The Active Network

Rich Graham

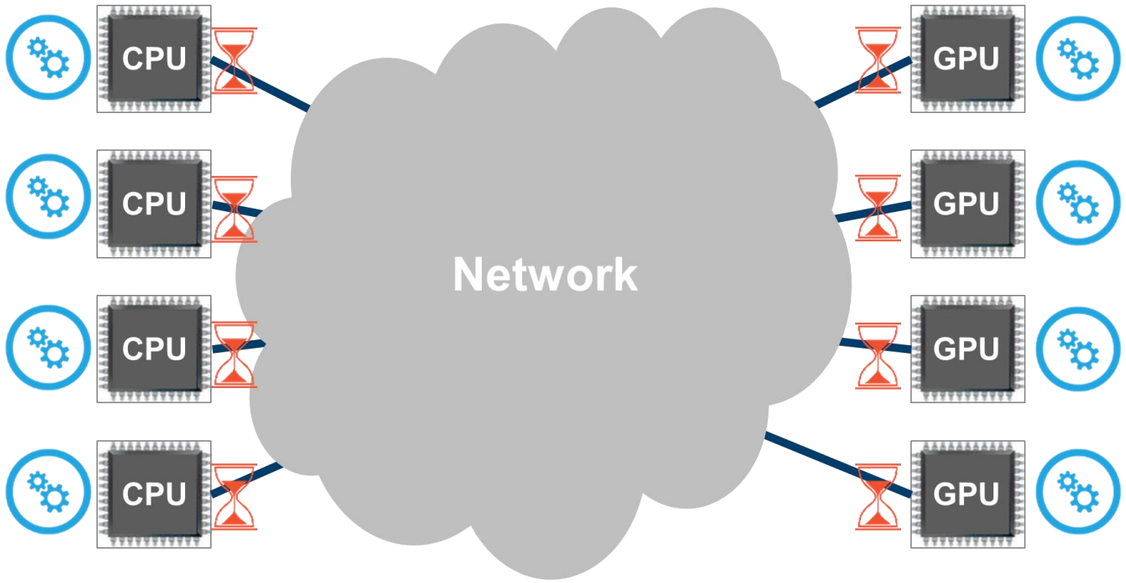
CCDSC – October 2016

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

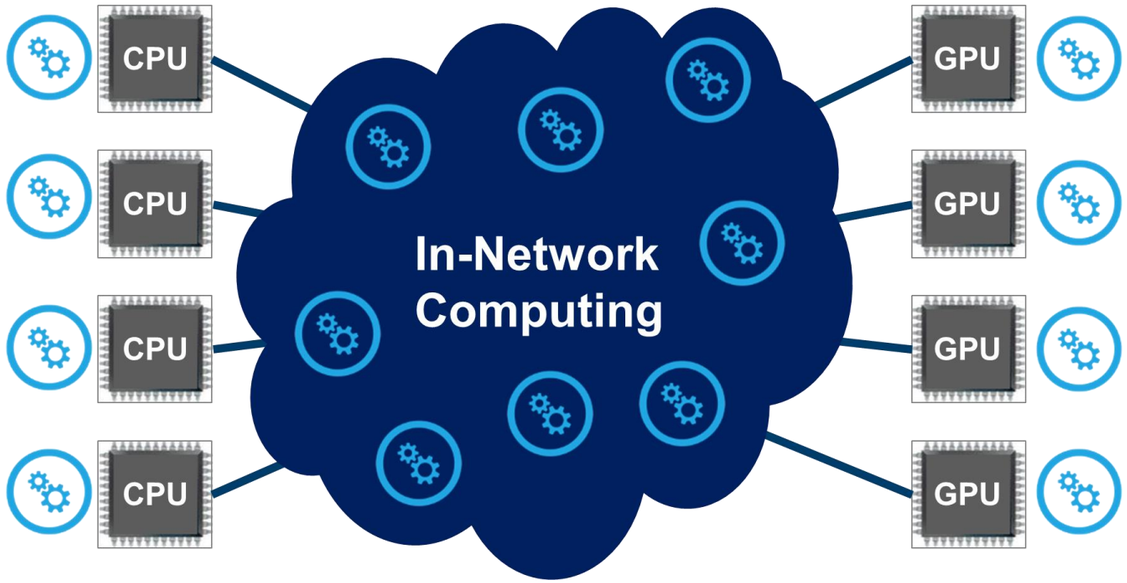
The Intelligent Interconnect to Enable Exascale Performance

CPU-Centric



Limited to Main CPU Usage
Results in Performance Limitation

Co-Design



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

Highest-Performance 100Gb/s Interconnect Solutions

Adapters — ConnectX®-5

100Gb/s Adapter, 0.6us latency
200 million messages per second
(10 / 25 / 40 / 50 / 56 / 100Gb/s)

Switch — SwitchIB™-2

36 EDR (100Gb/s) Ports, <90ns Latency
Throughput of 7.2Tb/s
7.02 Billion msg/sec (195M msg/sec/port)

Switch — Spectrum™

32 100GbE Ports, 64 25/50GbE Ports
(10 / 25 / 40 / 50 / 100GbE)
Throughput of 6.4Tb/s

Interconnect — LinkX™

Transceivers
Active Optical and Copper Cables
(10 / 25 / 40 / 50 / 56 / 100Gb/s)

VCSELs, Silicon Photonics and Copper

Software — HPC-X™

MPI, SHMEM/PGAS, UPC
For Commercial and Open Source Applications
Leverages Hardware Accelerations

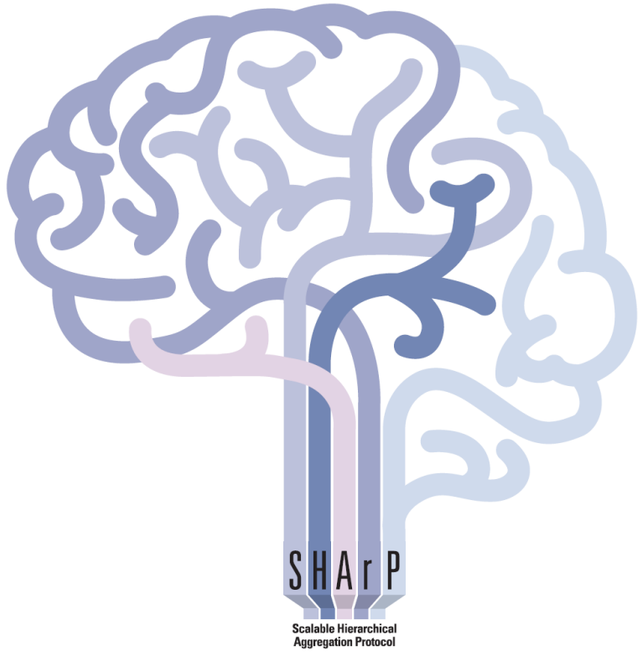
SwitchIB-2: Switch-Based Data Reduction

Scalable Hierarchical Aggregation Protocol

Accelerating Artificial Intelligence and HPC applications

Accelerating HPC Applications

- Significantly reduce MPI collective runtime
- Increase CPU availability and efficiency
- Enable communication and computation overlap

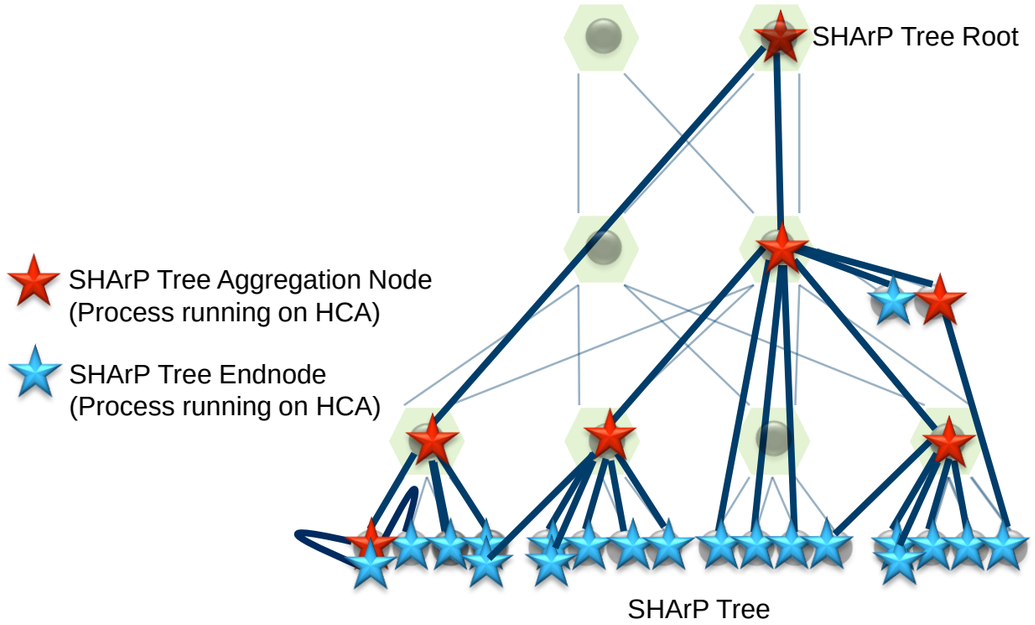


Enabling Artificial Intelligence Solutions to Perform Critical and Timely Decision Making

- Accelerating distributed machine learning
- Improving classification accuracy
- Reducing the number of batches needed for asynchronous training

Scalable Hierarchical Aggregation Protocol – MPI and OpenSHMEM

- Reliable Scalable General Purpose Primitive, Applicable to Multiple Use-cases
  - In-network Tree based aggregation mechanism
  - Large number of groups
  - Multiple simultaneous outstanding operations



Accelerating HPC applications

- Scalable High Performance Collective Offload
  - Barrier, Reduce, All-Reduce
  - Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
- Significantly reduce MPI collective runtime
- Enable communication and computation overlap

Accelerating Machine Learning Applications

- Prevent the ***many-to-one*** Traffic Pattern

SHArP
Scalable Hierarchical Aggregation Protocol

Current Performance Data* – OSU Allreduce 1PPN, 128 nodes

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

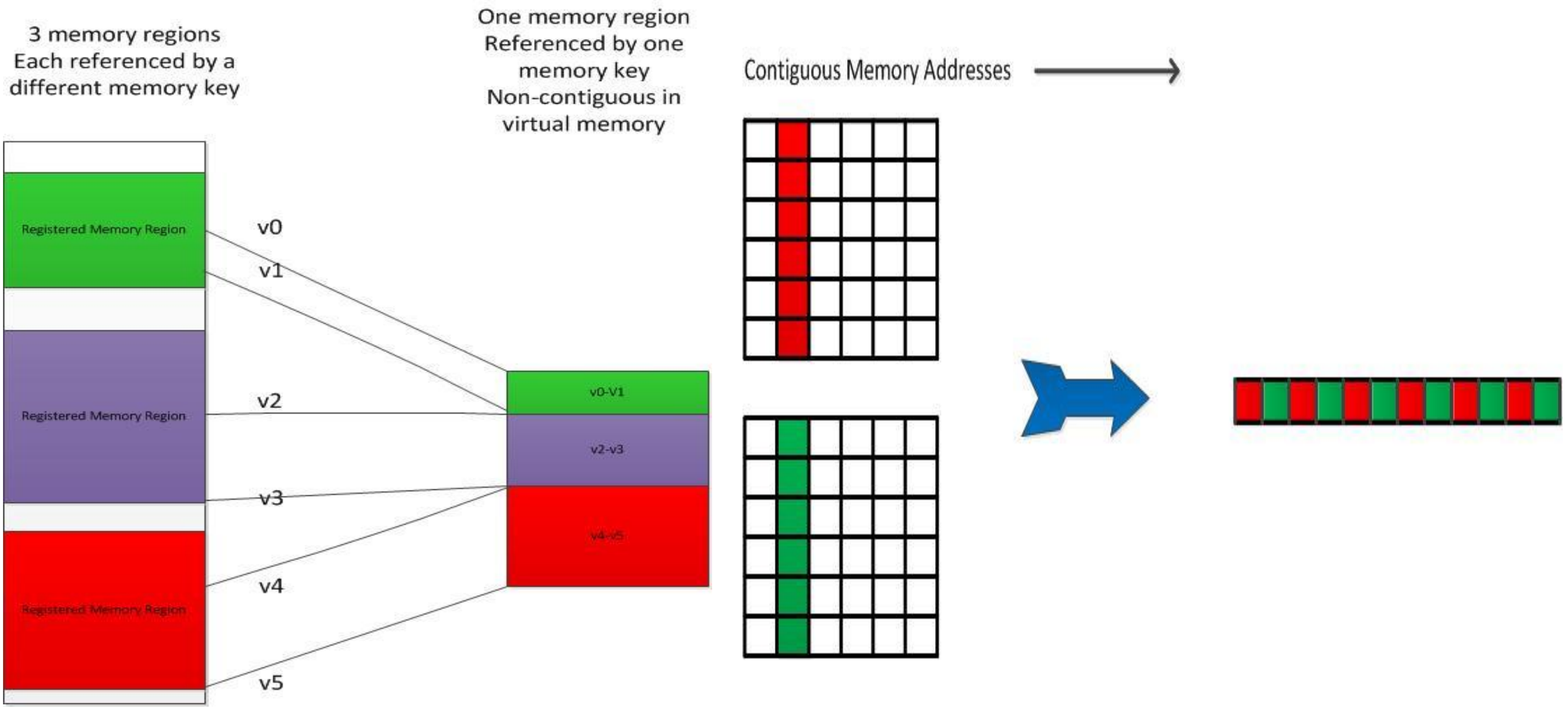
***Initial Numbers**

Mellanox Infiniband Scatter/Gather Capabilities

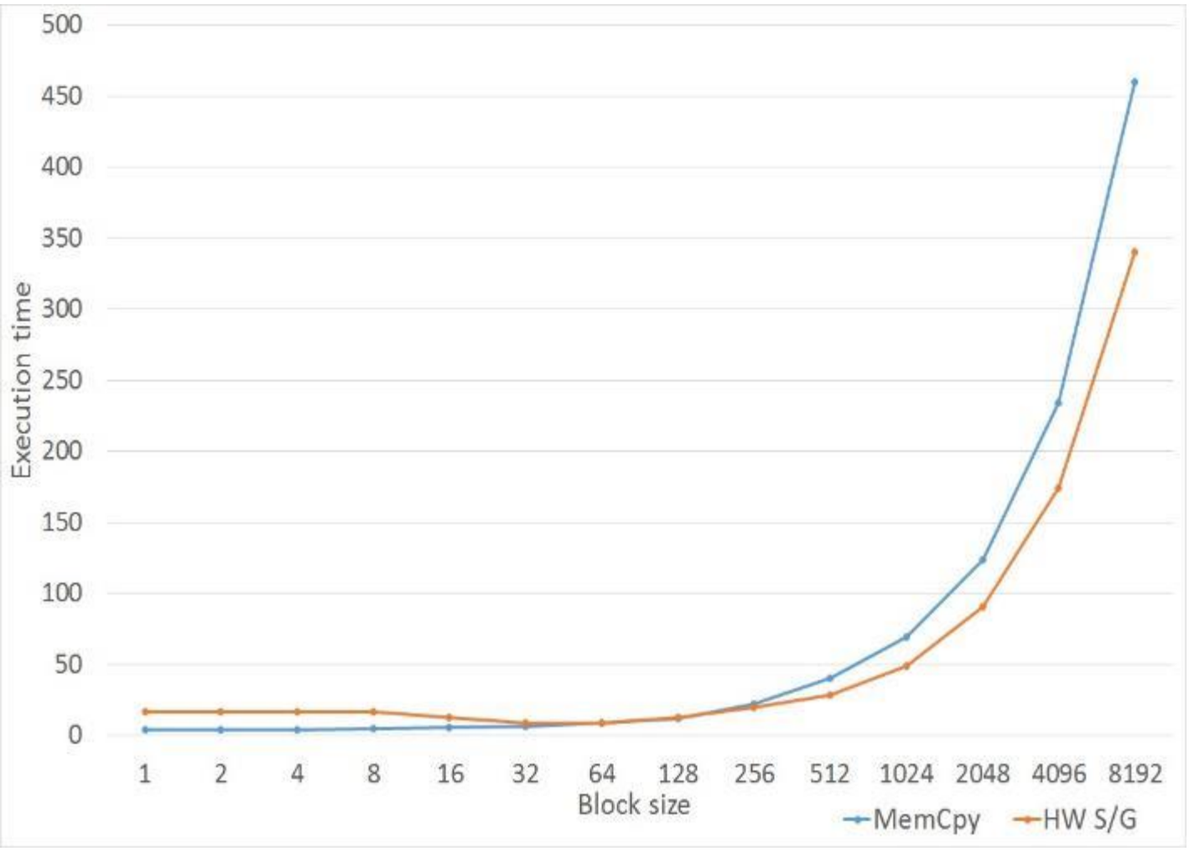
Scatter-Gather HCA Capabilities

- Scatter-Gather capabilities based on memory keys describing a non-contiguous memory region
  - Two types of memory keys
    - IOVec – like region description, just like Scatter-Gather Entry (SGE) list
    - Regular structure representation – base pointer, element length, stride, repeat count.(may interleave data from several independent buffers)
  - No CPU involvement in either gather or scatter (eliminate memory copies)
  - Support for non-contiguous memory asynchronous data transfers
- Memory-key based implies that any-place a memory key is used, not-contiguous data layouts may be use
  - Can do a put (form a non-contiguous region) to a non-contiguous region
- Memory descriptors are created by posting WQE's to fill in the memory key
  - If posted before operation to use the memory key, no need to wait CQE before posting the data operation using the memory key
- Feature supported since Connect-IB HCA (Connect-IB, ConnectX-4, ConnectX-5)

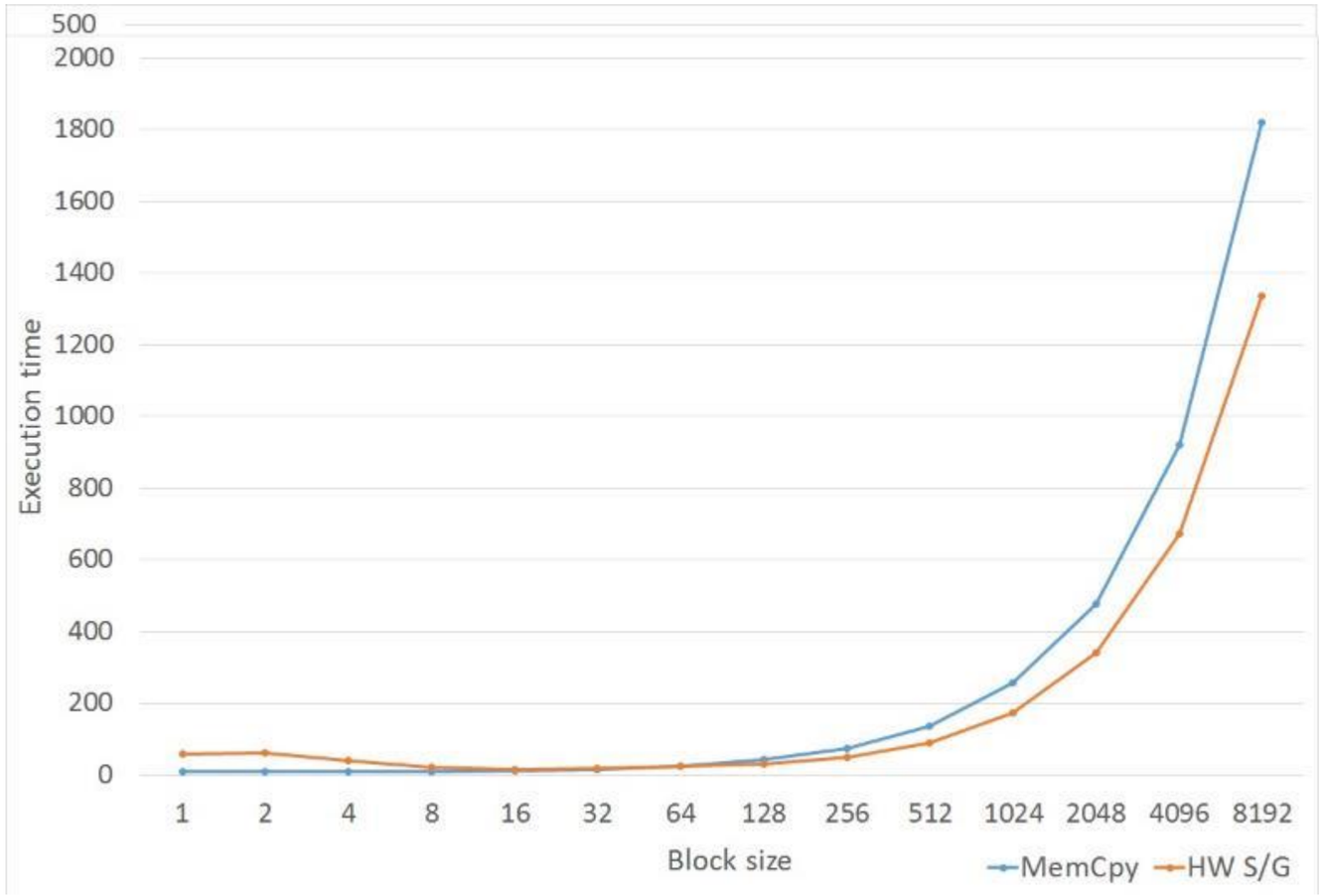
Scatter-Gather HCA Capabilities



Ping-Pong Latency : Fixed Block Count, Block Size Varied

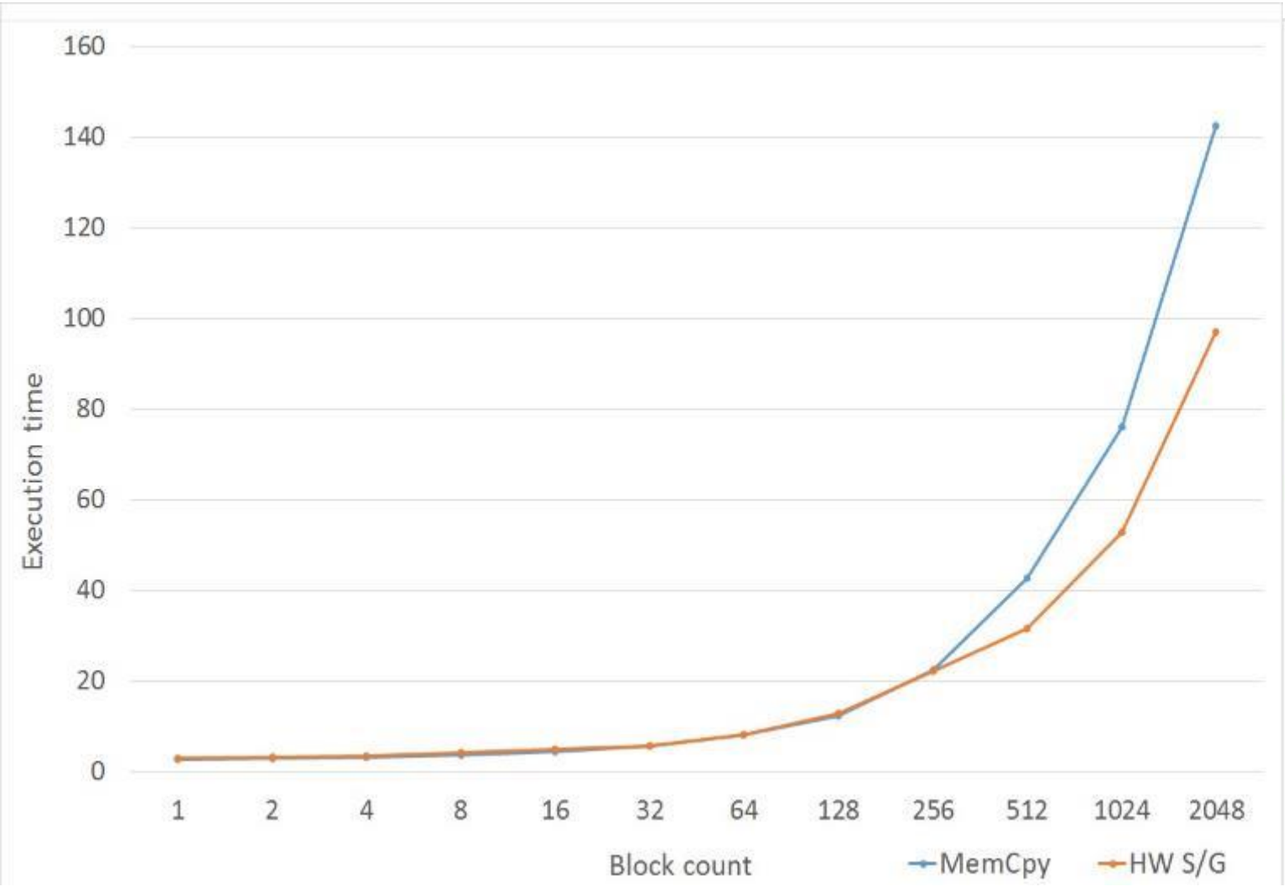


128 Blocks

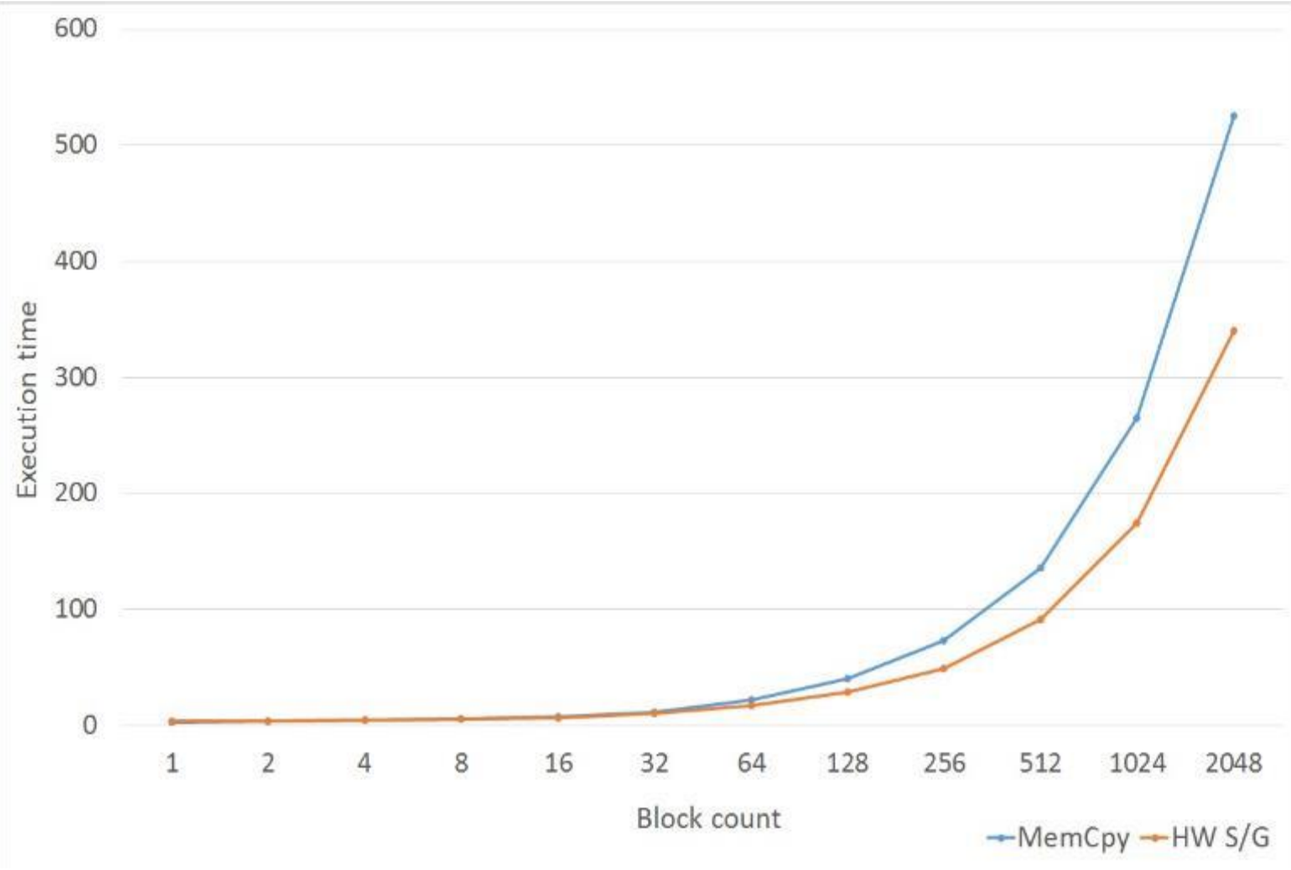


512 Blocks

Ping-Pong Latency : Fixed Block Size, Block Count Varied

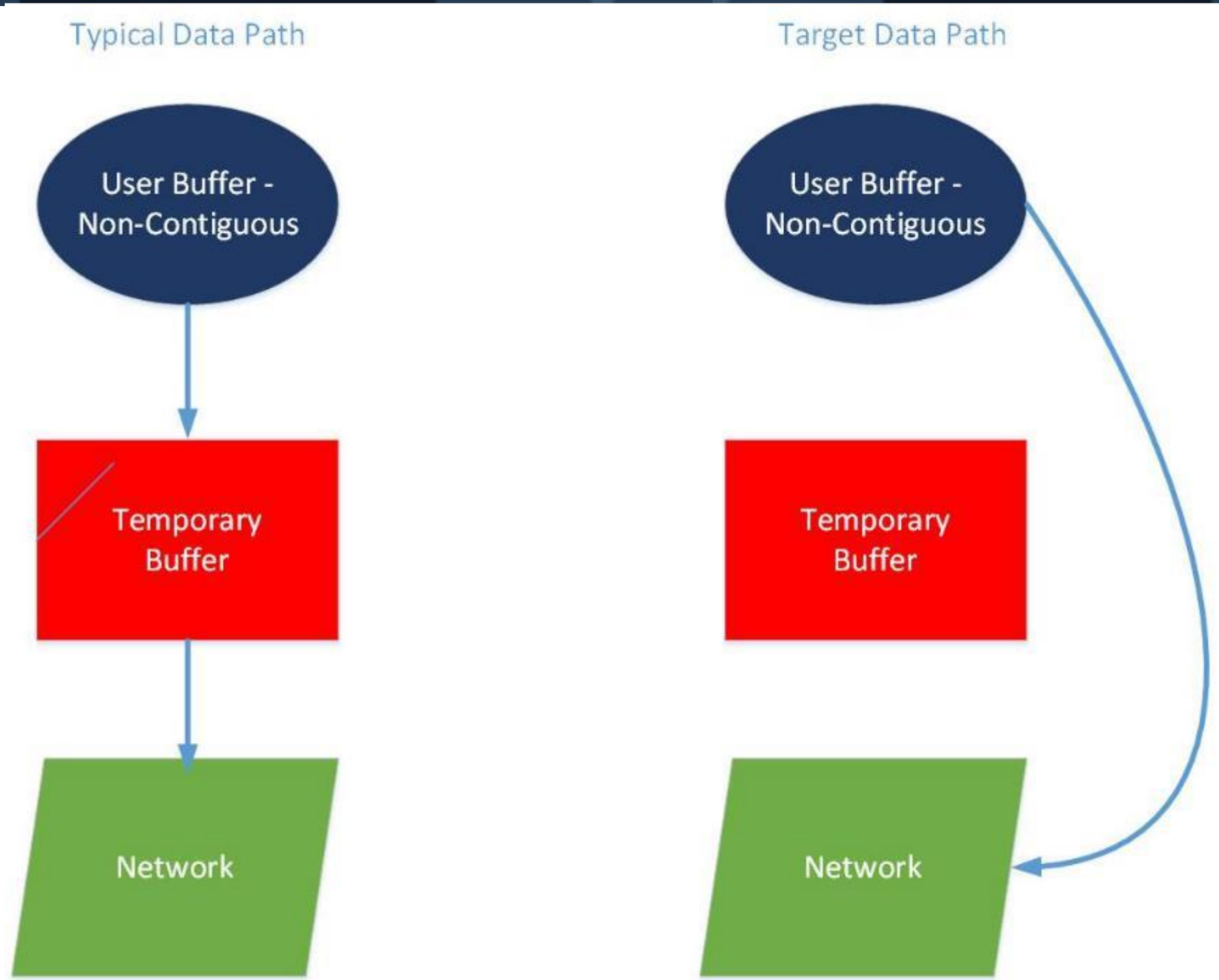


128 Bytes

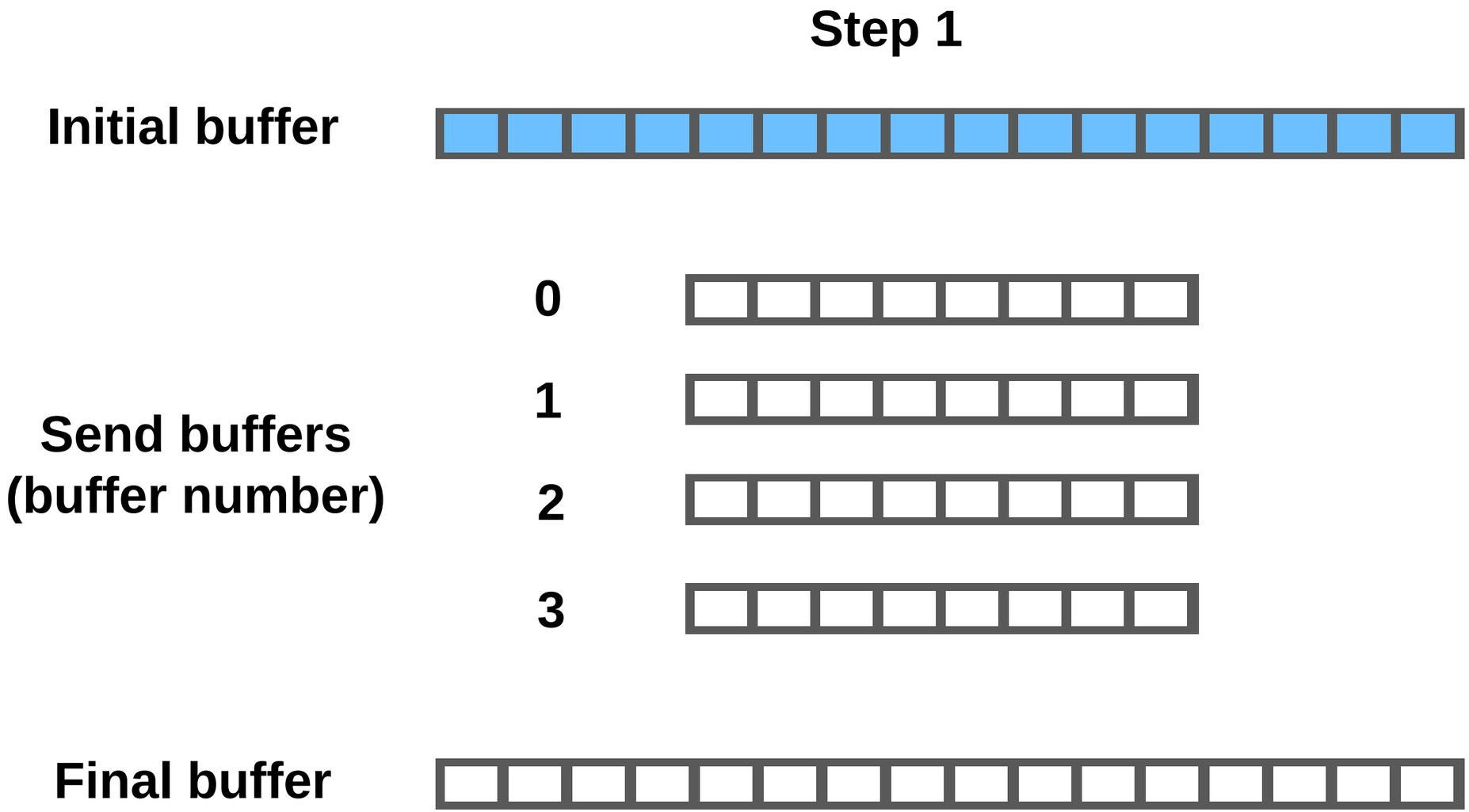


512 Bytes

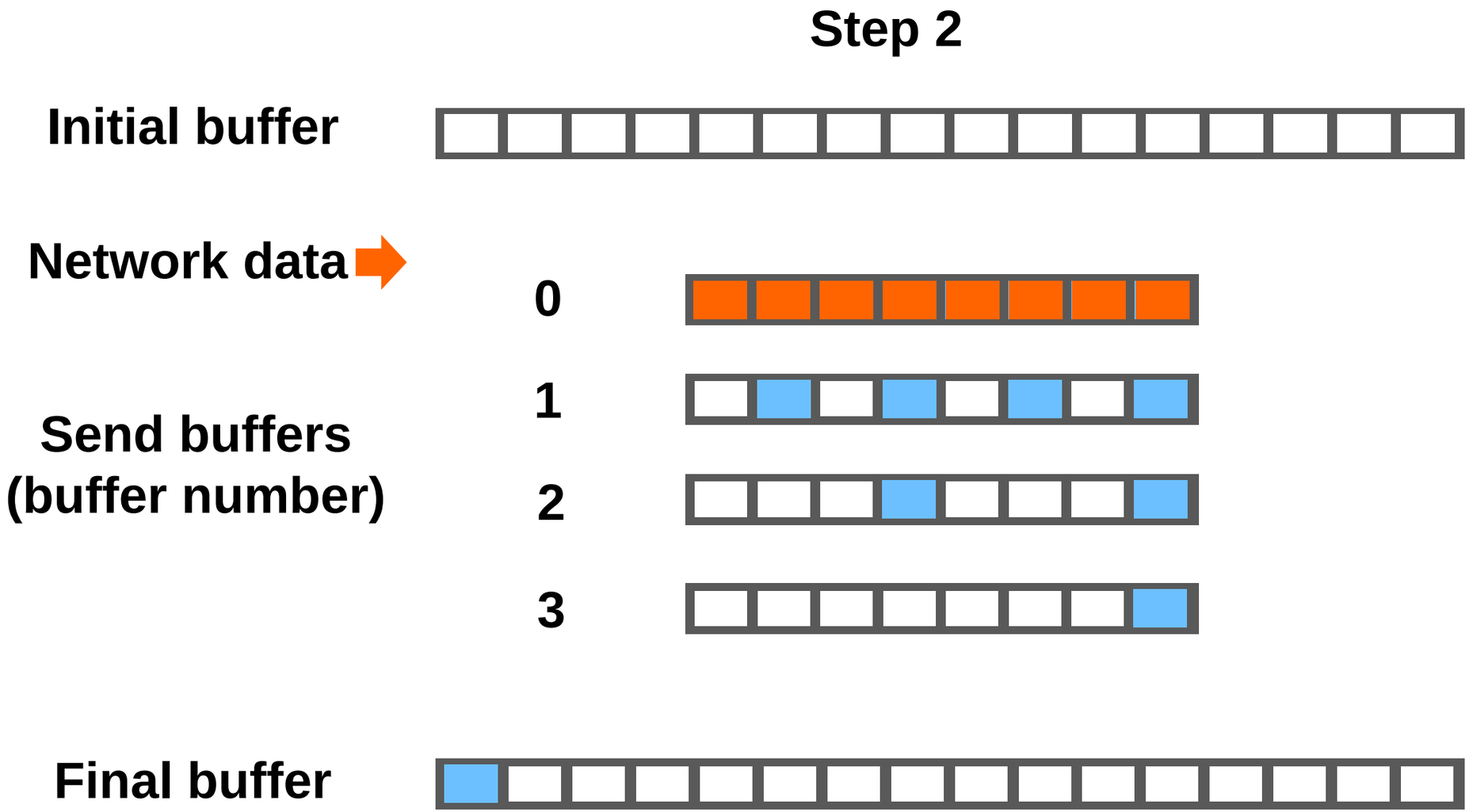
All-to-all collective operation



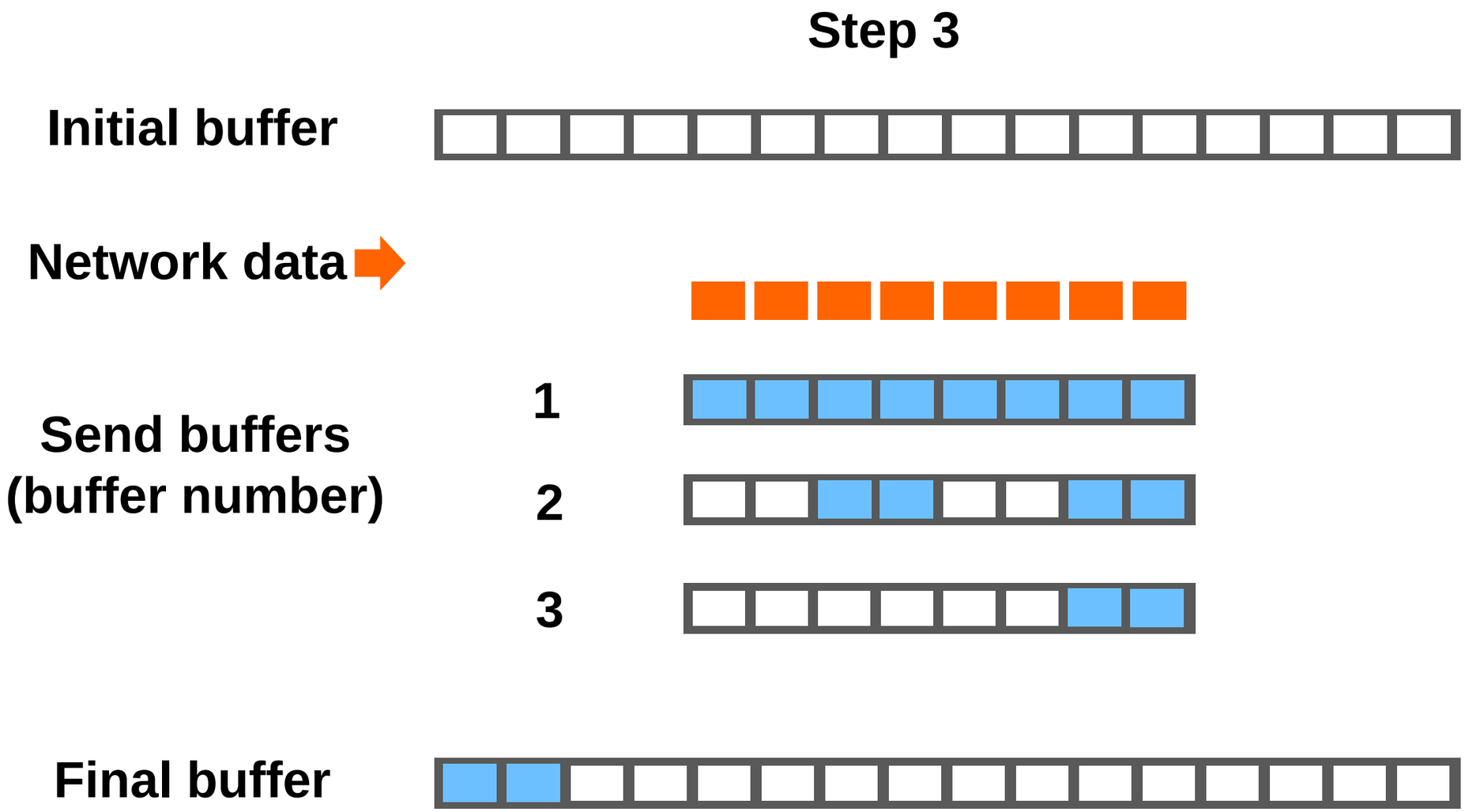
All-to-All Scatter Data Pattern



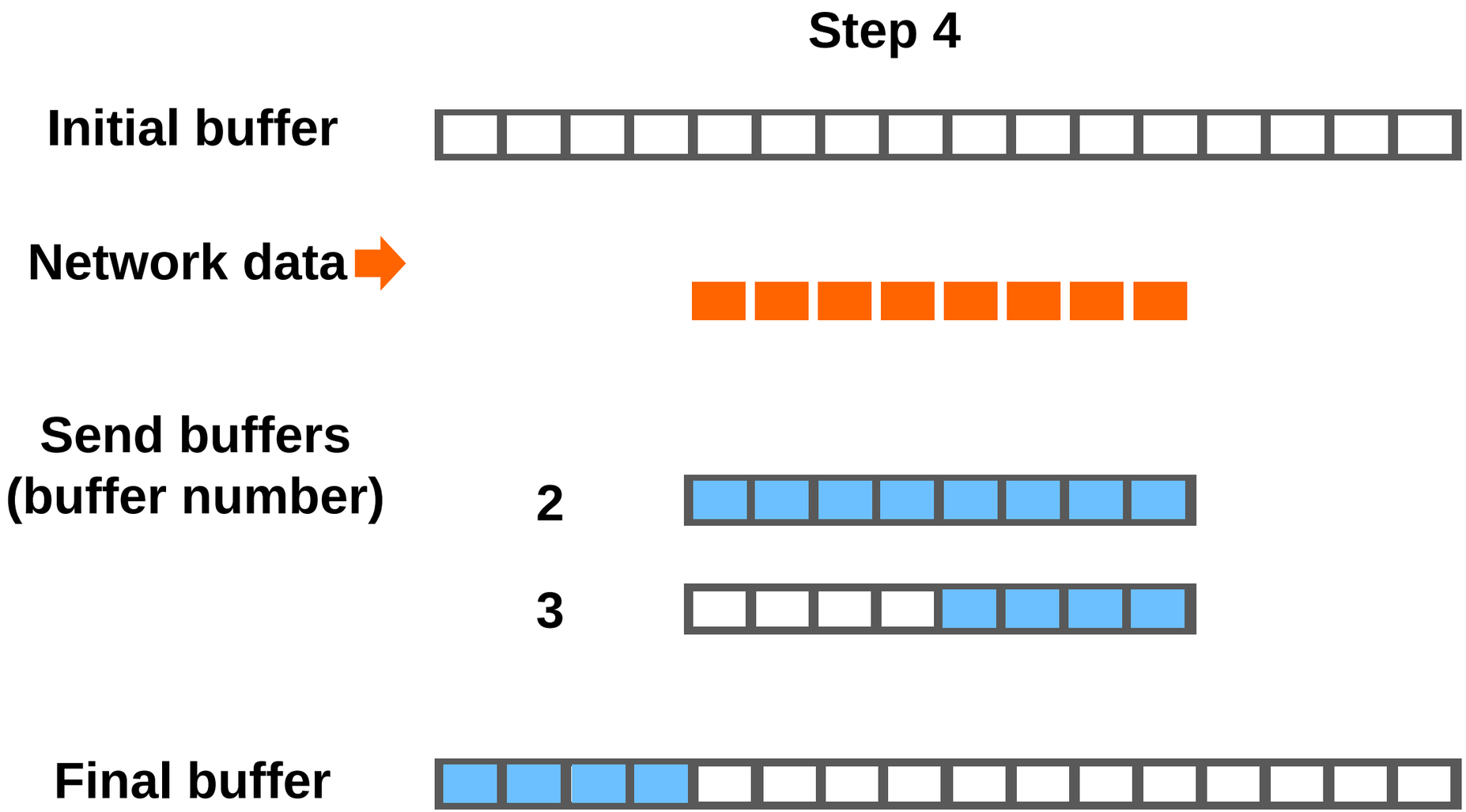
All-to-All Scatter Data Pattern



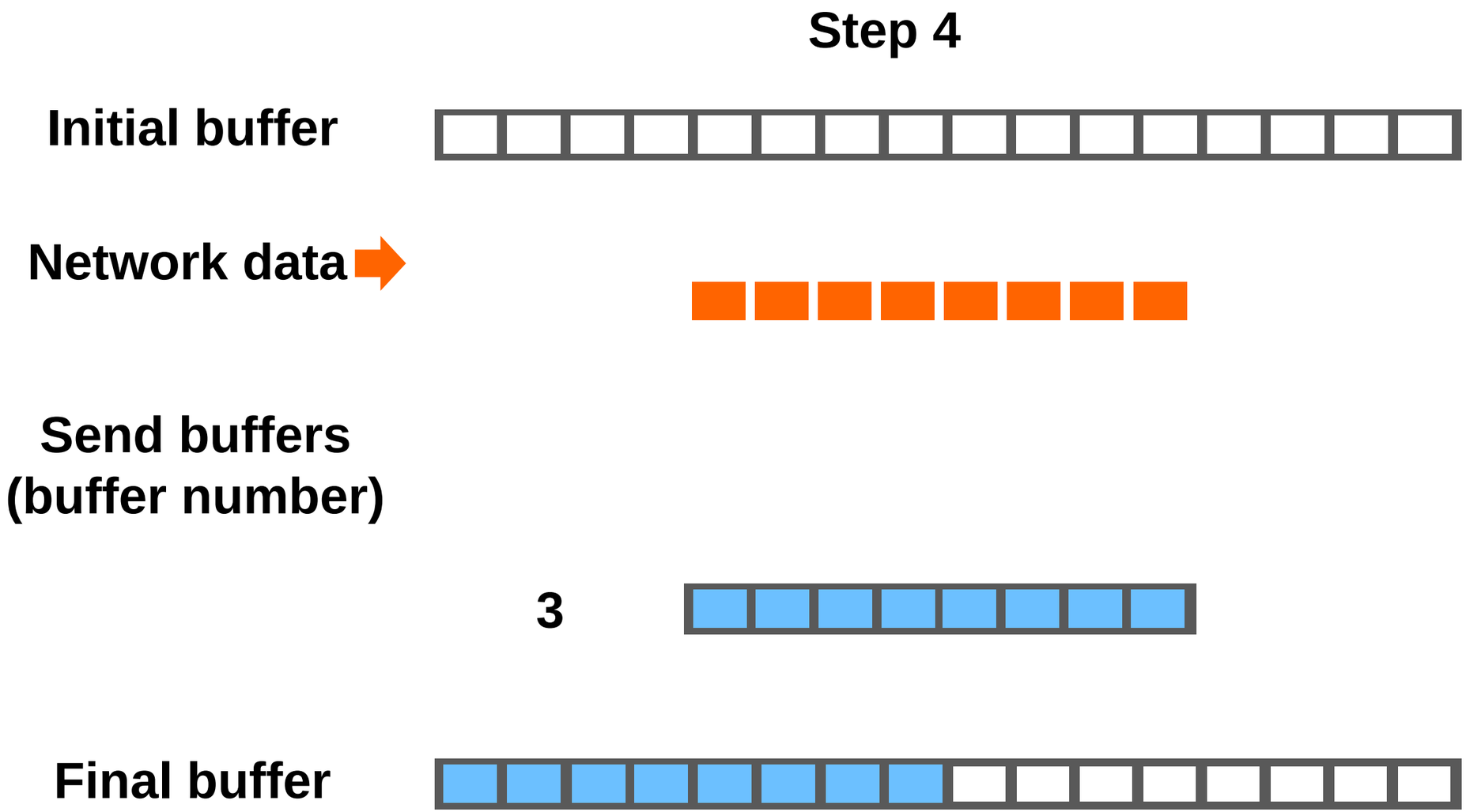
All-to-All Scatter Data Pattern



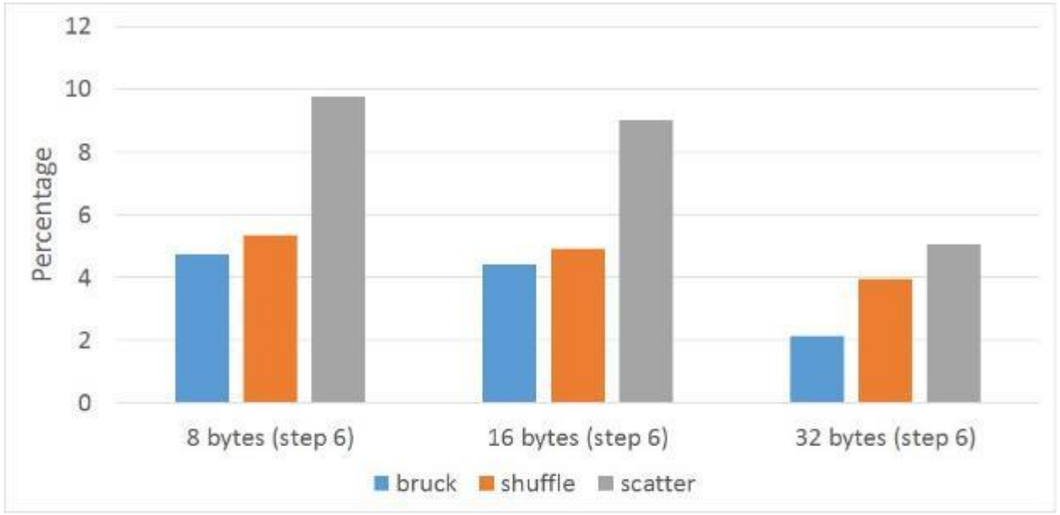
All-to-All Data Pattern



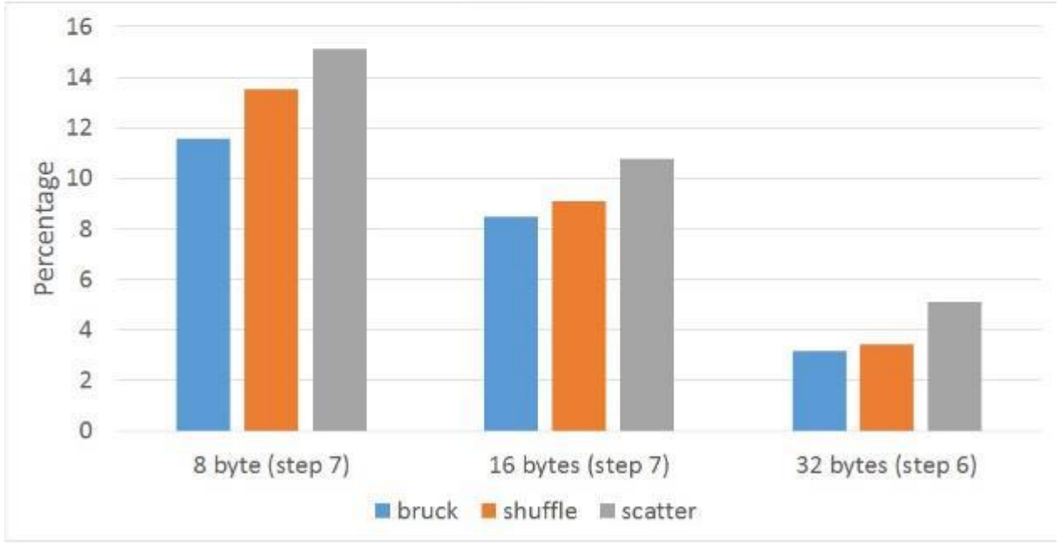
All-to-All Scatter Data Pattern



All-to-all Hybrid Algorithm: Speedup in percent of memory copy time



Thor



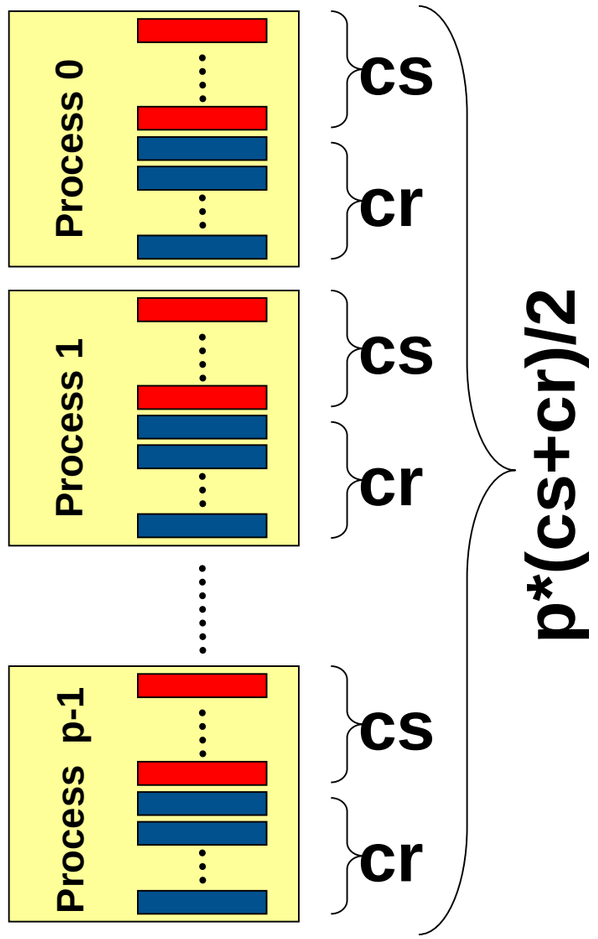
Jupiter

Dynamically Connected Transport

A Network Managed, Scalable Transport

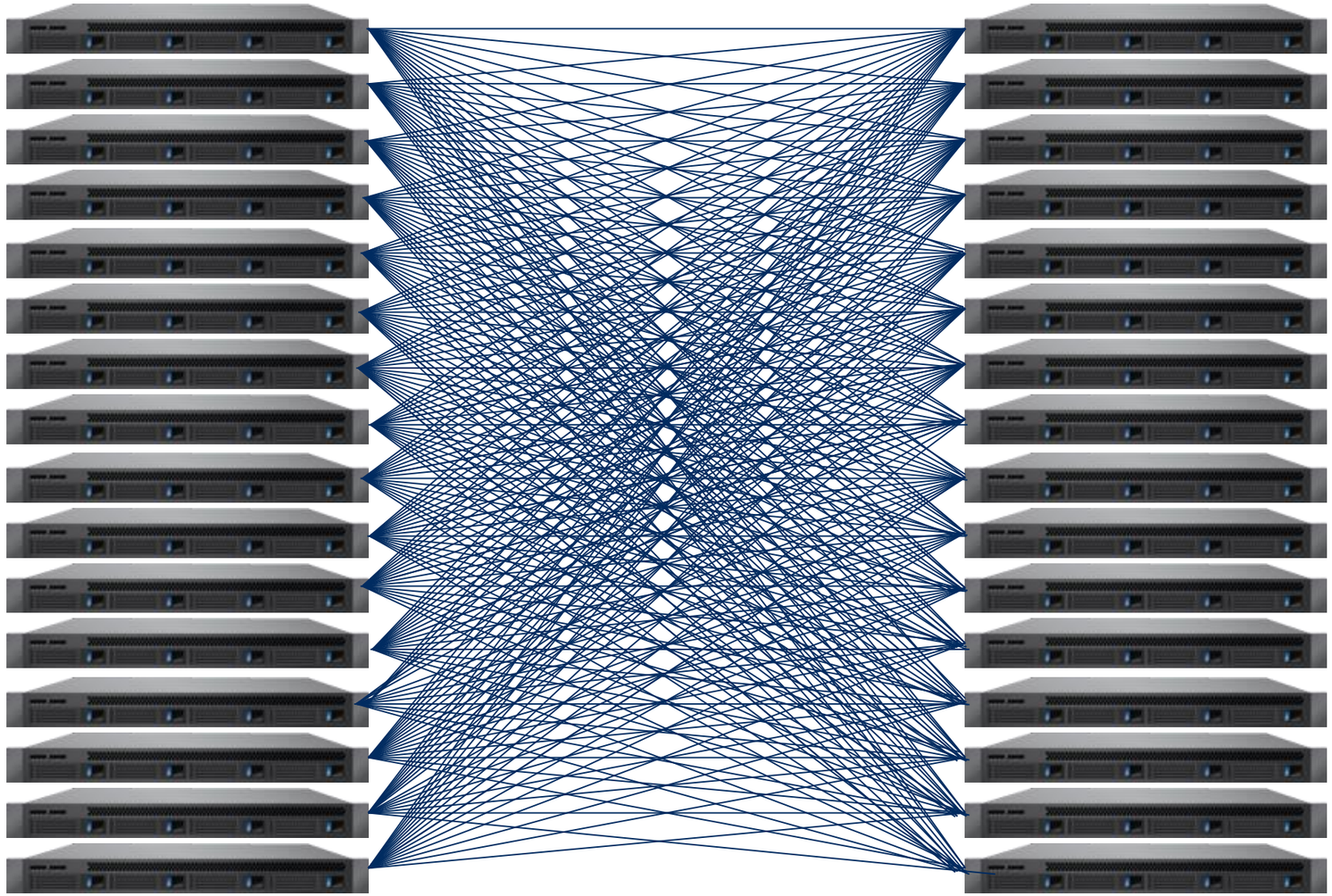
The DC Model

- Dynamic Connectivity
- Each DC Initiator can be used to reach any remote DC Target
- No resources' sharing between processes
  - process controls how many (and can adapt to load)
  - process controls usage model (e.g. SQ allocation policy)
  - no inter-process dependencies
- Resource footprint
  - Function of HCA capability
  - Independent of system size
- Fast Communication Setup Time

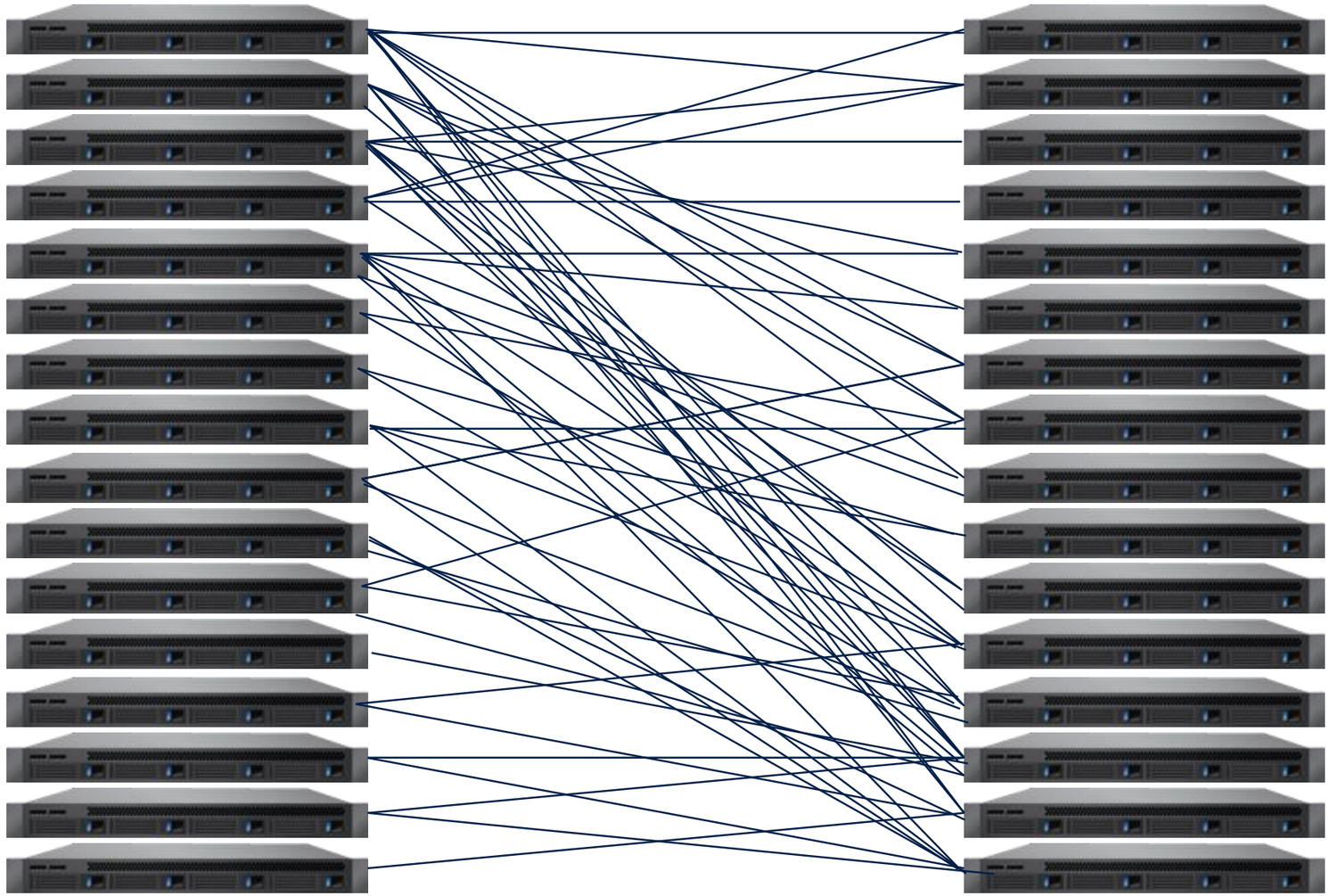


cs – concurrency of the sender
cr=concurrency of the responder

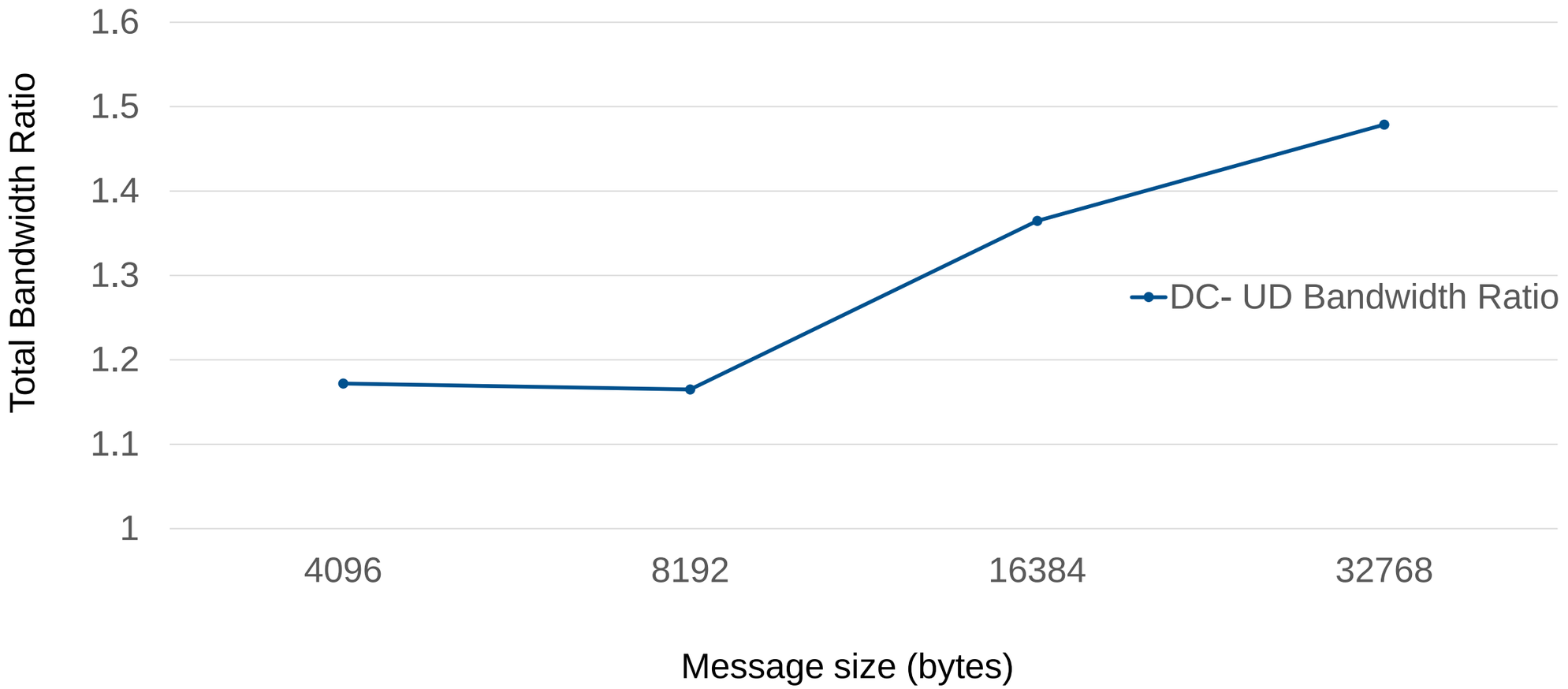
Reliable Connection Transport Mode



Dynamically Connected Transport Mode



All-To-All Performance



Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™