PARAMETER SERVER (II)

10-06-2018 YIQING MA

OUTLINE

- # First : Basic Knowledge PS (Story, Principle and Detail)
- # Second: Tutorial of PS (how to code Parameter Sever)

#	Third: State-Of-The-Art of Distributed learning frameworks
1	• Bringing HPC Techniques to Deep Learning (Ring all reduce)
2	• Horovod: fast and easy distributed deep learning in Tensorflow
3	• Blink: A fast NVLink-based collective communication library
4	• Parameter Hub: High Performance Parameter Servers for Efficient Distributed Deep Neural Network Training

OUTLINE

1. MPI
2. MapReduce
3. BSP ASP SSP
4. ParameterServer
5. All Reduce
6. Ring AllReduce
7. Rabit & Horvord
8. Poseidon
9. SFB(FactorBroadcasting)

MPI

MESSAGE PASSING INTERFACE

- MPI is a **communication protocol** for programming **parallel computers**. Both point-to-point and collective communication are supported.

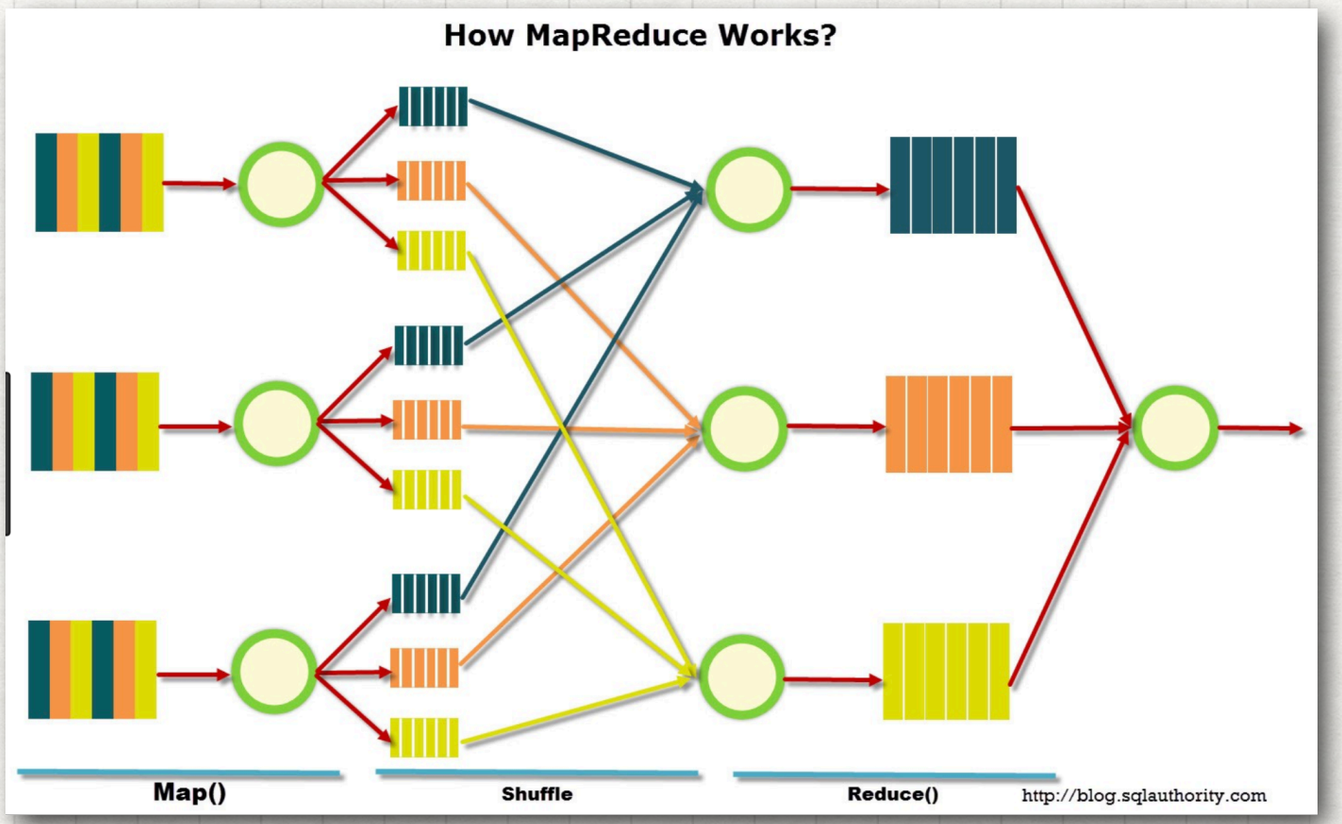
- MPI's goals are high performance, scalability, and portability. MP remains the dominant model used in **high-performance computing** today.

MPI. GRADIENT AGGREGATION（2007）

~~—~~—~~—~~ MAP REDUCE ~~—~~—~~—~~

MapReduce: Simplified Data Processing on Large Clusters
Jeffrey Dean and Sanjay Ghemawat
OSDI 2004

- MapReduce is a **programming model** and an associated implementation for processing and generating **big data** sets with **parallel**distributed algorithm on a **cluster**.

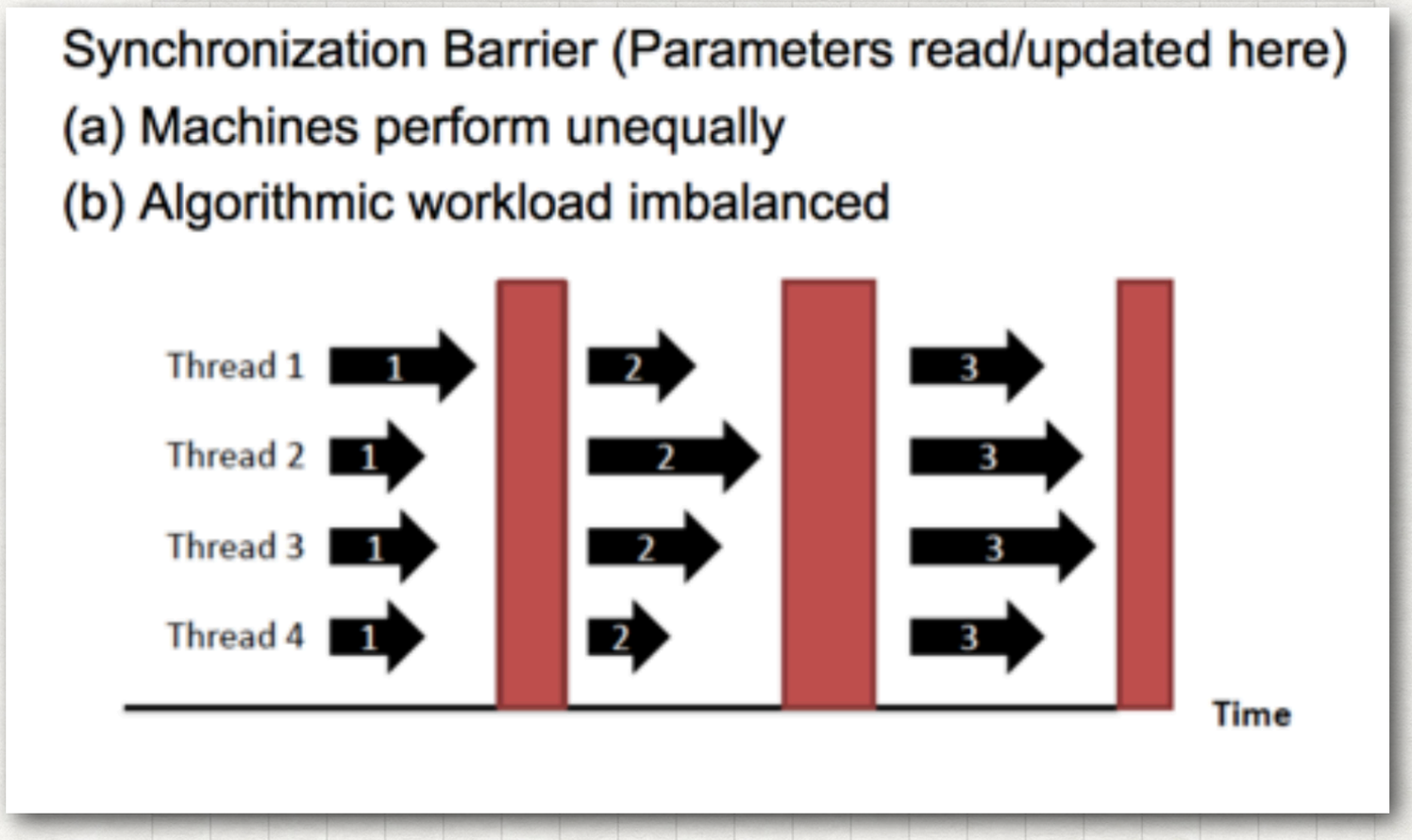


- Google File System + MapReduce => Apache Hadoop

BSP

BULK SYNCHRONOUS PARALLEL

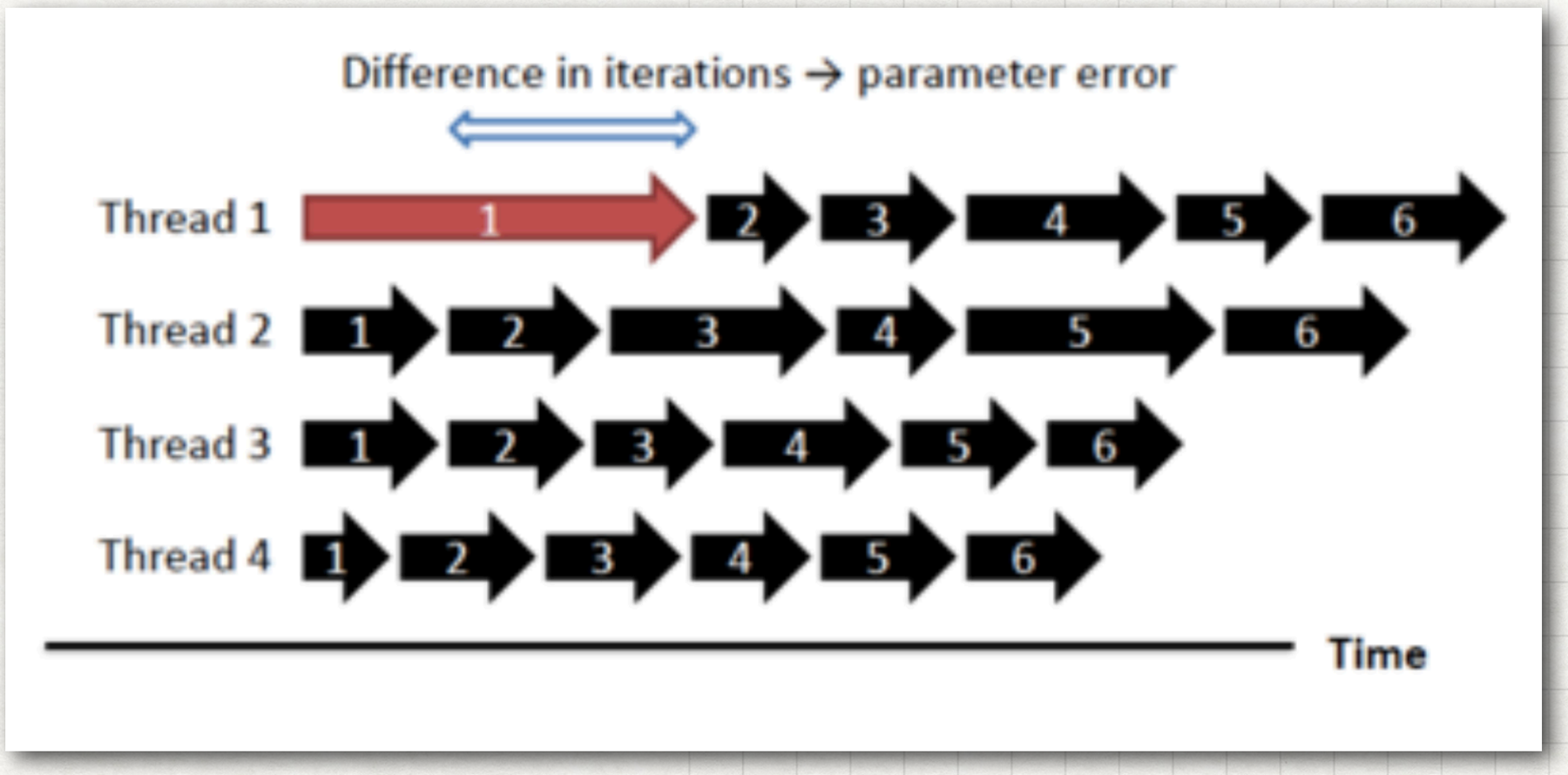
- The bulk synchronous parallel (BSP) **abstract computer** is a **bridging model** for designing **parallel algorithms**.



- The BSP model was developed by **Leslie Valiant** of **Harvard University** during the 1980s.
  The definitive article[1] was published in 1990.

ASP

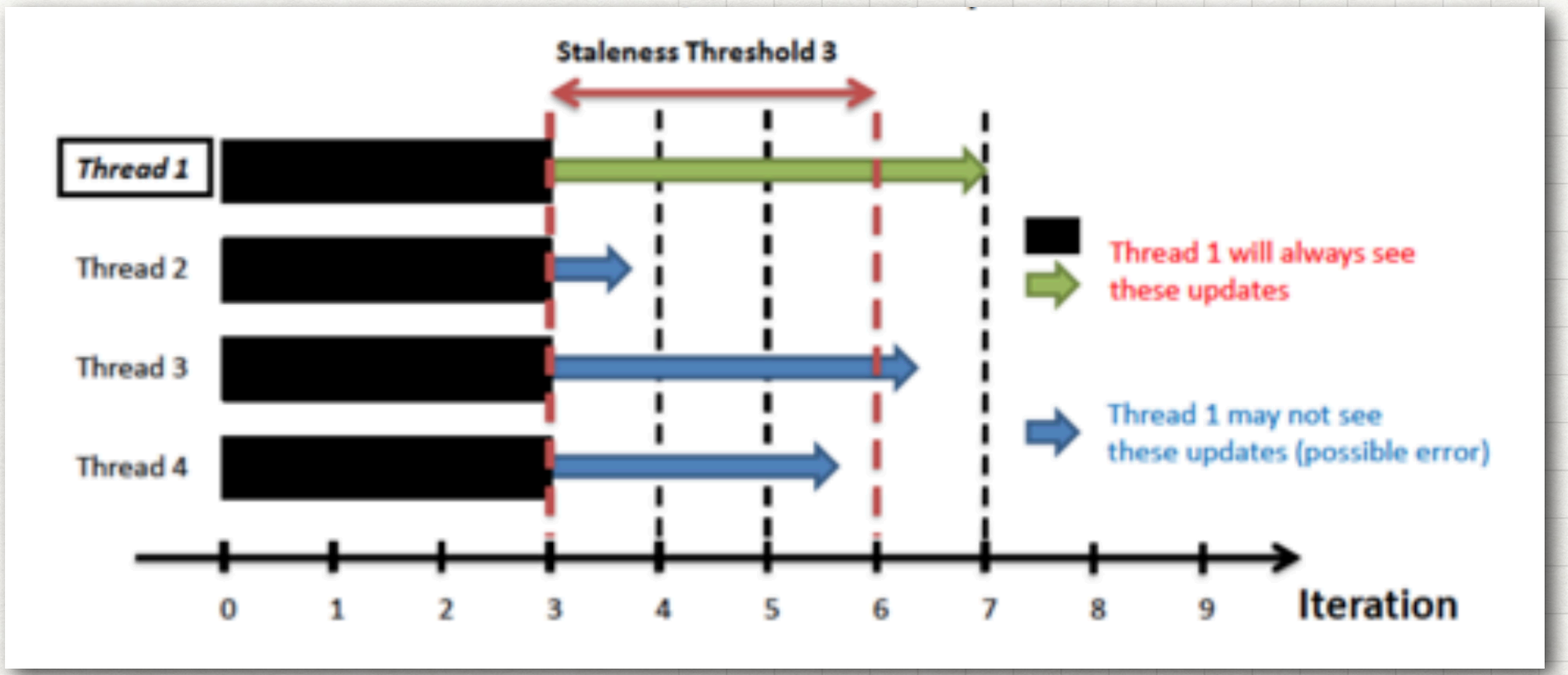
ASYNCHRONOUS



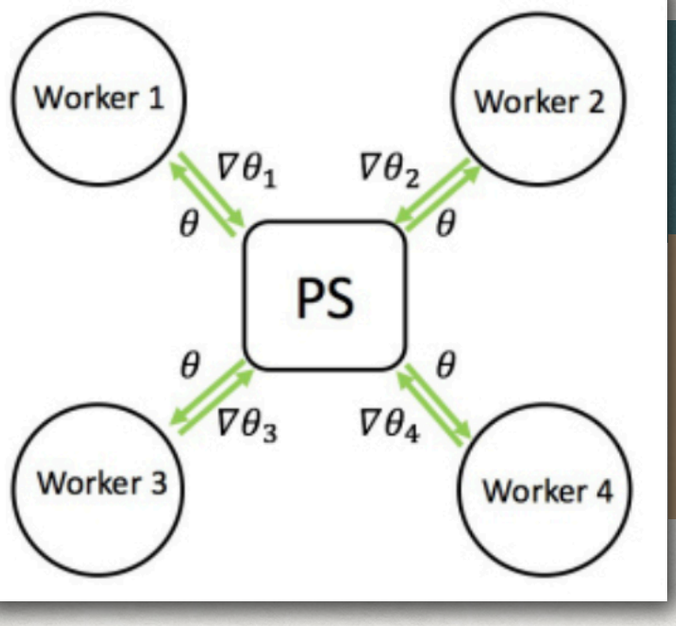
- The ASP model :
  Parameters read/updated at any time.
  Asynchronous is fast but has weak convergence guarantees

SSP

STALE SYNCHRONOUS PARALLEL(SSP)



- Fastest/slowest threads not allowed to drift > S iterations apart
- Protocol : Check cache first ; if too old , update from network.



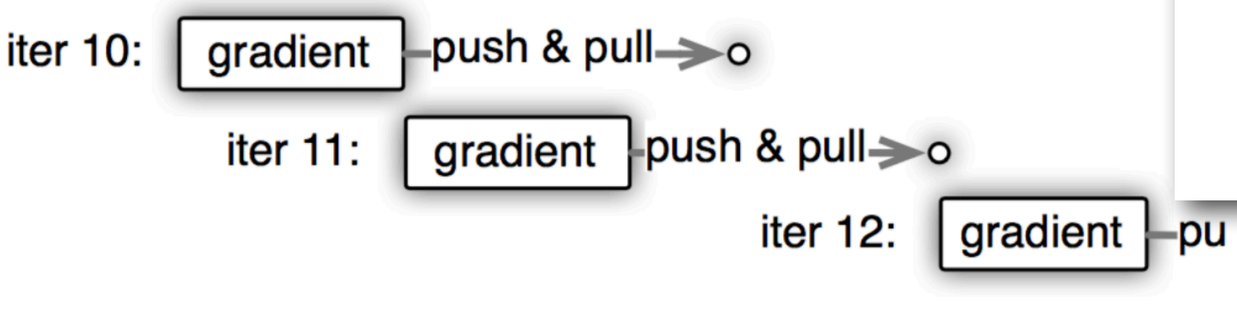
PARAMETER SERVER

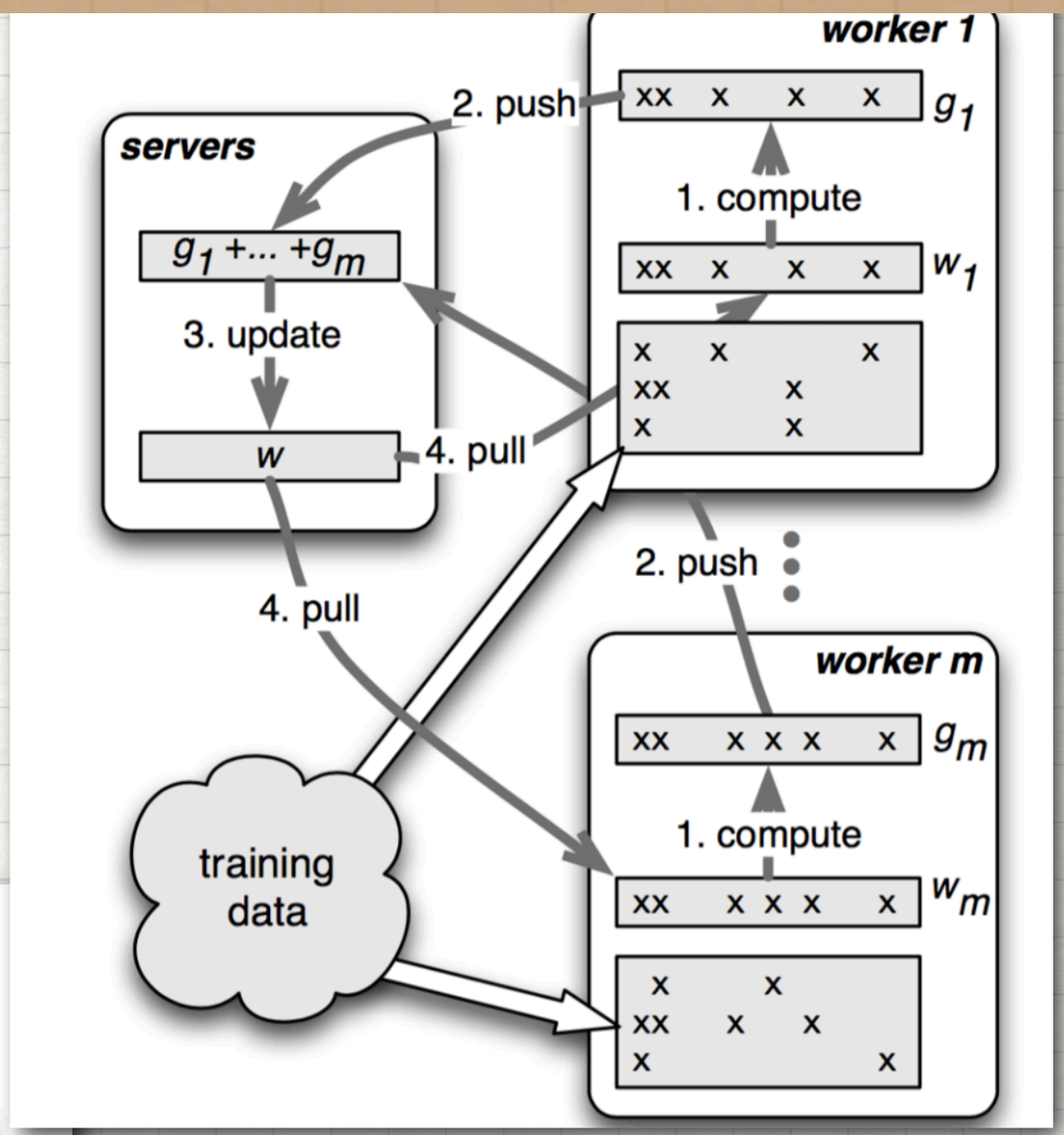
Scaling Distributed Machine Learning with the Parameter Server

Mu Li

OSDI 2014

- Parameter Server :
- 1.Bounded-delay & Asynchronous
- 2.Elastic Scalability
- .....





PARAMETER SERVER

Scaling Distributed Machine Learning with the Parameter Server

Mu Li

OSDI 2014

- **Parameter Server :**
  - 1.Bounded-delay & Asynchronous
  - 2.Elastic Scalability
  - ……

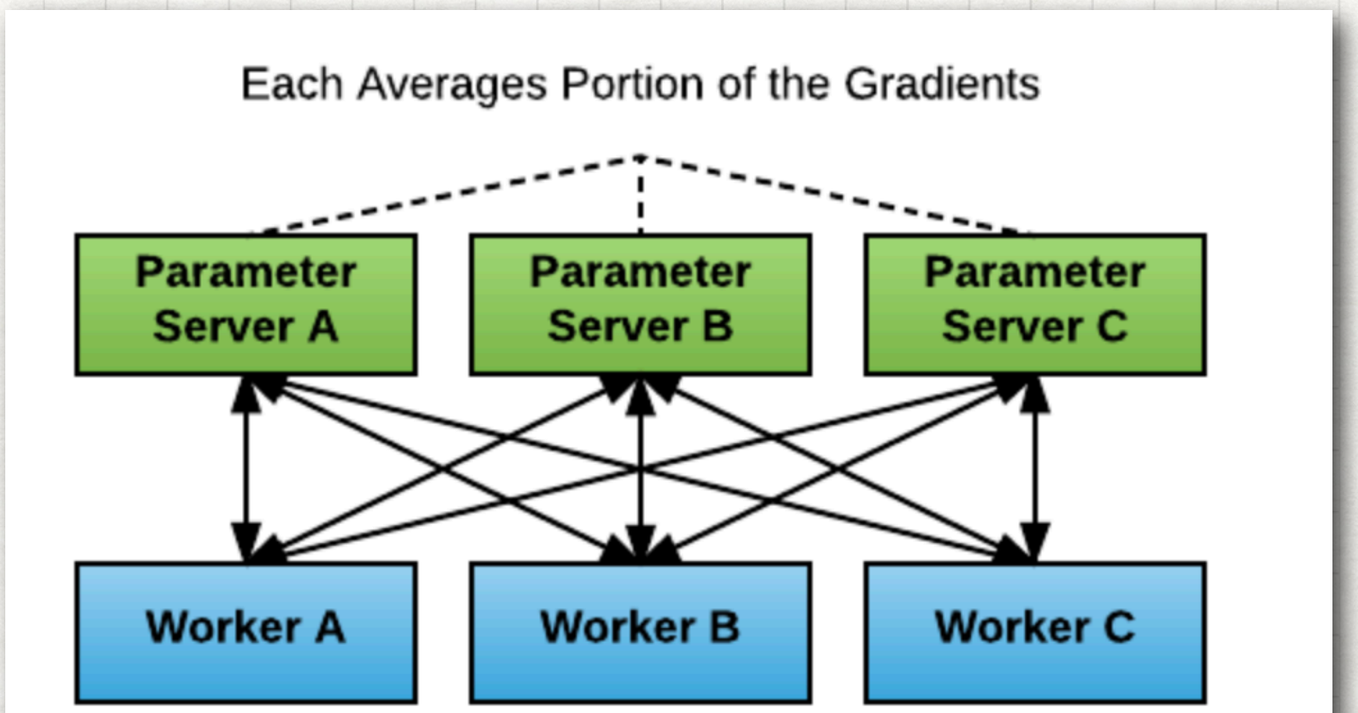
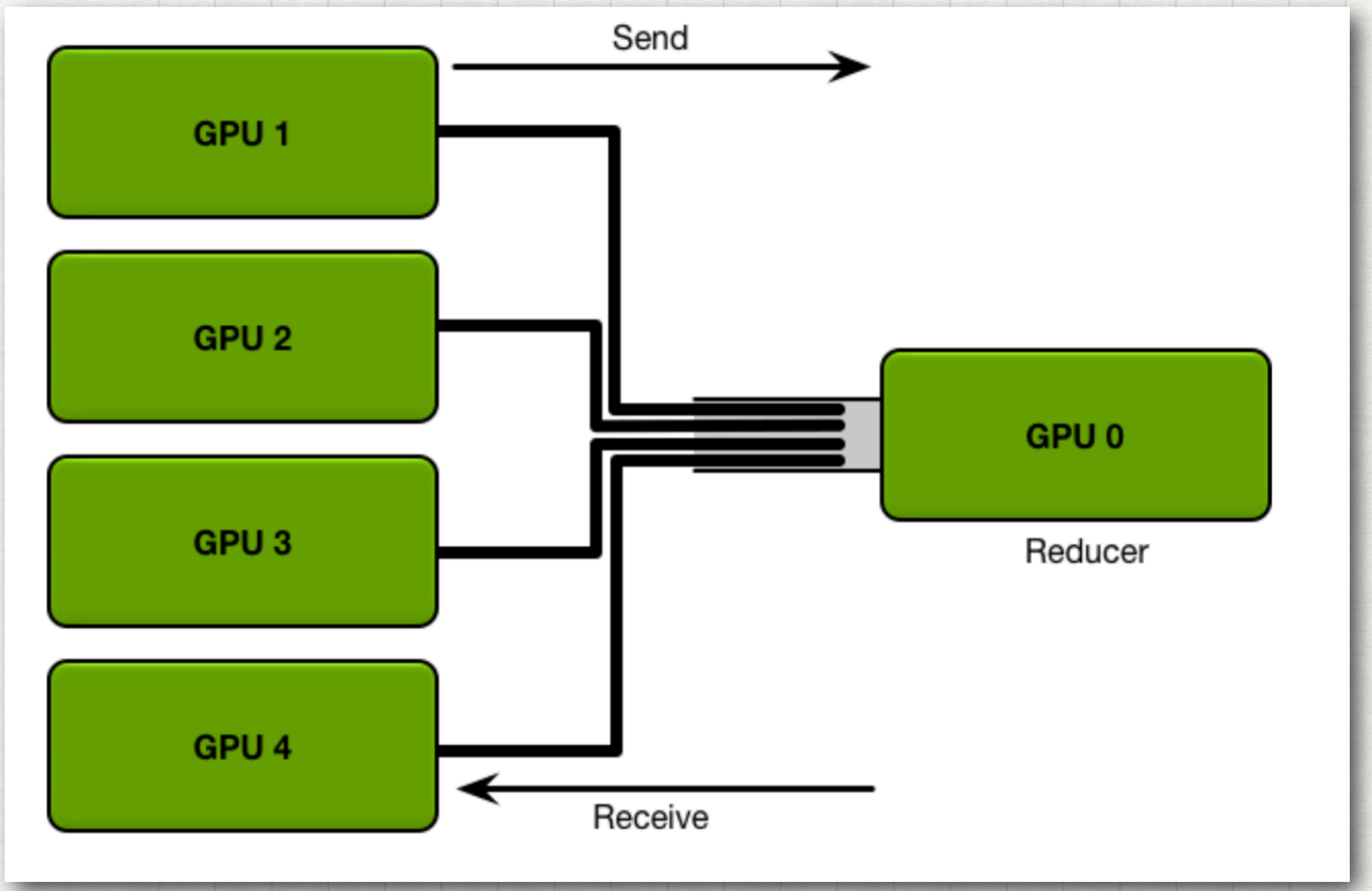


Figure 2: Too many parameter servers may saturate the network

ALLREDUCE

Bringing HPC Techniques to Deep Learning
Baidu Research

- Tree allreduce
- 300M parameters each is 4 Bytes. 1.2G data bandwidth 1GB/s.
- If we use 2 GPUs, delayed 1.2s, 10 GPUs, delayed 10.8s
  .....



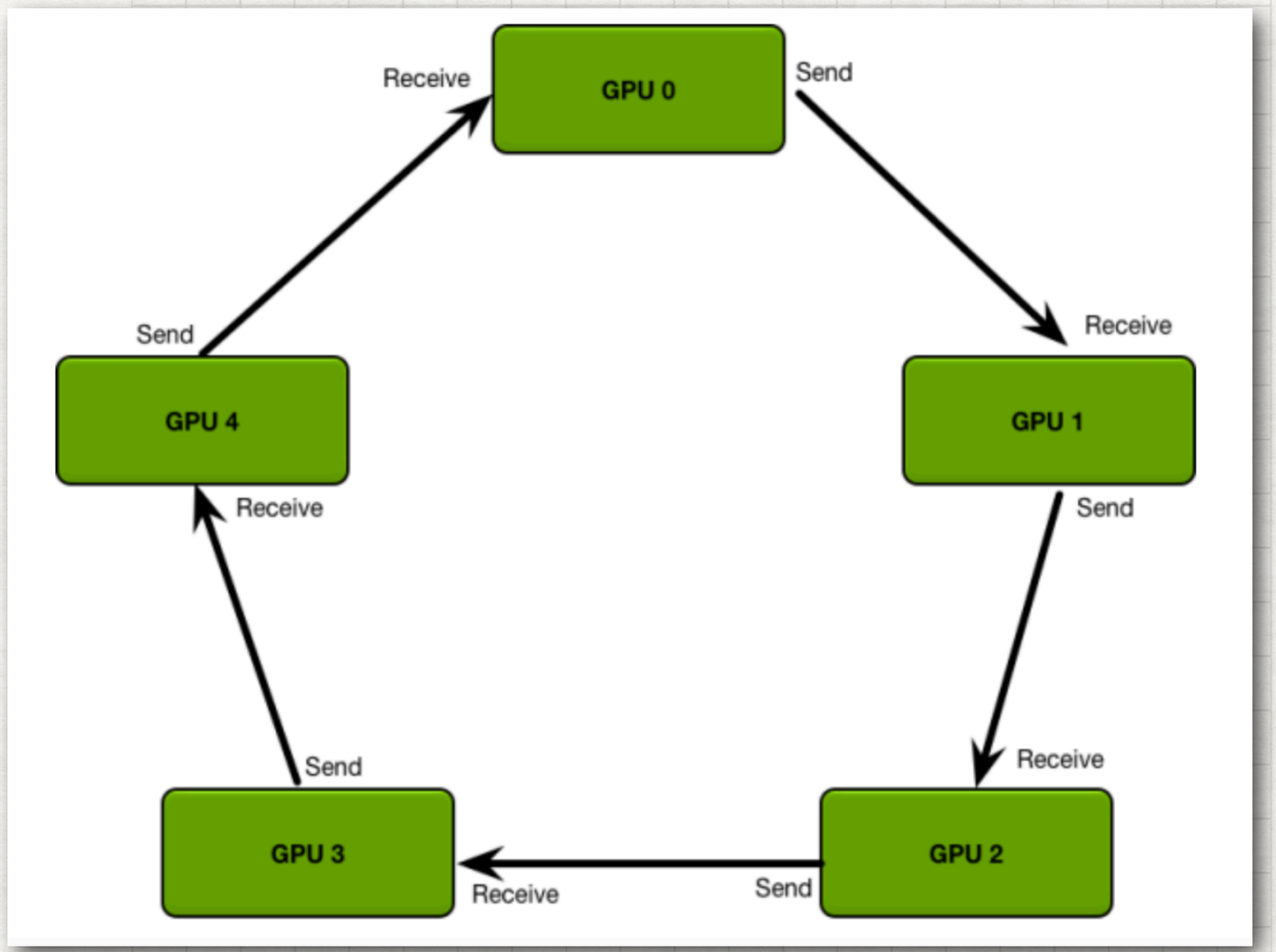
- TWO PROBLEMS

RING ALLREDUCE

Bringing HPC Techniques to Deep Learning
Baidu Research

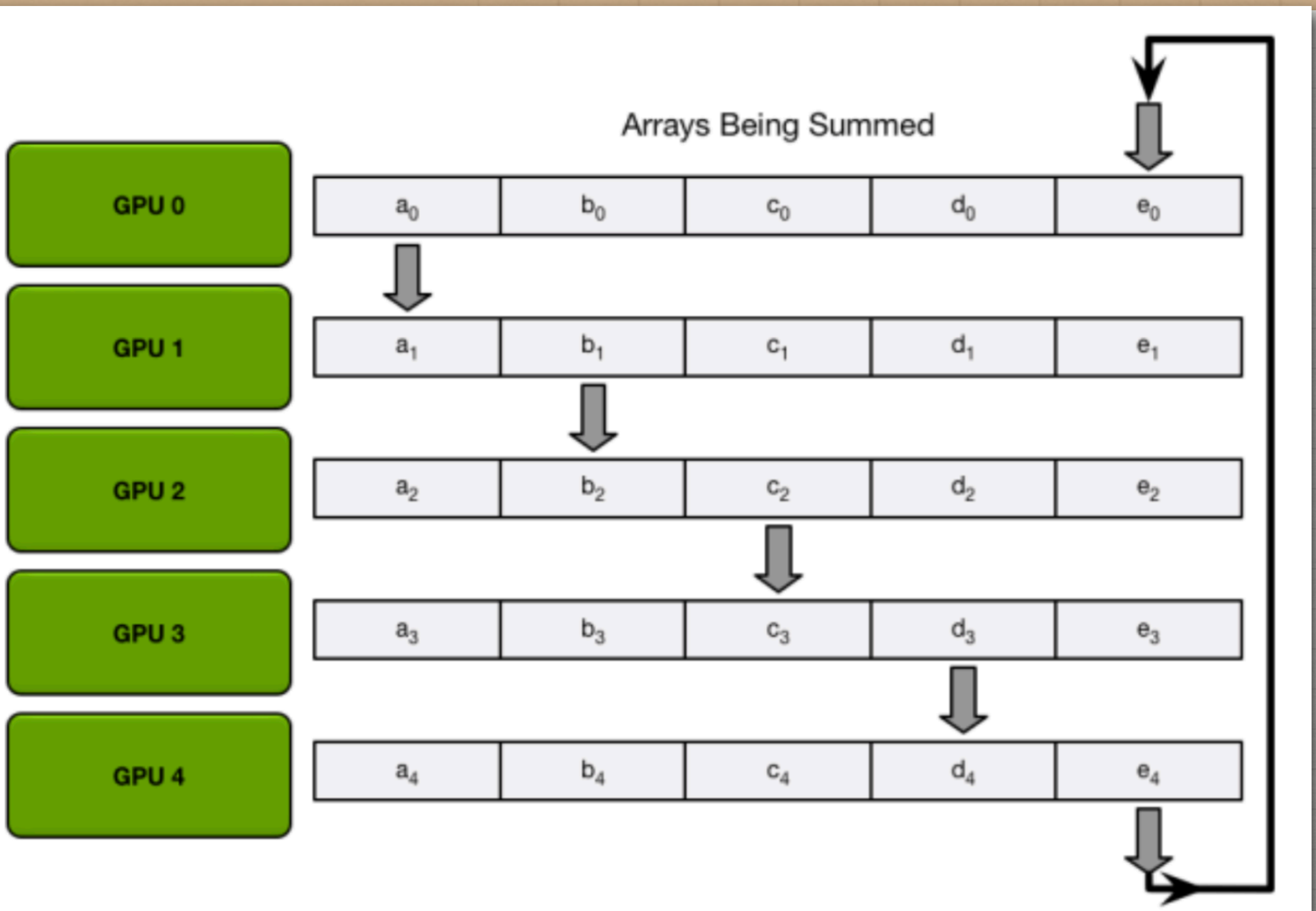
- GPUs have been placed in logical ring.
- GPU get data from left and sends to right.
- GPU #, Send, receive

0	Chunk 0	Chunk 4
1	Chunk 1	Chunk 0
2	Chunk 2	Chunk 1
3	Chunk 3	Chunk 2
4	Chunk 4	Chunk 3



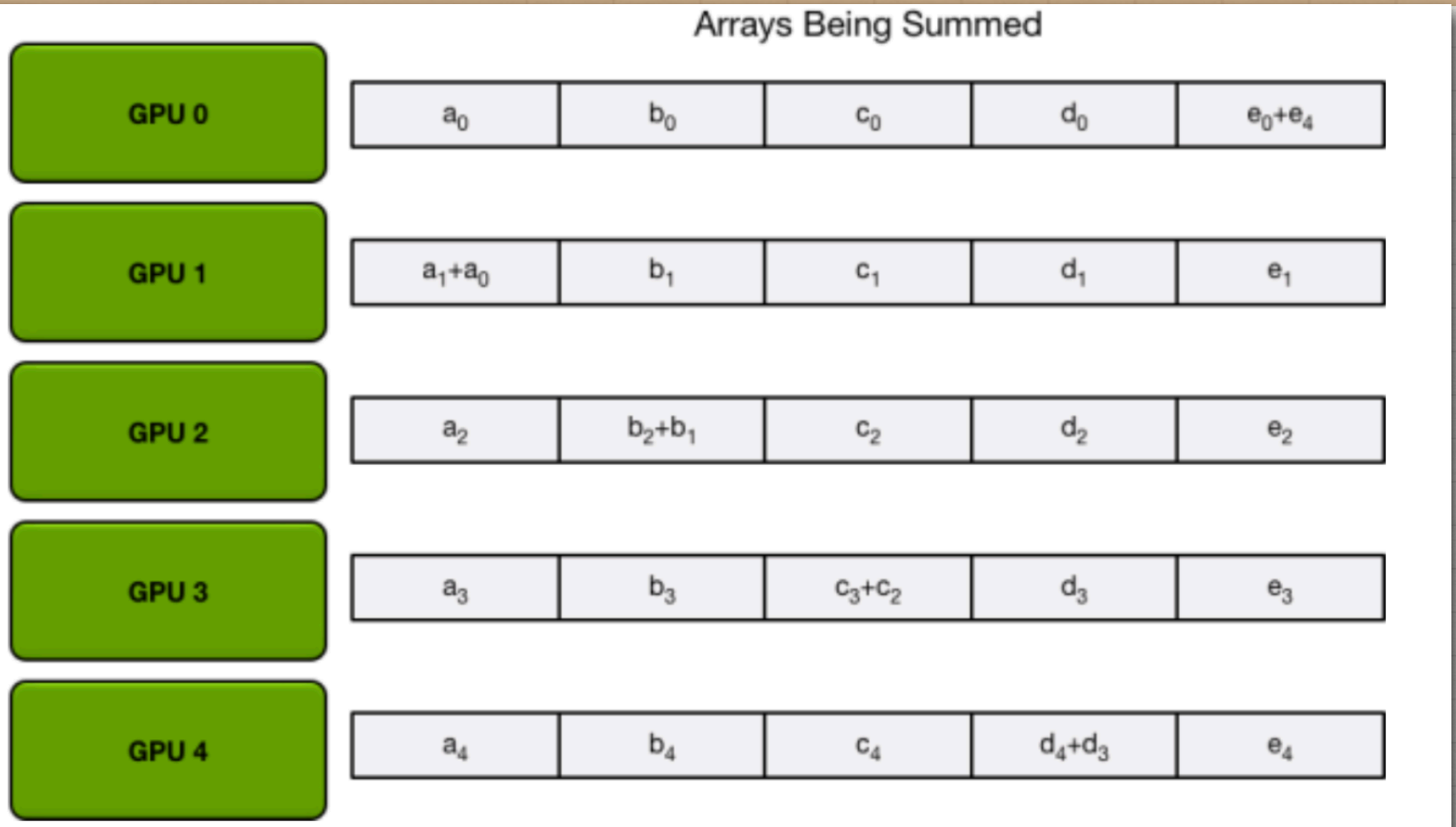
RING ALLREDUCE

Scatter - Reduce



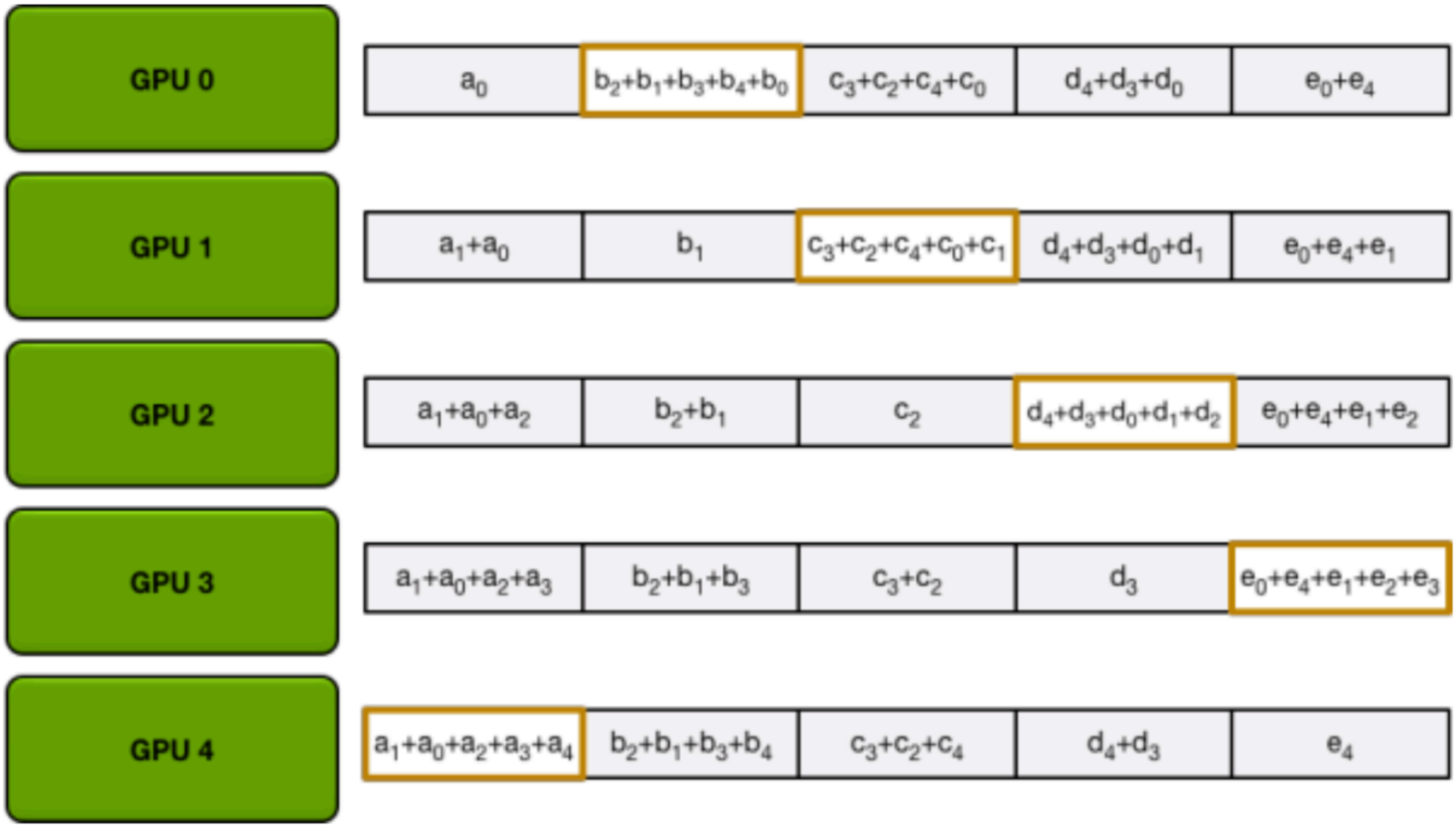
RING ALLREDUCE

Scatter - Reduce



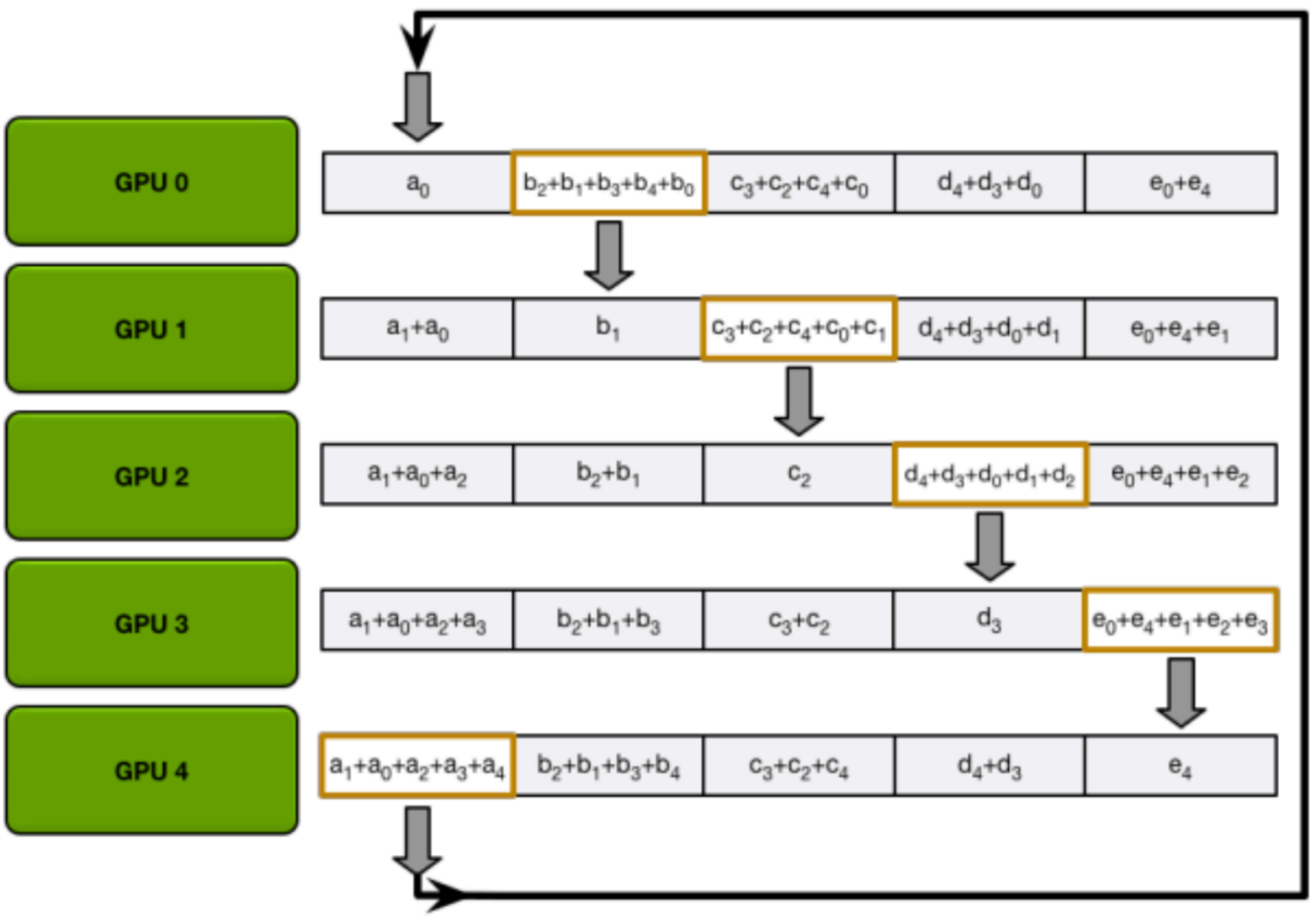
RING ALLREDUCE

Scatter - Reduce



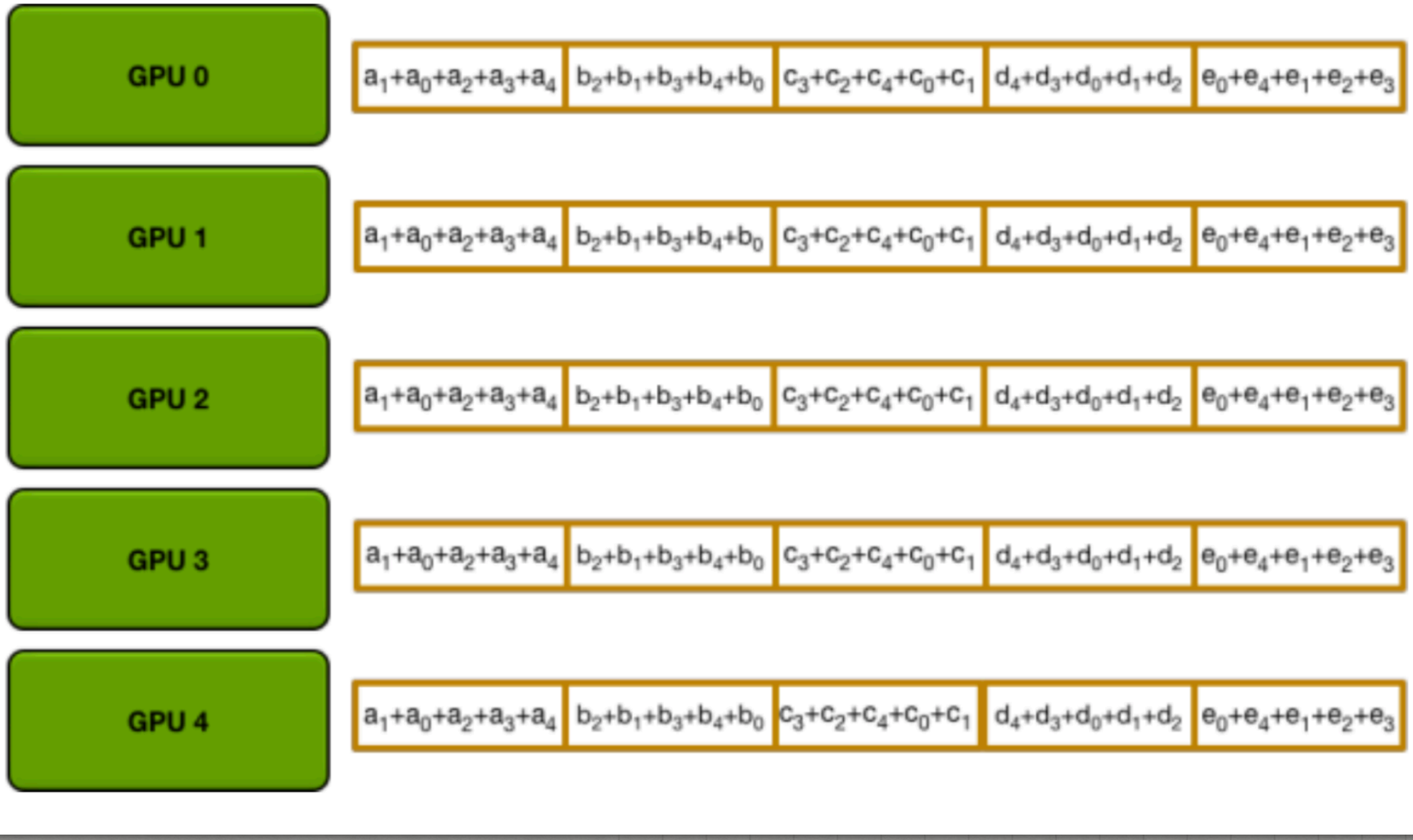
RING ALLREDUCE

ALL Gather



RING ALLREDUCE

ALL Gather



Communication Times = 2 * (N-1)

RABIT

RABIT: A Reliable Allreduce and Broadcast Interface

Tianqi Chen DMLC (2015)

- RABIT1 is an Allreduce library suitable for distributed machine learning algorithms that overcomes the drawbacks;
- it is fault tolerant and can easily run on top of existing systems.

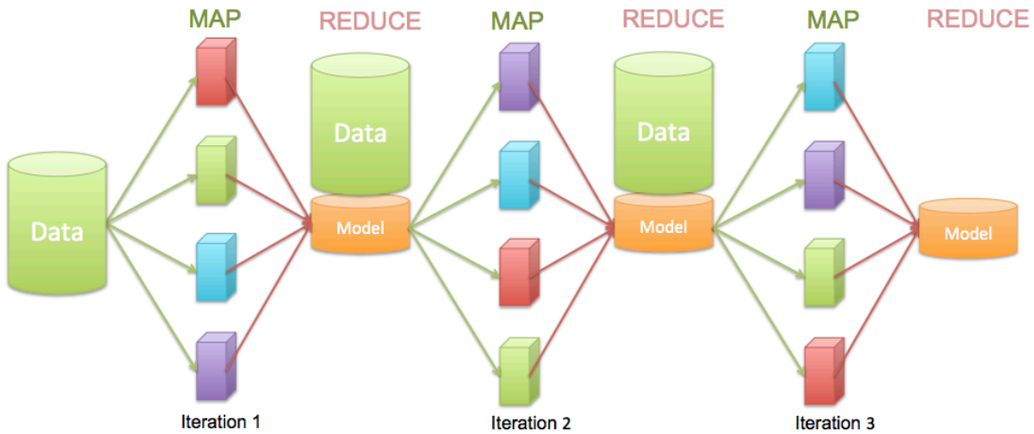


Figure 1: "Iterative" MapReduce

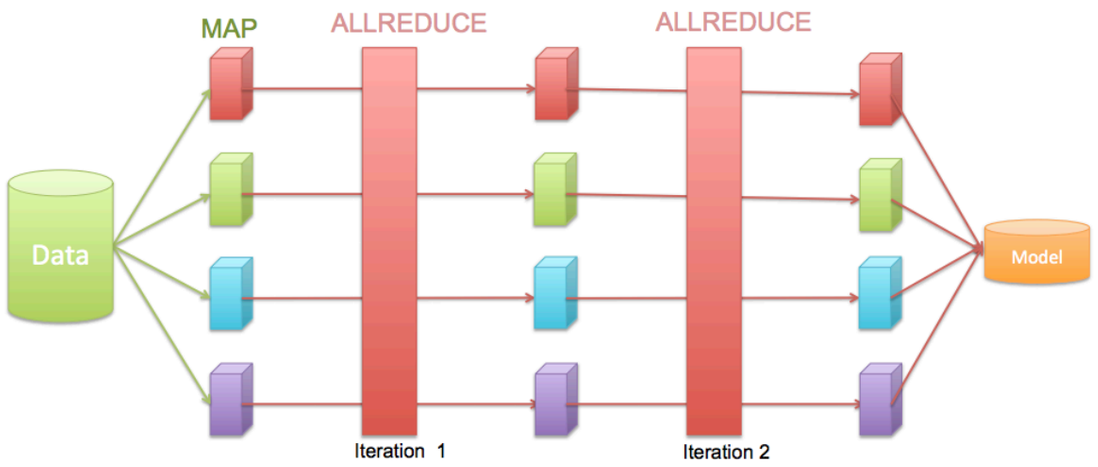


Figure 2: "Iterative" Allreduce

HORVORD

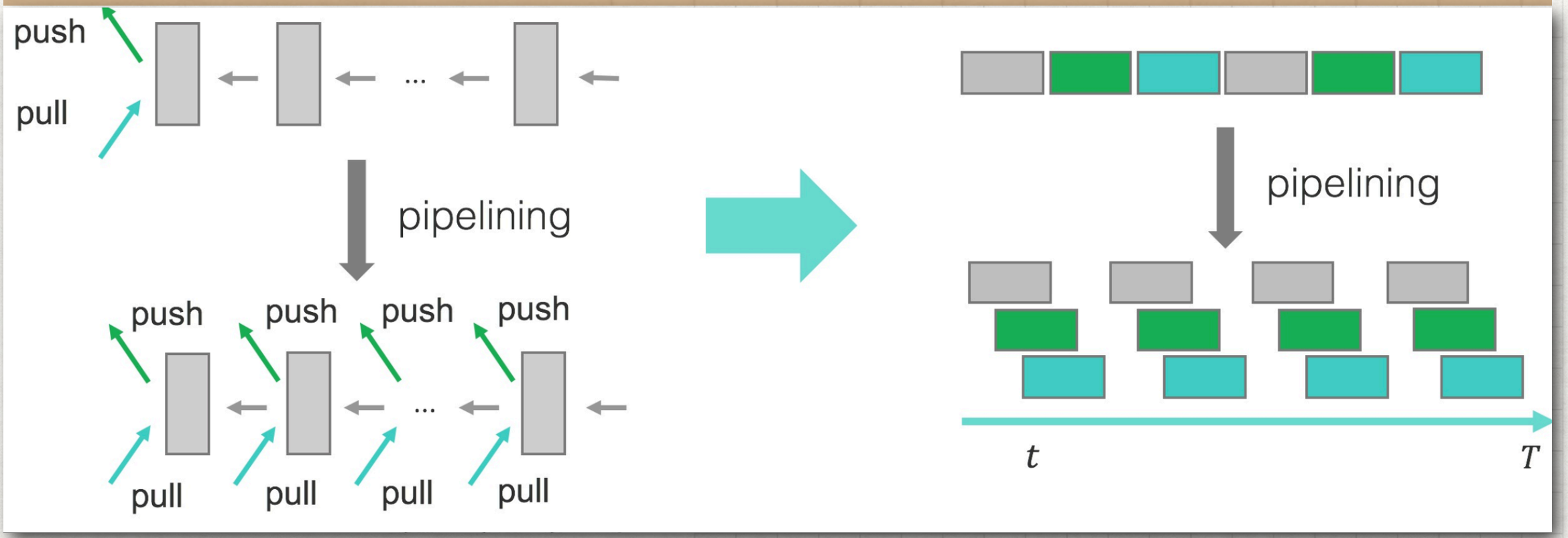
Horovod: fast and easy distributed deep learning in Tensorflow

Alexander Sergeev Uber(2018)

- Horovod is a distributed training framework.for TensorFlow, Keras, and PyTorch.
- Use Baidu Ring-all reduce algorithm.
- Replace it with Nvidia NCCL .NCCL 2 makes it achieves the best performance.

POSEIDON

Poseidon: An Efficient Communication Architecture for Distributed Deep Learning on GPU Clusters.
Hao Zhang, Eric Xing (2015)



- Poseidon:
  1. Build a pipeline to parallelize the computation of parameters and transmitting of parameters [layer by layer]
  2. hybrid communication: It can choose from PS or SFB to achieve the lower communication waist.

SUFFICIENT FACTOR BROADCASTING

Distributed Machine Learning via Sufficient Factor Broadcasting
Eric Xing (2015)

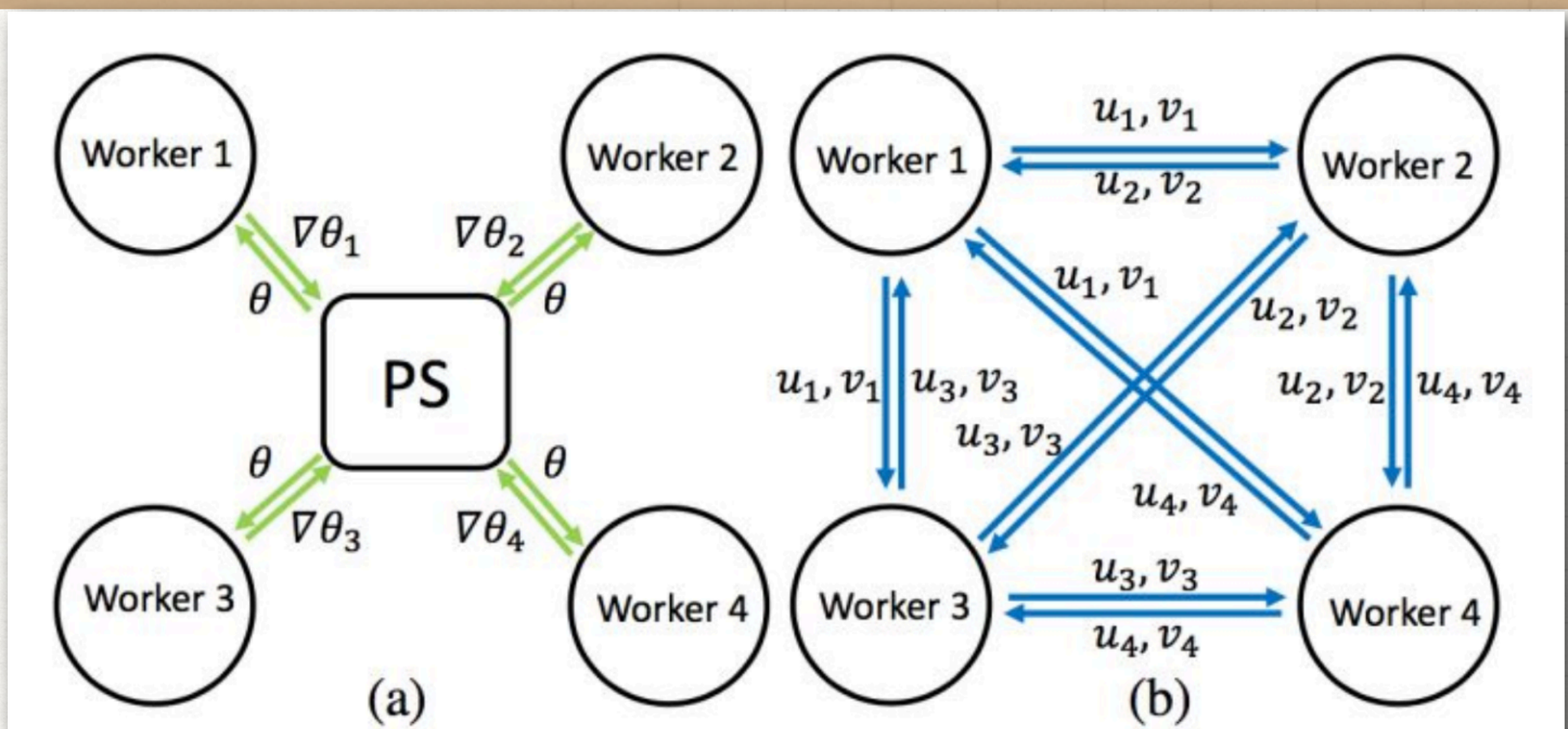


Figure 2: An illustration of (a) the parameter server and (b) sufficient factor broadcasting for distributed ML.

- (1) PS： Master-Server
  (2) SFB： P2P

$$\nabla\theta = \mu\nu^T$$