DPU-Based Hardware Acceleration: A Software Perspective

By Bob Wheeler
Principal Analyst

June 2021

www.linleygroup.com

DPU-Based Hardware Acceleration: A Software Perspective

By Bob Wheeler, Principal Analyst, The Linley Group

Data-processing units (DPUs) promise greater data-center efficiency, but low-level-programming requirements have hindered broad adoption. Nvidia aims to remove this obstacle using its DOCA framework, which abstracts the programming of BlueField DPUs. Furthermore, customers will be able to program future converged DPU+GPU hardware using the combination of DOCA and CUDA. Nvidia sponsored the creation of this white paper, but the opinions and analysis are those of the author.

A Path to DPU+GPU Convergence

Promising more-efficient data centers, data-processing units (DPUs) add another element to the heterogeneous-processing mix. DPUs are important to data-center disaggregation, allowing server processors to perform only compute tasks while the DPU handles data movement between networked compute and storage. Using DPU-based smart network-interface cards (NICs), cloud-service providers can save server-processor compute cycles for revenue-generating services. DPUs also handle network traffic more efficiently than a server processor, cutting data-center power. In storage systems, they can supplant a standard processor, handling the massive throughput of SSD arrays while consuming less power.

Several vendors now offer processors positioned as DPUs. After examining the product field, The Linley Group defines a DPU as a programmable network SoC that integrates all major functions from the network ports to the PCI Express (PCIe) interface. A high-bandwidth PCIe interface separates DPUs from programmable Ethernet switch chips as well as legacy embedded processors. Combined with an integrated data plane for high-rate packet processing, the PCIe interface suits DPUs to network-traffic termination in smart NICs and for connecting SSDs in storage-controller cards.

Nvidia entered the DPU field through its 2020 Mellanox acquisition, which brought with it the BlueField family. Now generally available, BlueField-2 serves in smart NICs and storage controllers, integrating up to 200Gbps Ethernet ports and a high-bandwidth PCIe interface. The chip integrates a high-performance eight-core Arm complex as well as a programmable line-rate data plane with in-line IPSec and TLS encryption. BlueField-2 includes a regular-expression (reg-ex) accelerator, which offloads string searches for applications such as intrusion detection, antivirus, and spam filtering. It also offers a public-key-cryptography engine, a true-random-number generator (TRNG), and secure boot. Its PCIe Gen4 x16 host interface can handle 200Gbps of network throughput.

At GTC 2021, Nvidia announced BlueField-3 for 2022 availability and disclosed its roadmap for BlueField-4, as Figure 1 shows. Using 16 Cortex-A78 cores, BlueField-3 will roughly quadruple compute performance relative to its predecessor while also doubling

network bandwidth. It will handle Ethernet and InfiniBand port speeds up to 400Gbps, and its PCIe Gen5 host interface will double the bandwidth available from a x16 slot.

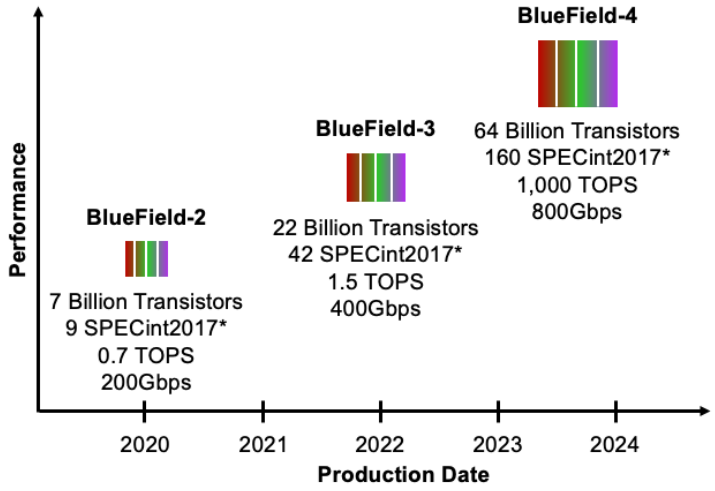

Figure 1. BlueField DPU roadmap. BlueField-3 will scale general-purpose-compute and network performance relative to its predecessor, whereas BlueField-4 will also add a GPU for AI acceleration. *SPECrate 2017 Integer.

Whereas BlueField-2 and BlueField-3 rely on their Arm cores for AI processing, BlueField-4 will integrate a GPU for AI acceleration. This will bring the chip's AI performance into the same class as leading-edge accelerators such as Nvidia's A100. In the meantime, the company plans to offer a BlueField-3X card using two chips to add a 75 TOPS accelerator, packing a DPU+GPU solution into a single PCIe slot. As developers add AI capabilities to cyber security, software-defined networking, cloud orchestration, and other applications, they can employ Nvidia DPU+GPU hardware and software to accelerate both AI and network processing.

BlueField Delivers Hardware Acceleration

Unlike server processors, DPUs are purpose-built for network-packet processing. Although architectures vary, most include programmable data planes as well as CPU cores for control-plane and application code. The DPU's dedicated data path is not only more efficient than using CPU cores, it's also far more performant.

As Figure 2 shows, the BlueField architecture essentially melds a NIC subsystem (based on ConnectX) with a programmable data path, hardware accelerators for cryptography, compression, and reg-ex, and an Arm complex for the control plane. In BlueField-3, the programmable packet processors comprise 16 cores handling 256 threads, enabling data-path processing with zero load on the Arm cores. In many applications, the data path handles known network flows autonomously, whereas the Arm cores handle exceptions such as new flows as well as control-plane functions.

Nvidia withheld BlueField-3 specifications, but BlueField-2's inline processing includes 100Gbps IPSec encryption and 200Gbps TLS encryption. Rated at 215 million packets per

second (Mpps), its programmable data path can implement a stateful flow table (SFT) as well as the NVMe-over-Fabrics (NVMe-oF) protocol. Given BlueField-4 will handle 400Gbps Ethernet and InfiniBand, we expect the company will double accelerator throughput as well.

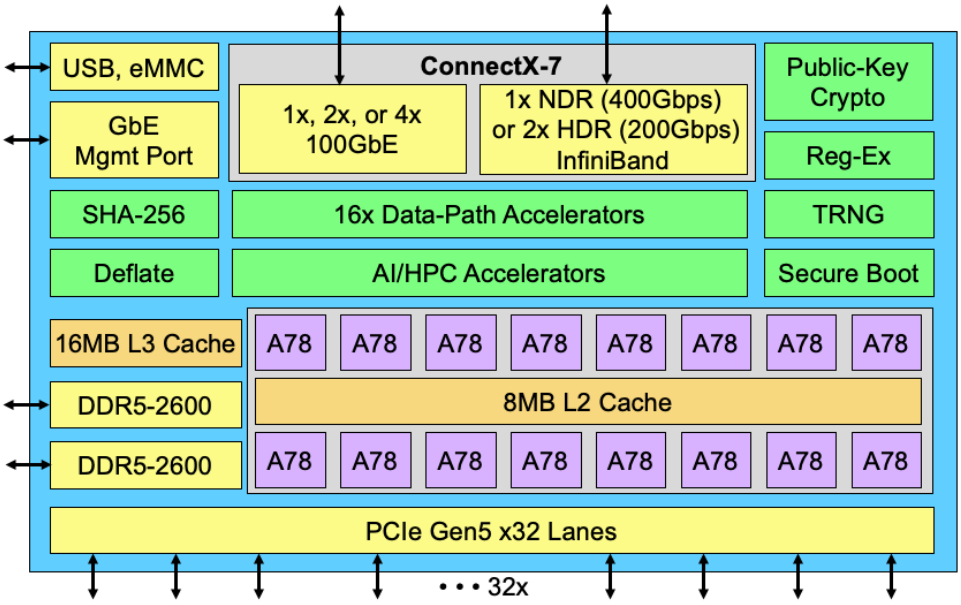

Figure 2. BlueField-3 DPU. The programmable data path combined with hardware-acceleration blocks enable line-rate processing without accessing the Arm complex.

For networking, the DPU accelerates advanced data-center SDN and network-function virtualization (NFV), including Open vSwitch, overlay protocols (such as VXLAN), network-address translation (NAT), load balancing, and fine-grained traffic management. For storage, the DPU accelerates RoCE (RDMA), NVMe-oF, data-at-rest encryption, data deduplication, distributed error correction, and data compression.

Whereas offloads free server-CPU cores for application processing, DPUs can also improve system security by isolating the application domain from the infrastructure-services domain. Cloud-service providers use this isolation to deliver bare-metal compute instances, hiding network and storage virtualization from the virtual server. This "air gap" between the host operating system and DPU-based services also protects the infrastructure from malware attacks launched on virtual servers within the data center.

DOCA Eases DPU Adoption

Although DPUs offer clear benefits, the requirement for customers to write low-level code limited early adoption to a narrow set of customers. To enable ISVs, service providers, and academia to adopt DPUs, Nvidia developed DOCA (data center on a chip architecture). DOCA is a framework of libraries, runtimes, and services built on a set of proven drivers. Some of the libraries relate to open-source projects, whereas others

are unique to Nvidia. Much like CUDA abstracts GPU programming, DOCA abstracts DPU programming to a higher level.

Figure 3 shows the DOCA 1.1 stack, which includes drivers, libraries, services agents, and reference applications. Nvidia delivers the stack through the combination of a DOCA SDK for developers and DOCA runtime software for out-of-the-box deployment. For example, ASAP² is the networking data-path driver and is supplied as a binary. It enables network-device emulation through VirtIO as well as low-level APIs that configure flow tracking and reg-ex accelerators. The security driver provides inline kernel offload for TLS. For storage, the SNAP driver provides NVMe virtualization, which presents remote block storage connected using NVMe-oF as if it were a local device.

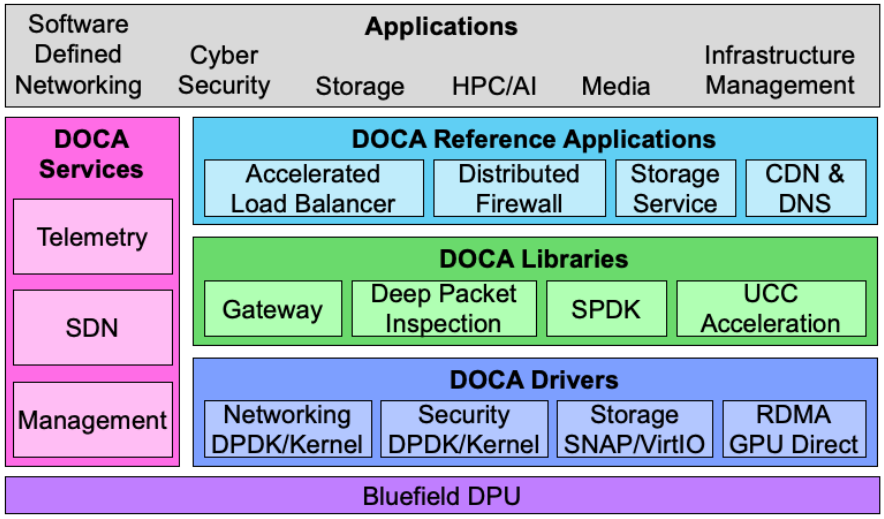

Figure 3. DOCA 1.1 stack. DOCA libraries provide application developers with high-level APIs, eliminating the need for low-level programming.

Moving up the stack, the Flow Gateway library implements a hardware accelerated gateway, building on the data path's SFT. Compared with the DPDK's generic flow API (rte_flow), the library provides higher level abstraction for gateway applications that filter and distribute network traffic. Similarly, the Deep Packet Inspection (DPI) library combines SFT and reg-ex acceleration, exposing a high-level API to the application layer. It enables unanchored searches of packet payloads against a compiled signature database.

For storage, DOCA supports the open-source Storage Performance Development Kit (SPDK), which provides user-space libraries. For HPC and AI, DOCA initially includes the Unified Collective Communication (UCC) library as a runtime component, with SDK support slated for a future release.

DOCA services include multiple telemetry agents, such as CollectX, NetQ, and WJH (What Just Happened), and a DPU Operations Tool for provisioning, management, and orchestration. Nvidia supplies standard OVS and OVS-over-DPDK applications for

high-performance network virtualization. It also provides source code for reference applications that implement load balancing and application recognition.

Beyond the 1.1 release, Nvidia will expand DOCA's application scope with additional drivers and libraries. To enable customer programming of the DPU data path, it plans to support the P4 language and the P4Runtime API. Planned libraries include time-as-a-service (TSDC) for telco and media applications, host introspection for out-of-band malware detection, and compression acceleration for the SPDK.

Although libraries eliminate low-level programming, many applications are not yet Arm-ready. To ease the transition, DOCA allows both development and deployment on an x86 host. An emulated-Arm container provides an x86-based development environment for DPU targets. For those developers unready or unable to port their application to the Arm architecture, however, Nvidia provides DOCA Runtime for x86. In this case, a gRPC client runs on the DPU and establishes a communications channel with the x86 runtime. The application can access DPU runtime components, and the developer needn't compile any Arm code.

DOCA Use Cases

An early DPU use case for networking is offloading virtual switching from the server CPU. By accelerating OVS using a BlueField DPU, Red Hat measured a 53x performance increase compared with using eight CPU cores for the function. It also performed a return-on-investment (ROI) analysis to show the cost savings of using a DPU in place of CPU cores. For a public-cloud data center hosting one-million virtual machines, it estimated an aggregate OVS switching requirement of 10 billion packets per second (10,000pps per VM). Red Hat then benchmarked standard OVS performance using the server CPU at only 43,750pps per core. Assuming 24 cores per server, the equivalent of 9,524 servers are needed just to handle networking.

To estimate ROI, Red Hat used a $8,500 server price and a $300 price premium for a DPU compared with a standard Ethernet NIC. For one-million cores, the total capital savings of using fewer DPU-enabled servers was $68.5 million. The DPU-accelerated servers, in aggregate, also consumed four megawatts less power, saving operating expenses as well. Given this analysis, it's easy to see why ecosystem partners including Red Hat and VMware are working with Nvidia to deliver in-box support for DPU-accelerated networking.

Using DPUs to offload OVS is straightforward, as the DPU data path can handle stateful tracking of known flows, whereas the Arm cores or host handle new flows. The data path's inline IPSec function can also offload encryption of OVS tunnels. By running the OVS control plane on the Arm complex, the DPU can fully offload vSwitch processing from the host. Other applications, however, can also benefit from BlueField's hardware accelerators for public-key encryption, reg-ex, compression, and hashing.

Figure 4 shows an example where the application is using the DPU as a lookaside accelerator as well as using it to offload the data path. A next-generation firewall (NGFW), for instance, can use SFT and inline IPSec in the data path, whereas it can

accelerate DPI and public-key cryptography through lookaside APIs. Other DPI use cases include URL filtering, intrusion detection, and application recognition. For ultimate offload and isolation, the ISV can port the application to the DPU's Arm complex, but even x86 applications can access all of the DPU's accelerators. Palo Alto Networks added DPU-based offload to its VM-Series NGFW while keeping the application on the x86 host and observed a 4–6x throughput increase, thereby enabling 100G Ethernet support.

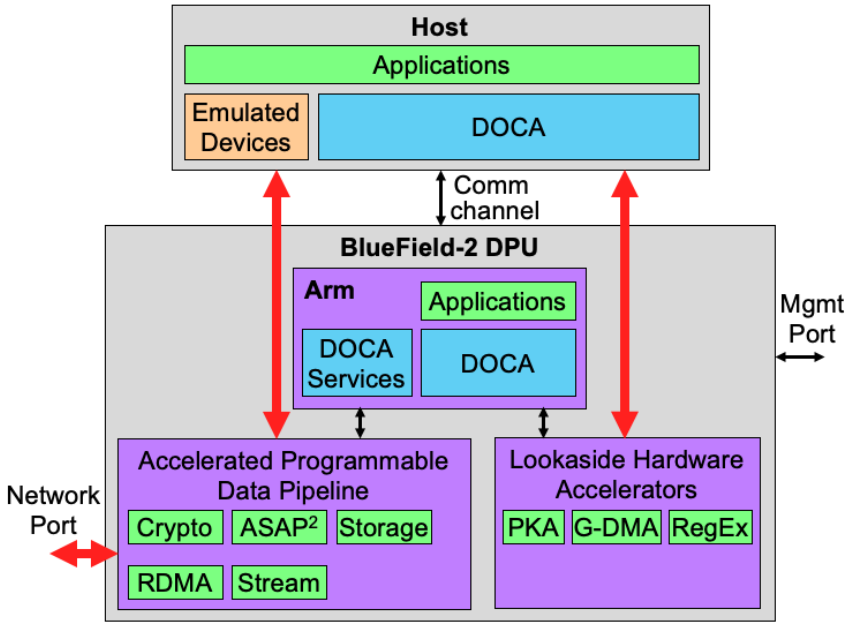

Figure 4. DPU-offload application example. Using DOCA, both x86-based applications running on the host and Arm-based applications running on the DPU can access the accelerated data path and lookaside accelerators.

Another cyber security example is Nvidia's Morpheus AI framework, which uses a natural-language-processing model to identify data leaks. Although the framework runs on a GPU-accelerated server, it uses distributed BlueField DPUs as sensors throughout the network. The DPU sends real-time-telemetry data to the Morpheus server, and the framework can respond to threats by pushing security policies to DPUs. By placing a DPU in every server, customers can implement micro-segmented security that's isolated from the server operating system and doesn't load the server CPU.

A unique use case for Nvidia's DPU is in storage virtualization and disaggregation. SNAP configures BlueField with a PCIe physical function (PF) for NVMe storage. To the server, this PF mimics a direct-attached SSD, but accesses actually go to networked storage. The DPU translates the native NVMe protocol into NVMe-oF packets and sends them to the storage target, such as an all-flash array. The DPU data path handles NVMe-oF command capsules, minimizing utilization of the Arm cores. Because NVMe-oF uses RDMA, the network's added latency is negligible compared with SSD access times. DOCA's SPDK library provides user-space APIs layered on top of SNAP's transparent storage virtualization.

Conclusion

It's easy to see why DPUs—SoCs purpose-built for networking—handle networking, security, and storage tasks more efficiently than server processors. On the other hand, the ubiquity of the x86 architecture makes it the preferred target for application developers. For example, network-security vendors that traditionally sold hardware can license their x86 application as virtual appliances running on virtual machines. Portability requires developers to use high-level APIs, avoiding any dependencies on underlying hardware. Conversely, adopting DPUs has required custom low-level code, creating a barrier to application developers.

With DOCA, Nvidia aims to remove this obstacle by providing a higher level of abstraction for DPU programming. By providing runtime binaries and high-level APIs, the framework allows developers to focus on application code rather than learning DPU-hardware intricacies. Although Arm servers are seeing early adoption in public clouds, many application developers have a large x86-code base and aren't ready for an Arm port. For these customers, Nvidia's DOCA Runtime for x86 removes the Arm-port hurdle, allow them to adopt DPUs now and optimize later.

For AI, there are similar tensions between running code on an x86 server processor and accelerating it using optimized hardware such as a GPU. Despite increasing competition, Nvidia remains the leader in AI acceleration due in part to the maturity and breadth of its CUDA software. Open-source neural-network frameworks essentially use CUDA as the default solution for acceleration. This AI leadership places Nvidia in a unique position to deliver converged DPU+GPU solutions comprising hardware plus an integrated development environment combining DOCA and CUDA.

Just as CUDA supports backward and forward compatibility across GPU generations, DOCA enables developers to begin working with DPUs now using BlueField-2, knowing their code will run seamlessly on BlueField-3 when it becomes available. Similarly, developers can adopt Nvidia GPUs such as the A100 PCIe card now, knowing their CUDA code will work on BlueField-4 in the future. Nvidia's vision is for DPUs to become the third leg of heterogeneous computing, complementing CPUs and GPUs. DOCA is critical to achieving that vision across a broad set of applications.

Bob Wheeler is a principal analyst at The Linley Group and a senior editor for Microprocessor Report. *The Linley Group offers the most-comprehensive analysis of microprocessors and SoC design. We analyze not only the business strategy but also the internal technology. Our in-depth articles cover topics including embedded processors, mobile processors, server processors, AI accelerators, IoT processors, processor-IP cores, and Ethernet chips. For more information, see our website at www.linleygroup.com.*

