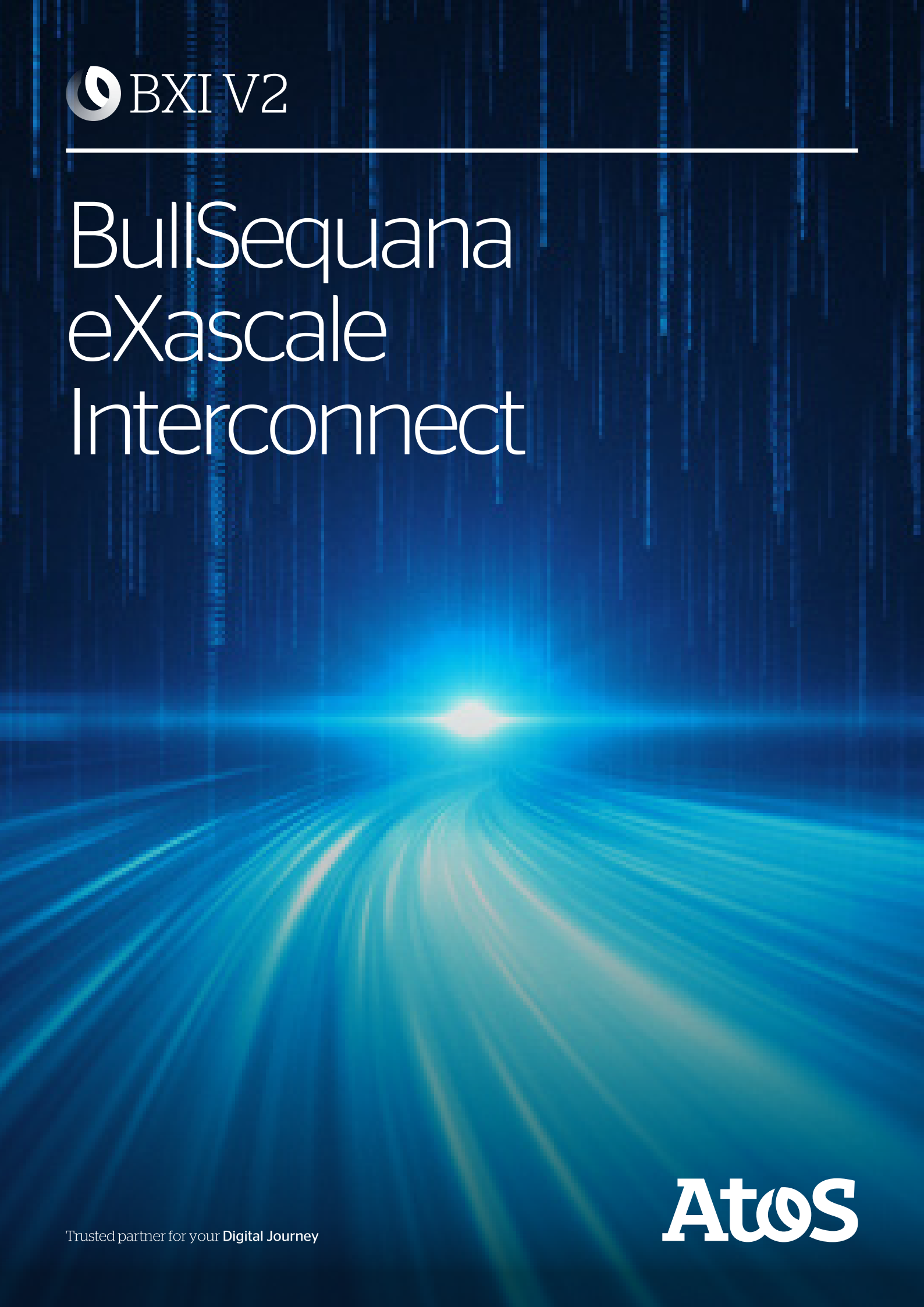

BXI V2
BullSequana
eXascale
Interconnect
Trusted partner for your Digital Journey
Atos

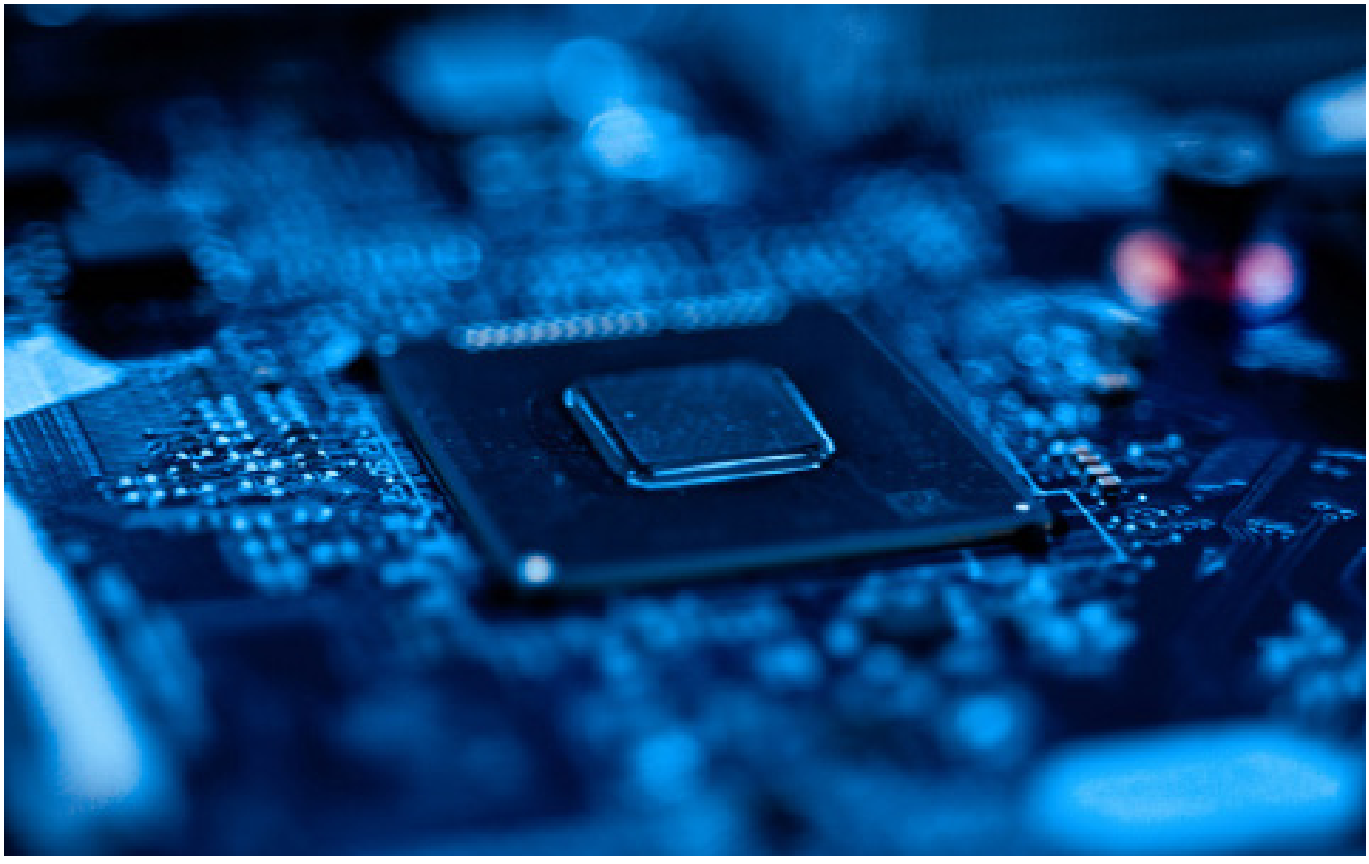
Summary

Exascale High Performance Computing (HPC) systems will provide one thousand times (1000x) higher performance than today's petascale supercomputers. This goal will be achieved through a large increase in the number of nodes/cores, of data volume and of data movement.
At such a scale, optimizing the network which is the system's backbone becomes a major contributor to global performance. Atos introduces the BullSequana eXascale Interconnect or BXI V2, a cornerstone of its eXascale program. BXI V2 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 each integrating 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 access to the global dataset for each thread.

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 it guarantees that an application will run everywhere even though it implies an important development effort. In fact, MPI requires 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 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 yet require less power consumption than traditional general purpose CPUs. They rely on SIMD (Single Instruction Multiple Data) 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 Inputs and Outputs (I/Os). To accelerate applications, 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 core frequency is not expected to rise. To match this requirement, not only will future processors embed more core 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 V2 (BullSequana eXascale Interconnect V2) network developed by Atos delivers the performance, scalability and efficiency required for Exascale applications performance.



BXI V2 overview

The BXI V2 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 V2 provides full hardware offload of communications, enabling CPUs to be fully dedicated to computational tasks while communications are independently managed by BXI V2.

The BXI V2 architecture is based on the Portals 4 communication library. The BXI V2 hardware primitives map directly to communication libraries such as MPI and PGAS; they also support Remote Direct Memory operations. BXI V2 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.

A beneficial aspect of communication offloading in hardware is that BXI V2 delivers high communication throughput even when the system is under heavy computational stress. Since the host processor is not involved in data movements, BXI V2 also minimizes the impact of communication on its internal cache, reducing cache misses and TLB (Translation Lookaside Buffer) invalidations.

For more flexibility, the BXI V2 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 X86 CPUs (e.g. Intel, AMD), ARM and hybrid nodes with GPU accelerators. With the 48 ports of the BXI V2 switch, large systems can be built with fewer elements, thus optimizing cost. With fewer hops in the data path, the communication latencies are reduced and congestion points are limited.

BXI V2 QoS allows for 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 prevent communication bottlenecks. End-to-end error checking and link level retry together with Error Correction Code (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 V2 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.

The BXI V2 network comes with a complete out-of-band software management stack. It provides all functions necessary for monitoring, routing and performance analysis. The runtime environment supports optimized MPI libraries and PGAS oriented languages for HPC applications. For global storage, BXI V2 features a native implementation of the Lustre parallel filesystem.

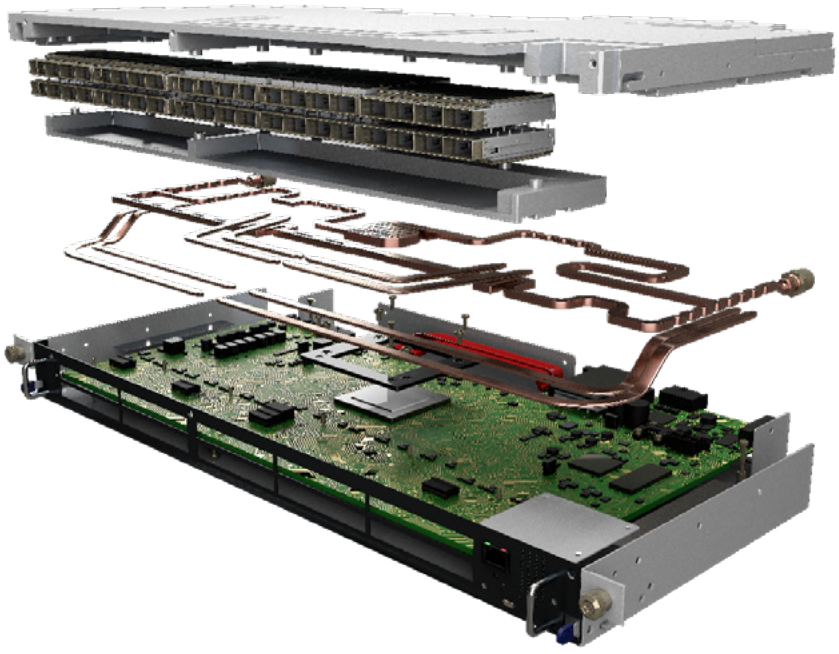


Figure 1: BXI V2 Switch for BullSequana XH2000

BXI V2 Network Interface Controller (NIC)

The BXI V2 NIC interfaces with the system node through a 16x PCI Express gen3 interface and with the BXI V2 fabric through a 100Gb/s BXI V2 port. The BXI V2 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 BullSequana XH2000 platform.

The BXI V2 V2 NIC provides dedicated communication FIFOs, DMA engines for sending and receiving, and matching engines for reception. IT also features rank (e.g. MPI, SHMEM, etc.) to physical address location as well as virtual to physical address translation with no need for memory pinning in the OS kernel. 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 V2 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 V2 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.

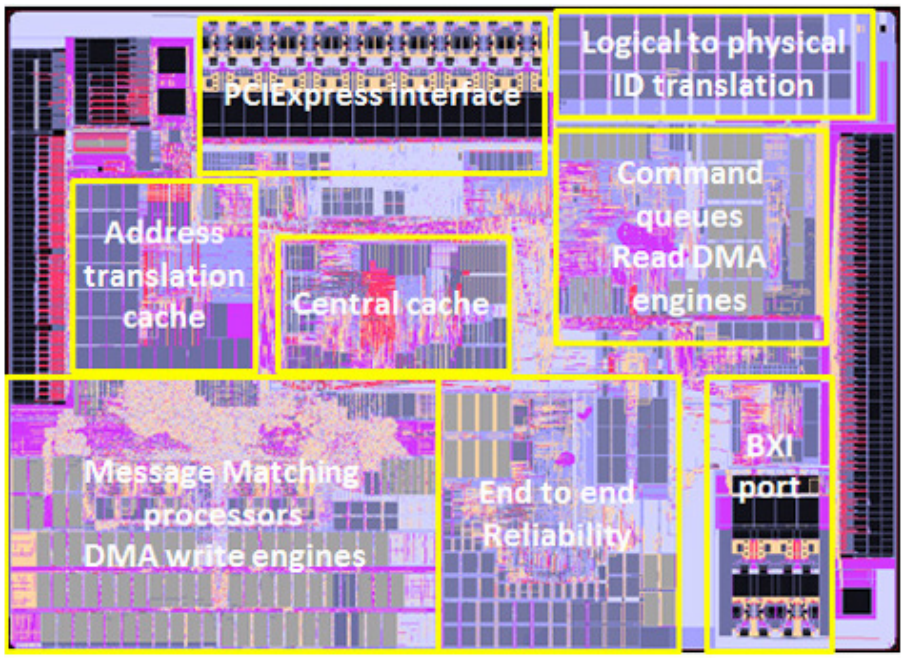


Figure 2: BXI V2 V2 NIC floor plan

OS Bypass capabilities

The BXI V2 NIC provides OS bypass capabilities. Sending or receiving a message is controlled by the NIC without Interruptions 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. OS Bypass ensures minimal CPU usage when posting a Portals command to improve the issue rate and latency. The BXI V2 NIC translates virtual addresses into physical addresses for incoming network requests, local reads and writes. The NIC translation unit sustains high message rates and 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.

BXI V2 provides full hardware offload of communications, enabling CPUs to be fully dedicated to computational tasks while communications are independently managed by BXI V2

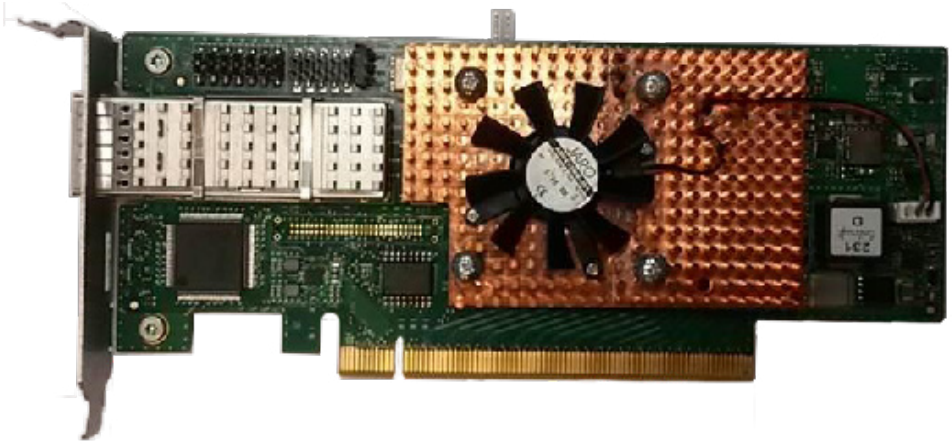
Communications offload in hardware

BXI V2's main feature 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 V2, 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, it requires to promptly match an incoming message with a MPI_Recv. With BXI V2, 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 Portals events.
- **Atomic units:** In addition to traditional remote read and writes, the BXI V2 NIC can also perform atomic operations, (conditional) swaps, and fetch/atomics. Such operations make 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 V2 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, allowing for 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 V2 fabric is designed to provide 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 a 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 permanent faults 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 V2 Switch ASIC

The BXI V2 switch ASIC is a low latency, non-blocking 48 ports crossbar. 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 V2 switch ASIC is directly embedded in the L1-L2 switch modules of the BullSequana XH2000 platform which targets large configurations. It is also packaged as an external switch with 48 QSFP connectors.

As seen in fig. 3, the die size is dimensioned by the number of SerDes which occupy the whole chip perimeter.

The high-radix BXI V2 switch simplifies the architecture of large scale systems since fewer components are required. For a given system size, this radix also helps reducing the network diameter and with fewer hops in the communication path, it shortens the latencies.

The BXI V2 switch features per-port destination-based routing, which allows for fabric flexibility and efficient adaptive routing. A default path is defined for each destination node in a full 64k entries routing table. In addition, the routing table 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, due to end-to-end protocol implemented in the BXI V2 NIC.

The BXI V2 switch is managed completely out-of-band through a microcontroller.

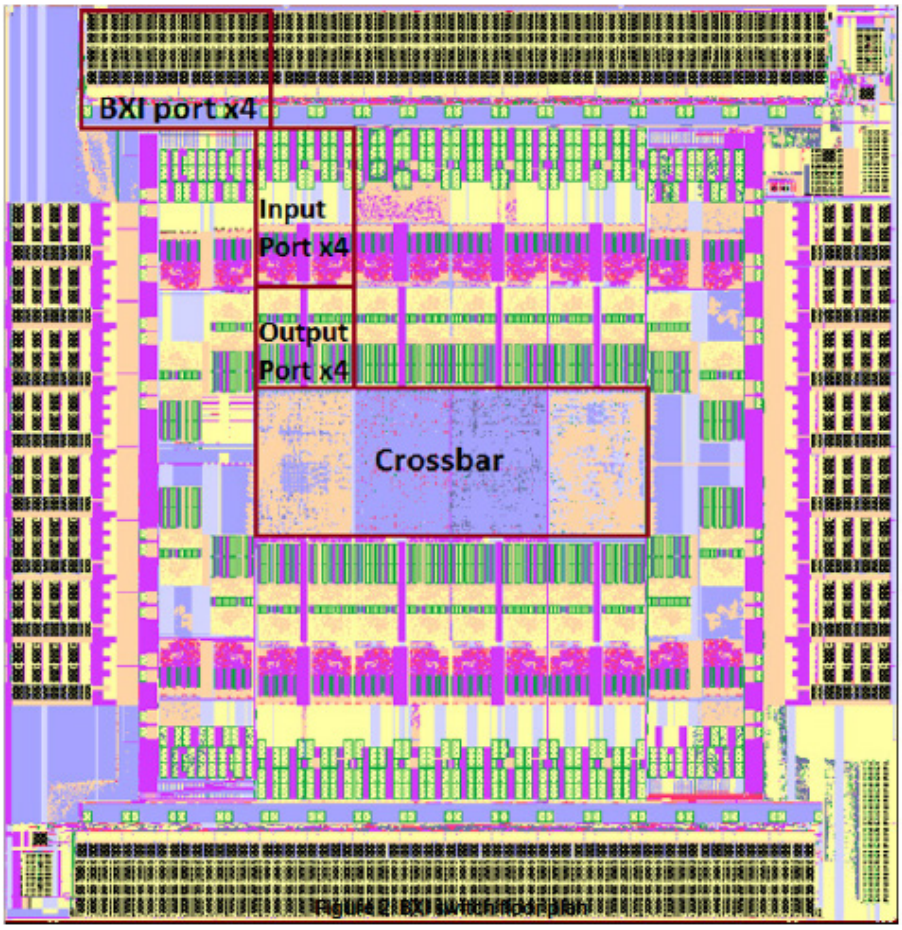


Figure 3: BXI V2 Switch floor plan

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 and 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 V2 switch, there is no restriction on the type of topology which can be implemented using the BXI V2 fabric. Fattrees can be implemented easily. In this case, all the VCs can be used for QoS 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 due to the switches' high radix, the availability of a high number of virtual channels and the per-port routing tables.

BXI V2 management software

The BXI V2 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 V2 traffic. It also guarantees that the BXI V2 components are accessible at all times even in the event of a partial or complete network failure. In BXI V2, Advanced Fabric Management (AFM) is responsible for monitoring and routing. IBMS handles Device Management and Performance Analysis tasks. Moreover, the solution is compliant with standard device management software due to an SNMP agent embedded in the switch firmware.

AFM

AFM reacts to errors and events emitted by the components, and are thus able to detect complex failure patterns and report synthetic information to the administrators. By querying the equipment on a regular basis, the BXI V2 monitoring modules detect both errors and unexpected situations. The BXI V2 fabric management can also be hierarchically organized, both to scale up to the largest configurations and to provide systemwide event correlations.

AFM 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 collected 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 providing the expected bandwidth to applications. As re-routing reaction times are critical, when a failure that would trigger a re-routing operation is detected, a highly performant mechanism is launched to compute new routes and to reroute the traffic around the failing component within an upper-bound maximum of 15 seconds. This mechanism is intended to react quickly and preserve a functional fabric.

IBMS

IBMS is in charge of Device Management and Performance Analysis that takes advantage of the error and performance counters available in the BXI V2 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 the switch level, according to rules that can be dynamically changed during production, depending on troubleshooting or profiling needs. Independent sampling rules can be defined for each switch. A sampling rule consists of a set of counters to sample and a sampling frequency. These measurements sampled directly at a high frequency level (10Hz by default) complement HPC applications profiling by providing fabric side information.

BXI V2 AFM and IBMS software are fully compliant within Supercomputer suite, a scalable, and robust software suite that meets the requirements including the need for enhanced security of even the most challenging high performance computing environments.

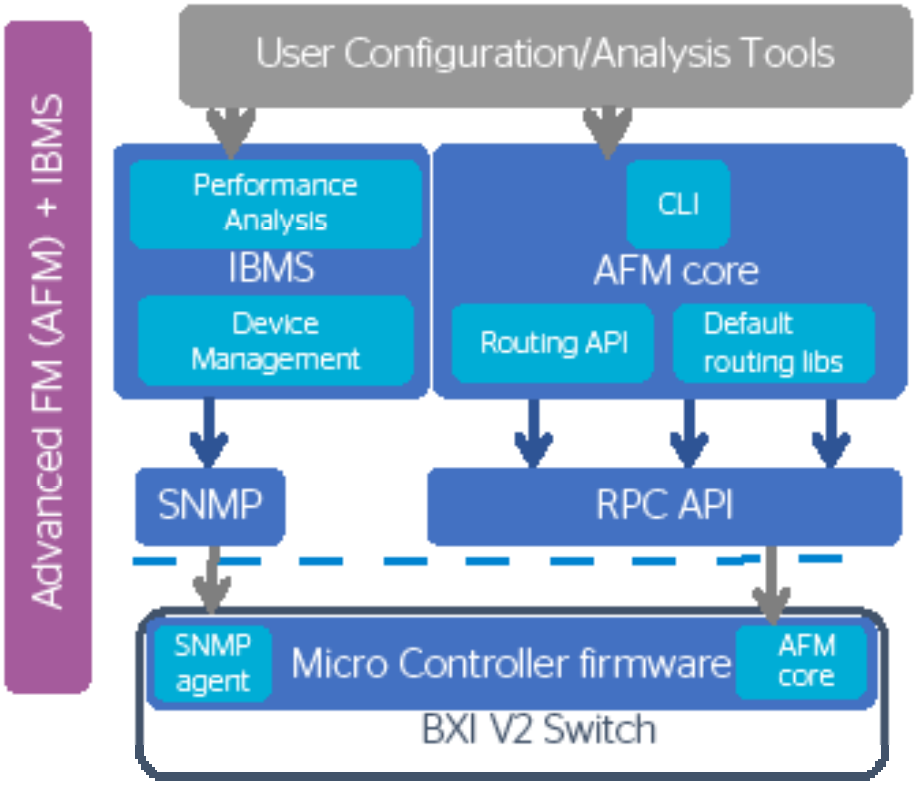


Figure 4: BXI V2 management software

BXI V2 application environment

The BXI V2 application environment, or Compute Stack, represented in figure 5 provides the software layer to interact directly with the BXI V2 network through native Portals4 interfaces using MPI, or PGAS communication libraries. All components are implemented using the native Portals 4 API.

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 PTLnet (IP over Portals) component makes it possible to have large-scale, efficient, and robust IP communication.

Communication components interact with the BXI V2 Kernel driver to gain direct access to the BXI V2 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 V2 NIC through a virtual memory mapping of the PCI address space, the total communication overhead from the application call to the hardware access is minimized.

Atos Open MPI offers a direct implementation of the MPI-1 two-sided (MPI_Send/MPI_Recv) and MPI2/3 RMA (one-sided) operations through the Portals 4 API, using the Put, Get, Atomic, and FetchAtomic primitives. Some synchronous (MPI-1) and asynchronous (MPI-3) triggered collective operations are already available thanks to Portals API and BXI V2 NIC. Using a chain of triggered operations, collective operations are offloaded to the BXI V2 NIC and allow a calculation resource saving.

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 separate dedicated paths within the NIC and switches. This separation ensures mutual protection between those two types of traffic, avoiding 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.

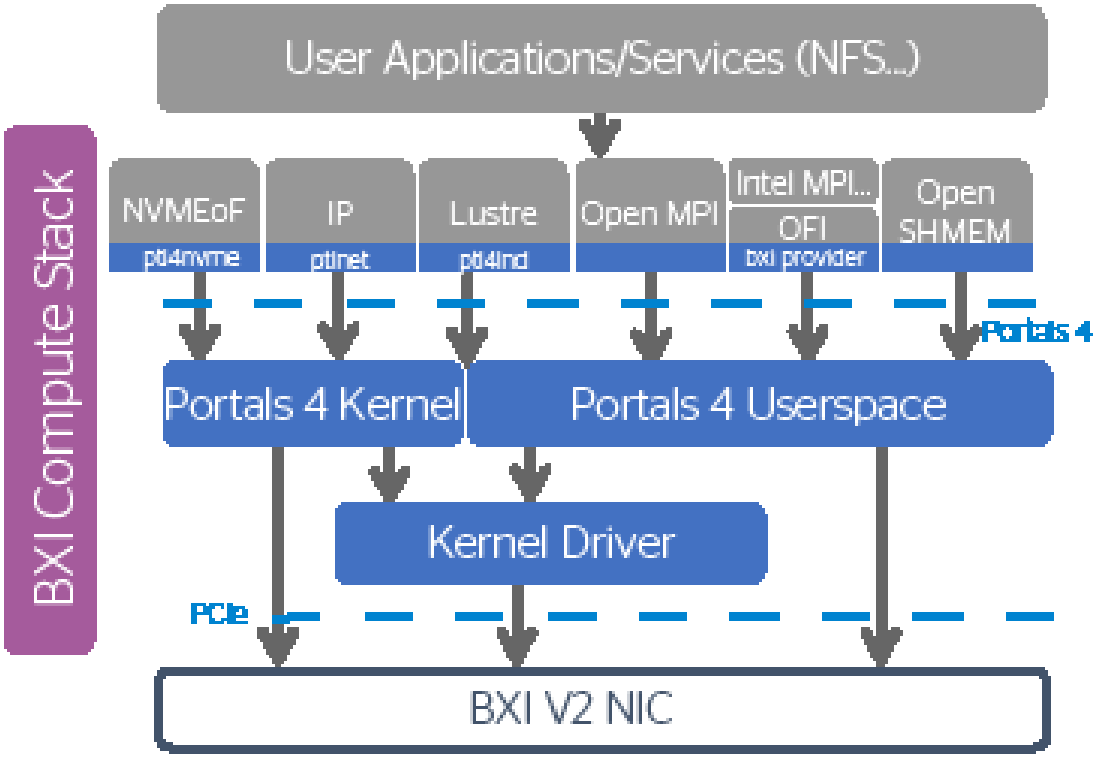


Figure 5: BXI V2 application environment

BXI V2 in the BullSequana XH2000 platform

In 2019 Atos introduced the versatile BullSequana XH2000 platform. With this platform, power efficient and densely packed cabinets can be assembled in different manners, depending on fabric size and cost constraints. A "cell" is composed of 1 cabinet with up to 96 nodes.

The BullSequana XH2000 eXascale platform can host a variety of compute nodes based on general purpose CPUs or GPUs. In addition to BXI V2, BullSequana XH2000 also supports other types of high performance networks such as InfiniBand HDR. It is cooled with the BullSequana Direct Liquid Cooling (DLC) solution, which can remove the heat generated in such a dense configuration with little overhead. To improve system reliability, BullSequana XH2000 supports redundant power supply units, redundant hydraulic heat exchangers, and two redundant management nodes in each cell.

The Interconnect Network is composed of the Interconnect Network switches and a uniquely designed network connection mid-plane.

Interconnect Network switches

Up to 10 direct liquid cooled Interconnect Network switches located at the top rear of the cabinet.

Network connection midplane

It is located at the center of the cabinet. It brings 3 major benefits:

- Flexibility to customize routing of the compute blades to the interconnect network switches,
- Possibility to mix different interconnect network speeds and/or technologies,
- Selection of the optimized interconnect network topology (Fat Tree, DragonFly+).

Topologies

BullSequana XH2000 supports two topologies:

- Fat Tree (Full and pruned): a proven network architecture that provides very good worst-case blocking,
- DragonFly+.

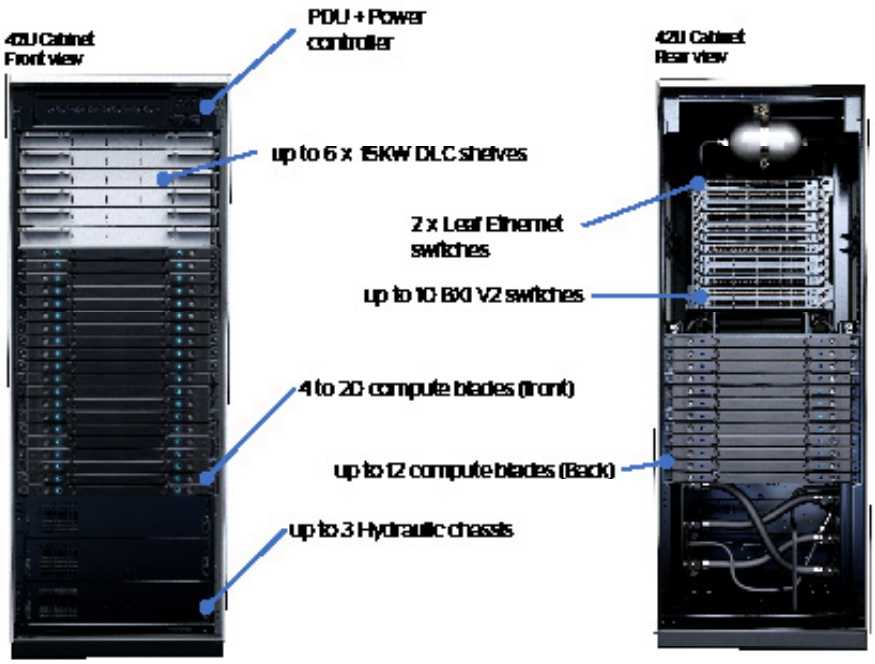
A hybrid computing solution

BullSequana XH2000 combines a broad variety of CPUs, accelerators, and high-speed interconnect networks to run mixed HPC workloads or dedicate a full GPU and High-speed Ethernet-based system to Deep Learning.

Highly flexible

Modular and scalable

- Combine different types of current and future computing technologies (GPUs, CPUs) and interconnect networks (InfiniBand HDR, HDR100, BXI V2, High-speed Ethernet) within one system to create a supercomputer that matches your needs perfectly,
- Select the desired network topology (Fat Tree or DragonFly+) and pruning ratio,
- BullSequana XH2000 scales from one-rack, up to exaflopic systems.



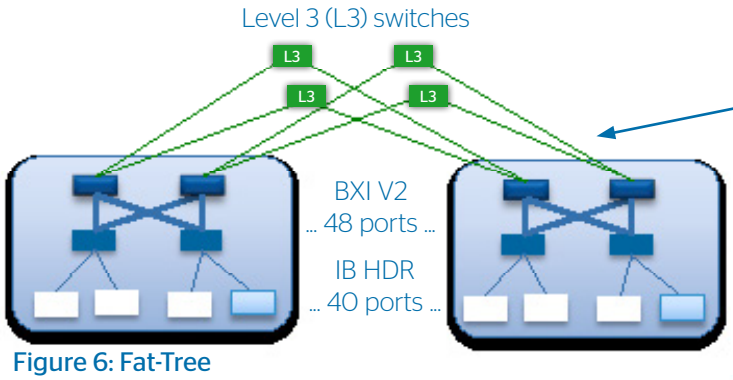


Figure 6: Fat-Tree
Medium size systems
3 levels = 13,8k nodes
4 levels= 64k nodes

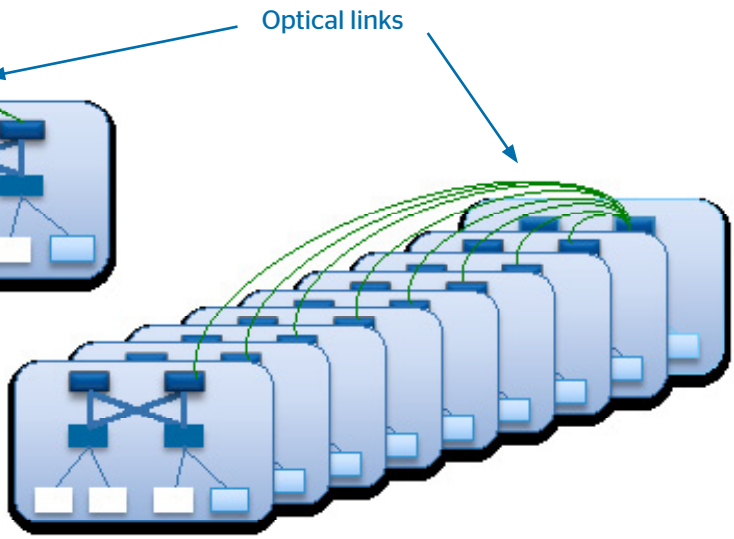


Figure 7: DragonFly+
Large size systems

Conclusion

The BXI V2 network provides the highest level of communication performance for HPC applications whether looking at bandwidth, latencies, or message rates. By offloading to the hardware components, BXI V2 communications do not interfere with computations. With a native interface to Portals4, MPI libraries are fully optimized for the complete range of communication primitives. Similarly, BXI V2 enables efficient implementations for SHMEM or PGAS environments. Finally BXI V2 offers a complete set of RAS features. As a result, BXI V2 allows HPC applications to scale to millions of threads.

Brochure

About Atos

Atos is a global leader in digital transformation with 110,000 employees in 73 countries and annual revenue of € 12 billion. European number one in Cloud, Cybersecurity and High-Performance Computing, the Group provides end-to-end Orchestrated Hybrid Cloud, Big Data, Business Applications and Digital Workplace solutions. The Group is the Worldwide Information Technology Partner for the Olympic & Paralympic Games and operates under the brands Atos, Atos|Syntel, and Unify. Atos is a SE (Societas Europaea), listed on the CAC40 Paris stock index.

The purpose of Atos is to help design the future of the information space. Its expertise and services support the development of knowledge, education and research in a multicultural approach and contribute to the development of scientific and technological excellence. Across the world, the Group enables its customers and employees, and members of societies at large to live, work and develop sustainably, in a safe and secure information space.

Find out more about us
atos.net
atos.net/careers

Let's start a discussion together

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

CT-200522-SKR-BR-BXIV2