MARVELL™

DPUs: Where Will They Go Next?

Manoj Roge
Sr Director, Processor Business Unit, Marvell

Linley Fall Processor Conference, November 2022

Forward-looking statements

Except for statements of historical fact, this presentation contains forward-looking statements (within the meaning of the federal securities laws) including, but not limited to, statements related to market trends and to the company's business and operations, business opportunities, growth strategy and expectations, and financial targets and plans, that involve risks and uncertainties. Words such as "anticipates," "expects," "intends," "plans," "projects," "believes," "seeks," "estimates," "can," "may," "will," "would" and similar expressions identify such forward-looking statements. These statements are not guarantees of results and should not be considered as an indication of future activity or future performance. Actual events or results may differ materially from those described in this presentation due to a number of risks and uncertainties.

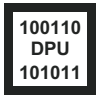
For factors that could cause Marvell's results to vary from expectations, please see the risk factors identified in Marvell's Quarterly Report on Form 10-Q for the fiscal quarter ended July 30, 2022, as filed with the SEC on August 26, 2022, and Marvell's Annual Report on Form 10-K for the fiscal year ended January 29, 2022, as filed with the SEC on March 10, 2022, and other factors detailed from time to time in Marvell's filings with the SEC. The forward-looking statements in this presentation speak only as of the date of this presentation and Marvell undertakes no obligation to revise or update publicly any forward-looking statements.

Agenda

1. System trends
2. OCTEON®: The original DPU platform
3. DPU use cases
4. OCTEON® solutions and benchmarks

DPU need driven by data centric applications

Data processing



DPU
OCTEON®

- Wired

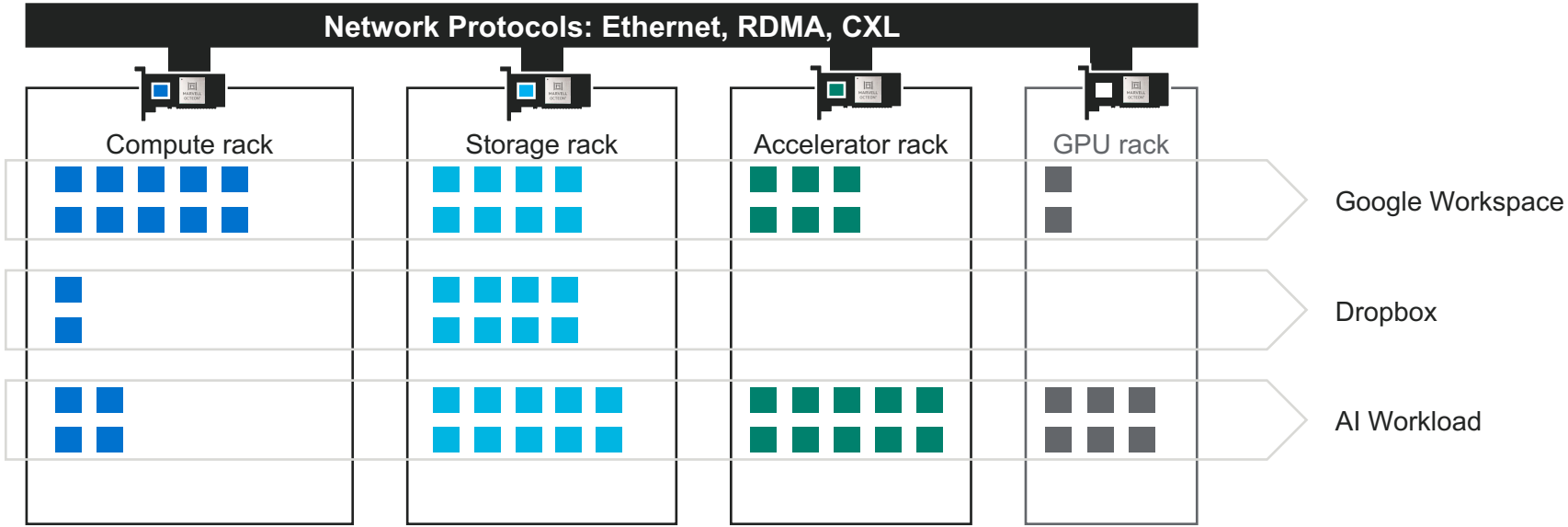
- Wireless

- Data Center

- Automotive

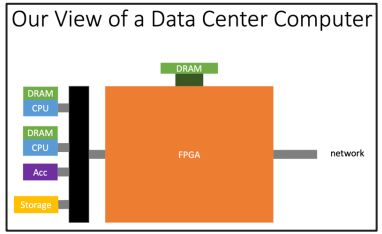
Multiple end markets for
OCTEON® DPU

Future Data Centers:
Composable, software-defined, hardware-accelerated



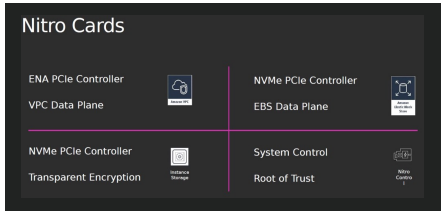
DPUs manage composability, accelerate workloads

DPU in every server!



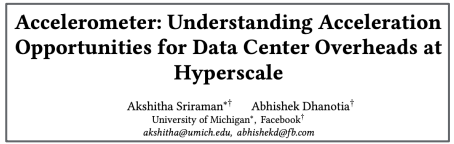
Microsoft

H2RC'16: "CPU is complexity offload engine for FPGA!"[1]



Amazon

2019 re:Invent – "All new instance launches use the Nitro System"[2]

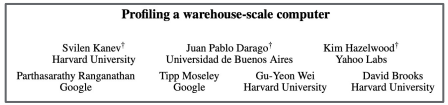


Accelerometer: Understanding Acceleration Opportunities for Data Center Overheads at Hyperscale

Akshitha Sriraman*† Abhishek Dhanotia†
University of Michigan*, Facebook†
akshitha@umich.edu, abhishekd@fb.com

Meta

"Microservices spend as few as 18% of CPU cycles executing core application logic"[3]



Profiling a warehouse-scale computer

Svilen Kanev† Harvard University
Juan Pablo Darago† Universidad de Buenos Aires
Kim Hazelwood† Yahoo Labs
Parthasarathy Ranganathan Google
Tipp Moseley Google
Gu-Yeon Wei Harvard University
David Brooks Harvard University

Google

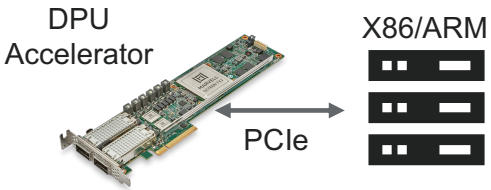
"Data center Tax" can comprise nearly 30% of cycles[4]

1: H2RC 2016 keynote.
2: AWS reinvent 2019
3: Accelerometer paper
4: Profiling a warehouse-scale computer paper

Transition to virtualization

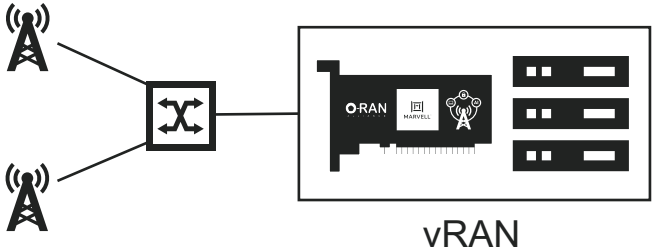
Traditional wireline appliances

Data center



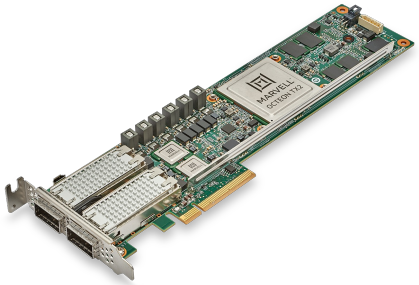
Traditional RAN/carrier

Data center



Applications require versatile mix of accelerators

Data center



Networking Storage Security AI/ML

5G Infrastructure

Baseband Security AI/ML Networking

OCTEON®: the original DPU platform

CAVIUM

OCTEON® multicore

LiquidIO®

OCTEON® Fusion

OCTEON® TX

OCTEON® TX2

OCTEON® 10

OCTEON® Next Gen

2005 | 2010 | 2012 | 2015 | 2019 | 2021

Industry's 1st DPU

First Marvell Arm DPU

6th Generation Marvell DPU

7th Generation Marvell DPU

Enterprise | Cloud | Wireless | Radio | Edge | Auto

OCTEON® 10 architectural overview

Scalable Compute

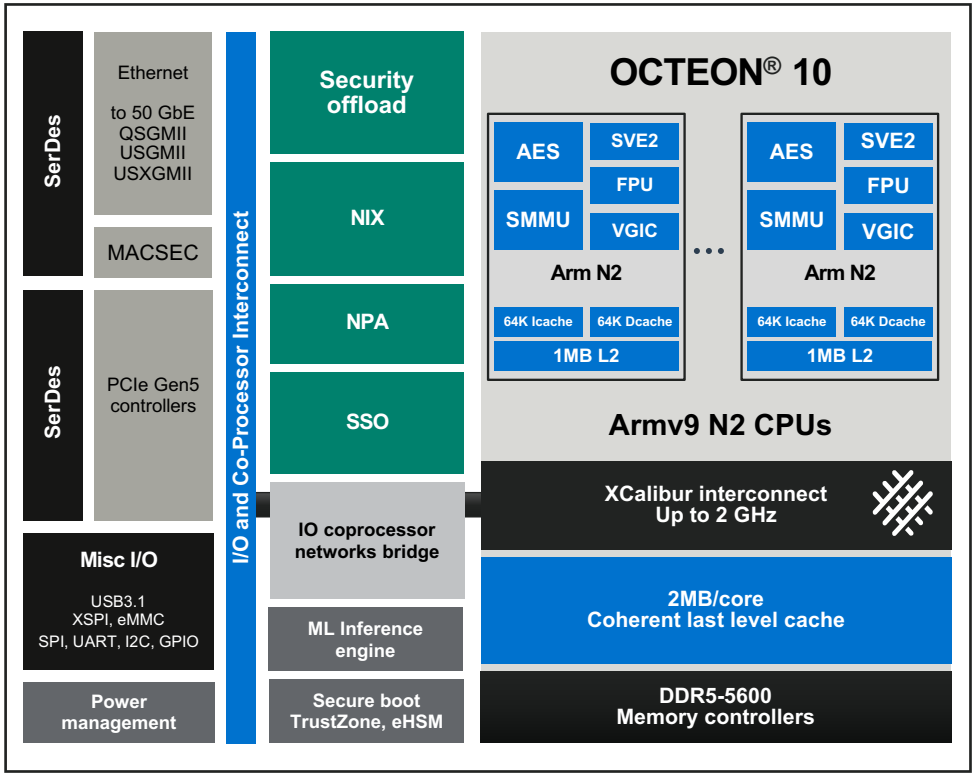
- ARMv9.0 64-bit Neoverse N2 cores

Memory subsystem and connectivity

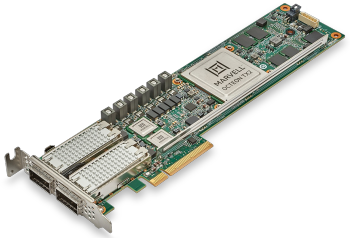
- 1MB/core L2, 2MB shared last level cache
- DDR5 w/ sideband-ECC and memory encryption
- XCalibur mesh interconnect

Hardware acceleration

- Highly-virtualized, software-friendly NIC
- Packet processing, QoS, hierarchical queues with shaper and WDRR scheduler
- Inline and Co-processor security (SSL/IPSec)
- Compression, Decompression
- Inline ML inference engine
- Secure boot + embedded hardware security module



Platform strategy



+

Virtualization

APIs and frameworks

Apps and accelerators

- DPU PCIe cards
- Robust and open source software support
- Partner ecosystem

Marvell PCIe accelerator cards

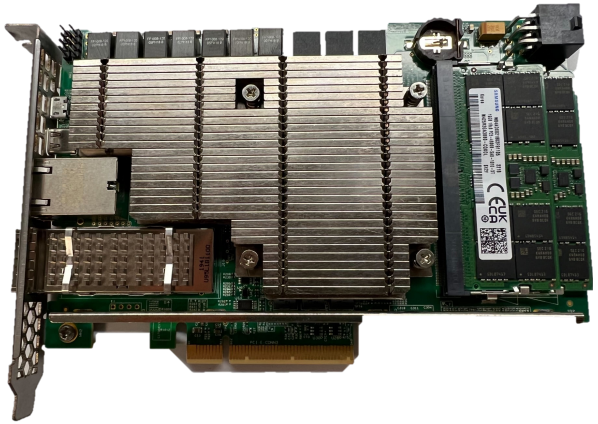
	CN98	CN106	CN103
Part Number	WA-CN98-A1-PCIE-4P100-R1	WA-CN106-A1-PCIE-2P100-R1	WA-CN103-A0-PCIE-4P50-R1
Port config	4x 100G PAM4	2x 100G PAM4	4x 50G PAM4
PCIe	Gen4	Gen5	Gen5
Core	36x ARM V8 TX2	24x ARM V9 N2	8x ARM V9 N2
Availability	Now	★	1Q CY23

Announcing
General availability!

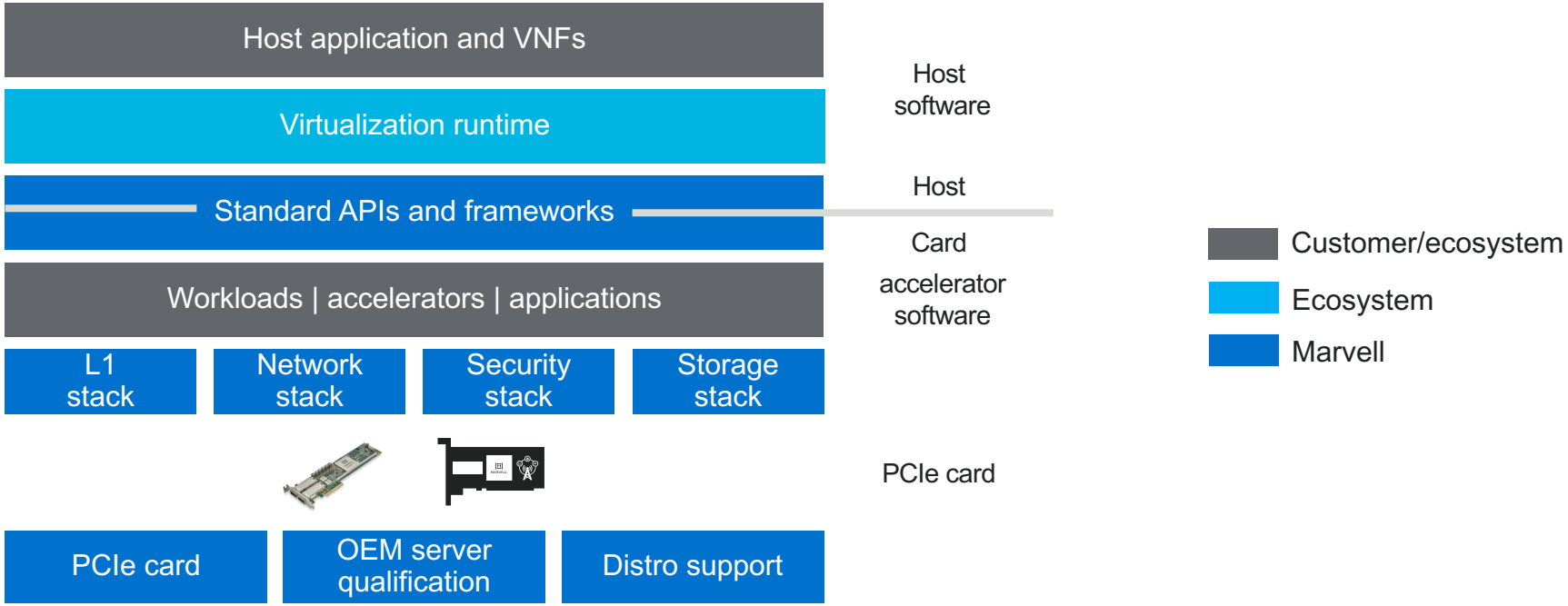
CN106 based PCIe card

Part number	WA-CN106-A1-PCIE-2P100-R1
Features	**Capability**
I/O	2 x 100G PAM4 PCIe Gen 5
Memory	6 x 40bit DDR5@ 5200MTs w/ECC, 8-40GB total
ARM cores	24 ARM N2, 2.5GHz, 100 SPECINT2017
Performance	120 MPPS, 120Gbps
IPSEC, RSA 2K, 1KB OpenSSL, TLS1.3 support	120Gbps IPSEC, 24Kops RSA 2K, 120Gbps 1KB OpenSSL
Hard ML block	Yes, 16TOPS

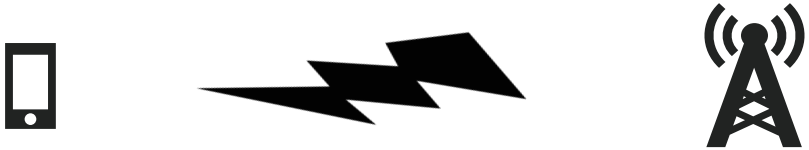
Availability	Date
SDK11 support	Now
Order in qty	Nov 2022



DPU solutions

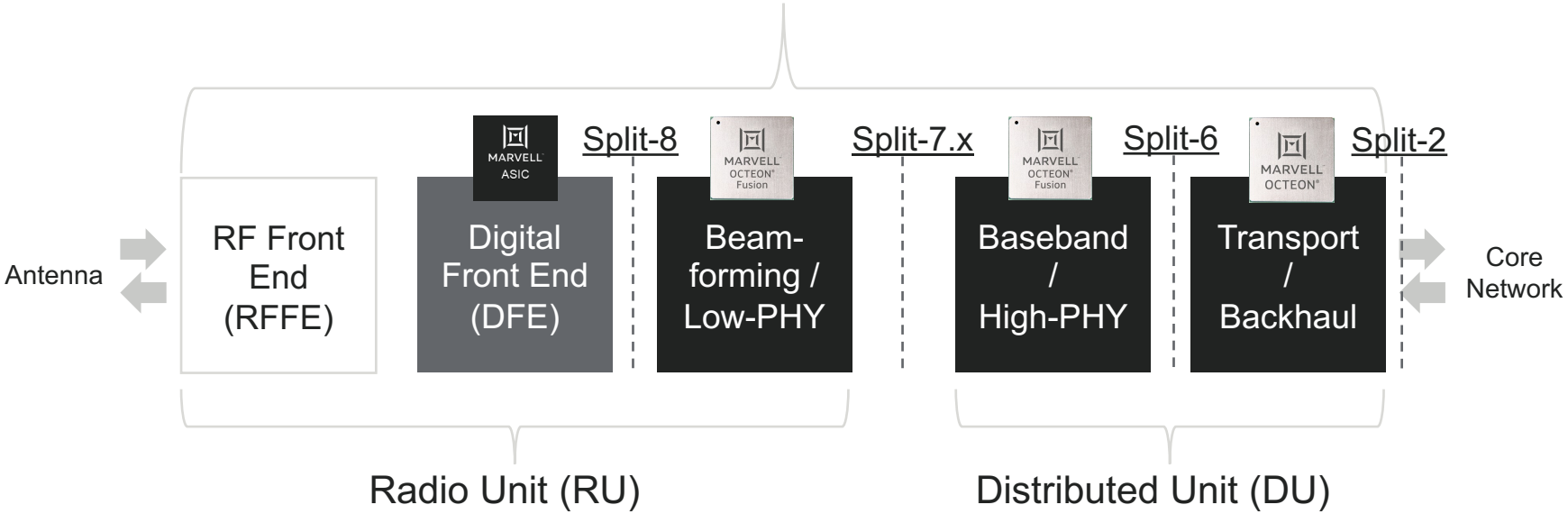


Industry-leading 5G infrastructure portfolio

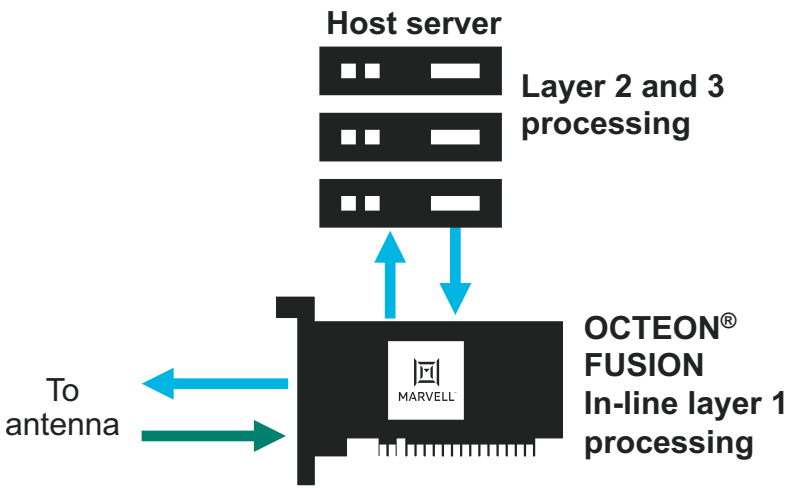


Four digital sockets

Marvell is major provider of all



5G O-RAN solution: open, scalable, best-in-class



news release

vodafone

25 October 2022

VODAFONE AND SAMSUNG COOPERATE WITH MARVELL TO ACCELERATE OPEN RAN PERFORMANCE AND ADOPTION

news release

vodafone

25 October 2022

VODAFONE AND NOKIA PARTNER TO ADVANCE OPEN RAN ECOSYSTEM IN EUROPE

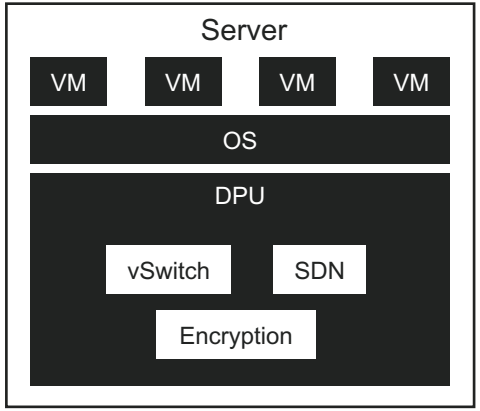
Vodafone and Nokia have agreed to jointly work on a fully compliant Open Radio Access Network (RAN) solution, marking a significant milestone for the mobile industry and a major boost to Europe's competitiveness.

The combination of Nokia's ReefShark advanced System on Chip (SoC) technology, developed in cooperation with Marvell, with standard Commercial-Off-the-Shelf (COTS) servers will enable the Open RAN system to reach functionality and performance parity with traditional mobile radio networks. Nokia's ReefShark SoC boosts the Layer-1 processing capability, which is necessary to connect many users to the mobile base station and support high levels of mobile data traffic.

Data Center use cases

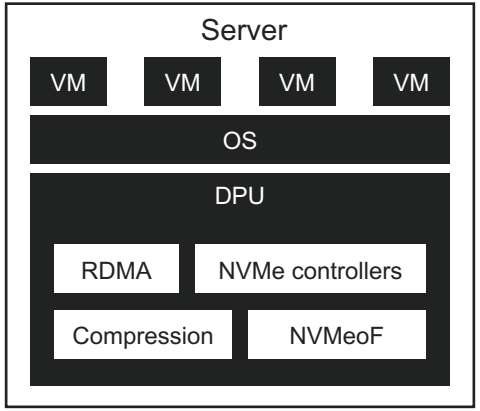
1 Network offload

Accelerate networking functions



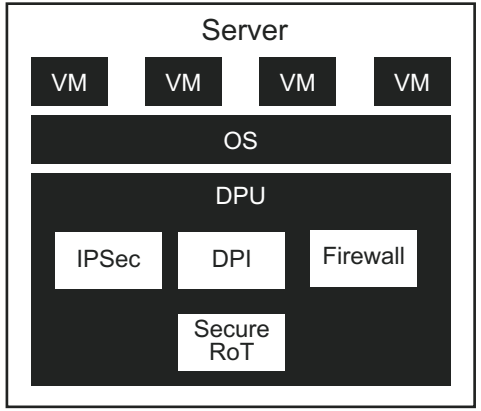
2 Storage offload

Accelerate storage functions



3 Security offload

Isolate tenants from host



DPU delivers performance, programmability & lower TCO

Enterprise use cases

SMB router, SDWAN, gateway

Control and data plane

High-end router, Firewall/security data plane

Data plane only

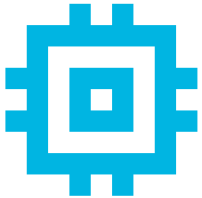
Switch, WLAN controller Line card/controller

Control plane only

Benchmarks

Compute performance

- SPECINT2006
- SPECINT2017
- CoreMark



Memory subsystem

- LMbench
- Stream

Packet processing

- TestPMD
- L3 Forwarding
- iPerf / Netperf

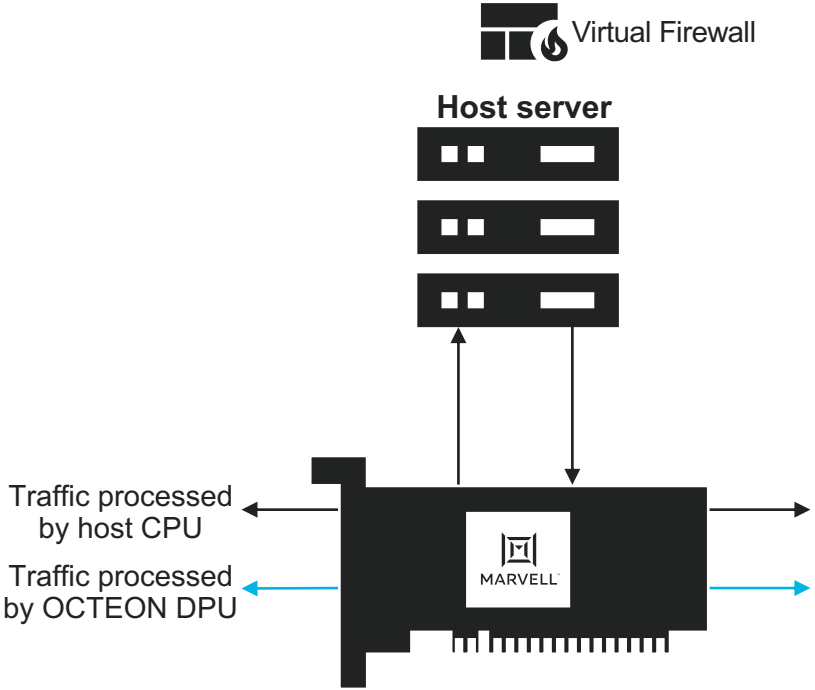
Application benchmark

- IPSec gateway
- kTLS
- Open offload
- SNORT/Hyperscan
- NVMEoF
- ML Inference

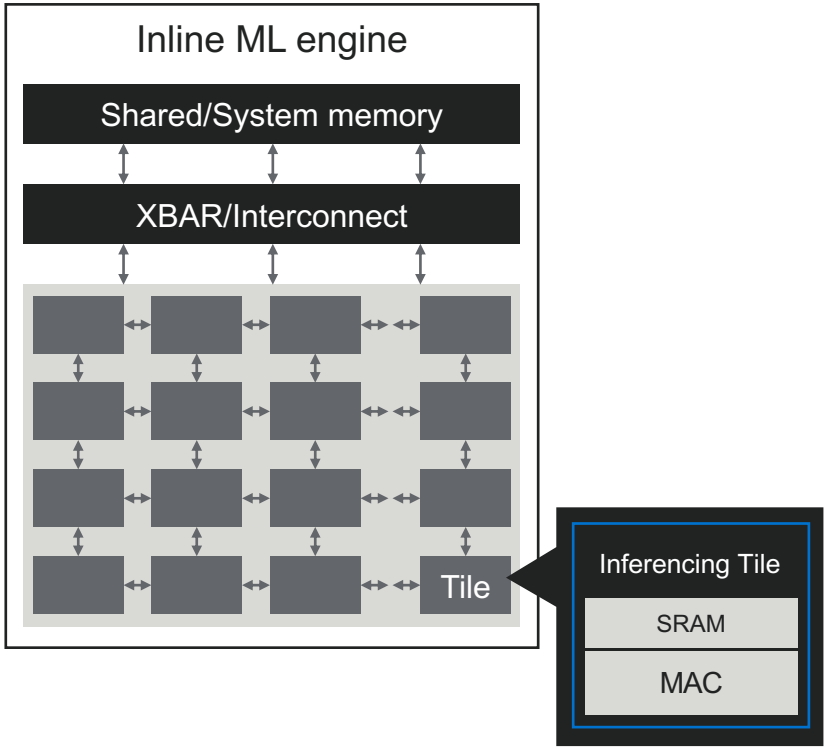
200Gbps open offload performance

- Full Firewall VM running on Host server
- Marvell DPU as fast path engine - maintain route/firewall cache based on OpenOffload

CN98xx 2Ghz, 2x100G						
Data plane cores	64B		512B		1518B	
	Mpps	Gbps	Mpps	Gbps	Mpps	Gbps
2	3.54	2.38	3.54	15.1	3.46	42.5
10	15.7	10.6	15.7	67.1	15.6	191.9
30	43.1	28.9	43.1	183.5	16.2	200



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

Summary

1. Most widely deployed DPU, shipping since 2005 for data center, enterprise and carrier use cases

2. Software-defined infrastructure requires hardware acceleration – OCTEON® portfolio has right accelerators to deliver best solution TCO

3. Unified software stack built on open source frameworks and benchmarks demonstrate leadership across broad workloads

MARVELL™

Thank You

MARVELL™

Essential technology, done right™