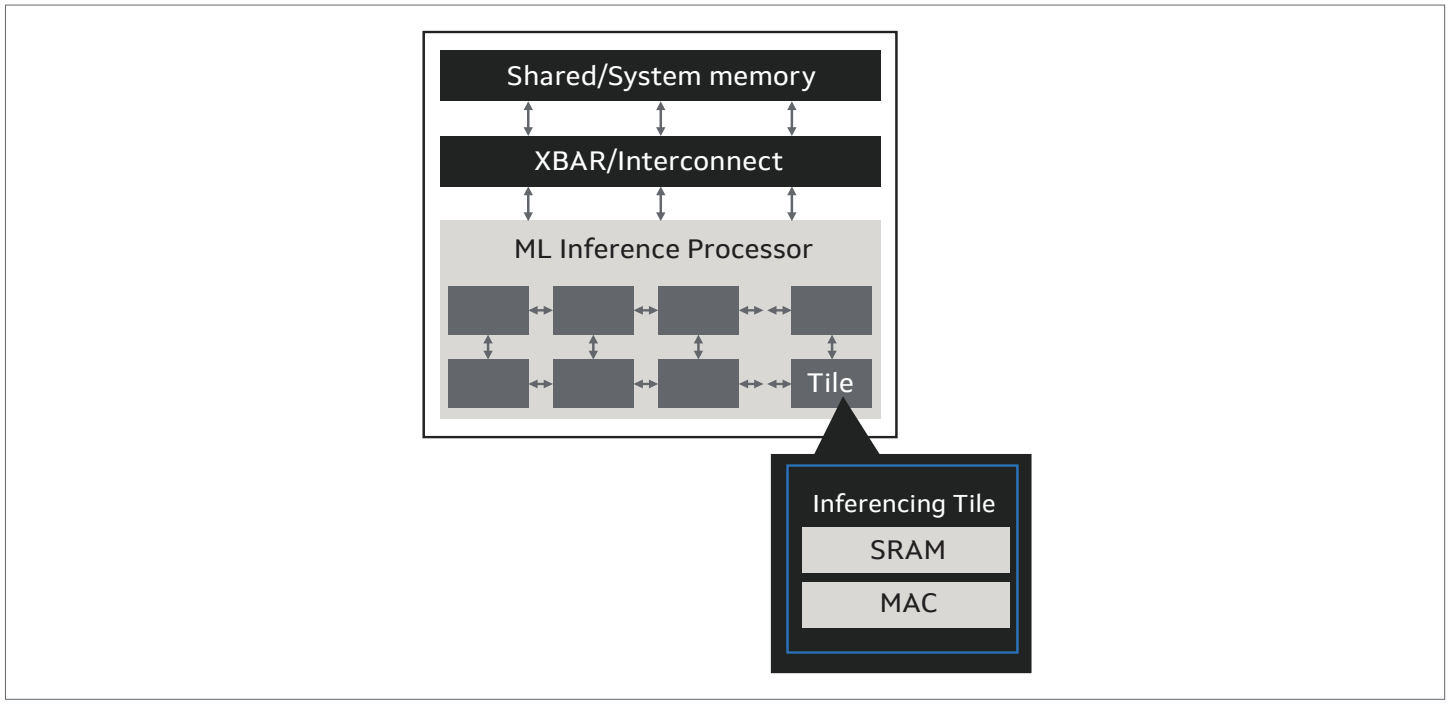
The combination of remote access, the drive to reduce costs and upgrade services, and the expansion of cloud network capabilities has led to a borderless (or perimeter-less) digital workplace and significantly changed the dynamics of building and managing connectivity and access. This disaggregation of data center resources has opened possibilities to add accelerators for these workloads into the cloud, so called Workload Accelerators. In addition to networking and security, new storage, video, and 5G workloads have been added to the list of infrastructure use cases, each needing workload acceleration. As a final kicker, machine learning inferencing capability can augment these workloads with anomalous fault detection, security and quality improvements, and deep application insight. These capabilities must be delivered with a comprehensive, optimized, open-sourced software framework.



Integrated Inferencing

Enter the 5nm OCTEON 10 DPU's integrated machine learning inference processor, which can be programmed with ML workloads in either a fully offloaded fashion, or as a hybrid offload of both engine-acceleration and OCTEON 10 Neoverse N2 core acceleration. Armv9 Neoverse N2 cores include optimized vector instructions for ML workloads, resulting in much higher performance over previous generation Arm processors. This hybrid integrated inferencing capability both eliminates the other required computing equipment in the network performing offline or near real-time machine learning jobs and transforms the jobs to real-time.

Integrated inferencing eliminates the extra data movement between network nodes dramatically reducing inference latency and reduces power and network requirements. It opens the door for applications requiring ML to be fully contained on the OCTEON 10 either as bare metal, or in VMs and Containers. The OCTEON 10's Machine Learning inference accelerator includes patented technology resulting in **best-in-class performance per watt** all in an easy-to-use, open, software suite.



High Performance

- > 100x improvement vs software
- Native INT8 and FP16 Matrix-Multiply-Add computation
- Accelerated Tanh and Sigmoid Activation functions
- 8 MB On-chip SRAM

Efficient

- Sub-2W Power
- Minimal I/O overhead
- Ultra-low latency

Robust

- Full-featured, optimized, open software suite
- Virtual Platform support
- Optimized for Infrastructure

Machine Learning Integrated Inferencing Workloads

Integrated inferencing benefits infrastructure workloads across multiple verticals. Here are some examples:

Networking

ML inferencing acceleration can be applied to software implementations of traditional networking equipment in switching, tunneling and edge, as well as in providing recommendations related to quality and telemetry metrics (jitter, latency, reliability) on network traffic. Network functions stand to benefit from using ML in combination with the traffic management, batch packet processing and scheduling accelerators in the OCTEON 10 data path.

Security

Distributed workforce and multi-service cloud applications demand traffic analysis (DPI, QoE, TM), security (firewalls, IDS/IPS) and network visibility to ensure data flows are mapped optimally are ideal targets for an integrated inferencing hardware accelerator. Using a software-based DPI in combination with ML to identify malicious traffic flows can be used for detecting bad network actors, as well as applying ML to network function traffic analytics to determine traffic SLAs of bandwidth, latency, drops and jitter are other emerging use cases of interest.

Storage

The third infrastructure accelerator class is storage – inextricably tied to network quality of service – by selecting methods for encryption and compression for data-at-rest and data-in-motion. Based on storage resources usage, the inferencing block can decide the choice of encryption and compression, tagging of storage blocks and the placement of data in hot, warm, and cold storage regions. This hardware-based integrated inferencing accelerator along with the over 3x compute performance of Neoverse N2 (vs A72) helps support legacy software-based storage applications as well as the newer computational storage use cases for edge deployments.

5G/Edge

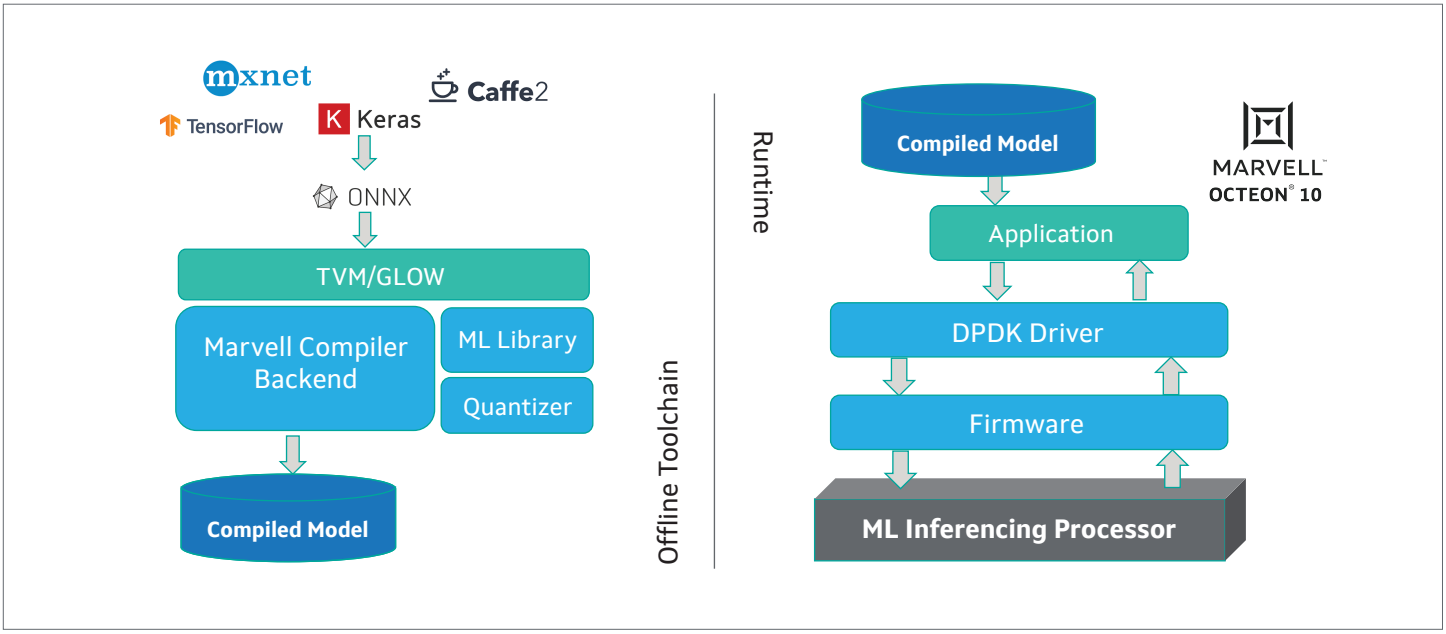
5G opens a new class of DPU offloads for which machine learning holds promise: Massive MIMO channel estimation, spectrum sensing, resource allocation, decision making under unknown network conditions are some of the use cases that are relevant. ML as a microservice at the edge is another use case for monitoring, surveillance and self-healing network use cases where the power-optimized OCTEON 10 DPU with full containerized virtualization software can fit in a role where a combination of CPU, DPU, GPU functions are performed in a single device.

Automotive

The automotive space for machine learning is very broad, not just limited to machine perception and self-driving, but also beneficial in driver assistance, safety systems, energy efficiency, analytics, and manufacturing. OCTEON 10's excellent inferencing performance and efficiency married with other OCTEON 10 workload accelerators make this a compelling choice.

Machine Learning Software

Machine Learning hardware is only as useful as the software running on it. Marvell's Machine Learning software suite includes a highly optimized, broad capability toolchain for compilation and execution of machine learning models on Arm Neoverse N2 CPUs and the OCTEON 10 ML Inference Processor. Software supports common machine learning formats and open, compilation and deployment frameworks.



The Marvell ML toolchain is optimized and integrated into ML compiler frameworks such as TVM and GLOW. Machine Learning models developed in these frameworks can be easily compiled for OCTEON 10 Neoverse N2 and/or the ML Inference Processor. These models can then be deployed on target hardware or tested and tuned on Marvell virtual platforms, including a functional and cycle-accurate ML inferencing processor emulator. Drivers are available that can be easily integrated into existing applications targeted at networking.

The OCTEON 10 DPU with Integrated Inferencing is sampling in 2H 2021.