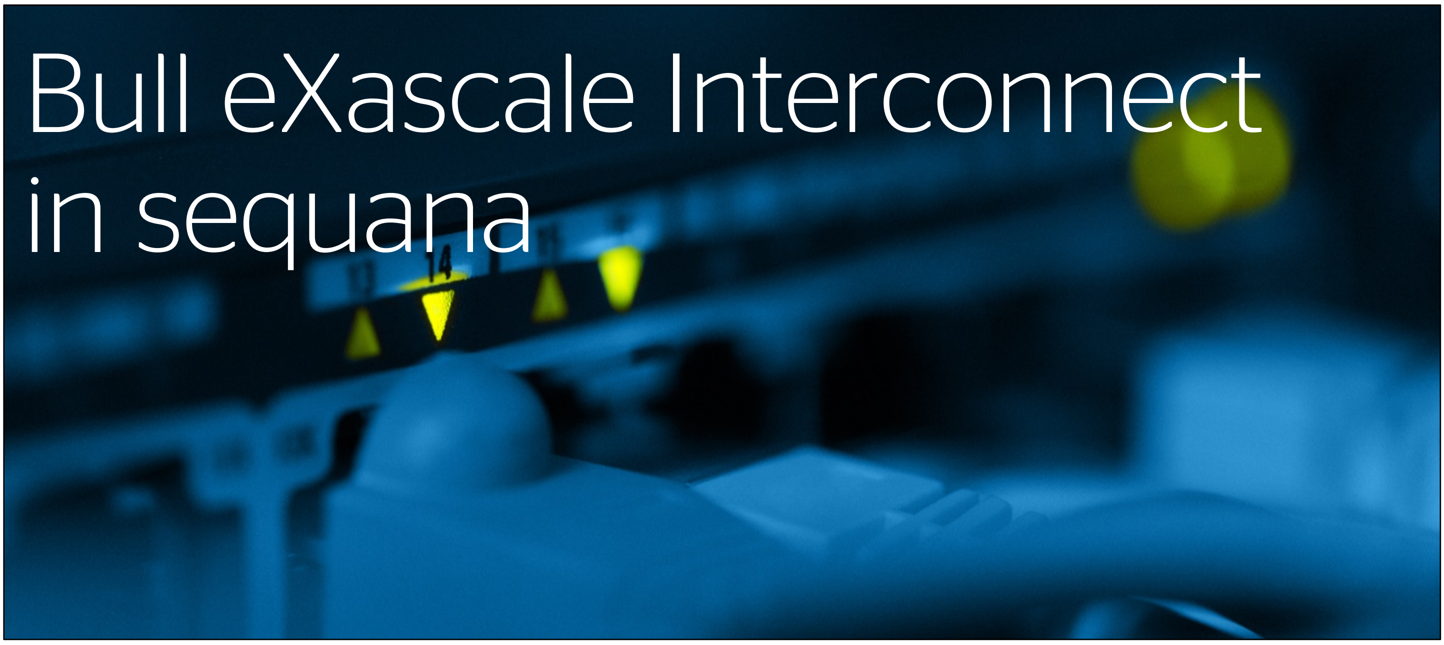
BXI

Bull eXascale Interconnect in sequana



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 Bull's 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 for the first version), it features:

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

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.

Bull sequana cell size and topology with BXI

With 48 ports per BXI switch, a non-blocking 2-level fat-tree provides 576 connections with 24 L1 switches and 24 L2 switches within a sequana cell.

As one compute node supports two NICs (Network Interface Controller), the maximal Bull sequana X1000 cell configuration is 288-nodes. The two NICs in a given node are connected to two different L1 switches for improve availability in case of NIC or switch failure.

A cost-efficient alternative is to have only one NIC per node, ie. only 12 L1 switches and 12 L2 switches.

Bull
atos technologies

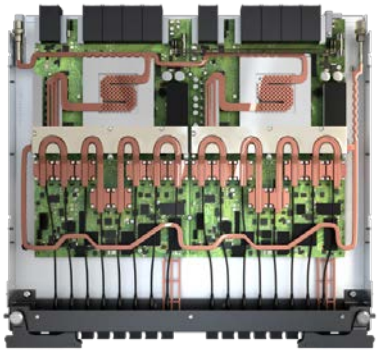
BXI Network Interface Controller ASIC

The BXI NIC is available as a mezzanine card for sequana 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 messaging
  - 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

BXI Mezzanine card for sequana blade

BXI switch ASIC



BXI L2 Direct Liquid Cooled switch for sequana

The BXI switch is available as L1/L2 switch modules in sequana with copper connectors (L1/L2) and as an external switch with 48 MPO optical connectors. The Bull sequana 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).
Large composite switches (288 or 576 ports) are built by combining two levels of ASICs.

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

Each port has a dedicated routing table, and for each destination, three output ports are defined: one deterministic route and two adaptive routes to balance traffic in case of congestion. With adaptive routing, incoming messages can be directed to less-loaded output ports.

For time-stamping accuracy which might be required for communication analysis, the BXI switch provides a global clock mechanism to synchronize all nodes in a system with a margin that does not exceed 1μs even for 32k nodes.

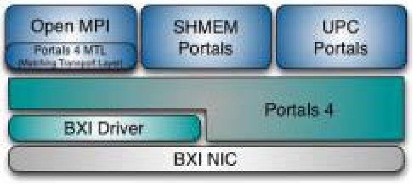
BXI fabric management

The fabric management is composed of routing solutions enabling quick reactions (<5 seconds) to failures and providing a stable platform for running applications (reliability for up to 25% connection failures on a 64K nodes fabric),

- Autonomous discovery of topology and miscabling checking,
- Local and global fabric events capture and management allowing quick reactions to complex fabric events,
- High frequency performance counters sampling (including histograms) provide accurate information on traffic. Four filters are available to define specific events, e.g. messages generated by a particular user, set of nodes, or traffic type.
- An extensive command line interface allows to easily configure and control the fabric or to monitor and diagnose possible failures.

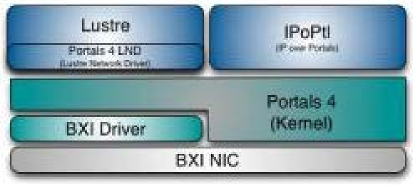
BXI application environment

BXI comes with a complete software stack to provide optimal performance and reliability to all traditional HPC components.



BXI Computing stack

- Parallel applications can take full advantage of the capabilities of the BXI network using MPI, SHMEM or UPC 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 IPoPtl (IP over Portals) component makes it possible to have large scale, efficient and robust IP communication for legacy software.



BXI Kernel services

For more information: Please contact hpc@atos.net

Your business technologists. **Powering progress**

bull.com