Processor Network Interface Working Group

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

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

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

Designing Processor Network Interface: Major Dimensions

- Architectural (Hardware and Software)Support
- On-chip NI and Off-chip NI
- On-loading Vs. Off-loading Vs. Hybrid

Prioritizing the Challenges

- (1) On-chip NI + Off-chip NI Integration
  - High, High, High (>$100M)
  - 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
  - Cache injection technique with message passing
    - If data is in cache, update it, do not flush it

  - (2) Light-weight communication protocol
    - High, Medium, Low (<$10M)
  - Designing extremely light-weight protocol
  - Put MPI on top of it with low overhead

Prioritizing the Challenges

- (3) Enhanced NIC design
  - High, High, Medium ($10-$100M)
  - 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)
- Better processing engines for MPI and PGAS (message rate)
- Virtualization of NIC resources

- (4) End-to-End Reliability Support
  - Medium, Medium, Low (<$10M)
- Critical for Exascale systems
- Efficient designs to keep network state information at appropriate places not to increase the NI complexity significantly
- Checksum support

Prioritizing the Challenges

- (5) Collectives
  - High, Low, Low (<$10M)
- Different on-loading/off-loading solutions for large-scale systems
- Optimal solution is desired without increasing the NIC complexity

- (6) Fine-grain synchronization
  - High, Medium, Low (<$10M)
- Good support from Processor NI to node-level NIC
- Support for efficient handling of out-of-order messages at the NIC (multipath, fence)

- (7) Connection Management Scalability
  - High, High, Medium ($10-100M)
  - No connection-oriented protocol for Exascale systems

Prioritizing the Challenges

- (8) Converged NI
  - Medium, Medium, Medium ($10-100M)
- 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)

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