High Performance Computing
2nd presentation

2017/10/10

Toshiki Tsuchikawa(B4)

Selected Paper

- S-Caffe: Co-designing MPI Runtimes and Caffe for Scalable Deep Learning on Modern GPU Clusters (PPoPP '17)

Authors: Ammar Ahmad Awan
and 3 other (The Ohio State University)

Outline

- Abstract
- Introduction
- Preliminaries
- Challenges and Requirements
- Proposed Architecture and Co-design
- Performance Evaluation
- Conclusion

Abstract

Abstract

- Most DL frameworks like Caffe,Torch,Tensorflow,and CNTK which have been limited to a single node.
- S-caffe is proposed to scale out DL frameworks and bring HPC capabilities to the DL arena.

Introduction

Introduction

- MPI(Message Passing Interface)
  - popular and widely used parallel programming model for large-scale application.
- CUDA
  - parallel computing platform for GPU(graphics processing unit) created by NVIDIA.
- the current DL frameworks have not used MPI +CUDA techniques.

Introduction

Modern DL framework's features

Deep Learning Frameworks	Distributed Adress Space Systems			
	Basic MPI Support	CUDA-Aware MPI	Overlapped Designs(NBC support)	Co-Designed with MPI runtimes
Caffe	×	×	×	×
FireCaffe	✓	Unknown	×	Unknown
MPI-Caffe	✓	×	×	×
CNTK	✓	×	×	×
Inspur-Caffe	✓	✓	×	×
S-Caffe	✓	✓	✓	✓

NBC・・・Non-blocking Collective

Preliminaries

Preliminaries

Cuda-Aware MPI

- Not Cuda-Aware MPI

```
// MPIランク0
cudaMemcpy （s_buf_h 、s_buf_d 、size 、cudaMemcpyDeviceToHost ）;
MPI_Send関数（s_buf_h 、サイズ、MPI_CHAR 、1 、100 、MPI_COMM_WORLD ）

// MPIランク1
MPI_RECV （r_buf_h 、サイズ、MPI_CHAR 、0 、100 、MPI_COMM_WORLD 、&状態)
cudaMemcpy （r_buf_d 、r_buf_h 、サイズ、cudaMemcpyHostToDevice ）;
```

Preliminaries

Cuda-Aware MPI

- Not Cuda-Aware MPI

```
// MPIランク0
cudaMemcpy （s_buf_h 、s_buf_d 、size 、cudaMemcpyDeviceToHost ）;
MPI_Send関数（s_buf_h 、サイズ、MPI_CHAR 、1 、100 、MPI_COMM_WORLD ）

// MPIランク1
MPI_RECV （r_buf_h 、サイズ、MPI_CHAR 、0 、100 、MPI_COMM_WORLD 、&状態)
cudaMemcpy （r_buf_d 、r_buf_h 、サイズ、cudaMemcpyHostToDevice ）;
```

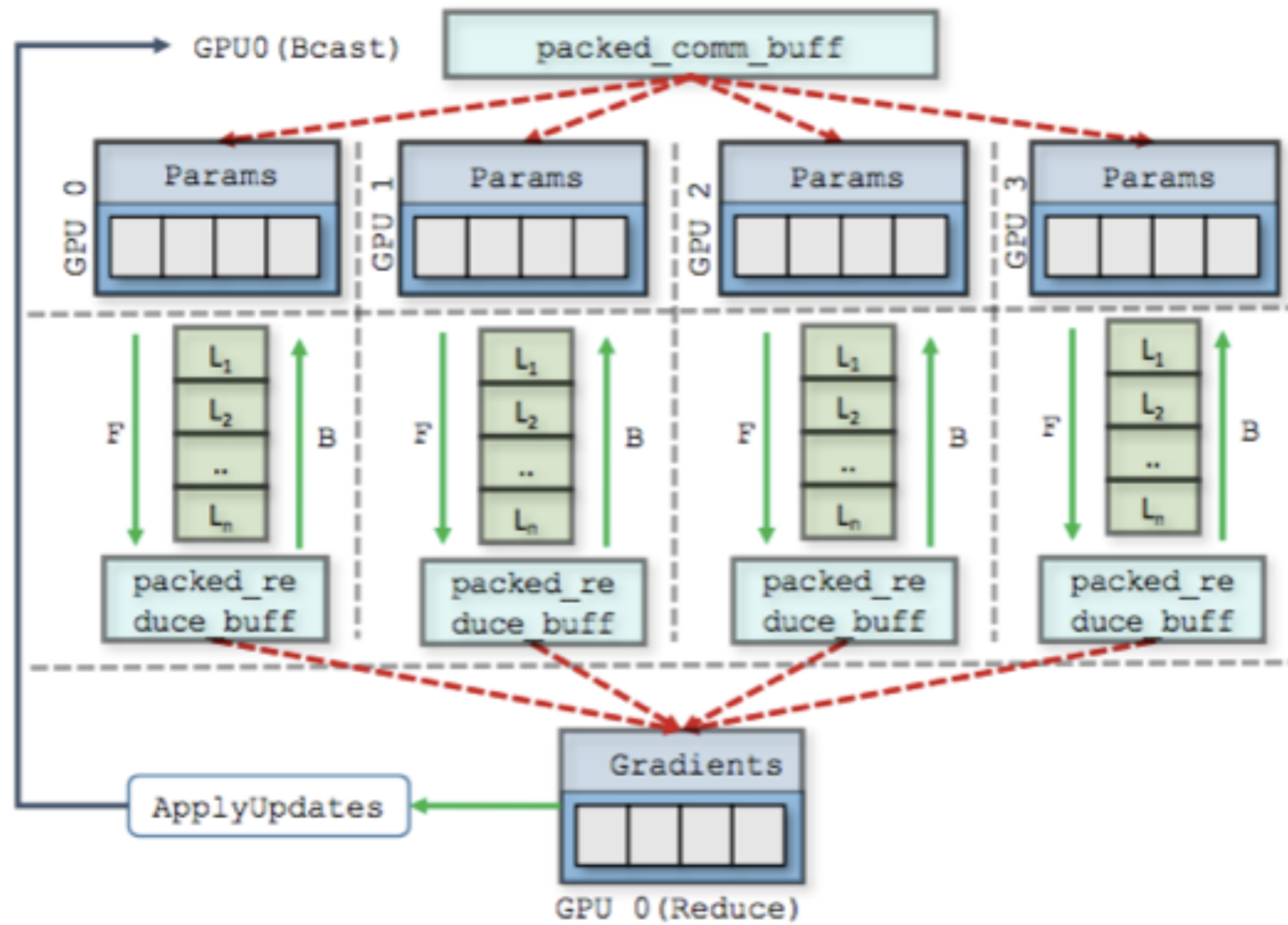
- Cuda-Aware MPI

```
// MPIランク0
MPI_Send関数（s_buf_d 、サイズ、MPI_CHAR 、1 、100 、MPI_COMM_WORLD ）

// MPIランクN-1
MPI_RECV （r_buf_d 、サイズ、MPI_CHAR 、0 、100 、MPI_COMM_WORLD 、&状態)
```

Preliminaries

- <u>Caffe Architecture</u>
  - forward pass generates a loss value by using parameter and Data
  - backward pass calculated parameter gradient
  - Applyupdates update parameter for the next iteration



Challenges and Requirements for Designing Scalable DL Frameworks

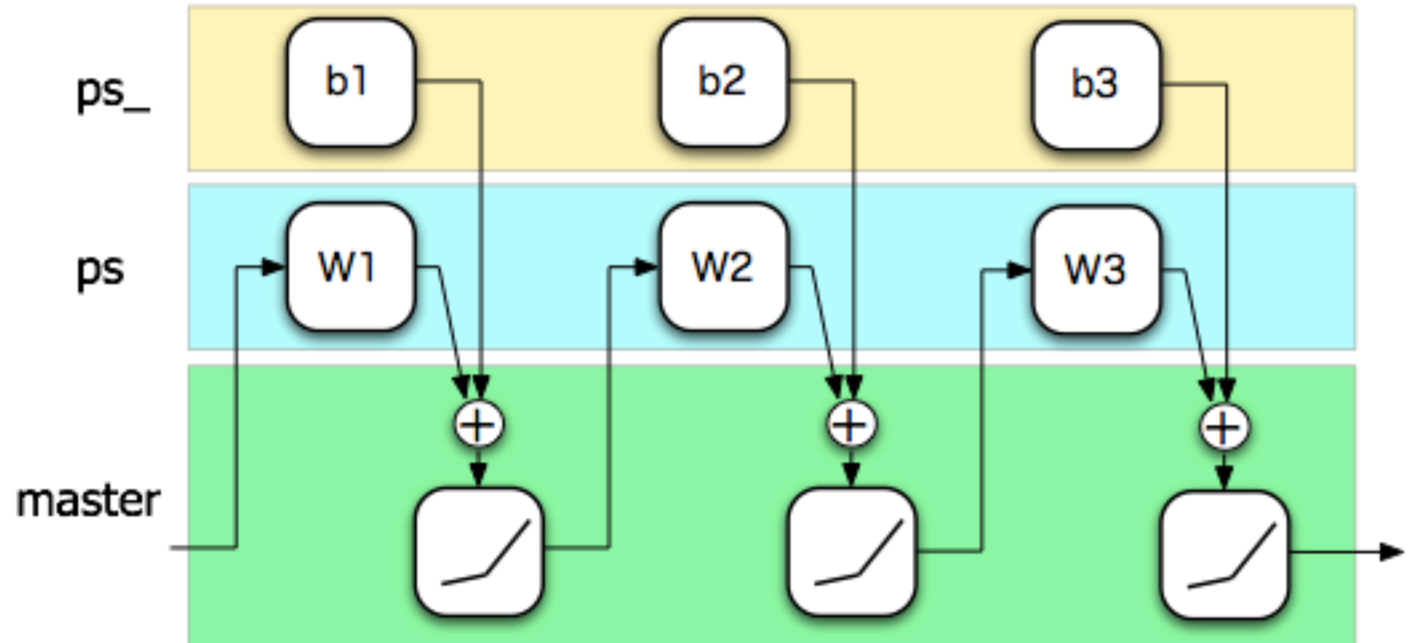
Challenges and Requirements

- S-Caffe adopts Data-parallel approach not Model-parallel

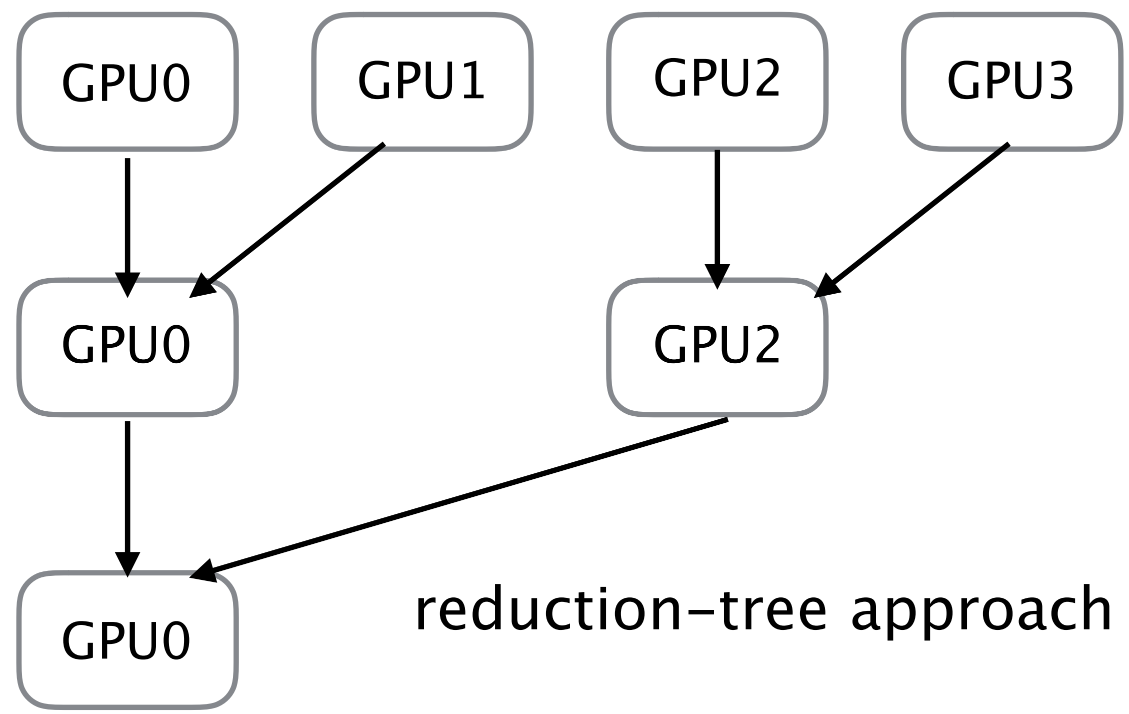
  Data-parallel
  - the same model is replicated for every processing element (a CPU core or a GPU), but is fed with different parts of the training data.

Challenges and Requirements

- Data-parallel has two different design choice. parameter-server approach and reduction-tree approach
- If many parameter, parameter-server approach to be the bottle-neck.
- S-caffe adopts reduction-tree approach.



parameter-server approach



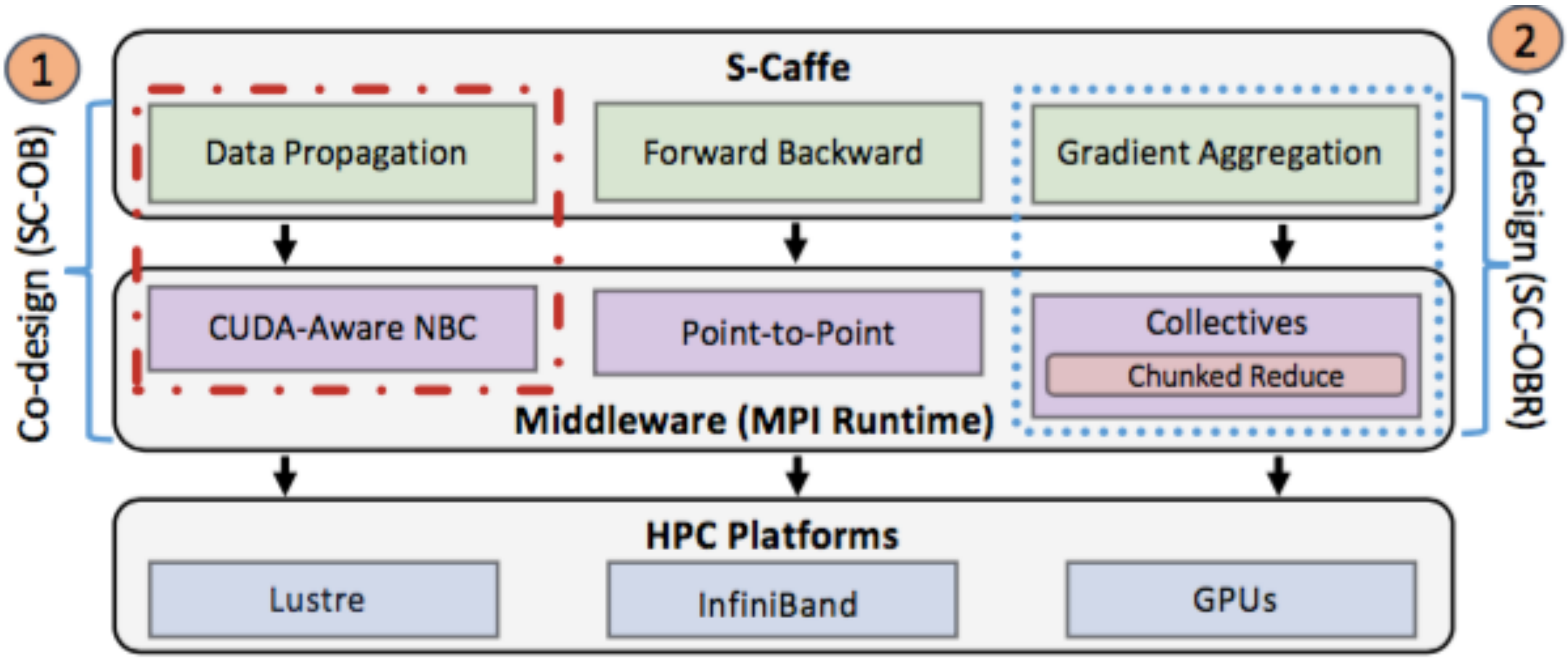
reduction-tree approach

Challenges and Requirements

- Other Challenge
  - Distributed Address-Space Design and Parallel Data Reading
    - Caffe has been designed for a single address space system
    - a single process can use multiple threads to take advantage of multiple GPUs in a node.
  - Exploiting Overlap of Computation and Communication

Proposed Architecture and Co-design

Proposed Architecture and Co-design



Proposed Architecture and Co-design

Basic CUDA-Aware MPI-design(SC-B)

- Avoids unnecessary copies between the CPU and GPUs by using CUDA-Aware MPI.
- Parallel Readers
  - LMDB does not scale for more than 64 parallel readers
  - it can be optimized for parallel file systems like Lustre.
  - Proposed design achieves scalability up to 160 GPUs.

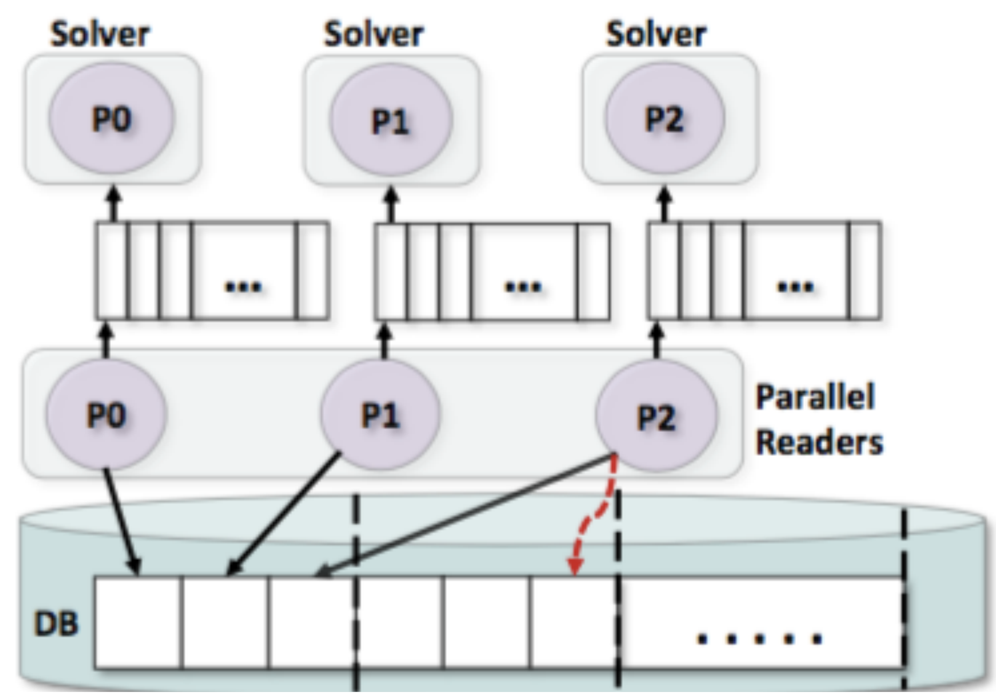
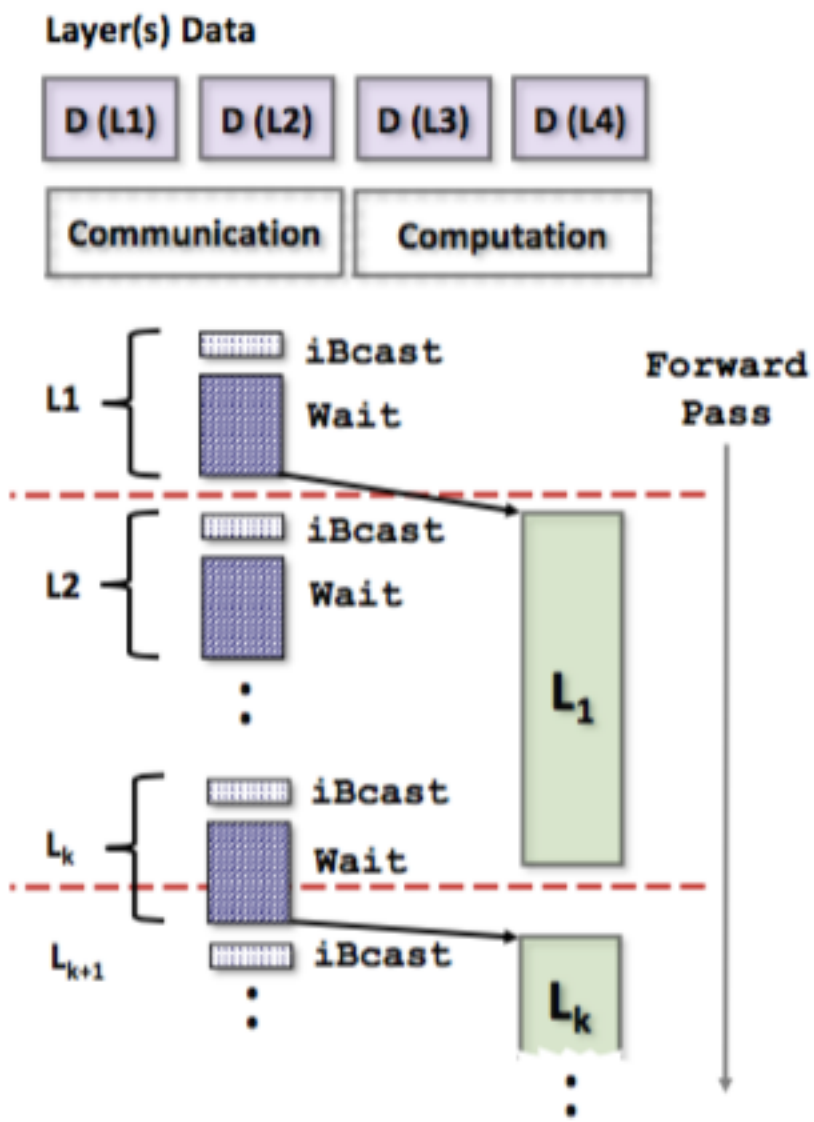
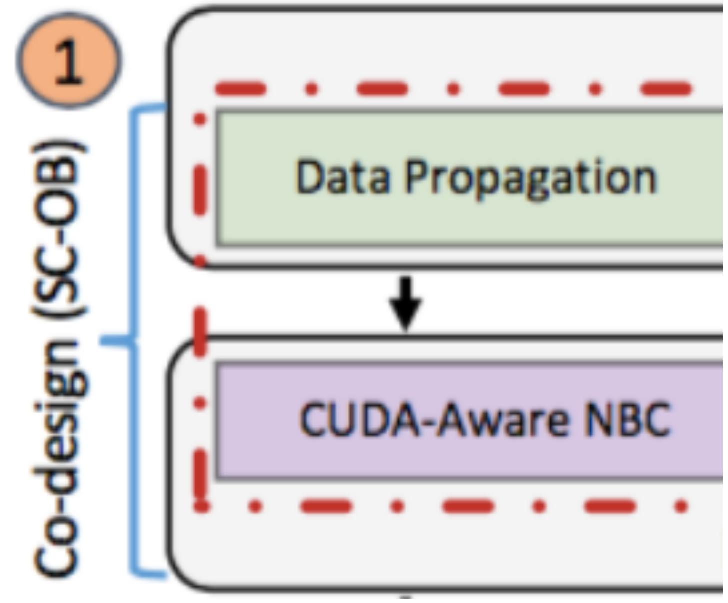


Figure 3. S-Caffe: Proposed Parallel Data Reader Design with Distributed Queues

Proposed Architecture and Co-design

Conventional method(SC-B)

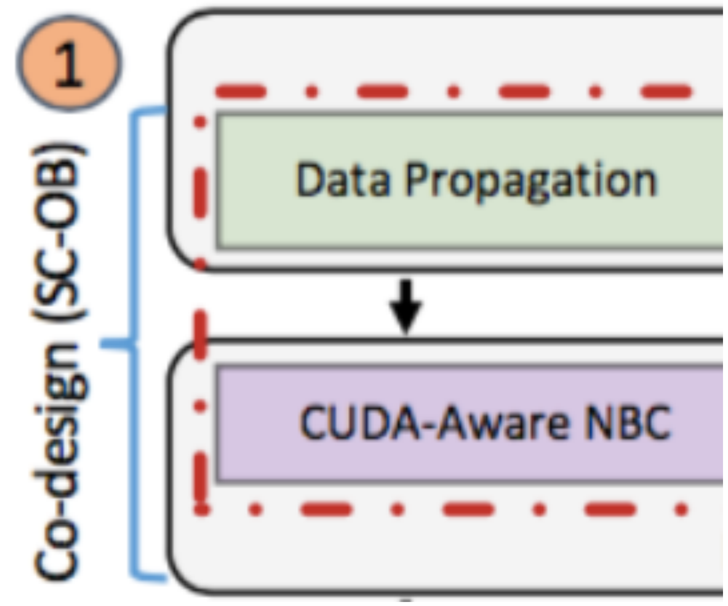


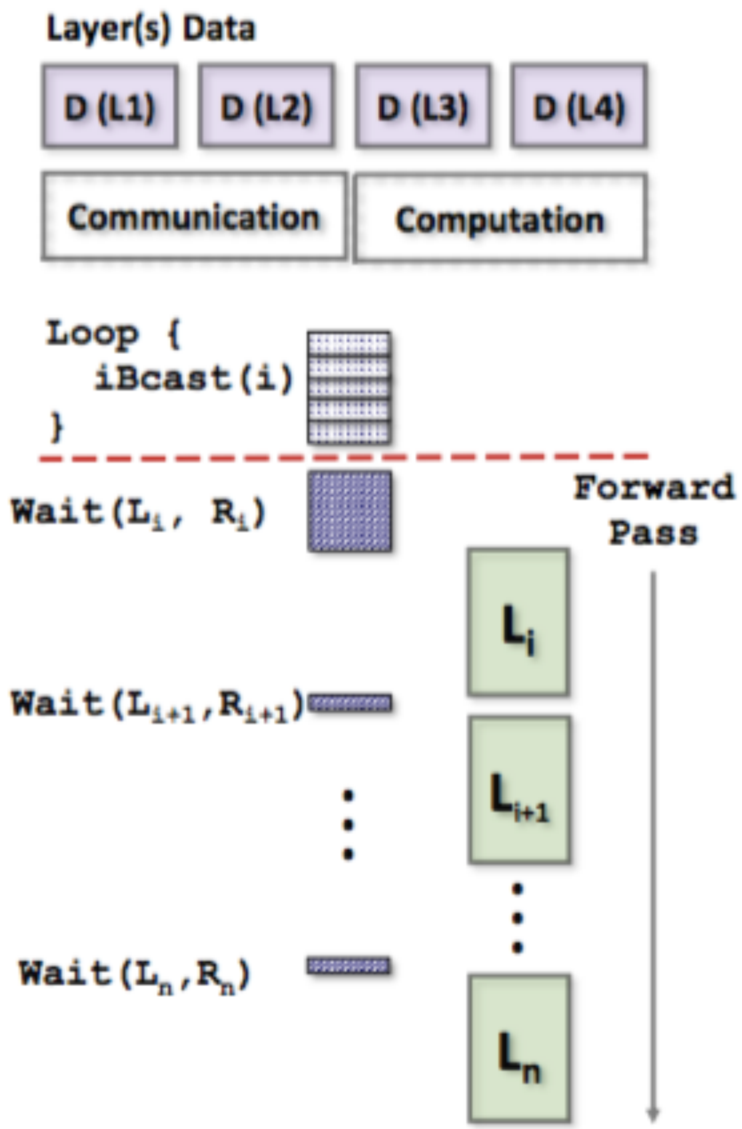


- Conventional design limits the asynchronous progress(because Wait called soon after called iBcast.)

Proposed Architecture and Co-design

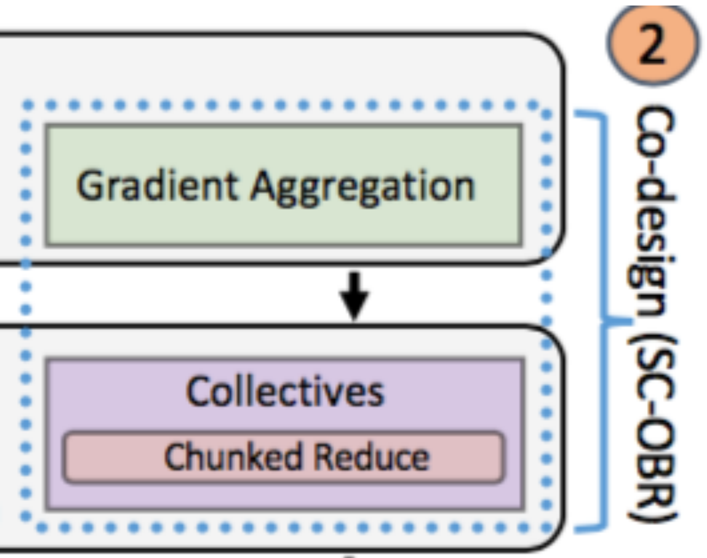
proposed method(SC-OB)



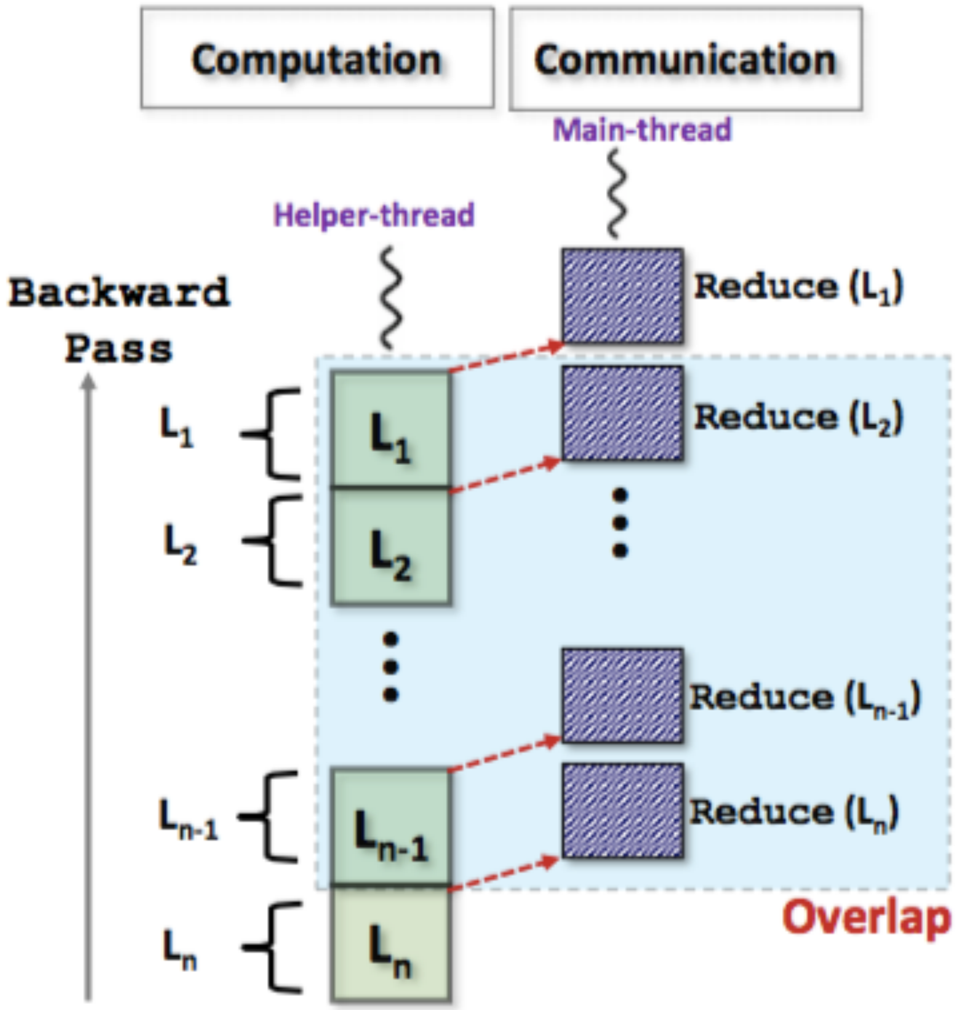


- This method starts all ibcast operations at the beginning
- Wait operation of ith Ibcast just before the ith Forward pass

Proposed Architecture and Co-design

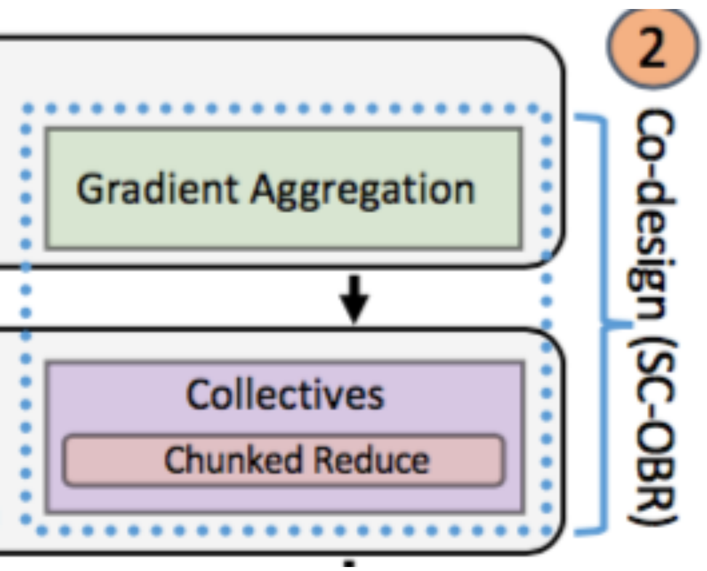


proposed method(SC–OBR)



- the n0th layer's reduce requires the completion of the n0th layer's computation.
- This method overlap n'th reduce and n−1'th computing.

Proposed Architecture and Co-design



DL-Aware Hierarchical Reduce(HR)

- In the example below lower level communication uses chunked chain algorithm and upper level communication Binomial Tree Algorithm.

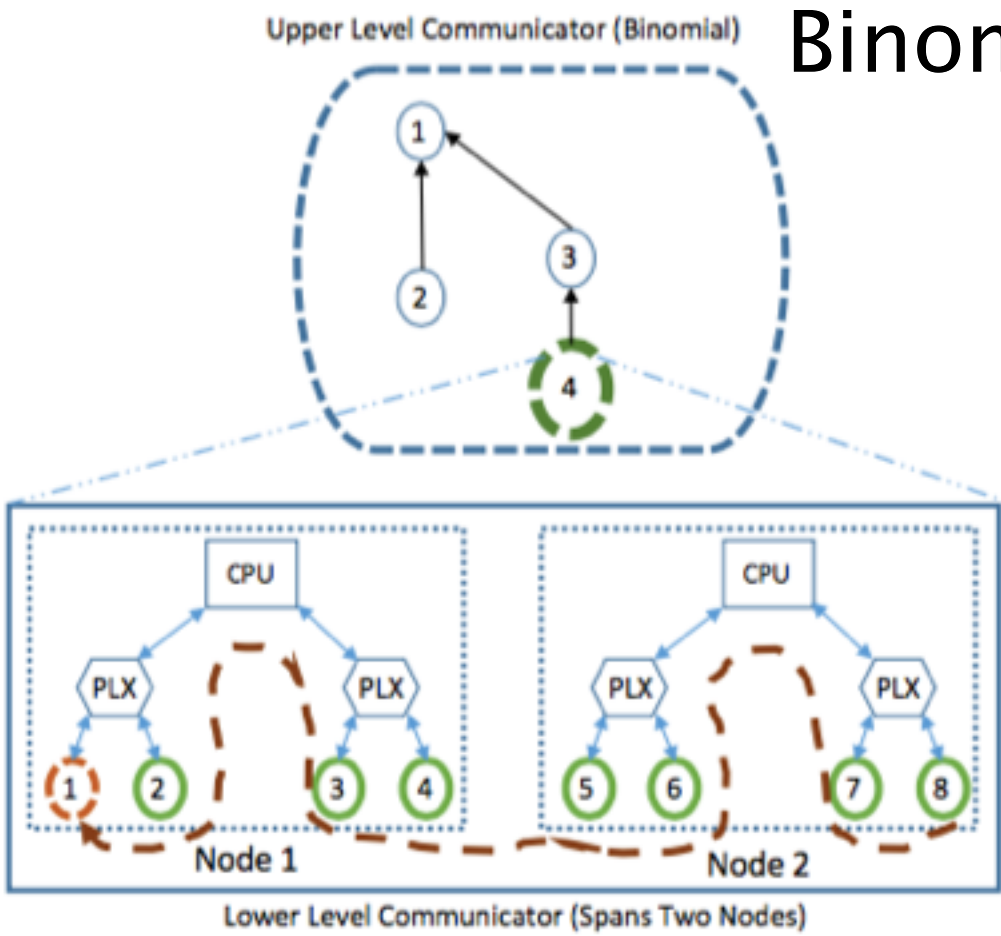


Figure 7. Hierarchical DL-Aware Reduction Design with a Chain-Binomial Combination

Performance Evaluation

Performance Evaluation

- DL model
  - GoogleNet, AlexNet
- Dataset
  ILSVRC 2012, CIFAR10
- Used GPU cluster(Cray CS-storm)
  Cluster-A
  - consisting of 12 hybrid nodes each containing 8 NVIDIA K-80 GK210GL GPUs.
  - total of 192 GPUs for the 12 nodes
  Cluster-B
  - consisting of 20 nodes each containing 1 NVIDIA K-80 GK210GL GPUs.

  (1 NVIDIA K-80 GK210GL contains two GPUs.)

Performance Evaluation

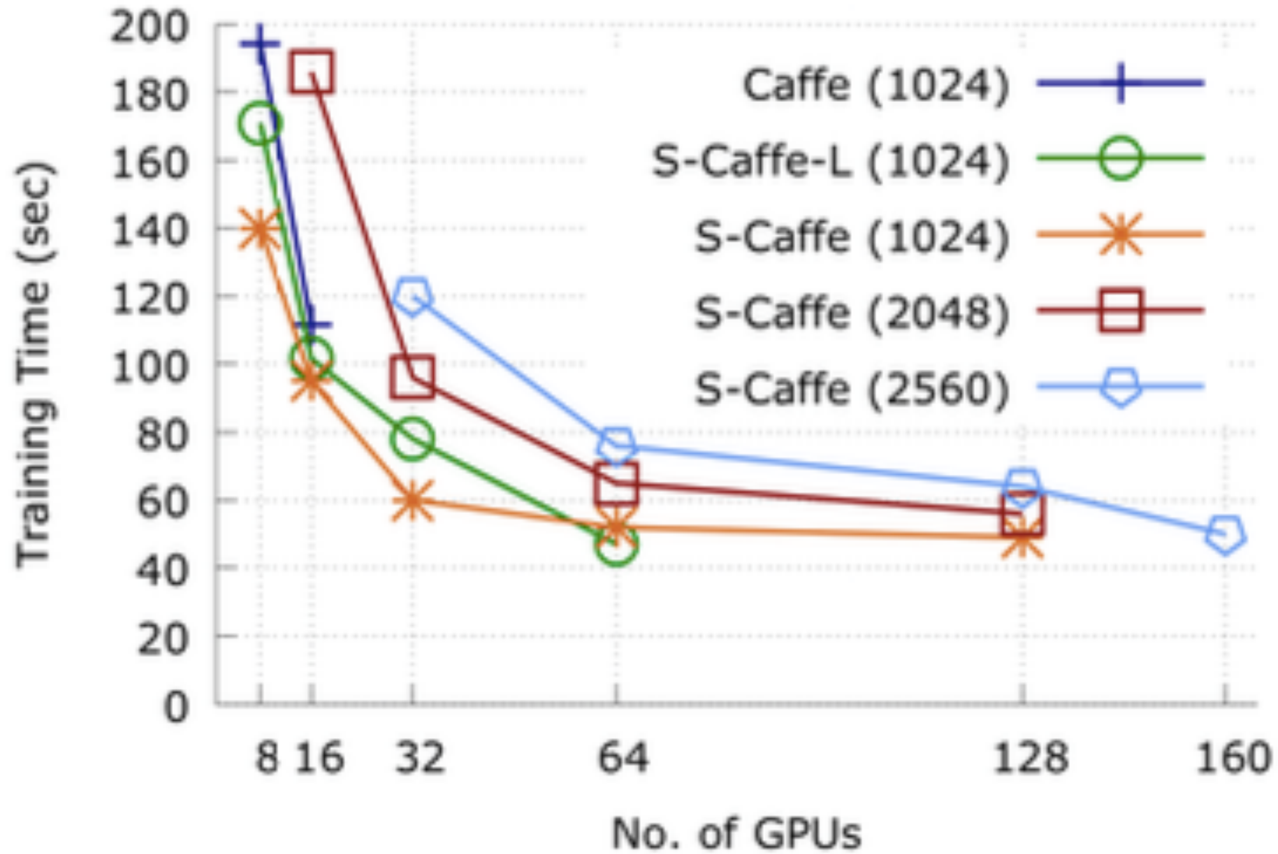


Figure 8. GoogLeNet: Comparison of S-Caffe (up to 160 GPUs) and Caffe (up to 16 GPUs) on Cluster-A

- () is batch-size.
- S-Caffe-L means that we have utilized LMDB database.
- S-Caffe utilizes Lustre file system.

Performance Evaluation

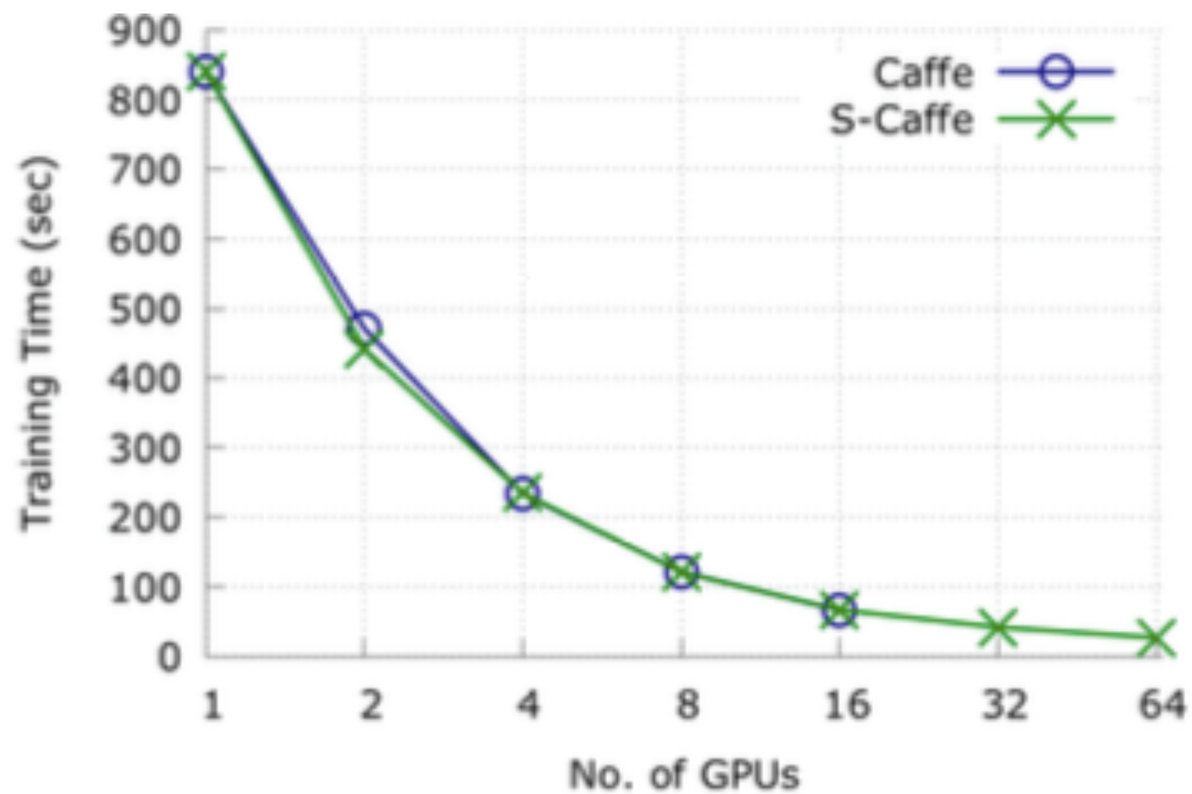


Figure 9. CIFAR10: Comparison of S-Caffe (up to 64 GPUs) and Caffe (up to 16 GPUs) on Cluster-A

- S-Caffe and Caffe is almost the same performance. It is present that S-Caffe does not suffer any overhead.

Performance Evaluation

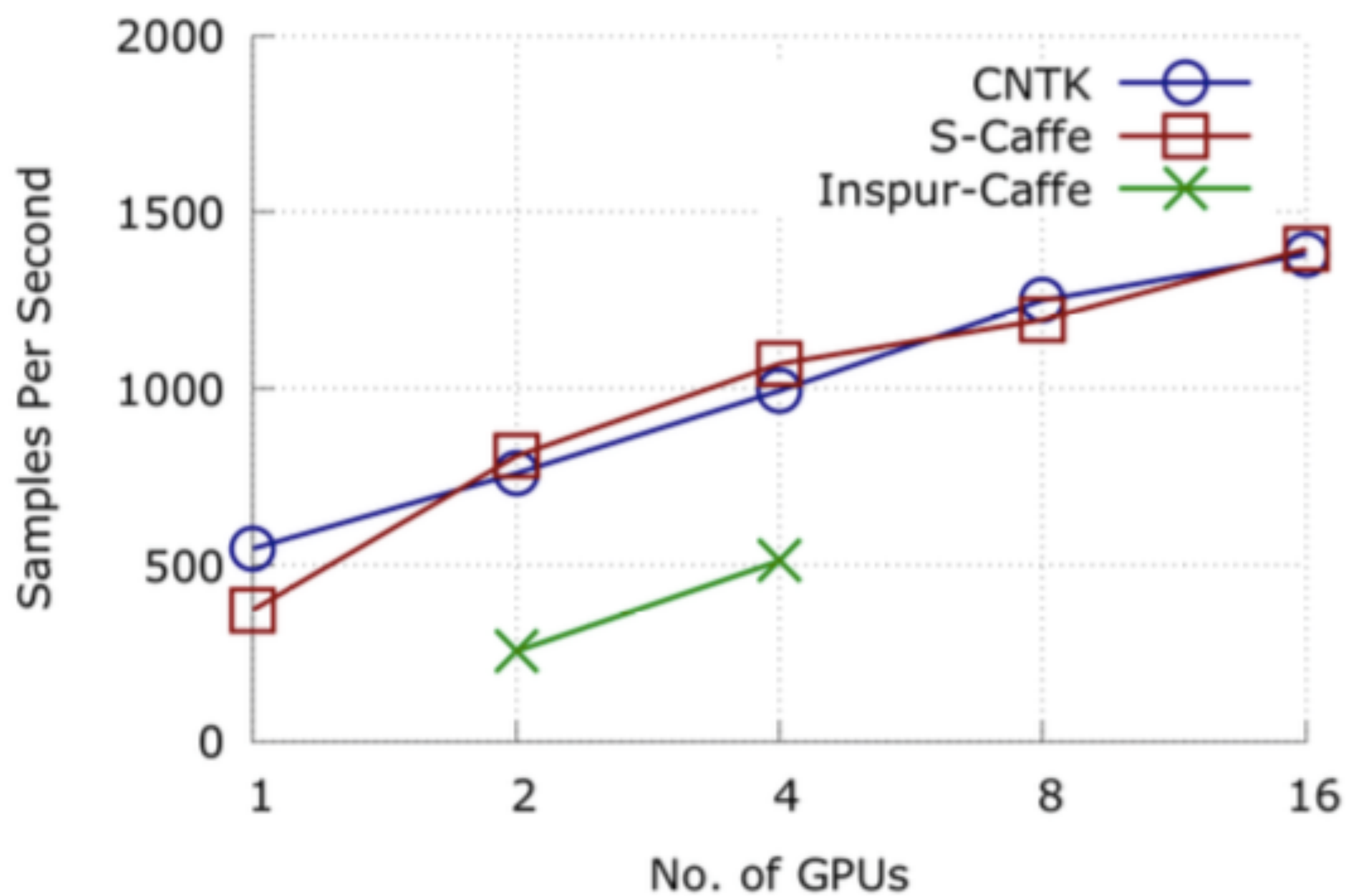
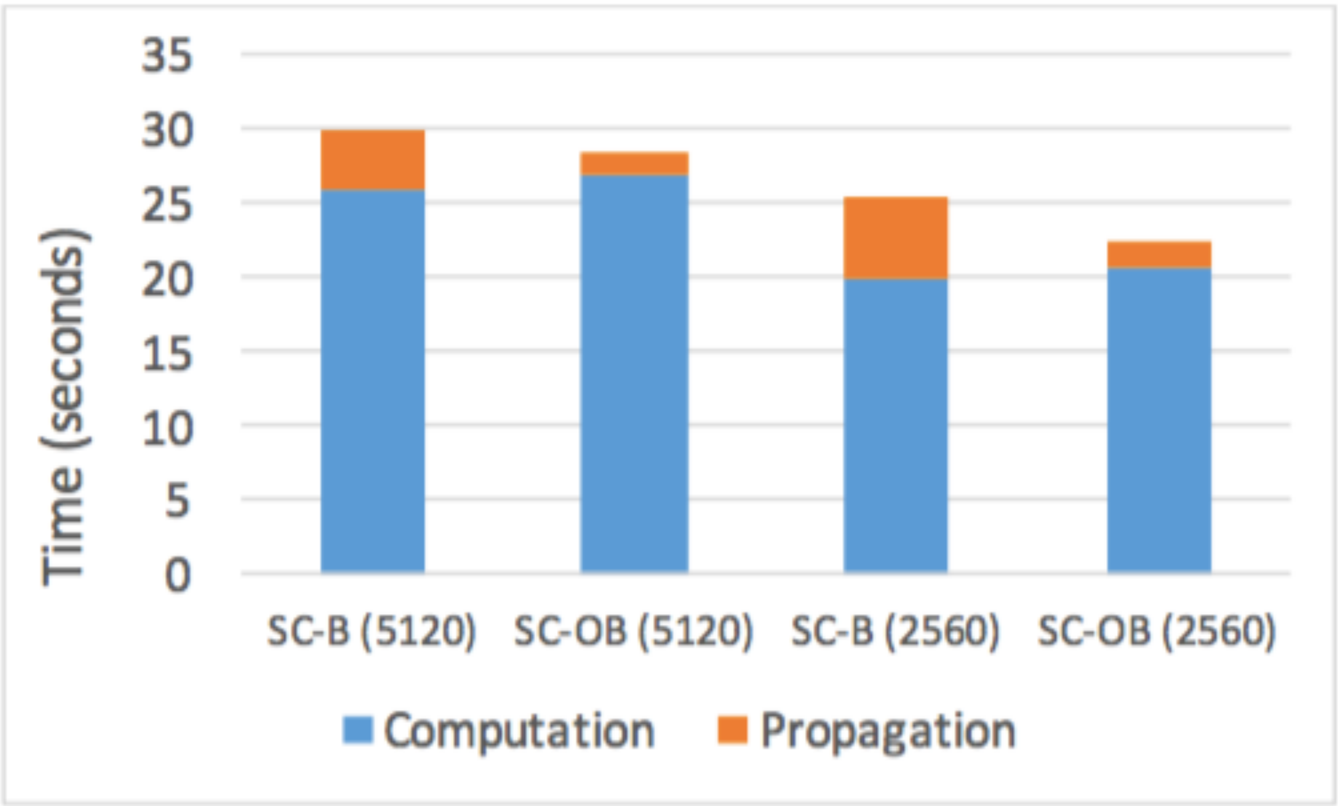


Figure 10. AlexNet: Comparison of S-Caffe, CNTK, and Inspur-Caffe (up to 16 GPUs) on Cluster-B

- Higher Samples Per Second donate better performance.
- Inspur-Caffe, which is an MPI-based parameter-server implementation.
- Microsoft CNTK, which is also an MPI-based framework.

Performance Evaluation



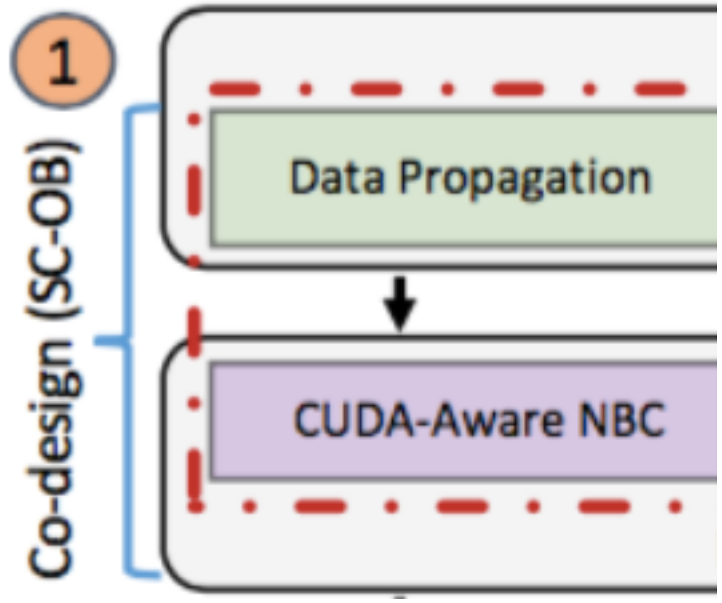
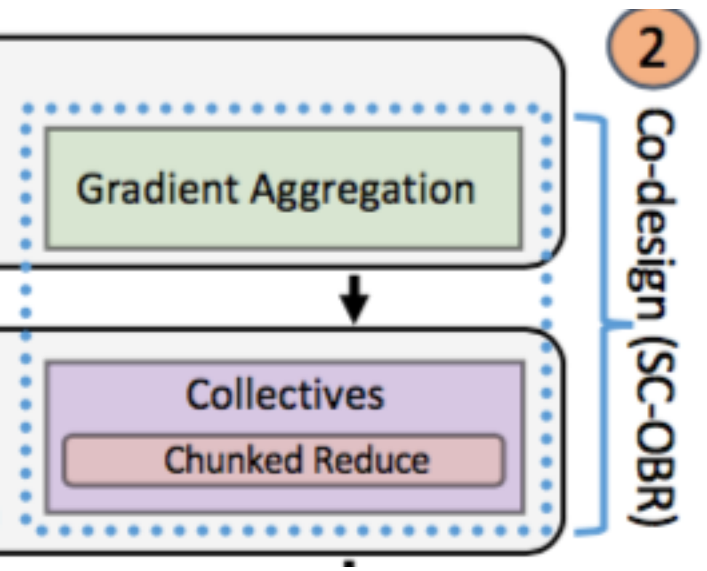


Figure 13. Comparison of SC-B with SC-OB

- SC-OB co-design provides an excellent overlap and hides the latency.
- SC-OB gives 15% improvement over SC-B.

Performance Evaluation



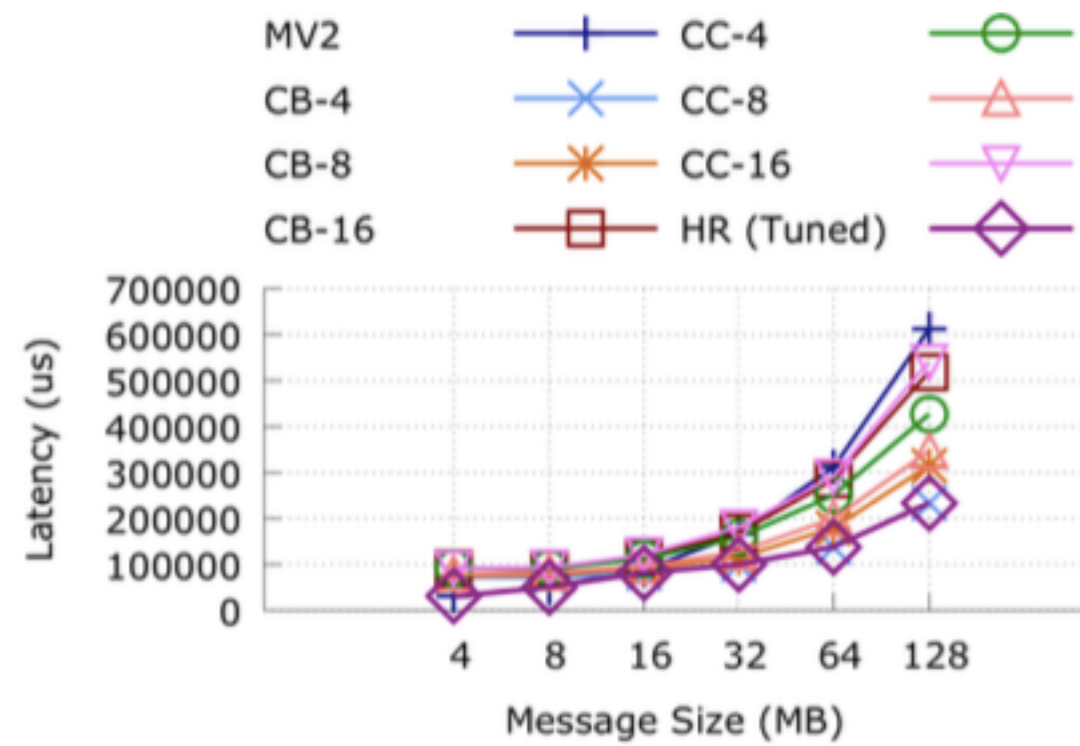
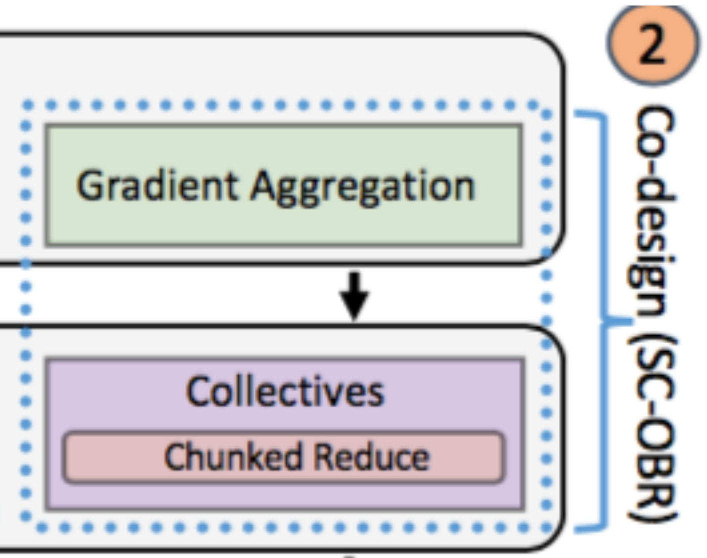


Figure 11. Performance for 160 Processes (GPUs): MVAPICH2, Chain-Binomial, Chain-Chain, and Proposed HR (Tuned) on Cluster-A

- HR (Tuned) is the new tuned design that builds on top of the tuning infrastructure in MVAPICH2.

Performance Evaluation



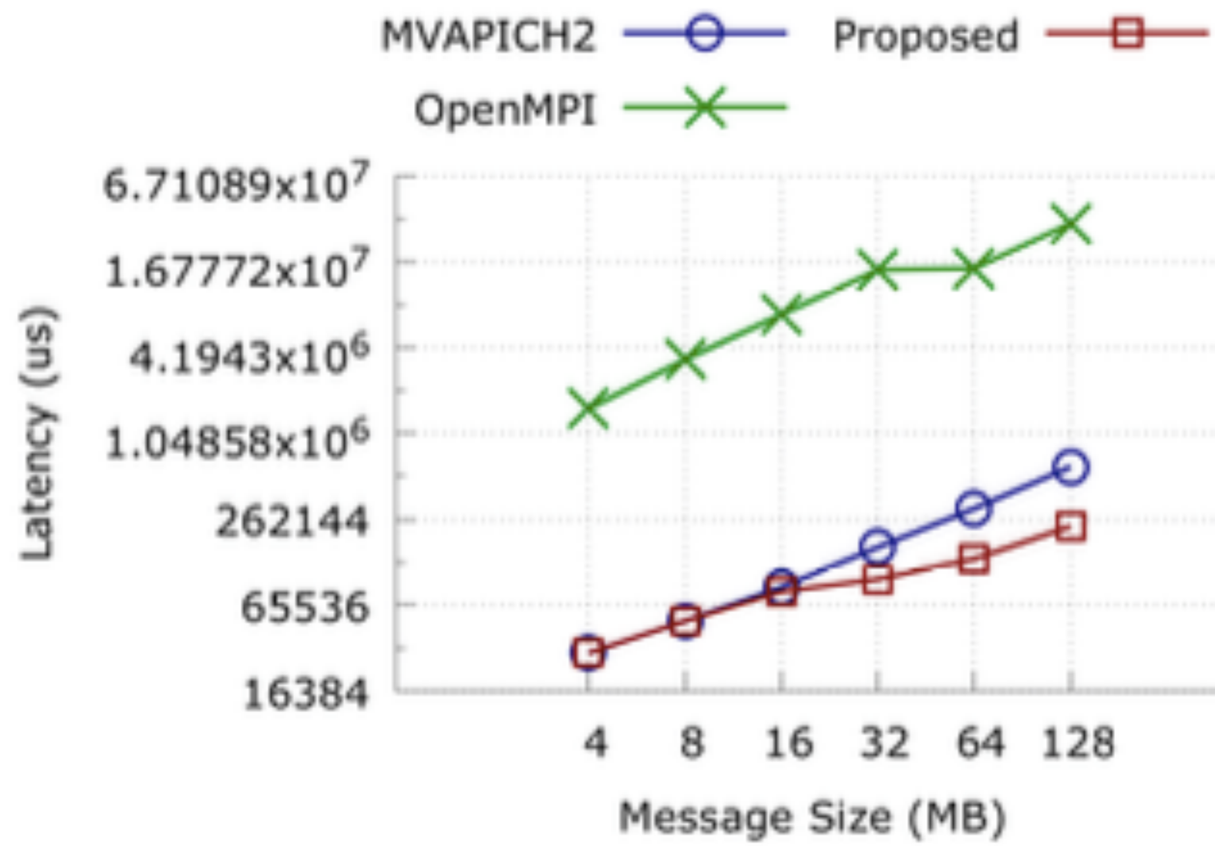
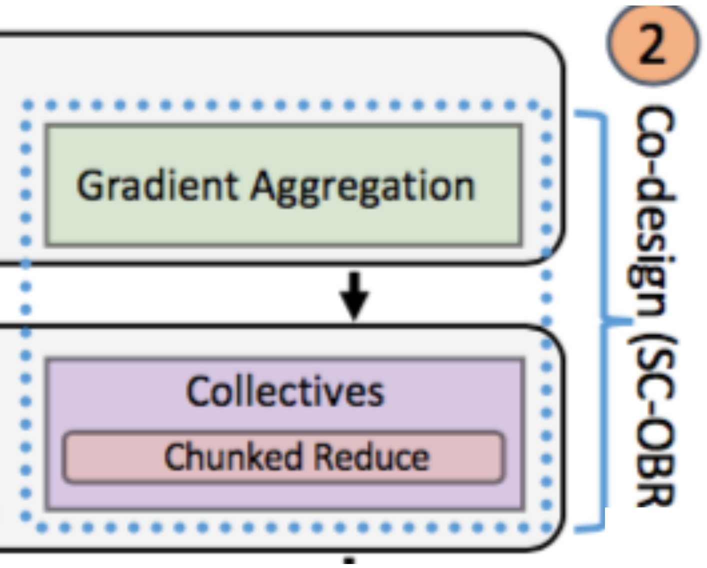


Figure 12. Performance Comparison: MVAPICH2, OpenMPI, and Proposed on Cluster-A

- HR (Tuned) uses OpenMPI v1.10.2 and MVAPICH2 2.2RC1
- Proposed system 3X faster than MVA- PICH2 and up to 133X faster than OpenMPI.

Performance Evaluation



(SC-OBR) and HR

Algorithm / Communicator	SC-B SC-B (+HR)	Aggregation Time	Total Time	Speedup for Aggregation	Overall Speedup
N/A	SC-B	40.6	113.6	1	1
CC-8	SC-B (+HR)	28.6	101.6	1.47	1.11
CB-4	SC-B (+HR)	19.8	92.8	2.04	1.22
CB-8	SC-B (+HR)	17.6	90.6	2.3	1.25

Table 2. Comparison of SC-B vs. SC-B with HR

20% improvement over SC-B for CaffeNet on 8 GPUs and 12% improvement for 16 GPUs.

Conclusion

Conclusion

- Data propagation co-design(SC-OB) give 15% improvement over the basic CUDA-Aware MPI design (SC- B).
- Gradient aggregation co-design(SC-OBR)+Hierarchical Reduce(HR) give 20% improvement for GoogLeNet based training on 160 GPUs.