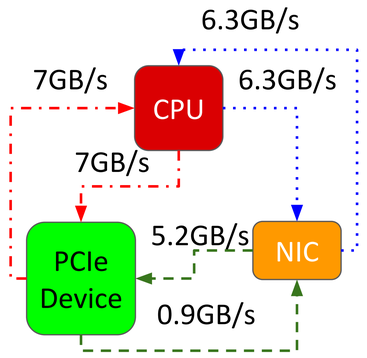
Designing High Performance and Energy-Efficient MPI Collectives for Next Generation Clusters

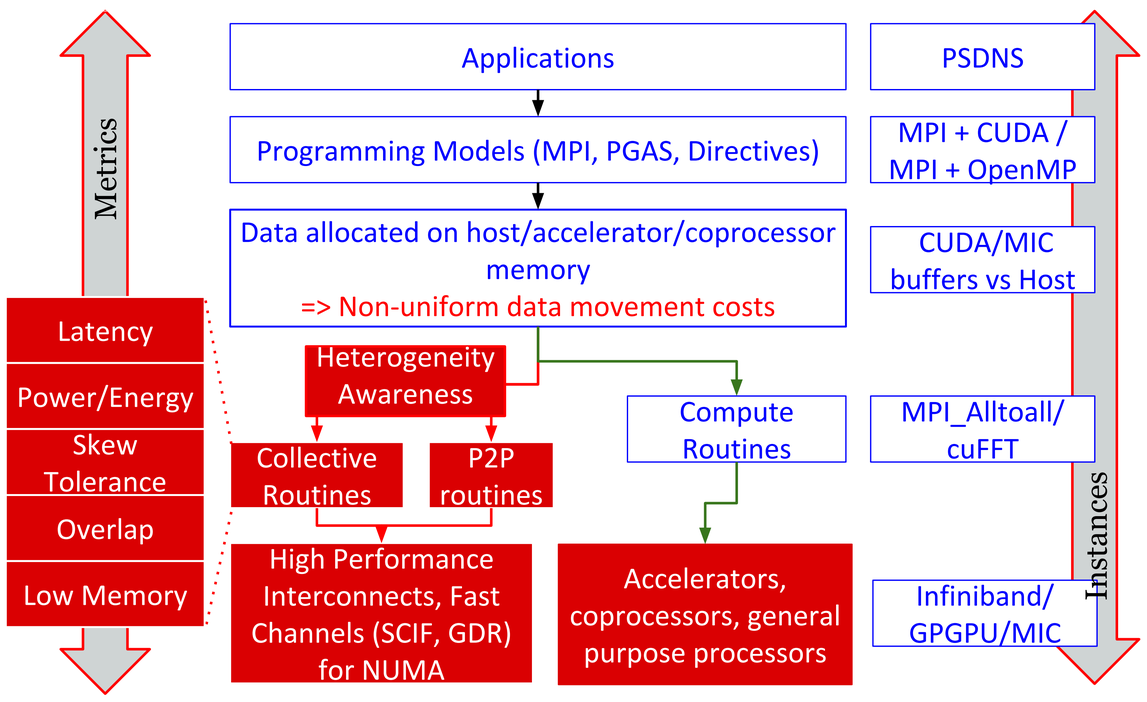
- Akshay Venkatesh (4th year Ph.D student), Advisor: DK. Panda

Research Challenges:

- Accelerators/coprocessors have throughput/watt but render communication paths ***heterogeneous*** and hence lead to ***non-uniform*** data movement costs making traditional collective algorithms suboptimal
- This work investigates the design of ***heterogeneity-aware*** algorithms to improve the latency, overlap, and power footprint of important MPI collectives.

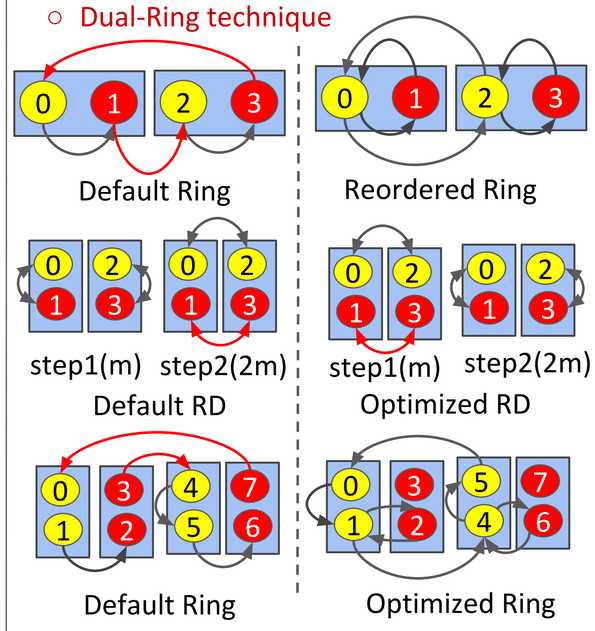


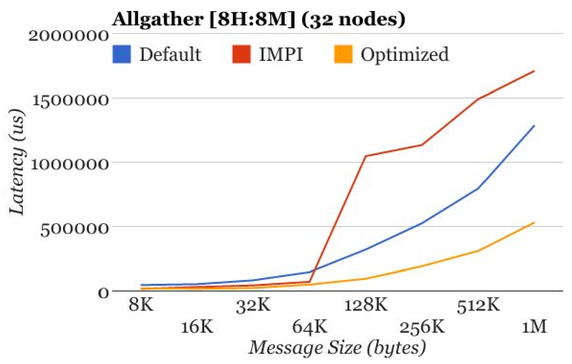
Framework:

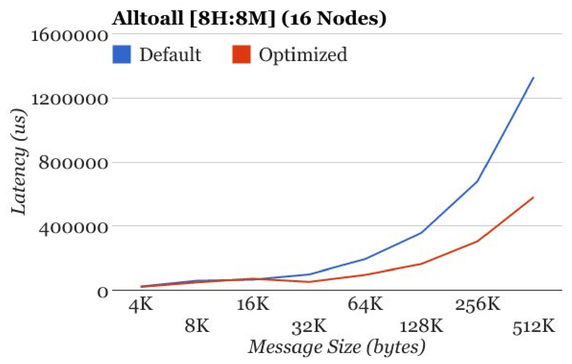


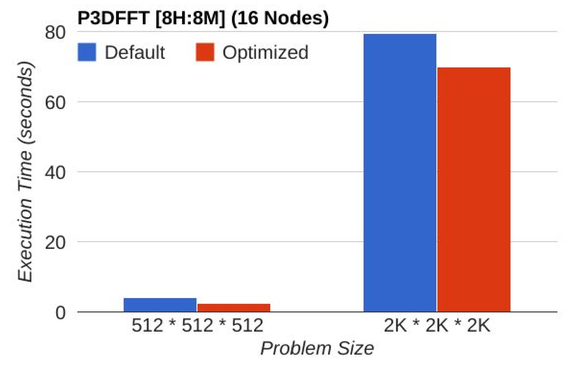
Optimized Allgather and Alltoall (MIC) [3]:

- State of the art algorithms like Bruck's Allgather and Alltoall, Recursive-doubling Allgather, Ring-allgather and Pairwise-alltoall assume uniformity
- To overcome bandwidth limitations, proposed methods use techniques such as:
  - Staging
  - Selective Rerouting
  - Order rescheduling
  - Ring Reordering
  - Dual-Ring technique





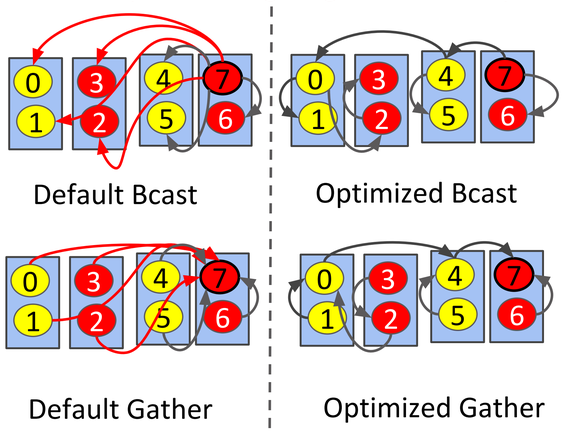


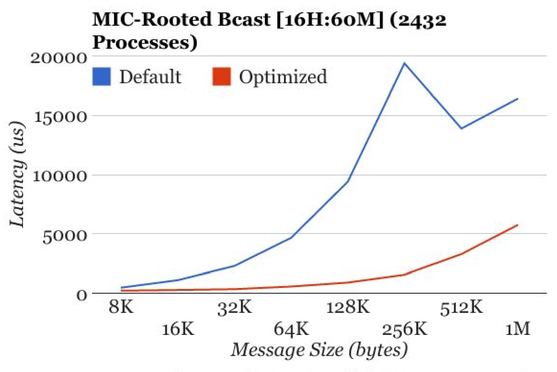


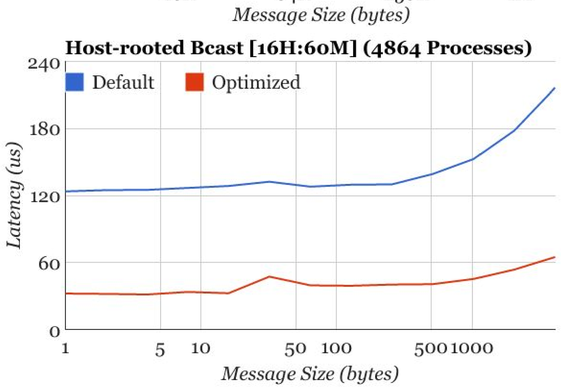
[3] High Performance Alltoall and Allgather designs for InfiniBand MIC Clusters IPDPS'14

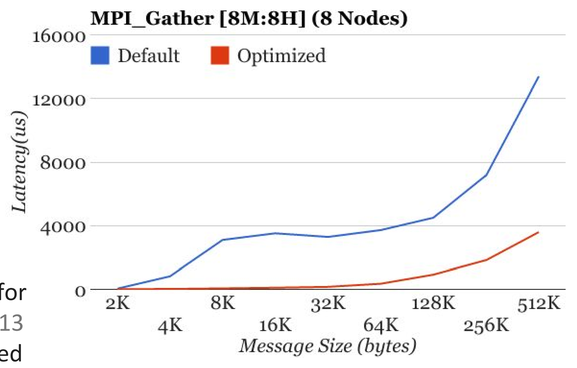
Optimized MIC-Bcast [1] and Gather [2] (MIC):

- Optimization leverages the presence of host processes on a MIC node in symmetric mode of execution
- When the root of the broadcast is located on the MIC, a pre-assigned process on the host process forwards the message to the rest of the system in hierarchical manner
- For the MPI_Gather operation, 3 techniques are used:
  - 3-level hierarchical approach
  - Pipelined approach
  - Overlapped 3-level approach
- Techniques leverage high bandwidth path





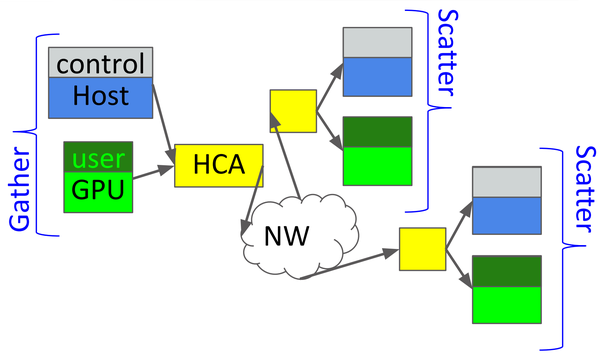


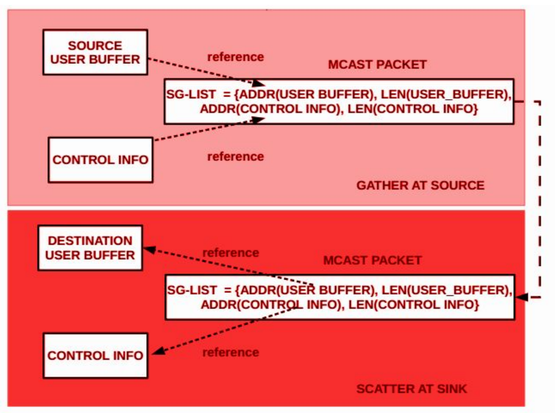


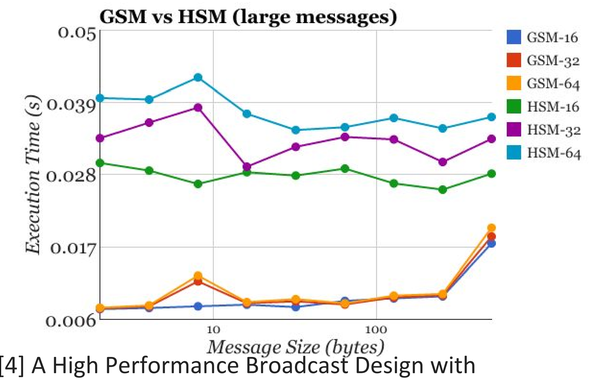
[1] Designing Optimized MPI Broadcast and Allreduce for Many Integrated Core (MIC) InfiniBand Clusters HOTI'13
[2] Optimized MPI Gather collective for Many Integrated Core (MIC) InfiniBand Clusters. XSCALE'13

GPU Buffer MCAST (GSM) for Streaming Applications [4]:

- Traditional approach involves staging GPU data on host memory (HSM)
- Detrimental to throughput oriented streaming class of applications
- Proposed work circumvents staging for multicast through *scatter-gather-list* abstraction from Infiniband and NVIDIA's *GPUDirect RDMA*



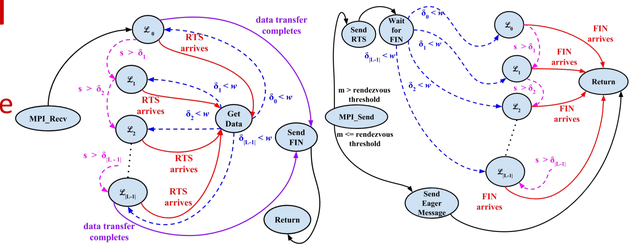


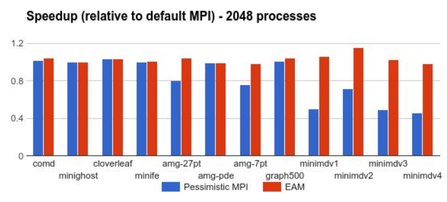


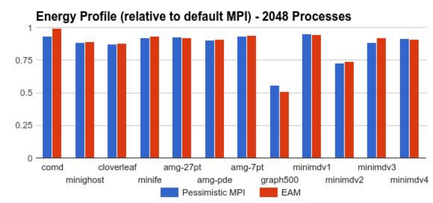
[4] A High Performance Broadcast Design with Hardware Multicast and GPUDirect RDMA for Streaming Applications on Infiniband Clusters HIPC'14

Energy-efficient Collectives and MPI runtime:

- Leverage point-to-point protocol and collective algorithm knowledge to apply energy-saving techniques in timely manner
- Application-oblivious and near zero performance degradation







[4] A Case for Application Oblivious Energy-Efficient MPI Runtime SC'15 (best student paper nominee)

Conclusions and future work:

- Published works circumvent the use of costly paths through reordering of operations to yield an average of 75% benefits for MPI Collectives and yield as much as 41% energy efficiency in MPI execution.
- Future works will aim at proposing efficient collectives with novel mechanisms such as GDR Async Technology to allow better collective overlap in GPU clusters..