Forum Teratec 2017 27 & 28 JUIN JUNE 2017 ECOLE POLYTECHNIQUE PALAISEAU / FRANCE

Boost HPC application performance thanks to hardware offload

Bull
atos technologies

27-28 June, 2017

Bull
atos technologies

Disclaimer

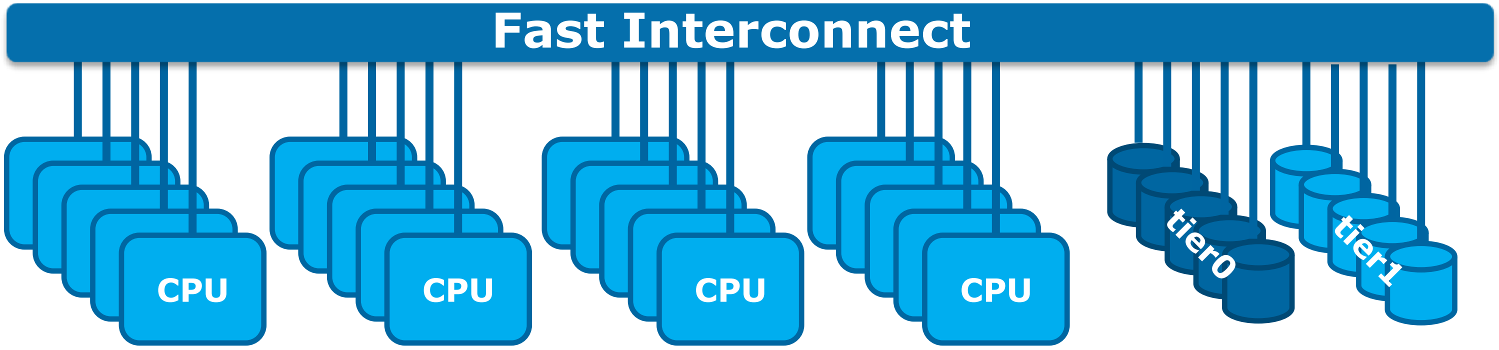
Atos, the Atos logo, Atos Consulting, Atos Worldgrid, Worldline, BlueKiwi, Canopy the Open Cloud Company, Yunano, Zero Email, Zero Email Certified and The Zero Email Company are registered trademarks of Atos. June 2017. © 2017 Atos. Confidential information owned by Atos, to be used by the recipient only. This document, or any part of it, may not be reproduced, copied, circulated and/or distributed nor quoted without prior written approval from Atos.

Atos may make changes to specifications and product descriptions at any time, without notice.

All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice. This is not a binding offer.

Atos hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

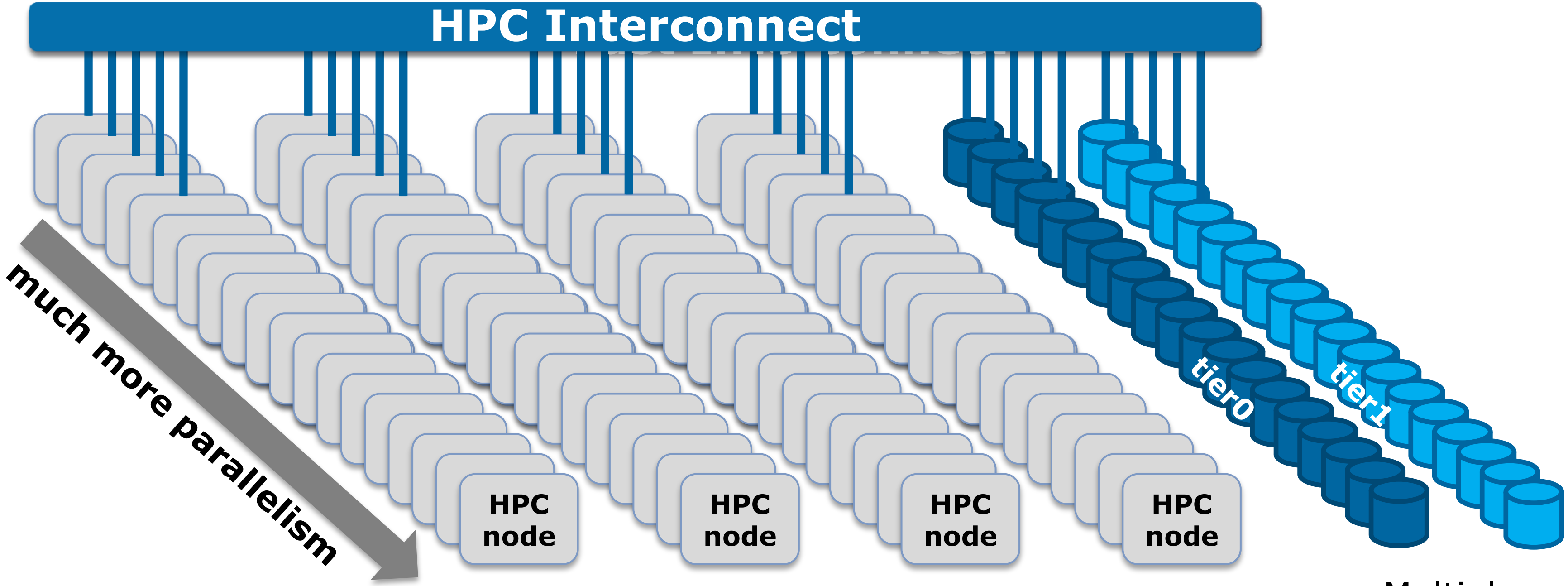
HPC systems are highly parallel Petaflops class featuring 1000s CPU nodes



1000-10,000 compute nodes
using CPUs, typically x86

Multiple
storage tiers

10-100 Pflops systems being deployed … with HPC specific Processing Units



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

Multiple
storage tiers

Network has to become intelligent

- **Smart interconnects**, based on the **ability to offload in hardware MPI semantics** from the host CPU, can be translated directly to **greater application performance**.
- Communications and computations progress completely independently.
- Performance is not impacted by heavy load on the host CPU.
- Point to point communication include MPI hardware matching.
- Triggered and atomic operations are used for protocol offloading (rendezvous, collectives, etc.).

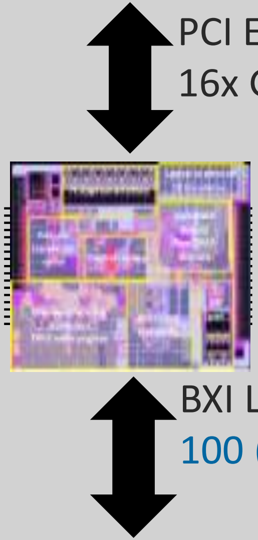
BXI – Interconnect overview

BXI

- **BXI 1st generation of Bull Exascale Interconnect**
- **BXI full acceleration in hardware for HPC applications**
- **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 Network is based on 2 ASICs

- ▶ One ASIC, in nodes, for NIC (**N**etwork **I**nterface **C**ontroller)

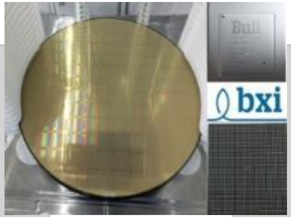


PCI Express
16x Gen 3

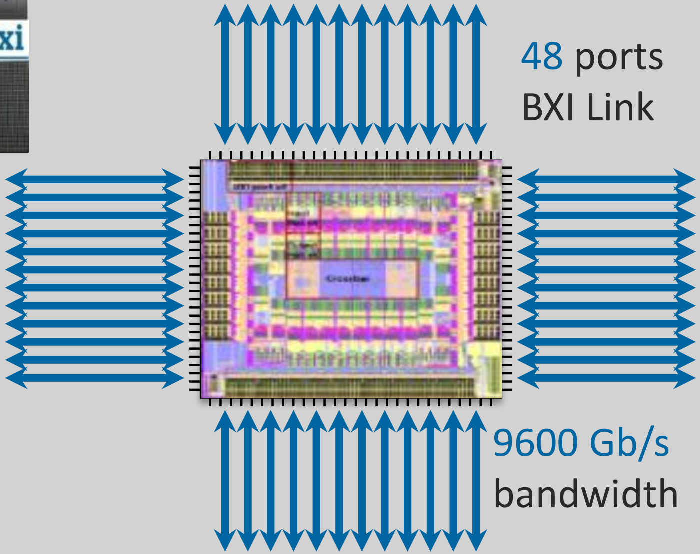
MPI Latency
<1 µs

Issue rate
> 100 Mmsg/s

BXI Link
100 (4x25) Gb/s



- ▶ One ASIC for Switch equipment

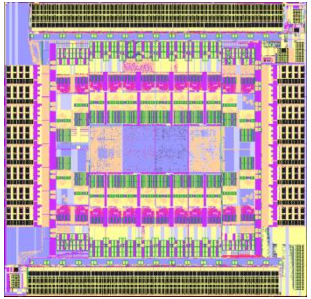


48 ports
BXI Link

9600 Gb/s
bandwidth

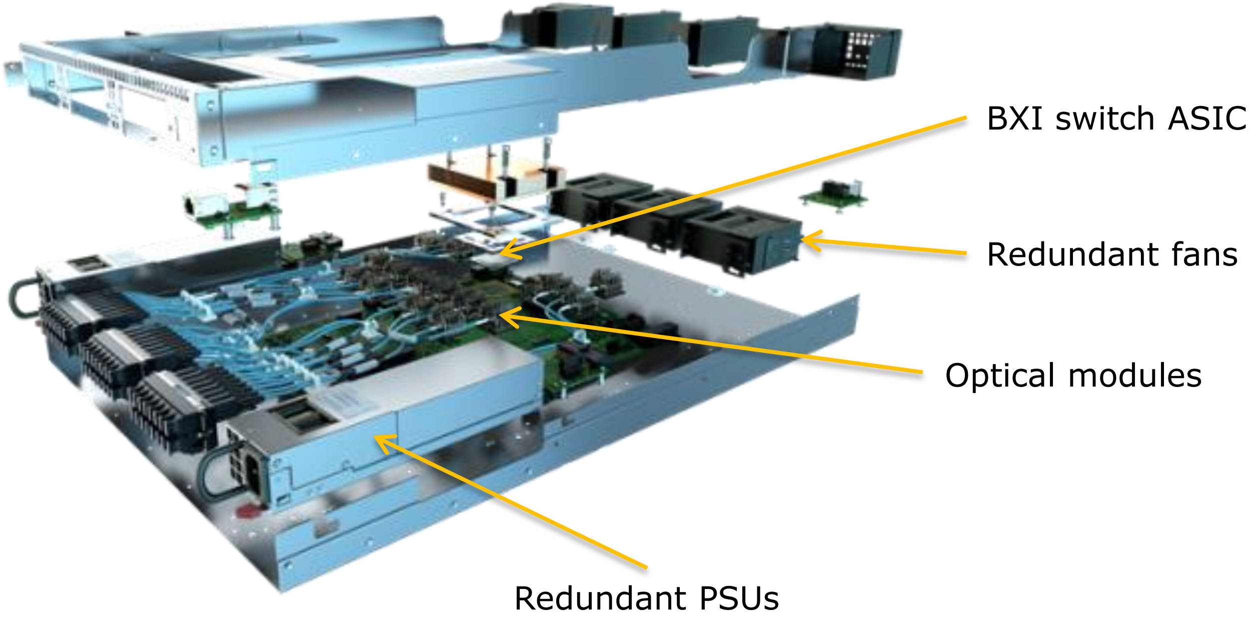
BXI – Switch

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 60W)
- **Techno**: TSMC 28nm HPM

BXI standalone switch: 1U, 48 optical ports



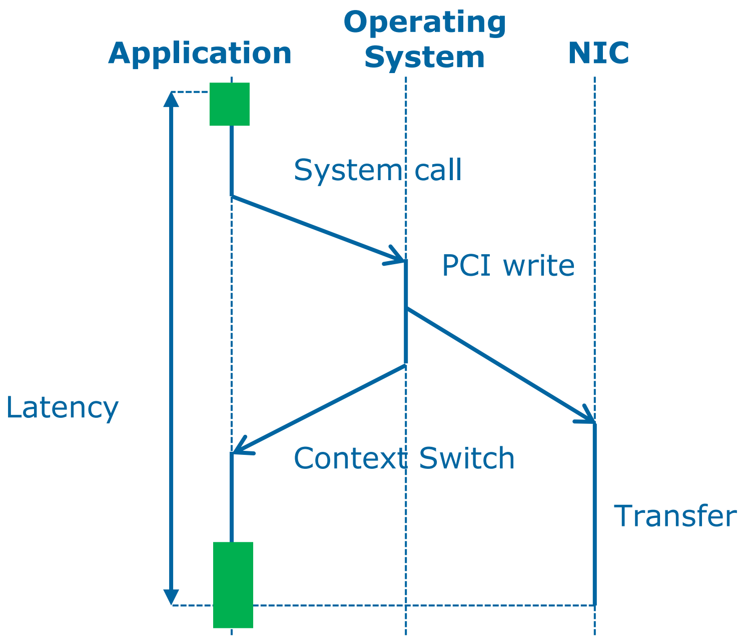
BXI – NIC
Main features

- **Implements in hardware Portals 4 communication primitives**
- **OS and application bypass**
- **Collective Operations offload in HW**
- **End-to-End reliability**
- **Load balancing & QoS with Virtual Channels**
- **Performance counters**

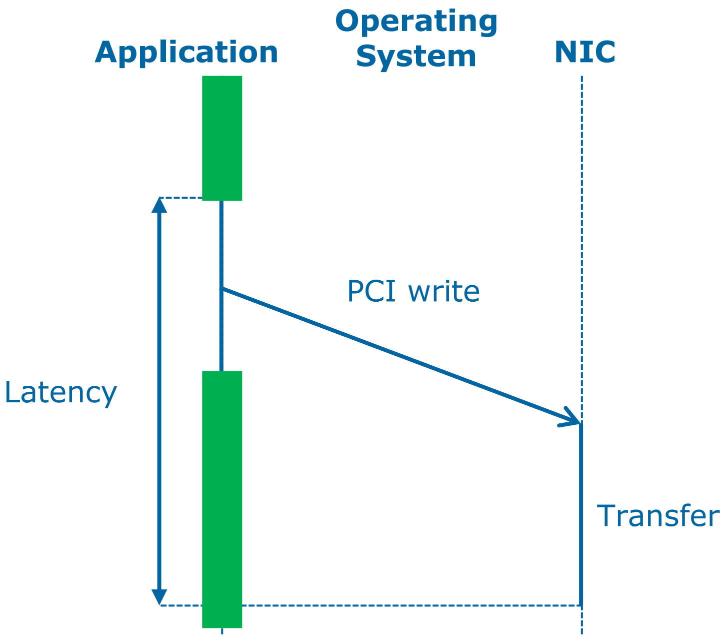
OS Bypass: the first step to offload (1/2)

Source node

Without OS Bypass

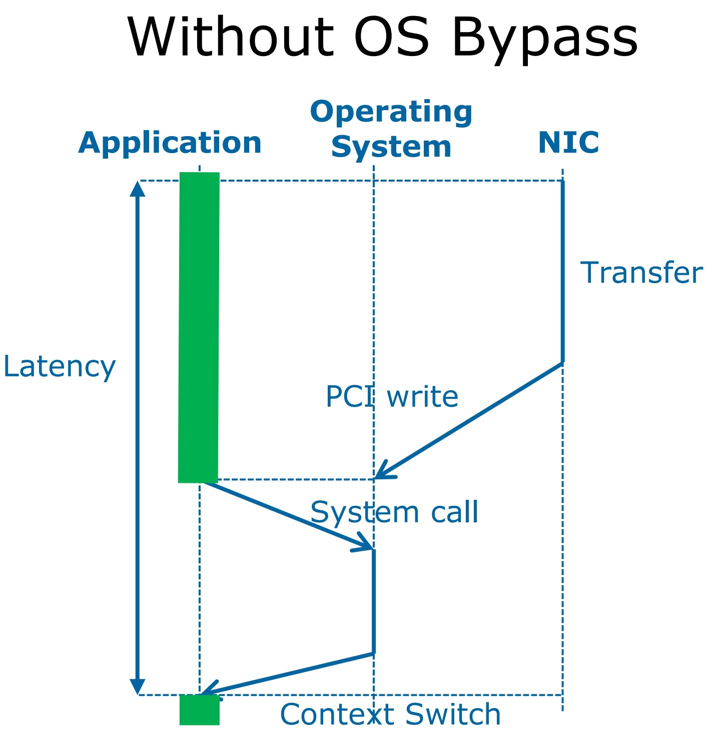


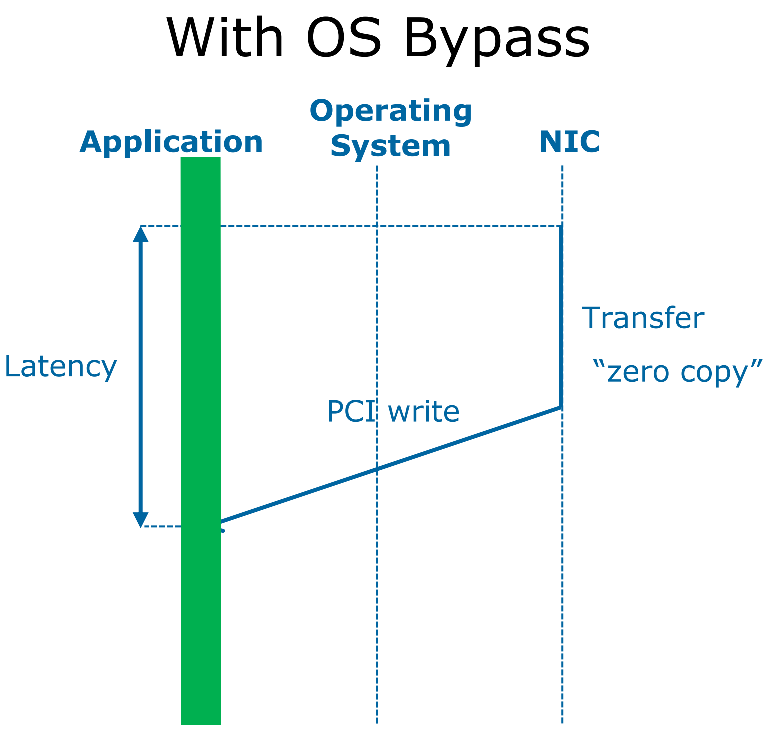
With OS Bypass



OS Bypass: the first step to offload (2/2)

Destination node





BXI - Offloading MPI communication

address V2P

size

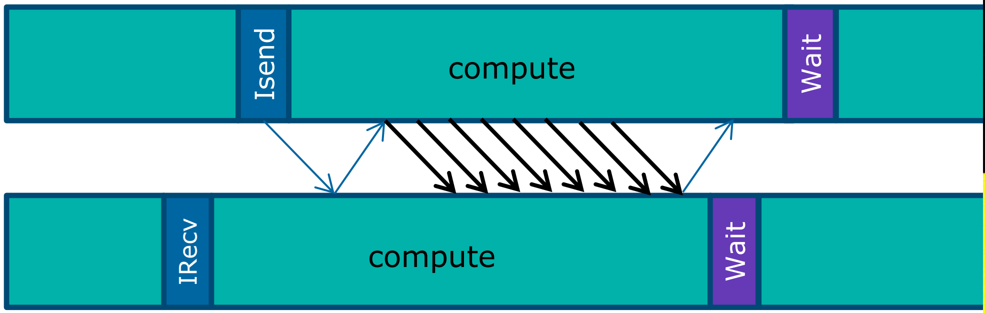
rank L2P

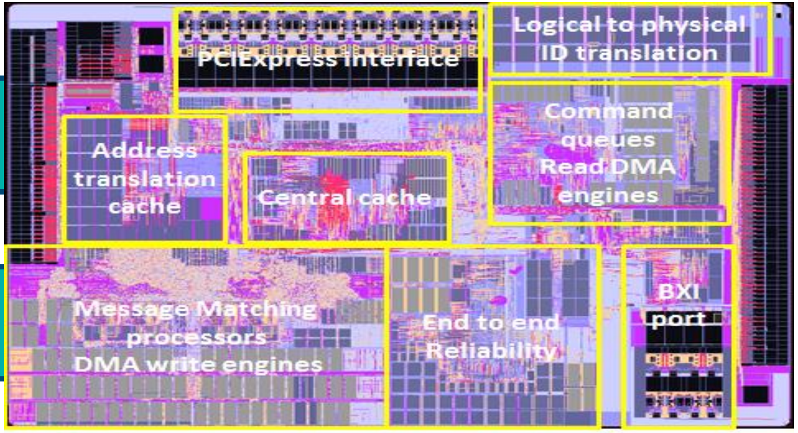
message order

portals

No impact on computation during data transmission

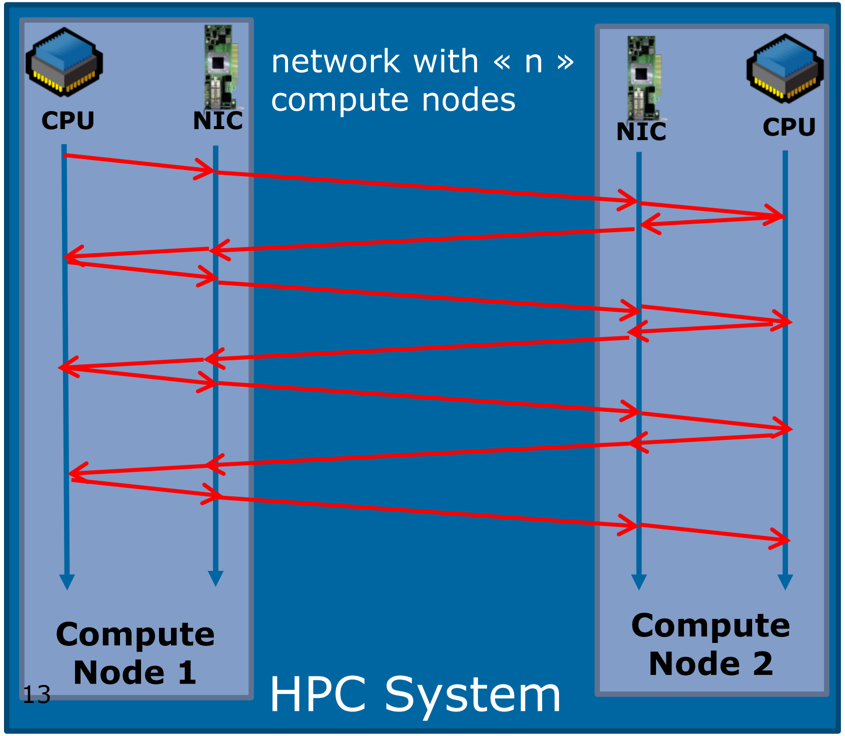
```
#include <mpi.h>

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

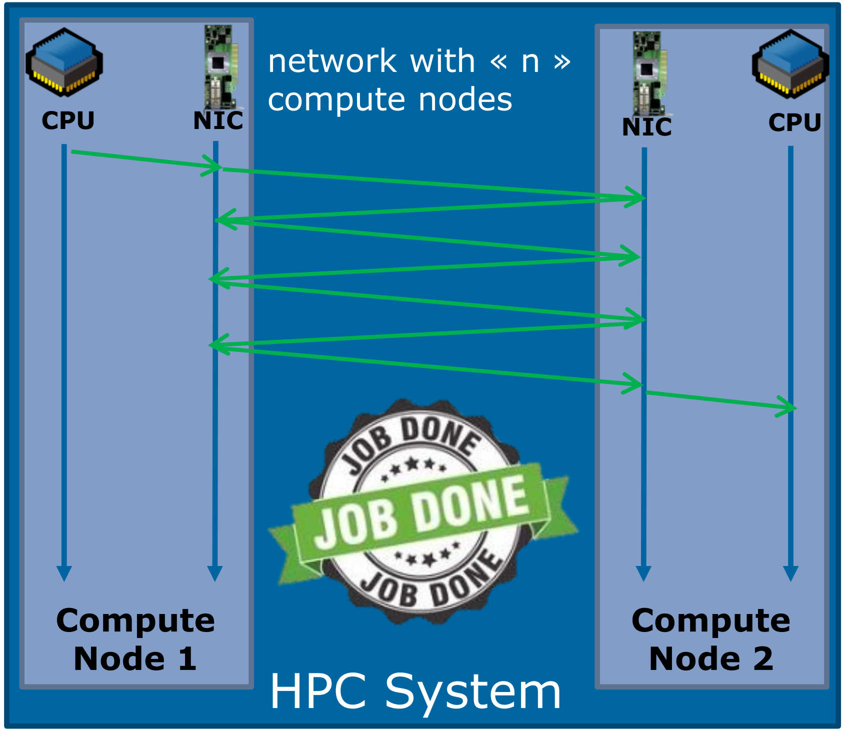
Offload Mechanism & Benefits

Conventional interconnect:
no offload on the NIC

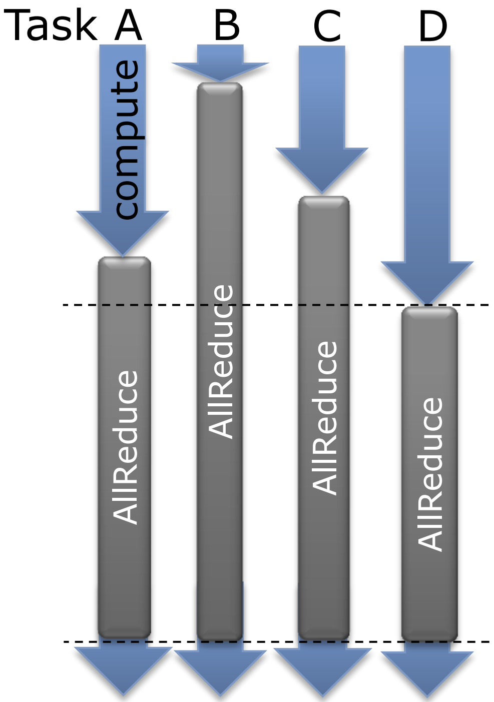


BXI

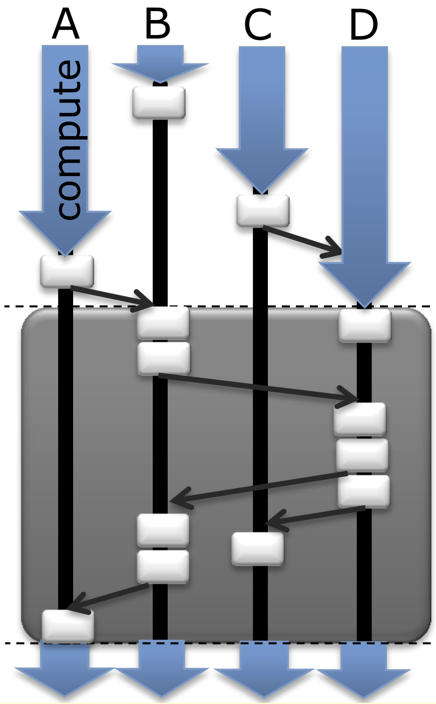
- ▶ Faster communications
- ▶ Less CPU solicitation



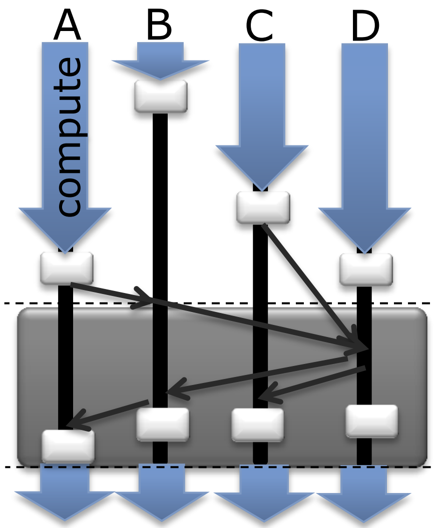
BXI - Offloading MPI communication



MPI-1 Collective



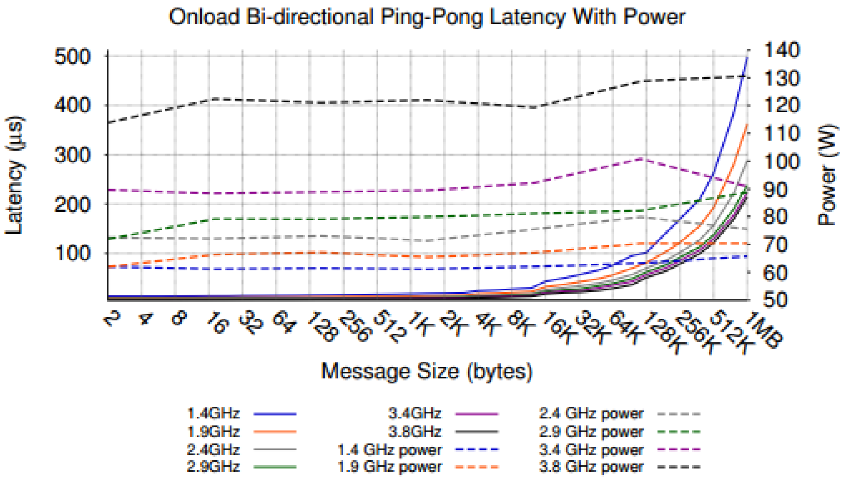
Non-triggered operations

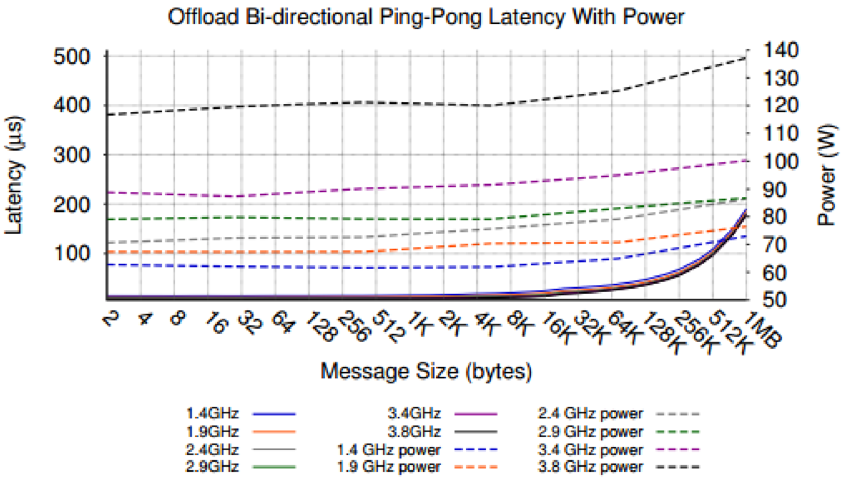


BXI "triggered operations"

Offload offers other benefits

- Offloaded networking approach:
  - is **less frequency sensitive** than onloaded networking,
  - provides **major power reductions**, particularly significant with large-scale systems.

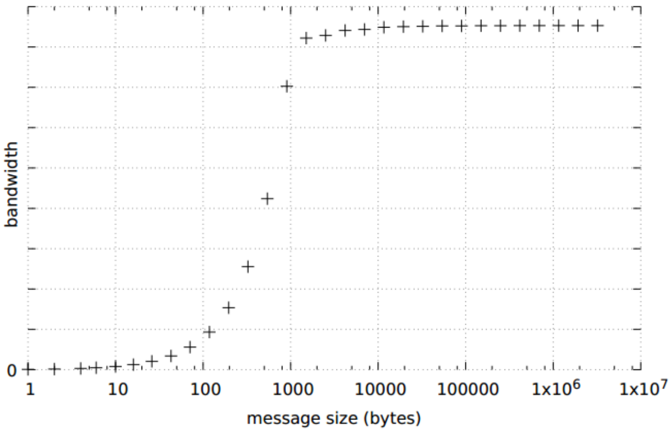


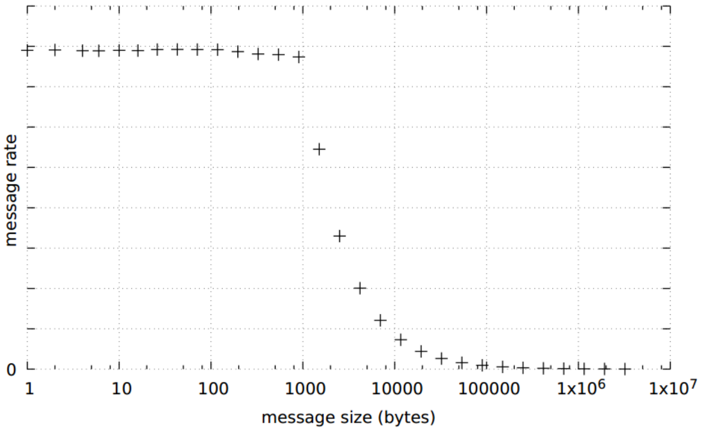


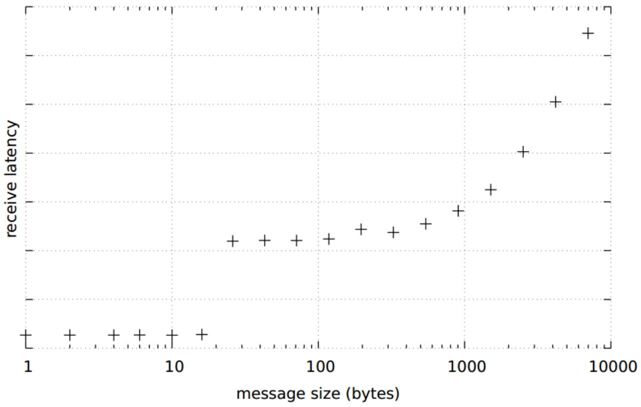
- Paper co-authored by Matthew Dosanjh, Ryan Grant, and Ron Brightwell entitled "Re-evaluating Network Onload vs. Offload for the Many-Core Era".

Performance

BXI







Conclusion and Next Steps

- BXI system 1st prototype on Top/Green500,
- BXI full system (> 8k nodes) in 2017,
- More installations coming,
- Growing BXI ecosystem,
- Next generation of BXI in preparation.

Thanks

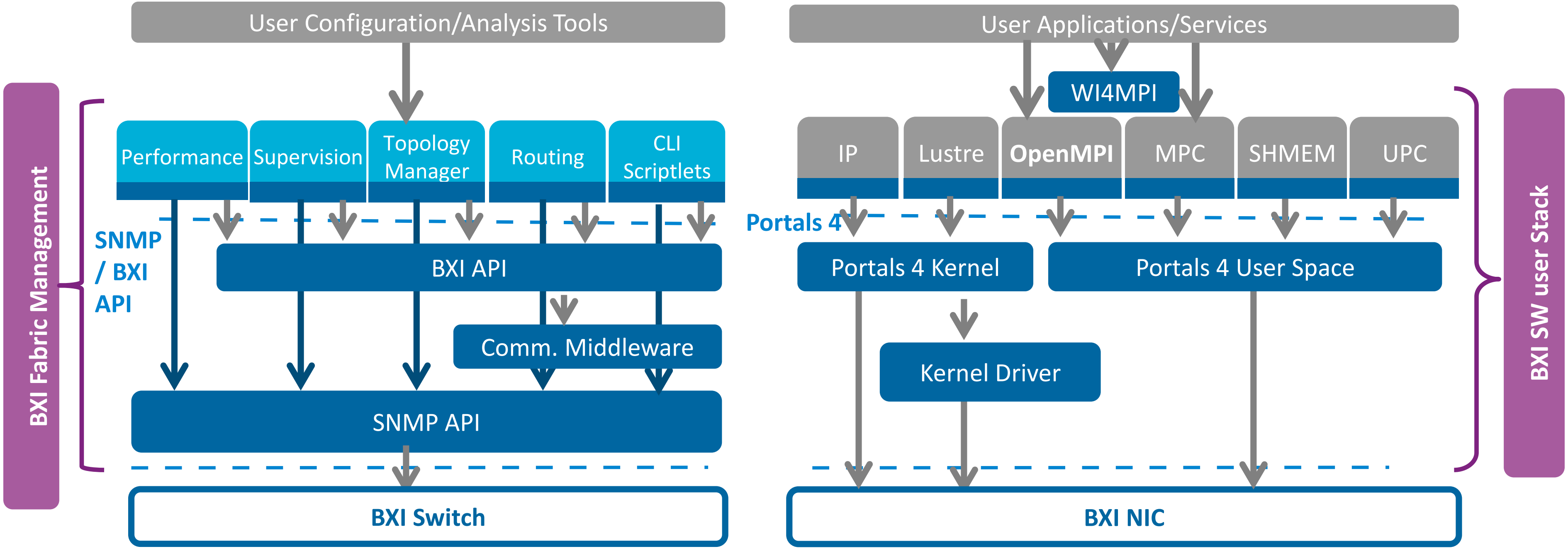
For more information please contact:
T+ +33 (1) 30 80 74 94
M+ +33 (6) 86 49 33 21
fabien.locussol@atos.net

Bull
atos technologies

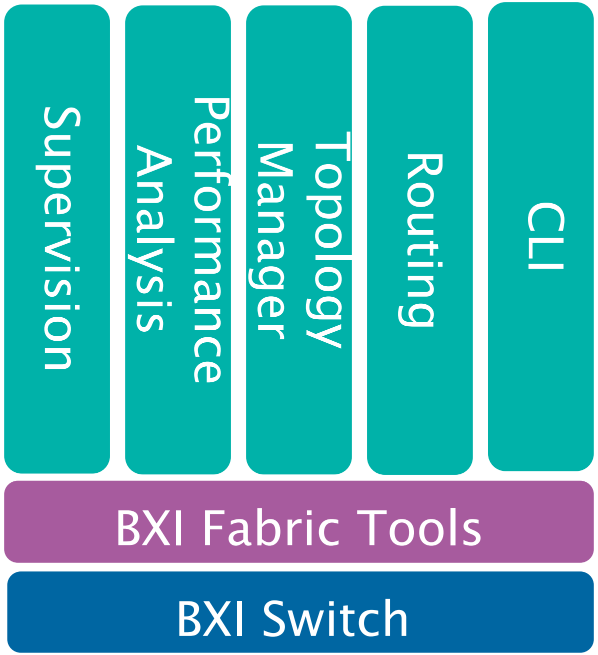
Backup

BXI Software Suite

Overview



BXI Fabric Management



Supervision

Performance Analysis

Topology Manager

Routing

CLI

BXI Fabric Tools

BXI Switch

- 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 fabric management

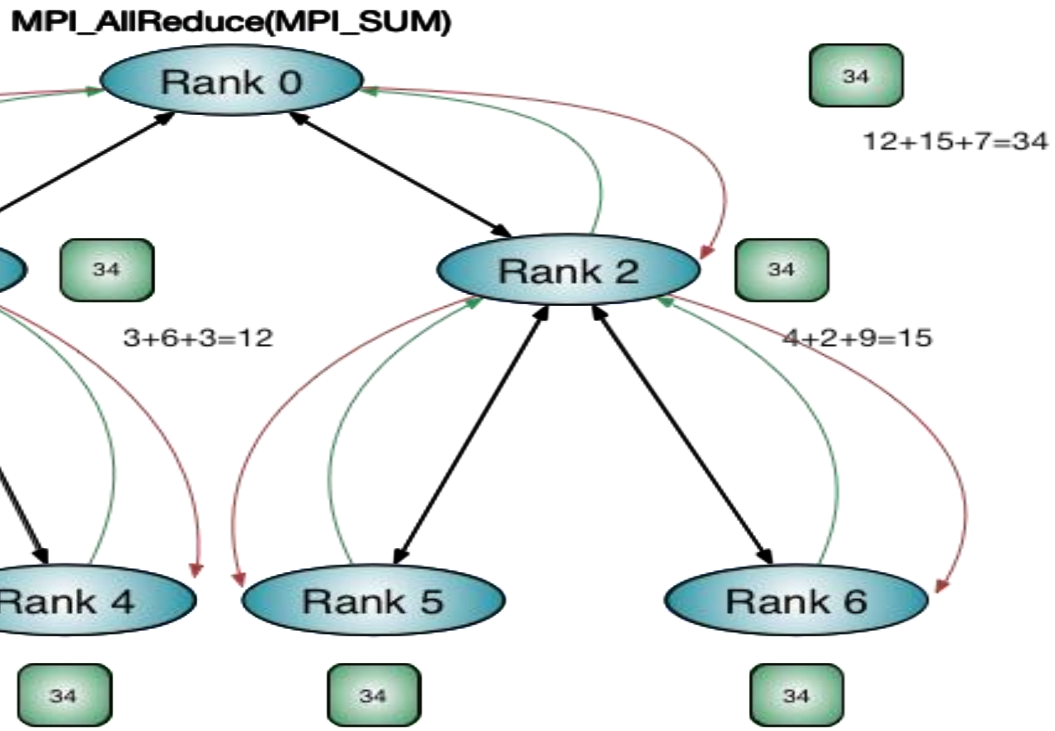
Fabric Monitoring/Profiling

- Administrators, support and users need information about fabric status
- Each class of users needs different kind of information with different purposes
  - **Administrator**: what happened? What is the status?
  - **Support**: is it working? Why not? Low level debug
  - **User**: how does my application use the interconnect? Reproducibility
- BXI provides counters and sampling:
  - Probes (set of counters + frequency) can be configured
  - 4 probes per switch maximum at the same time
  - Frequency up to **1Hz**
  - Threshold setting for monitoring warning
- Dynamic routing solutions ensure interconnect’s reliability and performances:
  - Topology specific and topology agnostic routing
  - Distributed routing between management nodes and switch embedded solutions

BXI - Offloading MPI communication

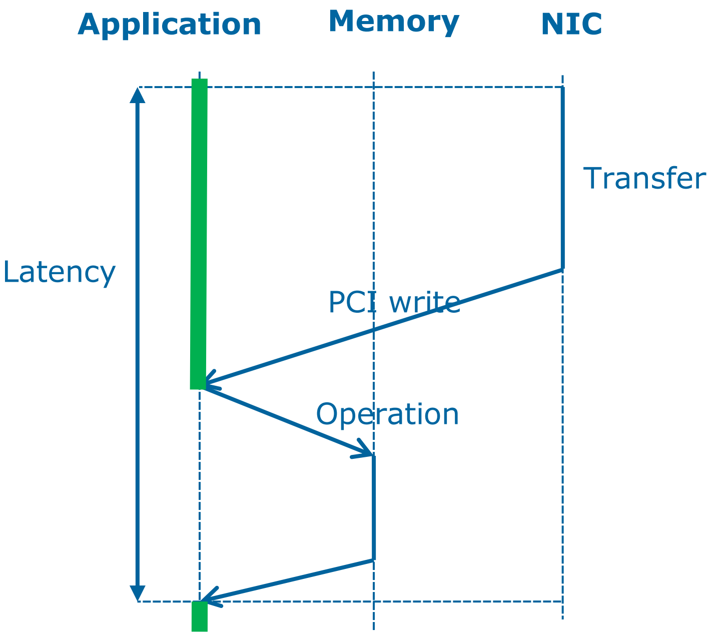
Ex. MPI_AllReduce principle

All in Hardware !!

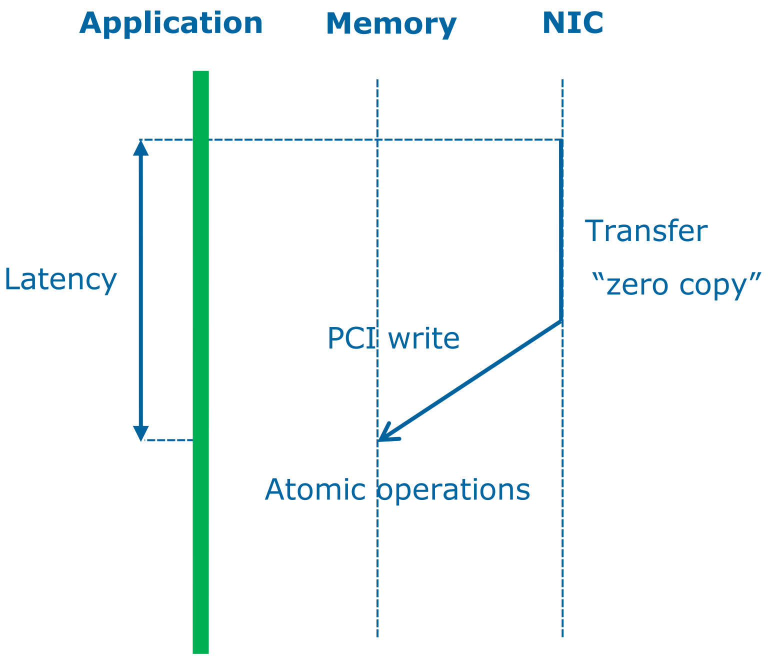


Application Bypass

Without Application Bypass



With Application Bypass (BXI)


Application
Memory
NIC
Transfer
"zero copy"
Latency
PCI write
Atomic operations