AI NETWORKING:

The Role of DPUs

Read the IDC Analyst Brief

Sponsored by Chelsio Communications

With research by:
Brandon Hoff
Research Director, Enabling Technologies:
Networking and Communication, IDC

ANALYST BRIEF

Sponsored by: Chelsio Communications

This Analyst Brief covers use cases for data processing units in CPU offload, networking, security, and storage and the benefits in datacenters for hyperscalers, cloud service providers, high-performance computing, and the enterprise.

DPUs Enable Secure, Flexible, and Efficient Semiconductor Solutions in the Datacenter and at the Heavy Edge

June 2025

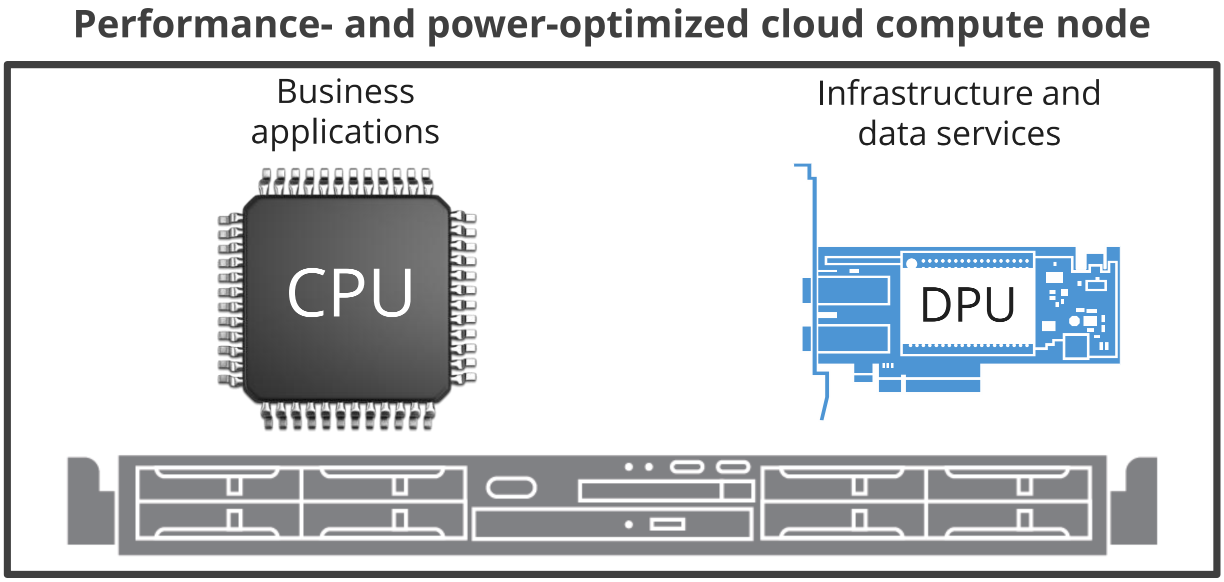
Written by: Brandon Hoff, Executive Analyst, Accelerate Compute and Networking

Introduction

The datacenter is embracing new semiconductor technologies to build an accelerated compute infrastructure that consists of CPUs, graphics processing units (GPUs), and data processing units (DPUs). A DPU is a programmable network semiconductor with compute capabilities (see Figure 1). It accelerates networking-intensive functions and microservices including networking (packet processing, load balancing), storage (NVMe over Fabrics [NVMe-oF]), security (transport layer security and IPsec encryption), virtualization (vSwitch offload, VM isolation, infrastructure management), and AI networking for accelerated compute.

FIGURE 1: ***DPU Architecture***

DPUs provide accelerated network functions and microservices in servers, networking, or storage systems and are used with CPUs.


Performance- and power-optimized cloud compute node
Business applications
CPU
Infrastructure and data services
DPU

Source: IDC, 2025

Networking in the datacenter and at the heavy edge is changing. Standard network interface cards (NICs) are not enough for new datacenter workloads and AI factories. DPUs provide improved performance and lower power consumption for accelerated compute. In many applications, DPUs get the network out of the way of datacenter workloads and large language models (LLMs).

Datacenters are quickly becoming revenue-generating assets, and semiconductors are a foundational ingredient to accelerate innovation. DPUs are a key piece of the datacenter.

Definitions

- **NICs or foundational NICs** support bandwidths from 100Mbps to 800Gbps. Advanced networking workloads are supported but limited to OS offloads (OS assist), such as RSS, SR-IOV, and overlay networks.
- **DPUs** can run either as the main processor for a network or storage device or as an offload for a system processor and OS. A DPU can also be a communications processor. DPUs can come in many different configurations, such as:
    - **Infrastructure processing units (IPUs)** are also referred to as DPUs. They are used to offload networking, security, storage connectivity, system monitoring, and hypervisor functionality. IPUs separate the compute environment from the control environment. IPUs/DPUs can also be referred to as SmartNICs and provide connectivity to meet the demand for data for AI processors and GPUs.
    - **Storage acceleration DPUs** are used in storage systems for deduplication, compression, erasure codes, NVMe connectivity, and NVMe over Fabrics. DPUs can be used to build Ethernet-connected just a bunch of disks (JBODs) without the need for a CPU and are used in the datacenters of many hyperscalers or cloud service providers.
    - **Network controller DPUs** are used for primary processors in dedicated networking products such as switches and routers.

Benefits

DPUs are a critical part of the datacenter and heavy edge infrastructure market and provide benefits in three key areas: AI networking, hypervisor offload, and system on a chip (SoC) for storage and networking devices.

Datacenter SmartNICs for AI Networking

DPUs and datacenter SmartNICs are required for high-performance networks addressing scale-out networks for accelerated compute and high-performance computing (HPC). According to IDC's research, AI networking is the second-largest semiconductor investment for AI factories behind accelerated compute, meaning GPUs and application-specific integrated circuits (ASICs). LLMs require large numbers of GPUs and ASICs to work in parallel and then share their data between them. The time it takes to send data between GPUs and ASICs can take up to 50–60% of the processing time for a LLM, essentially leaving expensive GPU resources idle. In-server DPUs with fat-tree networks can deliver optimal performance by managing congestion, in-order packet delivery, and elephant flows, to name a few key benefits. AI networks require DPUs or SmartNICs to optimize AI factory performance with the following features:

- Speeds at 200Gbps or higher

- Remote direct memory access for accelerated compute used in HPC and AI factories
- Near-zero packet loss
- Hardware-accelerated networking to minimize jitter
- Ability to meet Ultra Ethernet Consortium (UEC) features and functions
- Programmable to support custom networking designs

DPUs to Offload Datacenter Infrastructure Services for Bare Metal Performance

End users are continuously working to harness more compute, reduce system complexity, and reduce power. Datacenter offloads provide the following:

- **Virtualization offload:** As applications get larger, process more data, and can be moved anywhere, the hypervisor infrastructure workloads running on the server start to get in the way of revenue-generating compute workloads. Hypervisor workloads can become noisy neighbors or consume valuable CPU resources, creating bottlenecks. A key shift is that workloads and microservices that historically ran on x86 cores are now moving off them and onto dedicated hardware to improve performance, lower power demand, and accelerate innovation. This shift in compute for virtualization is very similar to the one that GPUs brought to the datacenter.
- **Security offload:** DPUs accelerate packet filtering and forwarding by offloading traffic from the host processors to dedicated hardware to provide intrusion prevention, advanced firewall security capabilities, distributed denial-of-service defense, and intelligent network filtering while improving overall performance.
- **Storage systems:** DPUs can provide advanced services, such as NVMe-oF target offloads for NVMe/TCP and NVMe/RoCE protocols. This leaves the storage system CPU for other services such as deduplication, compression, snapshots, and failover.
- **Inline field-programmable gate array (FPGA) support and integration:** DPUs also work with additional hardware offloads that FPGAs can provide for higher-performance encryption, compression, and additional networking functions at the very high data rates required in modern datacenters.

DPUs Can Operate as an SoC for Optimized and Cost-Reduced Appliances

DPUs operating as a system on a chip is a use case where DPUs provide lightweight compute resources and don't necessarily need a separate CPU to reduce costs. DPUs provide networking, infrastructure management, and connectivity for two key markets:

- **Storage devices:** Software-defined storage systems separate the software from the hardware to reduce storage costs, optimize fault domains, and build a robust ecosystem of simplified storage devices. Ethernet-connected JBODs provide flexibility in scalable, low-cost nodes for high-end storage, such as NVMe or storage class memory, volume storage using NVMe/SDDs and HDDs, and capacity storage (primarily HDDs). Ethernet-connected JBODs do not provide storage services such as tiering, failover, snapshots, or other advanced services since these are moved out of the storage system. The benefit is that the cost and complexity of the storage system are reduced, expensive

CPUs are not required, and the DPU provides the storage connectivity and compute required for these products. This approach is proven in large-scale hyperscale infrastructure deployments today.

- **Networking devices:** AI networking is a new datacenter network architecture that delivers the connectivity required for AI factories processing LLMs. Another proven approach is pulling the DPU out of the server and replacing it with a low-cost NIC, building the DPU into the switch. This offloads the AI networking services provided by the DPU from the accelerated server to the switch, improving system performance with frictionless insertion. This solution is also programmable and can support the new UEC standard.

Considerations

DPUs provide performance improvements for a diverse set of use cases, from accelerated compute and hybrid cloud to networking and storage. It is essential that architects and designers understand and leverage the benefits of DPUs from the datacenter to the heavy edge. In detail:

- **For datacenter architects:** The enterprise has been on a journey to cloudify operations for more than a decade. This has resulted in updates and changes to software architecture, such as deploying rack-optimized servers, embracing DevOps, and deploying CI/CD solutions. The next step is for the enterprise to upgrade its strategy to embrace cloud-optimized server semiconductor architectures that offload microservices to DPUs.
- **For accelerated compute architects:** Datacenter infrastructure is quickly shifting to token-generating factories that in turn, drive revenue. Lowering the cost and power requirements per token is a key performance indicator, and improving network performance and reducing network power consumption increase datacenter performance. In essence, each GPU requires a high-speed, network-optimized DPU for optimal datacenter performance.
- **For network system designers that build switches, routers, and security devices:** DPUs have proven to provide accelerated network services for high-performance network applications and network functions that don't need a CPU for lower-cost products.
- **For storage system designers that build storage systems:** DPUs are proven to provide target functionality and internode communications for high-performance, high-availability storage systems. DPUs can also be used in Ethernet JBOD solutions to reduce cost by removing expensive CPUs from the devices for software-defined storage solutions.

Conclusion

As datacenter infrastructure evolves to support AI, cloud-native applications, and distributed workloads, DPUs are emerging as a critical architectural component. First, DPUs offload infrastructure tasks — such as networking, storage, security, and infrastructure management — from the host CPU, enabling more efficient workload execution and reducing latency. Offloads not only improve CPU utilization but also enhance performance predictability for tenant applications, which is vital in multitenant and hyperscale environments. IDC views DPUs as central to the shift toward

IDC expects DPUs to be foundational to next-generation datacenter and heavy edge infrastructure.

composable and disaggregated infrastructure, where control and data planes are increasingly decoupled to optimize compute and network resource utilization.

Second, AI factories require high-performance networks to reduce GPU idle time and improve datacenter efficiency. DPUs enable hyperscalers, cloud service providers, HPC, and enterprise datacenters to move to Ethernet with the features required for GPU-GPU communication. DPUs are an essential part of the equation for how to build an AI factory.

Beyond operational efficiency, DPUs deliver performance improvements in security and workload isolation. By handling sensitive tasks such as encryption, firewall enforcement, and zero trust policy management in a dedicated hardware domain, DPUs help reduce the attack surface within modern datacenters. This is especially relevant for industries with stringent compliance requirements or workloads that span hybrid cloud and multicloud environments. DPUs also provide higher-performance solutions for storage devices by offering high-performance storage target features and internode communication for high-end storage arrays to low-cost Ethernet JBODs.

As AI clusters scale, datacenter workloads move into the hybrid cloud, and edge computing demands real-time performance with minimal overhead, IDC expects DPUs to become foundational to next-generation infrastructure, enabling scalable, secure, and agile datacenter architectures.

About the Analyst

Brandon Hoff, *Executive Analyst, Accelerate Compute and Networking*

Brandon Hoff leads IDC's Accelerated Compute and Networking infrastructure within IDC's Enabling Technologies team. Mr. Hoff covers technology trends, workloads, products, vendors, supply chain, and end-user adoption strategies in enterprise IT and datacenters of hyperscalers, cloud service providers (SPs), enterprise, HPCs, and telecommunications SPs.

MESSAGE FROM THE SPONSOR

Chelsio is a leading provider of high-performance Ethernet adapters and ASICs with advanced offload capabilities for storage, networking, and security. The company's T6 and upcoming T7 SmartNICs are designed for modern workloads including NVMe/TCP, iSCSI, RDMA, and TLS/IPsec acceleration. These JBOF-class adapters integrate protocol offloads and traffic management in hardware to deliver low-latency, high-throughput performance with reduced CPU utilization. Chelsio solutions are deployed globally in enterprise, cloud, and telco environments.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
blogs.idc.com
www.idc.com

IDC Custom Solutions produced this publication. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis that IDC independently conducted and published, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. This IDC material is licensed for external use, and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives.