Bull Exascale Interconnect (BXI)

un nouveau réseau pour le calcul de haute performance

Jean-Pierre Panziera

11-10-2016
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

Fast Interconnect

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

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

Multiple storage tiers

much more parallelism
HPC Interconnect

- **performant**
  - low latency
  - high message rate
  - high bandwidth

- **scalable**
  - 10,000s 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 \( \sim 1 \mu s \)
- Issue rate >100 Mmsg/s
- BXI Link 100 (4x25) Gb/s

switch ASIC

- 48 ports BXI Link
- 9600 Gb/s bandwidth

Lutetia

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

#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)
BXI offloading MPI communication in HW

With HW offload
BXI MPI AllReduce using Triggered and Atomic operations

\[ \text{TrigPut}(\text{result}) + 1 \]

\[ \text{CT} \]

\[ \text{ME}(\text{Ready}) + 1 \]

\[ \text{TrigAtomic}(\text{Data}) \]

\[ \text{Ch1} \]

\[ \text{Atomic}(\text{Data}) \]

\[ \text{TrigPut}(\text{result}) \]

\[ = 3 \]

\[ \text{CT} \]

\[ \text{ME}(\text{Result}) + 1 \]

\[ \text{P} \]

\[ \text{TrigPut}(\text{result}) \]

\[ \text{Put}(\text{Ready}) \]

\[ \text{P} \]

\[ \text{ch1} \]

\[ \text{ch2} \]

\[ \text{Put}(\text{Ready}) \]

\[ \text{Atomic}(\text{Data}) \]
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 x25,278 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 switch

Fabric Management Software

Mngt Node

10GbE

Ethernet Management Network

Fans, Sensors, ...

ARM µc

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

Quintin, Vignéras; Fault-Tolerant Routing for Exascale Supercomputer: The BXI Routing Architecture. HiPINEB’15
Quintin, Vignéras; Transitivity Deadlock-Free Routing Algorithms . HiPINEB’16
BXI Software compute stack
BXI

PCI adapter card and 48p standalone switch

- Optical cables (100Gb/s)
- BXI port
- BXI NIC ASIC
- 1U
- 48 port BXI switch
- 48p BXI switch ASIC
- Optical modules (PODs)
- Redundant Power Supplies
- Redundant Fans
“Sequana” – Embedded interconnect

Fast Interconnect layout
Sequana cells interconnection

Fat-Tree

L3 switches

Direct connections

... 48x ...

L3 switches
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
- First BXI system installed in Q4-2016
- Large BXI deployment (8+K nodes system) in 2017
Questions?

$10^{18}$

Bull exascale program

Bull

atos technologies