The BXI interconnect architecture

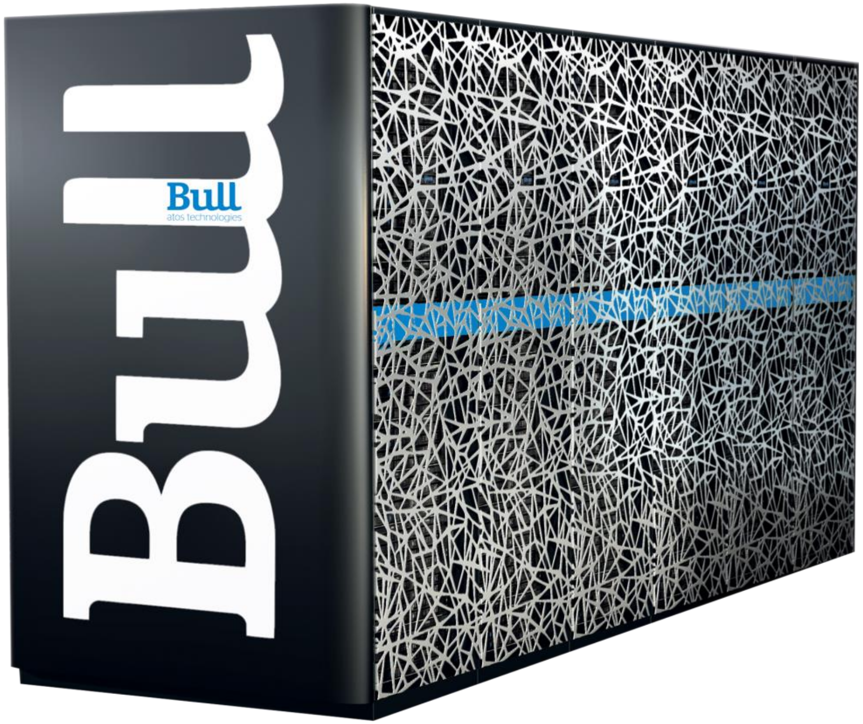
Saïd Derradji, Thibaut Palfer-Sollier, Jean-Pierre Panziera, François Wellenreiter

26/08/2015

agenda

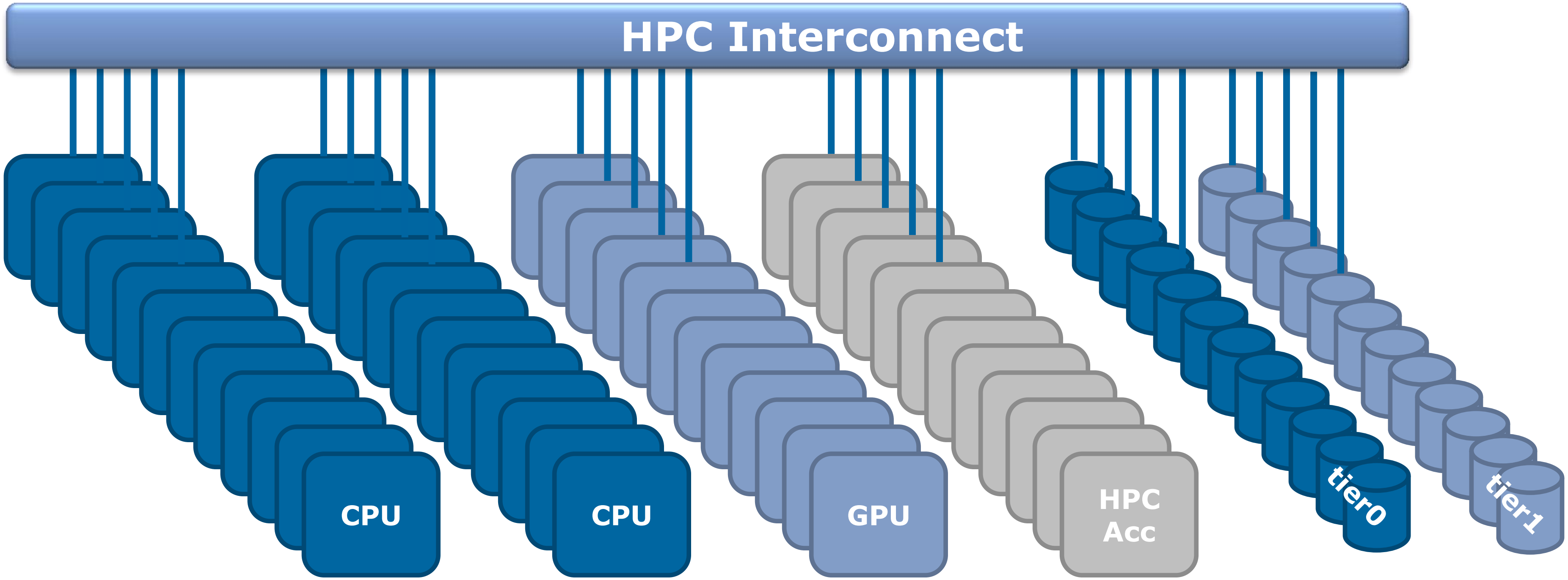
- Bull : Atos technology
- BXI overview
- BXI fabric
- BXI HW offloads
- BXI platforms
- BXI Performances
- Summary

Bull … Atos technologies



- Bull incorporated into Atos (2014)
- Atos WW IT company:
  - 86,000 employees
  - 66 countries
  - 9B € revenue (2014)
- HPC products keep the “Bull” branding
- Multiple Pflops HPC systems installed around the world
  - Europe: France, Germany, Netherlands, UK
  - Japan, Brazil
- Full spectrum expertise SW + HW:
  - system packaging, motherboards, ASICs

Interconnect is the backbone of HPC systems



1000s-10,000 compute nodes
CPUs, GPUs, HPC accelerators

Multiple
storage tiers

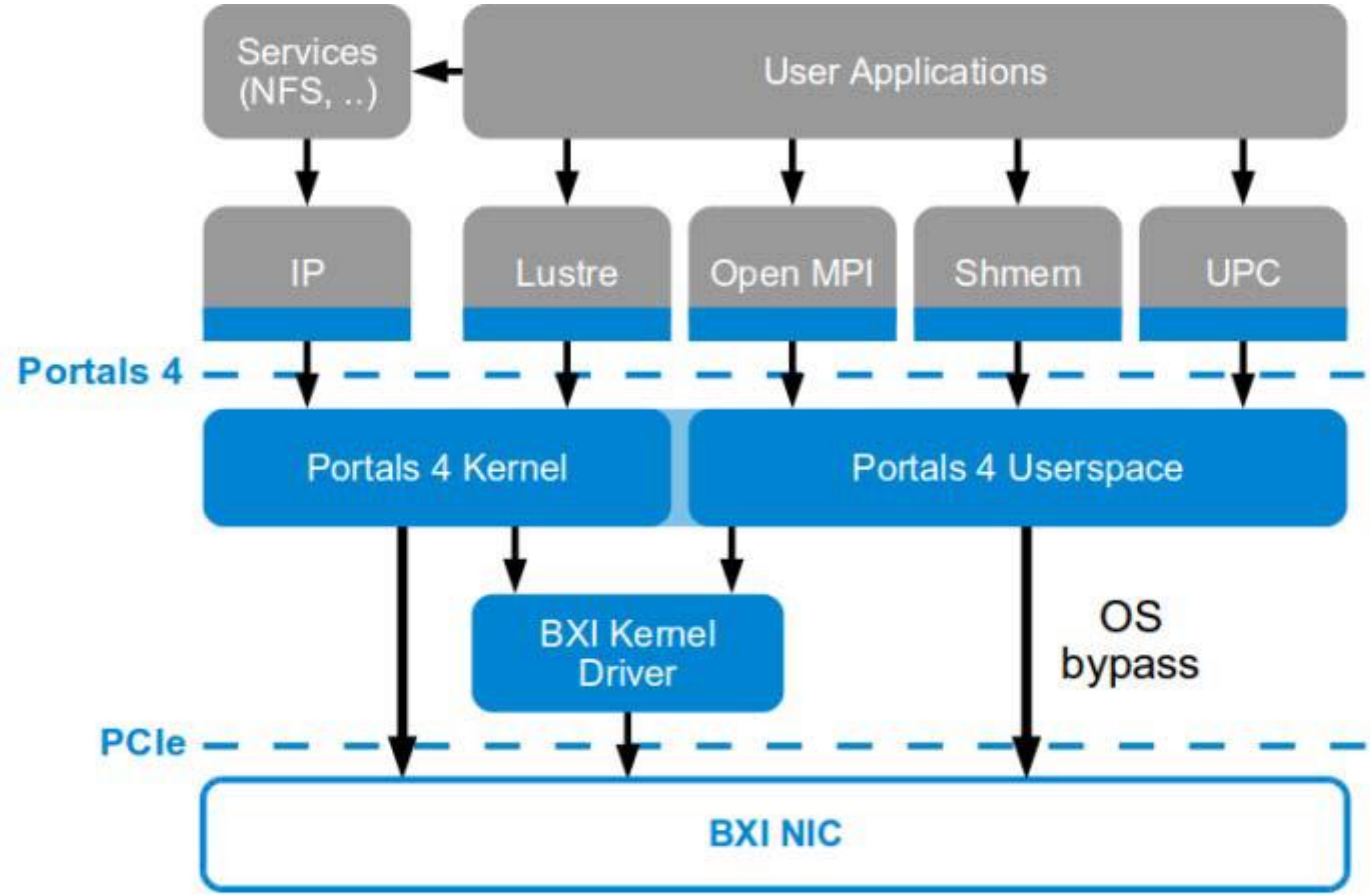
BXI Interconnect overview

- **BXI 1st generation of Bull Exascale Interconnect**
  - HW acceleration →sustained performance under heavy load
  - High Bandwidth, Low latency, High message rate at scale

- **BXI full acceleration in hardware for HPC applications**
  - based on Portals 4 a rich low level network API for message passing
  - HW support for:
    - MPI and PGAS communications over Portal 4 (send/recv, RDMA)
    - High performance collective operations

- **BXI highly scalable, efficient and reliable**
  - Exascale scalability → 64k nodes
  - Adaptive Routing
  - Quality of Service (QoS)
  - End-to-end error checking + link level CRC + ASIC ECC

BXI Software compute stack



BXI fabric

- A BXI port consists of 4 differential lanes signaling rate up to 25,278125 GT/s per direction -> an aggregate bandwidth of 100Gb/s
  - Width reduction to 3, 2 or 1 lane
  - Lane reversal and polarity reversal
  - Half signaling rate
  - Encoding 64b66b standard IEEE std 802.3

- Messages are composed of 32B flits, variable length. They can be up to 4GB.
  - A tail flit is used to recognize the end of a message
  - Network encapsulation: split into packets of 72B
    - 64 bytes payload = 2 independent flits of 32B
    - 8 bytes overhead added for link level reliability and control

BXI fabric features

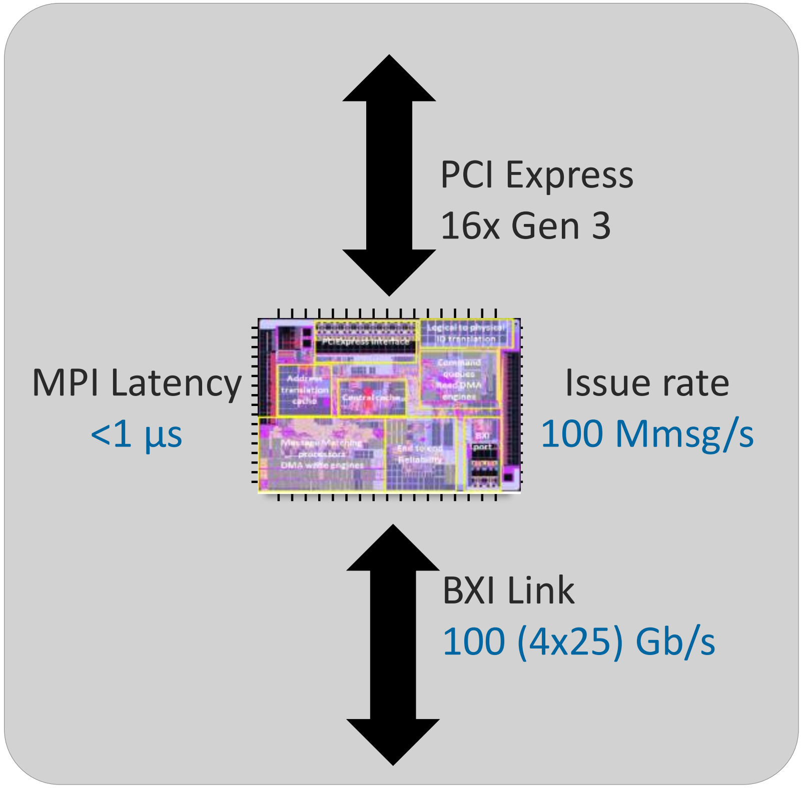
- Scalable up to 64K NICs
- Reliable and ordered network (end to end + Link level)
- Flexible with full routing table
  - Many topologies supported (fat tree, torus, flattened butterfly …)
  - Ease routing algorithm optimization
- Adaptive routing

- Extensive buffering implementing 16 virtual channels preventing deadlock and efficiently balancing traffic
- Quality of service with weighted round robin arbitration
  - highly configurable load balancing
  - Segregation of flows per destination
  - ensuring progress of short messages vs long messages

- High resolution time synchronization
- Out of band management

BXI Network is based on 2 ASICs

NIC ASIC


PCIExpress interface
Logical to physical ID translation
Address translation cache
Central cache
Command queues Read DMA engines
Message Matching processors DMA write engines
End to end Reliability
BXI port

PCI Express
16x Gen 3

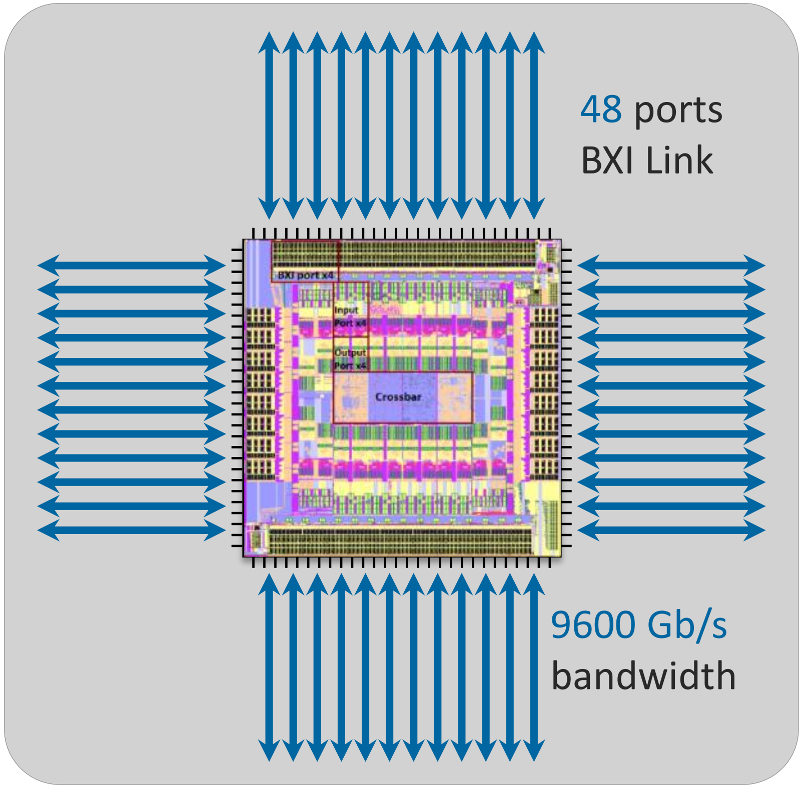
MPI Latency
<1 µs

Issue rate
100 Mmsg/s

BXI Link
100 (4x25) Gb/s

Lutetia

switch ASIC



48 ports
BXI Link

9600 Gb/s
bandwidth

Divio

NIC main features 1/2

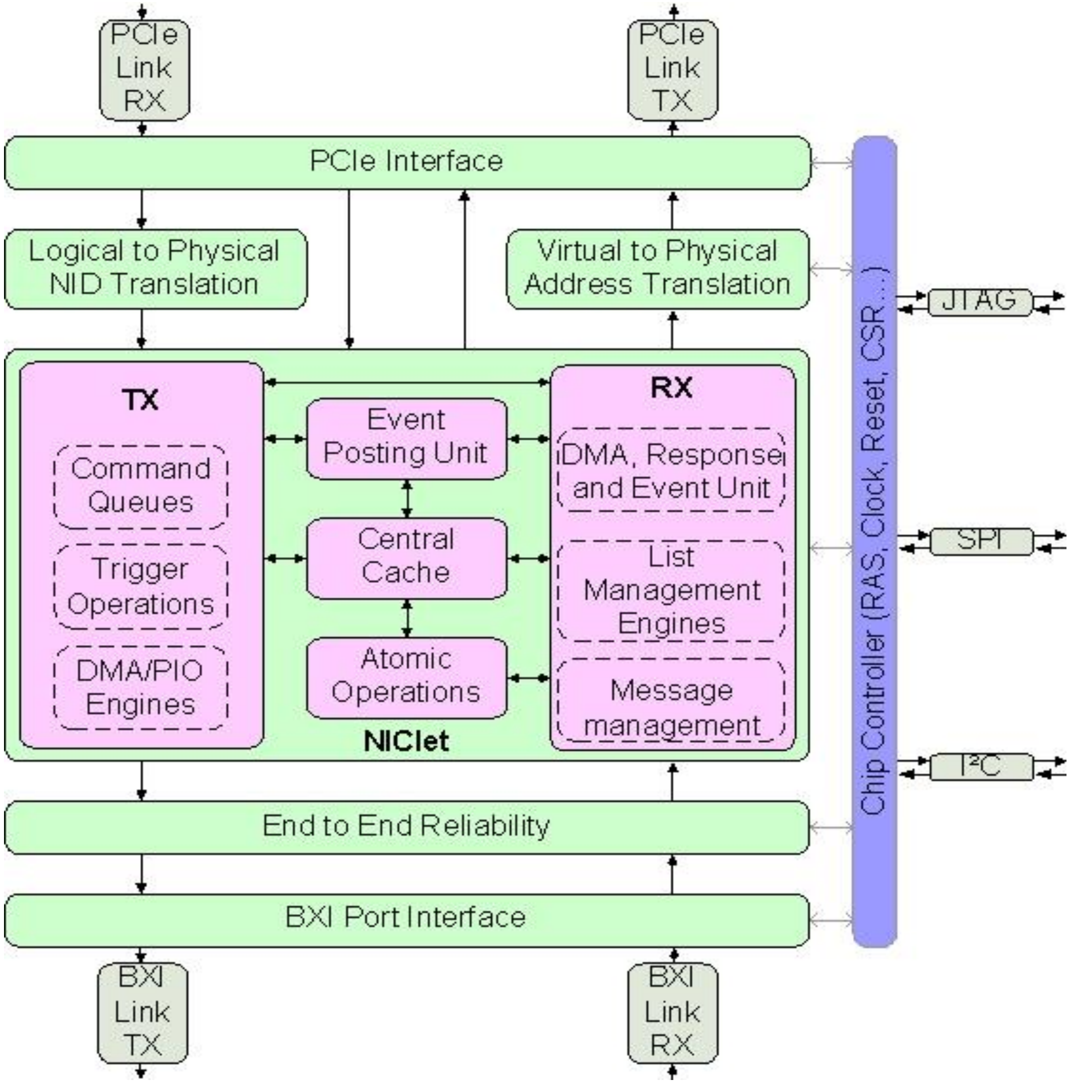
- **Implements in hardware the Portals 4 communication primitive**
  - Overlapping communications and computations by offloading to NIC
  - MPI two-sided messaging:
    - HW acceleration of list management and matching on the NIC
  - PGAS / MPI one-sided messaging:
    - use fast path inside the NIC

- **OS and application bypass**
  - Applications issue commands directly to the NIC, avoiding kernel calls
  - Reception controlled by NIC without OS involvement
  - Reply to a put or a get does not require activity on application side.
    - Logical to physical ID translation
    - Virtual to physical memory address translation.
    - Rendez-vous protocol in HW

NIC main features 2/2

- **Collective Operations offload in HW**
  - using Atomic and Triggered operations units
- **End-to-End reliability** recovery mechanism for transient and permanent failures
  - message integrity, 32bits CRC are added to each message (or each message chunk for large transfers).
  - message ordering required for MPI messages is checked with a 16 bit sequence number.
  - message delivery a go-back-N protocol is used to retransmit lost or corrupted messages.
- **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

NIC bloc diagram



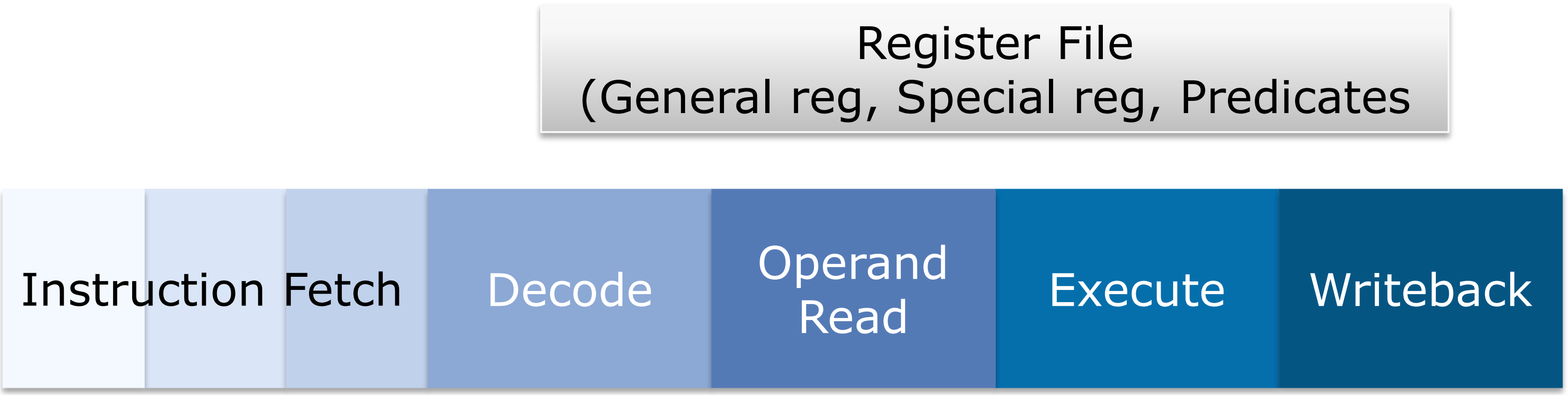
Bull
atos technologies

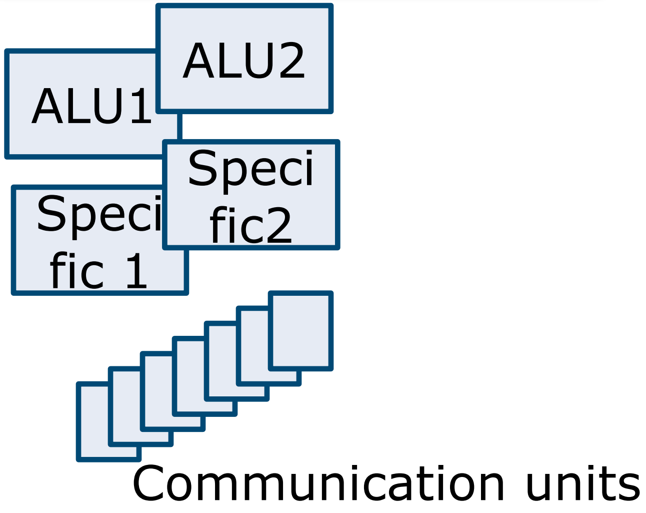
List Management Engine (LME)

Custom Application Specific Instruction Set Processor

- Portals4 list management offloaded to NIC.
- Implementing these functions in HW is fairly complex and not flexible.
- Custom ASIP called LME to handle processing of all Portals list related aspects:
  - Allocating and freeing of list resources, building and modifying lists.
  - Traversing lists to implement matching functions.
- multicore implementation
- Optimizing MPI two-sided communications
  - (MPI_Isend, MPI_Irecv, asynchronous)

LME diagram

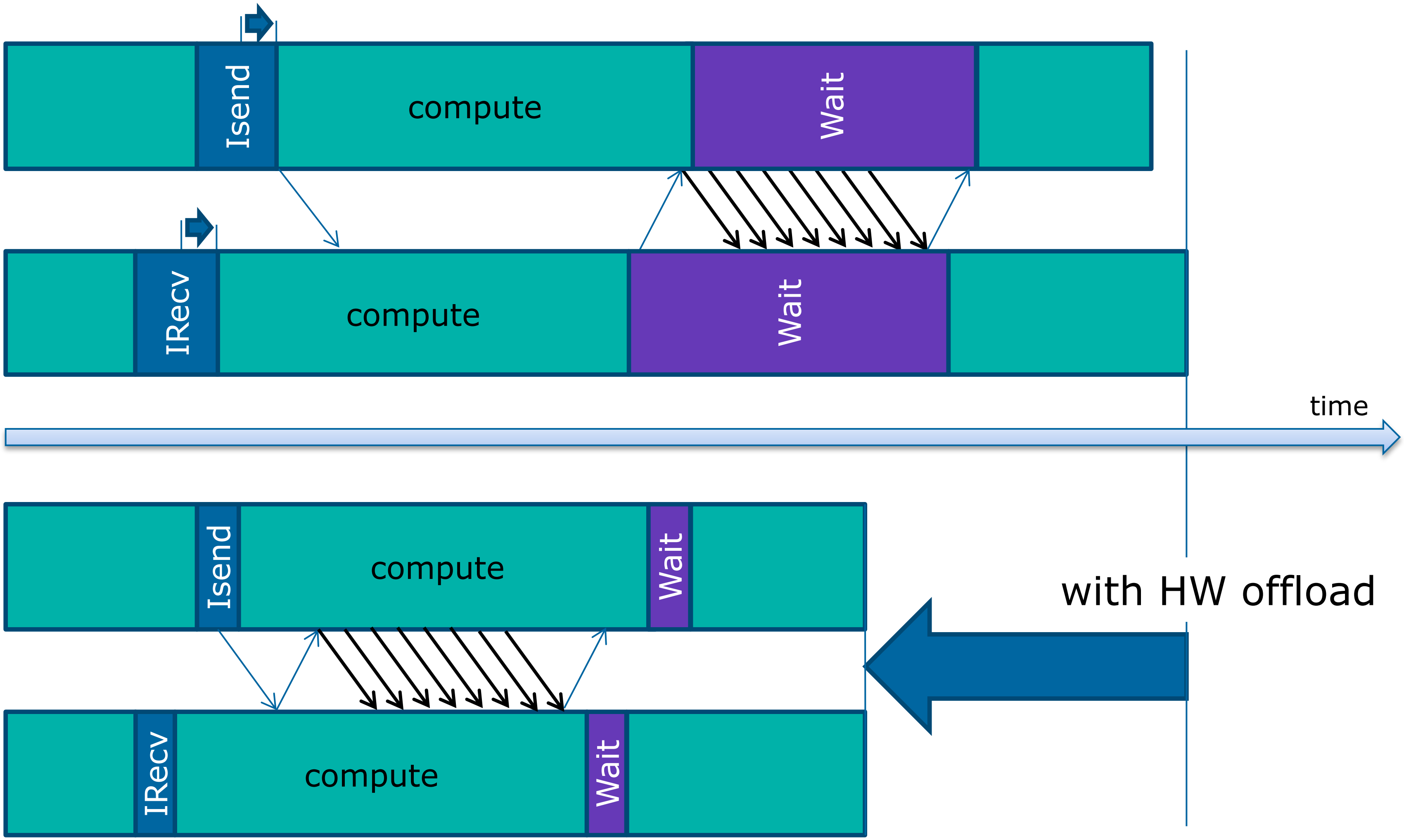




LME architecture:

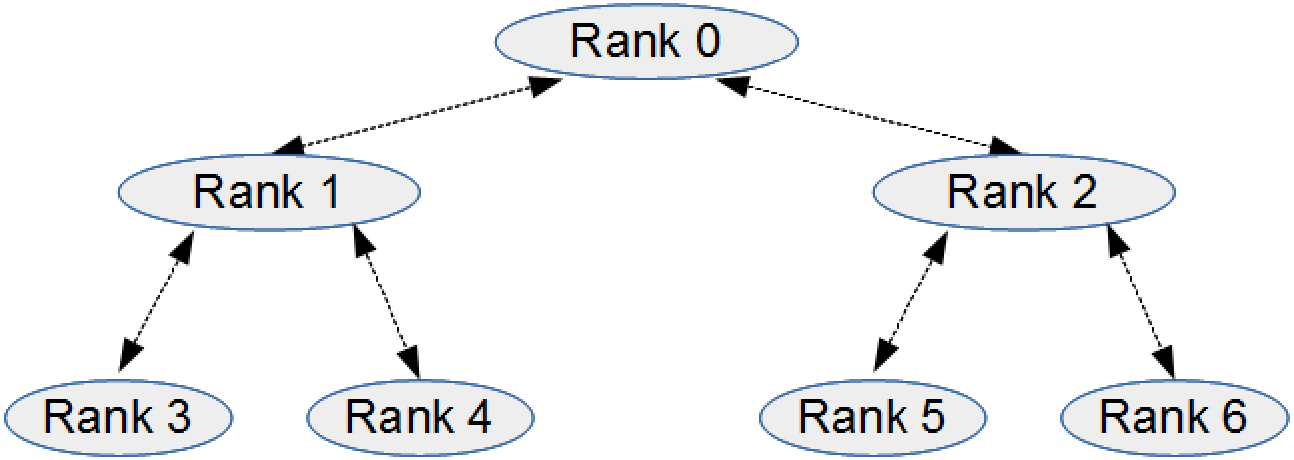
64b VLIW, with 2 operations per instruction.
1 cycle per instruction.
32 General Processing Registers, 64b wide.
32 Special registers
32 ALU operations. 2 units.
16 specific operations. 2 units.
An operation replaces a sequence of ALU op.
35 operations to send commands to interface modules
Pipe of 7 stages.

BXI: offloading MPI communication in HW

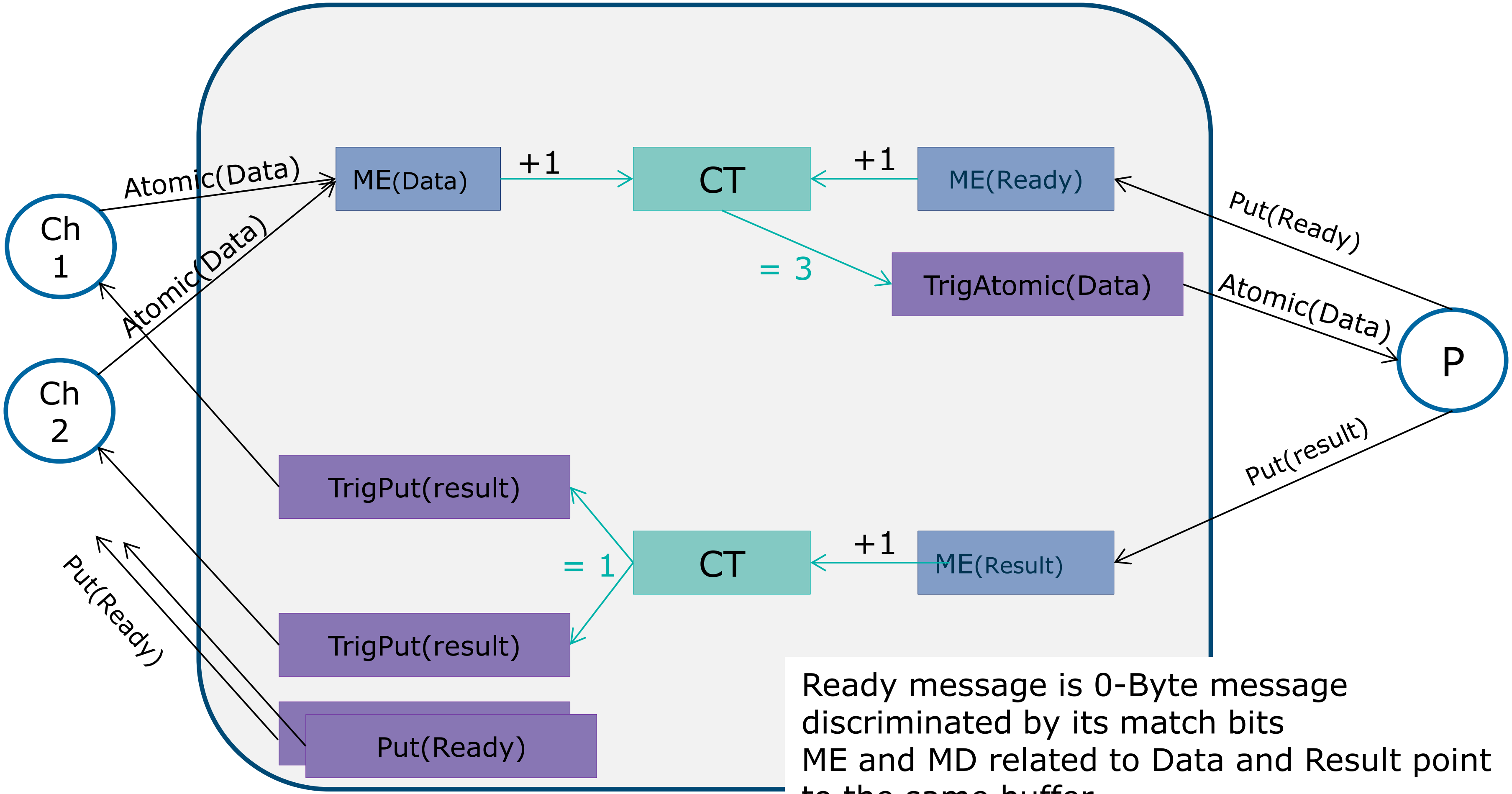


with HW offload

BXI: Offloading collective operations



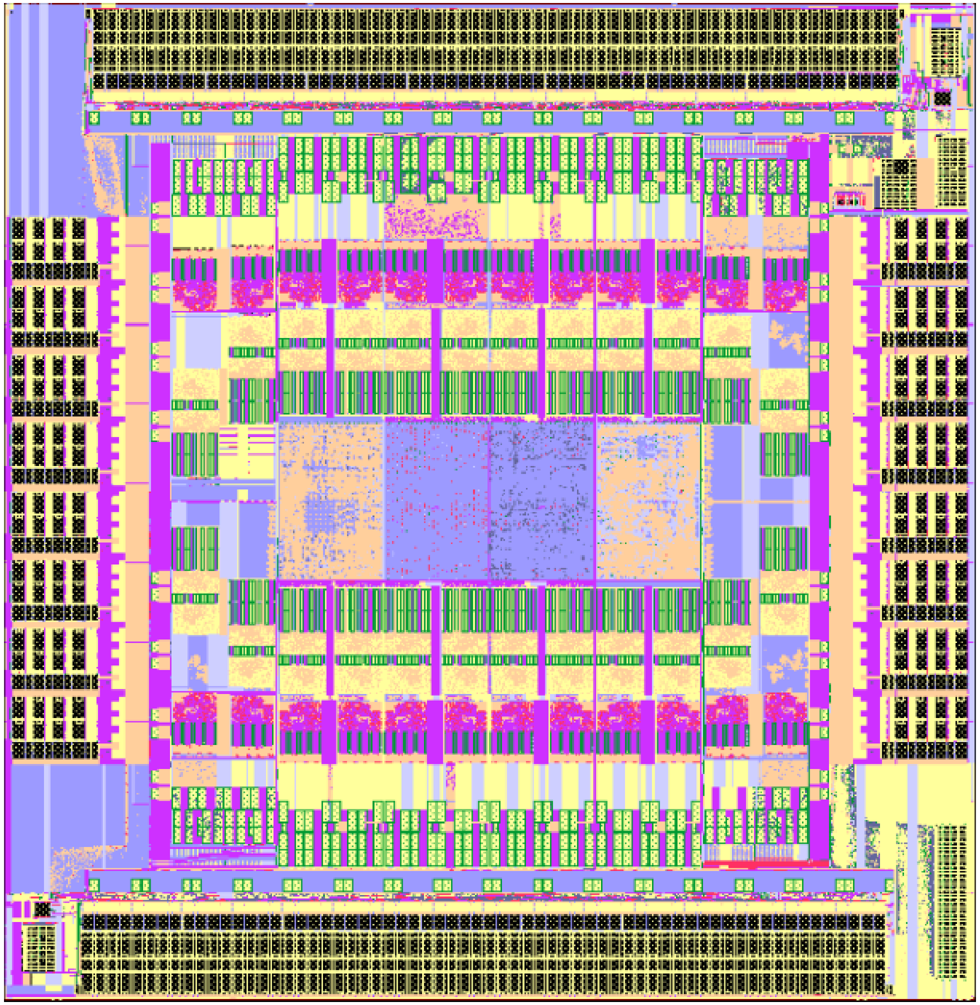
MPI IAllreduce implementation using Triggered and Atomic operations



Ready message is 0-Byte message discriminated by its match bits

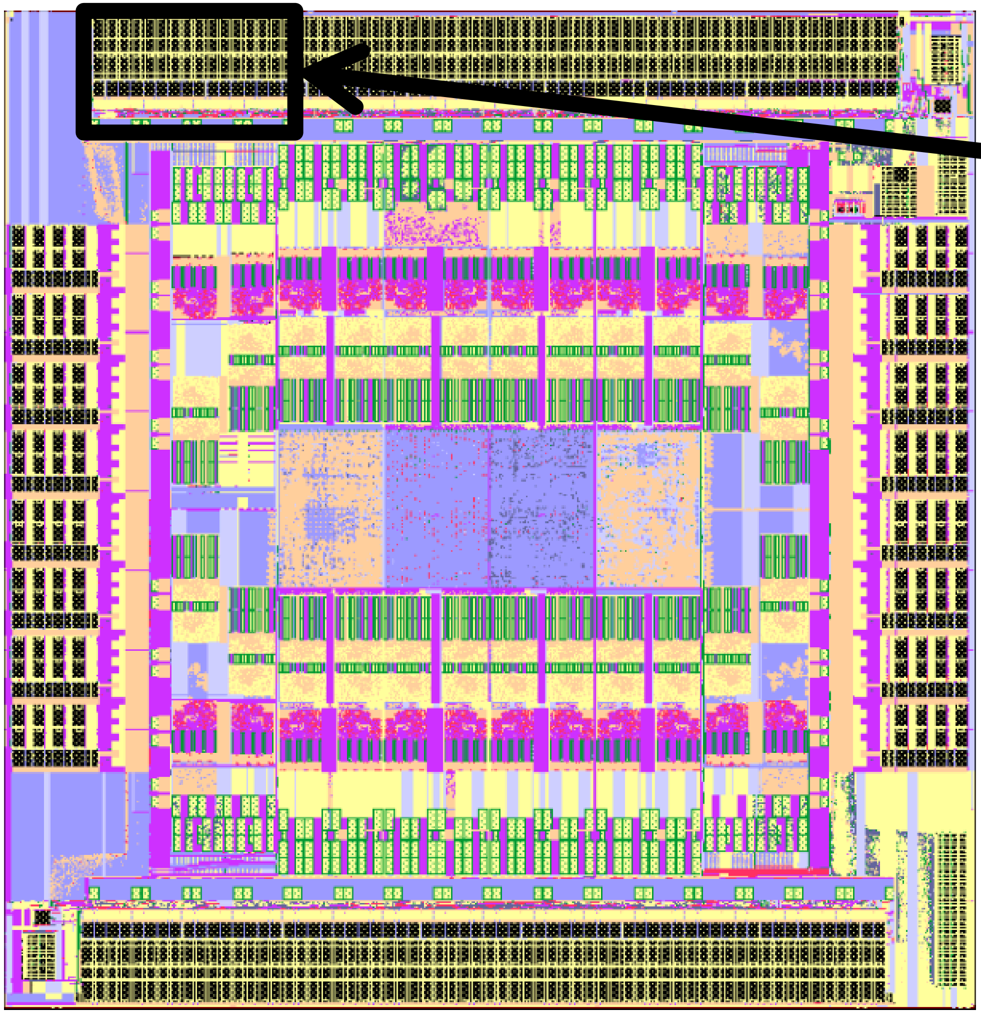
ME and MD related to Data and Result point to the same buffer

BXI Switch overview

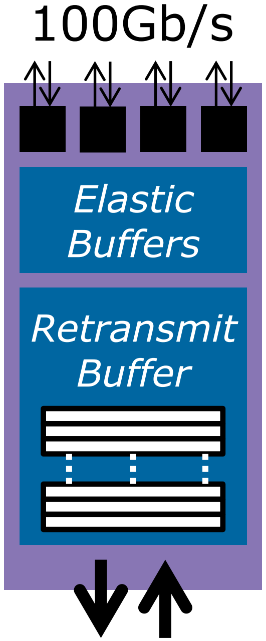


- 48 ports, 192 SerDes @ 25Gb/s
  - Total throughput : 9600 Gb/s
- Latency : 130ns
- Die : 22 x 23mm
- Package : 57.5 x 57.5mm
- Transistors : 5.5 billions
- TDP : 160W
  - Min power : 60W
- Techno : TSMC 28nm HPM

BXI Switch overview

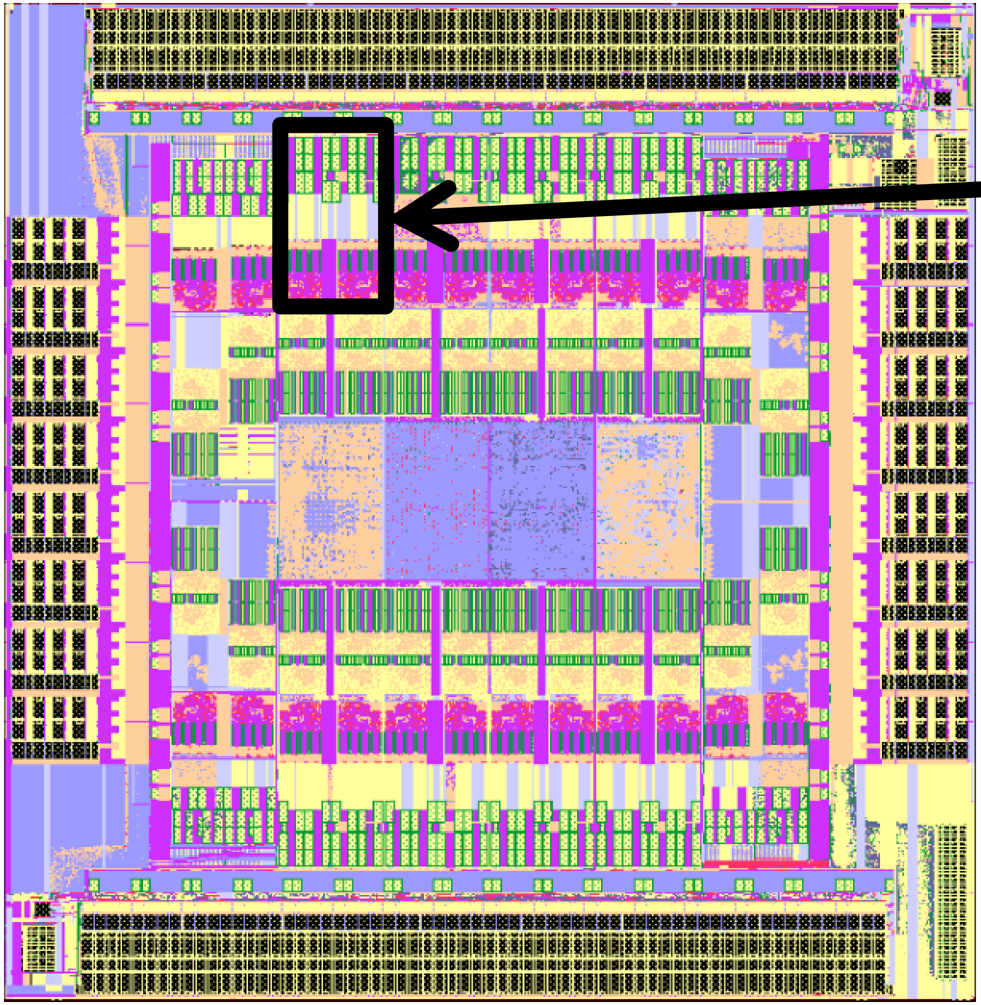


4 BXI ports
PHY + Link

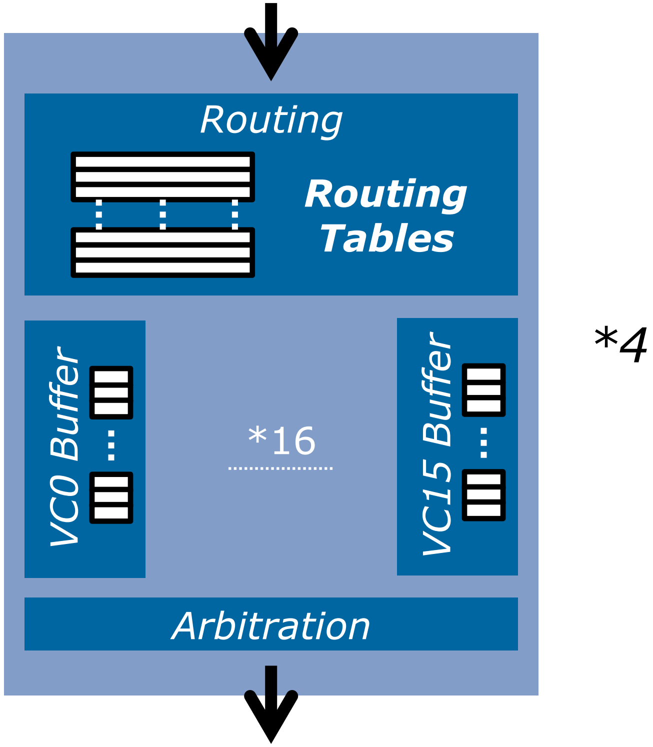


*4

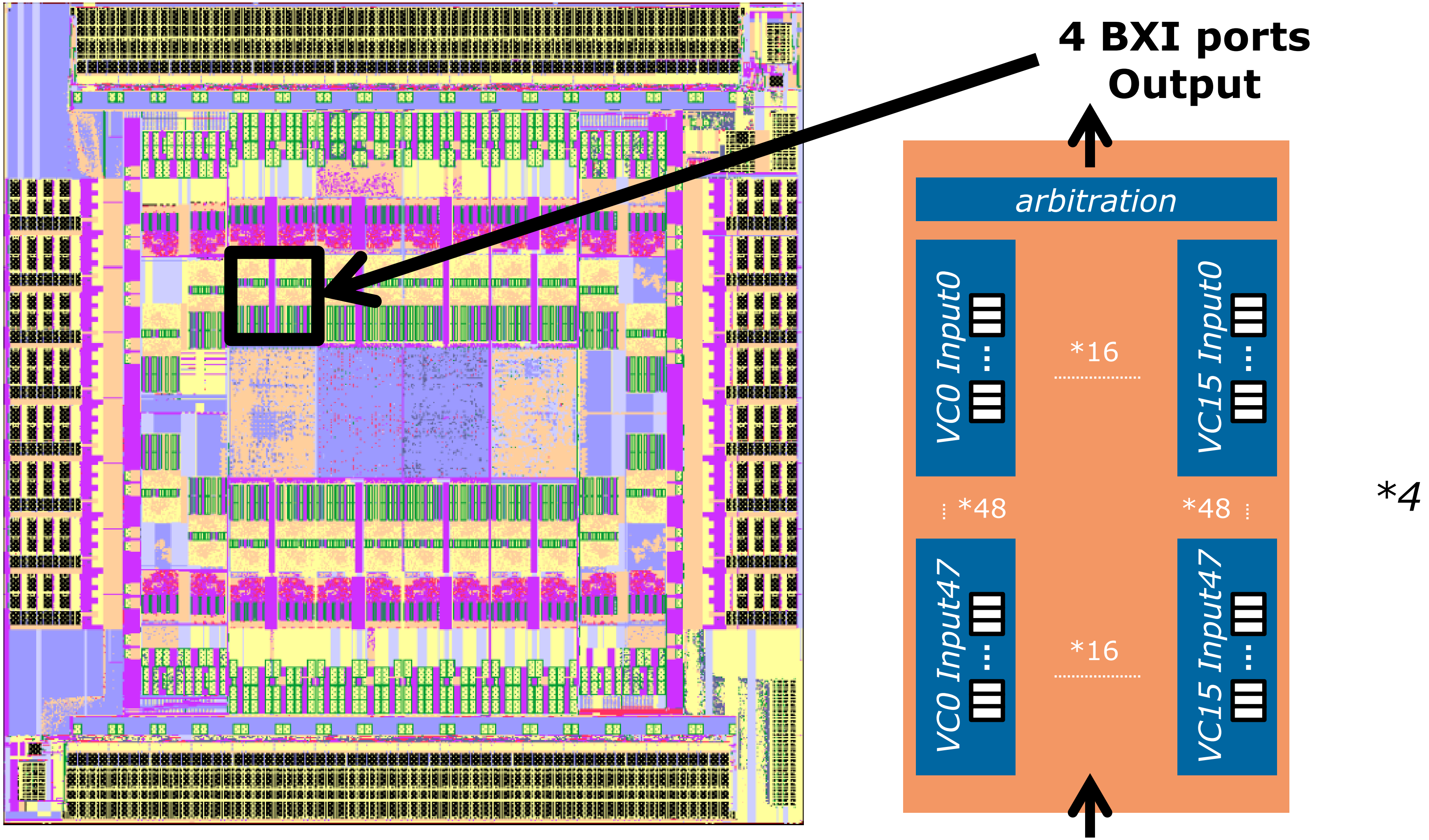
BXI Switch overview



4 BXI ports Input



BXI Switch overview



4 BXI ports Output

arbitration

VC0 Input0

*16

VC15 Input0

*48

*48

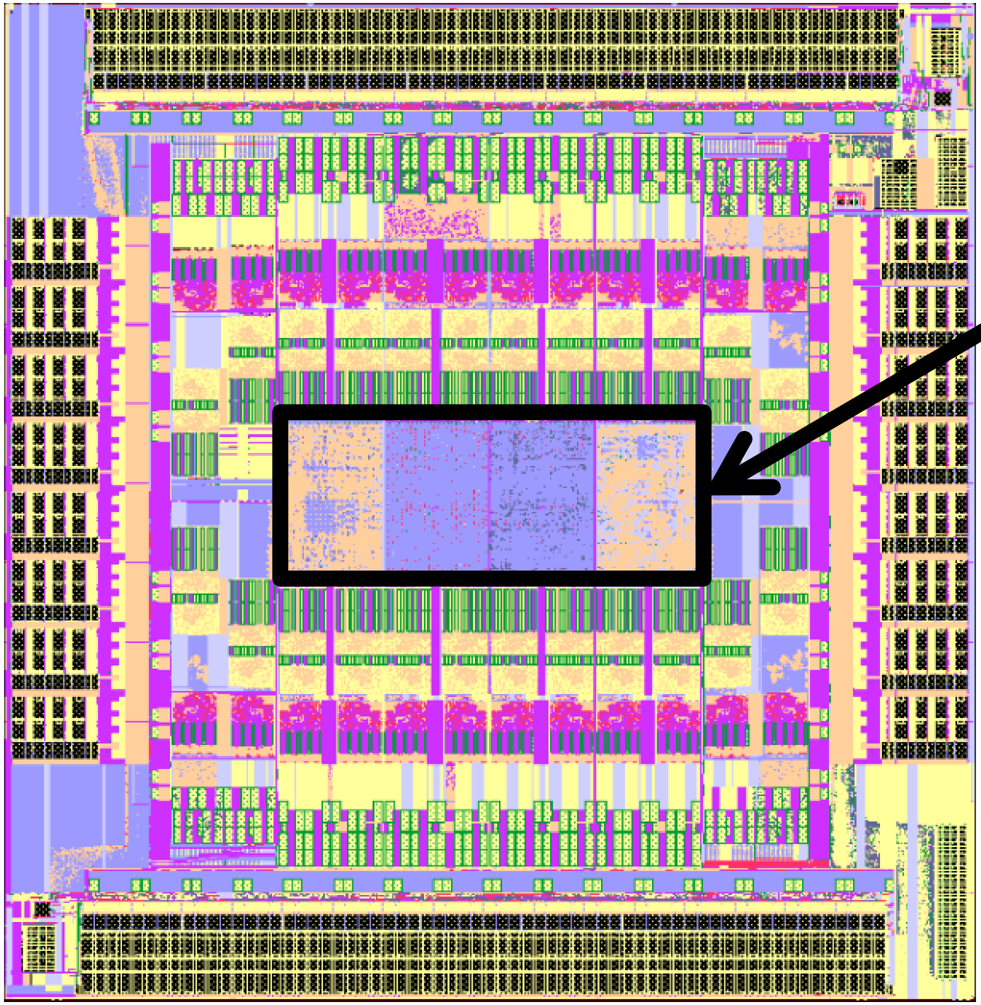
*4

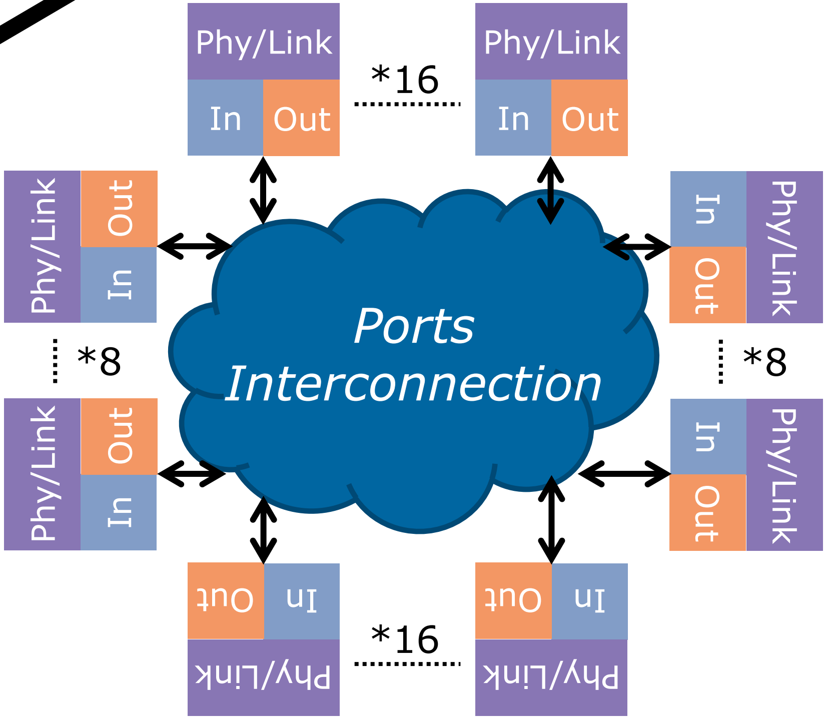
VC0 Input47

*16

VC15 Input47

BXI Switch overview

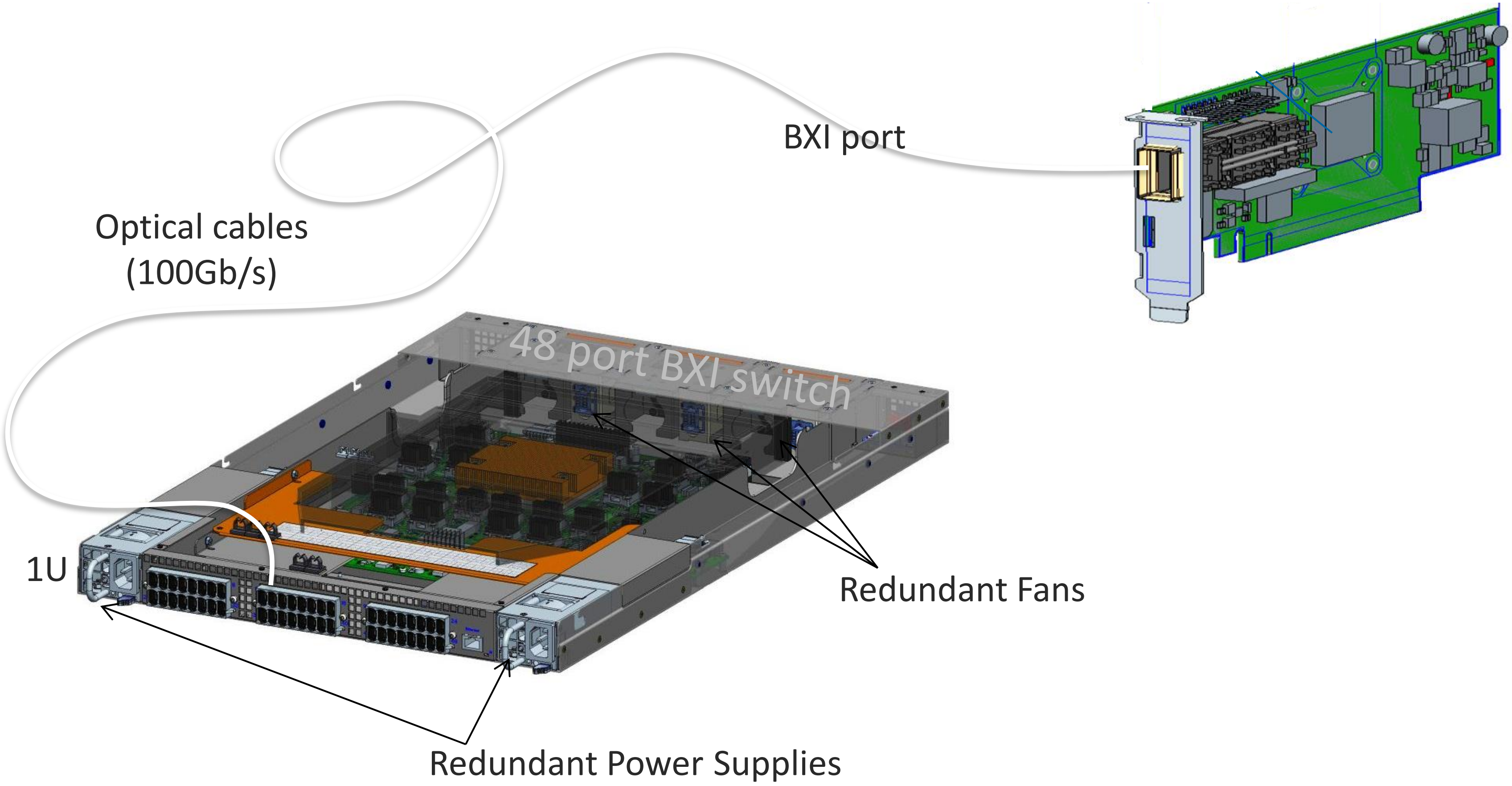




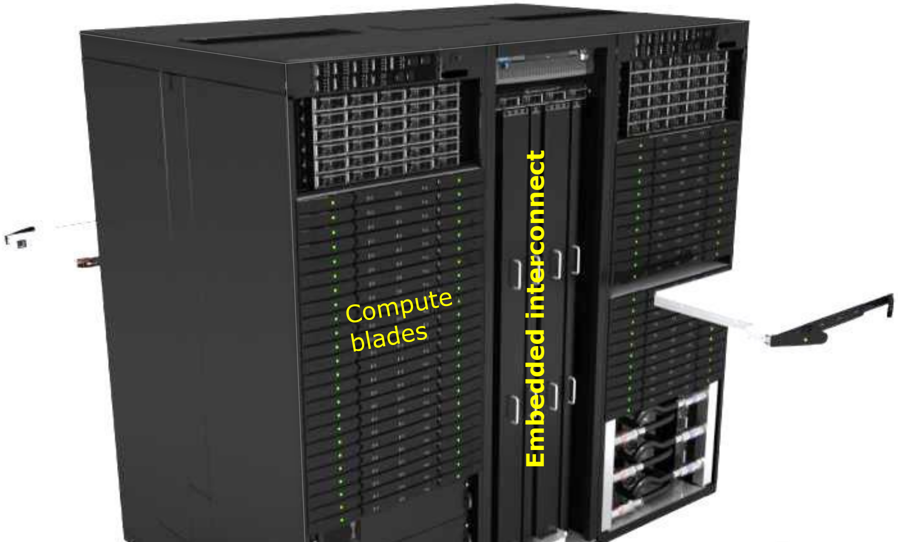
BXI Switch main features

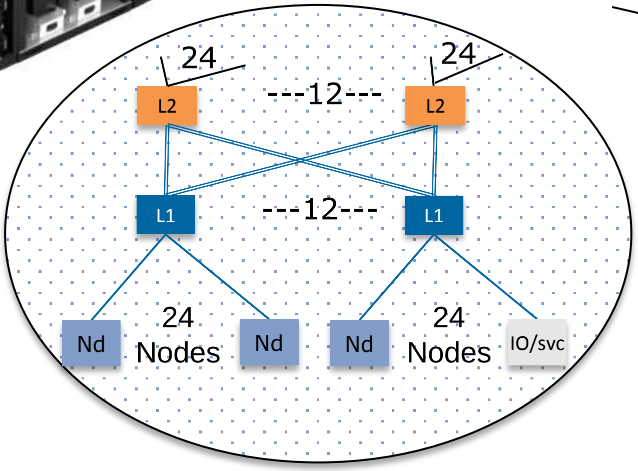
- Many topologies supported : Fat-Tree, Torus, Flattened Butterfly…
  - 16 VC, per-port routing, fine-grain adaptive routing
- Highly efficient arbitration scheme
  - Structured per {VC,Destination}, efficient wormhole switching
  - Highly configurable bandwidth balancing
- Per port traffic generator and checker
  - NIC-independent and highly configurable tool
- Per port performance monitoring
  - A set of fixed counters for the most common measurements
  - A set of highly configurable counters for user-specific measurements

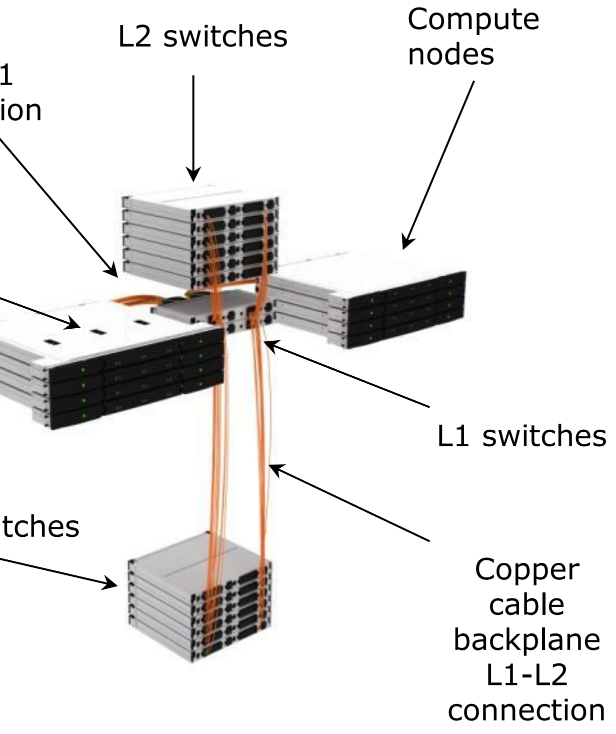
BXI
PCI adapter card and 48p standalone switch



“Sequana” – Embedded interconnect



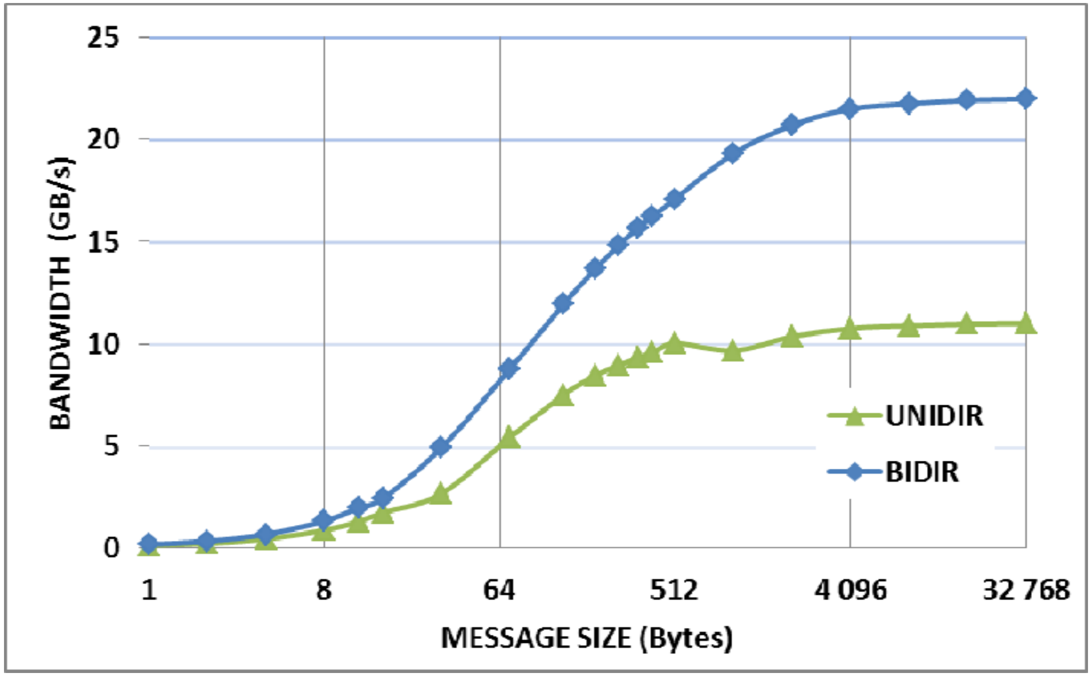




Fast Interconnect layout

BXI Performance estimations

- Message rate 110 M msg/s unidirectional 160 M msg/s bidirectional
- Latency < 1us
- Payload Bandwidth 11GB/s



BXI wrap up

- BXI is a new high performance interconnect for HPC
- BXI offloads communication primitives into the NIC
- BXI boosts MPI communications in HW
- Large radix (48p) switch ASIC
- Highly scalable, up-to 64k nodes
- BXI in production systems in 2016

Acknowledgement

- BXI development has been undertaken under a cooperation between CEA and Atos.
- The goal of this cooperation is to co-design extreme computing solutions.
- Atos thanks CEA for all their inputs that were very valuable for this research.

Questions ?

10^18
Bull
exascale
program

Bull
atos technologies

backup

NIC HW acceleration

- **Atomic operation**

- Integer and floating-point ALU to support portals atomic operations to better implement PGAS language and also improve MPI collectives

- Initiator sends data to target, where initiator and target data are used to perform operation op and put result in target memory. Op is a basic commutative arithmetic or logical operation.

- Examples of atomic operations supported:
  - Operation types: And, or, min, max, sum, product, swap, conditional swap...
  - Data types: signed-unsigned integer types, single-double floating points and complexes

NIC HW acceleration

- **Triggered operation**
- Counting event (on initiator or target side) can be used to trigger operations when a threshold is reached
- Operations can be: Triggered Put, Triggered Get, Triggered Swap...
- Up to 1K triggered operations can be stored inside each NICl
- Triggered operations enable to improve collectives management