BXI V2

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 BullSequana eXascale Interconnect V2 or BXI V2 introduces a paradigm shift in terms of performance, scalability, efficiency, reliability and quality of service for extreme workloads.

The BXI V2 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.

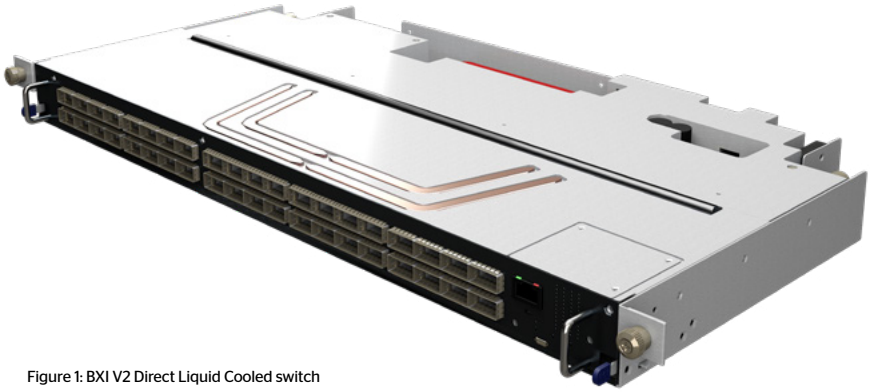


Figure 1: BXI V2 Direct Liquid Cooled switch

Getting rid of the communications overhead

The core feature of BXI V2 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 V2.

As a result, contrary to other commonly used networks, BXI V2 can deliver high communication throughput even when the system is under heavy computation stress.

BXI V2 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 V2 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 V2 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 V2 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 V2 adaptive routing capabilities dynamically avoid communication bottlenecks.

Reliability and resilience

For high reliability, BXI V2 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 V2 components

The BXI V2 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 V2 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 V2 components are detailed overleaf.

BXI V2 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 V2 fat-tree. To form large HPC systems, BullSequana XH2000 cells are interconnected using external BXI V2 switches.

BXI V2 Network Interface Controller (NIC)

The BXI V2 NIC is available as a mezzanine card for BullSequana XH2000 compute blades and as a standard PCIe card for other nodes. The BXI V2 NIC board interfaces each node to the BXI V2 interconnect.

- PCIe gen3 x16 link
  - NIC custom 4x BXI V2 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 V2 Standard PCI card

BXI V2 switch

The BXI V2 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 Atos 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 V2 switch ASIC is a low latency, non-blocking 48 ports crossbar. It features 48 BXI V2 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 V2 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 V2 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.

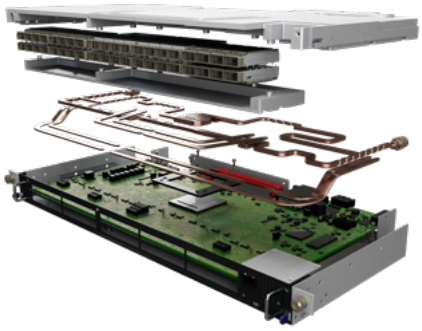


Figure 3: BXI V2 Direct Liquid Cooled switch for BullSequana XH2000

BXI V2 Software Suite

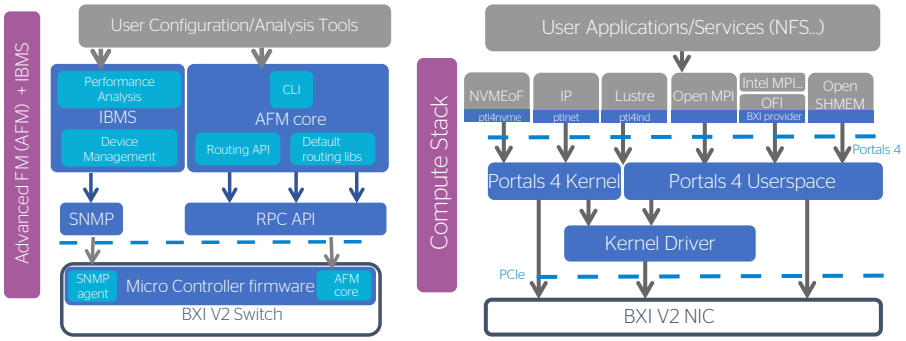
BXI V2 comes with the BXI V2 Software Suite, composed of the BXI V2 Compute Stack that allows users to program and run applications on BXI V2 systems, of the Advanced Fabric Management (AFM) software to install and configure a BXI V2 system and IBMS that is in charge of Device Management and Performance Analysis tasks.

BXI V2 AFM features the following:

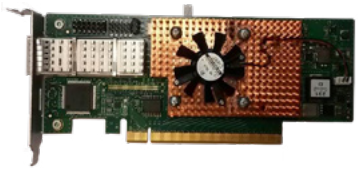
- Optimized routing algorithms enabling quick reactions (<15 seconds) to failures and providing a stable platform for running applications.
- Autonomous discovery of topology and miscabling checking.
- Local and global fabric events capture and management allowing quick reactions to complex fabric events.
- 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 V2 Compute Stack features the following:

- Parallel applications can take full advantage of the capabilities of the BXI V2 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 V2 Standalone Switch	BXI V2 DLC Switch for XH2000
			
Technology	Specs	Portals 4	
Ports	Capacity Type Max aggregate throughput Port-to-port latency	48x 32 x QSFP28 + 8 x QSFP-DD 9,6Tb/s 150ns	
Power supply	Power cord and connector Input voltage Redundancy Hot Plug (Yes or No) Idle power consumption Typical power consumption Max power consumption	C13/C14 100 to 240V, 50-60Hz, AC Yes, dual slots Hot plug 110W Copper: 220W – Optical: 320W Copper: 250W – Optical: 350W	 55V DC Yes N/A 80W Copper: 180W – Optical: 260W Copper: 220W – Optical: 300W
Operating conditions	Temperature range Non-operating/storage Humidity range Storage humidity	0°C to 35°C derating of 1°C/325m above 900m -40°C to 70°C 8% to 80% 5% to 95%	
Cooling	Reversible air flow Redundancy Fan	Yes (as an option) Yes Hot-swappable	N/A N/A N/A
Mechanical	Mounting Dimensions (W x D x H) Weight	19" rack within 1U chassis 17.52" x 21.61" x 1.71" 10.8kg (23.8Lb)	BullSequana XH2000 Rack 20.12" x 10.61" x 1.65" 8.4kg (17.6Lb)
Management	Interfaces LED indicators	1x RJ45 1Gbs Ethernet connector Reset button Recovery mode button Restore-ro-default button Per port: link status System LEDs: System, fans, power supplies Unit ID LED	Ethernet 1Gbs 1000B-X SideBand network Reset button, Recovery mode button Restore to default by SideBand Power LED Unit ID LED 48 LED port UP Recovery mode LED
Connectors		32 QSFP28 & 8 QSFP-DD	
Cable		Passive copper or active fiber	
Safety compliance		EC, IEC, UL and CSA	
EMC compliance		EC, FCC, ICES-03 and VCCI	
Environmental		RoHS II & WEEE directives, REACH regulation	
Warranty		**Standard warranty:** 1 year **Extended Warranty:** consult your local Sales representative	

		BXI V2 NIC Standard PCIe card	BXI V2 NIC Mezzanine Card
			
Technology		BXI 100Gb/s	
HFI specification	Device Type Advanced interrupts ASIC Virtual Lanes	End Point MSI-X Lutecia 1.3 Up to 16 virtual channels and 4 virtual networks	
System Interface		PCI Express Gen 3.0 x16	
Ports	Connectors Max Data Rate	zQSFP 100 Gbps – PCIe x16	TE STRADA Whisper 100 Gbps - PCIe x16
Software	Operating System	Red Hat Enterprise Linux 8.x	
Power Supply	Input voltage Typical power consumption Max power consumption	12VDC + 3,3DVC 18W 27W	12VDC + 3,3DVC 16W 27W
Operating conditions	Temperature range Non-operating/storage temperature Humidity range Storage humidity Thermal Solution	5°C to 38°C -40°C to 70°C 20% 60%, gradient 5%/h 20% 60%, gradient 5%/h Active Heat sink	10°C to 45°C -10°C to 70°C 8% to 80% non-condensing 5% to 95% non-condensing Water-cooled cold plate
Dimensions (W x H)	Card Standard WeightWeight Extra Bracket	Standard Low Profile: 2.7" (68,9mm) H x 6.6" (167,65mm) L / Single Slot PCIe Gen3 1,1Lb (500g) (with dissipator) Full Height 4.4» (111,1mm) standard PCIe bracket	4.6" (118mm) x 5.1" (130mm) N/A 500g (1,1Lb) (with cold plate) N/A
MTU size		Message up to 4GBytes (divided in 64Bytes packet)	
Management LED Indicators		Link status indicator (Green) Traffic indicator (Yellow blinking)	N/A
Cables		Passive copper or active fiber	Passive copper
Safety compliance		EC, IEC, UL and CSA	
EMC compliance		EC, FCC, ICES-03 and VCCI	
Environmental		RoHS II & WEEE directives, REACH regulation	
Warranty		**Standard warranty:** 1 year **Extended Warranty:** consult your local Sales representative	

For more information: **hpc@atos.net**

Atos, the Atos logo, Atos|Syntel, and Unify are registered trademarks of the Atos group. June 2020. © 2020 Atos. Confidential information owned by Atos, to be used by the recipient only. This document, or any part of it, may not be reproduced, copied, circulated and/or distributed nor quoted without prior written approval from Atos.