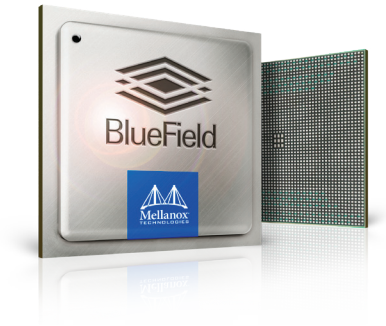
BlueField®

I/O Processing Unit (IPU)

Mellanox BlueField is a family of advanced I/O Processing Unit (IPU) solutions. Whether as a smart network adapter (SmartNIC) or as part of an embedded system, BlueField addresses diverse applications, including NVMe storage, security, networking, and machine learning.

BlueField integrates the Mellanox ConnectX®-5 network controller, 64-bit Arm cores, and PCIe switch into a single device, leveraging the broad Ethernet, InfiniBand and Arm ecosystems. As a co-processor, BlueField can offload the main CPU to overcome performance bottlenecks, controlling part of the resources on the platform or performing tasks not accessible by the main CPU. BlueField can also be leveraged as the sole processor in the platform.



BlueField combines up to 16 Armv8 A72 cores interconnected by a coherent mesh, DDR4 memory controllers, and a multi-port, RDMA-enabled, intelligent, Ethernet/InfiniBand adapter. The IPU supports 10/25/40/50/56/100Gb/s, an integrated PCIe switch with endpoint and root complex functionality, and up to 32 lanes of PCIe Gen 3.0/4.0.

At the heart of BlueField is the ConnectX-5 network offload adapter with RDMA and RDMA over Converged Ethernet (RoCE) technology, delivering cutting-edge performance for networking and storage applications such as NVMe over Fabrics. Advanced features include an embedded virtual switch with programmable ACL, transport offloads and stateless encaps/decaps of NVGRE, VXLAN, and MPLS overlay protocols.

The powerful Armv8 multicore processor array enables sophisticated applications and highly differentiated feature sets. By leveraging the vast ARM ecosystem, and software that is easily portable to and from x86, BlueField supports a wide range of markets, including Storage, Machine Learning, Networking and Security.

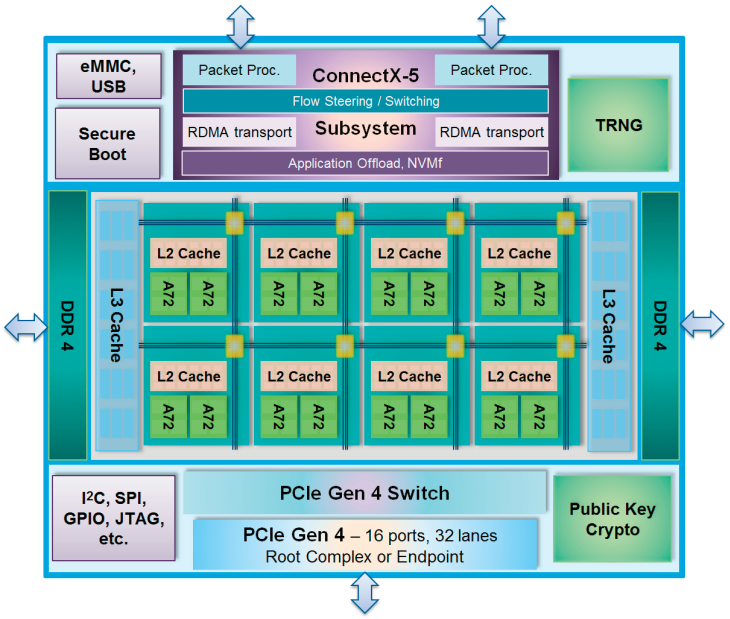


Figure 1. BlueField Architecture

BlueField for Storage Environments

Today's fast storage technologies drive storage OEMs to seek innovative, scalable and cost effective designs for their applications. BlueField's unique set of capabilities offers the ideal solution for storage platforms, such as NVMe over Fabrics (NVMe-oF) All-Flash Array (AFA) and storage controller for JBOF, server caching (memcached), disaggregated rack storage, scale-out direct-attached storage, and storage RAID.

Complete Storage Solution

BlueField brings the processing power of Arm cores for storage applications such as All-Flash Arrays using NVMe-oF, Ceph, Lustre, iSCSI/TCP offload, Flash Translation Layer, RAID/Erasure coding, data compression/decompression, and deduplication.

In high-performance storage arrays, BlueField functions as the system's main CPU, handling storage controller tasks and traffic termination. In other configurations, it may operate as a co-processor, offloading specific storage tasks from the host, isolating part of the storage media from the host, or enabling abstraction of software-defined storage logic using the BlueField Arm cores.

NVMe over Fabrics Capabilities

BlueField improves performance and lowers latency for storage applications thanks to advanced NVMe-oF capabilities. BlueField RDMA-based technology delivers remote storage access performance equal to that of local storage, with minimal CPU overhead, enabling efficient disaggregated storage and hyper-converged solutions.

Storage Acceleration

The BlueField embedded PCIe switch enables customers to build standalone storage appliances and connect a single BlueField to multiple storage devices without an external switch.

Distributed RAID

BlueField data path hardware supports advanced Erasure Coding offloading, enabling distributed Redundant Array of Inexpensive Disks (RAID). The Reed-Solomon engine introduces redundant block calculations, which, with RDMA, achieve superior performance and high availability.

Signature Handover

The BlueField embedded network controller enables hardware checking of T10 Data Integrity Field/Protection Information (T10-DIF/PI), reducing software overhead and accelerating delivery of data to the application. Signature handover is handled by the adapter on ingress and egress packets, reducing the load on the software at the Initiator and Target machines.

BlueField for Networking & Security

BlueField enables efficient deployment of networking applications, both in the form of a smartNIC and as a standalone network platform. Using a combination of advanced offloads and Arm compute capabilities, BlueField terminates network and security protocols inline.

BlueField SmartNIC

As a network adapter, BlueField offers the flexibility to fully or partially implement the data and control planes. As such, BlueField unlocks more efficient use of compute resources, now dedicated to run applications, as well as support for the growing demand for bandwidth per host. The programmability of the adapter ensures the ability to easily integrate new data and control plane functionality.

BlueField Security Features

The versatility offered by the BlueField adapter, including the integration of encryption offloads for symmetric and asymmetric crypto operations, makes it ideal to implement security applications. BlueField builds security into the DNA of the data center infrastructure, reducing threat exposure, minimizing risk, and enabling prevention, detection and response to potential threats in real-time.

Painless Virtualization

BlueField PCIe SR-IOV technology enables data center administrators to benefit from better server utilization while reducing cost, power, and cable complexity, allowing more Virtual Machines and more tenants on the same hardware.

BlueField also provides dedicated adapter resources and guaranteed isolation and protection for virtual machines (VMs) within the server.

Mellanox ASAP^2 - Accelerated Switching and Packet Processing® accelerates Open vSwitch (OVS) to deliver flexible, highly efficient virtual switching and routing capabilities.

Using a pipeline-based programmable embedded switch (eSwitch), as well as hairpin hardware capability, data can be handled by the Virtual Appliance with minimal server CPU intervention.

Overlay Networks

Data center operators use network overlay technologies (VXLAN, NVGRE, GENEVE) to overcome scalability barriers. By providing advanced offloading engines that encapsulate/de-capsulate the overlay protocol headers, BlueField allows the traditional offloads to operate on the tunneled protocols and also offloads NAT routing capabilities. With BlueField, data center operators can achieve native performance in the new network architecture.

BlueField for Machine Learning Environments

The BlueField IPU provides cost effective and integrative solutions for Machine Learning appliances. BlueField enables multiple GPUs to be connected via its PCIe Gen 3.0/4.0 interface. With its superior RDMA and GPUDirect® RDMA technologies, BlueField offers the most efficient data delivery for real-time analytics and data insights.

RDMA Acceleration

BlueField network controller data path hardware utilizes RDMA and RoCE technology, delivering low latency and high throughput with near-zero CPU cycles.

BlueField for Multi-GPU Platforms

BlueField enables the attachment of multiple GPUs through its integrated PCIe switch. BlueField PCIe 4.0 support is future-proofed for next-generation GPU devices.

Mellanox PeerDirect®

Mellanox PeerDirect is an accelerated communication architecture that supports peer-to-peer communication between BlueField and third-party hardware such as GPUs (e.g., NVIDIA GPUDirect RDMA), co-processor adapters (e.g., Intel Xeon Phi), or storage adapters. Mellanox PeerDirect provides a standardized architecture in which devices can directly communicate to other remote devices across the Mellanox fabric, avoiding unnecessary system memory copies and CPU overhead by copying data directly to/from devices.

GPUDirect RDMA Technology

The rapid increase in the performance of graphics hardware, coupled with recent improvements in GPU programmability, has made graphic accelerators a compelling platform for computationally demanding tasks in a wide variety of application domains. Since GPUs provide high core count and floating point operations capabilities, high-speed networking is required to connect between the platforms to provide high throughput and the lowest latency for GPU-to-GPU communications. GPUDirect RDMA is a technology implemented within both BlueField and NVIDIA GPUs that enables a direct path for data exchange between GPUs and the Mellanox high-speed interconnect.

GPUDirect RDMA provides order-of-magnitude improvements for both communication bandwidth and communication latency between GPU devices of different cluster nodes.

Table 1 - Part Numbers and Descriptions

OPN	Description
MT41M16T23A1-NDCR-TTEV	BlueField E-Series IPU, 16 Cores, Dual 100GbE/EDR VPI Ports, PCIe Gen4.0 x32, No Crypto
MT41M16T23A1-CDCR-TTEV	BlueField E-Series IPU, 16 Cores, Dual 100GbE/EDR VPI Ports, PCIe Gen4.0 x32, Crypto
Contact Mellanox	BlueField E-Series IPU, 8 Cores, Dual 100GbE/EDR VPI Ports, PCIe Gen4.0 x32, No Crypto
MT41M16P23A1-NDCR-TTEV	BlueField P-Series IPU, 16 Cores, Dual 100GbE/EDR VPI Ports, PCIe Gen4.0 x32, No Crypto
MT41M16P23A1-CDCR-TTEV	BlueField P-Series IPU, 16 Cores, Dual 100GbE/EDR VPI Ports, PCIe Gen4.0 x32, Crypto
Contact Mellanox	BlueField G-Series IPU, 8 Cores, Dual 100GbE/EDR VPI Ports, PCIe Gen4.0 x32, Crypto/No Crypto

Features

This section describes hardware features and capabilities. Please refer to the driver and firmware release notes for feature availability.

Powerful Processor Cores
- Up to 16 Arm v8 A72 cores (64-bit)
- Superscalar, variable-length, out-of-order pipeline
- Robust virtual memory system with TLBs, multiple page size support, and hardware virtualization
- Each core supports NEON™ 128b SIMD execution unit
- ArmVFPv4 single and double – precision floating point acceleration (IEEE 754)
- Per core 48KB I-cache and 32KB D-cache
- Cache coherent mesh interconnect of CPUs, I/O and memory – each tile contains 2 cores and 1 MB L2 cache
- Two banks of 6 MB L3 cache, sophisticated eviction policies

Encryption Acceleration
- Armv8 cryptography extensions: A64, A32, and T32 instructions for:
  - AES, SHA-1, SHA-224, and SHA-256
  - Finite field arithmetic used in algorithms such as Galois/Counter Mode and Elliptic Curve
- Hardware Public Key accelerator
  - RSA, Diffie-Hellman, DSA, ECC, EC-DSA and EC-DH
- True Random Number Generator with entropy source

Network Interfaces
- ConnectX-5 Virtual Protocol Interconnect® (VPI) adapter
- Ethernet: 10/25/40/50/56/100GbE port rates
- InfiniBand: SDR/DDR/QDR/FDR/EDR port rates
- Automatically identifies and operates on InfiniBand, Ethernet, or Data Center Bridging (DCB) fabrics
- Integrated PHYs seamlessly connect to all standard copper and fiber media

PCI Express Interface
- PCIe Gen 3.0/4.0, Endpoint or Root Complex
- Integrated PCIe switch with up to 16 downstream ports
- SR-IOV
- 2.5, 5.0, 8, 16GT/s link rate
- Auto-negotiates to x32, x16, x8, x4, x2, x1
- Support for MSI/MSI-X
- Configurable, user-programmable QoS guarantee for VMs

DDR4 DIMM Support
- DDR4 SoDIMM, UDIMM, RDIMM, NVDIMM-N and LRDIMM supported (with or without ECC)
- Up to 2 DIMMs per channel
  - AES, SHA-1, SHA-224, and SHA-256
  - Soldered, discrete RAMs
- Up to 90% bus utilization with multiple ranks
- ECC error protection support

Enhanced Features
- Hardware-based reliable transport
- Offloads of collective operations
- Offloads of vector collective operations
- PeerDirect RDMA (aka GPUDirect) communication acceleration
- Extended Reliable Connected transport (XRC)
- Dynamically Connected transport (DCT)
- Enhanced Atomic operations
- Advanced memory mapping support, allowing user mode registration and remapping of memory (UMR)
- On demand paging (ODP)
- Registration-free RDMA memory access

InfiniBand
- IBTA specification 1.3 compliant
- RDMA, Send/Receive semantics
- Hardware-based congestion control
- Atomic operations
- 16 million I/O channels
- 256 to 4Kbyte MTU, 2Gbyte messages
- 8 virtual lanes + VL15

Ethernet
- IEEE 802.3bj, 802.3bm 100 Gigabit Ethernet
- 25G Ethernet Consortium 25, 50 Gigabit Ethernet
- IEEE 802.3ba 40 Gigabit Ethernet
- IEEE 802.3ae 10 Gigabit Ethernet
- IEEE 802.3az Energy Efficient Ethernet
- IEEE 802.3ap based auto-negotiation and KR startup
- Proprietary Ethernet protocols (20/40GBASE-R2, 50/56GBASE-R4)
- IEEE 802.3ad, 802.1AX Link Aggregation
- IEEE 802.1Q, 802.1P VLAN tags and priority
- IEEE 802.1Qau (QCN) – Congestion Notification
- IEEE 802.1Qaz (ETS)
- IEEE 802.1Qbb (PFC)
- IEEE 802.1Qbg
- IEEE 1588v2
- Jumbo frame support (9.6KB)

Transport Offloads
- RDMA over Converged Ethernet (RoCE)
- TCP/UDP/IP stateless offload
- LSO, LRO, checksum offload
- RSS (also on encapsulated packet), TSS, HDS, VLAN insertion / stripping, Receive Flow Steering
- Intelligent interrupt coalescence
- OpenMPI, IBM PE, OSU MPI (MVAPICH/2), Intel MPI
- Platform MPI, UPC, Open SHMEM
- TCP/UDP, MPLS, VxLAN, NVGRE, GENEVE
- SRP, iSER, NFS RDMA, SMB Direct
- uDAPL

Hardware-based I/O Virtualization
- SR-IOV: Up to 512 Virtual Functions
- SR-IOV: Up to 16 Physical Functions per host
- Multi-function per port
- Mellanox Multi-Host® (up to 4 hosts)
- Address translation and protection
- Multiple queues per virtual machine
- Enhanced QoS for vNICs
- VMware NetQueue support
- Virtualization hierarchies
- Virtualizing Physical Functions on a physical port
- 1K ingress and egress QoS levels
- Guaranteed QoS for VMs

Overlay Networks
- Stateless offloads for overlay networks and tunneling protocols
- Hardware offload of encapsulation and decapsulation of NVGRE, VXLAN and GENEVE overlay networks
- Header rewrite supporting hardware offload of NAT

Advanced Boot Options
- Secure Boot (RSA authenticated)
- Remote boot over Ethernet/InfiniBand
- Remote boot over iSCSI
- PXE and UEFI

Management and Control Interfaces
- NC-SI, MCTP over SMBus, and MCTP over PCIe - Baseboard Management Controller interface
- SDN management interface for managing the eSwitch
- I2C interface for device control and configuration
- General Purpose I/O pins
- SPI interface to Flash
- eMMC memory controller
- MDC/MDIO master
- UART
- USB
- IEEE1588 time stamp real-time clock controls: PPS-Out, PPS-In
- LED chain
- Fan controller / Thermal shutdown
- JTAG IEEE 1149.1 and IEEE 1149.6

Software Development Toolchain
- Native and cross-compile GNU toolchain
- Performance analysis and profiling tools
- Compatible with Arm DS-5 and other commercial development and profiling tools

Software Support

Arm Environment
- BlueOS: Commercial grade Yocto-based Arm Linux distribution
- Commercial Linux distributions supported
- Delivered with OpenFabrics Enterprise Distribution (OFED)
- Arm-optimized versions of all Mellanox drivers and software stack
- Accelerated NVMe over Fabrics target stack
- Optimized Arm DPDK and ConnectX PMD

Connected Host (Network Adapter Environment)
- Linux
- Windows
- FreeBSD
- VMware
- OpenFabrics Enterprise Distribution (OFED)
- OpenFabrics Windows Distribution (WinOF-2)

Mellanox TECHNOLOGIES
350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com

Support
For information about Mellanox support packages, please contact your Mellanox Technologies sales representative or visit our Support Index page.

52964PB
Rev 3.1