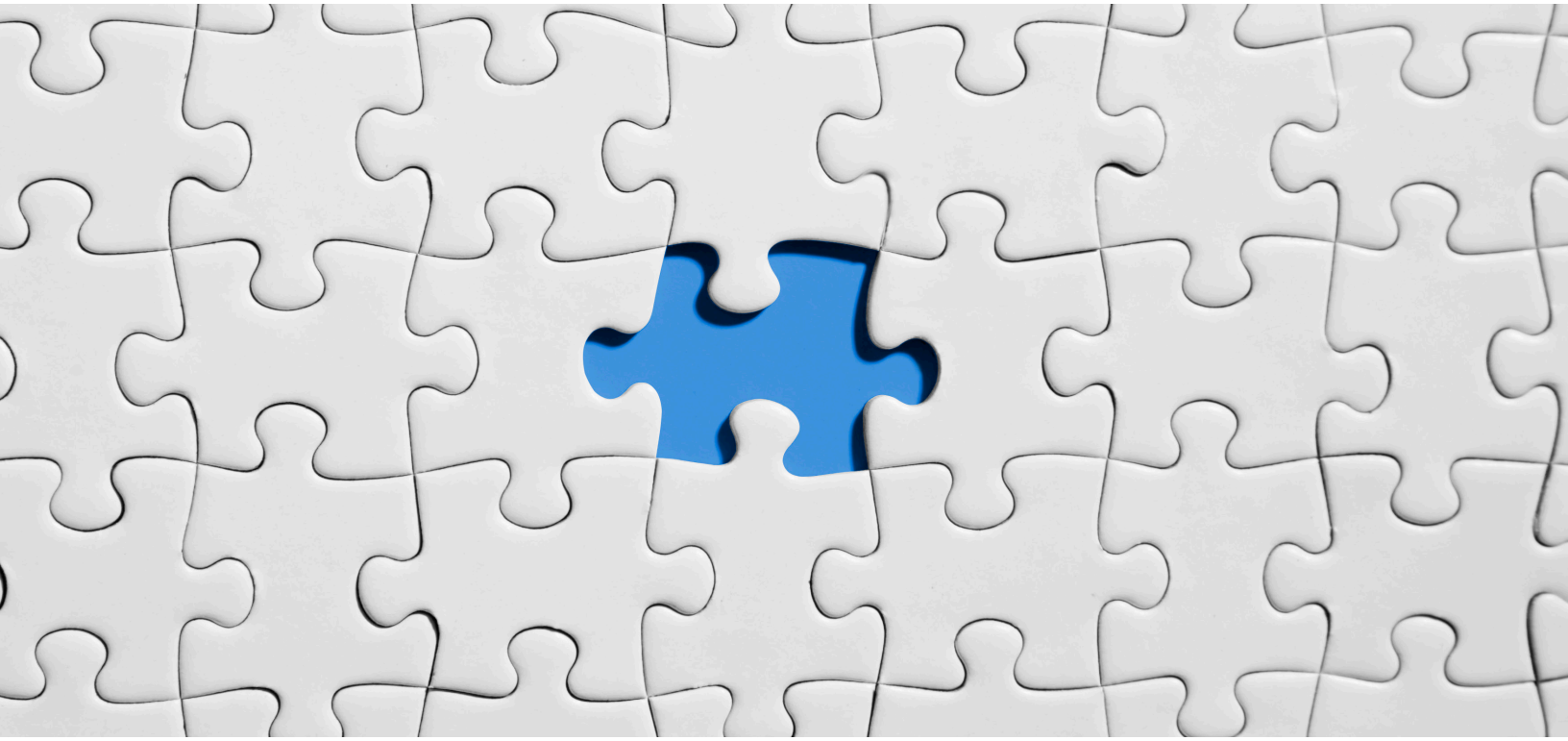
DELL Technologies

DPUS: ACCELERATION THROUGH DISAGGREGATION



Hussein Baligh
Senior Advisor, Sales Engineer Analyst
Dell Technologies
Hussein.baligh@dell.com

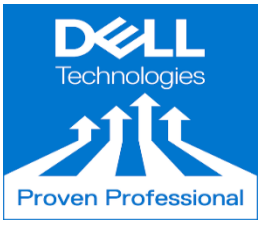
Ebrahim Serag
Sales Engineer Analyst
Dell Technologies
Ebrahim.serag@dell.com

Sameh Talaat
Senior Advisor, Customer/Technical Training
Dell Technologies
Sameh.talaat@dell.com

Yehia Gaballah
Sales Engineer Analyst, Data Protection Solutions
Dell Technologies
Yehia.gaballah@dell.com

DELL Technologies Proven Professional

Knowledge Sharing Article



The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

Table of Contents

Disclaimer: The views, processes or methodologies published in this article are those of the author. They do not necessarily reflect Dell Technologies' views, processes or methodologies.

1. Introduction

Year after year, Computing has been an extraordinary engine for innovation. The wide open doors for innovation made the ever-starving industry well fed for playing its expected role in implementing revolutionary ideas which took the field to a great extent of flexibility that allows the computing architecture to adapt to environments that are rapidly evolving with new technologies and demanding realities.

A perfect example of the above would be the history of the role of a CPU (Central Processing Unit). Initially serving as an all-purpose processor, it managed to do its intended job in the early days of compute systems. However, over time, the CPU was bombarded with extra processing tasks resulting from the evolution of the uses of compute systems. Introducing sophisticated visualization burdened the CPU with handling graphics-intensive computation while introducing high speed networking loaded the CPU with an extensive packet handling job.

"Jack of all trades, master of none" is a figure of speech used in reference to a person who has dabbled in many skills, rather than gaining expertise by focusing on one.[1]

Perhaps the above fittingly describes the situation of a traditional all-purpose CPU in a legacy compute architecture. Imagine a one-person IT help desk in a very small company. The so called “IT guy” is responsible for anything that might have the slightest hint of a link with technology. That won’t be limited to connectivity, software, and hardware issues but may even extend to include installing the projector on the ceiling or feeding the printer with paper in the copy room. In a larger company, the wide range of duties of that all-purpose IT guy will have to be distributed among a team of administrators that can focus on what they do best. For that, there will be a network administrator, a storage administrator, a backup administrator, etc.

Indeed, when it comes to the world of processing, segregation of duties was also a strategic decision. First, the GPU (Graphics Processing Unit) was introduced for two reasons. The most obvious one is offloading that burden from the CPU to let it focus its attention on application processing, but this is not the only reason. The other reason is handling the graphics computation much more efficiently. That gained efficiency is not achieved merely because of the added processing resources available in the GPU and it can’t be alternatively achieved by having faster traditional CPU. Instead, the actual reason behind that efficiency is that processing in the CPU and GPU serve different purposes and they are therefore optimized for what they were intended to do. A prime example to clarify this is how hyperthreading is implemented in those different processors. Hyperthreading is a technique to logically divide the processor into multiple slower processors by dedicating only a portion of the processor cycles to serve the threads resulting from a specific operation.

In a CPU, the main concern is that multiple applications may be running simultaneously on a compute system, so, hyperthreading can be applied for a better task scheduling. In such scenarios we can’t exaggerate in logically dividing the CPU so that the resulting logical CPUs are not too slow compared to the applications’ workloads. That is why hyperthreading would usually divide CPU cores by two only. In other words, for a few yet demanding applications running in parallel, we require a few yet high-speed CPUs to serve them.

In a GPU, the main concern is different. Though it is not required to execute different operations in parallel, it is required to execute the same operations millions of times as they have to be applied on a huge number of pixels; for instance, shifting pixels from one frame to the next while synthesizing the next frame to be displayed from a compressed movie file. As far as the GPU is concerned, what matters is having a large number of processors to do the job for all the pixels. In this context, hyperthreading logically divides the GPU into many slower processors unconcerned that each one of them would be too slow, since the task required of each is simple, and the main requirement is the processor count.

This comparison highlights that having a faster all-purpose CPU for everything can't be an alternative to the segregation of duties achieved by dedicating a processor to a specific kind of computation.

So, what is the next big thing? Today's compute systems deal with a very high speed of incoming and outgoing data transfers. Whether for accessing the network (especially after the introduction of 100 Gbps or faster networks), or for accessing storage (especially after the introduction of NVMe to access SSDs over PCIe interface as an alternative to SATA connected devices handled by the traditional slow AHCI controller). The high speed of the data transfers imposes a challenge on the CPU to process the data on time which draws its attention from serving the running applications, leading to performance degradation.

Earlier attempts at offloading some of the data processing burden was introduced by TOE NICs (TCP Offload Engine Network Interface Cards), an interface card equipped with a processor that could offload the TCP processing part such as handling acknowledgments, asking for re-transmission and re-ordering of packets. Later solutions invested more in the processing capabilities on the NIC allowing it to offload crypto functions as well. The term SmartNIC started to surface within the IT community to describe these devices, as well as some other vendor-specific terms. However, the industry lacked a standard vendor-neutral label until recently when the concept evolved into a more generic existence with a much broader range of benefits to computing architecture.

DPU (Data Processing Unit) was the term chosen to describe this game changer. The concept has evolved from a primitive NIC with a microcontroller to a fully equipped compute system with processors, memory, OS, storage resources and networking capabilities. It can be even a standalone system in a separate chassis connected over PCIe with the application server. One can imagine the performance when connecting that standalone DPU to storage resources. The application server would see it as a huge NVMe device connected over PCIe while the DPU will be dedicated to handle NVMe-oF (NVMe over Fabrics) communication and discover storage resources wherever they are.

Following the footsteps of the GPU, the DPU completes the strategy of segregation of duties, leaving the CPU practically as an application processing unit only that is fully attentive to executing applications' instructions to boost their performance.

As Jensen Huang, the CEO of NVIDIA puts it:

"*This is going to represent one of the three major pillars of computing going forward: The CPU for general-purpose computing, the GPU for accelerated computing, and the DPU, which moves data around the data center and does data processing,*"[2]

2. What is a DPU?

So, what exactly is a DPU?

Market research firm IDC categorizes DPUs as fitting into an emerging category that it calls functional offload coprocessors (FOCs). In coprocessor-based systems, one or more coprocessors are integrated into the architecture to take over some of the privileged operations instead of the CPU.[3]

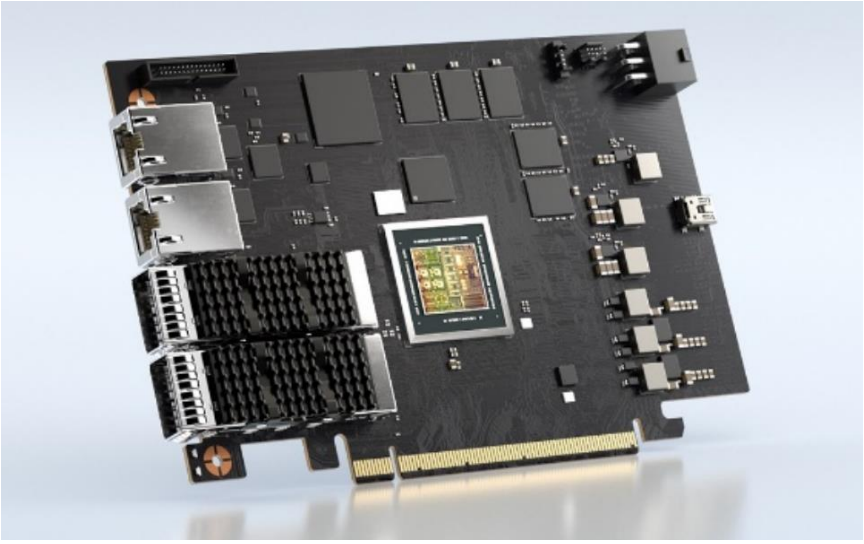


Figure 1: NVIDIA BlueField-2 SFF form factor, an example of a DPU [Ref 4]

A DPU is properly described as a System on a Chip (SoC), an integrated Circuit (IC) that is assembled of complete (or almost complete) computer systems. Such Systems (on Chips) include—but are not limited to—a CPU; for central processing of instructions, Memory; to store instructions or programs, and input/output ports; to connect to other components within the same device or otherwise.

Secondary storage is sometimes incorporated, as well as function accelerators. Processing units for specific applications (e.g. GPU) can also be incorporated in the same SoC as all the above.

Although there's much in common between Processing Units, DPUs are different from their predecessors (CPUs & GPUs).

A DPU should combine the following key pillars:

- Industry-standard, high-performance, software-programmable multi-core CPU
- High-performance network interface
- Flexible and programmable acceleration engines

Regarding the first pillar, one of the established architectures for DPUs are Arm architectures that are integrated with all SoC parts. This industry-standard architecture uses multi-core CPU that must insure high-performance and software-programmability. For the second pillar, a DPU's network interfaces should be able to parse, process and transfer data with efficiency, without creating bottlenecks, matching the speed of the high-performance network to attached GPUs and CPUs.[5]

The third Pillar is the most important of all three; it is the set of programmable acceleration engines that allow flexibility to offload applications that have burdened the CPUs in the past. Applications that will benefit from this acceleration include Artificial Intelligence and Machine Learning, Security, Telecommunications, Storage, and many others.

A SmartNIC (or Smart NIC) is an example of just such a DPU. You can see in the below diagram how all these components fit together.

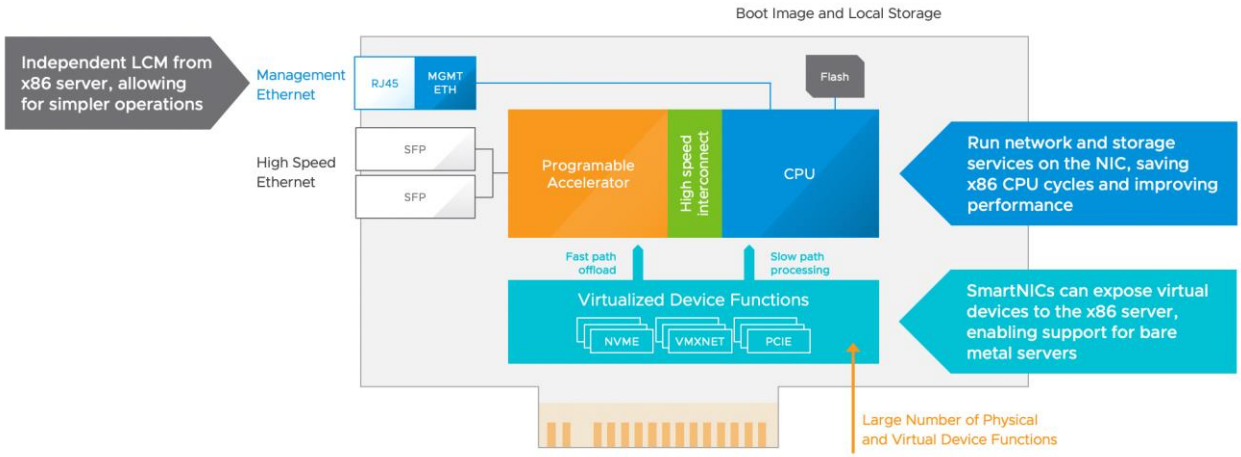


Figure 2: SmartNIC general layout and components [Ref. 6]

The DPU moves the compute function closer to where the data resides (data-centric architecture), replacing the outdated paradigm of moving data to where the compute is located (compute-centric architecture) freeing up CPU resources.

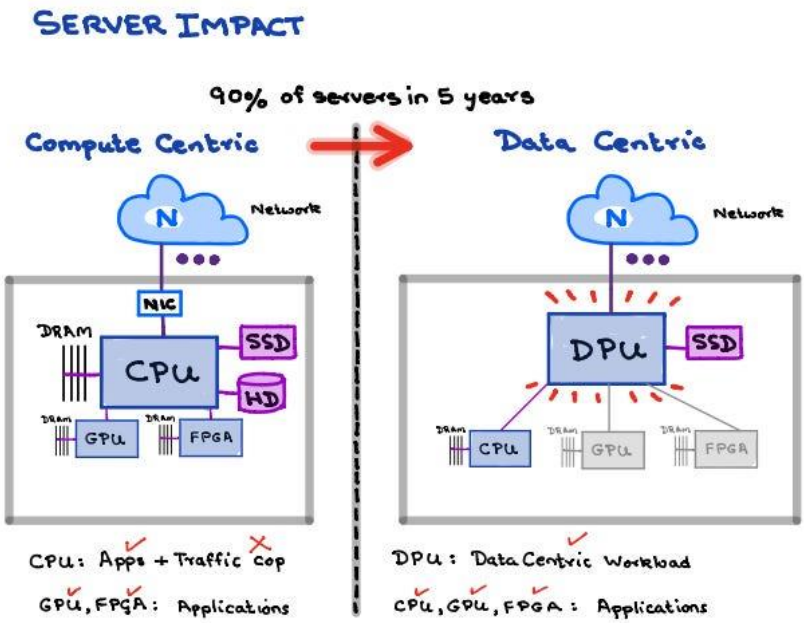


Figure 3: Compute-Centric vs Data-Centric architectures [Ref.7]

3. DPU Advantages

With that basic anatomy, DPUs can do some interesting things.[8]

On the networking side, the DPU can be used to offload secure connectivity in such a way that the host only sees a NIC while obfuscating the extra security layers. Cloud providers such as AWS use this to hide their proprietary networking details from their tenants. The DPU can also be used to offload NSX or OVS virtual switches, firewalls, or any other applications that need high-speed packet processing. Currently available DPUs support up to 200 Gigabits per second Ethernet and/or InfiniBand performance without the bottlenecks that are faced by applications at host CPU cores. This applies for traditional applications as well as modern accelerated Artificial Intelligence workloads in the presence of a GPU-accelerator.

For storage, the DPU in a host system can be presented as a standard NVMe device, while at the same time it can create a NVMe over Fabric (NVMe-oF) solution which uses actual NVMe storage from other servers in the data center. Similarly, the DPU can be directly connected to NVMe SSDs over PCIe and then exposed over the network to other DPUs in the data center, all without traditional host servers. DPUs can offload NVMe-oF Storage Direct, encryption, elastic storage, data integrity, compression, and deduplication, which allows remote storage performance with near latency of direct attached storage.

For compute, the DPU can be used to run a separate hypervisor than the main one on the server, thus making the x86 CPU or GPU compute or even FPGA just another type of resource to be pooled from multiple servers spanning across the whole data center, or simply to be attached directly to the DPU that exposes them to the network. In this way, any host in the data center can have access to the resources of the various processing engines, and any application on any host is free to take advantage of these accelerators, no matter where they physically reside.

For Cloud Service Providers, disaggregating data processing and flow of data in this way enables them to compose storage, networking and security anywhere within their infrastructure. Additionally, thanks to hardware accelerators, CSP operators are now able to present cloud tenants with bare-metal-as-a-service while properly controlling the server and keeping the environment safe and secure. Although this may sound theoretical to some, it has been practically proven already, for example being used by AWS, Alibaba and Baidu in their datacenters for years now.[9]

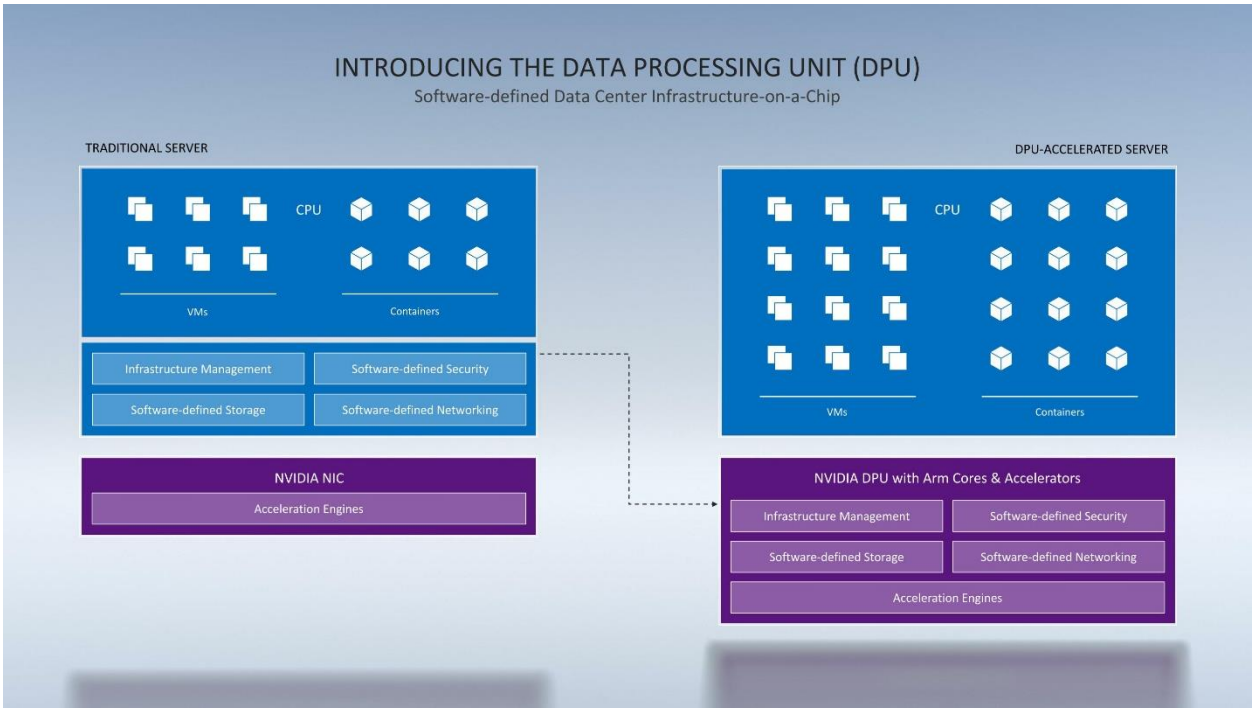


Figure 4: DPU-accelerated server offloads its software-defined infrastructure tasks off the CPU [Ref. 10]

Hyperscale Cloud providers have been the biggest adaptors and innovators in this area, using DPUs or Smart NICs to offload their stack on to the DPU, essentially running all their networking, security, management, monitoring, and telemetry on the DPU enclave, while freeing up the whole server to be provisioned to customers as a saleable bare-metal unit. In traditional data centers, it has been estimated that infrastructure management tasks take up to 30% of a server's CPU cores. For companies that make their money renting out CPU cores, offloading all this to the DPU and having the whole CPU be dedicated to the customer was a very attractive idea both financially and security-wise for multi-tenants.

With security-specific accelerators, DPUs offload security functions that are considered compute intensive. This not only makes such secure options available at no extra CPU consumption, it also minimizes latency; security functions are processed faster. Currently announced DPUs in the market guarantee security function acceleration and security at the edge of every server.

DPUs would be able to analyze data at the edge. This will allow storage, compute and security operations to be performed at low cost and low latency. High speeds and flow of data can be maintained. Another practical application of data analysis at the edge of current DPUs is storage for expanding workloads.

Current DPU manufacturers promote software development kits, making sure their equipment is categorized under Software-Defined Infrastructure. Developers can create high-performance, software-defined, cloud-native, DPU-accelerated services, leveraging industry-standard APIs.[3]

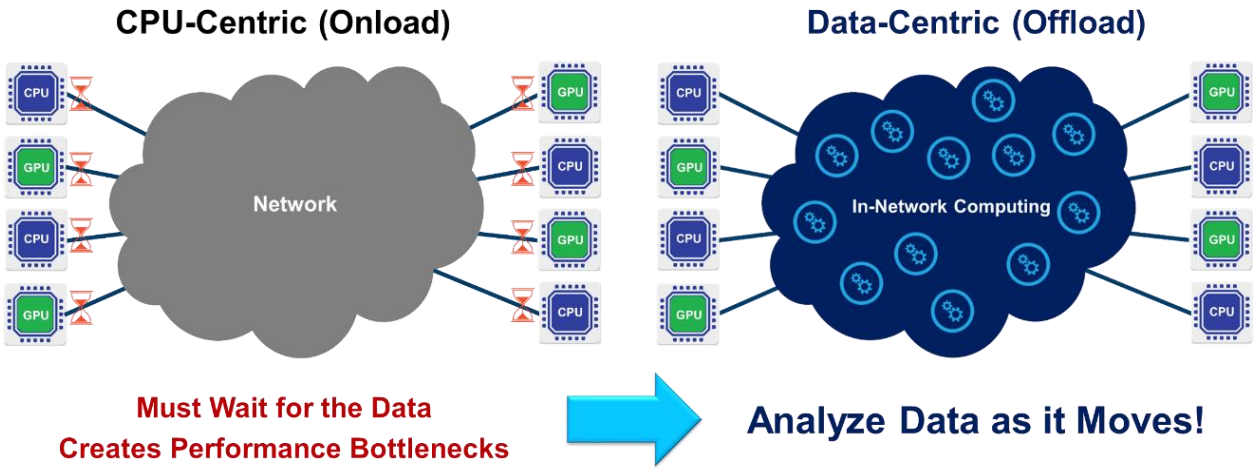


Figure 5: Figure 5: Compute-centric data center traffic operations vs Data-centric data center traffic operations enabled by DPUs [Ref. 11]

Another hidden advantage of releasing the resources from being blocked behind slow or overwhelmed CPUs is how DPUs enable disaggregation and pooling of all datacenter resources, delivering high-performance and low latency at scale. The DPU can enable heterogenous compute (e.g. CPUs, GPUs, FPGAs) and storage resources (SSDs) distributed across servers to be pooled together directly to maximize utilization, and in doing so increase TCO (Total Cost of Ownership). DPUs enable disaggregation and pooling of compute and resources at scale.[12]

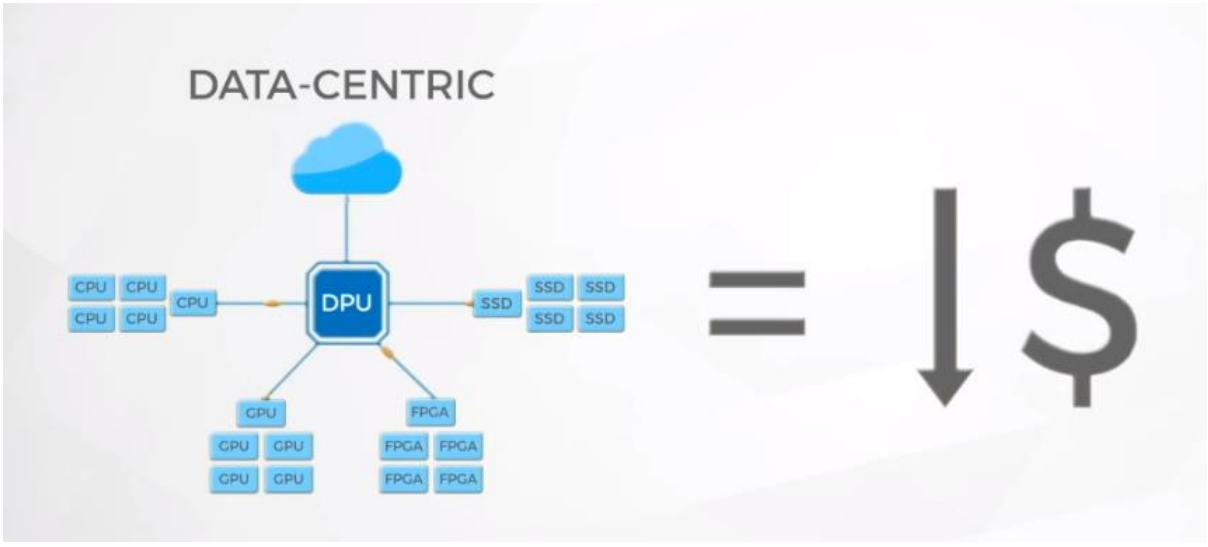


Figure 6: DPUs enable distributed heterogenous compute and storage resources to be pooled to maximize utilization and reduce TCO [Ref. 12]

4. Developing a DPU

4.1 What makes a DPU?

One would be wrong to assume that a strong embedded CPU at the heart of the DPU SoC is a complete (or even optimal) answer to all the above DPU requirements.

Unfortunately, some manufacturers and DPU developers have followed this brute approach of embedding a CPU at the edge of the server. We will argue that this is not efficient, or competitive.

We also point out how 100 Gigabit/sec packet processing can be a mountain of processing for a traditional x86 CPU. We ask those manufacturers as well as ourselves what chances an embedded CPU may have in performing better with respect to this common task of our time? We argue that the chances of the Embedded CPU to win this battle are minimal.

A Competitive DPU is one that leverages the application-specific Accelerator and components in a flexible way to satisfy all network data path processing. In such a case the embedded CPU is only responsible for control path initialization and exception handling. This, we believe, is a far better way to develop a DPU.

At the time of writing this, the recommended capabilities for network data path acceleration engines were as follows[13]:

1. Data packet parsing, matching and manipulation to implement an Open Virtual Switch (OVS)
2. RDMA data transport acceleration for Zero Touch RoCE (RDMA over Converged Ethernet)
3. TCP acceleration including RSS, LRO, checksum, etc.
4. Network virtualization for VXLAN and Geneve overlays and VTEP offload
5. Traffic shaping “packet pacing” accelerator to enable multimedia streaming, content distribution networks and the new 4K/8K Video over IP
6. Precision timing accelerators for telco cloud RAN such as 5T for 5G capabilities
7. Crypto acceleration for IPSEC and TLS performed inline, so all other accelerations are still operational
8. Virtualization support for SR-IOV, VirtIO and para-virtualization
9. Secure Isolation: root of trust, secure boot, secure firmware upgrades, and authenticated containers and application lifecycle management
10. Accelerators to bypass the CPU and feed networked data directly to GPUs (both from storage and from other GPUs)

The growing importance of Accelerator engines can be visible in the changes we see between 1st generation and 2nd generation DPUs.

Manufacturers are reducing the number of Embedded CPU cores and adding more Accelerator engines to offload more compute-intensive data path capabilities. These Accelerator engines will have data and metadata traveling between data center server CPUs, FPGAs, GPUs, NICs and NVMe SSDs.

Today’s data centers’ use of server CPUs as traffic cops will soon become obsolete due to their inability to match the high speed and low latency of the DPU.

4.2 Instruction Set Impact on DPU

Earlier we mentioned how the embedded CPU may not be the crown jewel of the DPU; in this section we will try to understand why the Arm architecture is more common in DPUs than x86.

First, let's get to know ARM: Advanced RISC Machine Processor.

Advanced RISC Machine (ARM) Processor is defined as family of CPUs that have wide usage in electronic devices such as smartphones, wearables, tablets and multimedia players.

RISC stands for Reduced Instruction Set Computer, which is a set of instructions "Reduced" in instruction length and complexity.

The ARM processor requires minimal number of instructions and operates on very low power with reduced circuit complexity, making it perfect for small-sized devices.

Advantages of ARM Processors[14]

1. **Affordability**
   ARM Processor is very affordable as it does not require expensive equipment for its manufacturing. Using it in DPUs keeps them at a competitive cost.

2. **Low Power Consumption**
   ARM Processors have lesser power consumption. They were initially designed for performing at lesser power. They even have fewer transistors in their architecture. This is considered a plus when addressing the environmental impact of DPUs and TCO of the devices.

3. **Faster**
   ARM performs at a ratio of one instruction per clock cycle. This will result in low latency of DPU operations when it is programmed correctly.

4. **Multiprocessing feature**
   ARM processors are designed so that they can be used in cases of multiprocessing systems where more than one processor is used to process information. Such is the case with most DPU architectures, where some DPUs scaled-up to 16 ARM cores.

5. **Load-store architecture**
   The processor uses load-store architecture that makes use of existing registers thus reducing memory interactions. It has separate load and store instructions that are used to transfer data between external memory and register bank.

6. **Simple Circuits**
   ARM processors have simple circuits; hence they are very compact and do not use up space on the chip. This enables DPU chip designers to utilize space in a more efficient way (i.e. more space for Accelerator engines).

It is important to note that some DPU manufacturers use ARM Architectures while others satisfy their embedded CPU requirements using MIPS or other processor architectures where the above advantages apply.

4.3 Possible DPU Components

In this section we will cover the components of the NVIDIA/Mellanox BlueField-2 DPU as a representative example of this new category of products.[15]

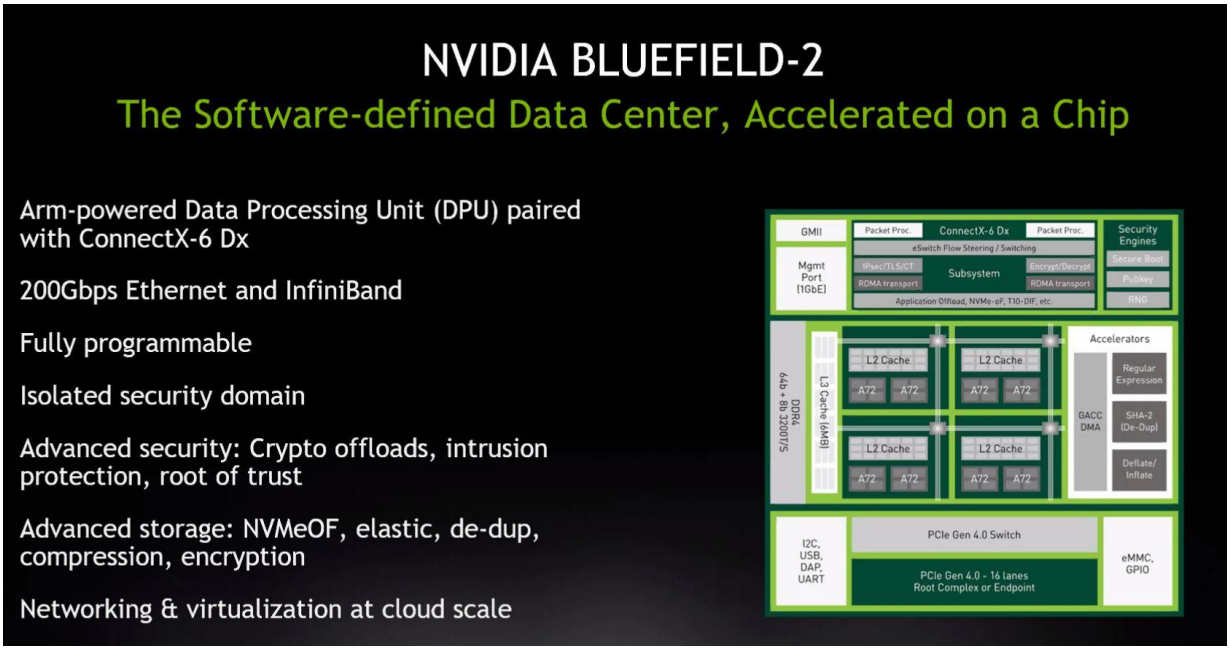


Figure 7: NVIDIA BlueField-2 DPU components & block diagram [Ref. 15]

Network Interfaces

- Ethernet - Dual ports of 10/25/50/100 Gb/s, or a single port of 200 Gb/s
- InfiniBand - Dual ports of EDR/HDR100 (100 Gb/s), or a single port of HDR (200 Gb/s) PCIe
- 8 or 16 lanes of PCIe Gen 4.0
- PCIe switch with up to 8 downstream ports

As expected, the Network Interfaces included are equipped to support and match the current network speeds and protocols.

ARM/DDR SUBSYSTEM

ARM Cores

- Up to 8 Armv8 A72 cores (64-bit) pipeline
- Arm NEON™ 128b SIMD execution unit
- 1 MB L2 cache per 2 cores
- 6 MB L3 cache with plurality of eviction policies

This manufacturer has chosen ARM Architecture, leveraging its advantages we had mentioned earlier. Note that the number of cores has decreased from the 16 cores of the 1st generation BlueField DPU.

DDR4 DIMM Support

- Single DDR4 DRAM controller
- 8 GB / 16 GB on-board DDR4
- ECC error protection support

Hardware Accelerators

Security

- Hardware root of trust:
  A hardware root of trust is the foundation on which all secure operations of a computing system depend. It contains the keys used for cryptographic functions and enables a secure boot process. The most secure implementation of a root of trust is in hardware making it immune from malware attacks.[16]

- Regular expression (RegEx) acceleration:
  For a CPU-based system, pattern matching is one of the most computation intensive parts. That is why using a Hardware Accelerator for this function is a great addition.[17]

- IPsec/TLS data-in-motion encryption:
  IPsec helps protect the confidentiality and integrity of your information as it travels across less-trusted networks.[18]
  Transport Layer Security (TLS) is a protocol which provides privacy between communicating applications and their users, or between communicating services.[19]

- AES key & data-at-rest encryption:
  The AES algorithm is a symmetric block cipher that can encrypt (encipher) and decrypt (decipher) information. Encryption converts data to an unintelligible form called ciphertext; decrypting the ciphertext converts the data back into its original form, called plaintext.[20]

- SHA 256-bit hardware acceleration:
  SHA-256 is one of the complex hash functions that are widely used. A hash is not 'encryption' – it cannot be decrypted back to the original text (it is a 'one-way' cryptographic function - and is a fixed size for any size of source text). This makes it suitable when it is appropriate to compare 'hashed' versions of texts, as opposed to decrypting the text to obtain the original version.[21]

- True random number generator (TRNG):
  The Random Number Generator is an essential IP core for all FPGA and SoC designs that target cryptographically secured applications.[22]

It is clear why these Security accelerators are essential when considering offloading security functions from CPUs. We took time to define each accelerated function to highlight its complexity and in turn highlight how much processing power is freed at the CPUs.

Storage

- NVMe SNAP
  NVMe SNAP makes networked storage look like a local NVMe SSD. This Accelerator offloads the Processing required to pull off such a feature.[23]

- NVMe-oF acceleration
  NVMe-over-Fabrics (NVMe-of) is a remote storage protocol specification which reduces the overheads of remote access to a bare minimum, thus greatly increasing the cost-efficiency of Flash disaggregation. This Accelerator enables NVMe-oF to meet latency expectations.[24]

- NVMe/TCP
  TCP can increase CPU usage because certain operations like calculating checksums must be done by the CPU as part of the TCP stack. This Accelerator offloads this duty from the CPU.[25]

- VirtIO-blk
  The virtio-blk device is a simple para-virtualization virtual block device. The FE places read, write, and other requests onto the virtqueue, so that the BE driver can process them accordingly.[26]

- Compression and decompression
- Data hashing and deduplication

Network

- RoCE, Zero Touch RoCE
- Stateless offloads
  - TCP/UDP/IP
  - LSO/LRO/Checksum/RSS/TSS/HDS
  - VLAN insertion/stripping
- SR-IOV
- VirtIO-net
- Multi-function per port
- VMware NetQueue support
- Virtualization hierarchies
- Ingress and Egress QoS
- Accelerations for tunneling technologies: NVGRE, VXLAN, Geneve
- Header rewrite (NAT)

These Network Hardware accelerators offload most – if not all – the CPU-draining network flow duties.

AI Accelerators

- GPUDirect / GDS (GPU Direct Storage)
  Enabling connection to GPUs in order accelerate AI operations

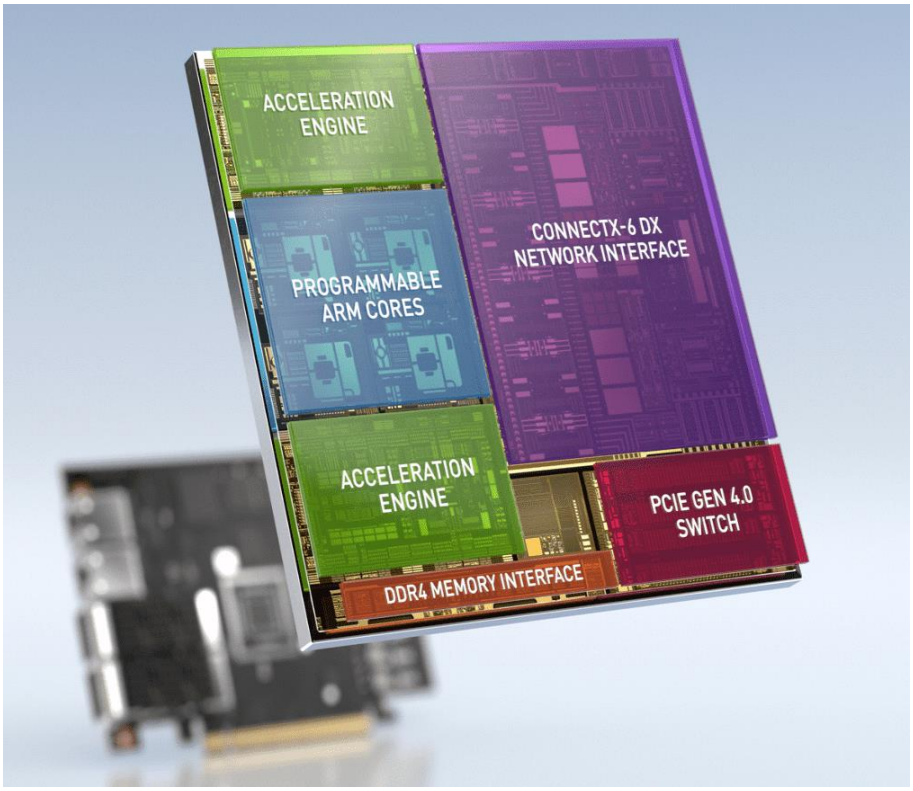


Figure 8: NVIDIA BlueField-2 Silicon Architecture [Ref. 15]

4.4 Other DPU Flavors

Flow Processors

Network Flow Processors (NFP) SoCs are used in some SmartNICs to provide in-field flexibility to cloud and data center operators. These chips use advanced IP blocks targeting data-intensive applications. NFP architecture focuses on distributed switch fabric that is also scalable, processing memory, accelerators and flow-processing cores implemented as Design Blocks and grouped into functional Logic Blocks. NFPs are programmable and use Run-to-Completion model. NFPs throughput ranges from 10Gb/s to 200Gb/s, all currently in production.[27]

Focus on FPGAs with no ARM cores

Some manufacturers focused on using an FPGA baseboard management controller (BMC) that manages the FPGA, much like ARM cores are used on other DPUs or SmartNICs to handle control-plane management. Such designs may not be categorized as DPUs (not a single SoC) but still tackle the same problem DPUs are here to fix.[28]

5. DPU Market Offerings: Vendors, Models, Hardware Comparison

5.1 Overview

Amazon AWS was a pioneer in this space with their AWS Nitro system, an early SmartNIC design that was architected to bring secure networking, storage, security, and system controllers to the AWS environment. This showed the need for commercial level DPUs in the modern data center.

Following AWS's success, others in the industry have worked to emulate and generalize this design to a broader audience, democratizing it for anybody to use. This set the bar for what a modern DPU (although the name hasn't been yet formalized) should be capable of and accordingly set the design specs.

As with all new technologies, there is still no one standard that every DPU manufacturer adheres to, but so far there are several commonalities in composition[8]:

- Industry-standard, high-performance, programmable multi-core CPU complex (usually ARM, but there are some MIPS-based solutions as well).
- Memory (mainly DDR4, but sometimes DDR5 or HBM).
- High-speed network connectivity (100GbE and up).
- High-speed packet processing acceleration, usually with P4 programmable logic, for parsing, encapsulation, decapsulation, lookup and transmit/receive acceleration, etc. (Usage of P4 makes it attractive to developers with networking backgrounds who would be familiar with it as a programming language for switches.)
- Programmable accelerators (for storage or cryptographic offload).
- PCIe Gen4 interfaces.
- Additional security and management features (e.g. for hardware root of trust).
- Able to run its own OS that is separate from the host system (usually Linux, with VMware ESXi now possible as per the announcement of Project Monterey).[29]

Adoption of the DPU presents an opportunity for data center networking folks to finally get a seat at the table inside servers with their own processor. To networking teams, data centers may be simplistically seen as data flows between where the data is stored, to where it will be processed, and how it should be routed. That may explain why development of many of the DPU solutions on the market is spearheaded by teams accustomed to data center networking infrastructure.[8]

5.2 Three Current DPU Solutions

In this section, we will look at some of the current DPU solutions out there and show some similarities and differences, and how well they align with the basic requirements as listed above. First is **Fungible F1**, then **Pensando Elba,** and lastly **NVIDIA-Mellanox BlueField-2** (which was briefly covered in section 4). These three are very different but all show the basic functionality of DPUs and how the industry players are innovating.

5.2.1 Fungible F1

First up is the Fungible F1 DPU. Fungible was one of the first companies to coin the name DPU for this new category of processors they were offering. The basic block diagram of the F1 DPU is shown below.

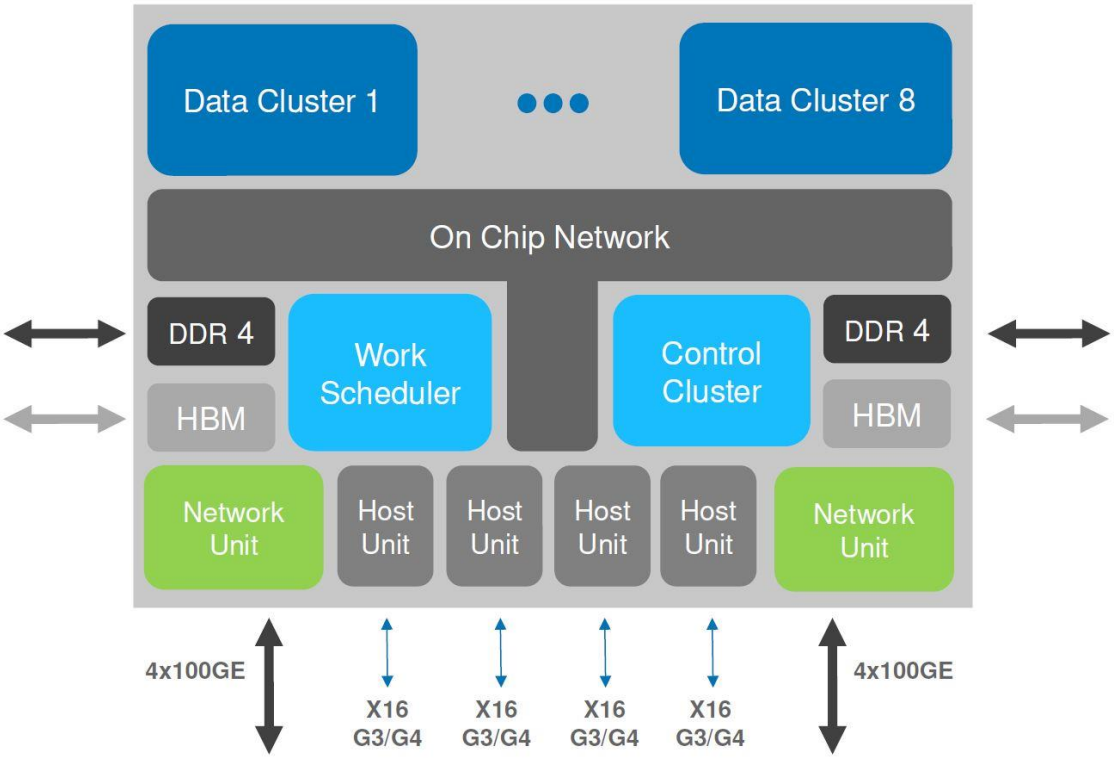


Figure 9: Fungible F1 DPU Block Diagram [Ref. 30]

Specifications:[31]

- **CPU core complex:** 8 data clusters of 4x SMT MIPS-64 cores
- **Memory controllers**: 2x DDR4 controllers plus support for 8GB HBM2
- **High-speed networking connectivity:** 2x 400Gbps network interfaces that are capable of up to aggregate 800Gbps or 8x 100GbE
- **High-speed packet processing acceleration:** P4-like language for parsing, encapsulation, decapsulation, lookup and transmit/receive acceleration
- **Accelerators:** Multiple accelerators including those for data movement
- **PCIe Gen4 lanes:** Four x16 host units that can run as root or endpoints
- **Security and management features:** 4-core x 2-way SMT control cluster with secure enclave, secure boot and a hardware root of trust. Other features are crypto engines and random number generation functions
- **Runs its own OS:** Linux

5.2.2 Pensando Elba

Pensando is a cloud startup founded by a group of well-known ex-Cisco engineers. As such, like the other two solutions, the Elba is a new processor coming from networking-oriented folks.[32]

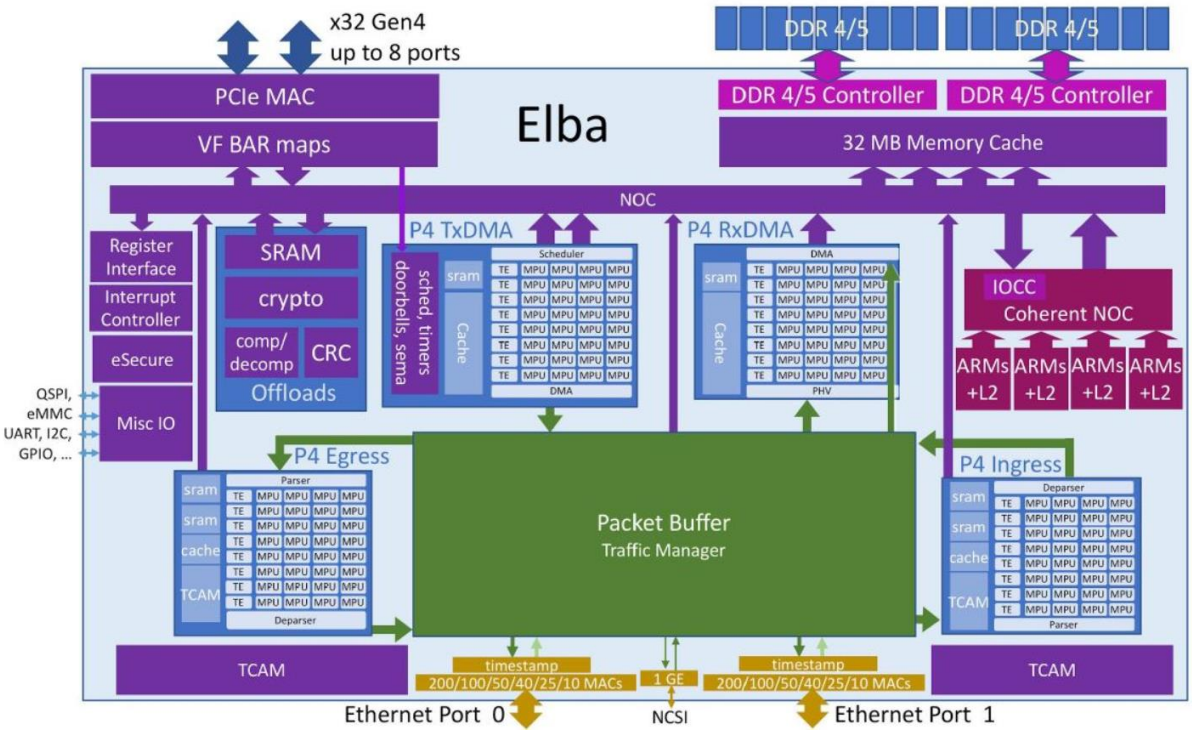


Figure 10: Pensando Elba Block Diagram [Ref. 33]

Specifications:[33]

- **CPU core complex:** 16 ARM A72 core complexes
- **Memory controllers**: Dual-channel DDR4/DDR5 memory support with 8–64GB. Pensando used HBM in previous models but switched to the cheaper and more flexible DDR.
- **High-speed networking connectivity:** 2x 200Gbps network interfaces
- **High-speed packet processing acceleration:** P4 programmable path
- **Accelerators:** For cryptography, compression, and data movement, among others.
- **PCIe Gen4 lanes:** 32x PCIe Gen4 lanes and 8 ports
- **Security and management features:** Hardware root of trust., with 1GbE interface for out-of-band management
- **Runs its own OS:** Linux with DPDK support, as well as VMware ESXi as per Project Monterey.[34]

5.2.3 NVIDIA BlueField-2

Lastly, we will review a unit with some history, since it started out as the first gen Mellanox BlueField, an NVMe-oF SoC solution that has been successfully used in some solutions such as the MiTAC HillTop NVMe-oF JBOF Storage solution. NVIDIA calls it a "seven billion transistor marvel", the BlueField-2 is a programmable data center infrastructure on a chip. The NVIDIA BlueField-2 was also known as Mellanox BlueField-2 before Mellanox's acquisition and renaming to NVIDIA Networking.[35]

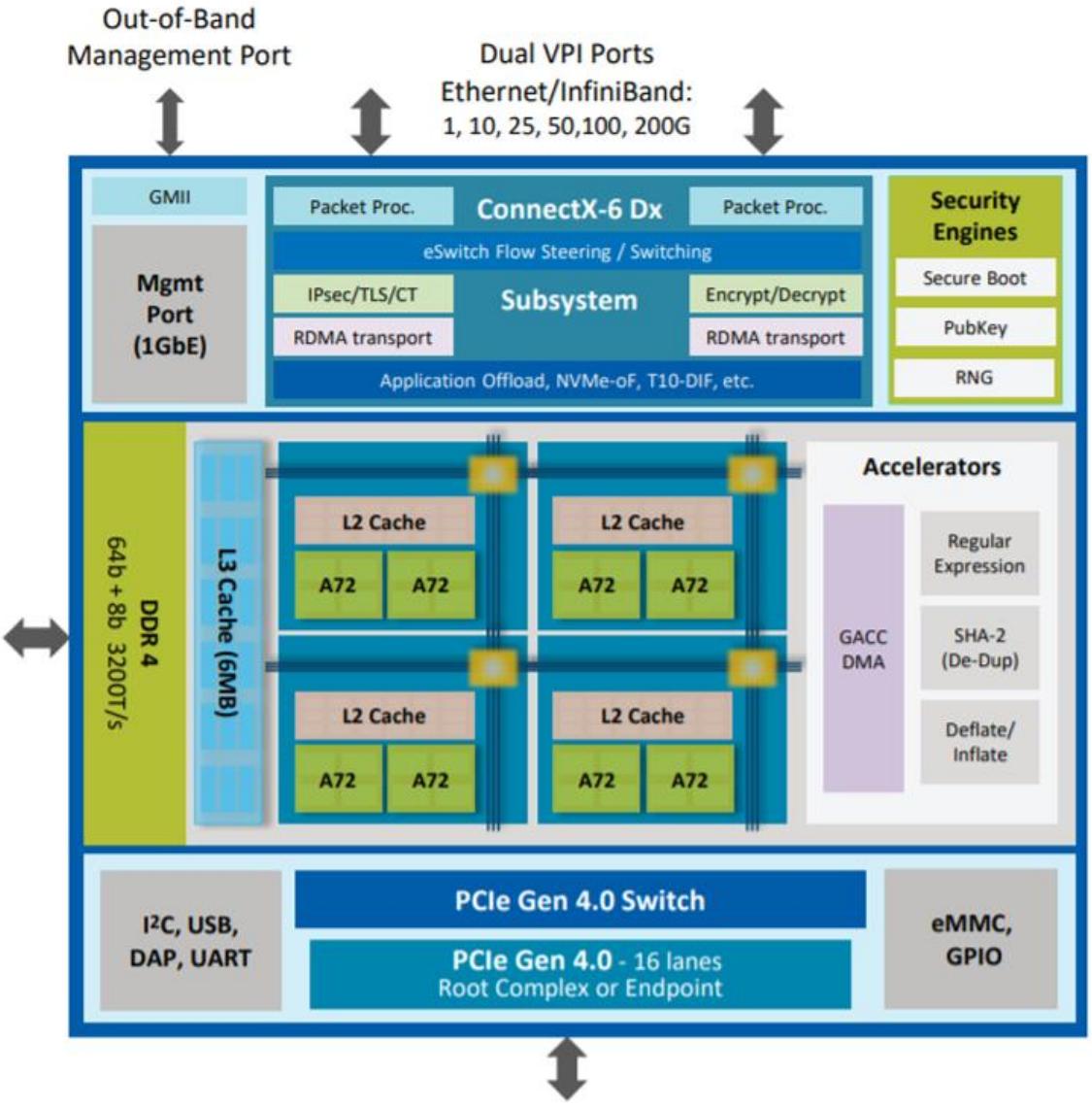


Figure 11: NVIDIA BlueField-2 Block Diagram [Ref. 8]

Specifications:[15]

- **CPU core complex:** 8x ARM A72 cores
- **Memory controllers:** 8Gb or 16Gb DDR4-3200 memory
- **High-speed networking connectivity**: 2x100Gbps or 1x 200Gbps Ethernet or InfiniBand, based on Mellanox ConnectX-6 Dx
- **High-speed packet processing acceleration:** Multiple offload engines and eSwitch flow logic similar to other ConnectX-6 Dx solutions
- **Accelerators:** For regex, dedupe, and compression algorithms plus crypto offloads.
- **PCIe Gen4 lanes:** 16 lane PCIe Gen3/4 PCIe switch
- **Security and management features:** Hardware root of trust, with 1GbE interface for out-of-band management.
- **Runs its own OS:** Many Linux distributions such as Ubuntu, CentOS, Yocoto, as well as VMware ESXi as per Project Monterey.[36]

We should remember looking at BlueField-2 that it was launched before the new Fungible and Pensando chips here were launched so it is almost like comparing a different generation of DPU.

5.3 Next Generation DPUs

When NVIDIA first announced its BlueField-2 DPU, it also presented the timeframe for its whole DPU roadmap for the next three years.[37]

Looking ahead to 2022, there will be a new generation of DPU: the NVIDIA BlueField-3 which will have 400Gbps networking. Since we know that a PCIe Gen4 link cannot handle 400Gbps, it is predicted that the BlueField-3 will support PCIe Gen5 and NVIDIA's propriety ConnectX-7 with 200Gbps/ 400Gbps networking. BlueField-3 is planned to deliver double the performance of today's BlueField-2, which means that NVIDIA will probably use updated and perhaps more ARM cores. Other DPU competitors are also pushing ahead with 16 core designs in this generation.

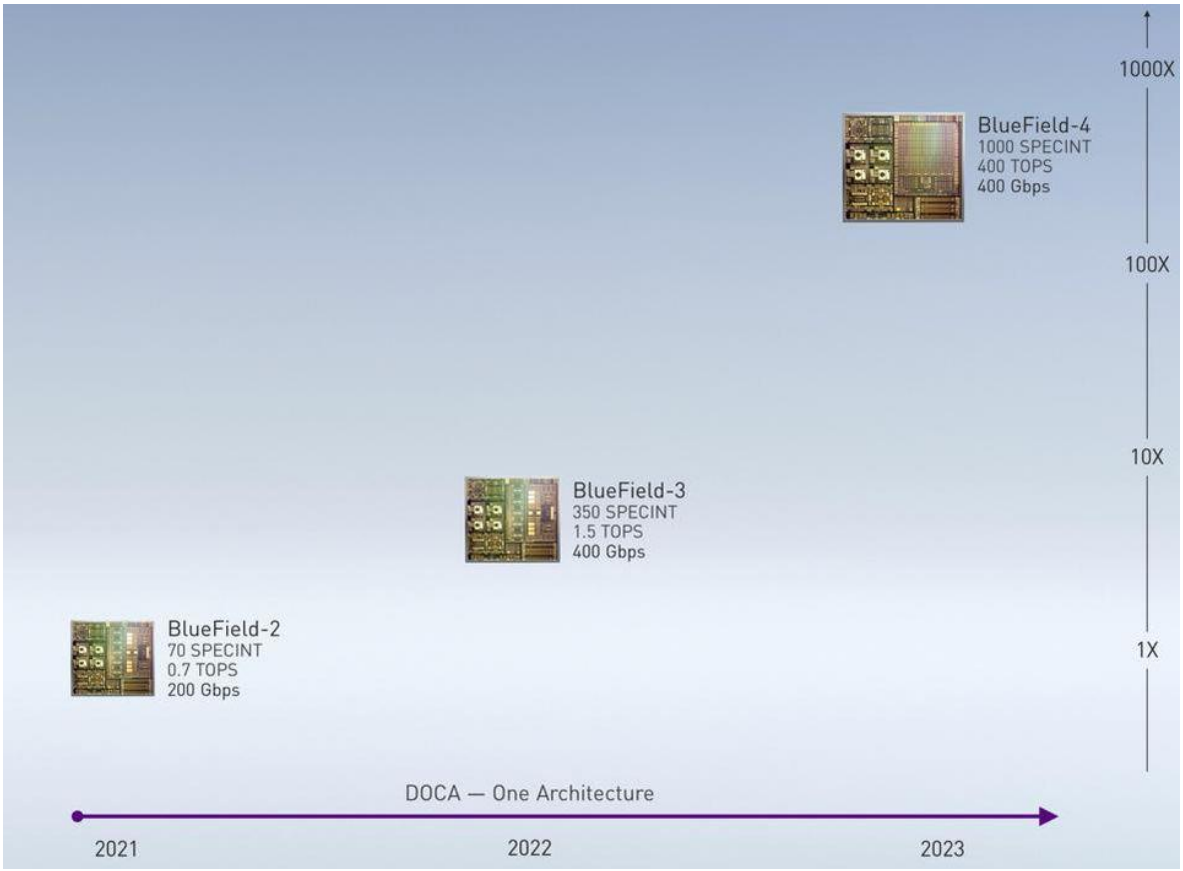


Figure 12: NVIDIA BlueField Family Roadmap [Ref. 38]

NVIDIA is also launching BlueField-3X in 2022 alongside BlueField-3. It is expected that BlueField-3X will include an updated GPU chip next to the DPU chip, with the chips linked together using PCIe Gen5 or perhaps Intel's new CXL standard.

In the year after that, we can see the "endgame" of what NVIDIA has been building up to since it first bought Mellanox and its BlueField chip in 2019/2020. The BlueField-4 DPU is planned for 2023 and seems to have no link between the DPU SoC and the GPU die, i.e. it will have an embedded GPU combining the DPU and GPU on the silicone onto a single hardware interface.

The performance numbers NVIDIA is promising for BlueField-4 will require capabilities far beyond today's x86 CPUs, almost equivalent to four current generation AMD processors. NVIDIA is claiming a 14x improvement in performance in only three years, which means that the CPU cores in the BlueField-4 SoC will be much more than just an offload engine.

Overall, these are exciting developments to look forward to. NVIDIA's vision is becoming clearer: it wants to deliver accelerated DPUs using GPU cores, ARM Cores, and Mellanox Networking tech all in a single package. For NVIDIA specifically, this means it can move its GPUs from being accelerators within servers to become network-connected resources managed by DPUs. For the industry as a whole, this is how large data centers can be transformed to the Data-Centric model we are moving towards.[38]

6. DPU Use Cases

Now that we have introduced what the DPU is and explained the theory behind its industry-changing capabilities, let's look at how the DPU is used in the real world.

6.1 Amazon AWS Nitro System

6.1.1 Overview

The AWS Nitro System has been called AWS's "secret sauce". Nitro was the core technology that enabled AWS to offer many valuable features such as bare-metal instances and to support a variety of hypervisors such as VMware on AWS. AWS proudly boasts that Nitro provides faster innovation, reduced costs, and increased security.[39]

But what is Nitro? AWS Nitro is a combination of software and hardware components that enhance the Amazon EC2 platform. Nitro is the underlying technology platform that AWS used to re-imagine their virtualization infrastructure, by offloading the hypervisor, network virtualization, and storage virtualization tasks to dedicated hardware to free up the main CPU.[40]

If that sounds familiar, it's because this dedicated hardware is one of the first examples of modern DPUs and their new place in cloud infrastructure.

The Nitro System is composed of three main components: Nitro Cards, Nitro Security Chips, and Nitro Hypervisor.[42]

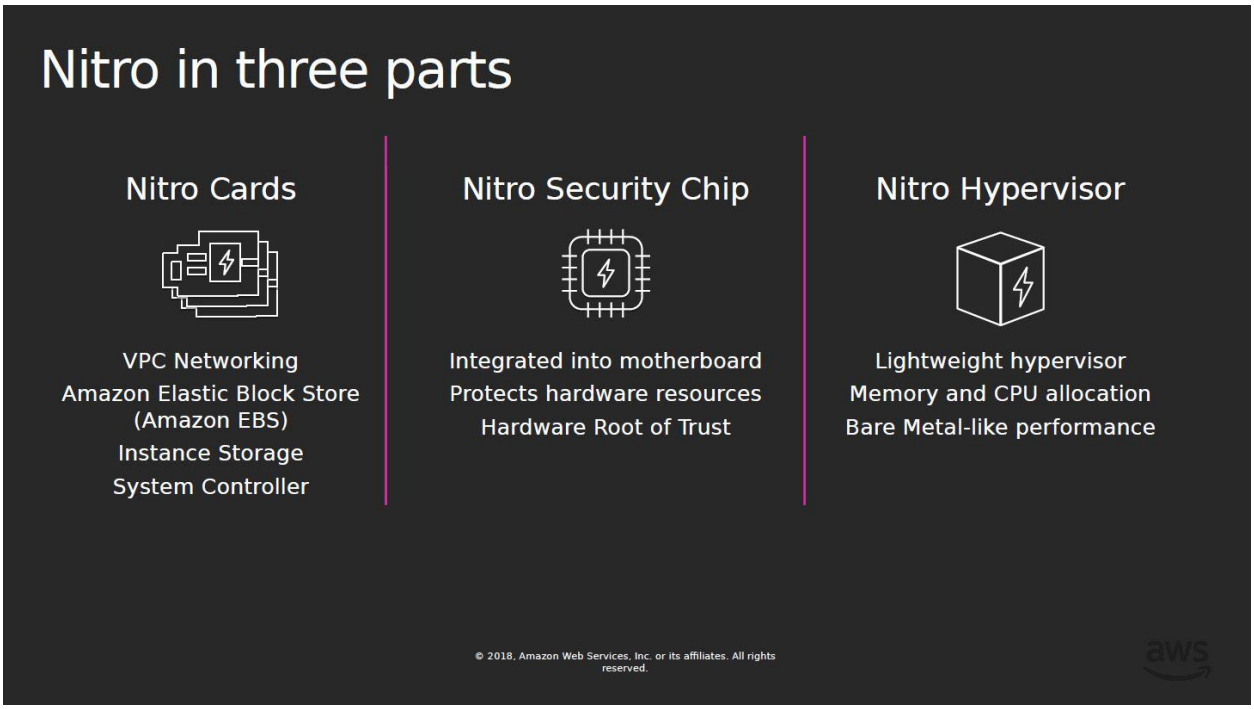


Figure 13: AWS Nitro System Components [Ref. 42]

Nitro Cards are a family of cards that offload and accelerate I/O for functions, such as Virtual Private Cloud (VPC), Elastic Block Store (EBS), and Instance Storage, which increases the performance of the overall system. These are the DPUs in the AWS architecture, and the template which other vendors have been using when developing their own DPU offerings.

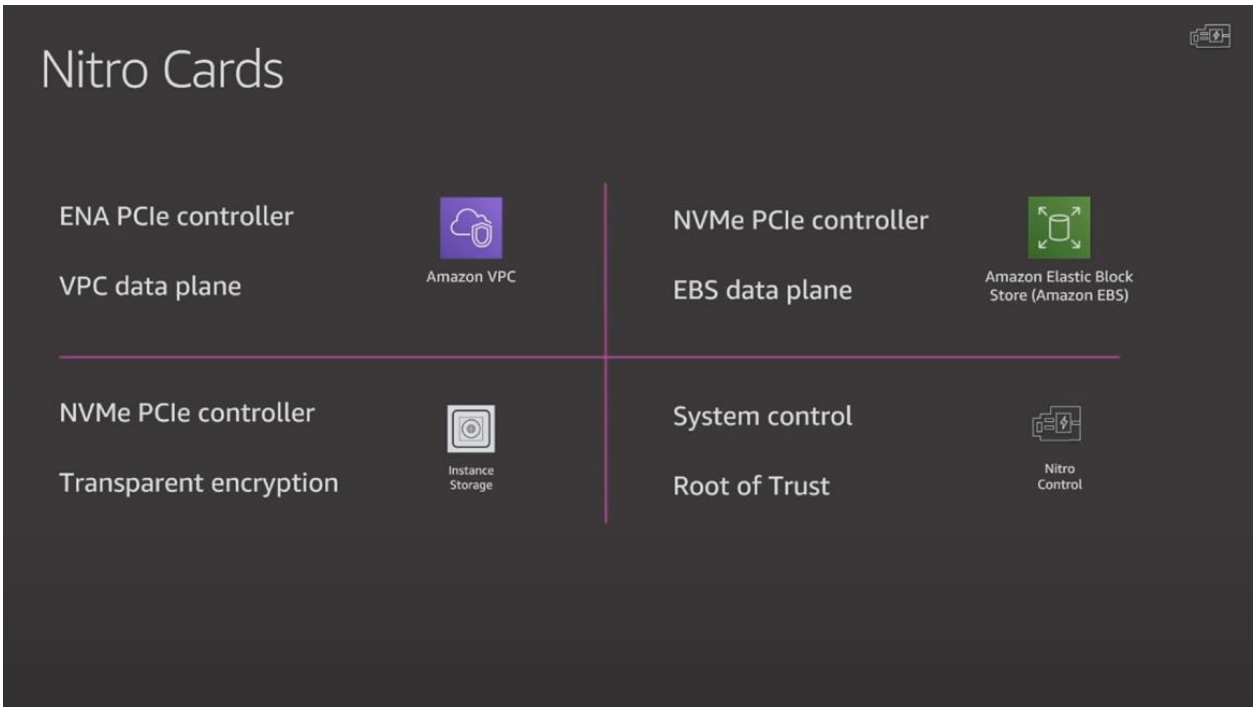


Figure 14: AWS Nitro family of cards [Ref. 42]

A brief overview of each of these cards or DPUs:[43]

- **Nitro Card for VPC** is essentially a PCIe-attached NIC or SmartNIC that handles the data plane tasks of encapsulation, security groups, enforcing limits, routing, and overall network acceleration.
- **Nitro Card for EBS** and **Nitro Card for Instance Storage** both handle storage acceleration tasks such as transparent encryption, SSD drive wear monitoring, and provides NVMe-oF access to remote storage volumes.
- **Nitro Card Controller** coordinates with all other Nitro cards and provides an API endpoint, implements the hardware root of trust using the Nitro Security Chip and supports instance monitoring functions such as measurement and attestation.

6.1.2 History

Werner Vogel, Amazon's CTO, laid out some of the history and reasons for the development of the Nitro System. After years of pushing traditional virtualization systems to their limits, it became clear to the AWS team that they needed to make a dramatic change to the architecture if they wanted to be able to keep up with the rising demands on performance, capacity, and security.[41]

Before Project Nitro, EC2 was built on a foundation that included a highly customized version of Xen hypervisor. However, it was found that in this architecture up to 30% of CPU resources was used for basic tasks such as CPU, storage, and network virtualization, leaving just 70% for the customer's applications.

AWS decided to offload the hypervisor and networking stack to a specialized hardware accelerator called Application Specific Integrated Circuit (ASIC). The change would be transparent to customers, who will benefit from the improved performance and lower costs.

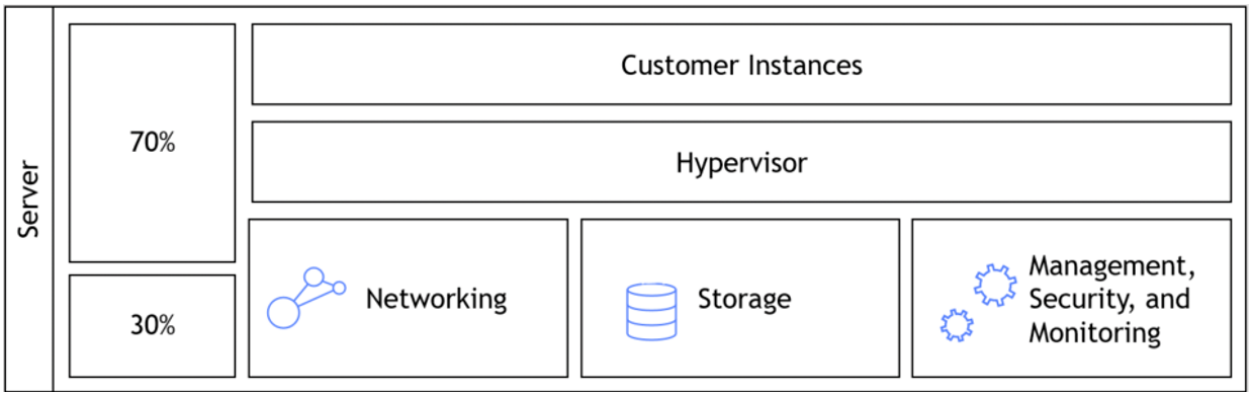


Figure 15: Pre-Nitro AWS EC2 instance host architecture for Xen Hypervisor [Ref. 41]

Thus, in 2013, AWS rolled out a totally new infrastructure that used dedicated software and hardware (based on AMD ASIC) to offload these basic networking and storage tasks and provide better value to their customers. Afterwards, AWS partnered with, and then acquired Annapurna Labs, which produces its custom-made chips and ASICs. It then replaced Xen with a wafer-thin lightweight KVM hypervisor that is highly optimized to operate with these ASICs that do all virtualization tasks. This combination is what's now called Project Nitro, and it delivered extremely fast virtual machines that almost have the same speed and performance of a bare metal server.

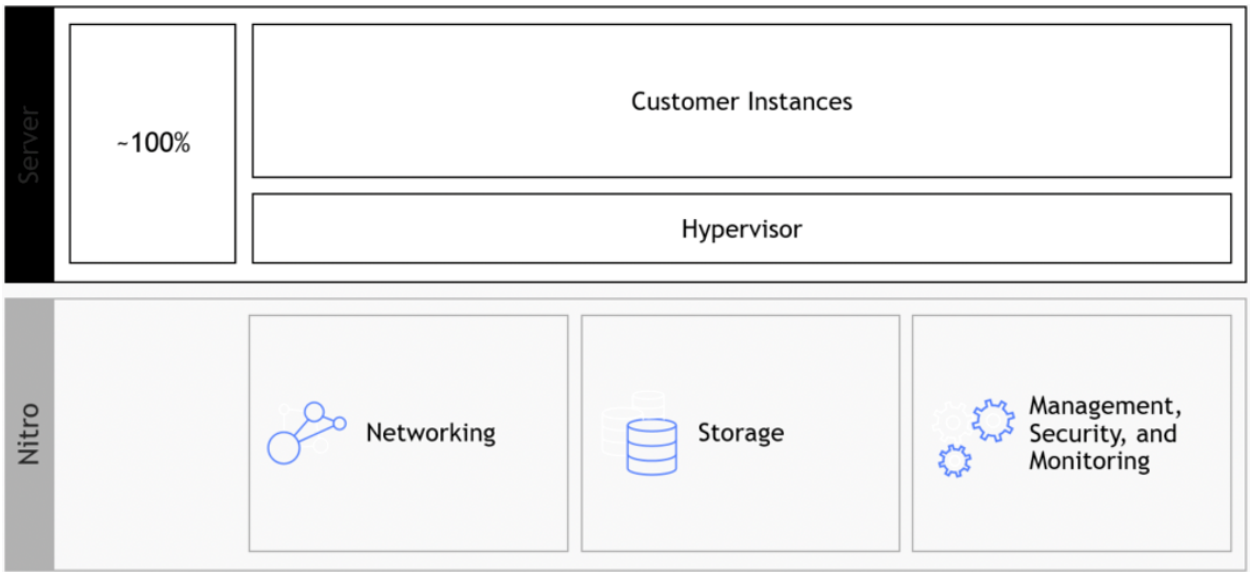


Figure 16: 2017 AWS Nitro System architecture [Ref. 41]

6.1.3 Advantages

The Nitro System does key network, server, security, and monitoring functions that frees up the entire main server for customer use. This allows EC2 instances to operate on all cores – there is no longer a need to reserve cores for "housekeeping" tasks such as storage or network I/O. This gives the customer better performance with less cost.

Nitro System also improves storage latency, which is achieved by offloading the overhead from local disk storage. Additionally, the Nitro System provides improved networking performance with up to 100 Gbps enhanced Ethernet networking.

Finally, security is improved since the Nitro System has a restricted API that makes it impossible to access customer data or change the system in unapproved ways.

The security enhancements provided by the Nitro system enabled AWS to develop a new offering it called AWS Nitro Enclaves which enables customers to create isolated secured compute environments to protect and process highly sensitive data. Since the Nitro Hypervisor is used on the edge cards (DPUs), this reduces the attack service area by providing CPU and memory isolation for the customer's Amazon EC2 instances where they can store sensitive data such as personally identifiable information (PII), healthcare, and financial data.[40]

6.2 VMware Project Monterey

6.2.1 Overview

At VMworld 2020, VMware announced its latest technology, code named Project Monterey, which fundamentally changes its hypervisor architecture, to respond to changing market dynamics and the evolving enterprise infrastructure landscape.[44]

Project Monterey utilizes SmartNICs or DPUs to deliver maximum performance, zero-trust security, and simplified operations. By leveraging DPUs to offload standard VMware components such as ESXi, VSAN, and NSX, Project Monterey extends VMware Cloud Foundation (VCF) to support bare metal operating systems and applications, whether in the data center, edge, or cloud. VMware is partnering with a broad set of SmartNIC and DPU vendors to deliver an integrated solution using devices such as Intel FPGA PAC N3000, NVIDIA BlueField and Pensando Distributed Services Card.[45]

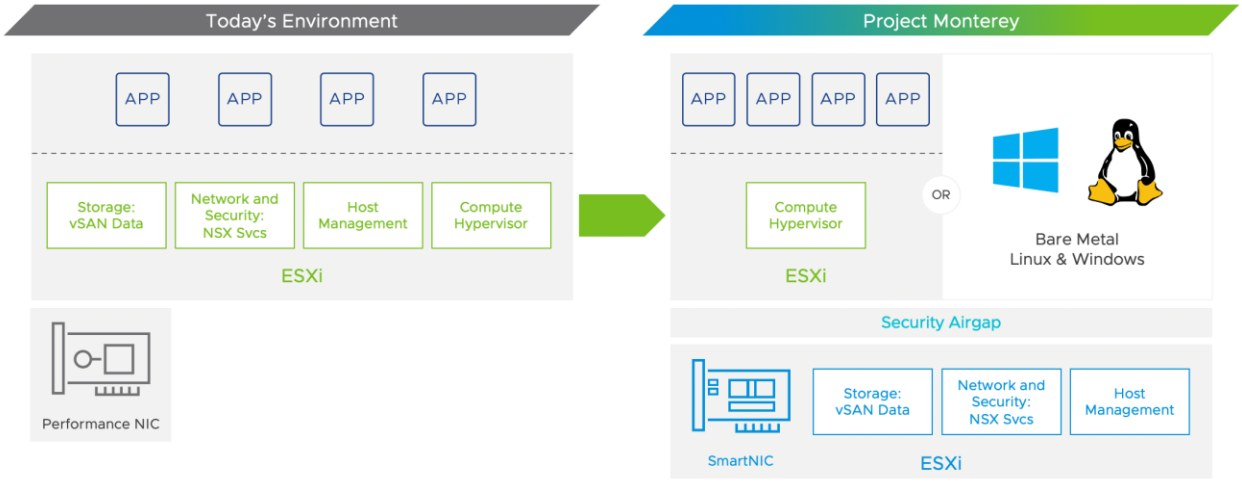


Figure 17: VMware Project Monterey offloads all virtualization tasks to the SmartNIC or DPU [Ref. 46]

Project Monterey delivers the following advantages:[46]

- **Network performance and security:** By offloading network and security functions to the DPU, the customer's scaling and performance challenges can be eliminated. VMware can achieve top performance with no core CPU overhead and can also deliver virtualized NSX network functions such as a distributed firewall with no network performance impact.
- **Cloud-scale storage and disaggregation:** Project Monterey accelerates storage functions such as compression, encryption, and erasure encoding by offloading them to the DPU without impacting the main server CPU performance, giving more flexibility and simplicity in delivering storage functionality.
- **Bare metal and composability:** Project Monterey finally provides enterprise customers and Cloud Service Providers with bare metal-as-a-service with vSphere, and can leverage VCF networking and storage services to bare metal workloads.

6.2.2 History

VMware was one of the leaders in the software-defined data center (SDDC) revolution. Unfortunately, the virtualization tasks for software-defined networking and software-defined storage required put a tremendous amount of pressure on the ever popular x86 processor. More pressure came from emerging trends such as machine learning, and all in all, the CPU is loaded by as much as 30% to manage tasks that do not belong to an application or a workload.

In 2019, VMworld announced Project Pacific, a re-architecture of VMware Cloud Foundation (VCF) to integrate Kubernetes into vSphere. Project Pacific led to VMware's Tanzu portfolio, a single platform that runs VMs and containers. Containerization leads to a microservices architecture that runs thousands of containers that additionally burden the x86 processor.

Project Pacific opened the floodgates to a stream of new, modern apps that need to run on VCF. These new applications are using more and more server CPU cycles, which brought the CPU to a breaking point. The solution is to offload basic hypervisor activities to specialized hardware accelerators, such as FPGAs, SmartNICs, and DPUs.[46]

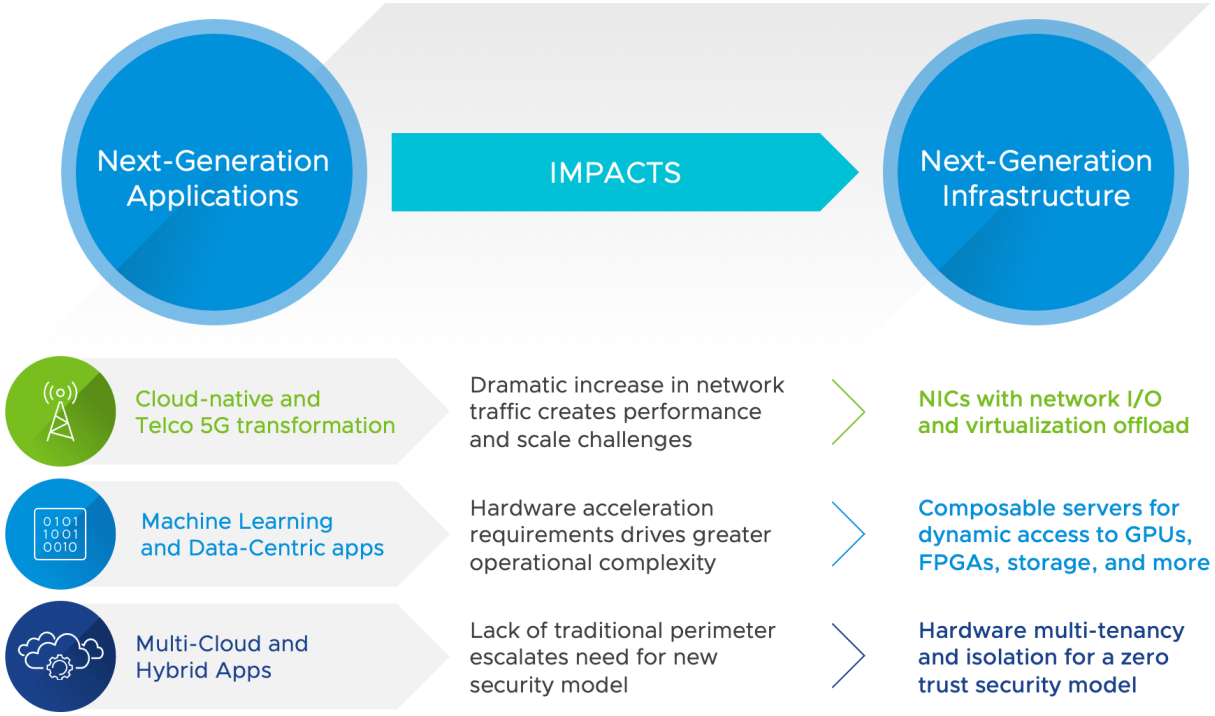


Figure 18: Next Gen applications require Next Gen infrastructure, VMware Monterey delivers it [Ref.46]

6.2.3 Advantages

Project Monterey can deliver the following key advantages, promising to completely transform customers' virtualized infrastructures:

- **ESXi on SmartNIC:** VMware is now capable of supporting ARM-based processors, which enables it to run on ARM-based SmartNICs and DPUs.
- **Two ESXi instances on the same server:** There can now be two ESXi instances running simultaneously, one on the main x86 CPU and the other on the DPU. The two ESXi instances can be managed separately for Cloud Service Providers (CSPs) who provide VCF-as-a-service.
- **Offloaded storage and network services:** You can run VSAN and NSX on the DPU. This improves storage and network I/O performance and reduces the load on the core CPU, leaving more CPU power for the apps.
- **Out-of-band host management:** The DPU ESXi can be used as a control point to independently manage the CPU ESXi, which allows us to improve Life Cycle Management and other functionality completely transparently to users and customers.
- **Security airgap:** Having an ESXi instance on the SmartNIC provides greater defense-in-depth. Even if the main CPU ESXi is somehow compromised, the DPU ESXi can still enforce proper network security and other security policies.
- **Zero-trust security model:** By offloading network security functions to DPU, VMware can provide comprehensive security capabilities without impacting application performance.
- **Bare metal OS support:** Since the DPU ESXi can manage the x86 CPU, we can now deploy any OS easily on the CPU, with no preference for ESXI over Linux or Windows. In essence, you can now have a bare-metal server that is managed through vSphere, which is very attractive to CSPs who can now offer bare metal server instances to their customers.
- **Improved performance:** by offloading network processing to SmartNIC and DPU, VMware can improve network bandwidth, reduce latency and free up core CPU cycles for better customer application performance.
- **Unified operations:** consistent operations across all apps, even those on bare metal OS. Simplified lifecycle management can dramatically reduce OpEx (operational expenses).
- **Virtualized device functionality:** DPUs can expose "virtual" devices on the PCI bus. These appear to the core CPU OS and applications as if they were actual hardware devices. This provides a level of flexibility not available before, essentially providing software-driven hardware.

- **Heterogenous HCI clusters:** In current Hyper Converged Infrastructure systems such as Dell EMC VxRail, there is a need for homogenous nodes so that any node has the same components as every other node. If you had a mixed-node cluster, with some nodes possessing certain hardware accelerators (such as FPGAs) while others don't, you can only use the acceleration abilities on the nodes on which they are installed. Thus, the cluster is usually designed for all nodes to have (FPGA) accelerators, which might not be needed by all applications and will only add costs. Project Monterey provides a much simpler solution: by exposing these hardware accelerators to the PCI bus, now all nodes in the cluster have access to their acceleration abilities, and any application on any node is free to take advantage of these accelerators, no matter where they physically reside. An example of such usage is shown in the diagram below.

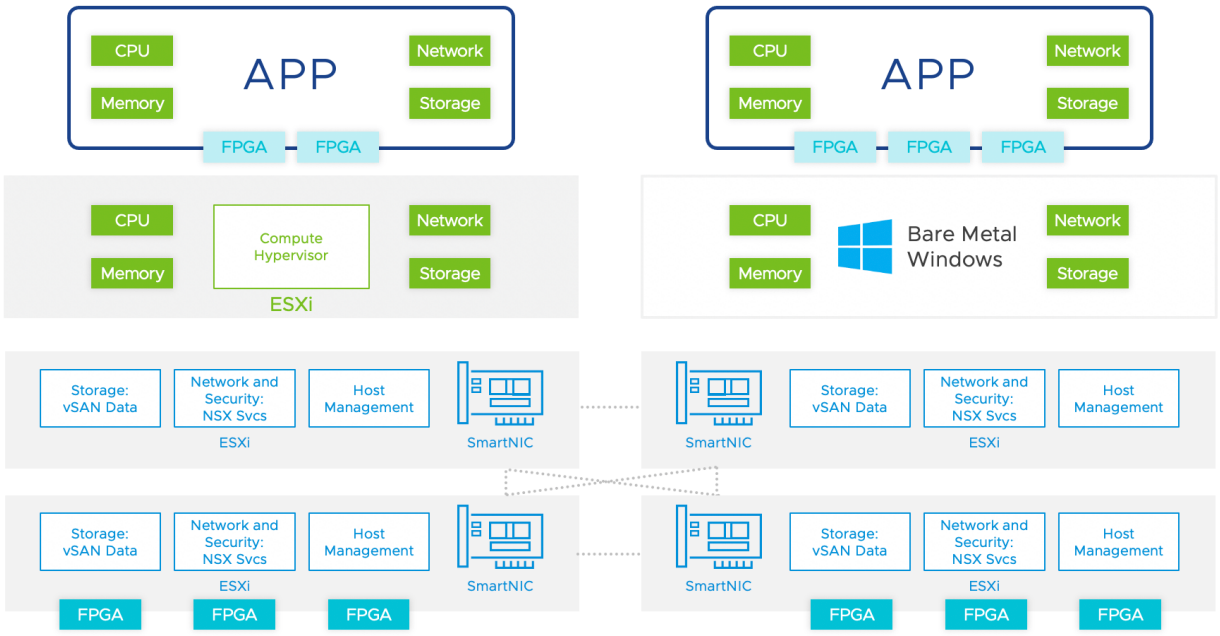


Figure 19: VMware Project Monterey allows applications in hosts without FPGAs to take advantage of resources in other hosts [Ref. 46]

7. Conclusion

If you ask people on the street what are the components of a computer, you might get a few different answers, but overall the answers would probably include: CPUs, disks, and memory; while the more tech-oriented would add GPUs and maybe NICs to the mix. Until very recently, an enterprise server would more-or-less have the same basic building blocks, but as time goes on and new trends in technology arise, a need for new components appears to handle these developments.

There is a pendulum in the computing industry that seems to swing periodically between consolidation, integration, and convergence on one side, and disaggregation on the other. For the past few years, data centers have been moving towards the hyperconverged model where compute, network, and storage are all in one box, with all tasks handled by the CPU. With the rise of the hyperscale data center, the pendulum shifted again to disaggregated data centers, with FPGAs and ASICs carrying out certain tasks suited for them, and now joined by the DPU to act as a bridge between all the different components.[47]

For decades, the x86 CPU has been the center of the data center, with everything connected to it and all data flowing through it. This means that CPUs were handling two roles at once: it's primary function of running apps, and the role of a data trafficker between the separate components. Until recently this was working perfectly well, with Moore's law doubling the processor speed to keep up with the demand. However, storage and network speeds are now skyrocketing, with hard disk storage being replaced with the much faster flash storage everywhere, and networking speeds rising from the paltry 1Gb and 10Gb of yesterday to the 400Gb and 800Gb of today and tomorrow's high-performance computing and hyperscale environments. What this means is that CPUs spend more time and energy handling their secondary duties – data processing – instead of the apps that they were built for.[48]

Enter the **Data Processing Unit (DPU)**. As the name implies, the DPU is designed for just that: data processing. The DPU moves data around the data center, but that does not just mean handling the networking. DPUs can be used to offload many of the tasks that don't have to go through the CPU, for example bypassing it entirely and connecting directly to high speed NVMe drives, or to the GPU, and presenting that data stream to another server in the environment. As such, the DPU becomes, as NVIDIA calls it, "*the third member of the data centric accelerated computing model*",[10] joining its older and more established siblings, the CPU (which still handles general purpose computing) and the GPU (for accelerated parallel computing).

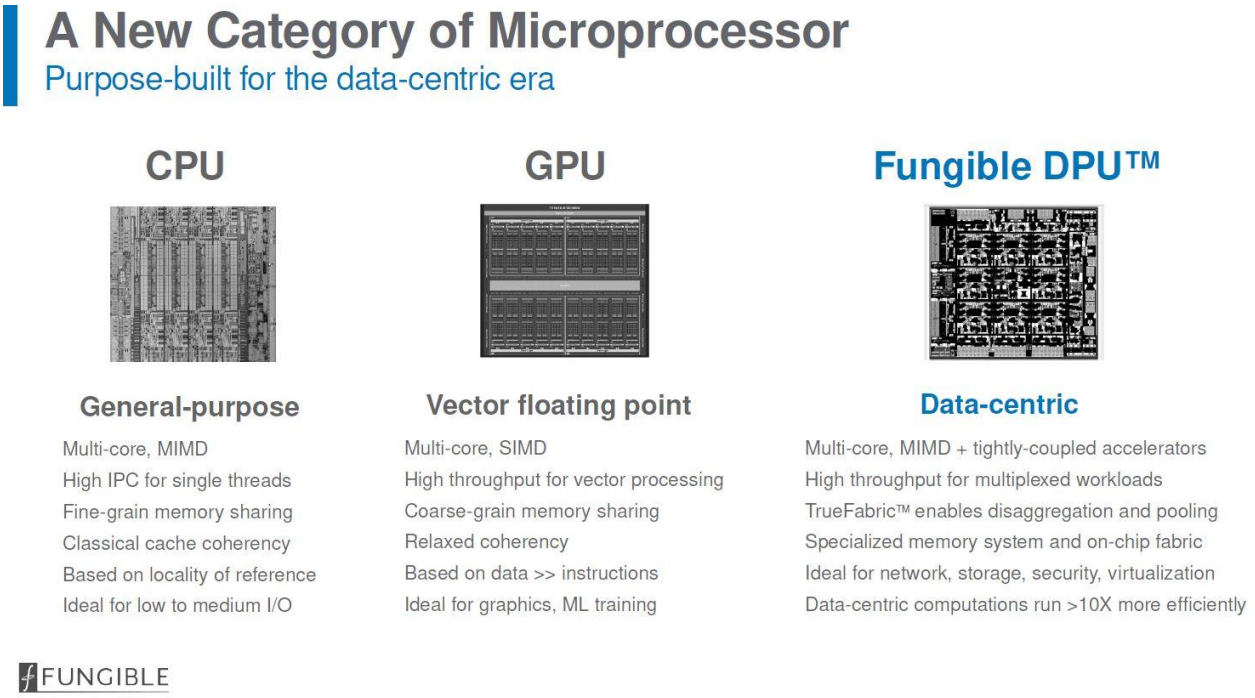


Figure 20: The DPU is the third pillar of the data center after CPU and GPU [Ref. 8]

The introduction of DPUs allow revolutionary use cases in multiple areas of the enterprise data center landscape, such as cloud provided bare-metal-as-a-service, offloaded virtualized high-speed networks, offloaded virtualized high-speed storage, and the disaggregation, pooling, and sharing of hardware accelerators, all of which make DPUs a very attractive prospect for customers like hyperscale web players, enterprises building out cloud architectures, and service providers.

This is an exciting time to work in tech. For years, the computing industry was dominated by the traditional CPU-centric architecture, with all data passing through the CPU that performs the task of trafficking data to the different components connected to it. The introduction of the DPU empowers us today to rethink data centers around their most precious commodity, the data itself, and thus puts us on the path to a totally new paradigm, to a truly data-centric architecture.

References

1. https://en.wikipedia.org/wiki/Jack_of_all_trades,_master_of_none
2. https://www.cadalyst.com/design-related-technologies/graphics-cards/what-dpu-third-pillar-computing-says-nvidia-75856
3. https://blogs.idc.com/2020/08/19/function-offload-coprocessors-a-new-era-of-decentralized-and-distributed-computing/
4. https://www.nvidia.com/en-us/networking/products/data-processing-unit/
5. https://blogs.nvidia.com/blog/2020/05/20/whats-a-dpu-data-processing-unit/
6. https://blogs.vmware.com/vsphere/2020/09/announcing-project-monterey-redefining-hybrid-cloud-architecture.html
7. https://www.businessinsider.com/fungible-series-c-softbank-pitch-deck-2019-6
8. https://www.servethehome.com/what-is-a-dpu-a-data-processing-unit-quick-primer/
9. https://www.theregister.com/2020/09/25/smartnic_dpu/
10. https://www.forbes.com/sites/janakirammsv/2020/10/11/what-is-a-data-processing-unit-dpu-and-why-is-nvidia-betting-on-it/
11. https://blog.mellanox.com/2020/03/welcome-to-the-dpu-enabled-data-revolution-era/
12. https://www.fungible.com/resources/dpu-a-new-category-of-programmable-processor/
13. https://www.fungible.com/articles/
14. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-arm-processor/
15. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/documents/datasheet-nvidia-bluefield-2-dpu.pdf
16. https://www.rambus.com/blogs/hardware-root-of-trust/
17. https://www.sciencedirect.com/science/article/abs/pii/S0141933112000610
18. https://www.ncsc.gov.uk/guidance/using-ipsec-protect-data
19. https://www.ncsc.gov.uk/guidance/tls-external-facing-services
20. https://www.nist.gov/publications/advanced-encryption-standard-aes
21. https://www.movable-type.co.uk/scripts/sha256.html
22. https://www.intel.com/content/www/us/en/programmable/solutions/partners/partner-profile/silex-insight/ip/true-random-number-generator--trng-.html
23. https://www.mellanox.com/sites/default/files/doc-2020/sb-mellanox-nvme-snap.pdf
24. https://dl.acm.org/doi/10.1145/3078468.3078483
25. https://blogs.oracle.com/linux/nvme-over-tcp
26. https://projectacrn.github.io/latest/developer-guides/hld/virtio-blk.html
27. https://www.netronome.com/media/documents/WP_Theory_of_Ops.pdf
28. https://engineeringjobs4u.co.uk/smartnic-architectures-a-shift-to-accelerators-and-why-fpgas-are-poised-to-dominate
29. https://www.servethehome.com/vmware-project-monterey-exsi-on-arm-on-dpu/
30. https://www.servethehome.com/fungible-f1-dpu-for-distributed-compute/
31. https://www.fungible.com/product/dpu-platform/
32. https://www.reuters.com/article/us-pensando-tech-idUSKBN1WV2R8
33. https://www.servethehome.com/pensando-distributed-services-architecture-smartnic/
34. https://pensando.io/pensando-vmware-and-project-monterey/

35. https://www.servethehome.com/mellanox-bluefield-2-ipu-smartnic-bringing-aws-like-features-to-vmware/
36. https://blogs.nvidia.com/blog/2020/09/29/vmware-bluefield2-dpus/
37. https://www.crn.com/news/components-peripherals/nvidia-wants-to-bring-new-bluefield-dpus-to-every-server-
38. https://www.servethehome.com/nvidia-shows-dpu-roadmap-combining-arm-cores-gpu-and-networking/
39. https://www.forbes.com/sites/janakirammsv/2019/03/10/how-an-acquisition-made-by-amazon-in-2016-became-companys-secret-sauce/?sh=6758517e2f67
40. https://aws.amazon.com/ec2/nitro/
41. https://www.allthingsdistributed.com/2020/09/reinventing-virtualization-with-aws-nitro.html
42. https://www.youtube.com/watch?v=rUY-00yFlE4
43. https://perspectives.mvdirona.com/2019/02/aws-nitro-system/
44. https://www.forbes.com/sites/janakirammsv/2020/10/20/vmwares-project-monterey-marks-a-major-shift-in-modern-infrastructure/?sh=1223d7c072af
45. https://www.servethehome.com/vmware-project-monterey-esxi-on-arm-on-dpu/
46. https://blogs.vmware.com/vsphere/2020/09/announcing-project-monterey-redefining-hybrid-cloud-architecture.html
47. https://www.nextplatform.com/2020/10/05/why-the-dpu-is-more-important-than-the-cpu-for-nvidia/
48. https://siliconangle.com/2018/11/30/junipers-founder-turns-to-custom-silicon-for-handling-the-data-tsunami-mayfield50/