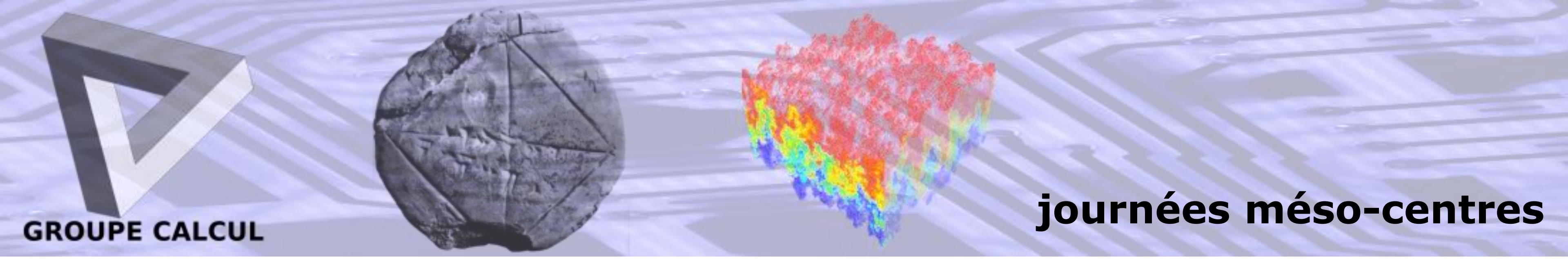


journées méso-centres

Bull Exascale Interconnect (BXI)

un nouveau réseau pour le calcul de haute performance

Jean-Pierre Panziera

11-10-2016

Bull
atos technologies

HPC applications

HPC applications are characterized by:

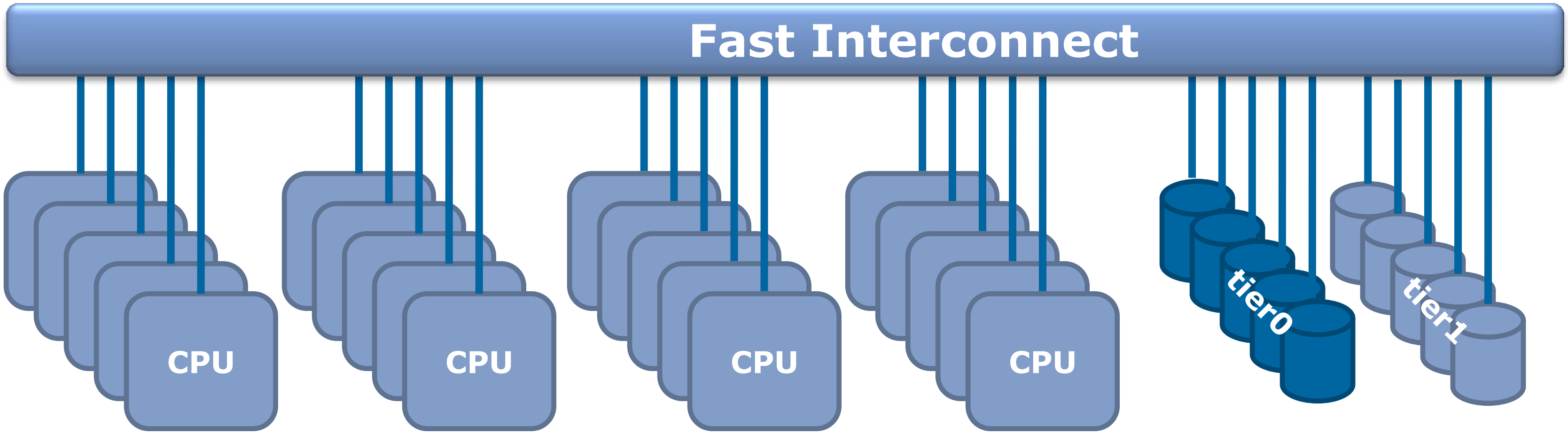
- X-large computing needs (TeraFlops, PetaFlops, ExaFlops…)
- X-large datasets

- large number of processors
- tight coupling between computing threads

- Many short MPI messages (latency)
- Large IO transfers (bandwidth)

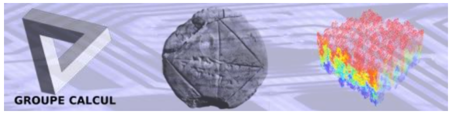
Efficient HPC Interconnect

HPC systems are highly parallel
Petaflop class featuring 1000s CPU nodes



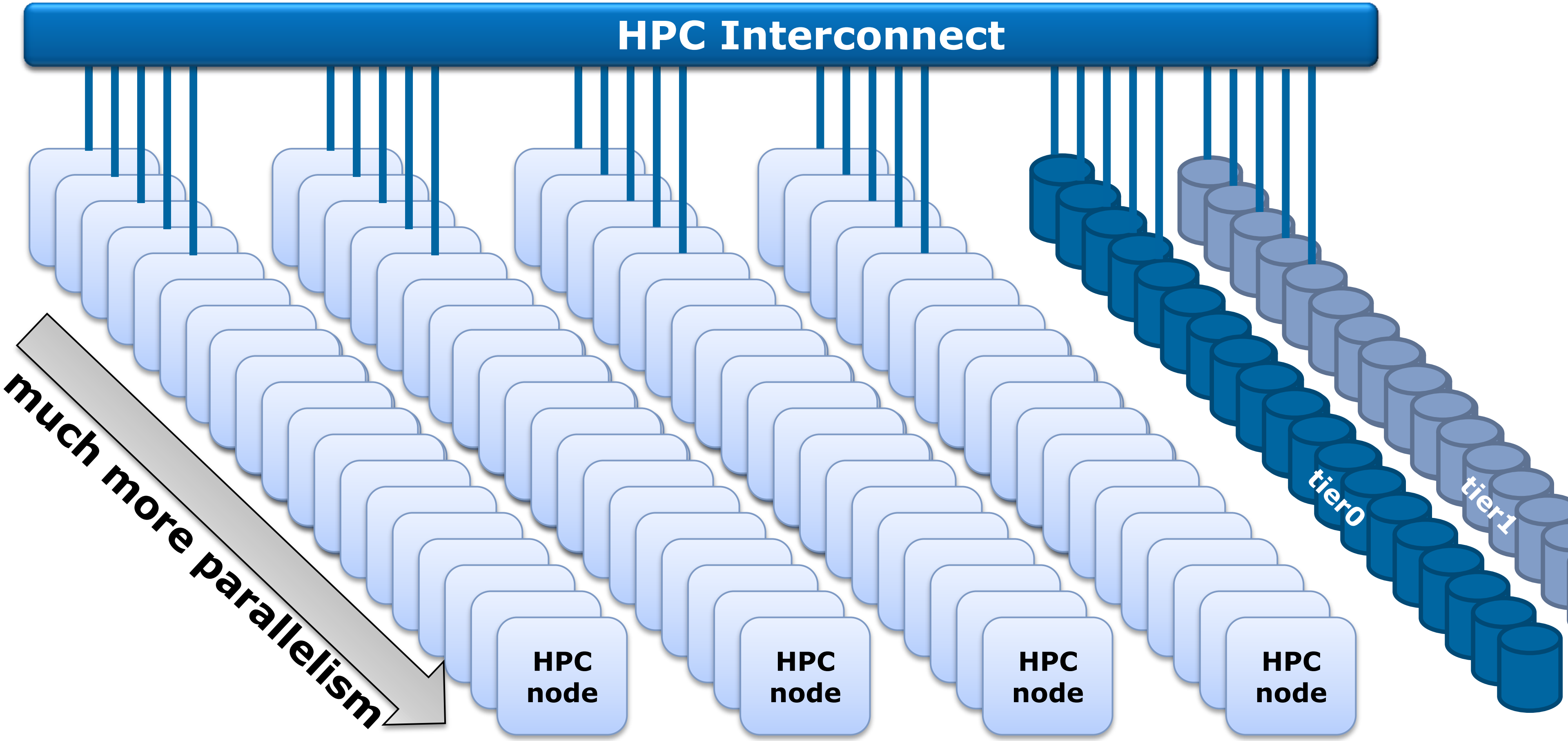
100-10,000 compute nodes using CPUs, typically x86

Multiple storage tiers



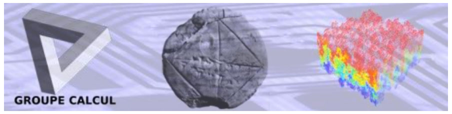
expecting 10-100 Pflops systems in 2016-17
… with HPC specific Processing Units

BXI



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

Multiple
storage tiers



Bull
atos technologies

HPC Interconnect

- **performant**
  - low latency
  - high message rate
  - high bandwidth
- **scalable**
  - 10,000**s** nodes
- **reliable**
  - fault tolerant
  - redundant
- **efficient**
  - handle simultaneously different flow types – small & big messages - MPI & IO
  - Adaptive routing
  - small memory footprint
  - link-level checking & retry, ECC protection
- Offload communications in **Hardware**
  - HPC cores are many but slow(er)

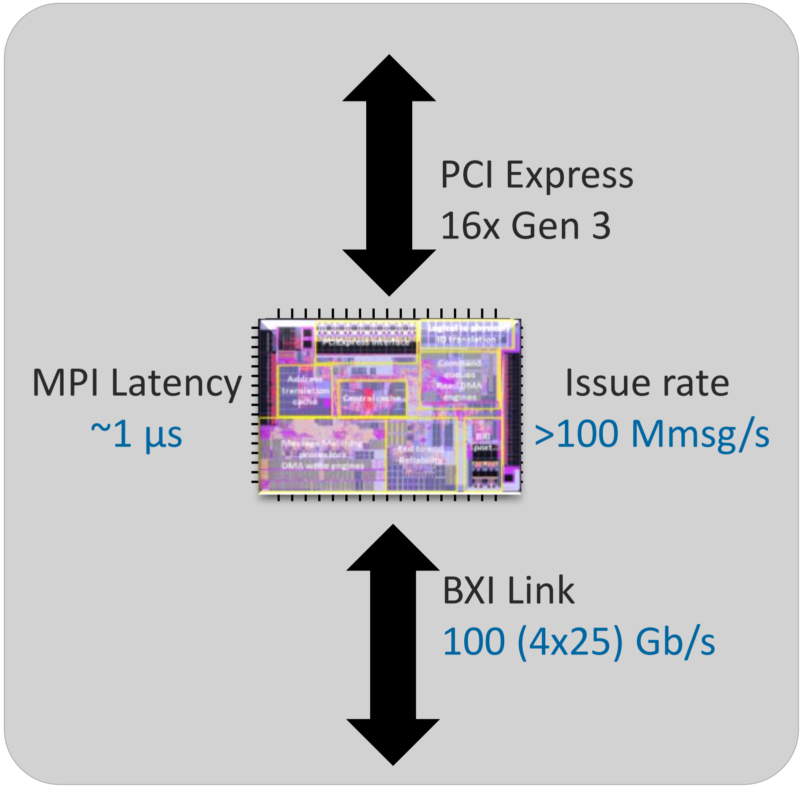
BXI overview

High Performance Interconnect for HPC

- **BXI: High Performance Interconnect for HPC**
  - Lowest latency, Highest message rate at scale, Highest Bandwidth
- **BXI full acceleration in hardware for HPC applications**
  - based on Portals 4 (Sandia), BXI provides full HW acceleration for:
    - **MPI** and PGAS communications (send/recv, RDMA)
    - High performance collective operations
- **BXI highly scalable, efficient and reliable**
  - Exascale scalability → 64k nodes (v1)
  - Adaptive Routing, Quality of Service (QoS)
  - End-to-end error checking + link level CRC + ECC in ASICs
- **BXI co-designed with CEA**

BXI Network is based on 2 ASICs

NIC ASIC



PCI Express
16x Gen 3

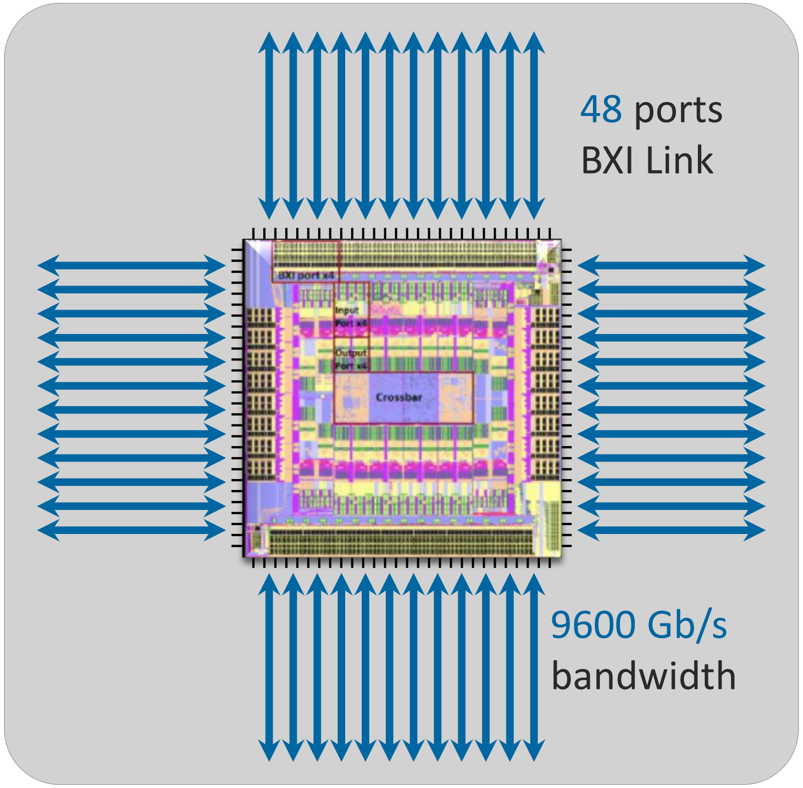
MPI Latency
~1 µs

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 error **counters** for Applications performance analysis

BXI
offloading MPI communication in HW

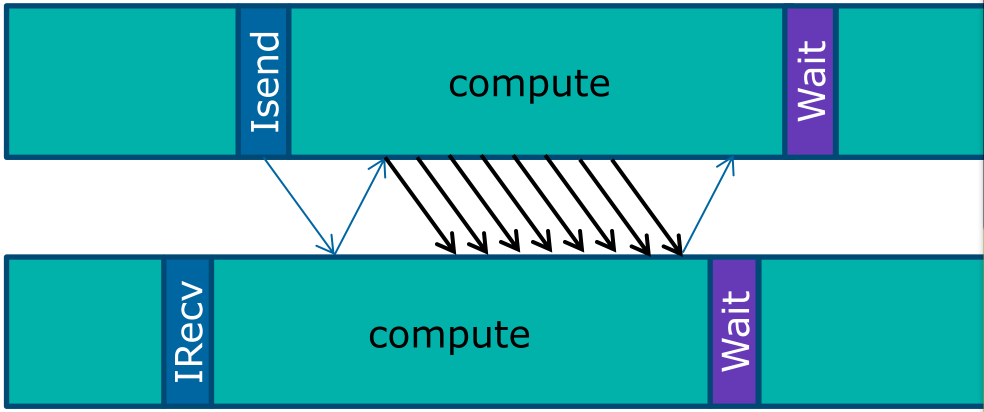
address V2P

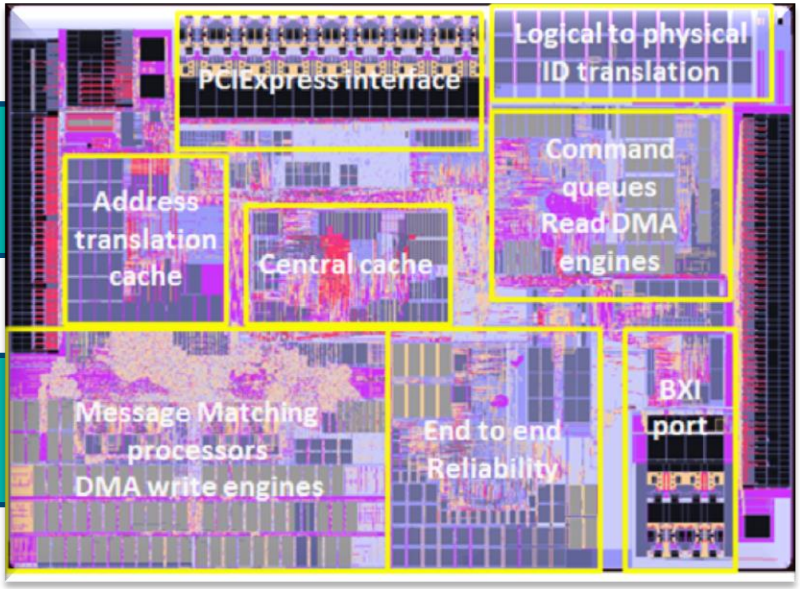
size

rank L2P

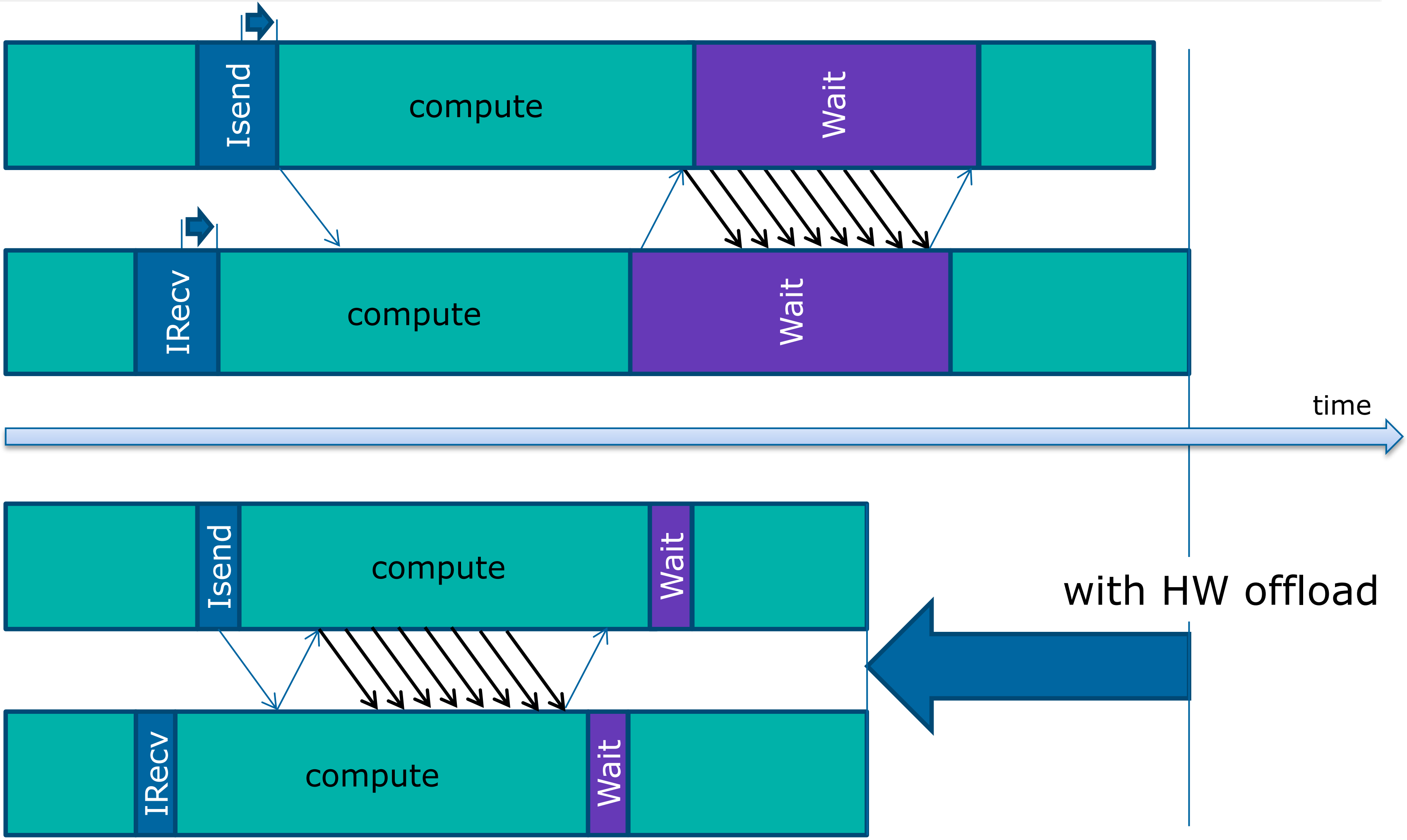
message order

```
#include <mpi.h>

int MPI_Isend( const void *buf, int count, MPI_Datatype datatype, int dest,    int tag, MPI_Comm comm, MPI_Request *request)

int MPI_IRecv(void *buf, int count, MPI_Datatype datatype, int source, int tag, MPI_Comm comm, MPI_Request *request)

int MPI_Wait(MPI_Request *request, MPI_Status *status)
```



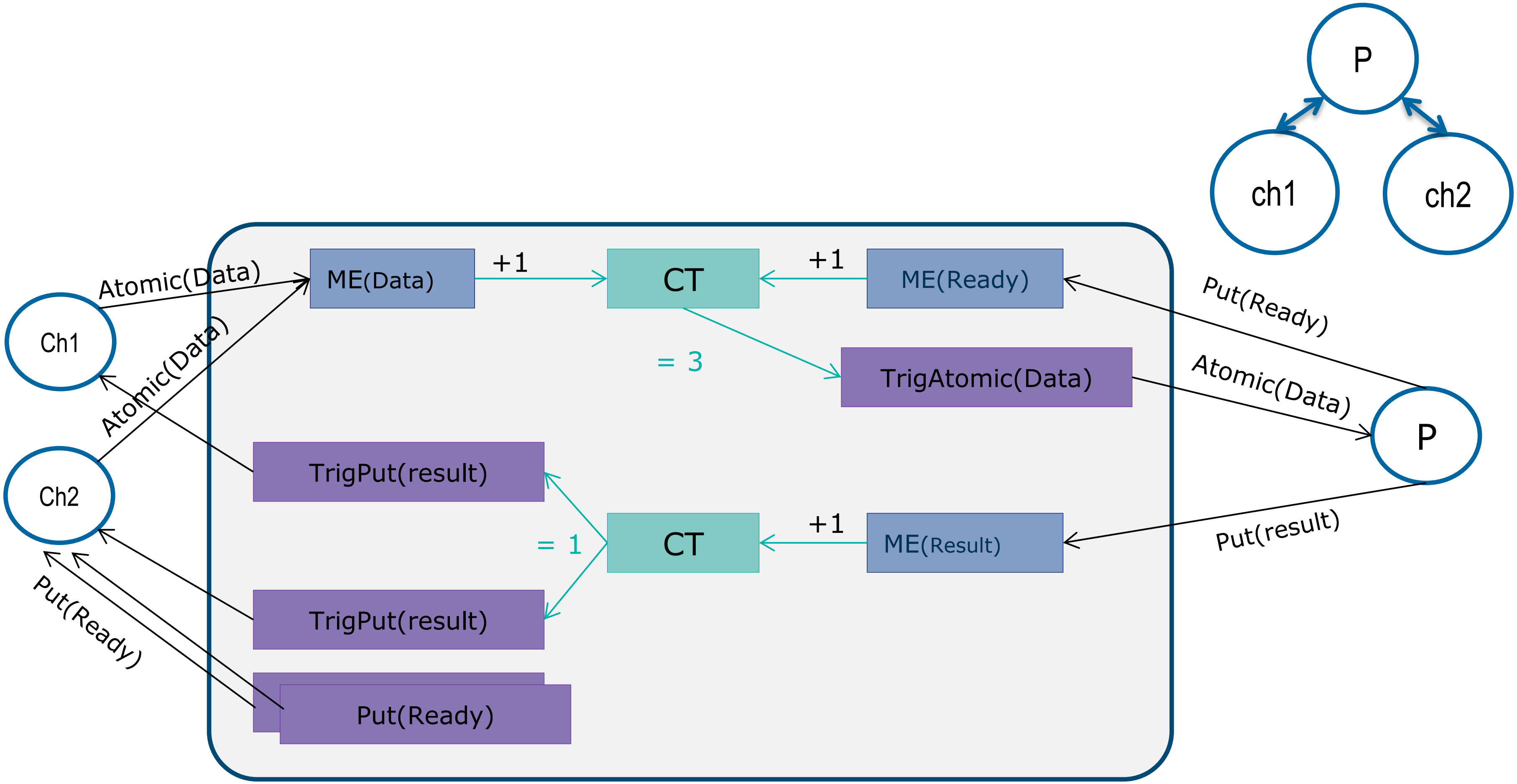

PCIExpress interface
Logical to physical ID translation
Command queues Read DMA engines
Address translation cache
Central cache
Message Matching processors DMA write engines
End to end Reliability
BXI port

BXI
offloading MPI communication in HW

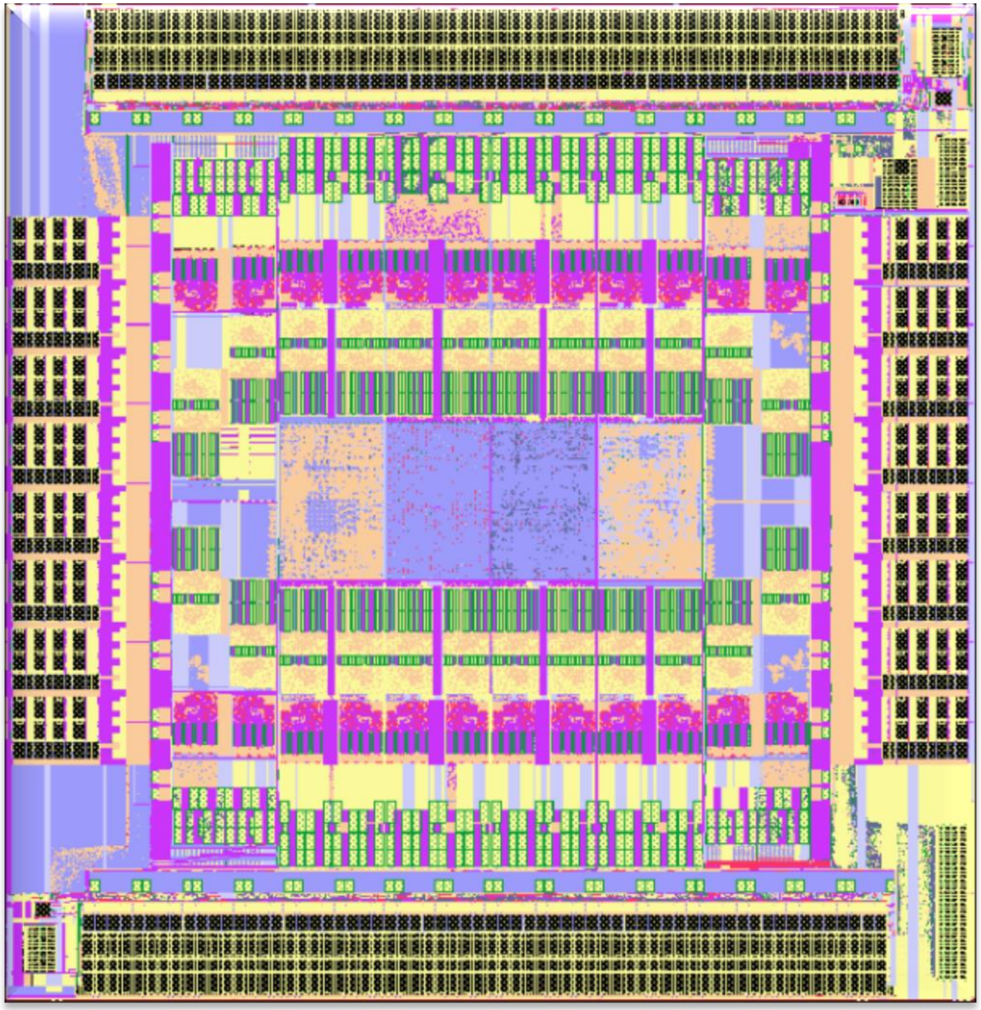


with HW offload

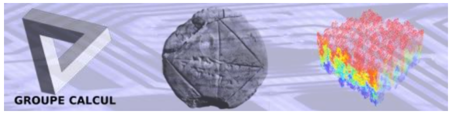
BXI MPI AllReduce
using Triggered and Atomic operations



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



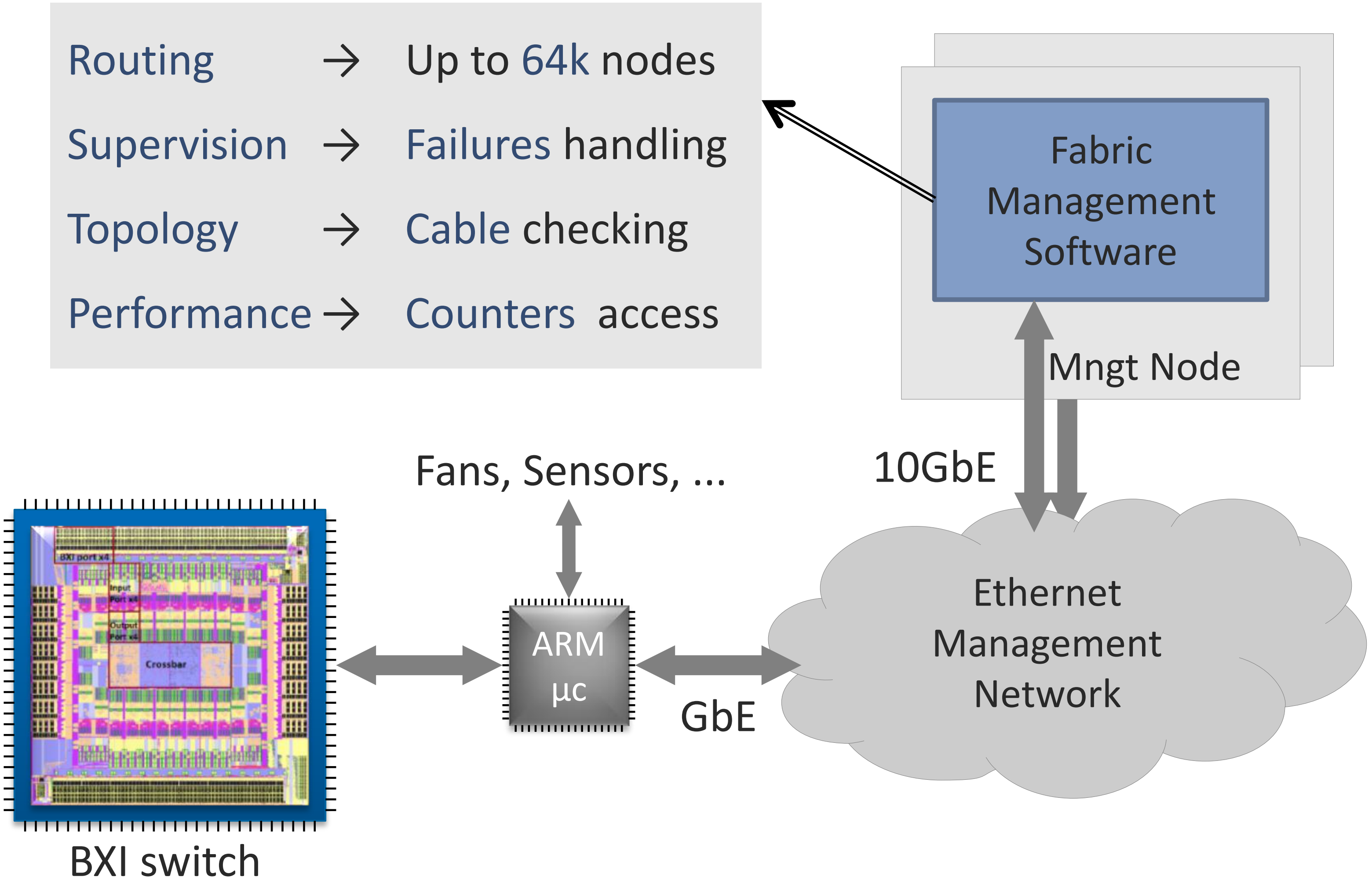
BXI fabric features

- Scalable up to **64K NICs**
- **100 Gb/s** links (4 lanes x25,278125 GT/s)
- **Reliable** and ordered network (end to end + Link level)
- **Flexible** with full routing table
  - Many topologies supported (**Fat-Tree**, Torus, Hypercube, **All-to-All**…)
  - Ease routing algorithm optimization
- **Adaptive routing**

- Extensive buffering implementing 16 virtual channels preventing deadlock and efficiently balancing traffic
- Quality of Service (**QoS**) 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**

Fabric Management Software



Routing → Up to 64k nodes

Supervision → Failures handling

Topology → Cable checking

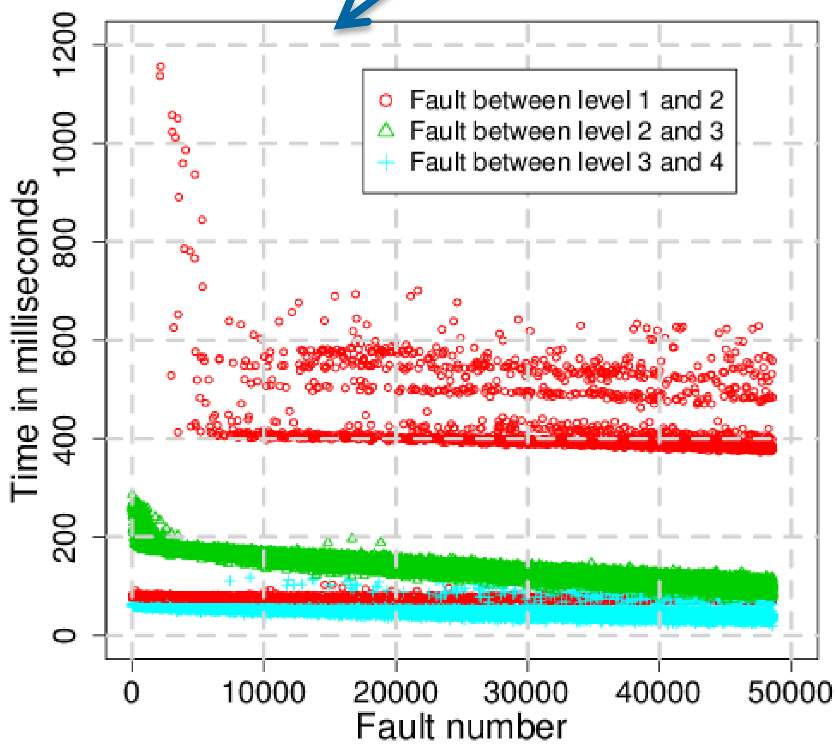
Performance → Counters access

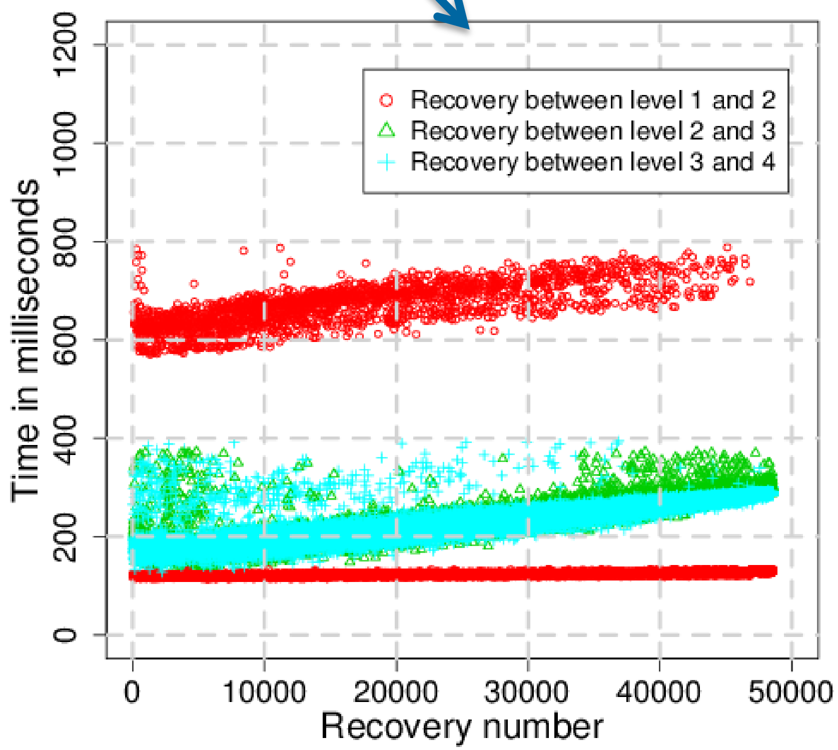
BXI Routing Online Mode Processing Time e.g. 64k nodes

routing table updates computed in < 1s

on link failure

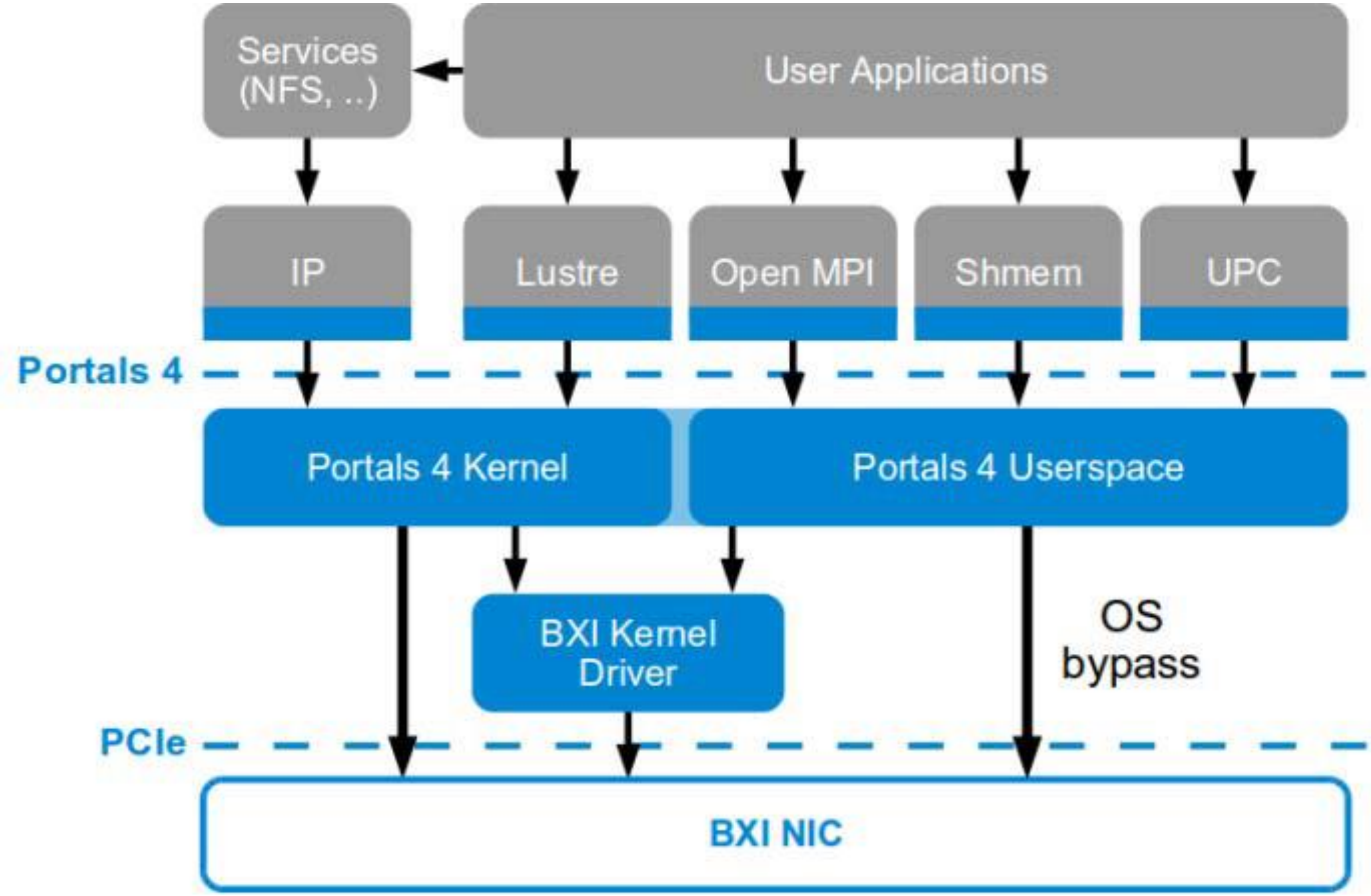
on link recovery

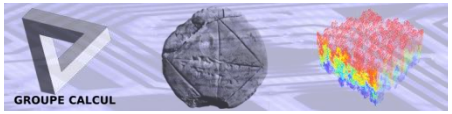




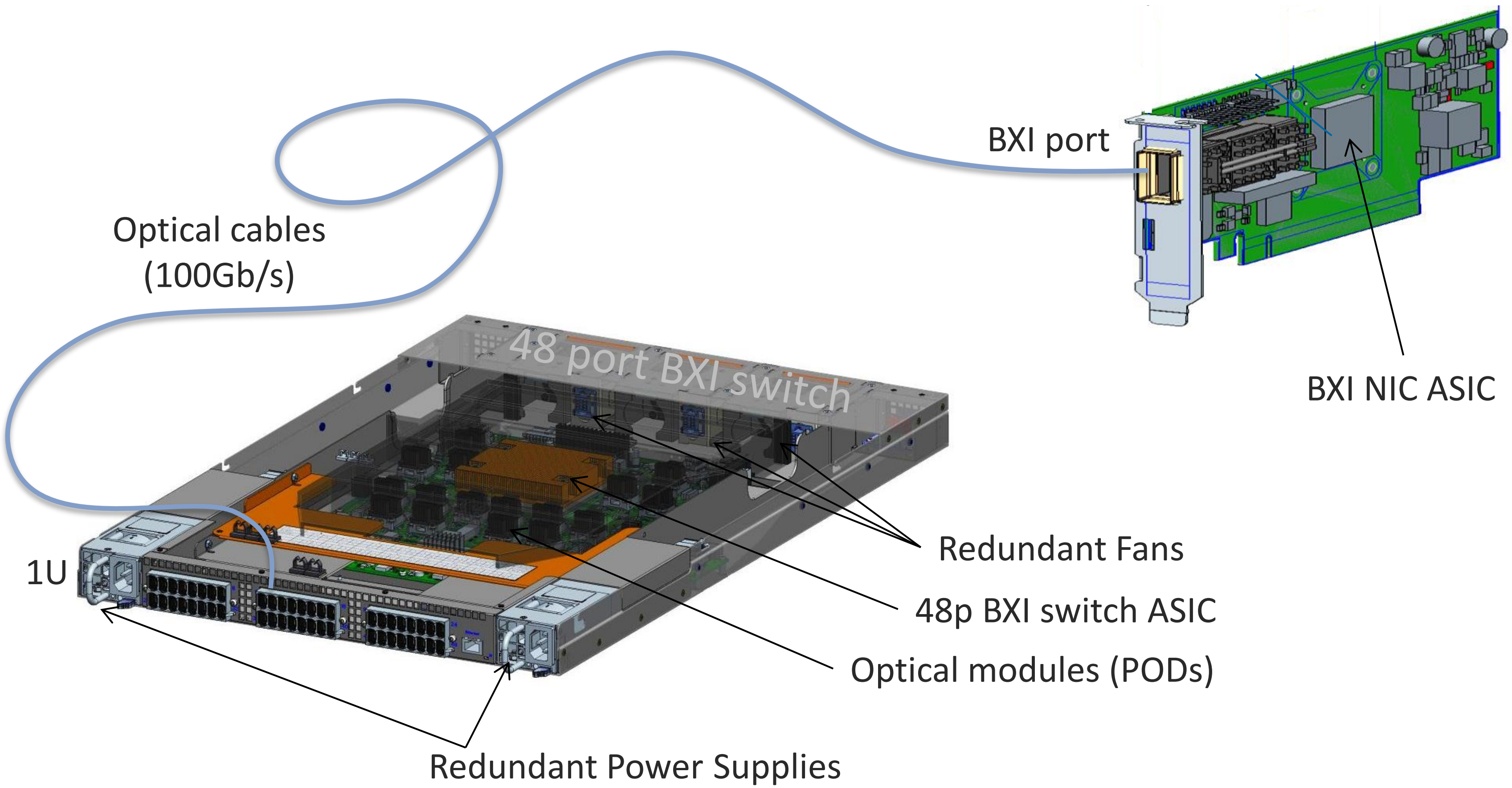
Quintin, Vignéras; Fault-Tolerant Routing for Exascale Supercomputer: The BXI Routing Architecture. HiPINEB'15
Quintin, Vignéras; Transitively Deadlock-Free Routing Algorithms . HiPINEB'16

BXI Software compute stack



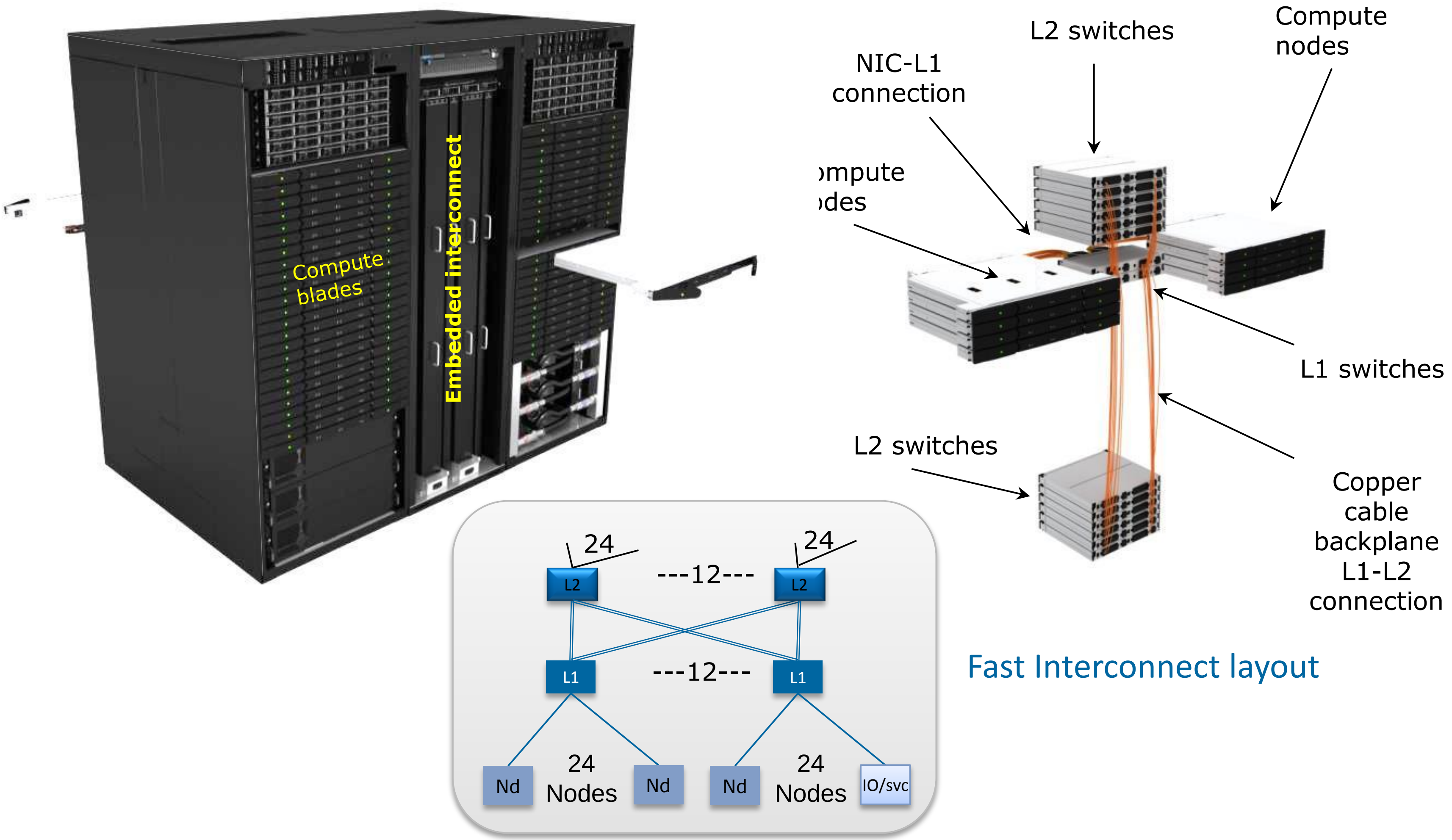


BXI
PCI adapter card and 48p standalone switch



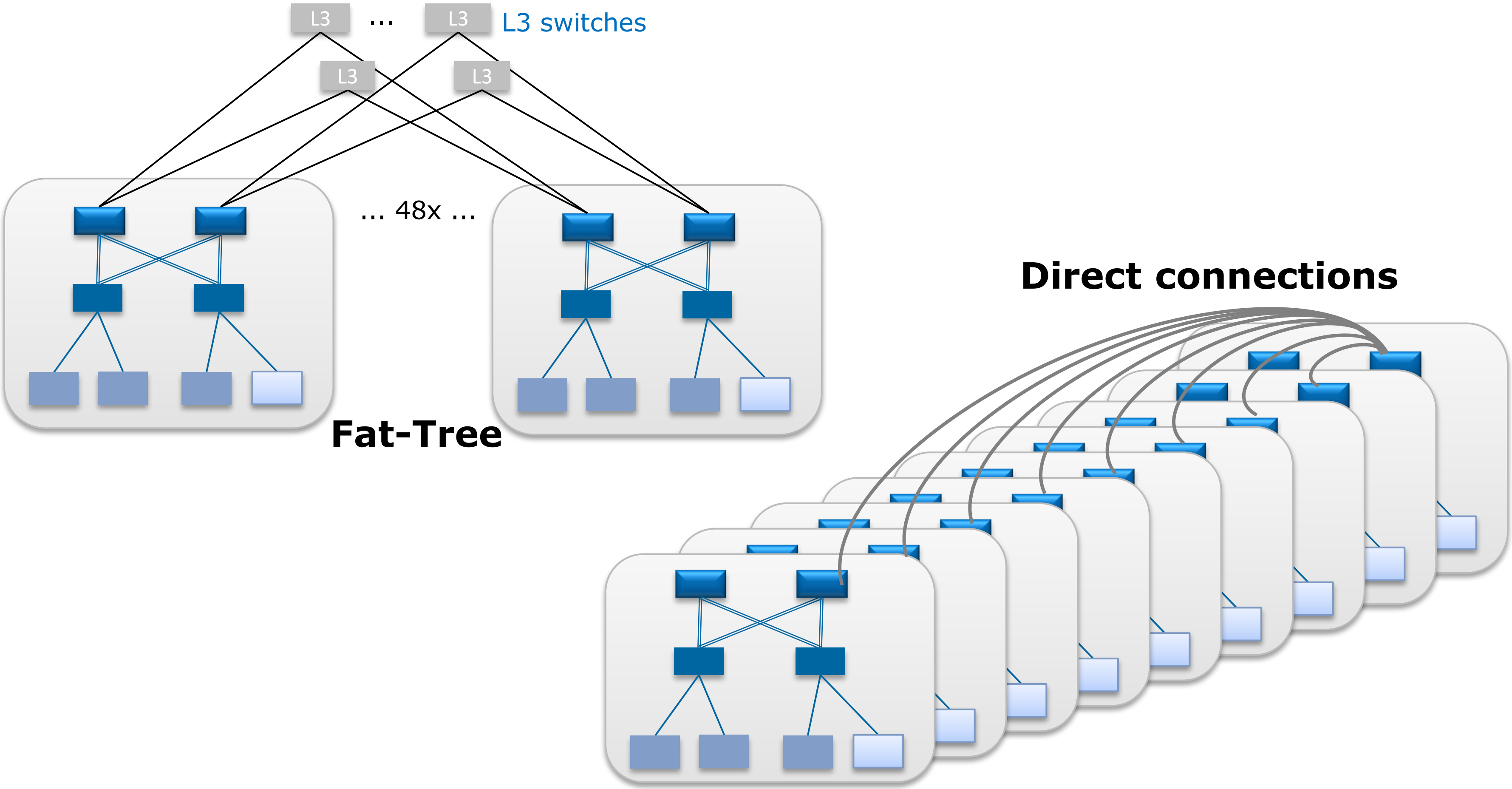
“Sequana” – Embedded interconnect

BXI



Fast Interconnect layout

Sequana cells interconnection



BXI wrap up

- BXI is Atos new High Performance Interconnect for HPC
- BXI offloads communication primitives into the NIC
- BXI boosts MPI communications in Hardware
- Highly scalable, up-to 64k nodes
- Fist BXI system installed in Q4-2016
- Large BXI deployment (8+K nodes system) in 2017

Questions?

10^18
Bull
exascale
program

Bull
atos technologies