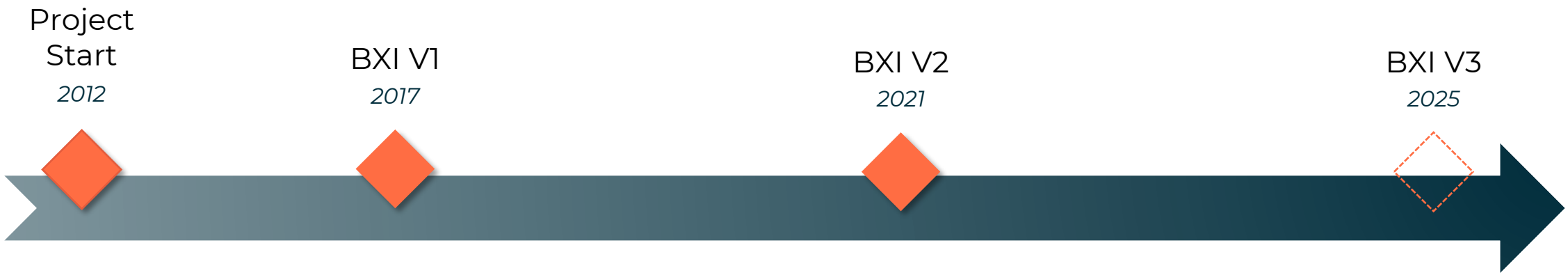
EVIDEN

What exciting features does the BXI v3 interconnect network offer to applications ?

Grégoire Pichon
Eviden / HPC SW R&D
11/2024

Quick Recap of the BXI Story

BullSequana eXascale Interconnect



Jun. 2018 **Tera-1000-2**
11.9 Tflops, rank: 14th

Tera-1000-2 - Bull Sequana X1000, Intel Xeon Phi 7250 68C 1.4GHz, Bull BXI 1.2, EVIDEN
Commissariat a l'Energie Atomique (CEA)
France

Jun. 2018 **JOLIOT-CURIE KNL**
1.3 Tflops, rank: 151th

JOLIOT-CURIE KNL - Bull Sequana X1000, Intel Xeon Phi 7250 68C 1.4GHz, Bull BXI 1.2, EVIDEN
CEA/TGCC-GENCI
France

Nov. 2021 **CEA-HF**
23.2 Tflops, rank: 14th

CEA-HF - BullSequana XH2000, AMD EPYC 7763 64C 2.45GHz, Atos BXI V2, EVIDEN
Commissariat a l'Energie Atomique (CEA)
France

Jun. 2024 **CEA-HE**
57.1 Tflops, rank: 17th

CEA-HE - BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail BXI v2, EVIDEN
Commissariat a l'Energie Atomique (CEA)
France

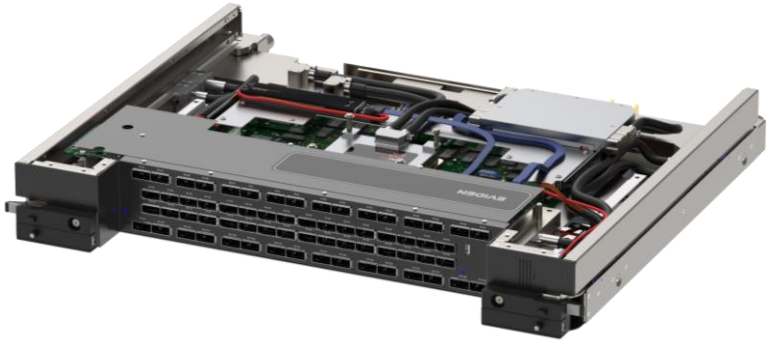
EVIDEN

BXI v3 Overview

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101175702. The JU receives support from the European Union's Horizon Europe research and innovation programme and France, Greece, Italy, Norway.

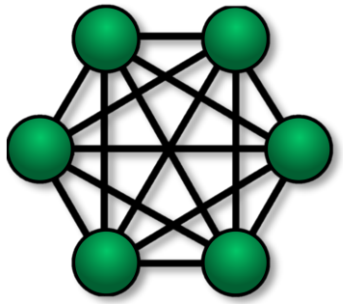
NIC

400 Gb/s bandwidth
220 Mmsg/s rate
1 µs latency
Many offload features
PCIe Gen5 x16
Mezzanine and PCIe boards



SWITCH

64 ports at 800 Gb/s
(128 ports at 400 Gb/s)
Air- and liquid-cooled
Full crossbar
200 ns traversing latency



FABRIC

Up to 64K NICs per fabric
Up to 128 interconnected fabrics
Fat-tree & Dragonfly+ topologies
Native interoperability with Ethernet frames
Adaptive routing, congestion mngt

2 Communication Flows
IP/Ethernet communications for Standard Ethernet services
Portals/Ethernet communications for HPC/AI applications and services

BXI v3 NIC Offloading Features

OFFLOADING

DMA
direct access to host memory through PCI
data reads and writes for payload transfers

Matching
hardware matching at message reception
MPI tags + advanced selection with match-bits/ignore-bits, source matching, size matching

Address Translation
virtual-to-physical address translation based on ATS/PRI requests
no need to register/pin memory before network transfers

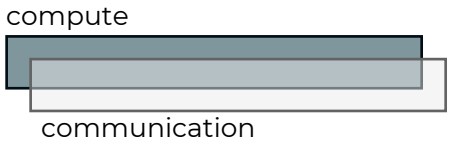
Large Command Queues
ability to post millions of commands into NIC queues
autonomous and fair progress of commands

Atomic and Triggered Operations
compute operations treated by atomic unit hardware bloc
data transfers triggered by network events
handle MPI reduce and collective operations

TCP Offload
checksum, segmentation (GSO), aggregation (GRO)

Reduce CPU workload related to network communications

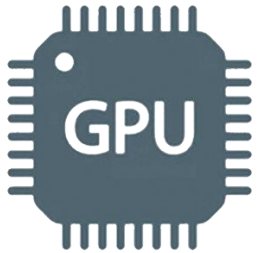
Improve compute and communication overlap



EVIDEN

ECLAT

BXI v3 features for Artificial Intelligence



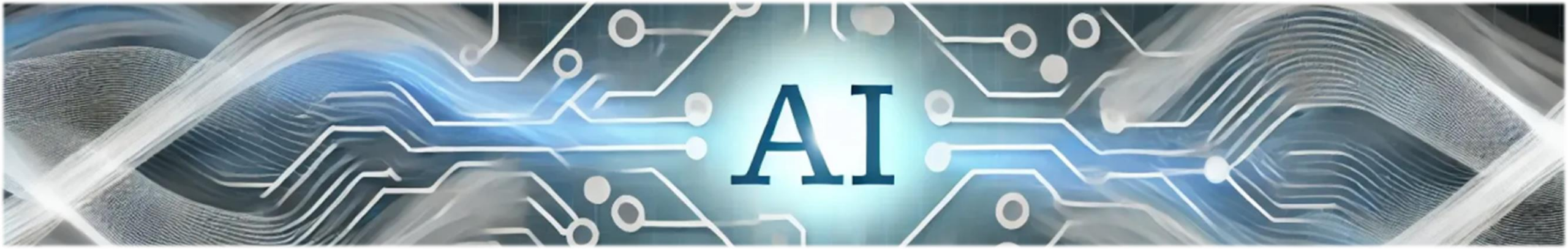
Performance
Direct transfers
from/to GPU memory
over the network

NVIDIA CUDA AMD ROCm

Software Stack Integration
Transport plugins for
Collective Communication Libraries
(NCCL, RCCL)

NIC Atomic Operation Unit
fp16/bfloat16

Optimized Reduce operations
Native support of half-precision
floating-point data types



BXI v3 Security Features

3 Traffic Classes

Independent communication handling/routing
in the NIC and in the Fabric
(Compute 1, Compute 2, Service)

Secured Firmware

Signed firmware
Secure boot
Both NIC and Switch firmware

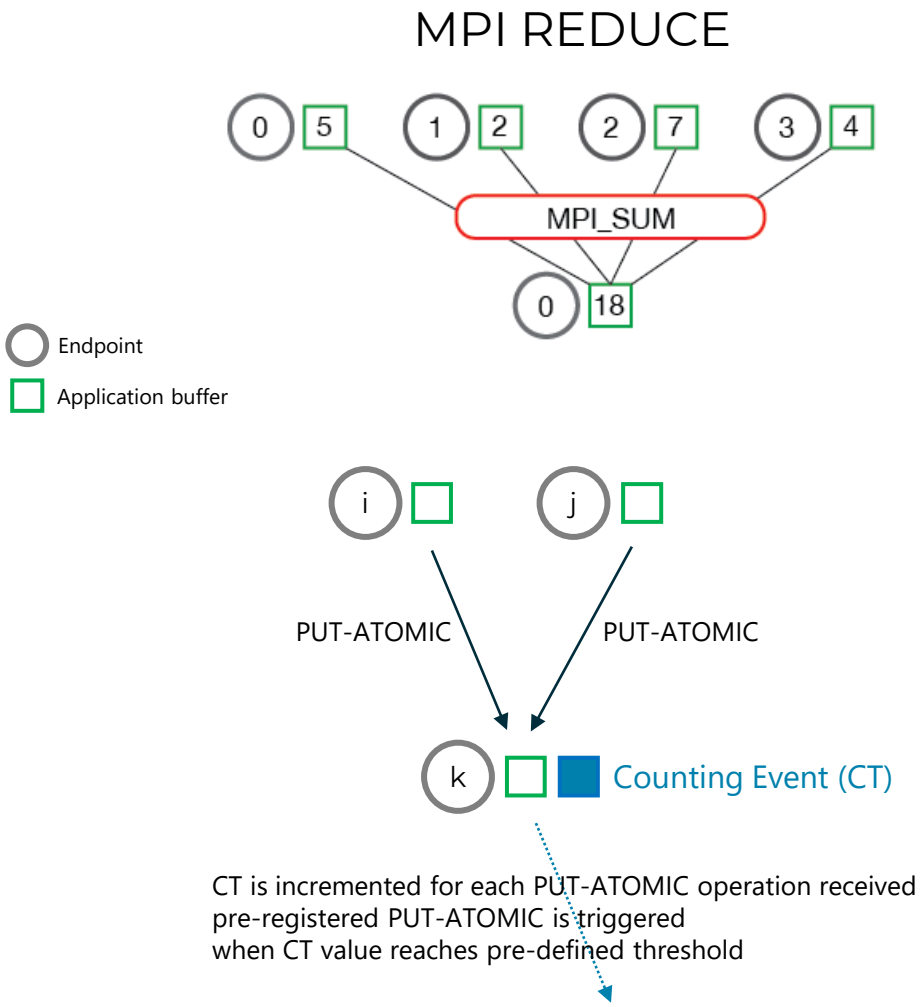
Communication Access Control

Per application / per user access control
Permission check at message reception
Based on Portals Usage-ID

Network Isolation

Traffic filtering at switch port level
Prevent abusive usage of Fabric resources
Based on pkey / Virtual Network

BXI v3 Collective Operations Offloading



Handling in MPI

- build a tree for the ranks in the MPI communicator
- break down collective into point-to-point operations
- post operations to the NICs using triggered operations
- return to the application for further computing

Handling in the NIC

- record delayed point-to-point operations
- process communication operations (put, get) and atomic operations (sum, max, min, prod)
- track network operation completions with “counting events” (CT)
- trigger next operations when condition on CT is reached

Application should

- Use non-blocking MPI collective primitives
- Continue simulation/model processing while communications autonomously progress

EVIDEN

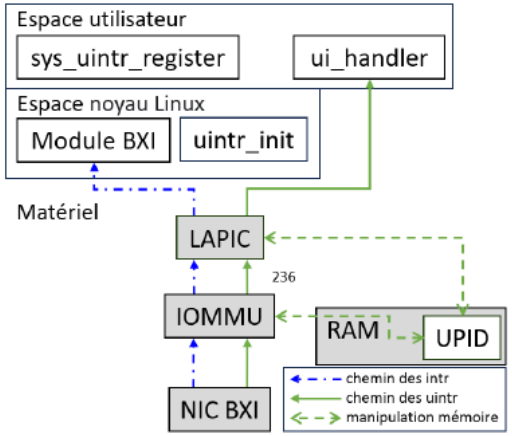
ECLAT

Event Notifications With User Interrupts

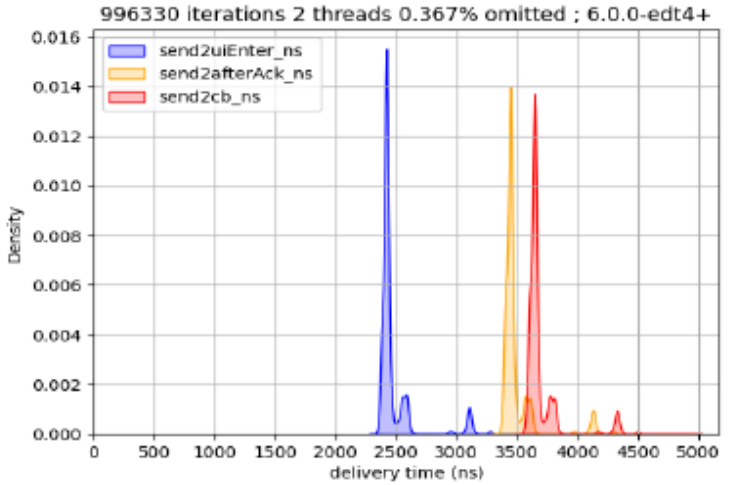
- Experimental work based on a new feature of the Intel processors (UINTR)
- Direct callback invocation from the device to the user-space : full kernel bypass

Benefits

- Prevent active polling and wasted CPU cycles
- Improve compute/communication overlap



User Interrupt Delivery Path (in green)



Latency between PtlPut and callback (in red)

Notification Mechanism	Latency
Active polling	1.3 µs
User interrupts	**3.7 µs**
Classical interrupts	11.3 µs

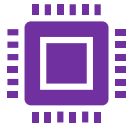
Half round-trip latency comparison

Charles Goedefroit. Interruptions en espace utilisateur délivrées par la carte réseau BXI. Compas 2024 - Conférence francophone d'informatique en Parallélisme, Architecture et Système, Jul 2024, Nantes, France. (hal-04693786)

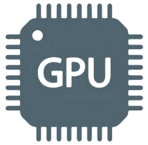
BXI v3 is design to run HPC/AI intensive workloads on Exascale systems

High Performance Level

Standard-Ethernet & HPC Communications



Offloading & Compute/Comm overlap



GPU Support

Full Software Stacks (MPI, NCCL/RCCL Lustre, OFI, ...)

Does it cover some of the SKA needs ?
Are there other requirements to consider ?

EVIDEN

Questions ?

EVIDEN

For more information please contact:
Grégoire Pichon
gregoire.pichon@eviden.com