DATA PROCESSING UNIT (DPU)
Technical overview

Alexander Petrovskiy, Staff System Engineer, NVIDIA Networking
August 2021
SERVER NETWORKING EVOLUTION: FROM NIC TO DPU
Software Defined Data Center Infrastructure on-a-Chip

Legacy NIC

Legacy Infrastructure

Accelerated SW-defined Infrastructure on CPU

SW-defined Infrastructure on DPU

Datacenter on a DPU
DEMYSTIFYING SMARTNICs AND DPUs
Software-Defined, Hardware-Accelerated

CONNECT-X SmartNIC
ASIC-based advanced NIC with Fully Accelerated Datapath for Secure Cloud, Telco and Enterprise

BLUEFIELD DPU
SoC-based DPU with full Data & Control Path Acceleration for Unified Cloud
## WHAT MAKES A SMARTNIC SMART?

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoCE</td>
<td>Selective Repeat</td>
</tr>
<tr>
<td>Resilient RoCE</td>
<td>IPsec Inline Offload Engine (aware/un-aware)</td>
</tr>
<tr>
<td>PTP HW Clock</td>
<td>Hardware Steering and Filtering</td>
</tr>
<tr>
<td>Accurate timestamp</td>
<td>Secure Firmware Update</td>
</tr>
<tr>
<td></td>
<td>Secure Boot (HW RoT)</td>
</tr>
<tr>
<td></td>
<td>Key Management</td>
</tr>
<tr>
<td></td>
<td>AES-XTS Storage Encryption Engine</td>
</tr>
<tr>
<td></td>
<td>TLS Inline Offload Engine</td>
</tr>
<tr>
<td></td>
<td>Connection Tracking</td>
</tr>
<tr>
<td></td>
<td>Hardware Steering and Filtering</td>
</tr>
<tr>
<td></td>
<td>AES-XTS Storage Encryption Engine</td>
</tr>
</tbody>
</table>

**x16 PCIe Gen 4.0:**
BLUEFIELD-2 DPU

Technical Overview

ConnectX-6 Dx inside

200 Gbps Ethernet & InfiniBand, NRZ & PAM4 modulation

8 ARM A72 CPUs subsystem in a Tile architecture
  - 8MB L2 cache, 6MB L3 cache in 4 Tiles
  - ARM Frequency up-to 2.75GHz

Fully integrated PCIe switch, 16 bi-furcated Gen4.0
  - Root Complex or End Point modes

1GbE Out-of-Band management port
**BLUEFIELD-3 DPU**

**Technical Overview**

ConnectX-7 inside

I/O

- 2x400Gbs (Active/Standby), 4x100Gbs Ethernet/InfiniBand
- 100G PAM4 serdes
- 400Gb/s bandwidth

Integrated PCIe switch
- Gen5.0 x32+2
- Multi-host - 8 hosts

Compute sub-system
- 16 Arm®A78 v8.2+ Hercules @2.3GHz
- SkyMesh™ fully coherent low-latency interconnect
- 8MB L2 Cache, 16MB LLC System Cache

Built-in accelerators

Advanced Memory sub-system
- Dual Channel 256GB DDR5-4800MT/s w/ ECC
- NVDIMM-N Support
- DDR memory encryption

1GbE Out-Of-Band management port

Self-hosted or Server-hosted
MOVING INFRASTRUCTURE SERVICES TO DPU

Software Defined Networking
- vRouter
- vSwitch
- NAT/Load Balancer
- VMs & Containers
- Telco/NFV

Software Defined Storage
- NVMe-oF
- Storage Direct
- Data Encryption
- Elastic Storage
- Compression
- DeDup

Software Defined Security
- Distributed IDS/IPS
- NG Firewall
- Root of Trust
- Micro Segmentation
- DDOS Prevention
DPU FOR UNIFIED CLOUD USE-CASE
Unified infrastructure for host Networking, Storage, Security and Management

Today’s Environment

- Network I/O
- Storage I/O
- Security & Crypto
- Host Mgmt

Hypervisor

Unified Datacenter

- CPU Scheduling
- Lightweight Hypervisor
- Bare Metal Server

Functional Isolation

- Network I/O
- Storage I/O
- Security & Crypto
- AI Acceleration
- Host Mgmt

BlueField-2 DPU
DPU - IS THE NEW NETWORK EDGE
Moving the Top-of-rack Into the Server
HBN FOR UNIFIED CLOUD

Modern Networking with Classic Controls

Network Administrator
- EVPN
- Crypto
- Ansible

Standard Linux Control Plane

Accelerated Data Plane

Offloaded VXLAN

BGP Peering

Server Administrator

Zero Trust

Bare Metal

Containers

Hypervisors

Microsoft

docker

kubernetes

vmware

NVIDIA CONFIDENTIAL. DO NOT DISTRIBUTE.
DPU CAPABILITIES AND FEATURES - NETWORKING
ASAP²: VIRTUAL NETWORKING AND SDN ACCELERATION

Accelerated Switching and Packet Processing

Virtual Switches / SDN Packet Processing Put Heavy Load on CPU

Hardware Accelerated Packet Switching with Zero CPU Utilization
Integrated with commercial and community partners
Leveraged to create Efficient Cloud Architectures
HOST NETWORKING ACCELERATION

Comparing two general approaches

- **SR-IOV**
  - Single root input/output virtualization
  - Native hardware access at the VM level
  - Every VM has direct access to the network adapter (through virtual function, VF interface)
  - Baremetal-like network performance in VM, zero CPU utilization
  - Guest awareness limitations: NIC driver in VM, VM Live Migration challenges

- **VirtIO**
  - Virtualization standard for network device drivers in Linux systems
  - VirtIO abstracts the hardware to the guest OS in SW
  - Poor network performance in VM, CPU is utilized to move packets
  - Guest un-aware: Virtio-net interface in VM, native VM Live Migration
  - Can be accelerated in NIC HW using vDPA
VF REPRESENTOR PORTS

SW Representation of SR-IOV NIC Virtual Function

- **VF Representor**
  - Net Device modeling of eswitch port and exposed through PF driver
  - VF and its representor works like Linux veth pair
  - Flow configuration (add/remove)
  - Works under switchdev mode
- Access from both kernel and DPDK
  - Multi Queue (RSS/TSO/CSUM)
  - Attach/Detach in DPDK
  - Multiple DPDK instances over VF representor
- With VF representor, vSwitch can work with SRIOV together and reduce CPU% consumed by virtio.
**EMBEDDED SWITCH (ESWITCH)**

*Flow-based Packet Processing and Steering Engine in SmartNIC/DPU*

- Flow based Classification and action
  - Hierarchical multiple layer tables
  - Table consists of classification and action
    - Action may point to next table
- Key fields example: Ethernet L2/IPv4/IPv6/TCP/UDP/Inner Packet (VXLAN/GENEVE/etc.)
- Actions example: Allow/Deny, Re-write (Route/NAT), Encap/Decap of headers, Meta Data set, Hairpin, Sample, Counter, etc.
NETWORKING OFFLOAD MODEL ON DPU

Full Control Plane and Data Plane offload

- Control plane and SW Datapath on DPU
- HW Datapath is accelerated as in SmartNIC
- Both SRIOV and VIRTIO interface to VM
- Advantages
  - Support virtualized network services for VMs, Containers and bare-metal cloud
  - Zero Host CPU utilization for networking services
    - All host resources (core and memory) can be used for VMs
  - Efficient packet forwarding in HW
  - Host isolation
  - DPU can offload extra I/O and management services
    - Storage (NVMeOF, Virtio-blk)
    - Security (Firewall, DPI, IPSec/SSL crypto)
    - Host infrastructure management (BMC, Barametal/VM provisioning)
ASAP\textsuperscript{2}: BLUEFIELD-2 TRAFFIC FLOW

Embedded CPU Configuration (Switchdev)
DPU CAPABILITIES AND FEATURES - STORAGE
DPU STORAGE

Storage Agility Meets Best-in-Class Hardware Acceleration

UNPARALLELED PERFORMANCE
- Dual 100Gbps or single 200Gbps
- Up to 5.4M IOPS @4KB
- Lowest latency
- NVMe-oF acceleration

STORAGE SECURITY
- Data-at-rest AES-XTS encryption
- Authentication services
- Protection between users

DISAGGREGATED STORAGE
- NVMe SNAP
- Virtio-blk SNAP
- Integrated data & control planes
SNAP: LOCAL STORAGE TO EMULATED STORAGE

**Physical Local NVMe Storage**
- ✓ Serving bare-metal and hypervisor/VMs
- ✗ Bound by physical SSDs capacity
- ✗ Under-utilized storage
- ✗ Scalability on demand
- ✗ Over-provisioning bound to compute node

**SNAP Drive Emulation**
- ✓ Serving bare-metal and hypervisor/VMs
- ✓ Over-provisioning, scaled to rack/cluster
- ✓ Saving OPEX and CAPEX
- ✓ OS-agnostic using inbox standard driver
- ✓ Supports all network transport types - NVMe-oF, iSCSI, iSER and even proprietary
- ✓ Accelerated data path* for VMs
- ✓ Live-migration with virtio-blk* and vDPA*
- ✓ Support for older OSs where only virtio-blk* is available

* Roadmap
BLUEFIELD-2 SNAP – NVME/VIRTIO-BLK
Framework for Storage virtualization software

- Emulate NVMe Local Storage
- Connected to Remote Cloud Storage
- Virtualized or Bare Metal Cloud
- OS Agnostic with RDMA inside

- Enabling two data paths - (1) offload with NVMe-oF(RDMA)* vs (2) SPDK
- Pluggable to Linux’s block devices (NVMe-of, iSCSI, iSER, etc)
- Provides infrastructure for Storage Application development
- Enabling End to End storage orchestration and integration
DPU CAPABILITIES AND FEATURES - SECURITY
DPU SECURITY SOLUTIONS

Integrated Security for modern data center needs

- **SECURED HARDWARE**
  - Secure FW upgrade
  - Root-of-Trust
  - Arm trust zone

- **ADVANCED L4-L7 SECURITY**
  - NG stateful firewall
  - Deep Packet Inspection
  - Host introspection

- **CRYPTO ACCELERATION**
  - Data-in-motion enc.
  - Data-at-rest enc.
  - Public Key Acceleration

- **PROGRAMMABILITY & ISOLATION**
  - Hardened Isolation
  - Micro-Segmentation
  - Programmable Networking
DPU SECURITY CAPABILITIES

Trust Shifts to the DPU

- Isolated Security Control Plane
- Deep Packet Inspection
- Stateful Firewall
- Inline Crypto Accelerators
- Root-of-Trust

DPU Security Services
- Micro-segmentation
- Next Generation Firewall
- DDoS Protection
- Intrusion Protection
- Anomaly Detection

Security Requires Full Isolation from the Host

Network Traffic

CPU

GPU
IPSEC: TRANSPARENT ENCRYPTION

Encryption/decryption at 100Gb/s bidirectional

Traditional Server
IPsec runs on CPU

DPU Accelerated Server
IPsec and vSwitch on DPU

Inline with other accelerators (tunneling, TLS, etc.)
Cipher: AES-GCM 128/256bit keys
Keys are stored encrypted in hardware
Encrypted RDMA

Control Plane
Software on Arm
Acceleration Engine
ACCELERATING NEXT-GENERATION FIREWALLS

Hardware-Accelerated Policy Enforcement

Accelerated Switching and Packet Processing (ASAP\textsuperscript{2}) enables seamless offload of packet filtering, steering, crypto and stateful connection tracking rules to the DPU HW.
DPU HIGH LEVEL SW ARCHITECTURE
Software-Defined, Hardware-Accelerated Infrastructure

Software-Defined Networking

Software-Defined Storage

Software-Defined Security

DPU SW and SDK (DOCA)
Open and Programmable API Framework
Easy, Flexible Programming of Infrastructure / Acceleration and Security

DPU HW
DPU SOFTWARE COMPONENTS

Bootloader - UEFI, ATF (Arm Trusted FW), ACPI

Linux Distro - CentOS reference drivers, Ubuntu commercial OS

Mellanox Drivers : OFED driver, ASAP2, NVME SNAP

Secure Boot and Secure Firmware Upgrade

OpenBMC for BMC Management

ConnectX-6 Dx firmware binary file
NVIDIA DOCA
Data-Center-Infrastructure-on-a-Chip Architecture

COMMUNITY of DEVELOPERS
SDK for ecosystem partners, academia, community

ACCELERATE TTM
Leverages open-source and industry standards (DPDK, P4); NGC-certified

COMPETITIVE EDGE
Best performance; out-of-the-box experience; libraries with special capabilities

LONG-TERM COMMITMENT
Backward and forward compatibility; consistency with performance improvements

DOCA is for DPUs what CUDA is for GPUs
ONE-STOP SHOP FOR DPU DEVELOPERS

- Developer Zone Program and Website
- SDK Manager Support
- Tools (Compilers, Benchmarks, etc.)
- DOCA Drivers and Libraries
- API References and Programming Guides
- Reference Applications per Use Case
- Accelerated Solutions Integration
DOCA SDK STACK

APPLICATIONS

- Networking
- Security
- Storage
- HPC/AI
- Telco
- Media

DOCA SERVICES
- Orchestration
- Telemetry
- Storage
- SDN
- DPU Management

DOCA LIBRARIES
- FLOW
- DPI
- Host Introspection
- Data Integrity
- SPDK
- HPC/AI
- UCX/UCC
- RiverMax
- TSDC

DOCA DRIVERS
- Networking: ASAP2, DPDK, P4, P4-RT
- Security: DPDK RegEx, DPDK SFT, Inline Crypto
- Storage: SNAP, VirtIO-FS, XTS Crypto
- HPC/AI
- RDMA
- FlexIO
- Comm Channel

DPU – BlueField and BlueField-X
JOIN THE DOCA DEVELOPER PROGRAM TODAY

DPU PARTNER ECOSYSTEM

Hybrid Cloud Compatibility  |  No Fork-Lift Upgrades  |  No Vendor Lock-In