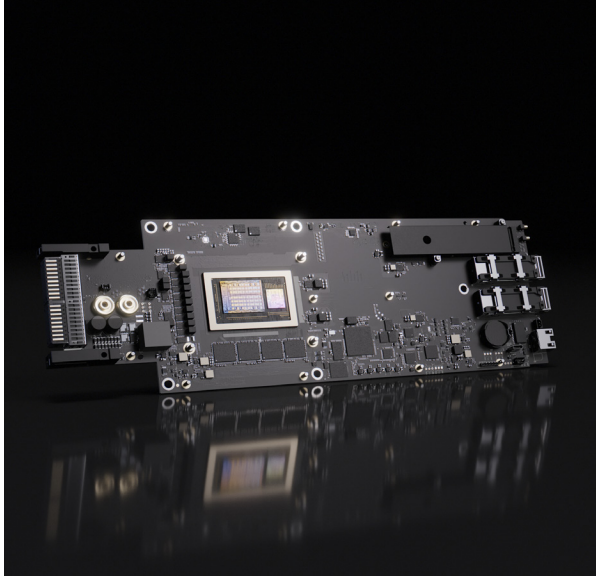
Datasheet

NVIDIA BlueField-4

The 800 Gb/s DPU for powering gigascale AI factories.



The NVIDIA® BlueField®-4 data processing unit (DPU) is an accelerated infrastructure platform purpose-built for gigascale AI factories, delivering throughput up to 800 gigabits per second (Gb/s). By combining the NVIDIA Grace™ CPU for high-performance computing with NVIDIA® ConnectX®-9 networking, BlueField-4 unlocks new levels of acceleration, efficiency, and cybersecurity, powering every layer of AI data center infrastructure.

The platform integrates powerful data processing, native service function chaining, built-in security acceleration, and massive I/O performance. BlueField drives efficiency across the data center as a data processing unit in AI compute nodes and as a high-performance storage controller that enables accelerated, secure, and software-defined access to data. AI-native organizations, service providers, and enterprises can now build multi-tenant AI infrastructure and data platforms with greater scale, achieving peak workload and power efficiency while supporting high-performance storage, rapid data access, and large-scale inference.

Accelerate AI Infrastructure

BlueField-4 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, orchestration, real-time threat detection, and data storage acceleration with higher efficiency and throughput.

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, separating infrastructure management from tenant workloads. Enterprises can further extend functionality by leveraging a broad range of accelerated infrastructure and cybersecurity applications from leading software providers.

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

Product Specifications

Supported Network Protocols

- Ethernet
- InfiniBand

Maximum Bandwidth

- 800 Gb/s

Network Speeds

- 800G per port, 200G SerDes

Compute

- 64x Arm Neoverse V2 cores

Memory

- 128GB LPDDR5

Host Interface

- PCIe Gen6 x16

Portfolio

- Server-hosted DPU:
  - PCIe form factor
  - Vera Rubin NVL72 form factor
- Self-hosted storage controller:
  - PCIe form factor

Powering Efficient AI Storage Platforms

NVIDIA BlueField-4 is an advanced processor to power software-defined, accelerated data storage platforms for AI factories. By providing storage acceleration for data movement, data protection, and security services, BlueField-4 enables AI storage builders to deliver low-latency, high-throughput access for inference, training, and agentic workloads.

With unmatched performance, storage processing, and built-in cybersecurity, BlueField-4 redefines infrastructure acceleration, empowering organizations to build and operate AI factories with efficiency, agility, and trust.

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	> Server-hosted DPU: • PCIe Gen 6 • NVIDIA SocketDirect > Self-hosted storage controller: • PCIe 5.0 x32
Compute and memory	> 64 Neoverse Arm V2 > 114 MB shared L3 cache > 128 GB on-board LPDDR5x > 512 GB on-board pluggable SSD > Programmable data path accelerator: • 16 cores, 256 threads • Heavy multi-threading applications acceleration • Generic device emulation • DOCA programmable
Storage	> BlueField SNAP file system/block storage emulation > VirtIO and NVMe emulated PCIe devices > NVMe-oF, S3 over RDMA, NVMe/TCP client and target controller accelerations > T10-Diff > AES-XTS 256/512-key data-at-rest acceleration > Real-time data inspection and threat detection > GPUDirect® Storage

Key Features	
Cybersecurity	> Real-time AI workload threat detection > Isolated trust domain > Connection tracking for stateful firewall > AES-GCM 128/256-key data-in-motion accelerations: IPsec, TLS, PSP > Platform security: • Secure boot with hardware root of trust • Secure firmware update • Flash encryption • Device attestation (SPDM 1.1)
AI networking	> RDMA and RoCEv2 > NVIDIA Spectrum-X™ Ethernet > Programmable RDMA transport > Advanced, programmable congestion control > NVIDIA Inference Transfer Library (NIXL) > NVIDIA GPUDirect RDMA > In-network computing
Cloud networking	> ASAP² Accelerated Switching & Packet Processing™ for SDN and VNF > Flexible overlay encapsulation (VxLAN, NVGRE, Geneve) > Programmable flexible parser: user-defined classification > Stateless TCP offloads
Management and control	> Advanced Secure Trusted Resource Architecture (ASTRA) > Integrated BMC > 1GbE out-of-band management port > Management capabilities using Redfish protocol > MCTP over SMBus and PCIe PLDM for:: • Monitor and control DSP0248 • Firmware update DSP0267 • Redfish device enablement DSP0218 • Field-replaceable unit (FRU) DSP0257 • Security Protocols and Data Models (SPDM) DSP0274
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

4686050. JAN26