Co-designing MPI Runtimes and Deep Learning Frameworks for Scalable Distributed Training on GPU Clusters

HiDL High-Performance Deep Learning http://hidl.cse.ohio-state.edu

MVAPICH

Ammar Ahmad Awan and Dhabaleswar K. Panda (Advisor)

awan.10@osu.edu, panda@cse.ohio-state.edu

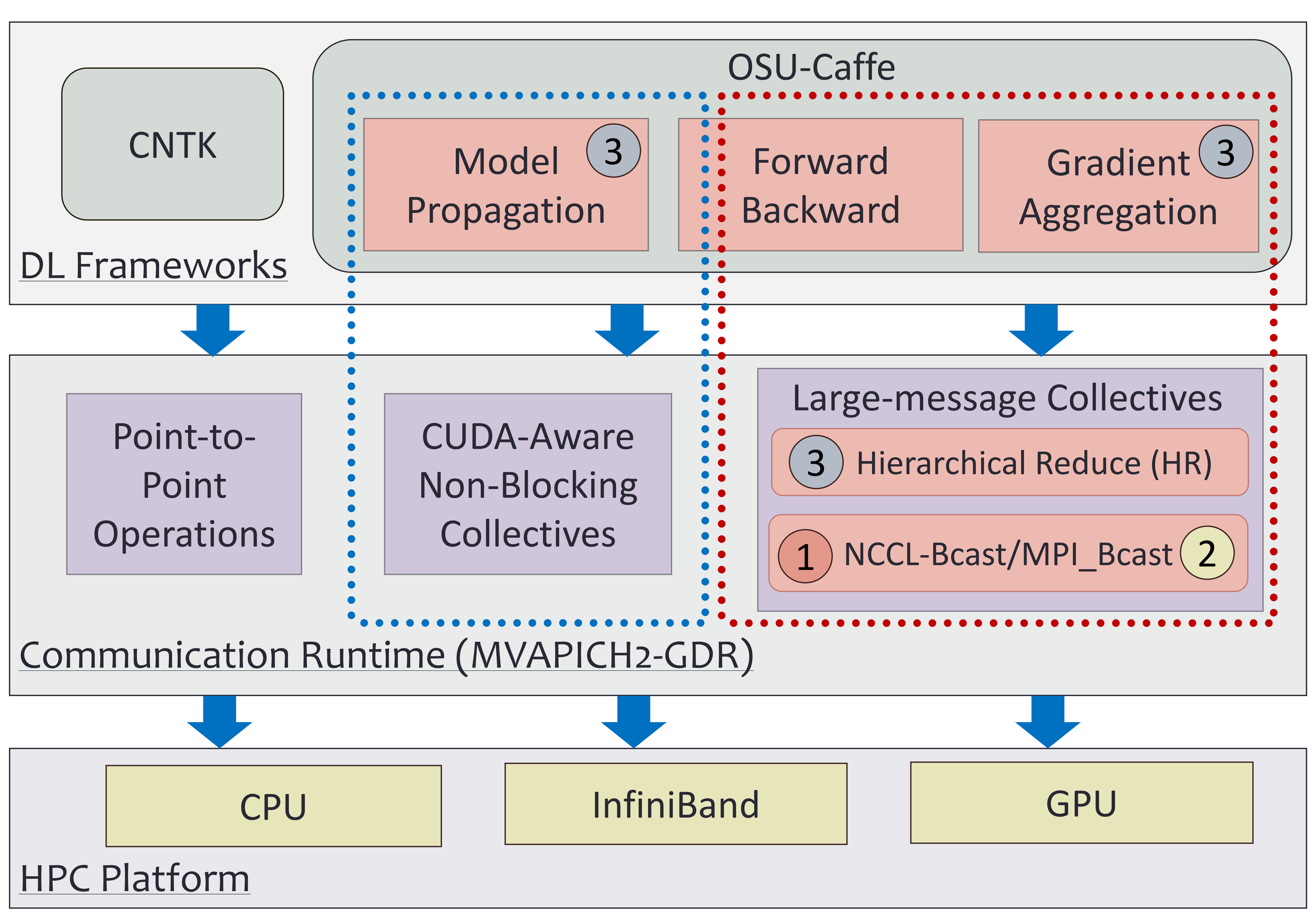
MOTIVATION

- Resurgence of Deep Learning (DL)
  - Availability of Large Datasets like ImageNet and massively-parallel modern hardware like NVIDIA GPUs
  - Emergence of DL frameworks (Caffe, TensorFlow, CNTK, etc.)
- Computability of Deep Neural Networks (DNNs)
  - ***Single GPU/node is not enough!***
  - ***Scale-up and Scale-out training: an emerging research area***

RESEARCH CHALLENGES

- Various Parallelization Strategies for DNNs
  - Model Parallelism / **Data Parallelism** (3)
- Alternative Implementation Styles
  - Parameter-Server approach / **Reduction-Tree approach** (1) (2) (3)
- Distributed Address-Space Design Constraints (3)
- Parallel Data Reading Mechanisms (3)
- Challenges for Communication Runtimes
  - Very Large GPU-based Buffers (1) (2) (3)
  - Overlap of Computation and Communication (3)

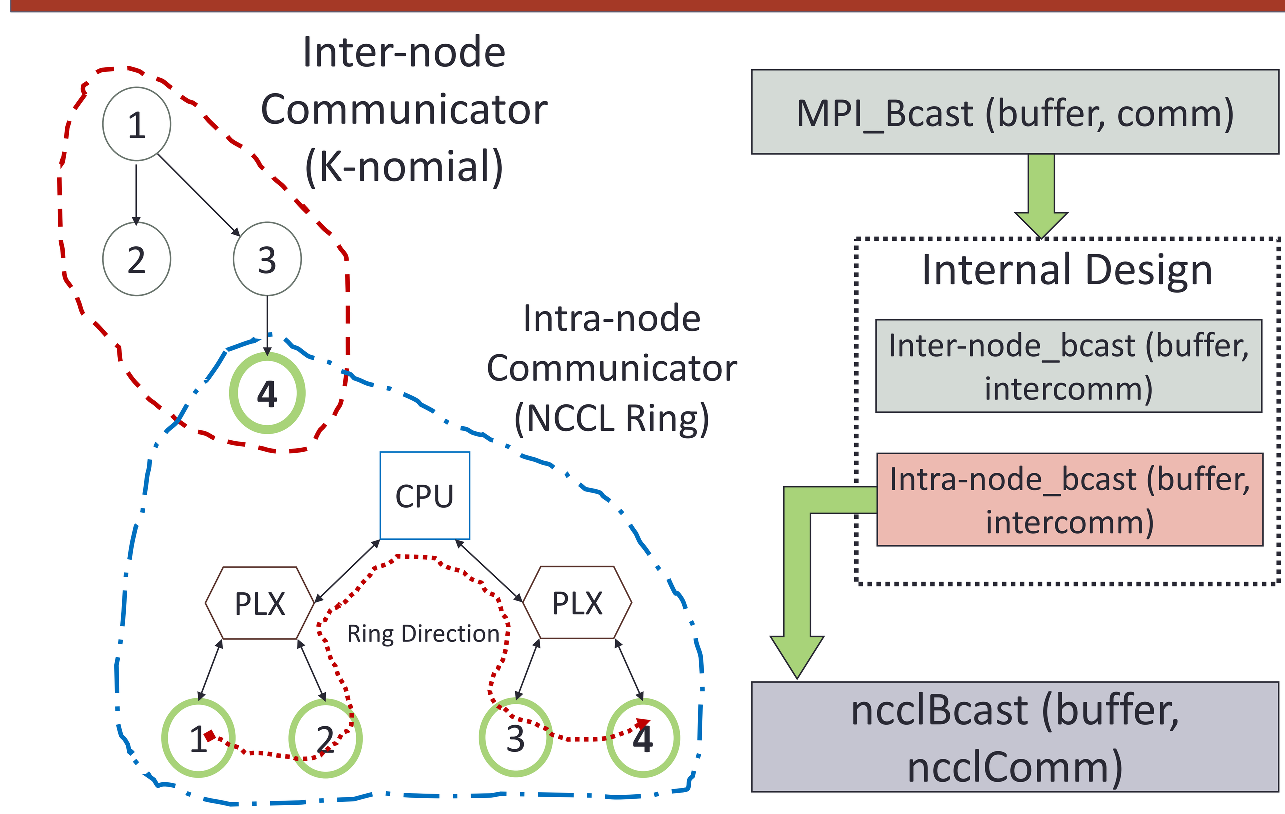
PROPOSED FRAMEWORK



SUMMARY OF CONTRIBUTIONS

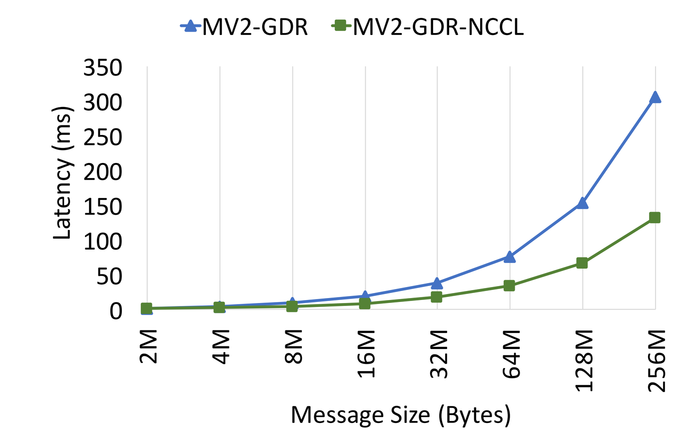
- Tackle the challenge of designing a **scalable** and **distributed** DL framework
- Efficient **Intra-node** and **Inter-node** training
- Proven **scale-out** for **GoogLeNet** up to 160 GPUs
- Support for Small (CIFAR10/MNIST) and Large Datasets (ImageNet)
- Optimized **Model Propagation** and **Gradient Aggregation**
- Various **Design Alternatives** to provide Optimal Performance for **Small** and **Large** scale training

PROPOSED SOLUTIONS AND PERFORMANCE EVALUATION

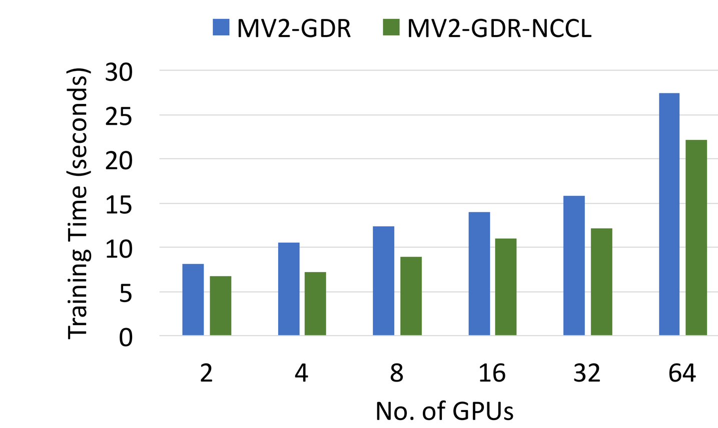


NCCL-based Hierarchical Design for Large-message MPI Bcast

- ***MPI_Bcast: Design Broadcast for DL Workloads using NCCL***
  - NCCL-augmented hybrid design in MVAPICH2-GDR for intra-node communication
  - Tuned inter-node communication using various algorithms like K-nomial Tree, Scatter-Allgather, etc.
  - Combine performance features of NCCL and MPI in a unified communication runtime
- ***Performance Benefits***
  - Up to ***2X improvement*** for micro-benchmarks
  - Up to ***38% improvement*** for VGG training with CNTK

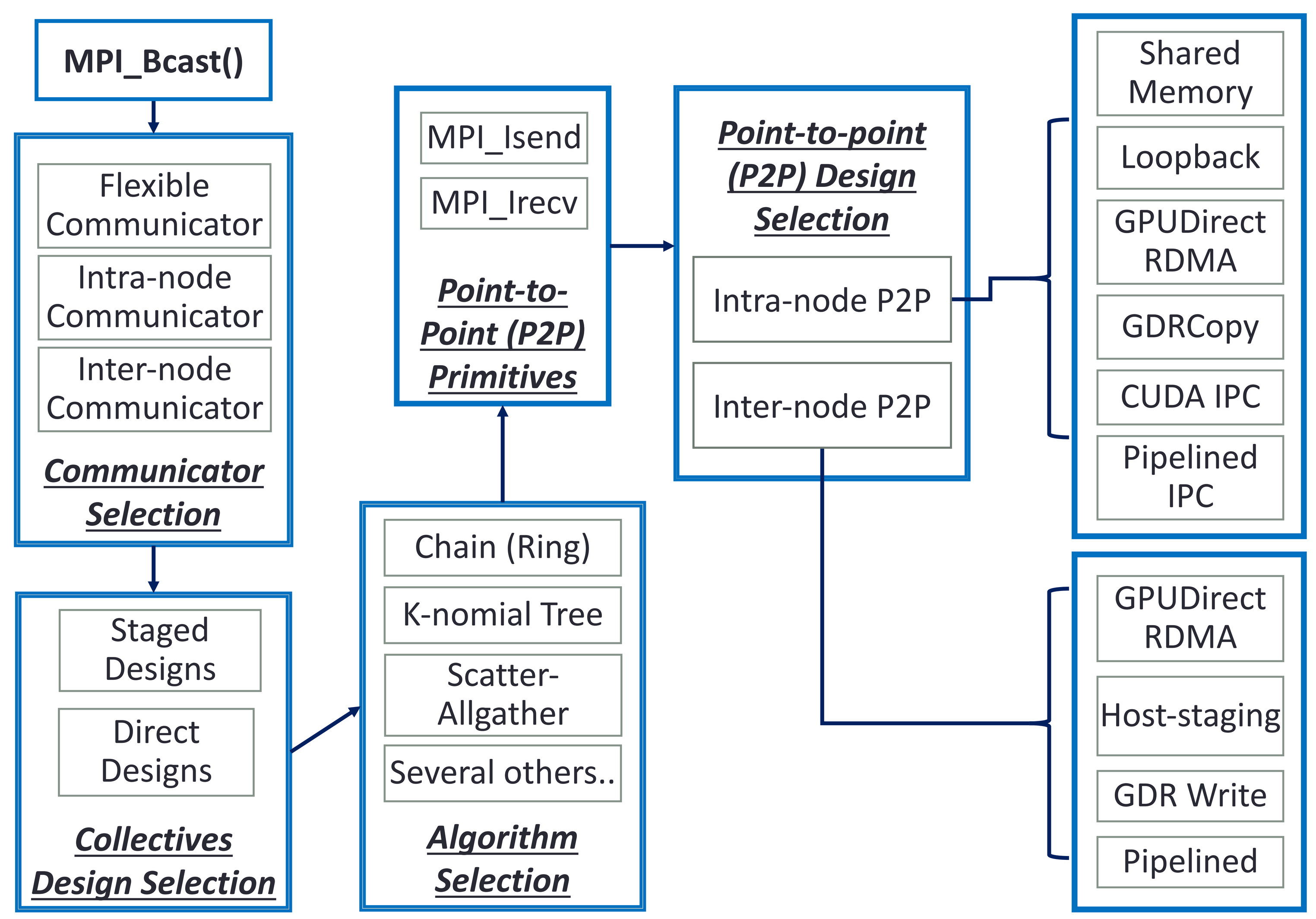


MPI Bcast Benchmark: 64 GPUs (8 nodes)



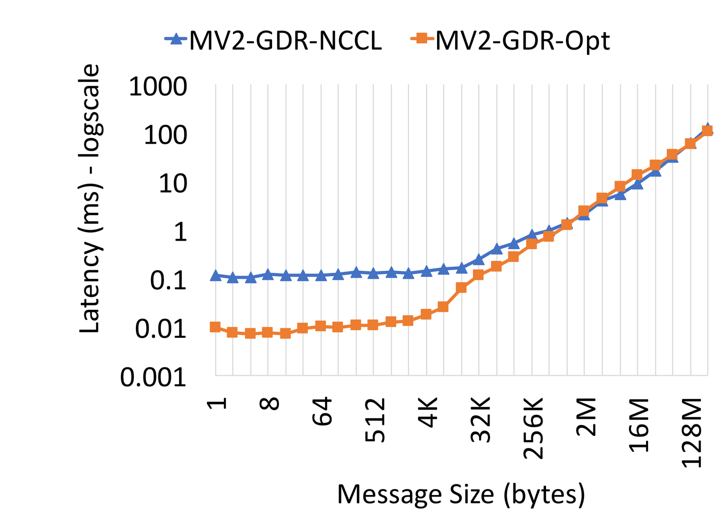
VGG Training with CNTK

(1) A. A. Awan, K. Hamidouche, A. Venkatesh, and D. K. Panda. Efficient Large Message Broadcast using NCCL and CUDA-Aware MPI for Deep Learning, ACM EuroMPI '16

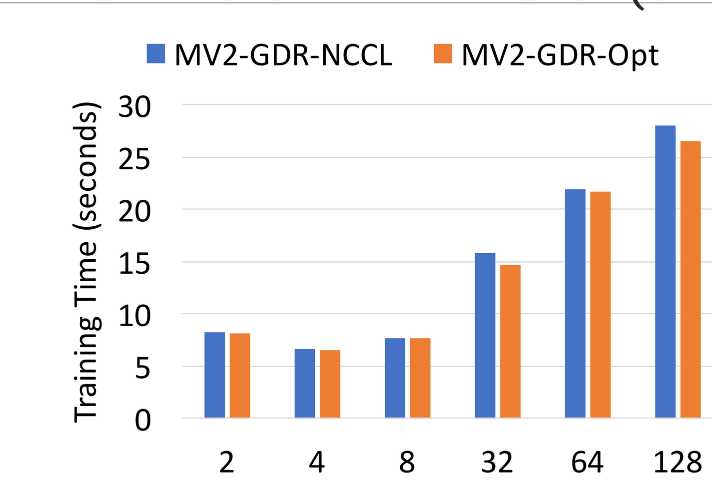


Pure MPI Design for DL-Aware MPI_Bcast

- ***MPI_Bcast: Design and Performance Tuning for DL Workloads***
  - Design ring-based algorithms for large messages
  - Harness a multitude of algorithms and techniques for best performance across the full range of message size and process/GPU count
- ***Performance Benefits***
  - Performance ***comparable or better*** than NCCL-augmented approaches for large messages
  - Up to ***10X improvement*** for small/medium message sizes with micro-benchmarks
  - Up to ***7% improvement*** for VGG training

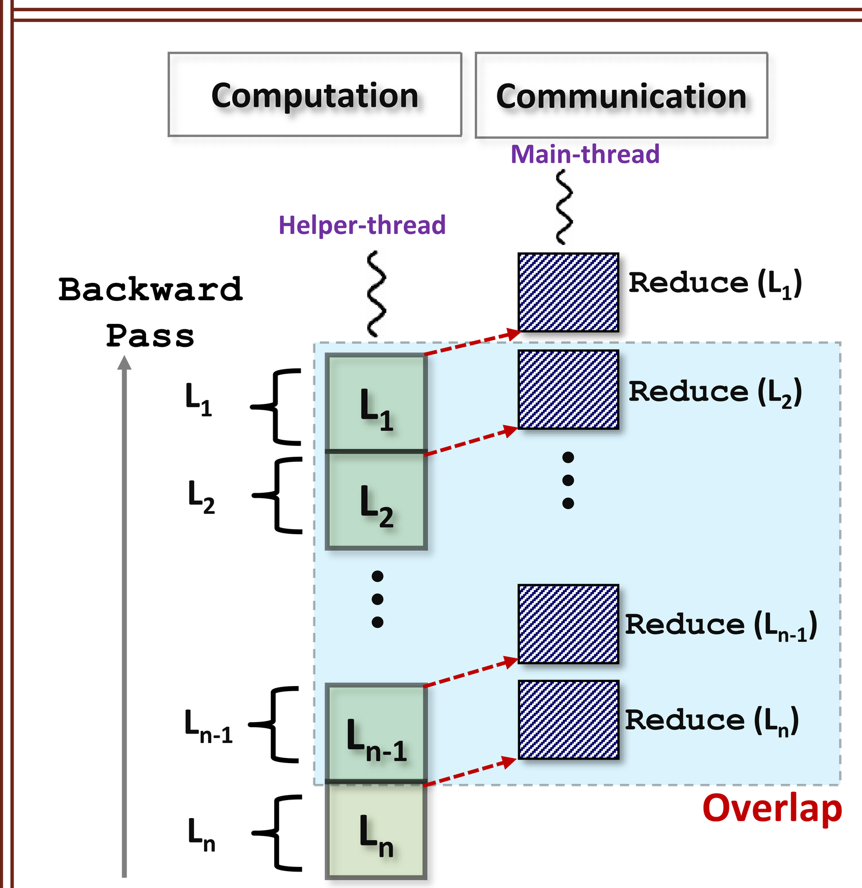


MPI Bcast Benchmark: 128 GPUs (8 nodes)



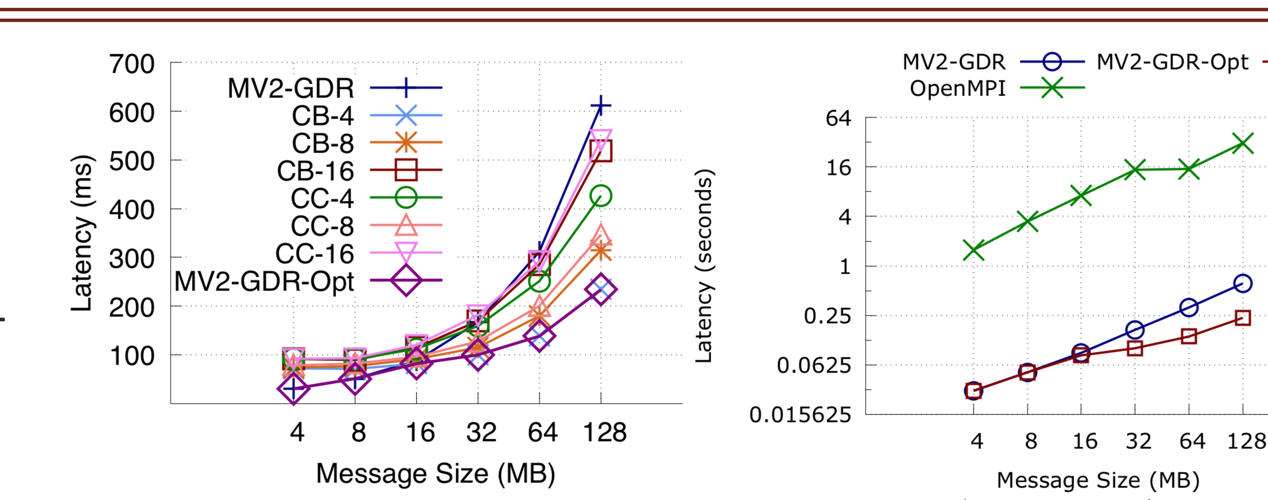
VGG Training with CNTK

(2) A. A. Awan, C-H. Chu, H. Subramoni, and D. K. Panda. Optimized Broadcast for Deep Learning Workloads on Dense-GPU InfiniBand Clusters: MPI or NCCL?, arXiv '17 (https://arxiv.org/abs/1707.09414)

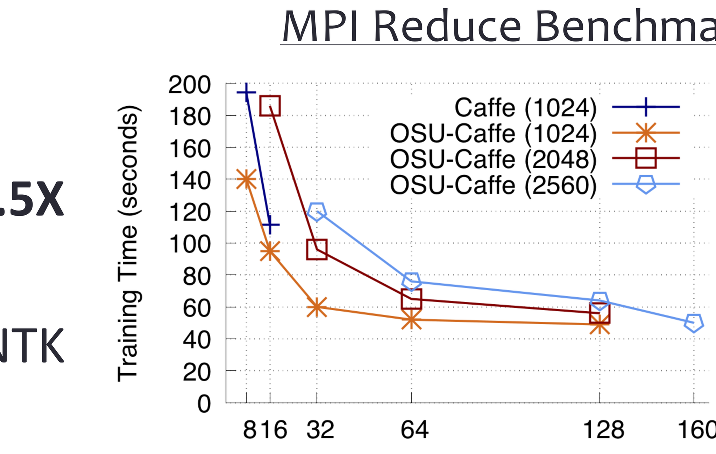


Layer-wise Overlapped Gradient Aggregation in OSU-Caffe

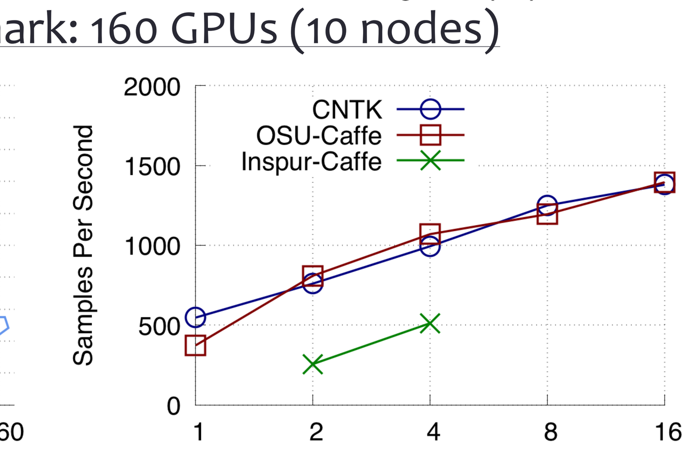
- ***OSU-Caffe: Co-Design MVAPICH2-GDR and Caffe***
  - Provide design principles to overlap DNN training with MPI communication
  - MPI_Reduce: Efficient GPU-based designs for large-message reductions
  - Delivers better or comparable performance to production-grade DL frameworks
- ***Performance Benefits***
  - MPI_Reduce: **130X** speedup over OpenMPI and **2.5X** improvement over MVAPICH2-GDR
  - OSU-Caffe: Better/comparable performance to CNTK for AlexNet training
  - OSU-Caffe: Scale-out to **160 GPUs** for GoogLeNet



MPI Reduce Benchmark: 160 GPUs (10 nodes)



GoogLeNet Training: Strong Scaling



AlexNet Training: Weak Scaling

(3) A. A. Awan, K. Hamidouche, J. M. Hashmi, and D. K. Panda. S-Caffe: Co-designing MPI Runtimes and Caffe for Scalable Deep Learning on Modern GPU Clusters, ACM PPoPP '17