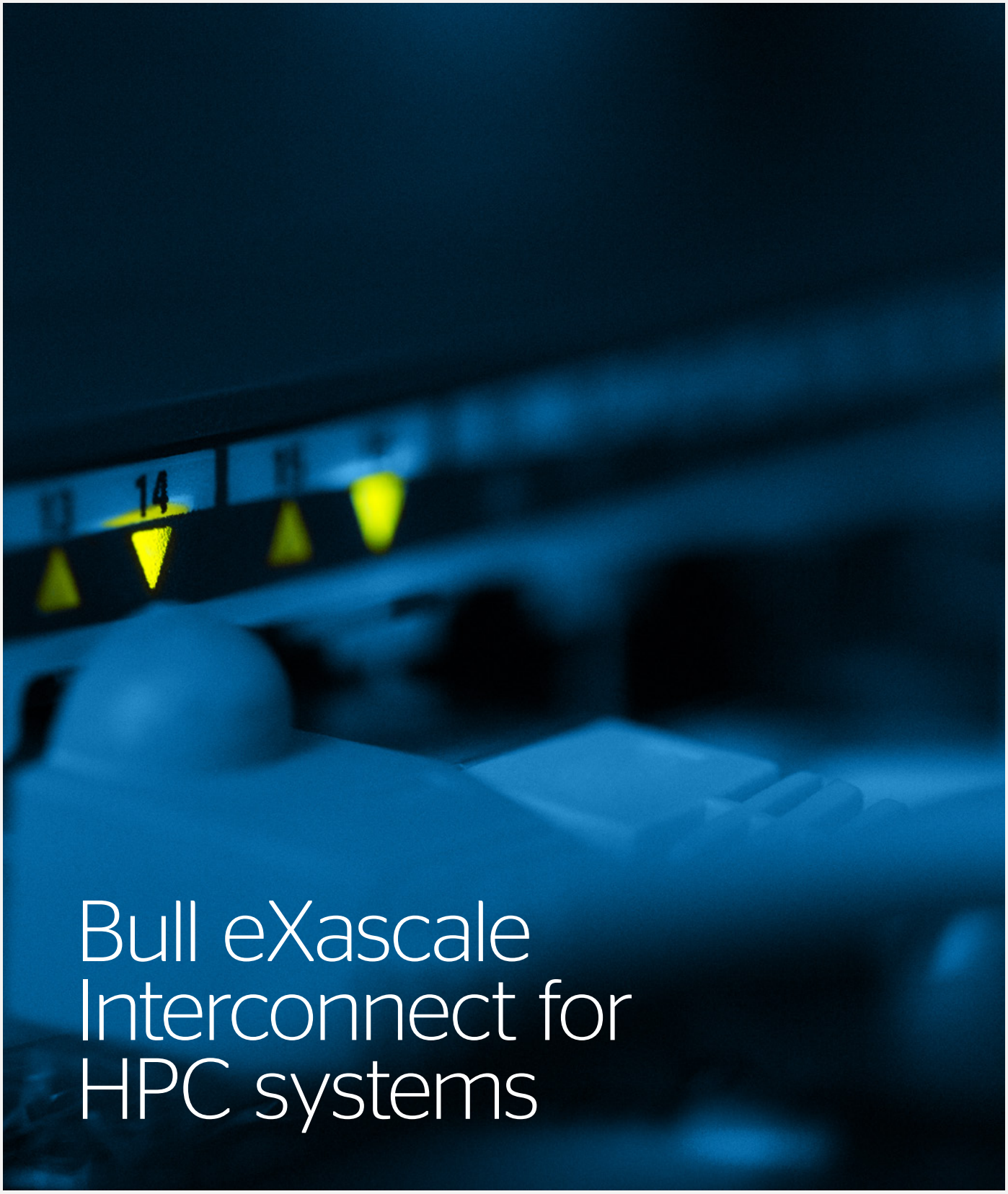
BXI



Bull eXascale Interconnect for HPC systems

Bull
atos technologies

Summary

Exascale High Performance Computing (HPC) systems will provide a thousand times (1000x) higher performance than today's petascale supercomputers. This goal will be achieved through an explosion 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. Atos introduces the **Bull eXascale Interconnect or BXI,** which is a cornerstone of its **Bull Exascale program.** BXI ambitions to introduce a paradigm shift in terms of **performance, scalability, efficiency, reliability** and **quality of service** (QoS) for extreme HPC workloads.

Scaling HPC applications to the next level of performance

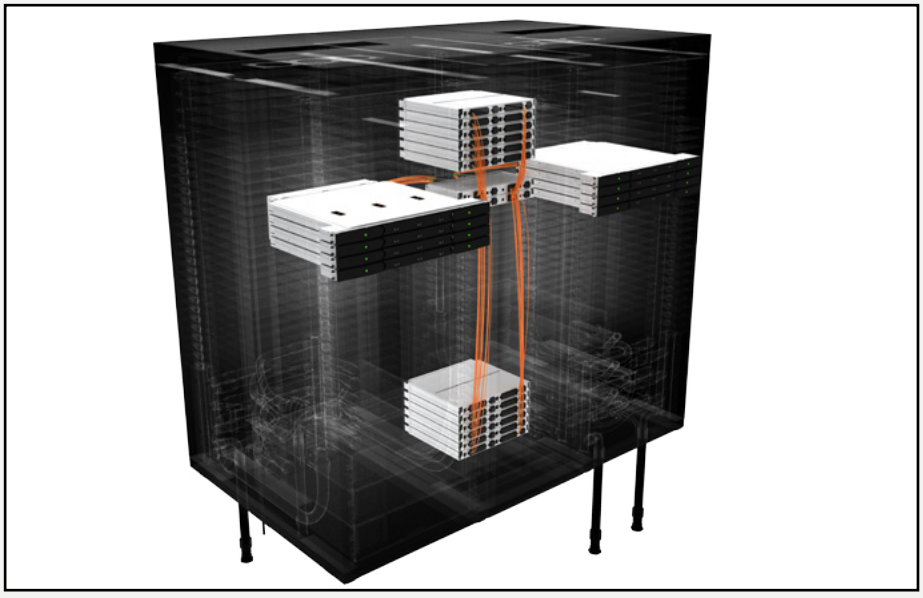
To deliver Petascale performance (1 Petaflops= 10^15 floating point operations per second), today's HPC applications rely on a high level of parallelism. Applications are distributed across thousands of processors and each processor integrates multiple/many cores (from 10 to 100); overall up-to a million threads might be used. A fast network is essential to interconnect all processing and storage nodes; it provides for each thread access to the global dataset.

The vast majority of HPC applications today use an explicit communication scheme: the **Message Passing Interface (MPI)** library. Since MPI is available on all type of platforms, using MPI guarantees that an application will run everywhere even though it implies an important development effort. In fact, MPI demands that the programmer identifies and explicitly codes all remote data access. In order to facilitate parallel programming, new schemes such as **PGAS (Partitioned Global Address Space)** have been proposed. An efficient HPC network must accelerate MPI communications and it must also provide support for the new PGAS programming model.

Another important characteristic of HPC systems is the increasing importance of accelerators such as **GPUs and Manycore** processing elements. These computing elements provide more computing performance for less power consumption than traditional general purpose CPUs. They rely on SIMD and vector architectures to deliver more flops, even though they run at a slower frequency to be more energy efficient. As a result, these computing elements (GPUs and Manycores) do not fare well for IOs. To accelerate communications, the HPC network must offload all communication tasks to the interconnection hardware.

Exascale (1 Exaflops= 10^18 floating point operations per second) will require 1000x more parallelism as computational cores frequency is not expected to rise. To match this requirement, not only future processors will embed more cores than today (up-to 100s), but the processor count will also increase. The Exascale system network must scale to tens of thousands of nodes.

The new BXI (Bull eXascale Interconnect) network developed by Atos delivers the performance, scalability and efficiency required for Exascale applications performance.



BXI overview

The BXI network scales up to 64k nodes; it features high-speed links (100 Gb/s), high message rates and low latency across the network. Most importantly, to boost the communication performance of HPC applications, BXI provides full hardware offload of communications, enabling CPUs to be fully dedicated to computational tasks while communications are independently managed by BXI.

The BXI architecture is based on the Portals 4 communication library. The BXI hardware primitives map directly to communication libraries such as MPI (Message Passing Interface) and PGAS (Partitioned Global Address Space); they also support RDMA operations. BXI accelerates all MPI communication types, including the latest MPI-2 and MPI-3 extensions such as **asynchronous collectives**. Furthermore, Portals 4 non-connected protocol guarantees a minimum constant memory footprint, irrespective of system size.

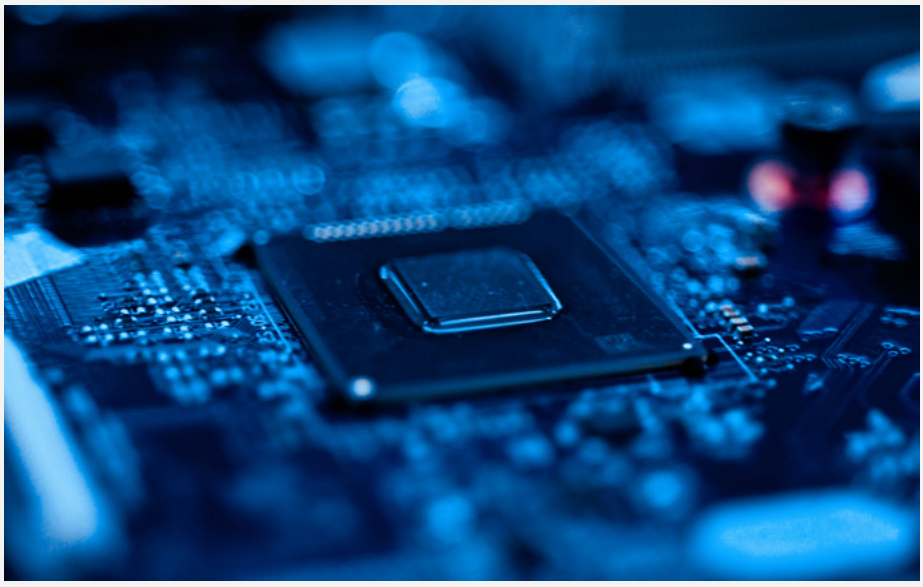
An interesting aspect of communication offloading in hardware is that BXI delivers high communication throughput even when the system is under heavy computational stress. Since the host processor is not involved in data movements, BXI also minimizes the impact of communication on its internal cache, reducing cache misses and TLB invalidations.

For more flexibility, the BXI network architecture is based on two separate ASIC components, a NIC and a switch. The NIC can be attached through a standard PCIe-3 interface to multiple types of nodes such as generic CPUs (e.g. x86), HPC-specific processors such as Intel® Xeon Phi™ and hybrid nodes with GPU accelerators. With the 48 ports of the BXI switch, large systems can be built with fewer elements, thus optimizing cost. Furthermore, with fewer hops in the data path, the communication latencies are reduced and congestion points are limited as well.

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.

End-to-end error checking and link level retry together with ECC enhance communication reliability and resilience without jeopardizing communication performance.

To improve reliability, an end-to-end error checking detection and recovery mechanism is integrated into the NIC. All data paths within the BXI chips are protected with ECC and eventual transient errors are detected in each link of the fabric; in such case the data is retransmitted locally. Overall, on a 32,000 node system, the expected rate of undetected errors is lower than one in 500 years.
Finally, the BXI network comes with a **complete out-of-band software management stack.** It provides all functions necessary for monitoring, routing and performance analysis via the open standard SNMP protocol. The runtime environment supports optimized MPI libraries and PGAS languages for HPC applications. For global storage, BXI features a **native implementation of the Lustre parallel filesystem.**



BXI Network Interface Controller (NIC)

The BXI Network Interface Connector (NIC) interfaces with the system node through a 16x PCI Express gen3 interface and with the BXI fabric through a 100Gb/s BXI port. The BXI NIC is available in a standard PCIe card form factor to interface with standard servers; it is also mounted on a mezzanine card to interconnect the nodes of the Bull sequana platform.

The BXI NIC provides dedicated communication FIFOs, DMA engines for sending and receiving, and matching engines for reception. It also features virtual to physical address translation with no need for memory pinning in the OS kernel, as well as rank (MPI, SHMEM ...) to physical location translation. The hardware acceleration makes it possible to have fast tracking and matching for two-sided communications, e.g. MPI send/receive, either blocking or non-blocking. Upon reception of a particular message, the data is copied to the target memory location without involving the host processors, thus allowing computational resources to be freed from communication tasks. **This hardware feature enables a high message throughput, even under heavy load, and without interfering with computation.**

With the Portals 4 non-connected model, the MPI library has a constant memory footprint, regardless of the number of MPI ranks. The BXI protocol also guarantees in hardware the ordering of Portals 4 operations, allowing for a direct MPI implementation with no extra software overhead in case of retransmissions. **The BXI NIC offers native support for PGAS applications and MPI-3 one-sided operations communications (put/get/atomics). PGAS non-matching operations use a dedicated fast path within the NIC, thus allowing the highest message issue rate and the best latency.**

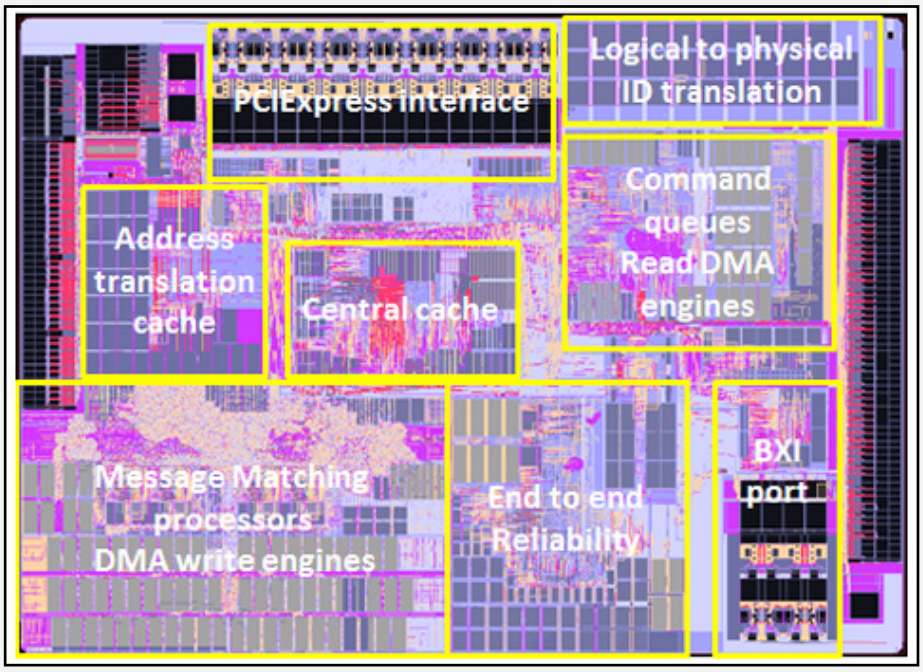


Figure 1: BXI NIC floor plan

OS Bypass capabilities

The BXI NIC provides **OS bypass capabilities.** Sending or receiving a message is controlled by the NIC without interrupts or OS involvement. Dedicated FIFOs permit direct access to the network from any computing task through Portals, so applications issue commands directly to the NIC, avoiding kernel calls. This ensures minimal CPU usage when posting a Portals command to improve the issue rate and latency. The BXI NIC translates virtual addresses into physical addresses for incoming network requests, local reads and writes. The NIC translation unit sustains high message rates, low latency communication even with a random access pattern to the host memory. The NIC maintains a cache of the active page table with each level of the page table hierarchy independently stored. Moreover, the translation unit pipeline supports a sufficient amount of ongoing translations to cover the latency of cache misses without impacting the bandwidth or the message rate achieved by the NIC.

Communications offload in hardware

The main feature of BXI is **communications offload in hardware.**

- **Logical to physical ID translation:** At the application level destination nodes are specified by their logical address within a job (MPI rank, SHMEM PE). In BXI, this address is translated into a physical address (node ID plus process ID) with the use of a low latency embedded RAM to improve performance and avoid cache misses in the host processor.

- **MPI matching:** Optimizing MPI two-sided communications such as MPI_Send/MPI_Recv and especially their asynchronous variants MPI_Isend/MPI_Irecv is crucial for many HPC operations. In particular, on the receiving side, this requires to promptly match an incoming message with a previously posted MPI_Recv. With BXI, the NIC receive logic expedites this matching in hardware. It also handles unexpected messages and prepares for quick matching when the corresponding MPI_Recv is posted. The NIC locally maintains processing lists for all pending MPI events.

- **Atomic units:** In addition to traditional remote read and writes, the BXI NIC can also perform atomic operations, (conditional) swaps and fetch/atomics. This makes it possible to have a rich remote memory API to implement distributed locks, hash tables or any classical algorithm run traditionally on multiple threads - at a much larger scale. The BXI Atomic Unit handles all MPI and PGAS data types for element sizes up to 256 bits.

- **Triggered operations:** Finally, using Portals 4 triggered operations, complex algorithms like collective operations can also be completely offloaded to the NIC, which permits an efficient MPI-3 implementation for non-blocking collectives operations. Triggered operations is a mechanism to chain networking operations. Most commonly used algorithms (trees, pipelines, rings ...) can therefore be "programmed" and executed in the NIC without the intervention of any host processor.

End-to-end reliability

The BXI fabric is designed to provide an exceptional reliability. In particular, the NIC implements an end-to-end protocol to cover for other possible transient errors and permanent errors as well (e.g. failure of a network link). To check message integrity, Cyclic Redundant Checks (CRC) error-detecting codes are added to each message. The message ordering required for MPI messages is checked with a 16 bit sequence number. Finally a **go-back-N protocol** is used to retransmit lost or corrupted messages. For such purpose, the NIC holds a copy of active messages in its central cache. Soft Error Rate (SER) mitigation: Transient errors (also called "soft" errors) may corrupt the state but will not result in any permanent fault in the silicon. Several strategies are employed in the NIC to minimize the overall SER impacts. To this end, **ECC protection** combined with **periodic patrol scrubbing** is employed in all memories of significant size and on all primary data paths. Finally, **"voting" flops** are used in the most sensitive parts of the design.

BXI switch ASIC

The BXI switch ASIC, the BXI, is a low latency, non-blocking 48 ports crossbar. As seen in fig. 2, the die size is dimensioned by the number of SerDes which occupy the whole chip perimeter. The switch ASIC is the first chip to integrate as many as 192 SerDes running at 25Gbps. It delivers a global bidirectional bandwidth of 9600Gb/s (48 ports * 100 Gb/s * 2).

The BXI switch ASIC is directly embedded in the L1-L2 switch modules of the Bull sequana platform which targets large configurations. It is also packaged as an external switch with 48 optical connectors.

The high-radix BXI switch simplifies the architecture of large scale systems since fewer components are required. For a given system size, this also helps reducing the network diameter and with fewer hops in the communication path, it shortens the latencies.

The BXI switch features per-port destination based routing, which allows for fabric flexibility and efficient fine-grain adaptive routing. A default path is defined for each destination node in a full 64k entries routing table. In addition, the routing table also specifies the alternative paths to be used for adaptive routing in case of congestion.

In case of a link failure, the routing tables can be dynamically reconfigured to avoid the failing path. The application may proceed without service interruption, thanks to end-to-end protocol implemented in the BXI NIC.

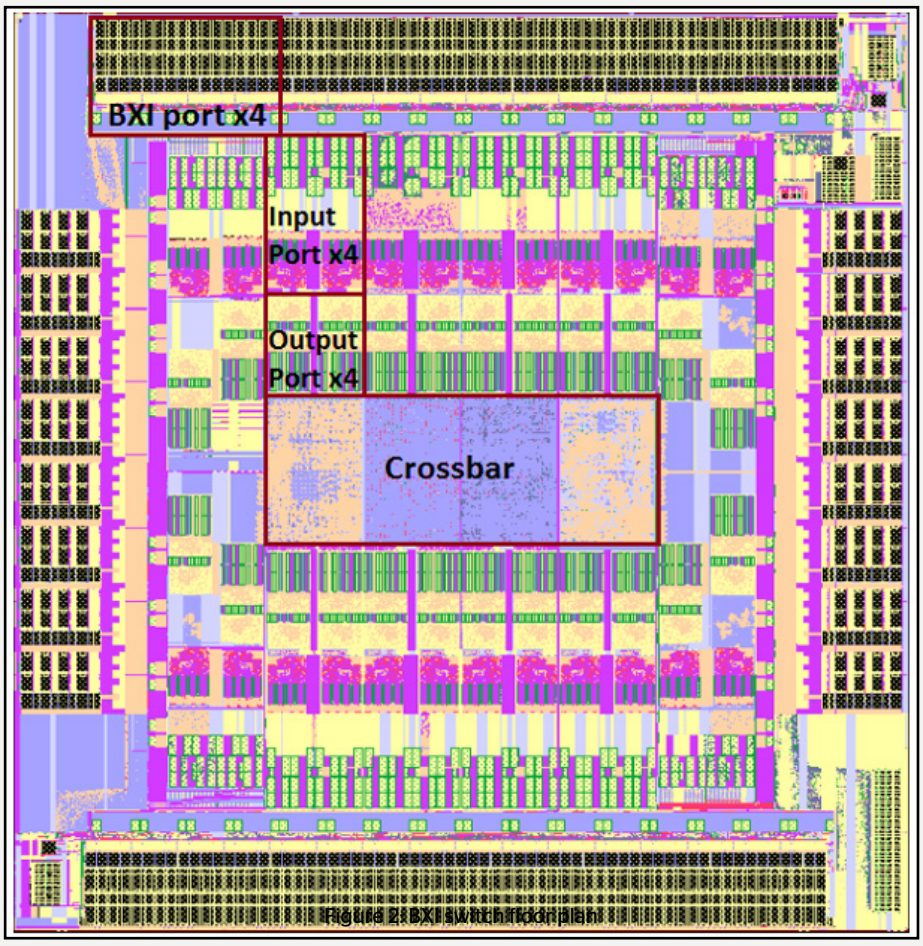


Figure 2 BXI switch floorplan

The BXI switch is managed completely out-of-band through a microcontroller.

Network performance monitoring

An extensive set of configurable counters is implemented in each port in both directions.

Some counters are specific to a given measured event, such as global message/flit counting and buffer/credit occupation. The input port buffers are monitored using a fully configurable 4-bin histogram.

In addition, **16 counters per port per direction can be programmed to indicate the events they should count.** Any header field can be masked or required to match. It is also possible to specify the virtual channel(s) to be tracked and the counting granularity (message or flit). For example, these counters can be configured to measure events corresponding to a given job or a specific source.

Multiple possible topologies

Thanks to the key features implemented in the BXI switch, **there is no restriction on the type of topology which can be implemented using the BXI fabric.** Fat-trees can be implemented easily. In this case, all the VCs can be used for quality of service policies. Moreover, the port-based routing tables allow for load-balanced adaptive routing and failures handling.

Directly connected topologies such as Torus, Hypercubes, Flattened-Butterfly, Dragonfly or Slimfly can also be implemented efficiently thanks to the switches' high radix, the availability of a high number of virtual channels and the per-port routing tables.

BXI management software

The BXI fabric is managed out-of-band through an Ethernet network for all management related operations, ranging from monitoring to routing tables updates. With such a setup, the management functions do not interfere with the BXI traffic. This also guarantees that the BXI components are accessible at all times even in the event of a partial or complete network failure. The BXI fabric management is responsible for the monitoring, the routing and performance analysis tasks.

Monitoring modules react to errors and events emitted by the components. Monitoring modules are also able to correlate errors and raw events. They are thus able to detect complex failure patterns and report synthetic information to the administrators. Yet, the monitoring modules must detect and handle silent failures and adopt a proactive behavior. By querying the equipment, on a regular basis, the BXI monitoring modules both detect silent errors and unexpected situations, such as misplaced fabric equipment (including cables). **The monitoring modules of the BXI fabric management can also be completely distributed and hierarchically organized, both to scale up to the biggest configurations and to provide system-wide event correlations.**

Monitoring modules can finely detect the state changes of the fabric elements. These changes are reported to the administrator and transmitted to the other fabric management modules, e.g. to trigger re-routing operations. The states of all fabric elements are stored in a distributed lightweight database which is resilient and scales up to the largest system sizes (64k nodes).

The routing module is responsible for computing and distributing the routing tables to the switches. It relies on the information gathered and updated by the monitoring modules, which guarantees that

the routing algorithms will compute optimal tables that will provide the expected bandwidth to applications. As re-routing reaction times are capital, when a failure that would trigger a re-routing operation is detected, a highly distributed mechanism is launched, at switch level, to reroute the traffic around the failing component within a five seconds upper-bound. This mechanism only exploits locally available information and is intended to react quickly and preserve a functional fabric, even though the resulting routing tables may not be optimal. Meanwhile, updated fully balanced routing tables are computed with fabric-wide knowledge before being re-dispatched.

The performance analysis module takes advantage of the error and performance counters available in the BXI ASICs. This extensive set of counters provides accurate information regarding traffic in each switch port. In order to scale up to the largest system sizes, the actual sampling is done at switch level, according to rules that can be dynamically changed during production, depending on troubleshooting or profiling needs. Four independent sampling rules can be defined for each switch. A sampling rule consists in a set of counters to sample and a sampling frequency. **These measurements sampled directly at high frequency (up to 1Hz) complement HPC applications profiling, by providing fabric side information.**

The BXI software is fully integrated in Bull supercomputer suite version 5, a scalable, open, and robust software suite that meets the requirements of even the most challenging high performance computing environments, including the need for enhanced security.

BXI application environment

The BXI application environment represented in figure 3 provides the software layer to interact directly with the BXI network through native Portals4 interfaces using MPI, SHMEM or UPC communication libraries. All components are implemented using the native Portals 4 API.

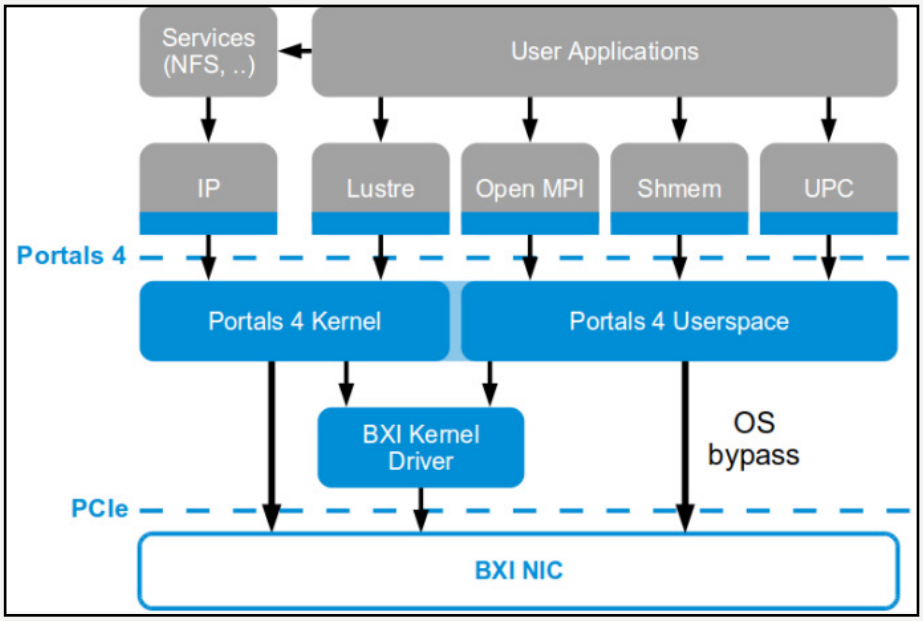


Figure 3: BXI application environment

Kernel services also use the Portals 4 kernel implementation. A Portals 4 LND (Lustre Network Driver) provides the Lustre parallel filesystem with a native access to Portals 4. Finally, the IPoPtl (IP over Portals) component makes it possible to have large-scale, efficient and robust IP communication.

Communication components interact with the BXI Kernel driver to get direct access to the BXI NIC. After the initialization phase, all communications go straight to the NIC in both directions without involving the kernel.

The Portals 4 API is a rich networking interface which has been designed to allow direct mapping to MPI and PGAS. Little extra software code is therefore needed to convert an MPI_Send or a SHMEM_PUT to a Portals Put. Since User space programs have direct access to the BXI NIC through a virtual memory mapping of the PCI address space, the total communication overhead from the application call to the hardware access is limited to its minimum.

The Portals API supports a direct implementation of the MPI-1 two-sided semantics (MPI_Send/MPI_Recv); it also permits the same direct approach for MPI2/3 RMA (One-sided) operations, using Put, Get, Atomic and FetchAtomic primitives. Portals Triggered operations are used to implement collective operations, **synchronous (MPI-1 or PGAS)** or **asynchronous (MPI3).** **Using triggered (chained) operations, collective operations are offloaded to the NIC and get optimal latency as well as full hardware acceleration for reduction operations.**

IP and Lustre use a dedicated interface to Portals within the kernel; it gives them direct access to the NIC with a full bandwidth, latency and issue rate. To improve general performance and avoid interferences between core services (IP, Lustre) and computing communication (MPI, PGAS), kernel and user flows are separated through QoS rules so that they use dedicated separate paths within the NIC and switches. This ensures mutual protection between those two types of traffic, to avoid the performance degradation usually seen during heavy I/O phases.

NIC management tools support NIC configuration and troubleshooting as well as performance analysis. A set of commands is available to the administrator to query the NIC state and configure it. Fabric Management software also use these tools to adjust NIC configuration (e.g. for QoS) and retrieve NIC-side performance counters.

BXI in the Bull sequana platform

Atos has introduced in 2015 the versatile Bull sequana platform. With this platform, power efficient and densely packed "cells" can be assembled in different manners, depending on fabric size and cost constraints. A "cell" is composed of 3 cabinets with up to 288 nodes

The Bull sequana exascale platform can host a variety of compute nodes based on general purpose CPUs or HPC accelerators such as Intel® Xeon Phi™ or GPUs. Besides BXI it also supports other types of high performance networks such as InfiniBand EDR. It is cooled with the Bull Direct Liquid Cooling (DLC) solution, which can remove the heat generated in such a dense configuration (up-to 180 kW for the 3-cabinet cell) with little overhead. To improve system reliability, Bull sequana supports redundant power supply units and redundant hydraulic heat exchangers, and two redundant management nodes in each cell.

Bull sequana embeds the BXI fabric configured as a non-blocking 2 levels fat-tree (fig. 5) **and an Ethernet management network.** The links between the NICs and the L1 switches are drawn horizontally, while the L1-L2 connections are laid out in the vertical "cable backplanes". All internal links use passive copper wire connections to minimize costs.

Figure 4: Bull sequana exascale platform 3 cabinet cell

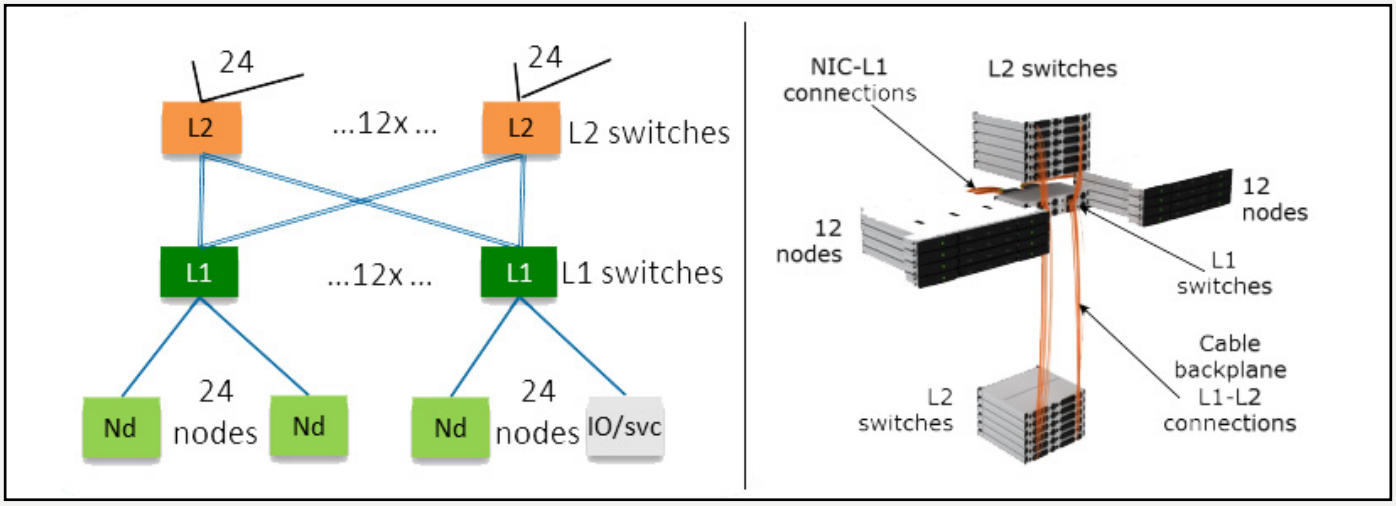


Figure 5: Bull exascale platform network logical configuration and physical layout

To form larger systems, cells are interconnected with active optical links. Each node can be configured with either 1 or 2 NICs. With 2 NICs per node, the cell-embedded fabric includes 576 (=242) end points, which is ideal with 48-port switches. Service nodes such as I/O gateways, administration or login nodes may replace compute nodes and plug in directly into the BXI fabric at L1 level without any need for extra switch infrastructure.

For medium size systems, up-to 2592 nodes, the cells can be connected directly to each other (fig. 6), without any other switch. This hybrid fat-tree/all-to-all topology is quite compact since the longest path in the network involves only 3 hops between switches; it also provides full global bandwidth.

For larger systems, cells can be connected using external BXI switches, and form a standard fat-tree (fig. 7). With a single level of external switches (L3 switches), this indirect topology allows for scaling up to 27k end-points in 48 cells. For such size, this small diameter topology (4 hops max between switches) compares favorably against other topologies such as Cray Aries dragonfly (5 hops) and InfiniBand fat-trees (6 hops). This configuration also features full bisection bandwidth. It is highly resilient and can sustain multiple components failure with only minimum impact on communication performance.

With the addition of a fourth level of switches, the maximum configuration supported by BXI is reached (64k nodes).

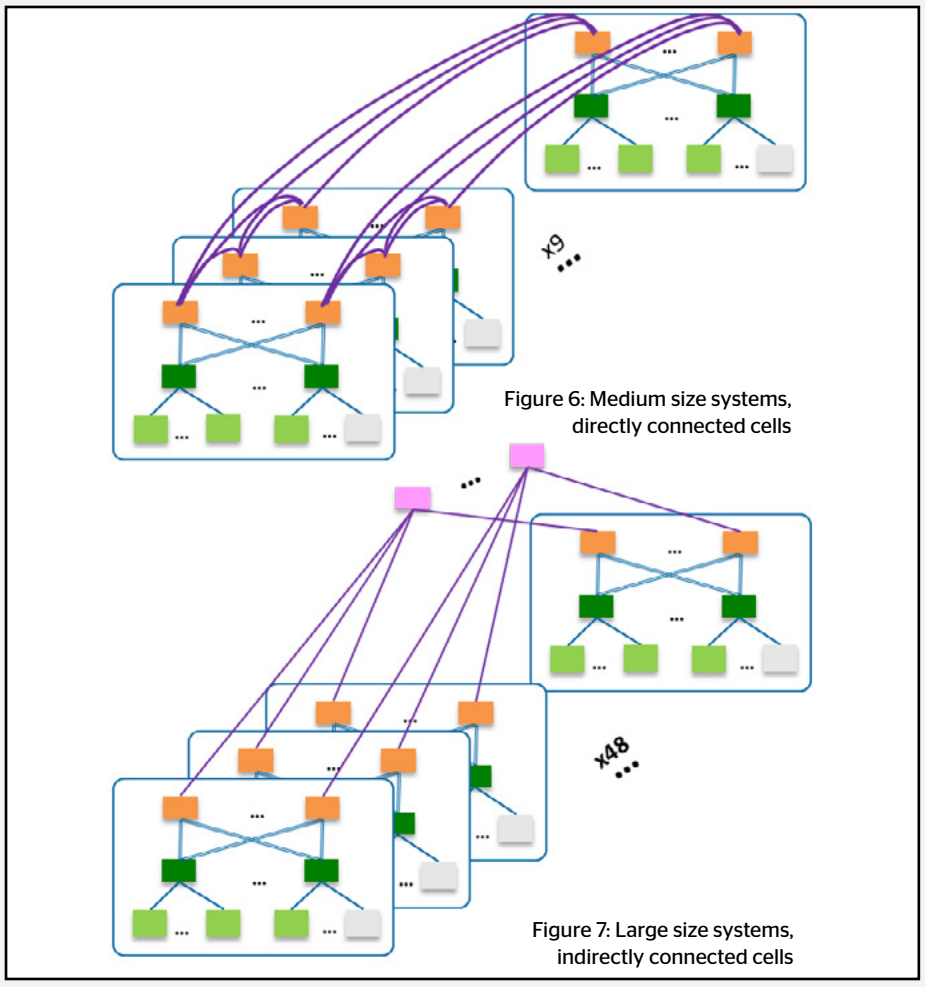


Figure 6: Medium size systems, directly connected cells

Figure 7: Large size systems, indirectly connected cells

Conclusion

The BXI network provides the highest level of communication performance for HPC applications whether looking at bandwidth, latencies, or message rates. Through offload to the hardware components, BXI communications do not interfere with computations. With a native interface to Portals4, MPI libraries are fully optimized for the whole range of communication primitives. Similarly, BXI enables efficient implementations for SHMEM or PGAS environments. Finally BXI offers a complete set of RAS features. As a result, on systems such as Bull sequana, BXI allows HPC applications to scale to millions of threads.

About Atos & Bull

Atos SE (Societas Europaea) is a leader in digital services with pro forma annual revenue of circa € 12 billion and circa 100,000 employees in 72 countries. Serving a global client base, the Group provides Consulting & Systems Integration services, Managed Services & BPO, Cloud operations, Big Data & Cyber-security solutions, as well as transactional services through Worldline, the European leader in the payments and transactional services industry. With its deep technology expertise and industry knowledge, the Group works with clients across different business sectors: Defense, Financial Services, Health, Manufacturing, Media, Utilities, Public sector, Retail, Telecommunications, and Transportation.

Atos is focused on business technology that powers progress and helps organizations to create their firm of the future. The Group is the Worldwide Information Technology Partner for the Olympic & Paralympic Games and is listed on the Euronext Paris market. Atos operates under the brands Atos, Atos Consulting, Atos Worldgrid, Bull, Canopy, Unify

For more information, visit **atos.net**

About Bull, the Atos technologies for the digital transformation

Bull is the Atos brand for its technology products and software, which are today distributed in over 50 countries worldwide. With a rich heritage of over 80 years of technological innovation, 2000 patents and a 700 strong R&D team supported by the Atos Scientific Community, it offers products and value-added software to assist clients in their digital transformation, specifically in the areas of Big Data and Cybersecurity.

Bull is the European leader in HPC and its products include bullx, the energy-efficient supercomputer; bullion, one of the most powerful x86 servers in the world developed to meet the challenges of Big Data; Evidian, the software security solutions for identity and access management; Trustway, the hardware security module and Hoox, the ultra-secure smartphone. Bull is part of Atos.

For more information: **www.bull.com**

For more information: **fr.directionmarketing@atos.net**
© Bull SAS – 2014 - RCS Versailles B 642 058 739 – All trademarks mentioned here in are the property of their respective owners. Bull reserves the right to modify this document at any time without prior notice. Some oers or parts of oers described in this document may not be available locally. Please contact your local Bull correspondent to know which oers are available in your country. This document has no contractual significance.
UK: Bull Maxted Road, Hemel Hempstead, Hertfordshire HP2 7DZ / USA: Bull 300 Concord Road, Billerica, MA 01821 / Bull – rue Jean Jaurès - 78340 Les Clayes-sous-Bois – France

IMPRIM'VERT®

This brochure is printed on paper combining 40% eco-certified fibers from sustainable forests - management and 60% recycled fibers in line with - current environment standards (ISO 14001).

Trusted partner for your **Digital Journey**

Atos