BXI

High performance interconnect for extreme HPC workloads

Exascale entails an explosion of performance, of the number of nodes/cores, of data volume and data movement. At such a scale, optimizing the network that is the backbone of the system becomes a major contributor to global performance. The interconnect is very likely going to be a key enabling technology for exascale systems. This is why one of the cornerstones of Atos exascale program is the development of our own new-generation interconnect. The Bull eXascale Interconnect or BXI introduces a paradigm shift in terms of performance, scalability, efficiency, reliability and quality of service for extreme workloads.

The BXI fabric is highly scalable (up to 64 000 nodes, it features:

- High-speed links (100 Gb/s)
- High message rate (up to 100 M msg/s)
- Minimal memory footprint and low latency components.

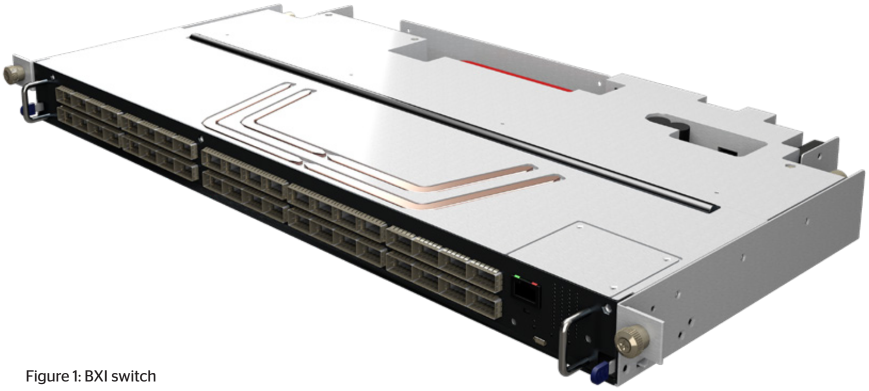


Figure 1: BXI switch

Getting rid of the communications overhead

The core feature of BXI is a full hardware-encoded communication management system, which enables CPUs to be fully dedicated to computational tasks while communications are independently managed by BXI.

As a result, contrary to other commonly used networks, BXI can deliver high communication throughput even when the system is under heavy computation stress.

BXI hardware primitives map directly to communication libraries such as MPI (Message Passing Interface) and PGAS (Partitioned Global Address Space). Thanks to this hardware acceleration, BXI delivers the highest level of communication performance for HPC applications, at full scale, characterized by high bandwidth, low latency and high message rates.

The BXI architecture is based on the Portals 4 communication library. This enables full optimization for all MPI communication types, including the latest MPI-2 and MPI-3 extensions and PGAS. The Portals 4 non-connected protocol guarantees a minimum constant memory footprint, irrespective of system size.

Quality of service

BXI quality of service (QoS) enables the definition of several virtual networks and ensures, for example, that bulky I/O messages do not impede small data message flow. In addition, BXI adaptive routing capabilities dynamically avoid communication bottlenecks.

Reliability and resilience

For high reliability, BXI implements both end-to-end and link-level error checking and retransmission. Furthermore, all ASIC parts feature ECC schemes for error detection and correction. These mechanisms ensure continuity of service in case of a transient or permanent failure (on link or switch).

BXI components

The BXI fabric relies on two types of ASICs as its building blocks, a **Network Interface Controller** (NIC) and a **switch**, and comes with its complete **software suite**.

BXI switches are managed through a distributed and out-of-band fabric management suite allowing to scale up to 64K nodes. Out-of-band management eliminates any interference of the management traffic with the applications traffic.

BXI components are detailed overleaf.

BXI in BullSequana XH2000

BullSequana XH2000 is the new HPC platform that optimizes energy efficiency thanks to its Direct Liquid Cooling technology (DLC). A BullSequana XH2000 cell comprises one rack with up-to 96 compute nodes interconnected with an embedded two-level BXI fat-tree. To form large HPC systems, BullSequana XH2000 cells are interconnected using external BXI switches.

BXI Network Interface Controller (NIC)

The BXI NIC is available as a mezzanine card for BullSequana XH2000 compute blades and as a standard PCIe card for other nodes. The BXI NIC board interfaces each node to the BXI interconnect.

- PCIe gen3 x16 link
  - NIC custom 4x BXI port to deliver 12,5GB/s per direction on the network
- Implements in hardware the Portals 4 communication primitive
  - Overlapping communications and computations by offloading to NIC
  - MPI two-sided building blocks: matching in hardware and V2P
  - PGAS / MPI one-sided messaging
- OS and application bypass
  - Reception controlled by NIC without interrupts or OS involvement
  - Reply to a put or a get does not require activity on application side
- Collective Operations offloads Accelerations in HW
- End-to-End reliability recovery mechanism for transient and permanent failures
- Allocates Virtual Channels: Separating different type of messages to avoid deadlocks and to optimize network resources usage (load balancing and QoS)
- Offers performance and errors counters for Applications performance analysis

Figure 2: BXI standard PCI card

BXI switch

The BXI switch is available as an internal switch module in BullSequana XH2000 with copper connectors and as an external switch with 48 ports.

The BullSequana XH2000 switch modules are cooled with an enhanced version of the Bull patented Direct Liquid Cooling (DLC) solution, a proven cooling technology that minimizes global energy consumption by using warm water up to 40°C.

The BXI switch ASIC is a low latency, non-blocking 48 ports crossbar. It features 48 BXI ports at 100Gb/s. The chip aggregate bandwidth is 1200GB/s (48 ports * 12.5GB/s/ direction * 2 directions).

The 16 virtual channels available with BXI can be used to avoid message dependent dead locks and to improve communications efficiency with QoS.

To balance traffic in case of congestion, each port has a dedicated routing table, and one predefined deterministic route and up to 47 adaptives routes, covering the maximum physical options. With adaptive routing, incoming messages can be directed to less-loaded output ports.

The BXI switch provides an integrated benchmarking & troubleshooting tool called Traffic Generator.

This 100% hardware implementation (integrated into the ASIC) is available on each switch port, 100% configurable via CLI (duration, target, message size, virtual channel, age and allocated bandwidth).

It enables checking of link integrity, behaviour & performance, and application bandwith allocation.

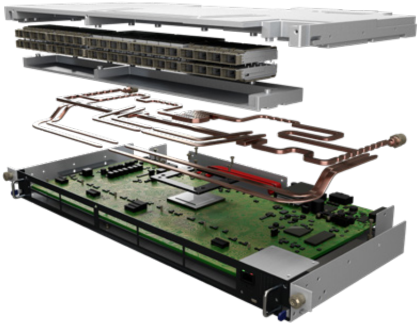


Figure 3: BXI Direct Liquid
Cooled switch for BullSequana XH2000

BXI Software Suite

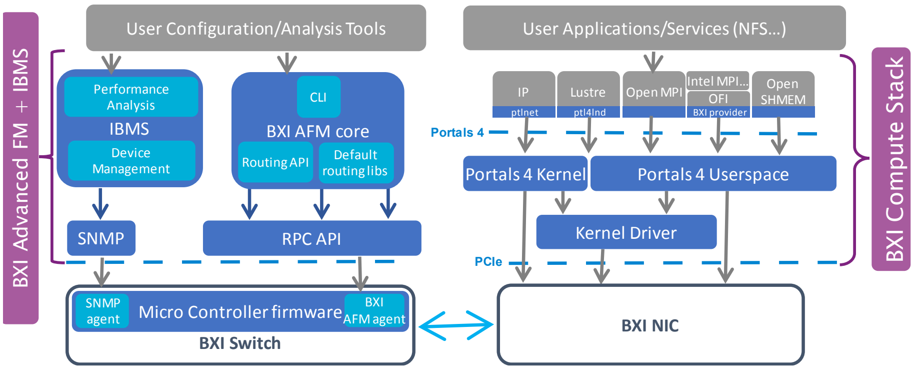
BXI comes with the BXI Software Suite, composed of the BXI Compute Stack that allows users to program and run applications on BXI systems, of the BXI Advanced Fabric Management (AFM) software to install and configure a BXI system and IBMS that is in charge of Device Management and Performance Analysis tasks.

BXI AFM features the following:

- Optimized routing algorithms enabling quick reactions (<15 seconds) to failures and providing a stable platform for running applications.
- Autonomous discovery of topology and miscabling checking.
- Local and global fabric events capture and management allowing quick reactions to complex fabric events.
- A Command Line Interface (CLI) allows to easily configure and control the fabric or to monitor and diagnose possible failures.
- Moreover, the solution is compliant with standard device management software thanks to an SNMP agent embedded in the switch firmware

BXI Compute Stack features the following:

- Parallel applications can take full advantage of the capabilities of the BXI network using MPI or Open SHMEM communication libraries.
- All components are implemented directly using the Portals 4 API.
- Kernel services are also implemented using the kernel Portals 4 implementation.
- A Portals 4 LND (Lustre Network Driver) provides the Lustre parallel filesystem with a direct / native access to Portals 4.
- The PTLnet (IP over Portals) component makes it possible to have large scale, efficient and robust IP communication for legacy software.


BXI Advanced FM + IBMS
User Configuration/Analysis Tools
Performance Analysis
IBMS
Device Management
CLI
BXI AFM core
Routing API
Default routing libs
SNMP
RPC API
SNMP agent
Micro Controller firmware
BXI AFM agent
BXI Switch
User Applications/Services (NFS...)
IP
ptlnet
Lustre
ptl4lnd
Open MPI
Intel MPI...
OFI
BXI provider
Open SHMEM
Portals 4
Portals 4 Kernel
Portals 4 Userspace
Kernel Driver
PCIe
BXI NIC
BXI Compute Stack