# AMD PENSANDO™ SALINA DPU

AMD
PENSANDO

## OVERVIEW

The AMD Pensando™ "Salina" DPU represents a revolutionary advancement in cloud data center architecture as the third generation of AMD's Pensando DPUs. This innovative solution offers a comprehensive suite of software-defined services at the compute edge, delivering significant performance improvements over its predecessors. The Salina DPU  is uniquely positioned to meet the escalating demands of AI workloads, providing essential SDN, security, storage offloads, and telemetry services for AI front-end networks and enterprise cloud with its cutting-edge features, the AMD Pensando "Salina" DPU transforms cloud data center architecture, establishing itself as a vital component of next-generation AI-driven infrastructure for AI factories, public and enterprise clouds and revolutionizing how data centers operate and process information.

**AMD PENSANDO™ SALINA DPU**
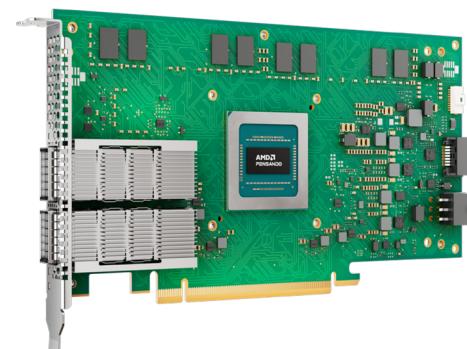


## KEY FEATURES

| SPECIFICATIONS | |
|---|---|
| **FORM FACTOR** | • Full-height Half (8") length PCIe |
| **HOST INTERFACE** | • Single Host: 16-lane PCIe Gen5.0, Multi-Host up to two |
| **NETWORK PORTS** | • 2 ports QSFP112 (NRZ/PAM4) |
| **PORT CONFIGURATIONS** | • 2 x 400 GE<br>• 4 x 200 GE<br>• 4 x 100 GE<br>• 4 x 50 GE<br>• 4 x 25 GE<br>• NRZ (10/25G)<br>• PAM4 (50/100/200/400G)<br>• Quad 200/100/50/25 GE |
| **RoCE SUPPORT** | • RoCEv1 and RoCEv2  support on L2/L3 networks |
| **MANAGEMENT** | • MCTP over SMBus |
| **MEMORY** | • 64 GB DDR5 6400 MT/s memory with ECC |

AMD PENSANDO

## ENHANCED OBSERVABILITY

- Flow-based packet telemetry
- Stateful connection statistics
- Latency metrics, drop statistics
- Session threshold alerting
- ERSPAN (bi-directional)
- NetFlow/IPFIX

## ADVANCED SECURITY

- Stateful firewall
- Connection Tracking (L4)
- Security groups, Stateless and reflexive ACLs
- VPN termination (IPsec)
- NAT, PAT
- TLS/DTLS encryption, TLS Proxy

## ENCRYPTION

- Bulk Cryptography
- Public Key Encryption (PKE)
- Authentication
- Compression and Decompression (LZRW1-A)
- Inline IPsec & DTLS
- Secure hashes and CRC blocks are available to chain in a programmable fashion

## CLOUD NETWORKING

- Routing (BGP), ECMP, VPN Overlays
- Flexible Encapsulation (VxLAN, NVGRE, Geneve, IP-in-IP, GRE)
- Load balancing
- Multi-Tenancy

## STORAGE ACCELERATION

- RDMA, RoCEv2, UEC Ready RDMA
- NVMe virtualization
- NVMe-oF with RDMA or TCP transport
- Data-at-rest encryption (AES-CTS)
- Compression
- Deduplication: SHA2 128/256/512, CRC32, Azure CRC64, Adler32, and M-Adler32
- Checksum Acceleration (CRC64/32)

# AGILE PLATFORM FOR CLOUD PROVIDERS

The DPU is a cutting-edge software-defined platform designed to optimize cloud and AI front-end infrastructure performance. It enables developers to create customized data plane services that achieve 400G throughput with microsecond-level latencies, effectively scaling to millions of flows. The platform supports high-level programming languages such as P4, C++, and C, facilitating rapid development and deployment of new features and services.

## USE CASES

- Multi-tenant SDN, virtual networking with flexible encapsulations (VXLAN, NVGRE, Geneve etc.)
- Routing, segment routing, MPLS, BGP/eBGP
- Comprehensive observability and troubleshooting of network, storage, and security functions
- Flow capture/mirroring (bidirectional ERSPAN, per-packet IPFIX)
- Network load balancing, including TCP/TLS termination
- Storage virtualization and offload (NVMe-oF over TCP or RoCE)
- Data compression/decompression

## ENABLING CLOUD INFRASTRUCTURE

- Virtual private cloud (VPC)
- Security groups, firewall, DDoS protection
- Transit gateway and VPC peering
- NAT gateway
- Metering, rate control and QoS marking
- SR-IOV for workload traffic/resource isolation
- IPsec/VPN gateway

# DEPLOYMENT CONFIGURATION

The DPU offers versatile deployment configurations to meet diverse operational needs:

**SERVER EDGE**

In the server edge configuration, DPUs are installed in every server within the data center, running network, security, and storage services related to the workloads on the hosting server. This approach offers horizontal scalability, as additional servers with DPUs can be added to the infrastructure. DPUs can operate in two modes:

• **Host-Mode:** Delivers networking, Security and storage traffic to/from the host, implementing configured services for that traffic.
• **Bump-in-the-Wire Mode:** Connected between the switch and a regular NIC, forwarding traffic from one port to another while implementing required networking services and traffic policies with security requirements.

**DPU HOST-MODE DEPLOYMENT**

In host mode, DPUs are physically installed in servers and equipped with necessary drivers. This configuration allows bidirectional data traffic flow between the host (bare metal or virtual machines) and the DPU. Management can be conducted over the network in-band directly from the cloud provider management systems via the AMD Pensando platform's gRPC-based API or custom management APIs.

### DATA PROCESSING UNITS DEPLOYED IN STANDARD SERVERS
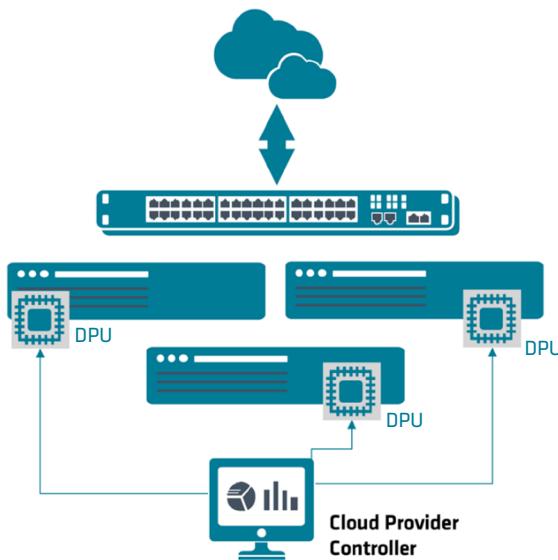


### DATA PROCESSING UNITS DEPLOYED IN AI/GPU SERVERS

 (AMD PENSANDO logo, top right)
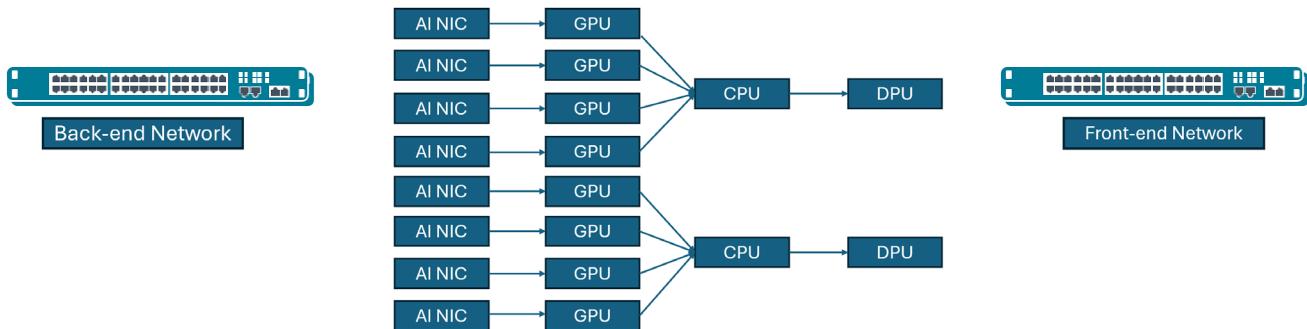
## DPU BUMP IN THE WIRE MODE DEPLOYMENT

This deployment mode positions DPUs as inline network devices, ideal for bare-metal solutions. The PCIe interface is solely used for power, eliminating the need for host drivers or agents. Management is conducted over the network, ensuring streamlined operations.

**IN-LINE DISTRIBUTED SERVICES ("BUMP IN THE WIRE")**
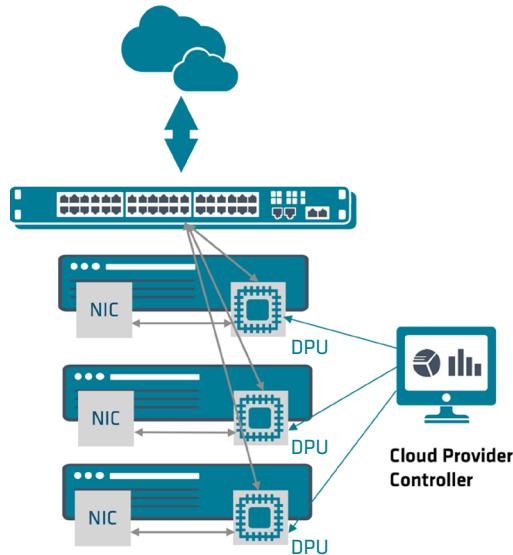


## DPUS IN SDN APPLIANCE DEPLOYMENT

An alternative approach involves using multiple DPUs in a dedicated appliance, providing services to a group of servers. Traffic is redirected to the appliance between the servers and the network, offering a cost-effective solution for environments where bandwidth and latency requirements are less stringent or only a portion of the workloads require security enforcement, encryption, or other services. This approach is horizontally scalable by adding more DPU-based appliances as workload processing needs to grow. A pair of DPUs installed within the same appliance or different appliances can provide high availability for the redirected workloads.

**MULTIPLE DPUS DEPLOYED IN AN APPLIANCE**
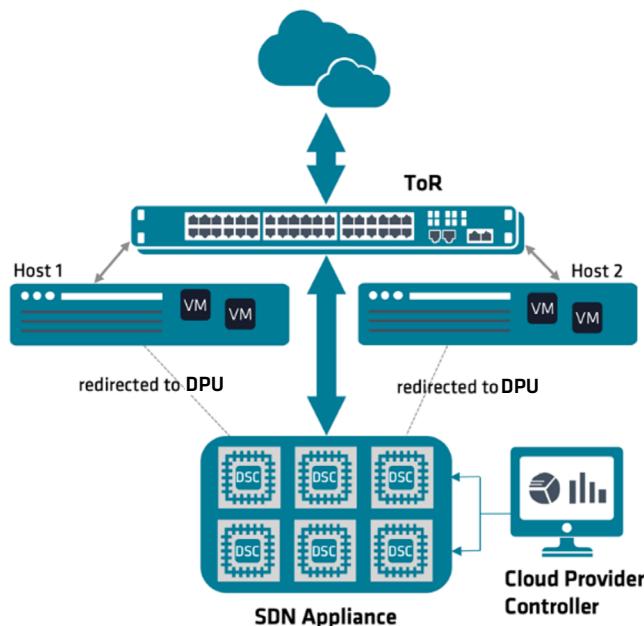
## DPU NVME DEPLOYMENT

DPUs˜ may be deployed on compute servers, emulating NVMe virtual functions (VFs) with other network and security services. The host operating system can then consume these in bare-metal workloads or by guest operating systems in pass-through mode in VM workloads. The NVMe VFs will be managed by standard inbox NVMe drivers. The NVMe driver issues NVMe commands over PCIe, which are terminated fully in the DPUs and originated as NVMe/TCP after encryption (AES-XTS: 128- or 256-bit keys with Header/Data Digest), NVMe/PCIe to NVMe-oF PDU conversion and any applicable load balancing across NVMe-oF target servers, achieving up to 2+M IOPS

**DPUS DEPLOYED TO EMULATE NVME VFS**
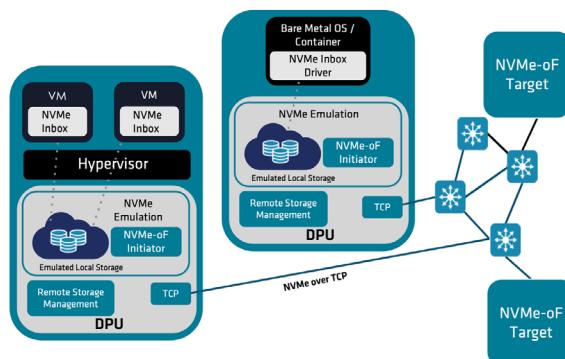


| ADAPTER CARD PORTFOLIO AND FORM FACTORS | |
|---|---|
| ORDERABLE PART NUMBER (OPN) | • DPUs-2Q400-64S64E64P |
| HOST INTERFACE (PCIE) | • PCIe Gen5.0 x16 |
| FORM FACTOR | • Full Height Half (8") length |
| BANDWIDTH | • 400Gb/s |
| NETWORK PORTS | • 2 x QSFP112 |

All adapters are shipped with the tall bracket mounted and a short bracket as an accessory

## ENDNOTES

1. Image shown is for illustrative purposes only; actual product may vary. Card shown with heatsink removed.

## DISCLAIMERS

## COPYRIGHT NOTICE