Scalable Hierarchical Aggregation Protocol (SHArP): A Hardware Architecture for Efficient Data Reduction

Gil Bloch

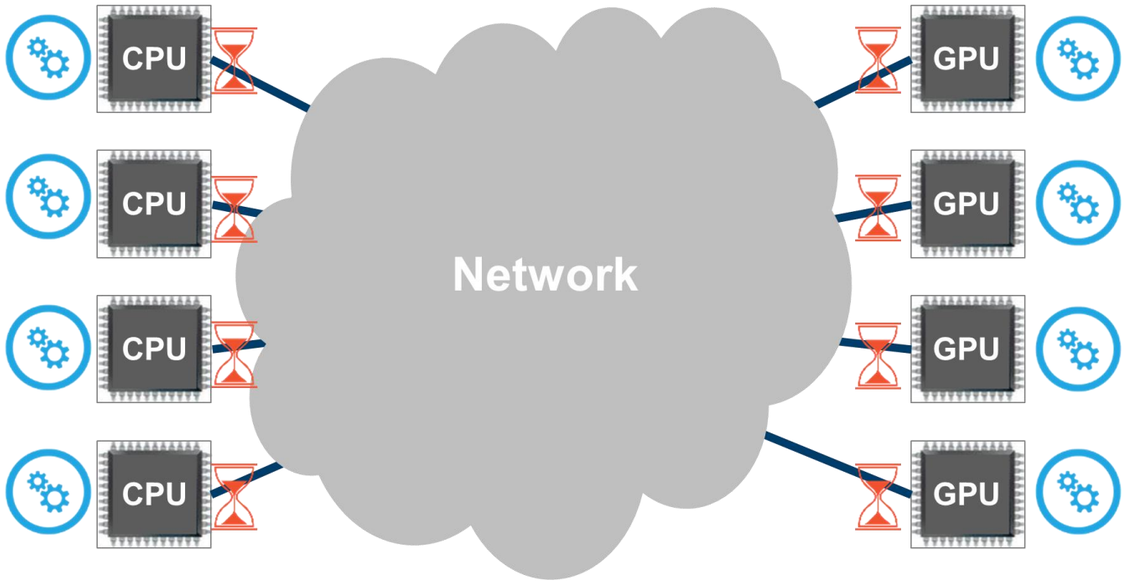
Supercomputing Frontiers Singapore 2017

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

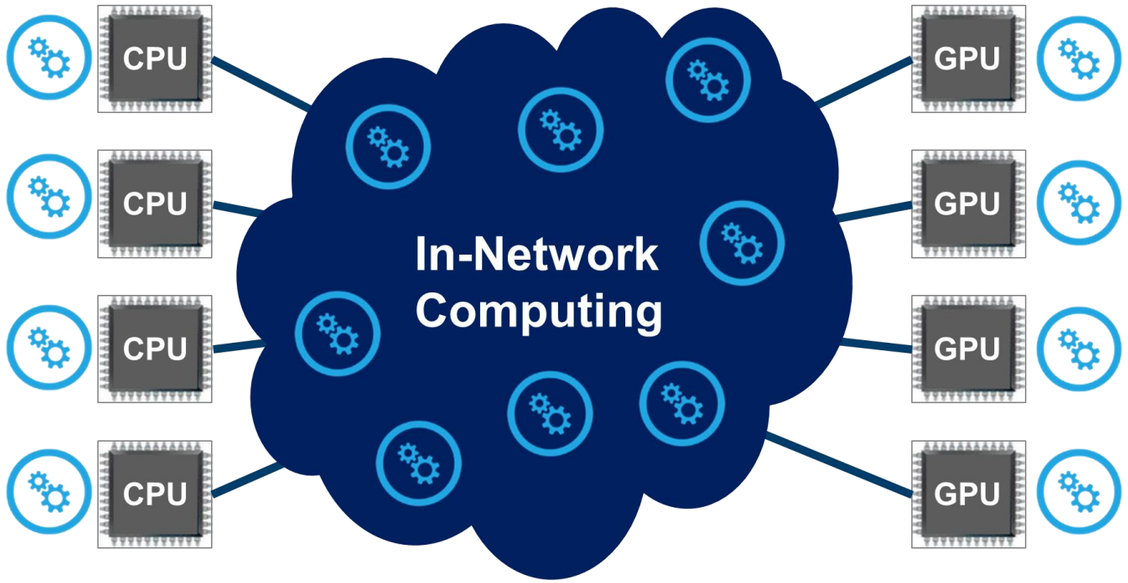
SHArP is a component of the Intelligent Interconnect

CPU-Centric



Limited to Main CPU Usage
Results in Performance Limitation

Data-Centric



Creating Synergies
Enables Higher Performance and Scale

Must Wait for the Data
Creates Performance Bottlenecks

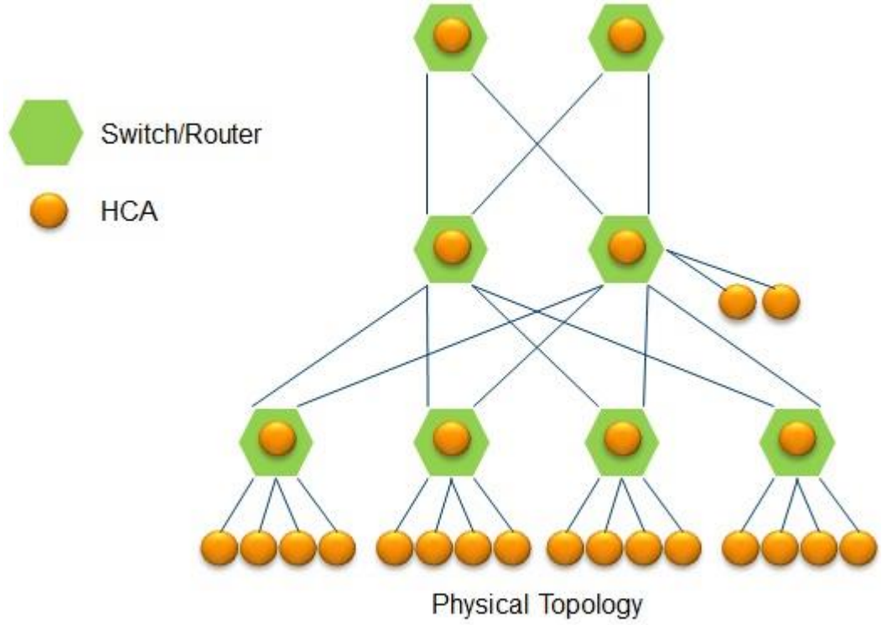
Work on The Data as it Moves
Enables Performance and Scale

Scalable Hierarchical Aggregation Protocol (SHArP) : Goals

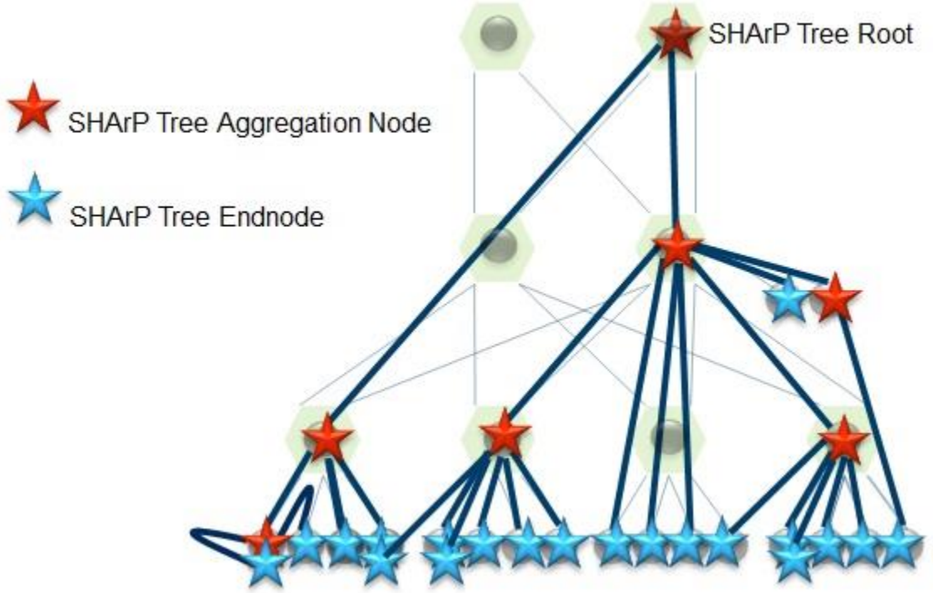
- Optimize data aggregation (reductions)
- Offload these operations to the network
- Minimal data paths
- Virtual topology overlaid over physical topology
- Protocol sits above network transport layer

Scalable Hierarchical Aggregation Protocol (SHArP) Architecture

Network Hardware
Topology (Physical Tree)



SHArP Logical
Tree



SHArP Implementation Characteristics – SwitchIB-2

- Reliable Scalable General Purpose Primitive, Applicable to Multiple Use-cases
  - In-network Tree based aggregation mechanism
  - Large number of groups
  - Multiple simultaneous outstanding operations
- Scalable High Performance Collective Offload
  - Barrier, Reduce, All-Reduce, Broadcast and more
  - Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
  - Integer and Floating-Point, 32 / 64 bit
  - Repeatable results
- Topology Agnostic
  - Virtual topology allows mapping onto multiple physical topologies – no need for aggregation nodes to by directly connected
- Hardware level enforcement of resources

Allreduce Latencies of Various Implementations (uSec)

Message Size [B]	HPC-X SHArP Based	HPC-X Host Based	MVAPICH-2 Host Based
0(barrier)	2.91	5.25 (80.4%)	11.47 (290%)
8	2.83	6.01 (112%)	11.90 (320%)
16	2.79	5.95 (56.9%)	11.27 (304%)
32	3.94	6.19 (57.1%)	11.69 (197%)
64	3.02	6.80 (125%)	12.03 (298%)
128	4.30	7.69 (78.8%)	14.08 (227%)

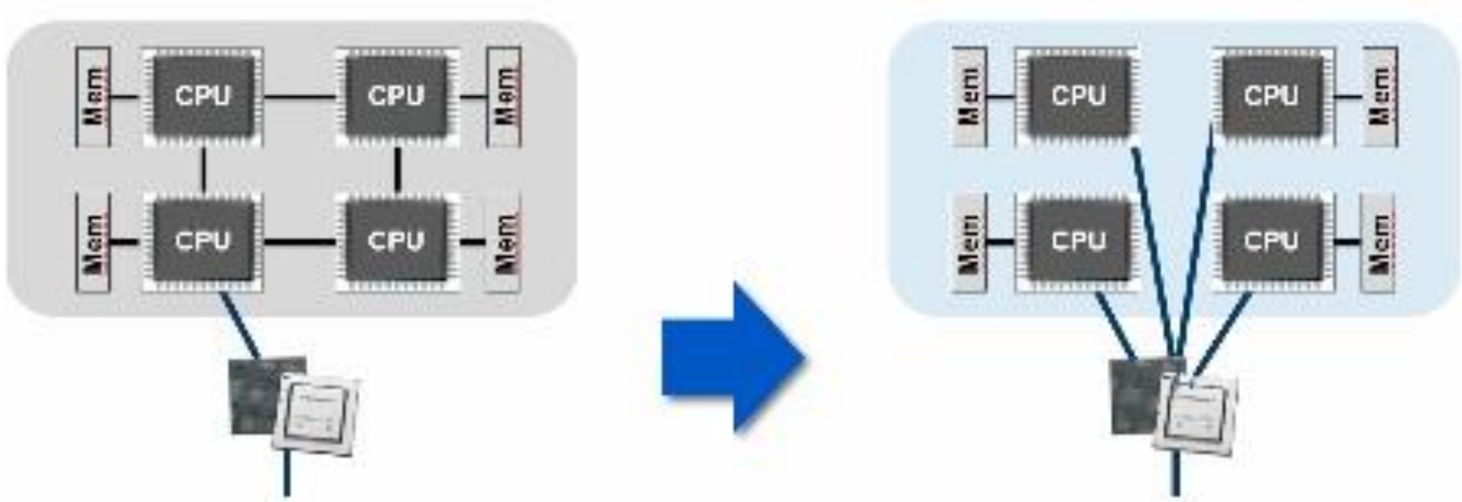
Pipelining SHArP Operations (uSec)

- SwitchIB-2 support limited vector size
  - 256 bytes
  - 256 outstanding operations
- Implementing pipelining of reduction operations

Message Size [B]	SHArP based	Host Based	SHArP improvement factor
8	2.76	5.82	2.11
16	2.76	5.91	2.14
32	2.86	6.04	2.11
64	3.01	6.76	2.25
128	3.24	7.37	2.27
256	3.50	8.99	2.57
512	4.06	11.11	2.74
1024	5.49	18.04	3.29
2048	8.44	33.61	3.98
4096	14.48	46.93	3.24

Multi Socket Server

- Many servers implement more then one CPU socket
  - Connected with the CPU coherent bus
- Using single network adapter for multiple CPUs
  - Connected into one of the PCIe slots
  - Closer to one of the CPUs, but require to traverse the CPU-CPU bus for a remote CPU
- With Multi-Host technology
  - Connected to multiple CPUs
  - Short distance to all CPUs



OpenFOAM Application Performance – IcoFOAM test case

Number of Processes	HPC-X with HCOLL no SHArP	Default Open MPI	HPC-X with HCOLL with SHArP
28	1892	1892(0%)	1892 (0%)
56	832	866 (4.09%)	824 (1.05%)
112	426	469 (9.96%)	424 (0.59%)
224	249	285 (14.3%)	248 (0.39%)
448	159	194 (22.4%)	155 (2.50%)
896	127	167 (31.4%)	121 (4.75%)
1792	118	180 (52.6%)	116 (1.65%)
2688	136	206(51.3%)	127 (7.18%)
3584	174		151 (15.2%)

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™