NVIDIA DATA CENTER PROCESSING UNIT (DPU) ARCHITECTURE

Idan Burstein, DPU Principal Architect
DATA CENTER IS THE NEW UNIT OF COMPUTING

Monolithic
Software-Defined Datacenter
Disaggregated, Micro-Services, Scaled Out
GPU-Accelerated Computer
DPU-Accelerated Data Center Infrastructure

E V O L U T I O N O F T H E D A T A C E N T E R


Z E R O T R U S T S E C U R I T Y
NAIVELY MOVING WORKLOADS TO NIC CPUS DOESN’T WORK

Traditional Server - 30 Total Cores

- Server Cores for Applications
- Server Cores for Infrastructure
- Regular NIC

Server with Non-Accelerated DPU Offload - 36 Total Cores

- Server Cores for Applications
- DPU Cores for Infrastructure
- Regular NIC

18 DPU Cores Replace 12 Server CPU Cores — No Gain in Performance or Efficiency
Not compatible for higher bandwidth without requiring significant system modification

Shift CPU Workload to DPU Cores
DPU MUST INCLUDE HARDWARE ACCELERATION

Today's Environment

![Diagram showing current environment with CPUs and cores for VMs and Containers.]

Offload, Accelerate, Isolate

![Diagram showing enhanced environment with NVIDIA DPU, Arm Cores, and Accelerators. Arm Cores Run Control Plane or Security Workloads Requiring Domain Isolation.]

NVIDIA DPU with Arm Cores & Accelerators

- Infrastructure Management
- Software-defined Security
- Software-defined Storage
- Software-defined Networking
- Acceleration Engines

DPU Accelerators and 8 Arm Cores Replace 20 to 120 CPU Cores — HUGE Efficiency Gain
NVIDIA DPU ROADMAP
Exponential Growth in Data Center Infrastructure Processing

- **BlueField-2**
  - 7B Transistors
  - 9 SPECint*
  - 0.7 TOPS
  - 200 Gbps

- **BlueField-3**
  - 22B Transistors
  - 42 SPECint*
  - 1.5 TOPS
  - 400 Gbps

- **BlueField-4**
  - 64B Transistors
  - 160 SPECint*
  - 1000 TOPS
  - 800 Gbps

DOCA — ONE DEVELOPMENT ARCHITECTURE

* SPECint2k17-rate
NVIDIA BLUEFIELD-3 DPU
First 400Gb/s Data Processing Unit

22 Billion Transistors
400Gb/s Ethernet & InfiniBand Connectivity (1-4 Ports)
Pcie Switch Gen 3/4/5 x32+x4
400Gb/s Crypto / Security Acceleration
2x370M PPS, 2x40M PPS at scale of millions of flows
18M IOP/s Elastic Block Storage
300 Equivalent x86 Cores
16C A78 ARM 42 SPECINT2k17-rate
128b DDR5-5600
NVIDIA DOCA
Enabling Broad BlueField Partner Ecosystem

Software Development Framework for BlueField DPUs
Software Compatibility for Generations of BlueField DPUs
Offload, Accelerate, and Isolate Infrastructure Processing
Support for Hyperscale, Enterprise, Supercomputing and Hyperconverged Infrastructure
DOCA is for DPUs what CUDA is for GPUs
**BLUEFIELD-3 PROGRAMMABLE ENGINES**

| ARM | 16 Arm A78 cores  
|     | Fully programmable OS  
|     | Apps/services, service chaining  
|     | Control Path / Slow Path  
|     | Memory to Memory Accelerators  |

| Datapath Accelerator | 16 cores, 256 threads  
|                      | Programmability through DOCA  
|                      | Heavy multi-threading application acceleration  |

| ASAP² | Programmable packet processor flow pipeline  
|       | Flow table based  
|       | Data Path  |

---

*Infrastrucure Applications*
- Containers
- DOCA Framework
- Open APIs and Services
- NVIDIA BlueField DPU
- BlueField Operating System
- NVIDIA BlueField-3/3X DPU
Server Class CPU subsystem
- Data center operating system control plane
- Isolated memory subsystem optimized for networking

NIC subsystem
- Isolated boot domain, real time OS
- Accelerating data path at line rate

PCIe subsystem
- Flexible EP/RP assignment, PCIe switching, NTB, p2p communication, emulated devices, optimized for IO

Data acceleration
- Accelerating ARM workload
DPU ACCELERATED SWITCHING AND PACKET PROCESSING
Programmable Data Path | Software-Defined Orchestration

**Accelerated**
- ✓ Virtio-Net/Other Emulation
- ✓ QoS & scheduling
- ✓ Telemetry and statistics
- ✓ Micro Segmentation / IPS / IDS / WAF
- ✓ Encryption (Ipsec / MACsec)
- ✓ Tunneling (VXLAN / GRE)
- ✓ NAT
- ✓ Routing
- ✓ ACL

**Software Defined**
- ✓ eSwitch management
- ✓ Connection Establishment
- ✓ Key Association
- ✓ Monitoring & Stats

![Diagram of DPU architecture and features](image-url)
100G OVS-DPDK - VXLAN & CONNECTION TRACKING

Faster Performance | Lower CAPEX

BlueField-2 P-series DPU, 100GbE Single Port Card | AMD EPYC 7742 64-Core Processor
12

BLUEFIELD-2 100G IPSEC TCP PERFORMANCE
Faster Performance | Lower CAPEX

- BlueField-2 P-series 100GbE Single port
- Intel(R) Xeon(R) CPU E5-2687W v4 @ 3.00GHz
- Host OS : RHEL 8.3 (Ootpa)

**Throughput**
IPSec + VXLAN, single tunnel, 8 processes

- 88.9 Gbps
- 21.2 Gbps

**Server Cores Utilization**
IPSec + VXLAN, single tunnel, 8 processes

- 100% utilization
- ~50% less cores

**Performance per Core**
IPSec + VXLAN, single tunnel, 8 processes

- ~6X

- Each SW core does : 21.2 / 14.2 = ~1.5G
- To reach ~90G in SW we need ~60 cores
- To reach ~90G w/ DPU we need ~10 cores
- SW cores do : 21.2 / 14.2 = ~1.5G
- With DPU SW cores do : 88.9 / 10.3 = ~9G
- Core performance increased by magnitude 6X
# DPU ACCELERATED STORAGE PROCESSING

**Programmable Data Path | Software Defined Orchestration**

<table>
<thead>
<tr>
<th>Accelerated</th>
<th>Software Defined</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ NVMe / Virtio-Block Emulation</td>
<td>✓ LVM / RAID control plane</td>
</tr>
<tr>
<td>✓ Data Reduction, Erasure Coding</td>
<td>✓ Key Association</td>
</tr>
<tr>
<td>✓ Data at Rest Crypto &amp; Integrity</td>
<td>✓ QoS Monitoring &amp; Stats</td>
</tr>
<tr>
<td>✓ RDMA - RoCE / InfiniBand, TCP</td>
<td>✓ Namespace/controller management</td>
</tr>
<tr>
<td>✓ Data-in-Flight Encryption</td>
<td></td>
</tr>
</tbody>
</table>

---

**Diagram**

- **Bare Metal Host**
- **Control Path**
  - ARM
  - Hypervisor
  - SDS
  - SDN
- **Data Path**
  - Virtual PCIe Switch
  - NVMe
  - IO Processors
  - RDMA/TCP Pipeline
  - ASAP2 Pipeline
  - InfiniBand/Ethernet
- **Network**
- **BlueField DPU**
STORAGE NVME-OF PERFORMANCE
Latency Determinism | IOPs Efficiency

Latency CDG for 4K Reads and Writes with PFC CC

Noise added by TCP to 4KB WRITE
Noise added by TCP to 4KB READ

RoCE Maintains the latency determinism of local SSD access

RoCE IO Processing @ 4KB is x4-5 Lower
DPU ENABLES CLOUD-NATIVE SUPERCOMPUTING

Multi-Tenancy with Zero-Trust Security

Collective offload with UCC accelerator
Smart MPI progression
User-defined algorithms
1.4X higher application performance
IMPROVING NON-BLOCKING MPI PERFORMANCE

44% Faster for MPI iAlltoall, 36% Faster for MPI iAllgather

Source: Courtesy of Ohio State University MVAPICH Team and X-ScaleSolutions
HIGHER APPLICATION PERFORMANCE

Up to 35% App Performance, MPI Collectives Offload

P3DFFT Performance

<table>
<thead>
<tr>
<th>Grid Size</th>
<th>Time (Microseconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>Not Offloaded</td>
</tr>
<tr>
<td>1,024</td>
<td>Not Offloaded</td>
</tr>
<tr>
<td>2,048</td>
<td>DPU Accelerated</td>
</tr>
<tr>
<td>4,096</td>
<td>DPU Accelerated</td>
</tr>
</tbody>
</table>

32 servers, Dual Socket Intel® Xeon® 16-core CPUs ES-2697A V4 @ 2.60 GHz (32 processes per node), NVIDIA BlueField-2 HDR100 DPUs and ConnectX-6 HDR100 adapters, NVIDIA HDR Quantum Switch QM7800 40-Port 200Gb/s HDR InfiniBand, 256GB DDR4 2400MHz RDIMMs memory and 1TB 7.2K RPM SATA 2.5" hard drive per node.

Courtesy of Ohio State University MVAPICH team and X-SCALE Solutions
DPU ISOLATES GEFORCE NOW CLOUD GAMING

Isolated and Secured Infrastructure | More Concurrent Users

GeForce NOW Pod

Game Engine 40ms
Encode 5ms
10ms
User Input 10ms
Decode 5ms
Render 10ms
<100ms

Game Engine 40ms
Encode 5ms
10ms
User Input 10ms
Decode 5ms
Render 10ms
<100ms

BlueField-2 DPU

Game Graphics Driver Video
SDDC Security Telemetry
NAT | DDOS | Reverse Proxy

Ethernet NIC

Game Graphics Driver Video
SDDC Security Telemetry
NAT | DDOS | Reverse Proxy

Driver Accelerator Streamer Accelerator
SDDC Security Telemetry
NAT | DDOS | Reverse Proxy
DPU ENABLES FULLY INLINE 5G NETWORK PROCESSING
Offload, Isolate, Accelerate 5G Infra

Accelerate 5G or AI - Fully fungible - Fully programmable
Support for CUDA, DOCA - Toolchains, SDKs, Libraries
Secure, Isolated, Accelerated data processing
No need to move data back and forth from accelerators to host memory
Domain specific acceleration for 5G, AI, Network Security
Independent DPU time domain (5T for 5G)
Fully optimized data path
SUMMARY

NVIDIA DPU Enables the Data Center as the New Unit of Computing

- The CPU can no longer do it all
- Must offload & isolate server infrastructure tasks to a DPU
- Effective DPU must offer hardware acceleration and security isolation
- To enable such effective DPU, need to develop broad software eco-system to utilize hardware acceleration across variety of disciplines (e.g. HPC, AI/ML, Storage, Networking, Security) - DOCA
- NVIDIA DPU & DOCA is a computing platform with rich stack optimized ideal for AI, bare metal cloud, cloud supercomputing, storage, gaming, 5G wireless, and more
- NVIDIA is committed to line rate performance every generation.