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

Ammar Ahmad Awan and Dhabaleswar K. Panda (Advisor)
Department of Computer Science and Engineering,
The Ohio State University
awan.10@osu.edu, panda@cse.ohio-state.edu

ABSTRACT

Deep Learning frameworks like Caffe, TensorFlow, and CNTK have brought forward new requirements and challenges for communication runtimes like MVAPICH2-GDR. These include support for low-latency and high-bandwidth communication of very-large GPU-resident buffers. This support is essential to enable scalable distributed training of Deep Neural Networks on GPU clusters. However, current MPI runtimes have limited support for large-message GPU-based collectives. To address this, we propose the S-Caffe framework; a co-design of distributed training in Caffe and large-message collectives in MVAPICH2-GDR. We highlight two designs for MPI_Bcast, one that exploits NVIDIA NCCL and the other that exploits ring-based algorithms. Further, we present an MPI_Reduce design that provides up-to 2.5X improvement. We also present layer-wise gradient aggregation designs in S-Caffe that exploit overlap of computation and communication as well as the proposed reduce design. S-Caffe provides a scale-out to 160 GPUs for GoogLeNet training and delivers comparable performance to CNTK for AlexNet training.

1 INTRODUCTION

Deep Learning (DL) is undergoing a phenomenal growth spike where CS and Machine Learning (ML) communities alike are rapidly adopting DNNs to solve an array of classical problems. As DNN training becomes too large to be performed on a single GPU, several DL toolkits and frameworks like Caffe, TensorFlow, and Cognitive Toolkit (CNTK) have resorted to parallel/distributed training on multiple GPUs and compute nodes. To this end, there are new opportunities and challenges for the HPC community to rethink and enhance MPI-based communication runtimes like MVAPICH2-GDR [3] and enable low-latency, high-bandwidth communication for GPU-resident buffers. The first challenge is *"how to design the distributed training workflow in Caffe using MPI?"* To address this, we propose in S-Caffe [5] that the DL model (or its parameters) can be shared among distributed trainers (GPUs) using an MPI_Bcast while gradient aggregation can be achieved through MPI_Reduce. The second challenge is *"how to re-design collectives like MPI_Bcast and MPI_Reduce for DL workloads?"*. To improve the existing support in MVAPICH2-GDR, we take a co-design approach where we develop large-message collectives for GPU buffers in MVAPICH2-GDR and exploit them in S-Caffe using layer-wise propagation and aggregation schemes.

2 PROPOSED DESIGNS

We describe two MPI_Bcast designs to support large-message communication in Caffe and CNTK in Section 2.1 and 2.2. Next, we describe co-design of layer-wise gradient aggregation in S-Caffe and large-message reductions in MVAPICH2-GDR.

2.1 Design of MPI_Bcast for DL Workloads using NVIDIA NCCL

MVAPICH2-GDR provides good performance for small and medium sized GPU buffers but little exists that deals with very-large sized buffers. To address this, we propose hierarchical collective designs [6] to improve the communication latency of MPI_Bcast. We exploit NVIDIA NCCL, a special-purpose GPU-based library, to deal with large-message communication between GPUs within a node. However, NCCL 1.x designs work within a node only. Thus, we incorporate a NCCL communicator object inside the intra-node communicator of MVAPICH2-GDR to improve intra-node performance. Further, to scale-out on multiple nodes, we exploit the inter-node communicator and select the best algorithm based on the message size and process count.

2.2 Design and Performance Tuning of MPI_Bcast for DL Workloads

NVIDIA NCCL can be exploited to provide large-message communication. However, such designs can lead to both performance degradation as well as maintenance issues because of the complex interactions of NCCL and MPI communicators, process groups, and multiple CUDA contexts. Thus, we propose a pipelined chain design for MPI_Bcast [4] that provides better or comparable performance to NCCL-based approaches discussed in [6].

2.3 S-Caffe: Co-Design of MVAPICH2-GDR and Caffe

S-Caffe [5] is our proposed DL framework that builds on the strengths of Caffe and MVAPICH2-GDR. We enhance the workflow in Caffe to maximize the overlap of computation and communication to provide efficient strong scaling. We perform multi-stage data propagation using the proposed MPI_Bcast designs and explore advanced gradient aggregation schemes that allow layer-wise reductions in a non-blocking manner using a helper-thread design. We propose the Hierarchical Reduce (HR) design that provide efficient large-message reductions for GPU buffers.

2.4 Public Availability

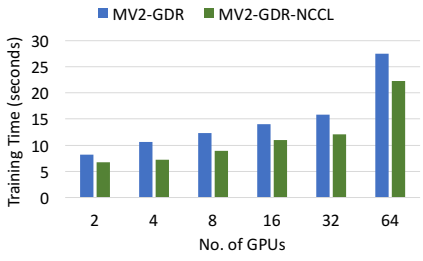
S-Caffe designs have been made available under the HiDL project [1]. The proposed large-message collectives have been made available with the MVAPICH2-GDR 2.2 release [3].

3 PERFORMANCE EVALUATION

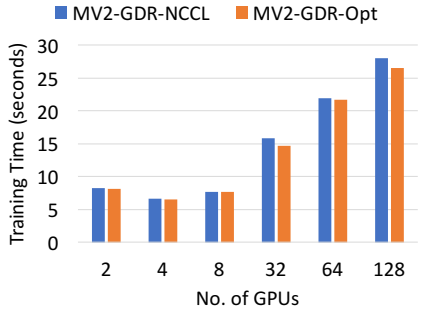
We used a Cray CS-Storm based GPU cluster called KESCH [2] for our experiments. KESCH has 12 nodes and each node has eight NVIDIA K-80 GPUs.

Proposed MPI_Bcast and MPI_Reduce: NCCL-based MPI_Bcast *(labeled MV2-GDR-NCCL)* designs provide up-to 3X speedup over

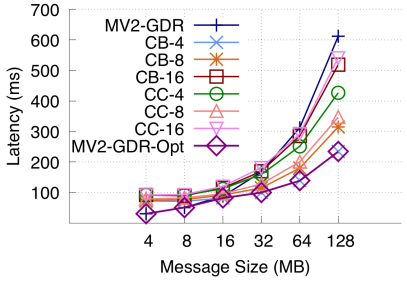
MVAPICH2-GDR (2.1) *(labeled MV2-GDR)* for the MPI_Bcast micro-benchmark. The detailed results for different configurations can be seen from [6]. The performance benefits observed for CNTK are shown in Figure 1(a). On average, the proposed designs provide 20-25% faster training for the VGG network. Similarly, the ring-based tuned designs *(labeled MV2-GDR-Opt)* provide up-to 10X speedup over NCCL-based MPI_Bcast for small and medium messages and comparable performance for large messages. The detailed micro-benchmark results appear in [4]. The application level benefits observed with CNTK for VGG training are presented in Figure 1(b). The tuned designs provide up to 7% improvement over NCCL-based solutions for VGG training on 128 GPUs. For MPI_Reduce, several combinations of algorithms have been presented in Figure 1(c). In essence, the proposed HR-tuned scheme provides up to 2.5X speedup over MVAPICH2-GDR (2.1).



(a) CNTK: MV2-GDR vs. MV2-GDR-NCCL up to 64 MPI Processes (1 per GPU) executing on 8 nodes



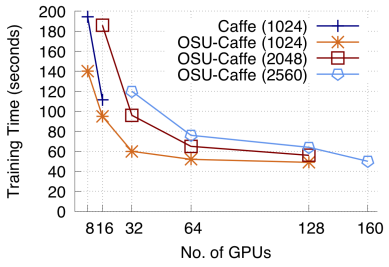
(b) CNTK: MV2-GDR-NCCL vs. MV2-GDR-Opt up to 128 MPI Processes (1 per GPU) executing on 8 nodes



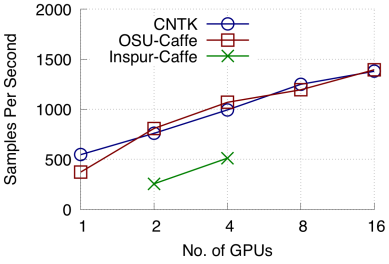
(c) MPI_Reduce: MV2-GDR vs. MV2-GDR-Opt for 160 Processes (GPUs)

Figure 1: Performance Benefits for Proposed MPI_Bcast and MPI_Reduce Designs

S-Caffe (OSU-Caffe): We present strong scaling numbers for GoogLeNet training on a 160 GPUs in Figure 2(a). We report speedups of 2.5X over 32 GPUs for GoogLeNet training on 160 GPUs. Figure 2(b) provides a weak-scaling performance comparison for S-Caffe, Inspur-Caffe, and CNTK for AlexNet training. A different 16-node cluster with one K-80 GPU per node was used. Due to technical difficulties, we could not run Inspur-Caffe for configurations other than two and four GPUs. S-Caffe and CNTK provide comparable performance for AlexNet training.



(a) Strong Scaling Results for GoogLeNet (ImageNet) up to 160 GPUs



(b) Weak Scaling Results for AlexNet: Comparison of S-Caffe, CNTK, and Inspur-Caffe (up to 16 GPUs)

Figure 2: Performance Evaluation of S-Caffe

4 FUTURE WORK

We plan to design and enhance MPI_Allreduce to cover the full spectrum of collectives being used by various other DL frameworks and toolkits.

REFERENCES

[1] 2017. High Performance Deep Learning (HiDL). http://hidl.cse.ohio-state.edu/. (2017).

[2] 2017. KESCH: Cray CS-Storm System. http://www.cscs.ch/computers/kesch_escha/index.html. (2017).

[3] 2017. MVAPICH2: MPI over InfiniBand, 10GigE/iWARP and RoCE. https://mvapich.cse.ohio-state.edu/. (2017).

[4] A. A. Awan, Ching-Hsiang Chu, H. Subramoni, and D. K. Panda. 2017. Optimized Broadcast for Deep Learning Workloads on Dense-GPU InfiniBand Clusters: MPI or NCCL?. In *eprint arXiv:1707.09414*. https://arxiv.org/abs/1707.09414

[5] 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. In *Proceedings of the 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP '17)*. ACM, New York, NY, USA, 193–205. http://doi.acm.org/10.1145/3018743.3018769

[6] A. A. Awan, K. Hamidouche, A. Venkatesh, and D. K. Panda. 2016. Efficient Large Message Broadcast using NCCL and CUDA-Aware MPI for Deep Learning. In *Proceedings of the 23rd European MPI Users' Group Meeting*. ACM, 15–22.