Processor Network Interface Working Group

Keith Underwood (Intel)
Dhabaleswar K. (DK) Panda (OSU)

Assumptions
(Based on Earlier presentations and Discussions)

- Programming Models
  - Message Passing (MPI), will be prevalent
    - Includes MPI + OpenMP model too
  - PGAS models, will be there
- Number of sockets
  - ~32K
- Multiple cores/socket
  - 2011: 8, 16-32
  - 2015: 32, 64-128
  - 2019: 128, 256-512

System Interconnects

	2011		2015		2019	
System Size Sockets Peak PF TF/Socket	32,768 32 1.0		32,768 200 6.1		32,768 800 25.0	
	Expect	Want	Expect	Want	Expect	Want
NIC B/W (B/F)	0.01 - 0.1	1.0	0.005 - 0.03	1.0	0.025 - 0.25	1.0
Link B/W (B/F)	0.01 - 0.1	1.0	0.005 - 0.03	1.0	0.025 - 0.25	1.0
MPI Latency (ns)	750 - 1500	500	500 - 1000	400	400 - 750	300
MPI Throughput (M Msg/s)	20	50	80	300	300	1200
Load/Store (M Msg/s)	75	400	150	1,600	300	6400
Load/Store Latency (ns)	300	100	300	100	300	100

Sandia National Laboratories

Clarifications on the Numbers in the Previous Slide (during discussion)

- Numbers are indicated as per socket
- The "want" column means you push hard to get things ahead (and it is what is needed to get each machine balanced and stable). The "expect" column is without extra work.
- Perhaps "want" is more like "desired".

Exascale Issues for Processor Network Interface Major Categories

- Performance
- Power
- Reliability and Resilience
- Scalability
- Manufacturability
- Cost

Typical Architectural Support in Network Interface

- Point-to-point
  - Two-sided
  - One-sided
    - Load/Store, RDMA, Active Message
  - Atomics
- Fine-grain pt-to-pt communication with low overhead
- Collectives
  - Allreduce, Broadcast, Reduce, AlltoAll, ..
  - Optimization for small messages
- Datatype Support
- Memory Registration
- Flow Control
- Reliability
- Resilience and Fault Tolerance
- Quality of Service (QoS)
- Instrumentation for Performance Monitoring
- Virtualization

Designing Processor Network Interface: Major Dimensions

- On-chip NI and Off-chip NI
  - Number of NIs per node
    - Balance between off-chip vs. on-chip
- On-loading Vs. Off-loading Vs. Hybrid

Challenges (Based on the discussion)

- On-chip NI + Off-chip NI Integration
  - Extending the NI from NOC to system area network
  - How the processor vendors will do integration to the NIC is a big challenge
  - Concurrency is a big challenge
- Converged NI
  - To handle memory, I/O and storage traffic
  - Match with memory subsystem performance
  - Set of small NICs vs. a bigger NIC
  - Keeping an eye on developments such as PCI Express Gen3
  - Adding additional semantics/operations (such as atomics)
  - Enhanced NIC design
  - Not having the interface NIC coherent and TLB coherent means you violate two principles essential for PGAS languages. Need closer integration.
  - Better processor-NIC interface (to allow efficient ordering and concurrency)
  - Cache injection technique with message passing
  - If data is in cache, update it, do not flush it

Challenges (Based on the discussion)

- Collectives
  - Different on-loading/off-loading solutions for large-scale systems
  - Optimal solution is desired without increasing the NIC complexity
- Fine-grain synchronization
  - Good support from Processor NI to node-level NIC
  - Support for efficient handling of out-of-order messages at the NIC (multipath, fence)
- On-Loading NI functionality to processor
  - If there is a possibility, Ethernet (not IP) will be the first one to do so
  - May not be good for small message transfers
- Connection Management
  - No connection-oriented protocol for Exascale systems
- End-to-End Reliability Support
  - Critical for Exascale systems
  - Efficient designs to keep network state information at appropriate places not to increase the NI complexity significantly
- Light-weight communication protocol
  - Designing extremely light-weight protocol
  - Put MPI on top of it with low overhead

Additional Observations/Comments

- Latency targets for MPI and Load/Store for 2019 (under Want column) are aggressive and can not be achieved in a realistic manner

Characterization of Challenges

- Probability that the challenge will not be solved by relying on current technology trends: high, med, low.
- Impact that the lack of a solution for this challenge will have on the ability of the HPC community to build an Exascale computer by 2016: high, med, low. That is, HIGH means that if we don't have a solution for this problem, there is no workaround solution that will allow us to build the Exascale system.
- Approximate NRE cost for a solution: high (greater than $15M), med ($5-10M), low (less than $5M).

Tasks (for Tuesday Afternoon)

- Need to prioritize the list of challenges
- Characterize the challenges
  - Probability, Impact and Cost