Accelerating Deep Learning with MVAPICH

OSU Booth Talk (SC '17)

Ammar Ahmad Awan, Hari Subramoni, and Dhabaleswar K. Panda

Network Based Computing Laboratory

Dept. of Computer Science and Engineering

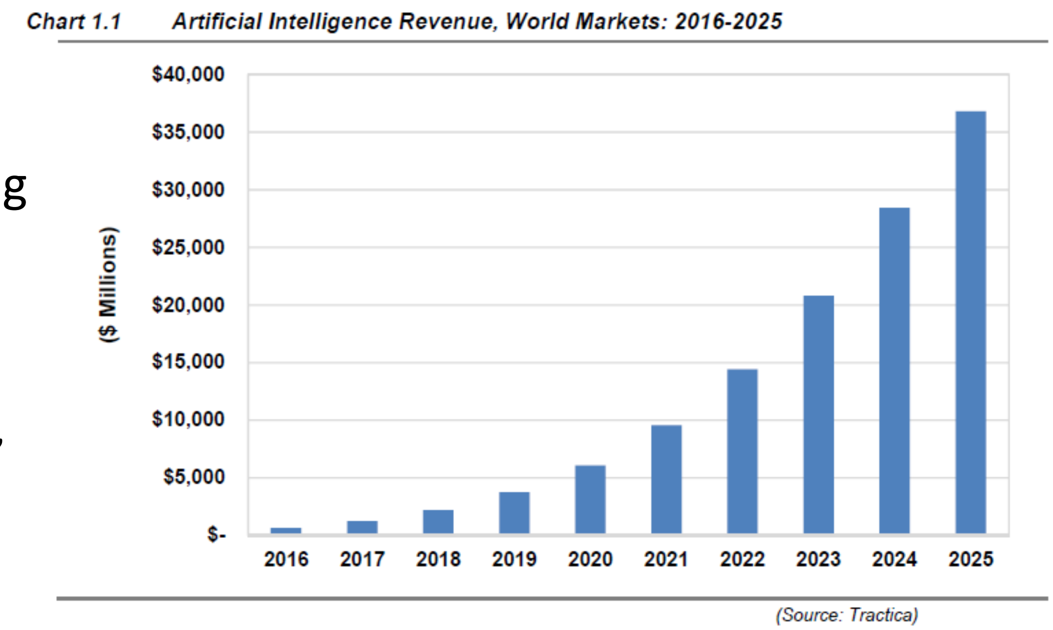
The Ohio State University

Agenda

- **Introduction**
  - **Deep Learning Trends**
  - **CPUs and GPUs for Deep Learning**
  - **Message Passing Interface (MPI)**
- Co-design Efforts
  - OSU-Caffe
  - NCCL-augmented MPI Broadcast
  - Large-message CUDA-Aware MPI Collectives
- Characterization of Deep Learning Workloads
  - CPUs vs. GPUs for Deep Learning with Caffe

DL Frameworks and Trends

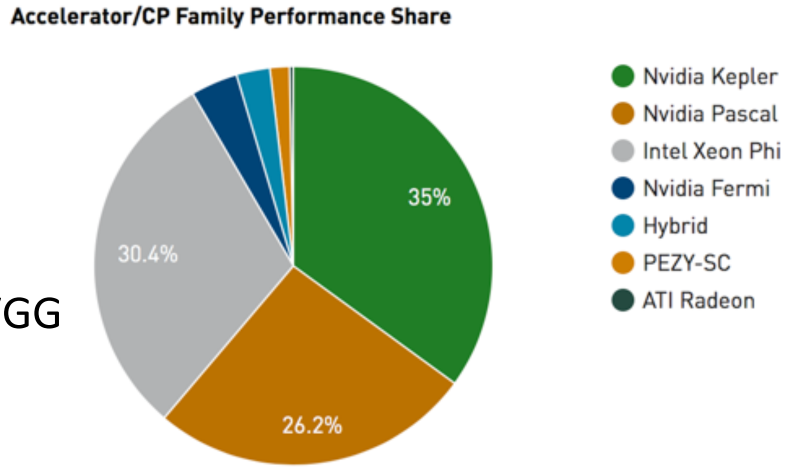
- **Caffe**, TensorFlow, CNTK and many more..
- Most frameworks are exploiting GPUs to accelerate training
- Diverse applications – Image Recognition, Cancer Detection, Self-Driving Cars, Speech Processing etc.



https://www.top500.org/news/market-for-artificial-intelligence-projected-to-hit-36-billion-by-2025/

GPUs are great for Deep Learning

- NVIDIA GPUs have been the main driving force for faster training of Deep Neural Networks (DNNs)
  - The ImageNet Challenge - (ILSVRC)
  - 90% of the ImageNet teams used GPUs in 2014*
  - DL models like AlexNet, GoogLeNet, and VGG
  - A natural fit for DL due to the throughput-oriented nature
  - GPUs are also growing in the HPC arena! → https://www.top500.org/statistics/list/

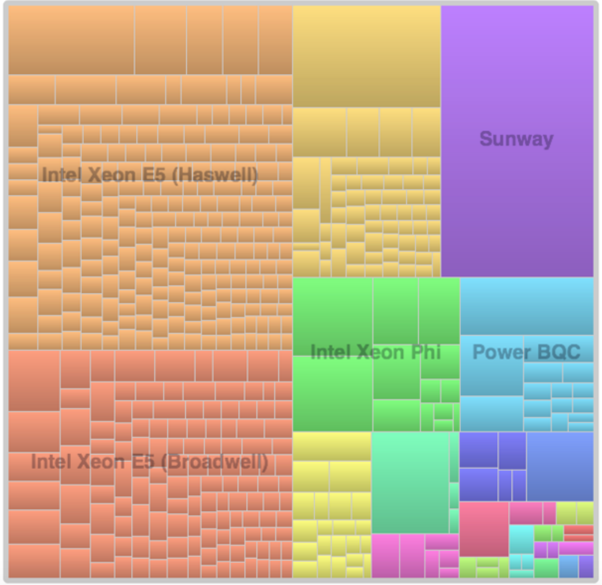


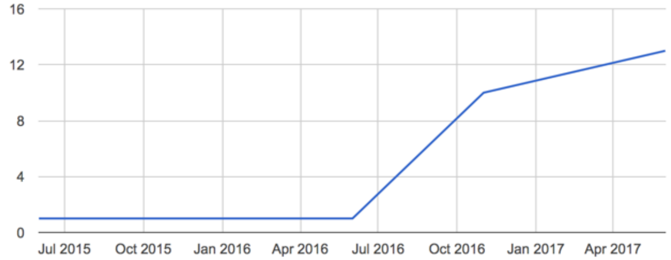
*https://blogs.nvidia.com/blog/2014/09/07/imagenet/

https://www.top500.org/statistics/list/

And CPUs are catching up fast

- Intel CPUs are everywhere and many-core CPUs are emerging according to Top500.org
- Host CPUs exist even on the GPU nodes
  - Many-core Xeon Phis are increasing
- Xeon Phi 1st generation was a co-processor
- **Unlike** Xeon Phi 2nd generation, which is a self-hosted processor!
- Usually, we hear CPUs are ***10x – 100x*** slower than GPUs? [1-3]
  - ***But can we do better?***





System Count for Xeon Phi

1- https://dl.acm.org/citation.cfm?id=1993516
2- http://ieeexplore.ieee.org/abstract/document/5762730/
3- https://dspace.mit.edu/bitstream/handle/1721.1/51839/MIT-CSAIL-TR-2010-013.pdf?sequence=1

What to use for scale-out? (Distributed training of Neural Nets.)

- What is Message Passing Interface (**MPI**)?
  - a de-facto standard for expressing distributed-memory parallel programming
  - used for communication between processes in multi-process applications
- ***MVAPICH2 is a high performance implementation of the MPI standard***

- **What can MPI do for Deep Learning?**
  - MPI has been used for large scale scientific applications
  - Deep Learning can also exploit MPI to perform high-performance communication
- **Why do I need communication in Deep Learning?**
  - If you use one GPU or one CPU, you do not need communication
  - But, one GPU or CPU is not enough!
  - DL wants as many compute elements as it can get!
  - ***MPI is a great fit – Broadcast, Reduce, and Allreduce is what most DL workloads require***

Overview of the MVAPICH2 Project

- High Performance open-source MPI Library for InfiniBand, Omni-Path, Ethernet/iWARP, and RDMA over Converged Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Started in 2001, First version available in 2002
  - MVAPICH2-X (MPI + PGAS), Available since 2011
  - ***Support for GPGPUs (MVAPICH2-GDR) and MIC (MVAPICH2-MIC), Available since 2014***
  - Support for Virtualization (MVAPICH2-Virt), Available since 2015
  - Support for Energy-Awareness (MVAPICH2-EA), Available since 2015
  - Support for InfiniBand Network Analysis and Monitoring (OSU INAM) since 2015
  - **Used by more than 2,825 organizations in 85 countries**
  - **More than 432,000 (> 0.4 million) downloads from the OSU site direc**
  - Empowering many TOP500 clusters (June '17 ranking)
    - **1st, 10,649,600-core (Sunway TaihuLight) at National Supercomputing Center in Wuxi, Chin**
    - 15th, 241,108-core (Pleiades) at NASA
    - 20th, 462,462-core (Stampede) at TACC
  - Available with software stacks of many vendors and Linux Distros (RedHat and Su
  - **http://mvapich.cse.ohio-state.edu**
- Empowering Top500 systems for over a decade
  - System-X from Virginia Tech (3rd in Nov 2003, 2,200 processors, 12.25 TFlops) ->
  - Sunway TaihuLight (1st in Jun'17, 10M cores, 100 PFlops)

Deep Learning Frameworks – CPUs or GPUs?

- There are several Deep Learning (DL) or DNN Training frameworks
  - Caffe, Cognitive Toolkit, TensorFlow, MXNet, and counting....
- Every (almost every) framework has been optimized for NVIDIA GPUs
  - cuBLAS and cuDNN have led to significant performance gains!
- But every framework is able to execute on a CPU as well
  - So why are we not using them?
  - Performance has been “terrible” and several studies have reported significant degradation when using CPUs (see nvidia.qwiklab.com)
- But there is hope, actually a lot of great progress here!
  - And MKL-DNN, just like cuDNN, has definitely rekindled this!!
  - Coupled with Intel Xeon Phi (Knights Landing or KNL) and MC-DRAM, the landscape for CPU-based DL looks promising..

The Key Question!

How to efficiently scale-out a Deep Learning (DL) framework and take advantage of heterogeneous High Performance Computing (HPC) resources like GPUs and Xeon Phi(s)?

Research Challenges

Various datasets and networks handled differently in DL frameworks

Possible strategies to evaluate the performance of DL frameworks

Performance trends that can be observed for a single node

Performance behavior for hardware features

Computation and communication characteristics of DL workloads?

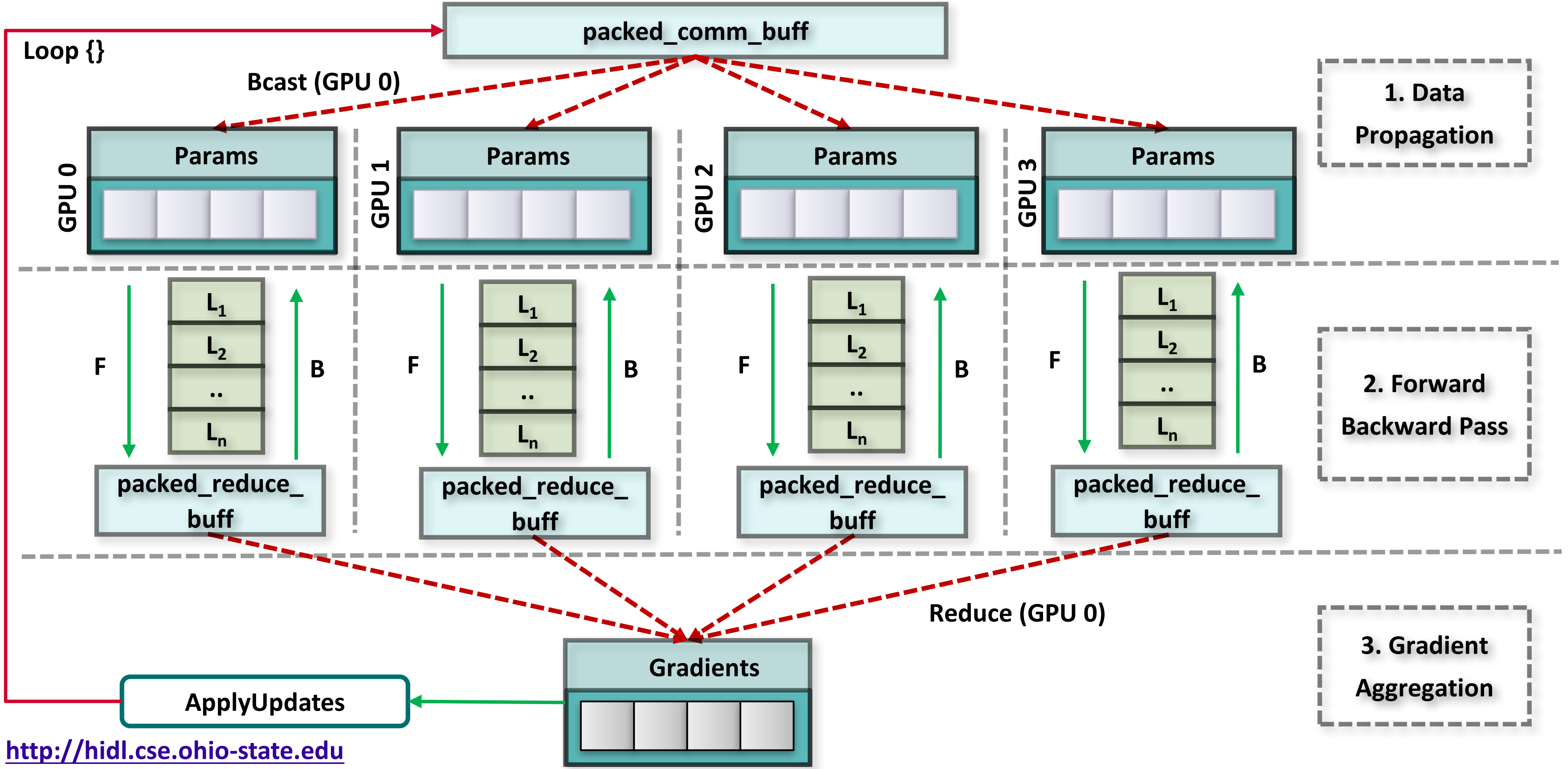
Scale-out of DNN training for CPU-based and GPU-based DNN training

Let us bring HPC and DL "together"!

Agenda

- Introduction
  - Deep Learning Trends
  - CPUs and GPUs for Deep Learning
  - Message Passing Interface (MPI)
- **Co-design Efforts**
  - **OSU-Caffe**
  - **NCCL-augmented MPI Broadcast**
  - **Large-message CUDA-Aware MPI Collectives**
- Characterization of Deep Learning Workloads
  - CPUs vs. GPUs for Deep Learning with Caffe

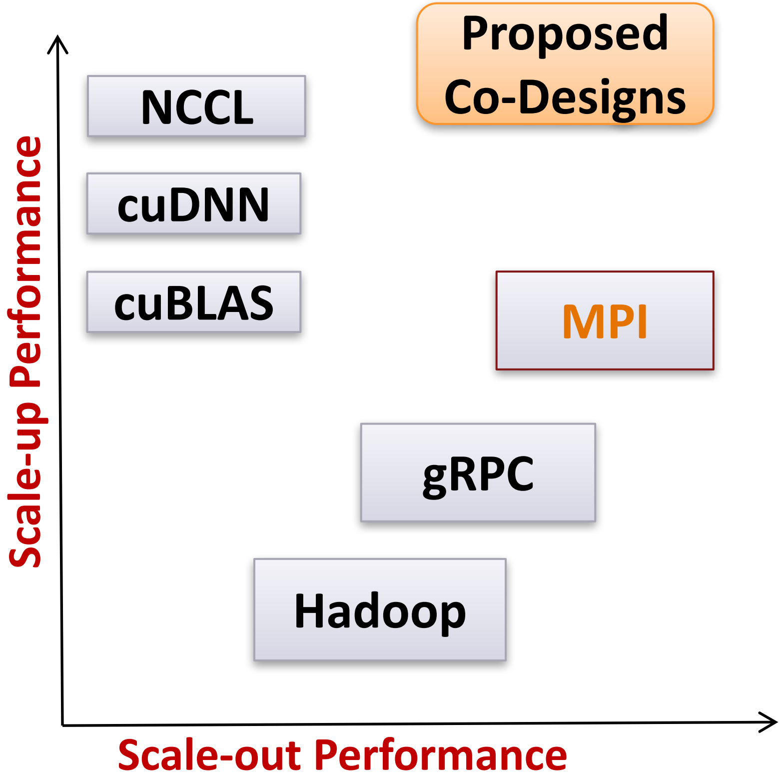
Caffe Architecture



http://hidl.cse.ohio-state.edu

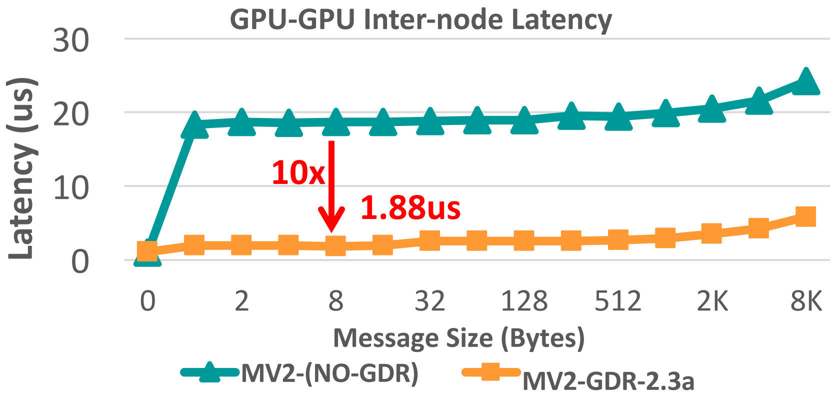
OSU-Caffe: Co-design to Tackle New Challenges for MPI Runtimes

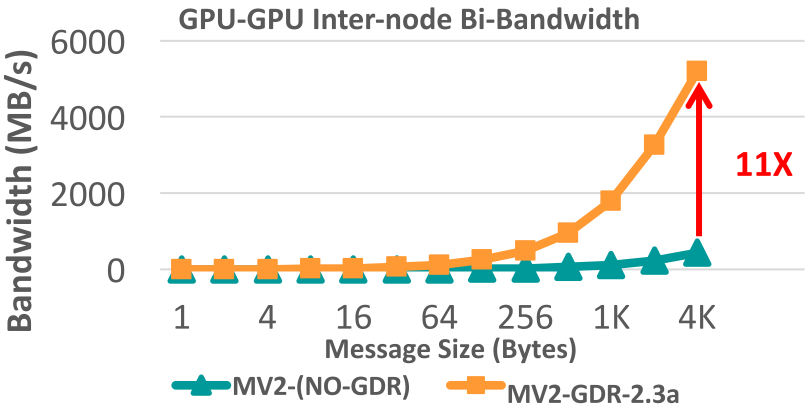
- Deep Learning frameworks are a different game altogether
  - Unusually large message sizes (order of megabytes)
  - Most communication based on GPU buffers
- Existing State-of-the-art
  - cuDNN, cuBLAS, NCCL --> **scale-up** performance
  - CUDA-Aware MPI --> **scale-out** performance
    - For small and medium message sizes only!
- Proposed: Can we **co-design** the MPI runtime (**MVAPICH2-GDR**) and the DL framework (**Caffe**) to achieve both?
  - Efficient **Overlap** of Computation and Communication
  - Efficient **Large-Message** Communication (Reductions)
  - What **application co-designs** are needed to exploit **communication-runtime co-designs**?

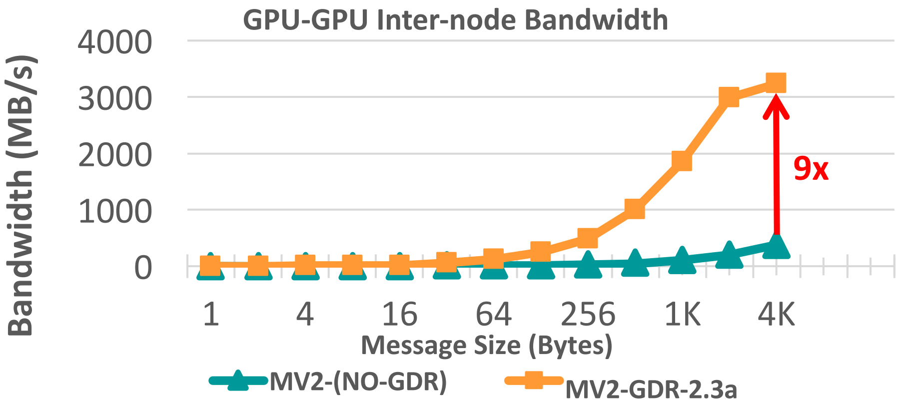


A. A. Awan, K. Hamidouche, J. M. Hashmi, and D. K. Panda, S-Caffe: Co-designing MPI Runtimes and Caffe for Scalable Deep Learning on Modern GPU Clusters. In *Proceedings of the 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (PPoPP '17)

MVAPICH2-GDR: Scale-out for GPU-based Distributed Training





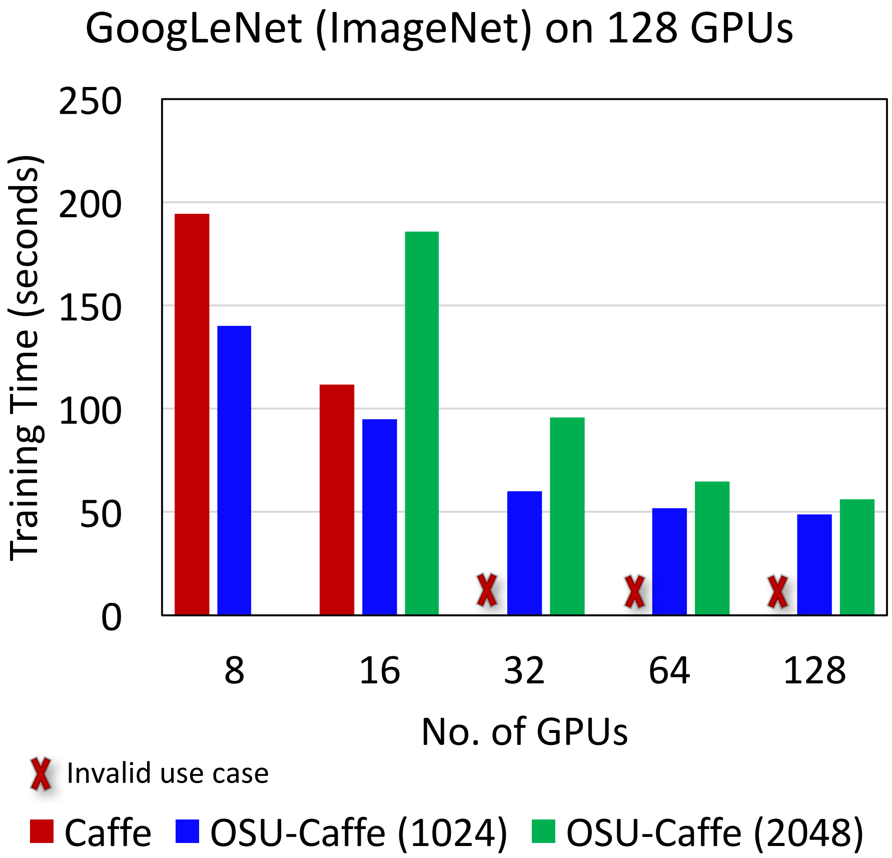


MVAPICH2-GDR-2.3a
Intel Haswell (E5-2687W) node - 20 cores
NVIDIA Volta V100 GPU
Mellanox Connect-X4 EDR HCA
CUDA 9.0
Mellanox OFED 4.0 with GPU-Direct-RDMA

MVAPICH2-GDR: Performance that meets Deep Learning requirements!

OSU-Caffe 0.9: Scalable Deep Learning on GPU Clusters

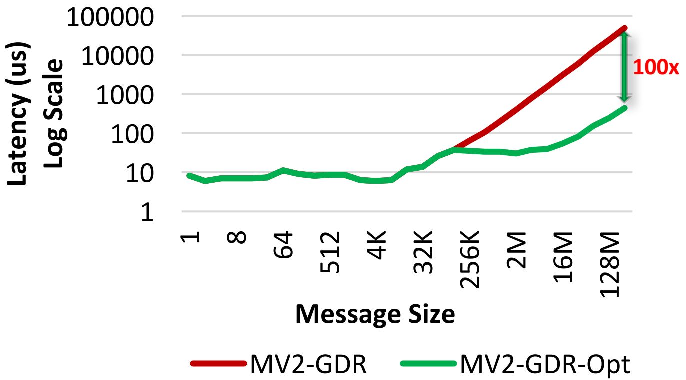
- Caffe : A flexible and layered Deep Learning framework.
- Benefits and Weaknesses
  - Multi-GPU Training within a single node
  - Performance degradation for GPUs across different sockets
  - Limited Scale-out
- OSU-Caffe: MPI-based Parallel Training
  - Enable Scale-up (within a node) and Scale-out (across multi-GPU nodes)
  - Scale-out on 64 GPUs for training CIFAR-10 network on CIFAR-10 dataset
  - Scale-out on 128 GPUs for training GoogLeNet network on ImageNet dataset



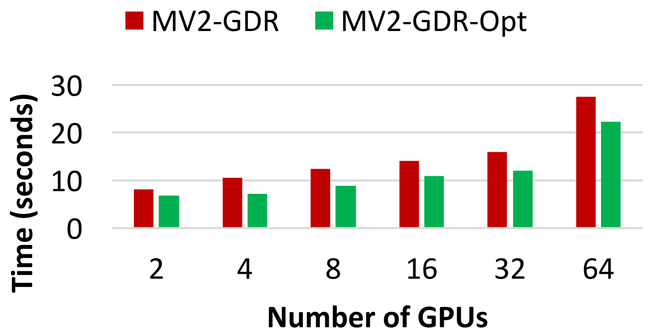
OSU-Caffe 0.9 available from HiDL site

Efficient Broadcast for MVAPICH2-GDR using NVIDIA NCCL

- NCCL has some limitations
  - Only works for a single node, thus, no scale-out on multiple nodes
  - Degradation across IOH (socket) for scale-up (within a node)
- We propose optimized MPI_Bcast
  - Communication of very large GPU buffers (order of megabytes)
  - Scale-out on large number of dense multi-GPU nodes
- Hierarchical Communication that efficiently exploits:
  - CUDA-Aware MPI_Bcast in MV2-GDR
  - NCCL Broadcast primitive



Performance Benefits: OSU Micro-benchmarks



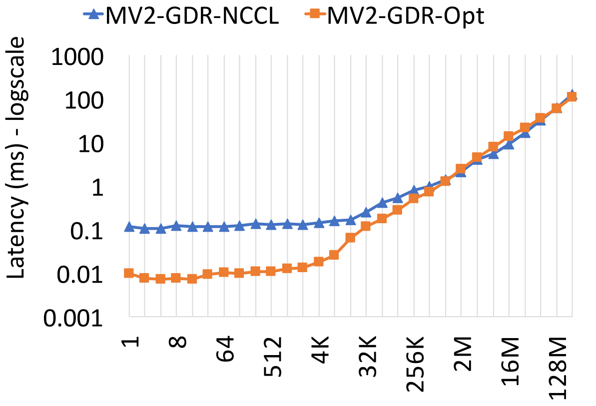
Performance Benefits: Microsoft CNTK DL framework

(25% avg. improvement)

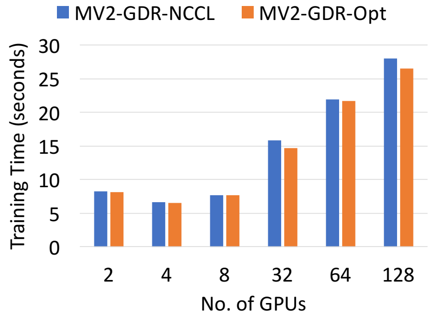
Efficient Large Message Broadcast using NCCL and CUDA-Aware MPI for Deep Learning, A. Awan , K. Hamidouche , A. Venkatesh , and D. K. Panda, The 23rd European MPI Users' Group Meeting (EuroMPI 16), Sep 2016 [Best Paper Runner-Up]

Pure MPI Large Message Broadcast

- 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)

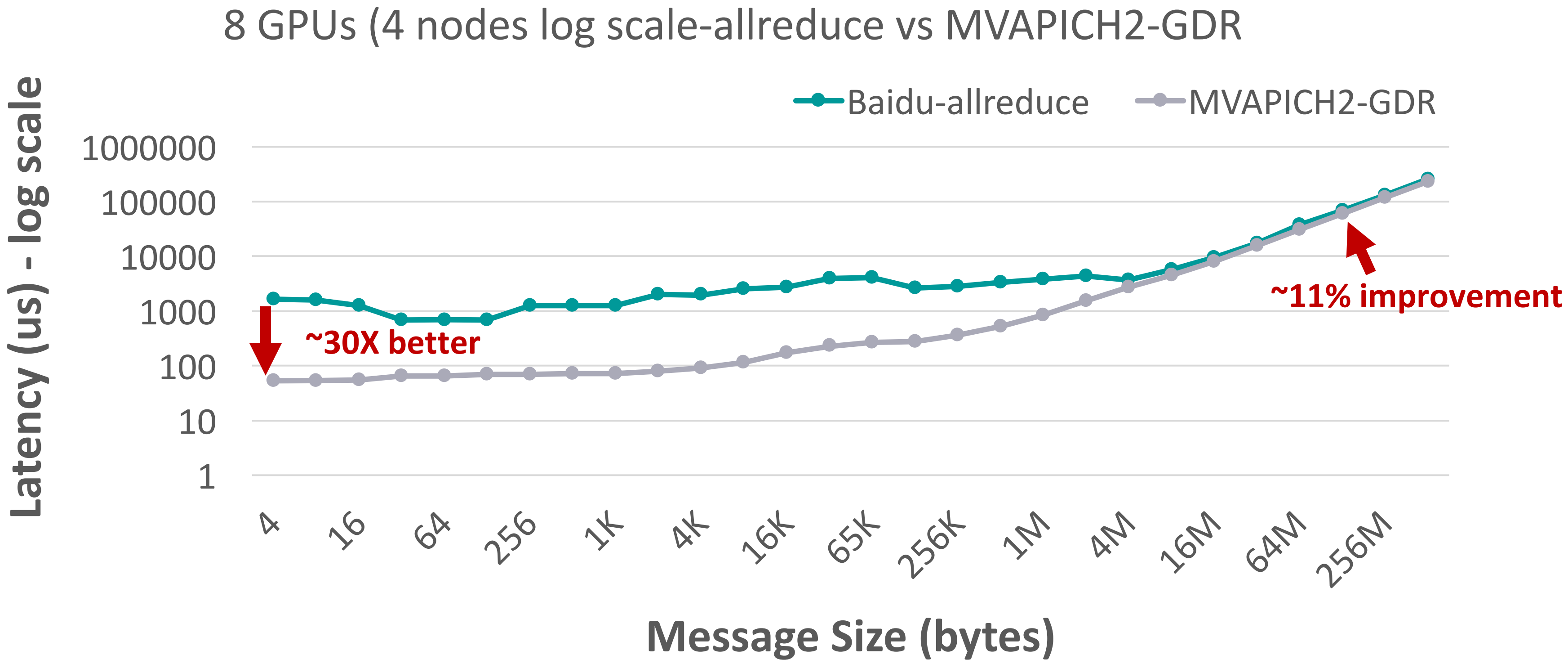


VGG Training with CNTK

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)

Large Message Allreduce: MVAPICH2-GDR vs. Baidu-allreduce

- Performance gains for MVAPICH2-GDR 2.3a* compared to Baidu-allreduce

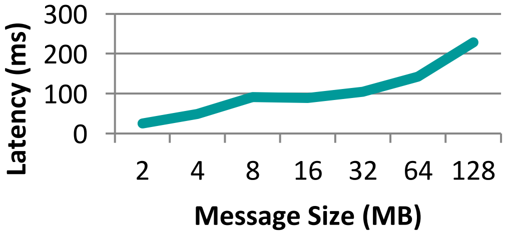


**Available with MVAPICH2-GDR 2.3a*

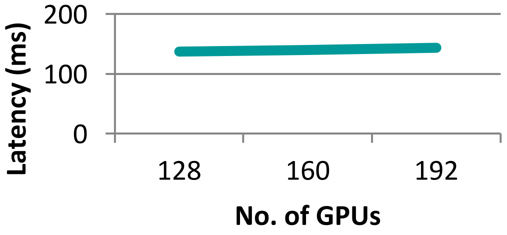
Large Message Optimized Collectives for Deep Learning

- MVAPICH2-GDR provides optimized collectives for **large message sizes**
- Optimized Reduce, Allreduce, and Bcast
- **Good scaling with large number of GPUs**
- **Available in MVAPICH2-GDR 2.2 and higher**

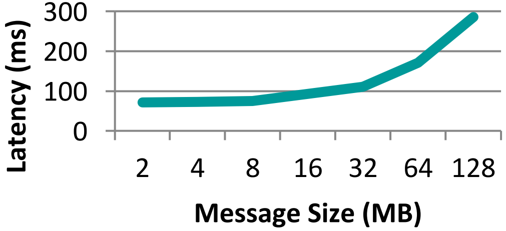
Reduce – 192 GPUs



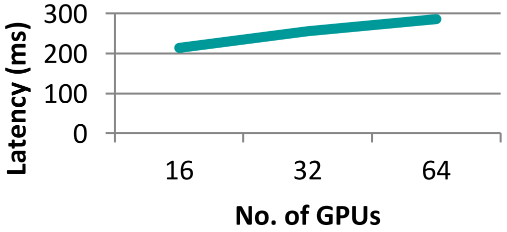
Reduce – 64 MB



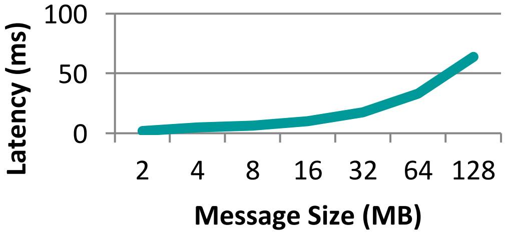
Allreduce – 64 GPUs



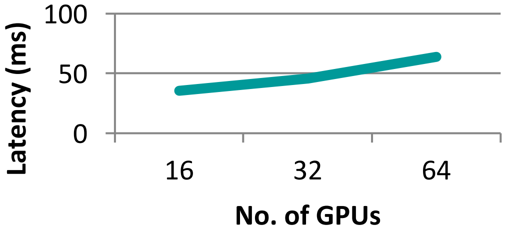
Allreduce - 128 MB



Bcast – 64 GPUs



Bcast 128 MB

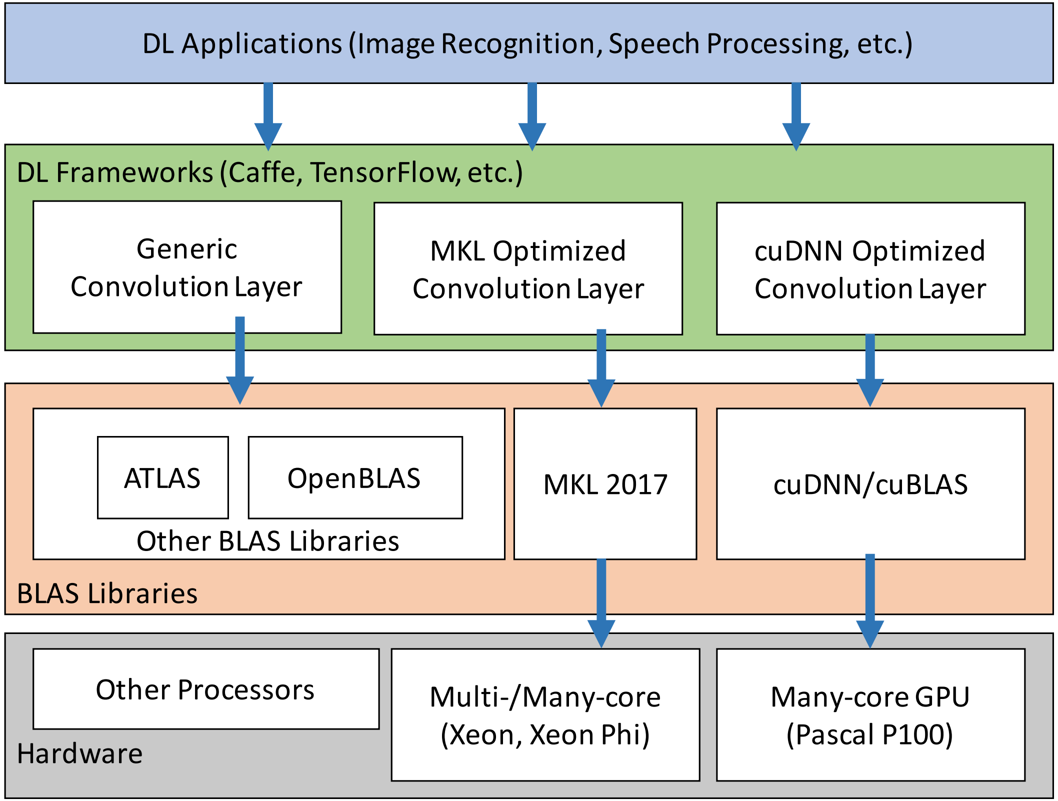


Agenda

- Introduction
  - Deep Learning Trends
  - CPUs and GPUs for Deep Learning
  - Message Passing Interface (MPI)
- Co-design Efforts
  - OSU-Caffe
  - NCCL-augmented MPI Broadcast
  - Large-message CUDA-Aware MPI Collectives
- **Characterization of Deep Learning Workloads**
  - **CPUs vs. GPUs for Deep Learning with Caffe**

Understanding the Impact of Execution Environments

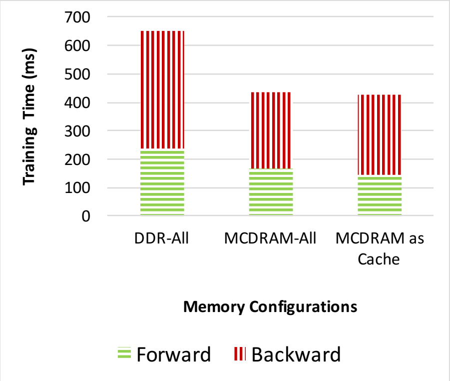
- Performance depends on many factors
- Hardware Architectures
  - GPUs
  - Multi-/Many-core CPUs
  - Software Libraries: cuDNN (for GPUs), MKL-DNN/MKL 2017 (for CPUs)
- Hardware and Software co-design
  - Software libraries optimized for one platform will not help the other!
  - cuDNN vs. MKL-DNN



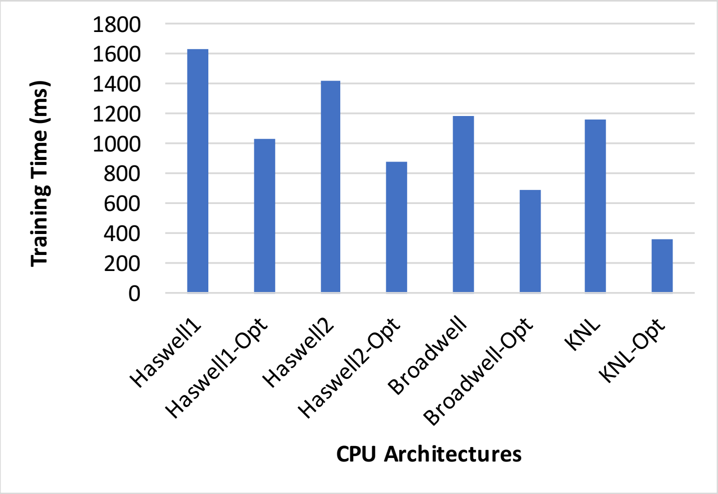
A. A. Awan, H. Subramoni, D. Panda, "An In-depth Performance Characterization of CPU- and GPU-based DNN Training on Modern Architectures" 3rd Workshop on Machine Learning in High Performance Computing Environments, held in conjunction with SC17, Nov 2017.

Impact of MKL engine and MC-DRAM for Intel-Caffe

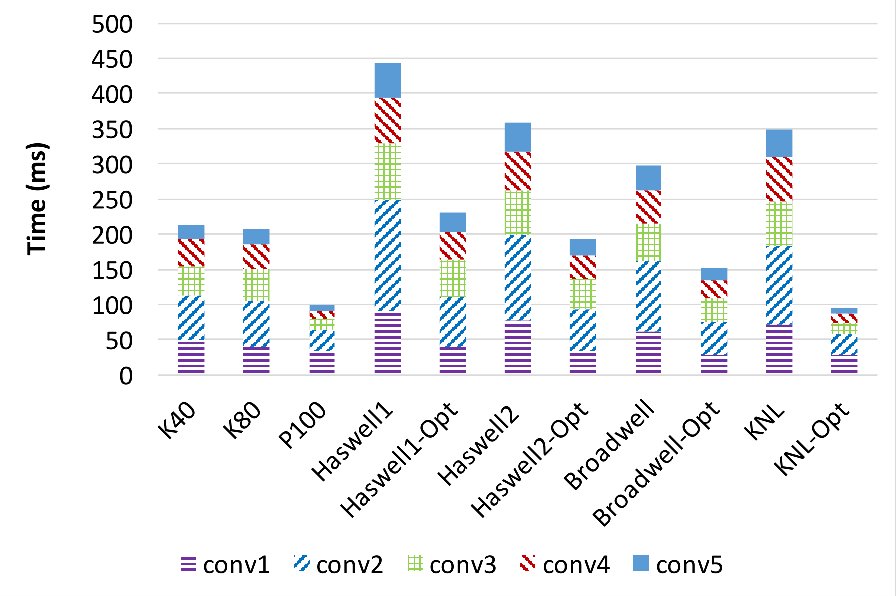
- We use ***MCDRAM as Cache*** for all the subsequent results
- On average, DDR-All is up to ***1.5X slower*** than MCDRAM

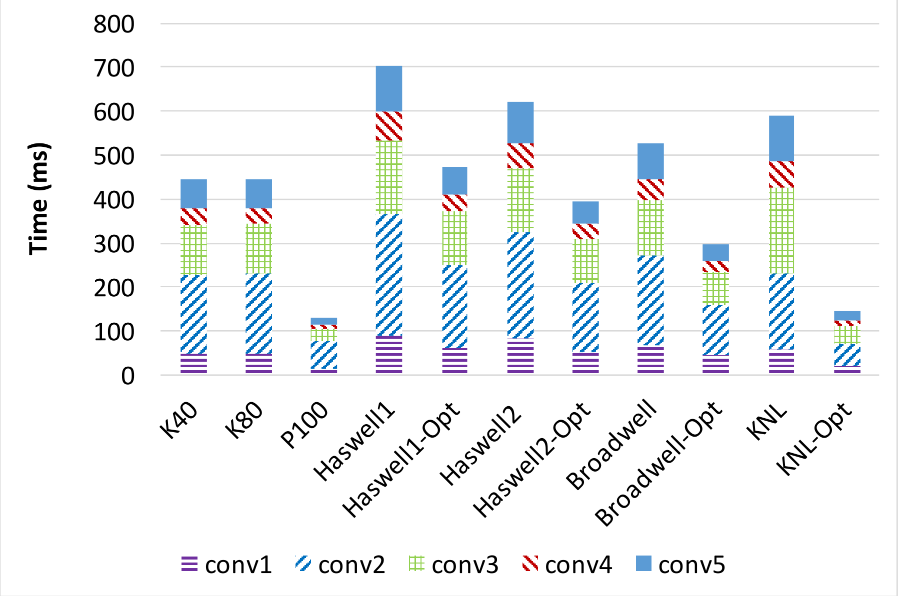


- MKL engine is up to ***3X better*** than default Caffe engine
- ***Biggest*** gains for ***Intel Xeon Phi*** (many-core) architecture
- Both Haswell and Broadwell architectures get significant speedups (***up to 1.5X***)



The Full Landscape for AlexNet Training

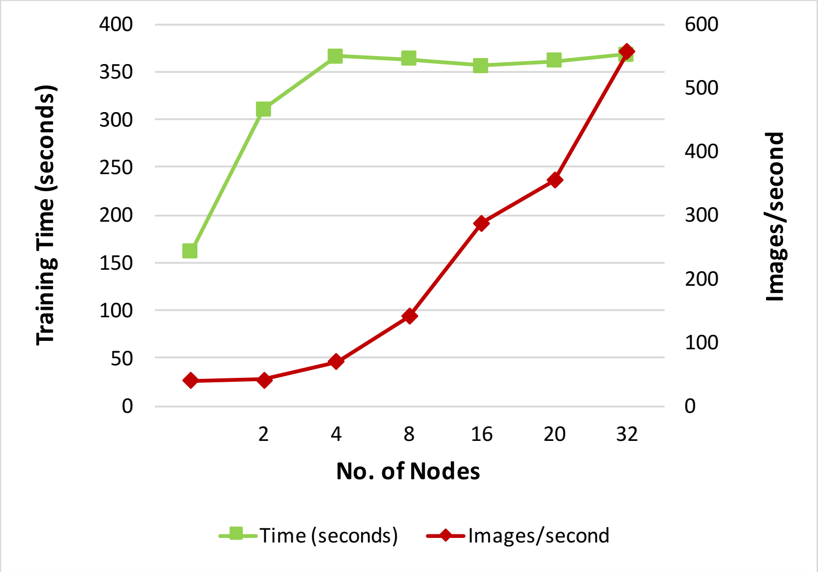




- Convolutions in the Forward and Backward Pass
- ***Faster Convolutions → Faster Training***
- Most performance gains are based on ***conv2*** and ***conv3***.

Multi-node Results: ResNet-50

- All results are ***weak scaling***
  - The batch size remains constant per solver but increases overall by:
  - *Batch-size * #nodes or*
  - *Batch-size * #gpus*
- Images/second is a derived metric but more meaningful for understanding scalability
- Efficiency is another story [1]
  - ***Larger DNN architectures → Less scalability due to communication overhead***



ResNet-50 Intel-Caffe

1. *Experiences of Scaling TensorFlow On Up to 512 Nodes On CORI Supercomputer*, Intel HPC Dev. Con., https://www.intel.com/content/www/us/en/events/hpcdevcon/overview.html

Summary

- Deep Learning is on the rise
  - Rapid advances in software, hardware, and availability of large datasets is driving it
- Single node or single GPU is not enough for Deep Learning workloads
- We need to focus on distributed Deep Learning but there are many challenges
- MPI offers a great abstraction for communication in DL Training tasks
- A co-design of Deep Learning frameworks and communication runtimes will be required to make DNN Training scalable

Thank You!

awan.10@osu.edu

http://web.cse.ohio-state.edu/~awan.10

Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

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

HiDL
High-Performance Deep Learning

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

MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

The High-Performance MPI/PGAS Project
http://mvapich.cse.ohio-state.edu/

Please join us for other events at SC '17

- Workshops
  - ESPM2 2017: Third International Workshop on Extreme Scale Programming Models and Middleware
- Tutorials
  - InfiniBand, Omni-Path, and High-Speed Ethernet for Dummies
  - InfiniBand, Omni-Path, and High-Speed Ethernet: Advanced Features, Challenges in Designing, HEC Systems and Usage
- BoFs
  - MPICH BoF: MVAPICH2 Project: Latest Status and Future Plans
- ACM SRC Posters
  - Co-designing MPI Runtimes and Deep Learning Frameworks for Scalable Distributed Training on GPU Clusters
  - High-Performance and Scalable Broadcast Schemes for Deep Learning on GPU Clusters
- Booth Talks
  - The MVAPICH2 Project: Latest Developments and Plans Towards Exascale Computing
  - Exploiting Latest Networking and Accelerator Technologies for MPI, Streaming, and Deep Learning: An MVAPICH2-Based Approach
  - Accelerating Deep Learning with MVAPICH
  - MVAPICH2-GDR Library: Pushing the Frontier of HPC and Deep Learning

Please refer to http://mvapich.cse.ohio-state.edu/talks/ for more details