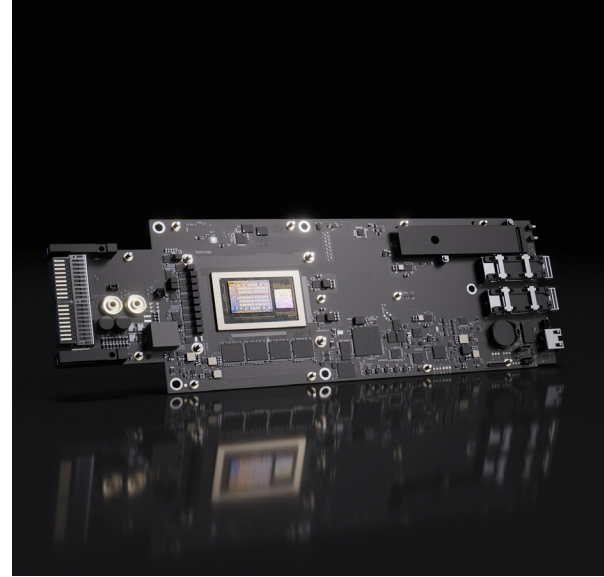




NVIDIA BlueField-4 DPU

The 800 Gb/s DPU powering agentic AI factories.



The NVIDIA® BlueField®-4 data processing unit (DPU) is an accelerated infrastructure platform purpose-built for gigascale AI factories, delivering up to 800 gigabits per second (Gb/s) of throughput. As an integral component of the NVIDIA Vera Rubin platform, the BlueField-4 DPU offloads, accelerates, and isolates infrastructure services while providing high-performance connectivity for AI infrastructure. Combining the NVIDIA Grace™ CPU for high-performance computing with NVIDIA® ConnectX®-9 networking, BlueField-4 unlocks new levels of acceleration, efficiency, and cybersecurity for agentic AI factories.

The BlueField-4 DPU integrates advanced data processing, native service function chaining, built-in security acceleration, and massive I/O performance. Cloud providers, AI-native providers, telecom operators, and enterprises can now scale secure, multi-tenant AI infrastructure and data platforms, while maximizing workload and power efficiency, and supporting large-scale inference.

Accelerate AI Infrastructure

The BlueField-4 DPU is an open, programmable platform powered by NVIDIA DOCA™, designed to accelerate, protect, and scale modern AI infrastructure. It natively runs DOCA microservices, delivering prebuilt, containerized applications for AI networking, runtime security, data storage acceleration, and orchestration with higher efficiency and throughput. BlueField-4 works seamlessly with NVIDIA Spectrum-X™ Ethernet for secure, accelerated connectivity in AI factories.

The BlueField-4 platform introduces the Advanced Secure Trusted Resource Architecture (ASTRA), designed to extend security and manageability in select NVIDIA Rubin platforms with NVIDIA ConnectX-9 SuperNIC™ cards. ASTRA enables secure bare-metal compute instances with zero-trust tenant isolation and full software-defined infrastructure control.

The BlueField architecture spans the full NVIDIA AI infrastructure portfolio, from computing systems to scale-out data center designs. It serves as the infrastructure control plane in AI factories and cloud environments, isolating infrastructure management from tenant workloads. Enterprises can further extend platform capabilities by leveraging a broad ecosystem of accelerated infrastructure and cybersecurity applications from leading software providers.

Product Specifications

Supported Network Protocols

- > Ethernet
- > InfiniBand

Maximum Bandwidth

- > 800 Gb/s

Network Speeds

- > 800G per port, 200G SerDes

Compute

- > Grace CPU with 64x Arm Neoverse V2 cores

Memory

- > Up to 128 GB LPDDR5x

Host Interface

- > PCIe Gen6 x16

Portfolio

- > PCIe form factor
- > Custom form factors for NVIDIA programs

Powering Agentic AI Storage

While the BlueField-4 DPU serves as the infrastructure processor for AI factories, the NVIDIA BlueField-4 STX storage processor extends the BlueField-4 platform to AI-native storage. It powers data platforms with accelerated data processing, high-performance NVMe storage, and in-silicon security.

Integrating the NVIDIA Vera CPU and NVIDIA ConnectX-9 SuperNIC, BlueField-4 STX delivers high-throughput data movement, efficient storage access, and secure data services for demanding AI storage workloads. Featured within the NVIDIA STX reference architecture, the BlueField-4 STX storage processor supports emerging AI-native storage workloads, including NVIDIA CMX context memory storage, AI data platforms, and storage infrastructure for AI factories. See the BlueField-4 STX storage processor [datasheet](#) for more information.

Key Features*

Network Interface	> Bandwidth: 800 Gb/s > 200/100/50 Gb/s PAM4 and 25 Gb/s NRZ > Up to 800 Gb/s per port > Support up to 8 split ports
PCIe Interface	> PCIe Gen 6 > NVIDIA Socket Direct™
Compute and Memory	> Grace CPU with 64 Neoverse Arm V2 > Up to 128 GB on-board LPDDR5x > Up to 512 GB on-board SSD > Programmable data path accelerator: • 16 cores, 256 threads • Heavy multi-threading application acceleration • Generic device emulation • DOCA-programmable
Storage Initiator	> BlueField SNAP file system/block storage emulation > VirtIO and NVMe-emulated PCIe devices > NVMe-oF, S3 over RDMA, and NVMe/TCP client acceleration > T10-Diff > AES-XTS 256/512-bit key data-at-rest acceleration > Real-time data access control > GPUDirect® Storage

Key Features*

Cybersecurity	> Real-time AI workload threat detection > Isolated trust domains > Stateful firewall connection tracking > AES-GCM 128/256-bit key data-in-motion acceleration: IPsec, TLS, PSP > Platform security: • Secure boot with hardware root of trust • Secure firmware updates • Flash encryption • Device attestation (SPDM 1.1/1.2)
AI Networking	> RDMA and RoCEv2 > NVIDIA Spectrum-X Ethernet > Programmable RDMA transport > Programmable congestion control > NVIDIA Inference Transfer Library (NIXL) > NVIDIA GPUDirect RDMA > In-network computing
Cloud Networking	> DOCA-Accelerated Programmable Steering and Packet Processing > Flexible overlay encapsulation (VxLAN, NVGRE, Geneve) > Programmable flexible parser: user-defined classification > Stateless TCP offloads
Management and Control	> Integrated DPU BMC > 1GbE out-of-band management port > Redfish-based management (HTTPS/TLS) > Advanced Secure Trusted Resource Architecture (ASTRA) > NVIDIA System Management (NSM) over MCTP > PLDM over MCTP over USB and PCIe for FW update (DSP0267), monitor and control (DSP0248), Redfish device enablement (DSP0218), and FRU (DSP0257)
Network Boot	> InfiniBand or Ethernet > Preboot eXecution Environment (PXE) boot > Unified Extensible Firmware Interface (UEFI)

* This section describes hardware features and capabilities. For feature availability, refer to the [NVIDIA DOCA release notes](#).

Ready to Get Started?

To learn more about the NVIDIA BlueField platform, visit:
nvidia.com/dpu or contact an NVIDIA sales representative:
nvidia.com/networking-contact-sales

© 2026 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, BlueField, ConnectX, DOCA, GPUDirect, Grace, Socket Direct, Spectrum-X, and SuperNIC are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated. 5193854. JUN26

