PAUL G. ALLEN SCHOOL

Parameter Hub

A Rack-Scale Parameter Server for Efficient Cloud-based Distributed Deep Neural Network Training

Liang Luo, *Jacob Nelson,* Luis Ceze, *Amar Phanishayee* and Arvind Krishnamurthy

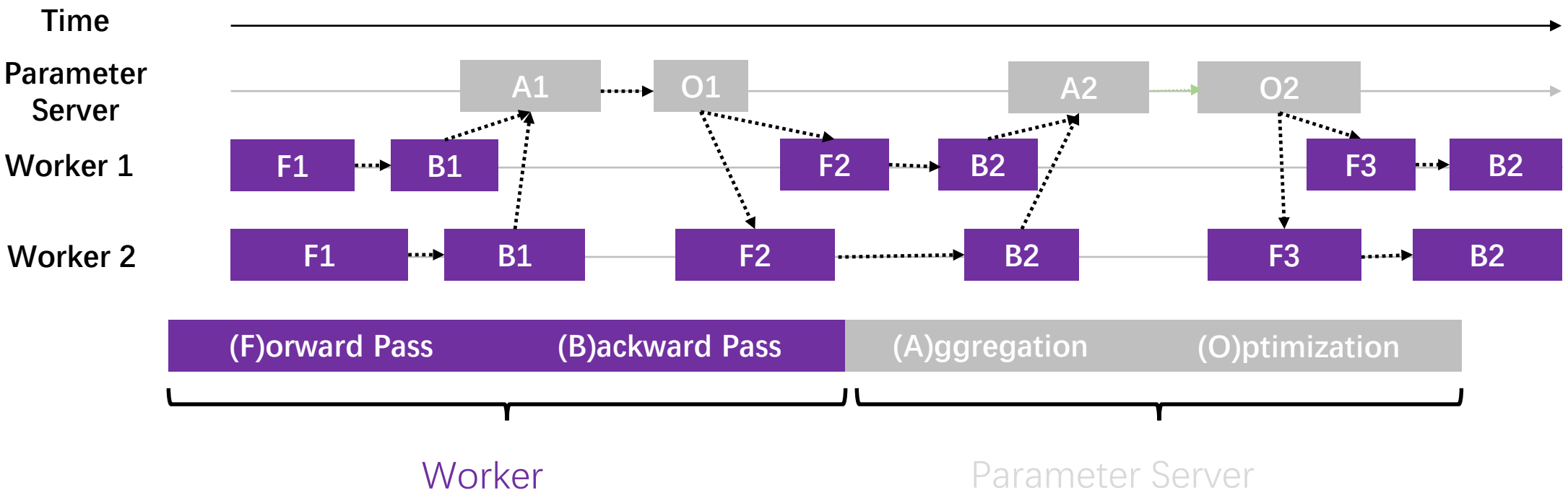
sampl tvm Microsoft

- DNN training is computationally expensive
- Needs to train it in distributed fashion
- People use cloud for DDNN training

Major cloud providers all have an ecosystem for cloud-based DDNN training.

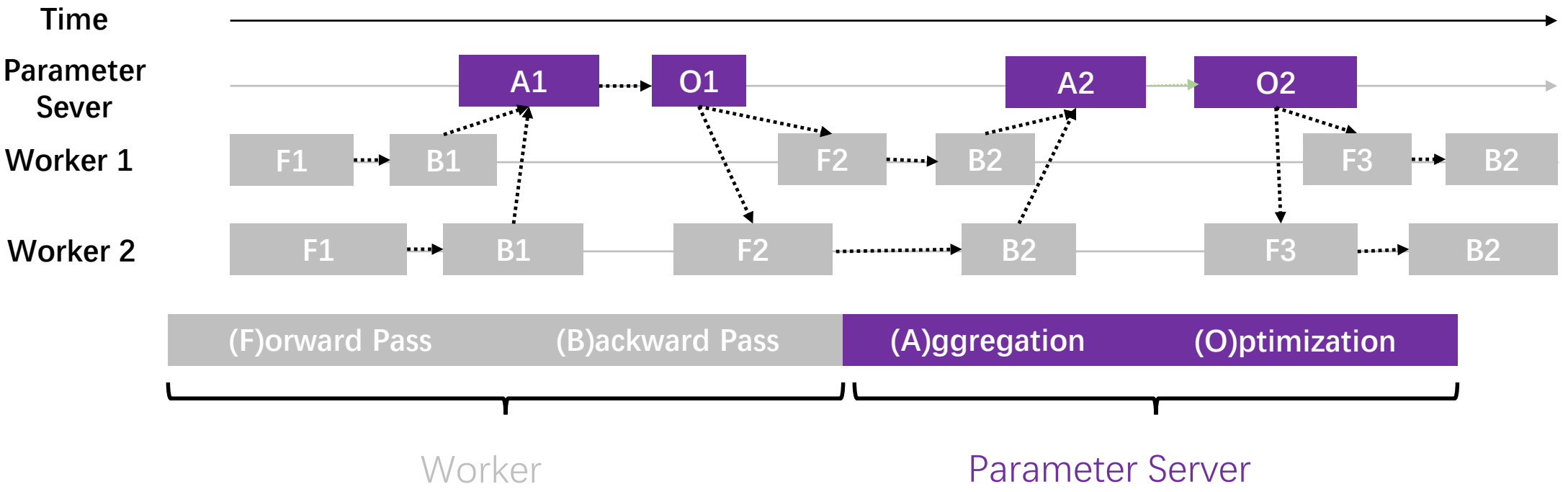
Distributed Training

INDEPENDENT FORWARD/BACKWARD PASSES + COORDINATED PARAMETER EXCHANGE



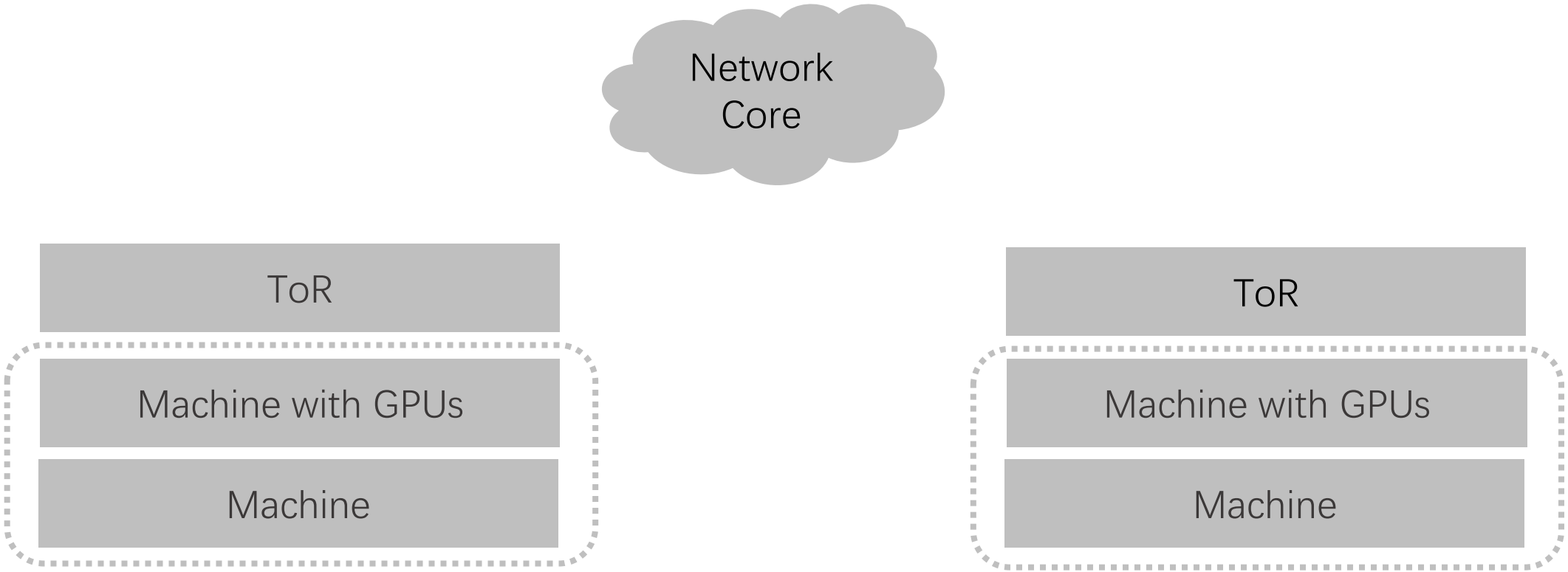
Distributed Training

INDEPENDENT FORWARD/BACKWARD PASSES + COORDINATED PARAMETER EXCHANGE



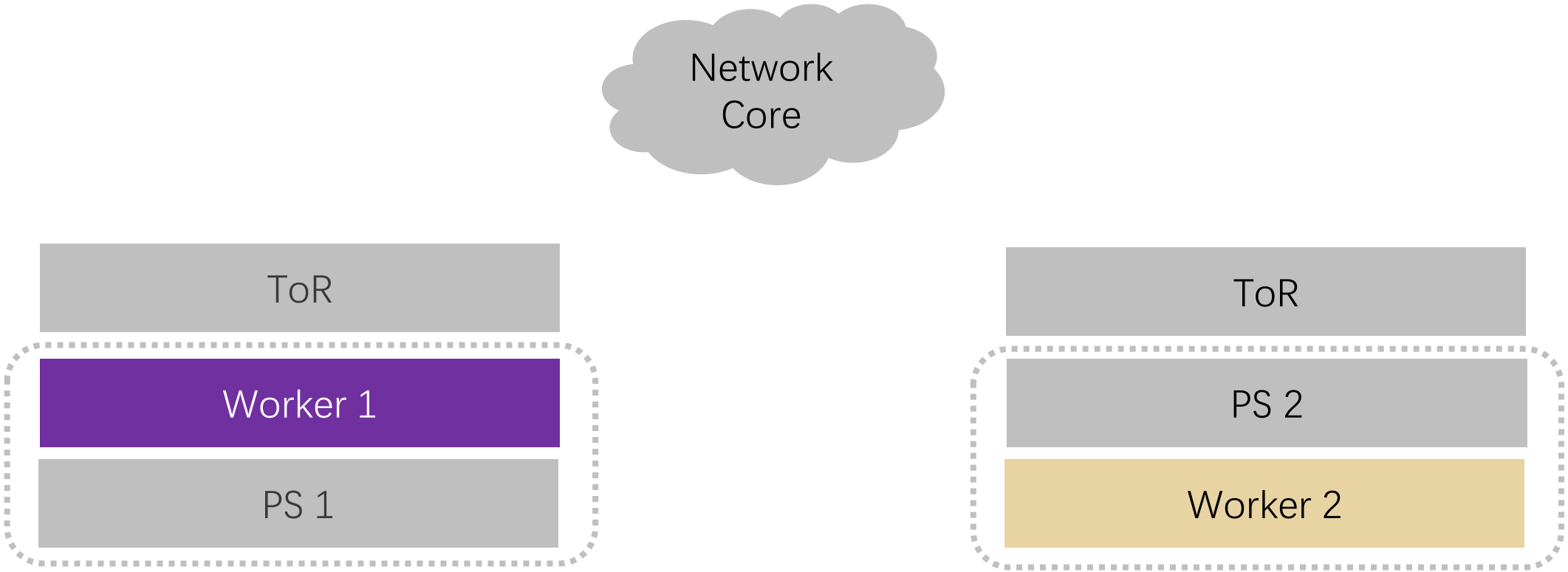
Cloud-based Distributed Training Today

IN THE CONTEXT OF THE CLOUD



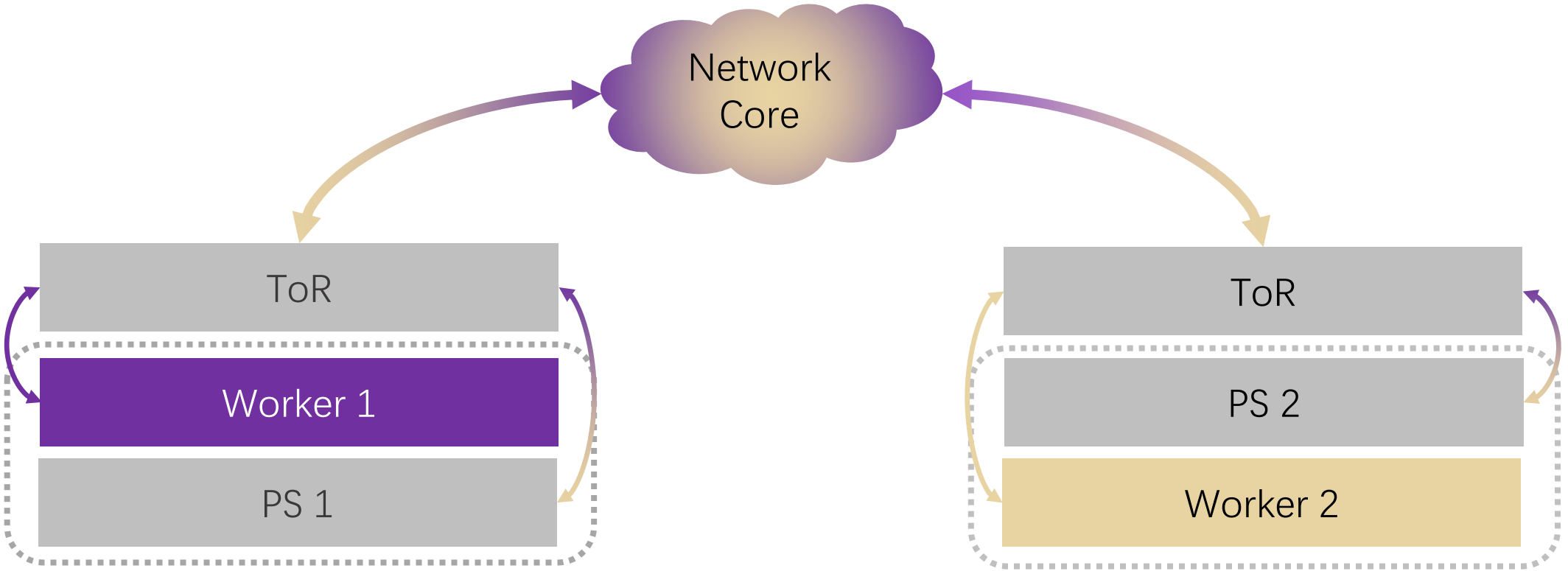
Cloud-based Distributed Training Today

FORWARD AND BACKWARD PASSES IN WORKER



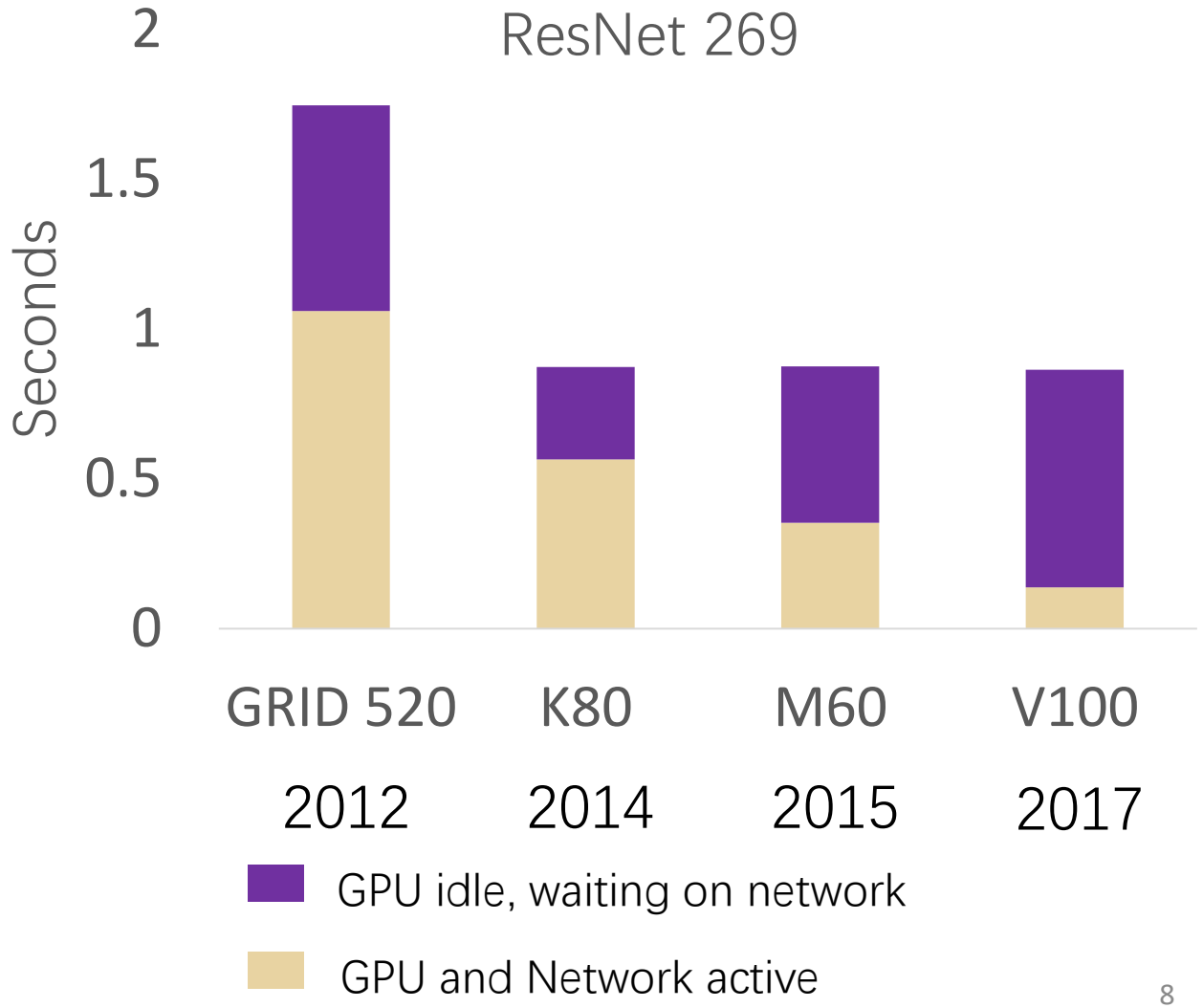
Cloud-based Distributed Training Today

AGGREGATION AND OPTIMIZATION IN PS

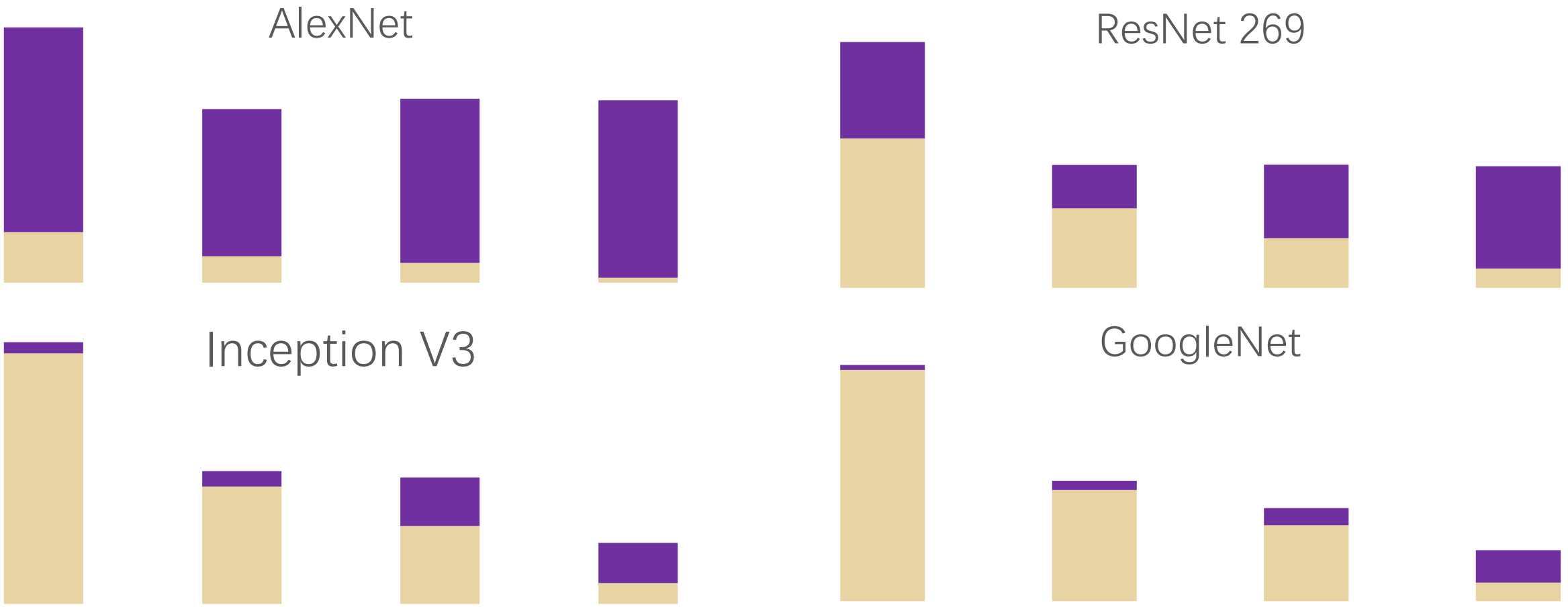


DDNN training is communication bound

- Problem gets worse over time: shifting bottleneck.
- With modern GPUs most of the time is spent on communication.
- Making GPUs faster will do little to increase throughput
- Wasting compute resources.

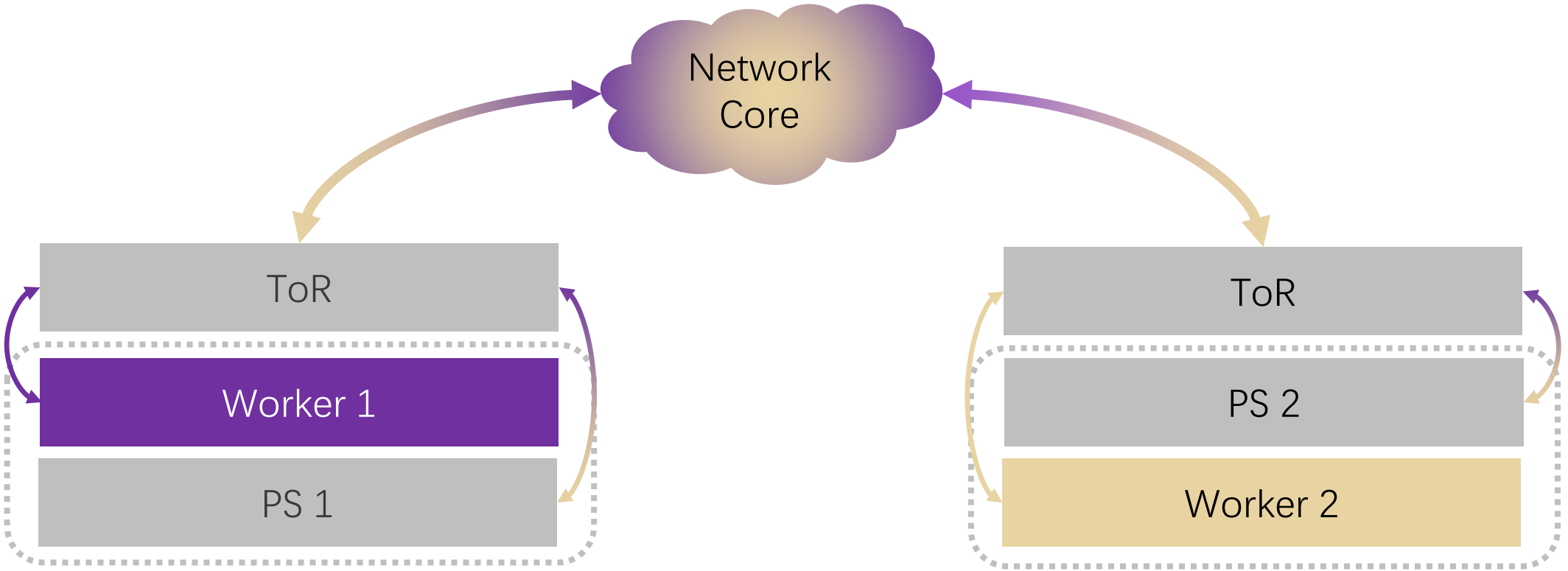


DDNN training is communication bound



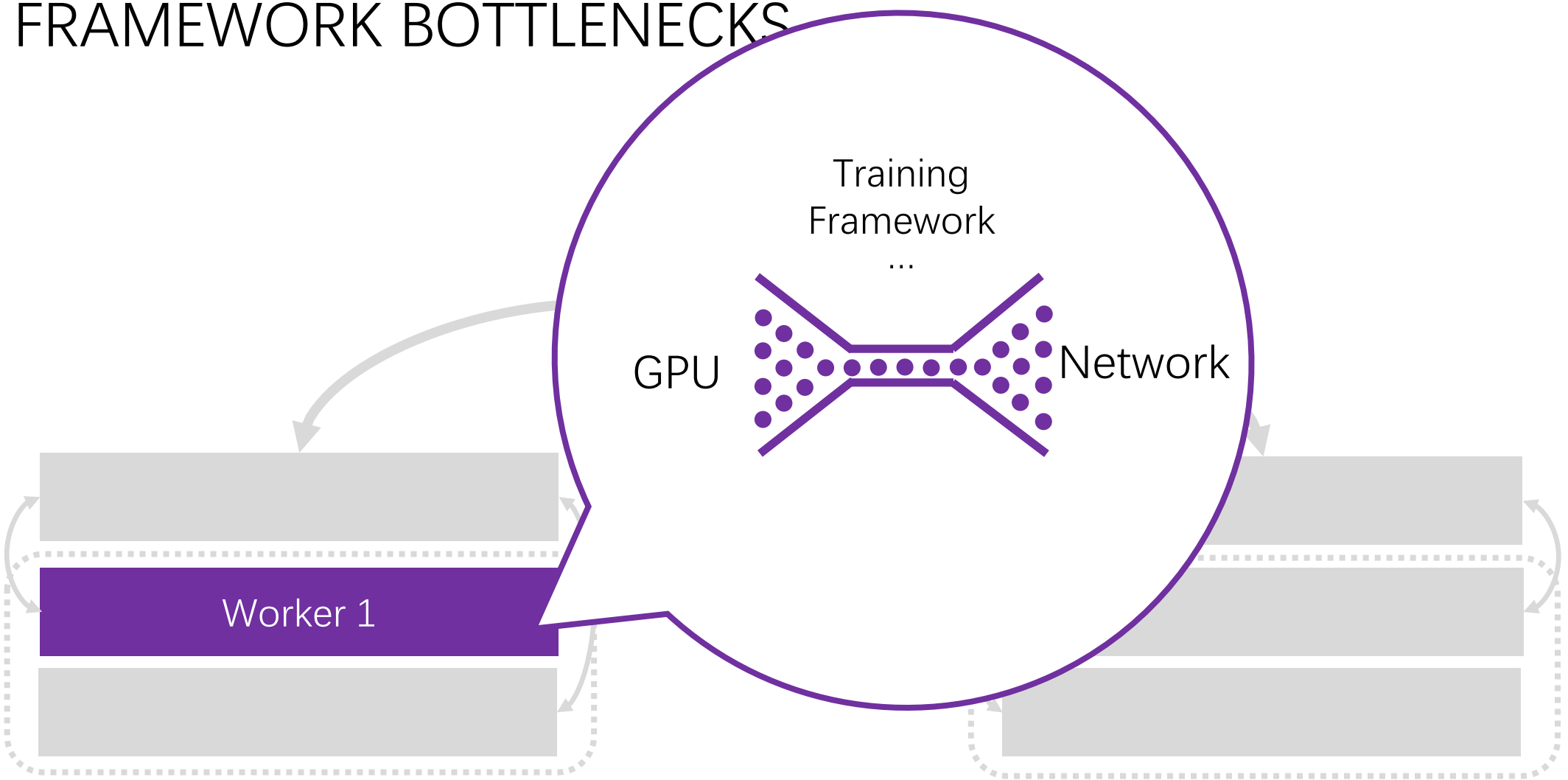
Bottlenecks in Cloud-based DDNN training

MAPPING OF TRAINING WORKLOAD TO THE CLOUD IS INEFFICIENT.



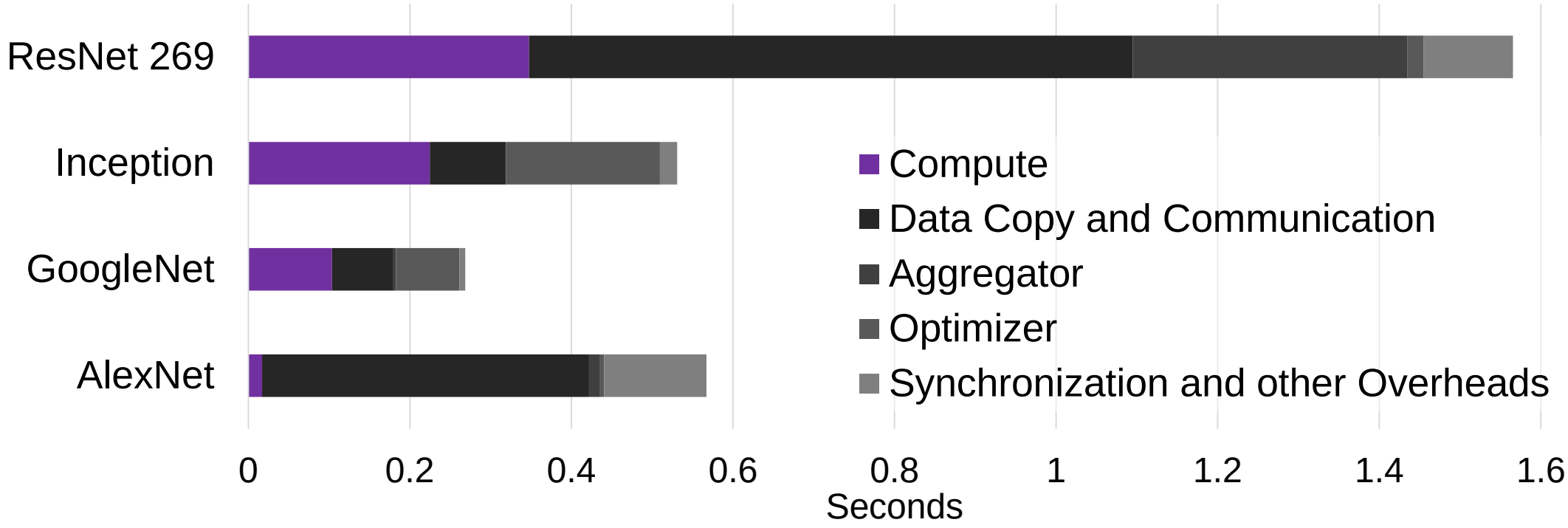
Bottlenecks in Cloud-based DDNN training

FRAMEWORK BOTTLENECKS



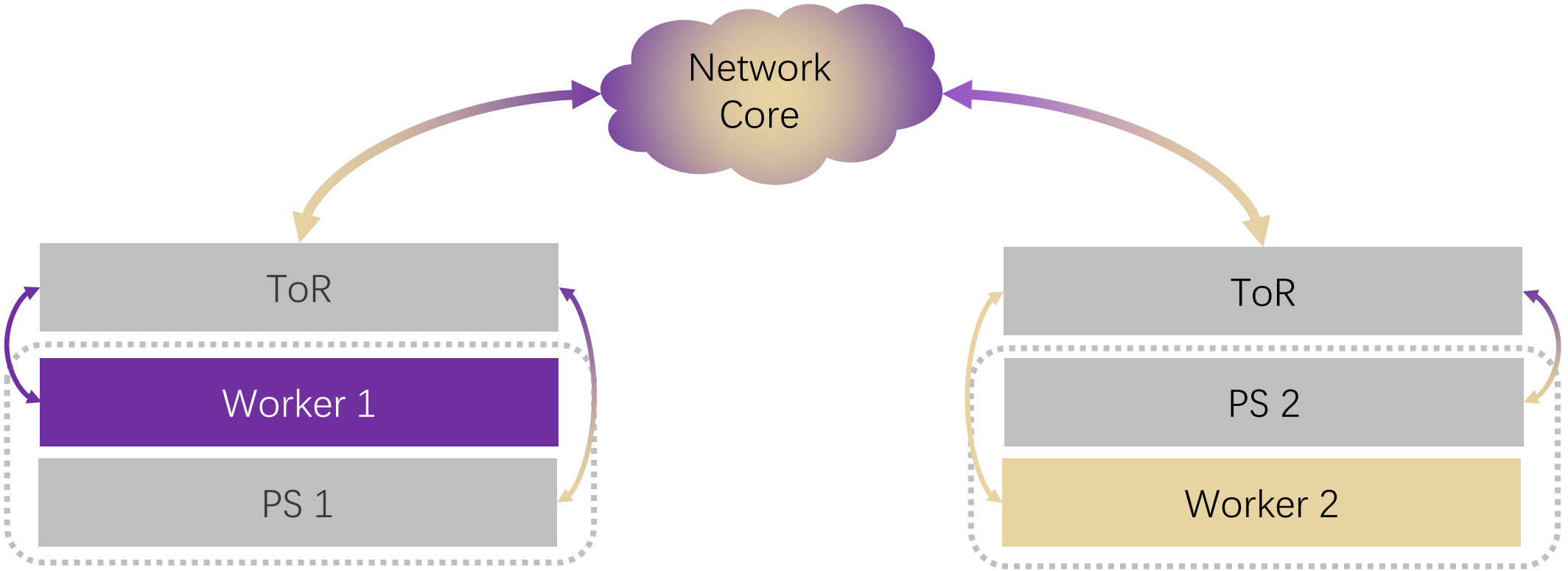
Bottlenecks in Cloud-based DDNN training

FRAMEWORK BOTTLENECKS



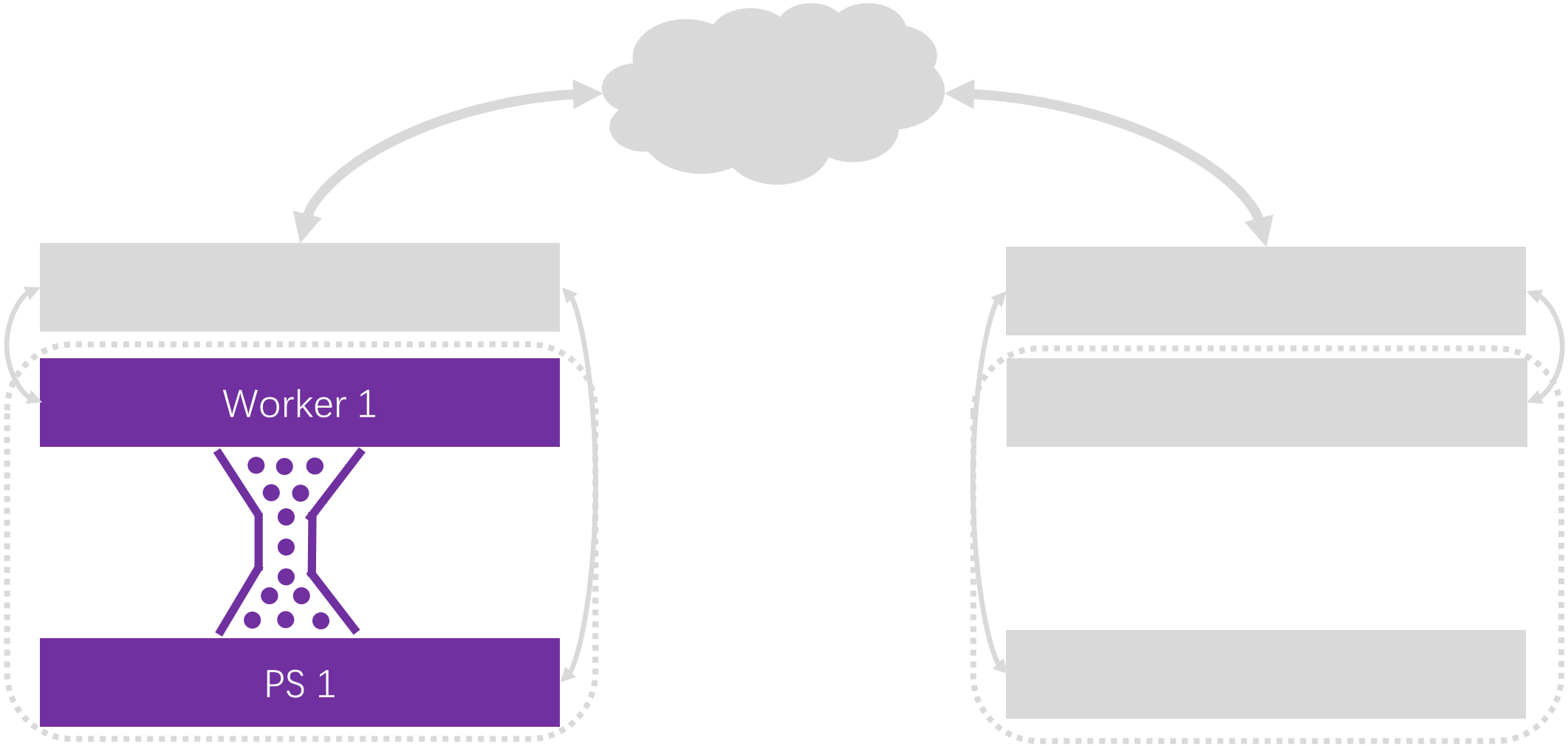
Bottlenecks in Cloud-based DDNN training

MAPPING OF TRAINING WORKLOAD TO THE CLOUD IS INEFFICIENT.



Bottlenecks in Cloud-based DDNN training

BANDWIDTH BOTTLENECK

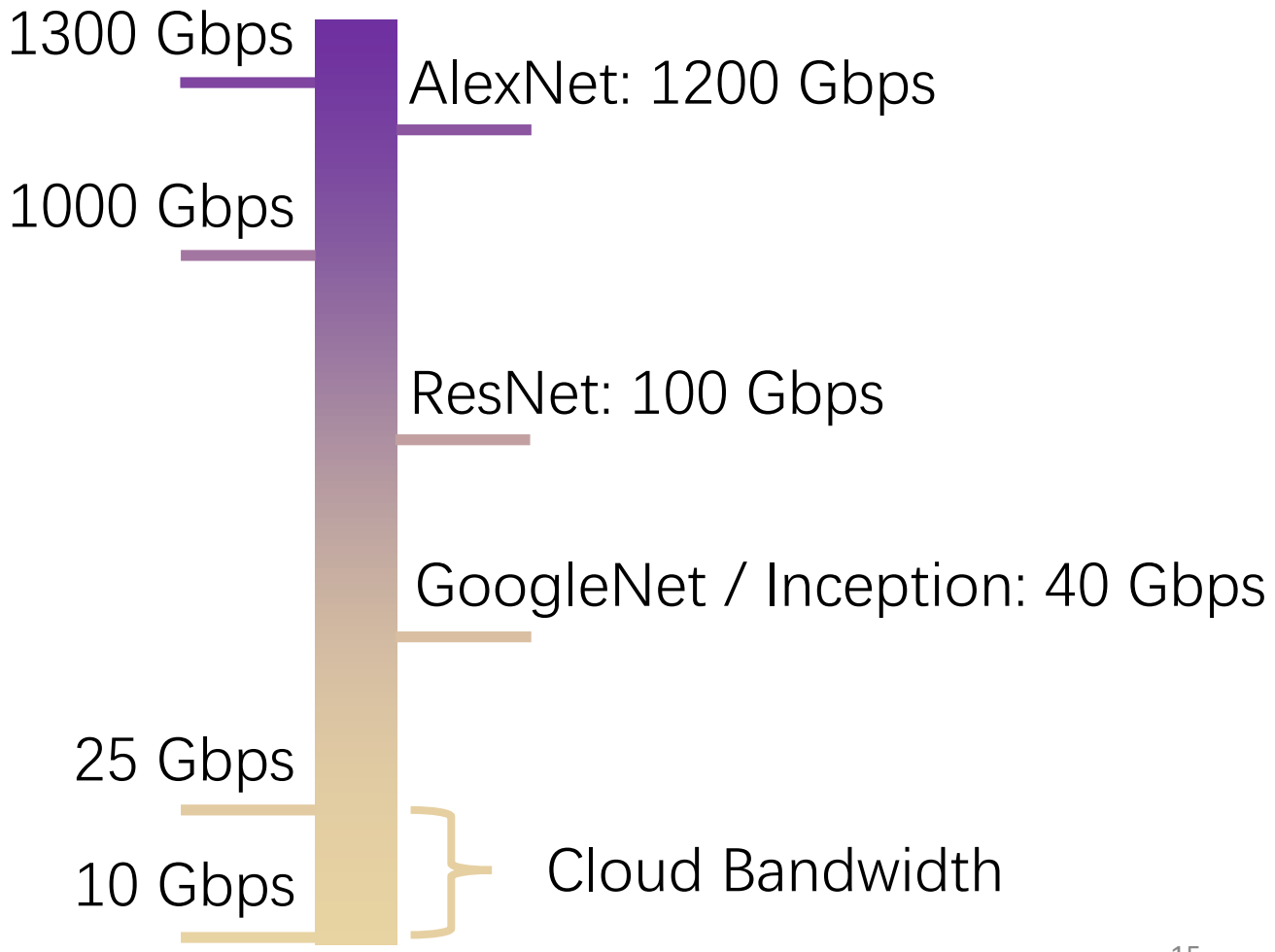


Bottlenecks in Cloud-based DDNN training

INSUFFICIENT BANDWIDTH

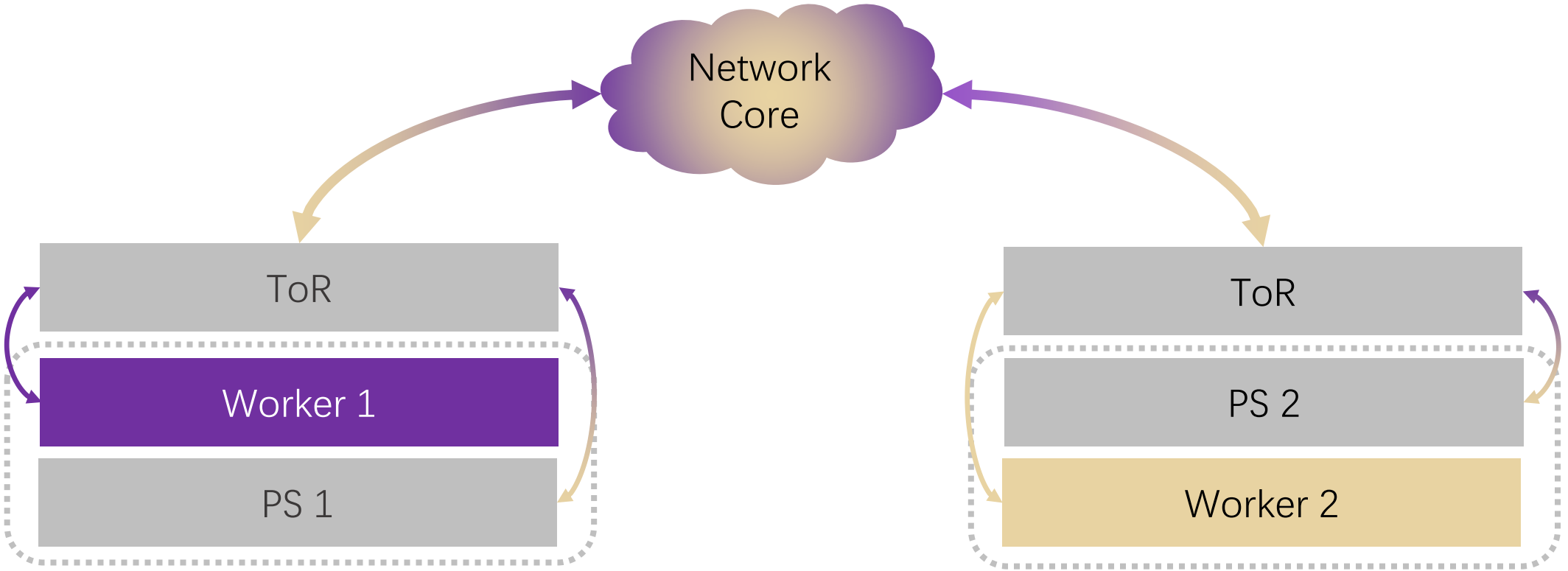
Minimum bandwidth required for each of the popular NNs for communication to not bottleneck computation?

8 workers, GTX 1080 Ti, central parameter servers. MxNet



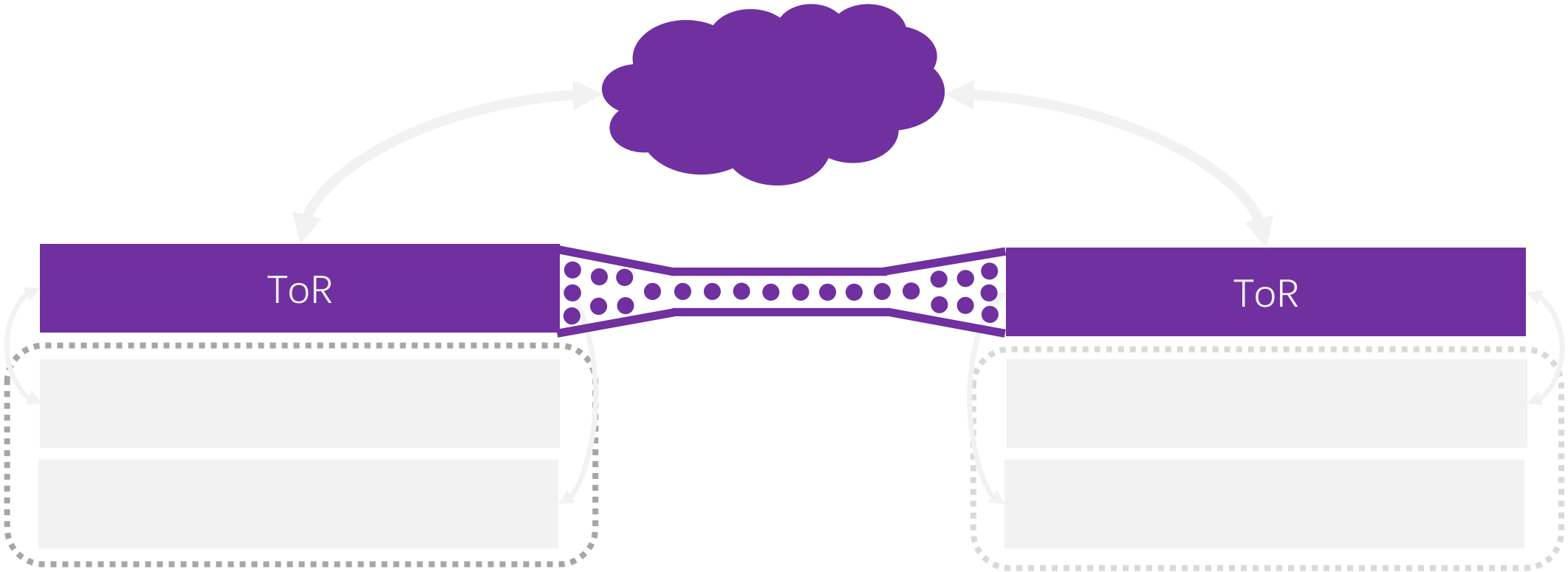
Bottlenecks in Cloud-based DDNN training

MAPPING OF TRAINING WORKLOAD TO THE CLOUD IS INEFFICIENT.



Bottlenecks in Cloud-based DDNN training

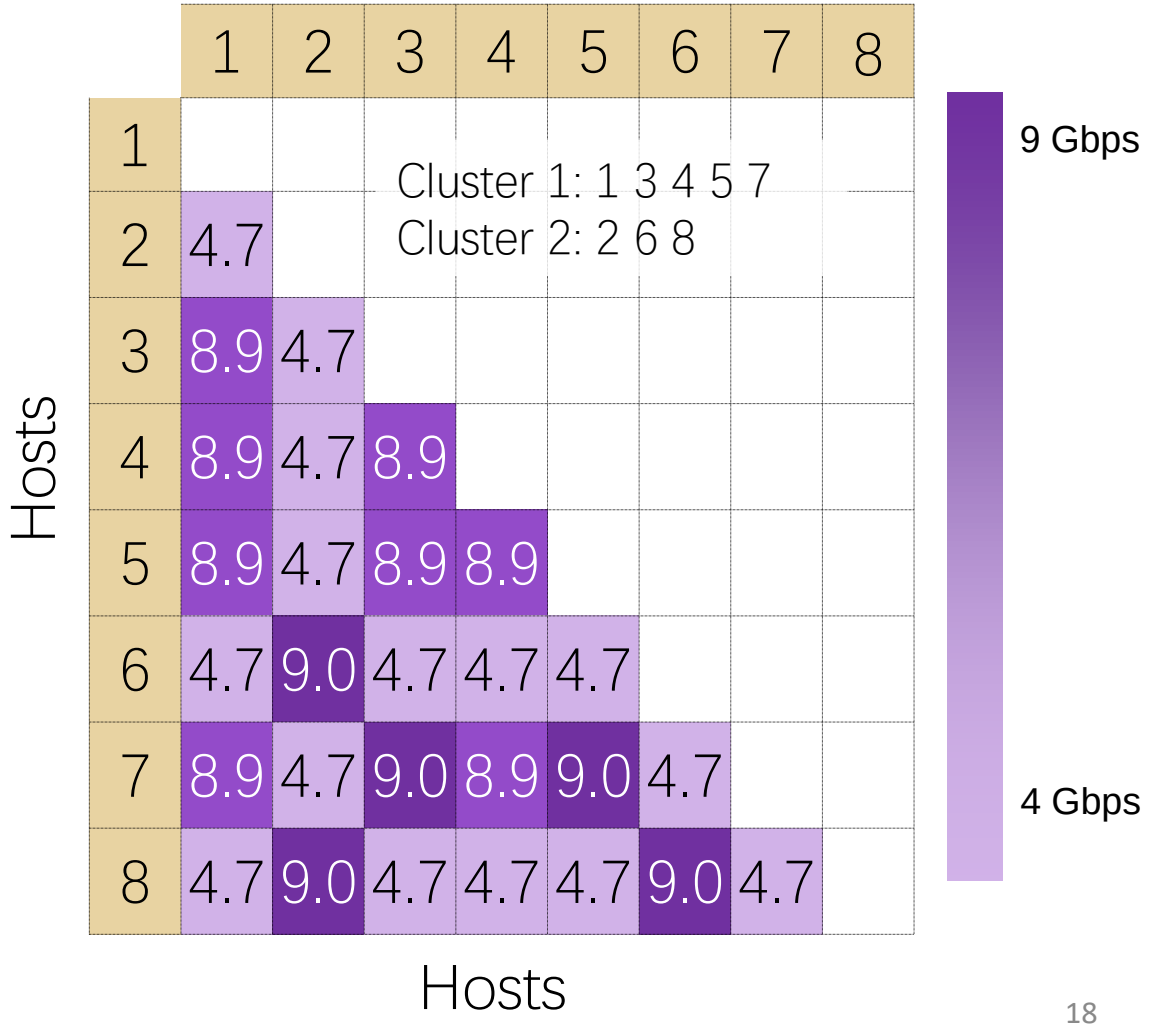
DEPLOYMENT-RELATED OVERHEAD



Bottlenecks in Cloud-based DDNN training

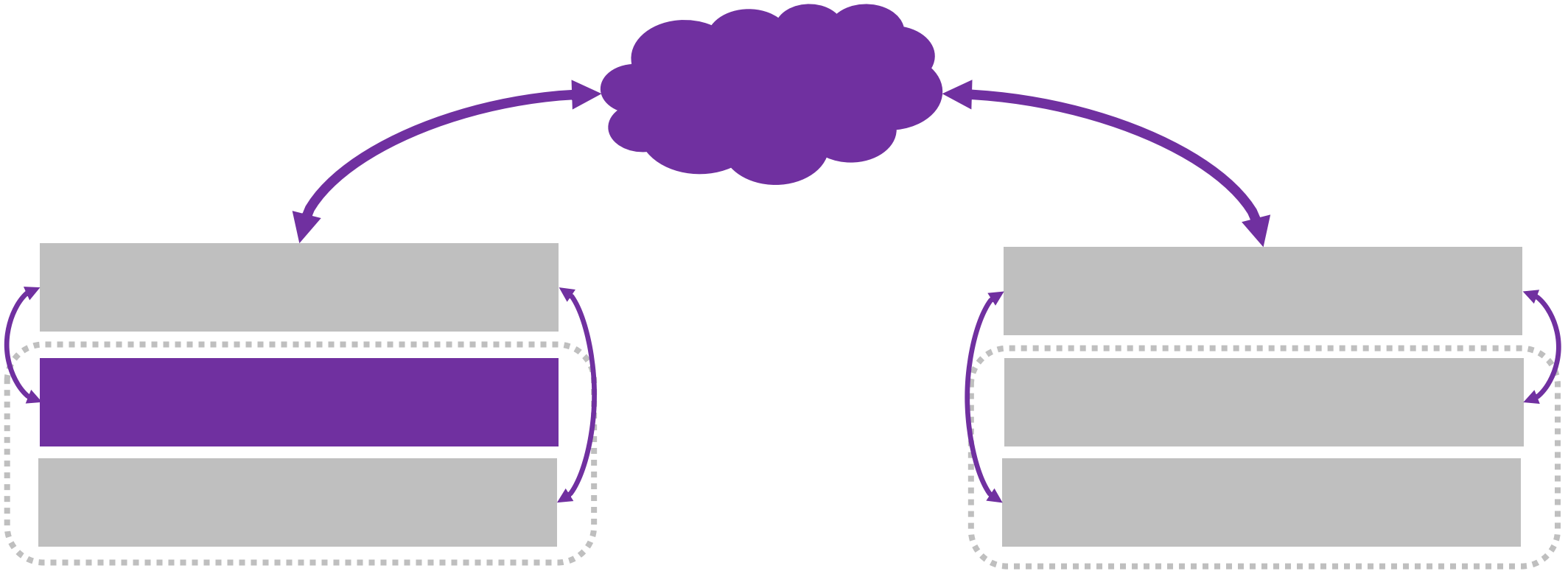
DEPLOYMENT-RELATED OVERHEAD

- Transient congestion, or oversubscription by design
- Cross-rack communication cost is higher than Intra-rack communication.
- Comm. bottlenecked by slowest link.



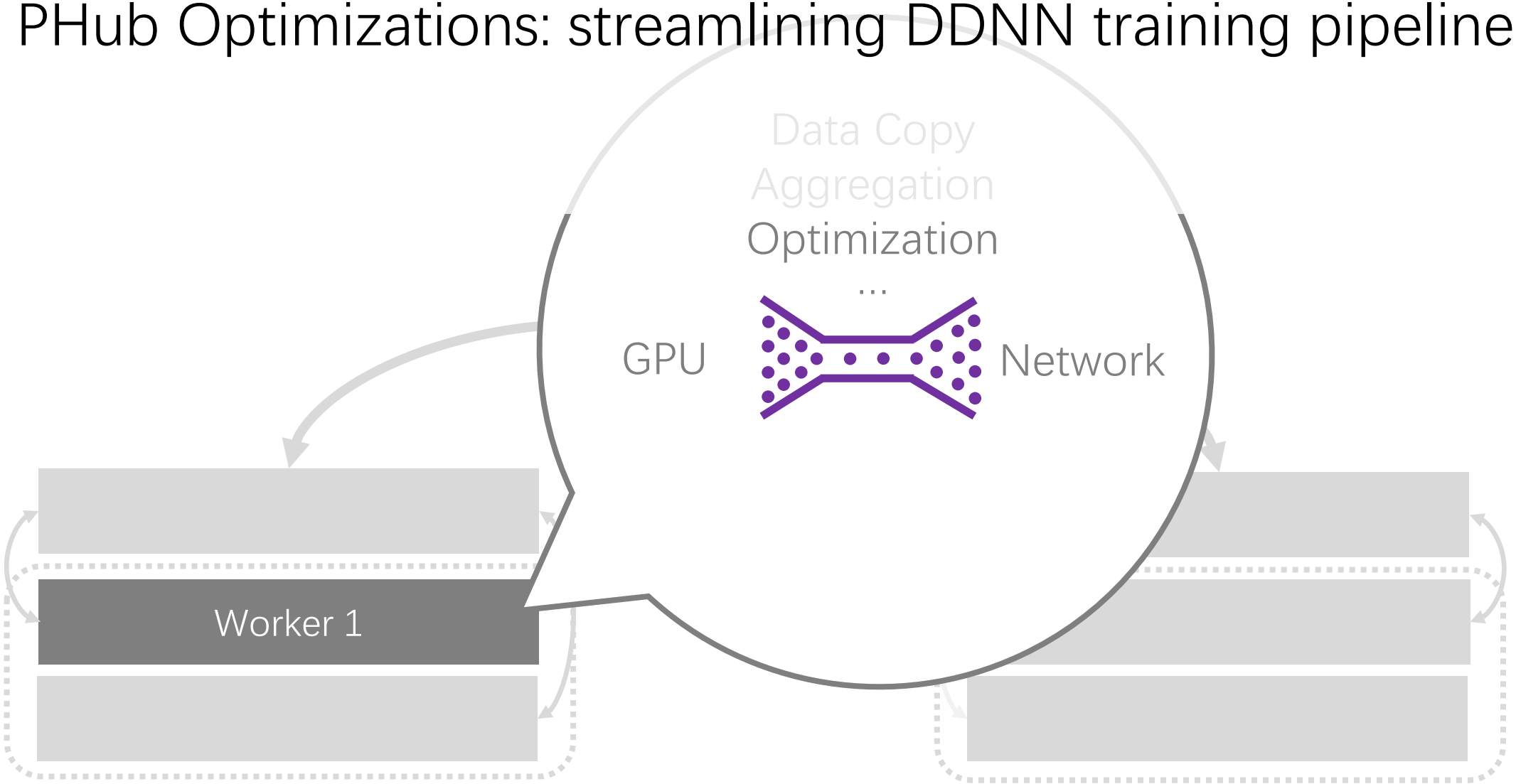
Parameter Hub Optimizations

CODESIGNING SOFTWARE, HARDWARE AND CLUSTER CONFIGURATION FOR EFFICIENT CLOUD-BASED DDNN TRAINING



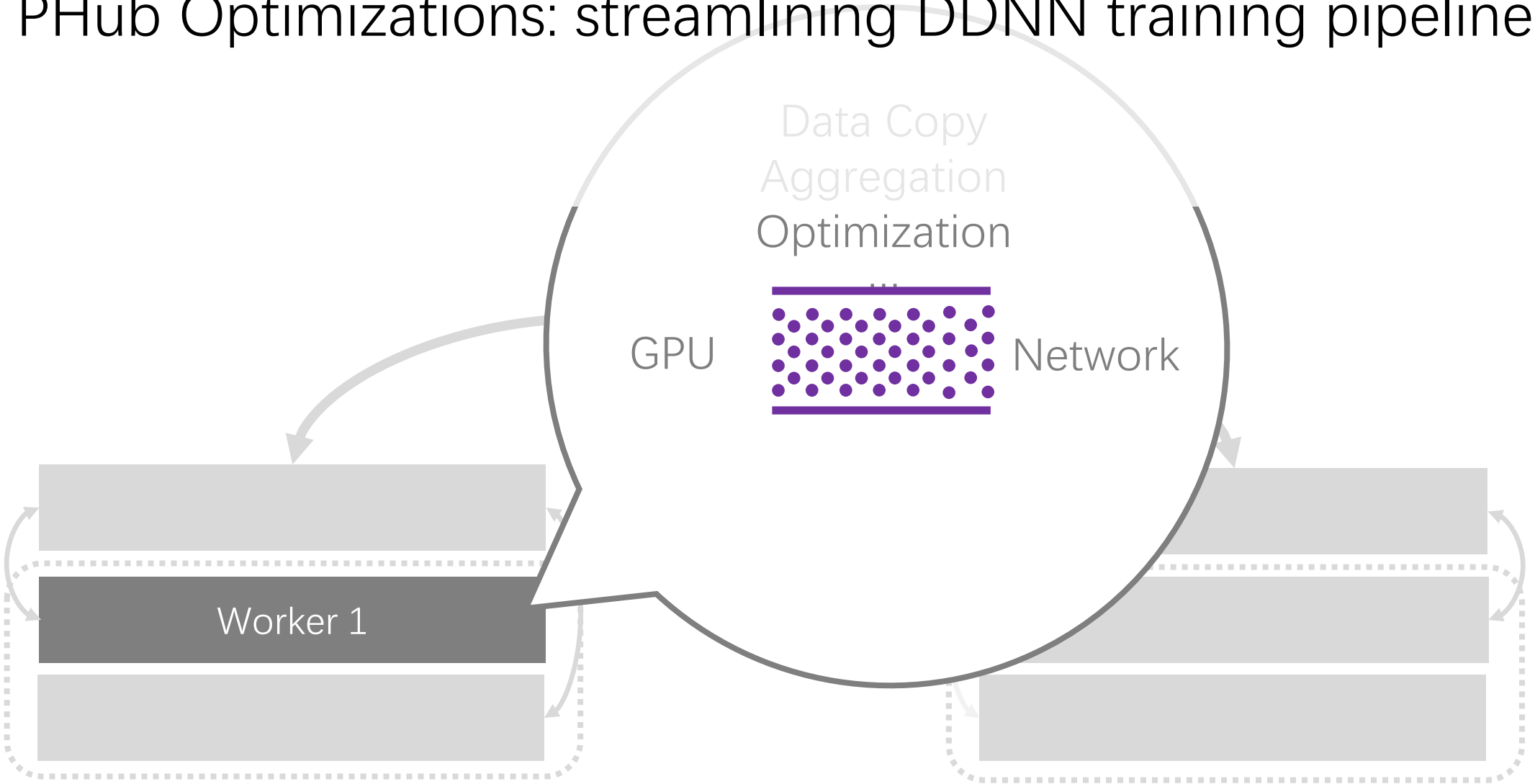
Eliminating framework bottlenecks:

PHub Optimizations: streamlining DDNN training pipeline

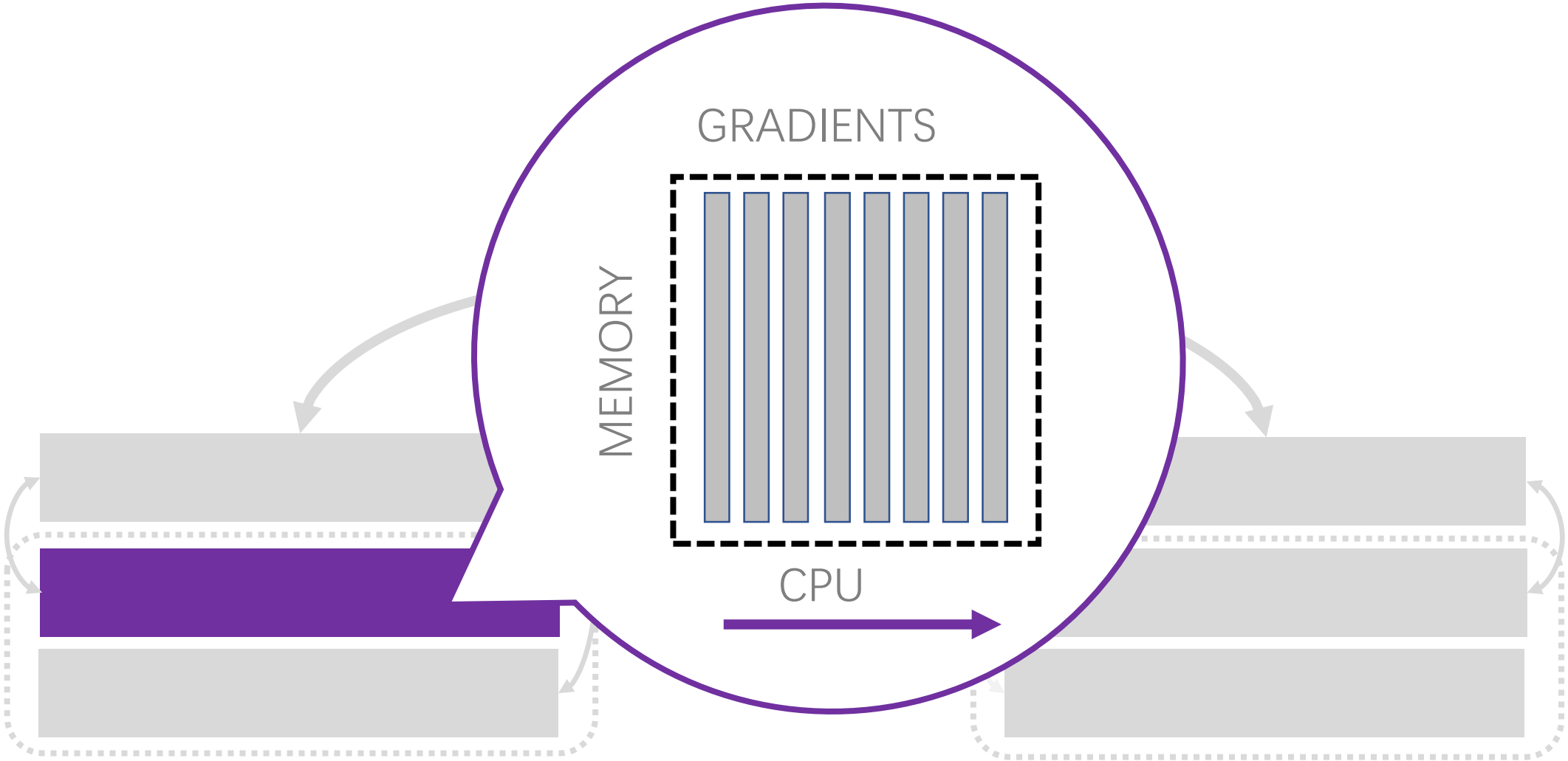


Eliminating framework bottlenecks:

PHub Optimizations: streamlining DDNN training pipeline



Software Optimizations



Software Optimizations

GRADIENT AGGREGATION AND OPTIMIZATION

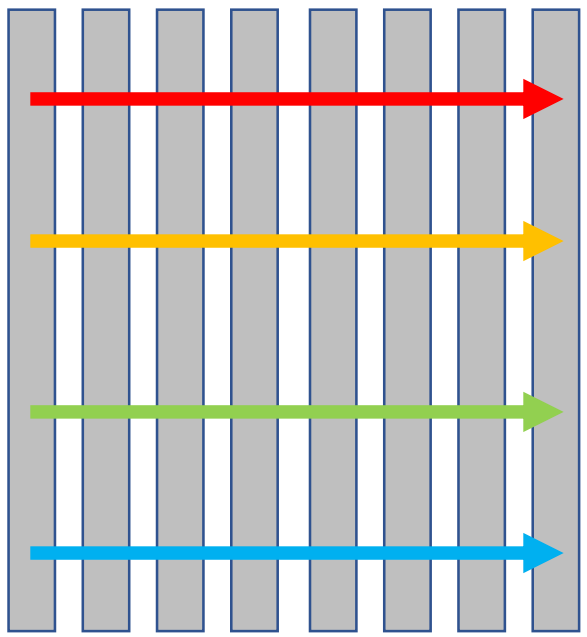
Requires synchronization.

Each core reads the input Q from different workers and writes to different locations to the output queue

Great locality. No synchronization

For each input Q, launch a series of threads for aggregation. This is used in MxNet. (Wide Aggregation)

Great locality. No synchronization



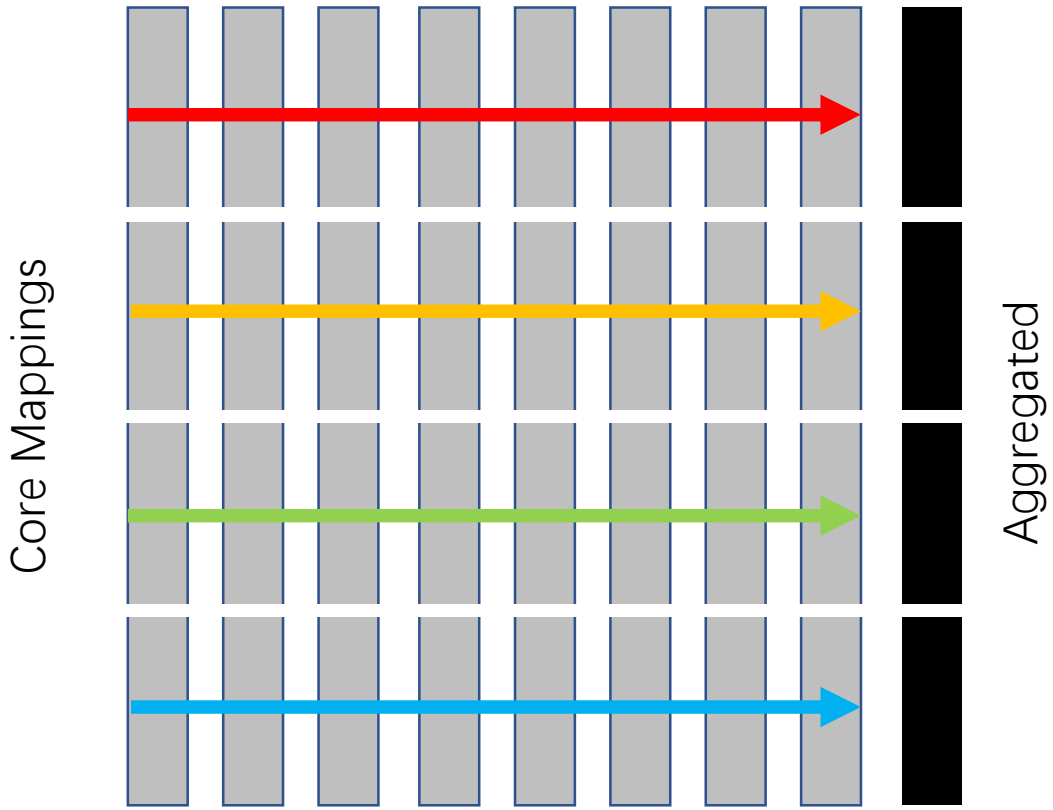
Sequentially aggregates the same portion of gradients within each queue. (Tall Aggregation)

Too much coherence and synchronization

Organize processors into hierarchy. Perform NUMA aware tree reduction.

Software Optimizations

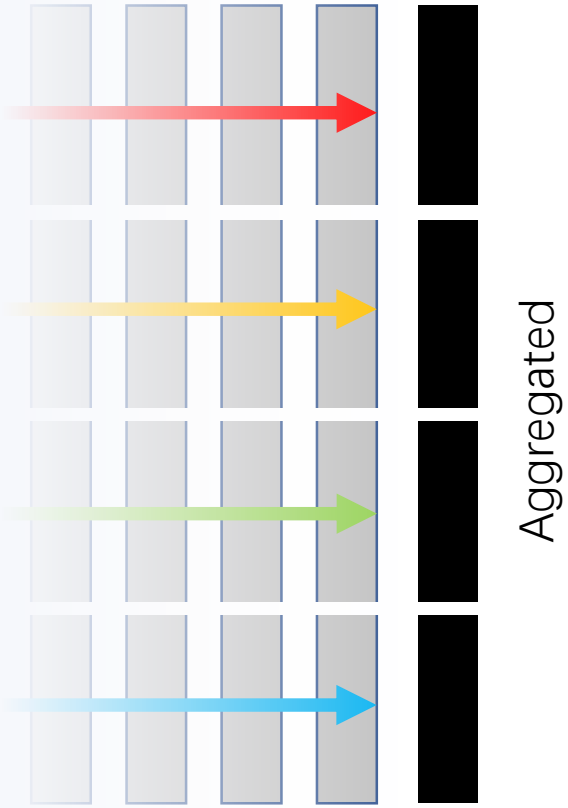
TALL AGGREGATION AND OPTIMIZATION



- Chunk a gradient into a series of virtual gradients deterministically.
- A virtual gradient is mapped to a particular core on the server.
- Virtual gradients are transferred independently.
- A chunk is only processed by a single core : maintaining maximum locality.

Software Optimizations

TALL AGGREGATION AND OPTIMIZATION

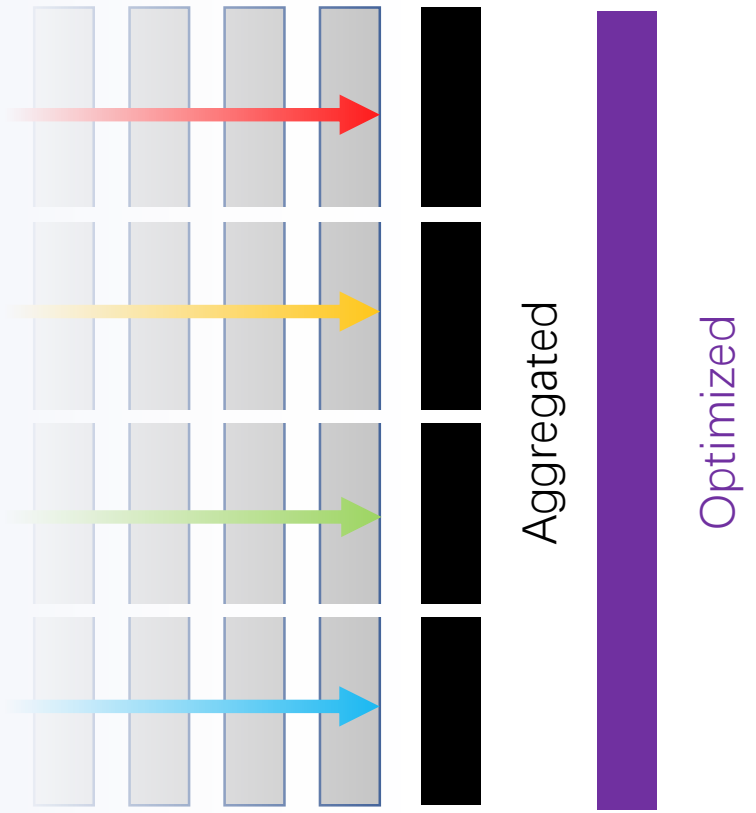


When Aggregation is done, PHub:

- PHub optimizes a chunk with the same core that aggregates that chunk.

Software Optimizations

TALL AGGREGATION AND OPTIMIZATION



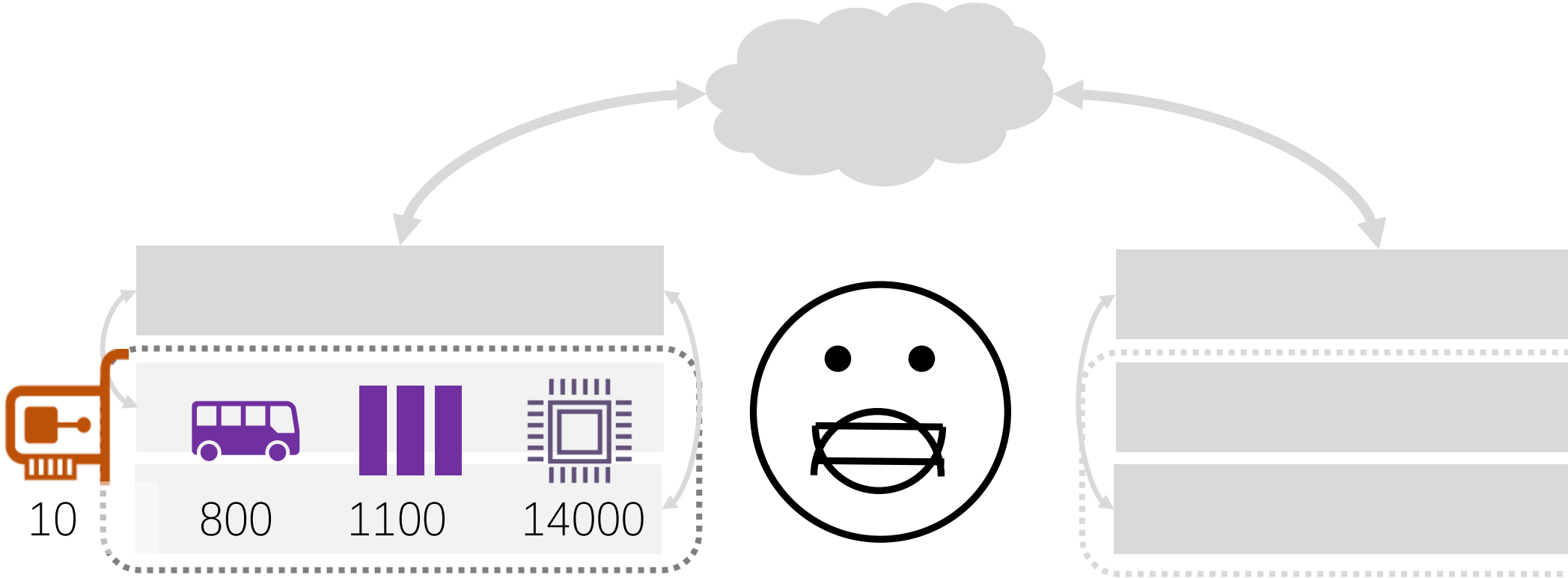
When Aggregation is done, PHub:

- PHub optimizes a chunk with the same core that aggregates that chunk.
- Allows overlapping of aggregation, optimization and gradient transmission.

Software Optimizations

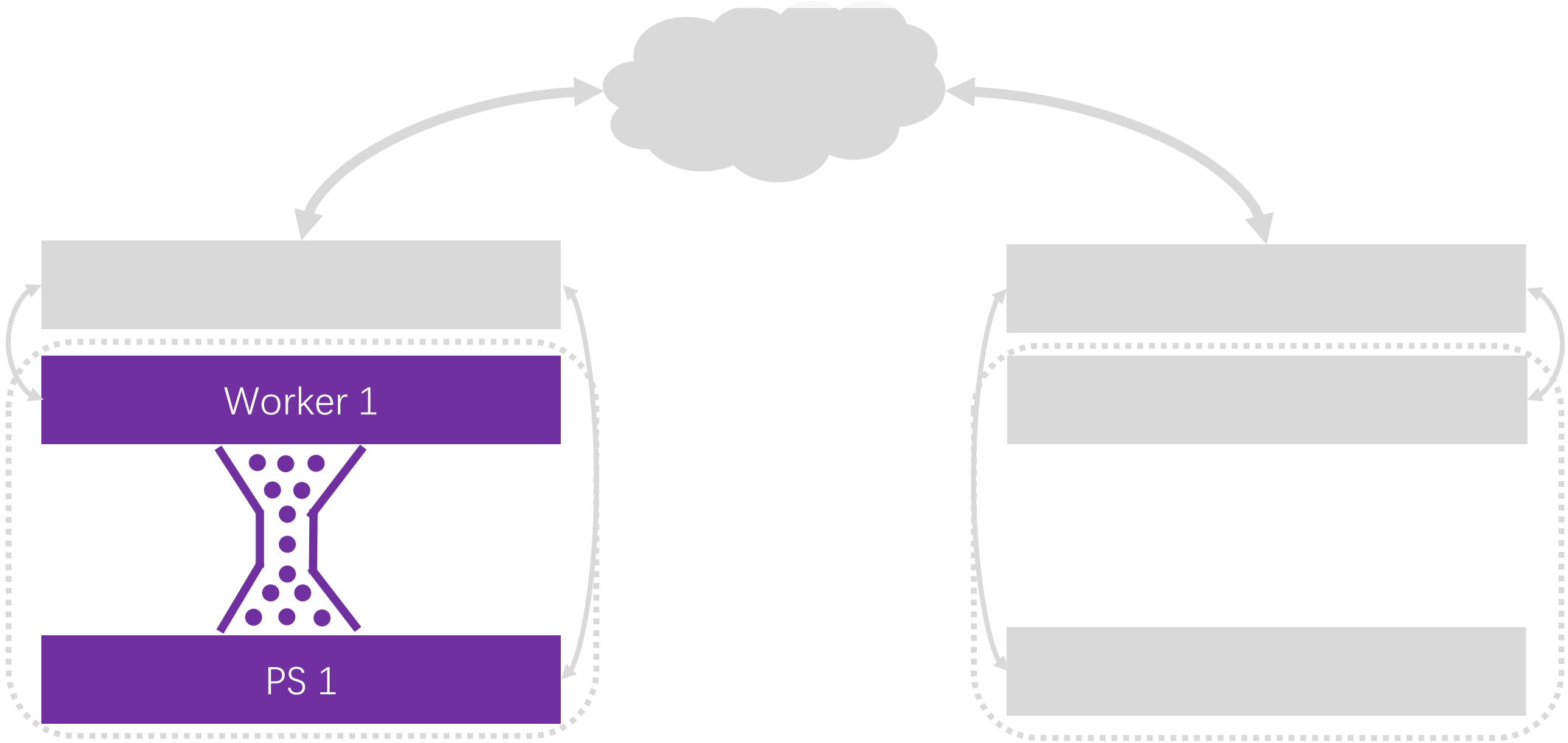
NOT ENOUGH ON THEIR OWN!

Typical server configuration is unbalanced



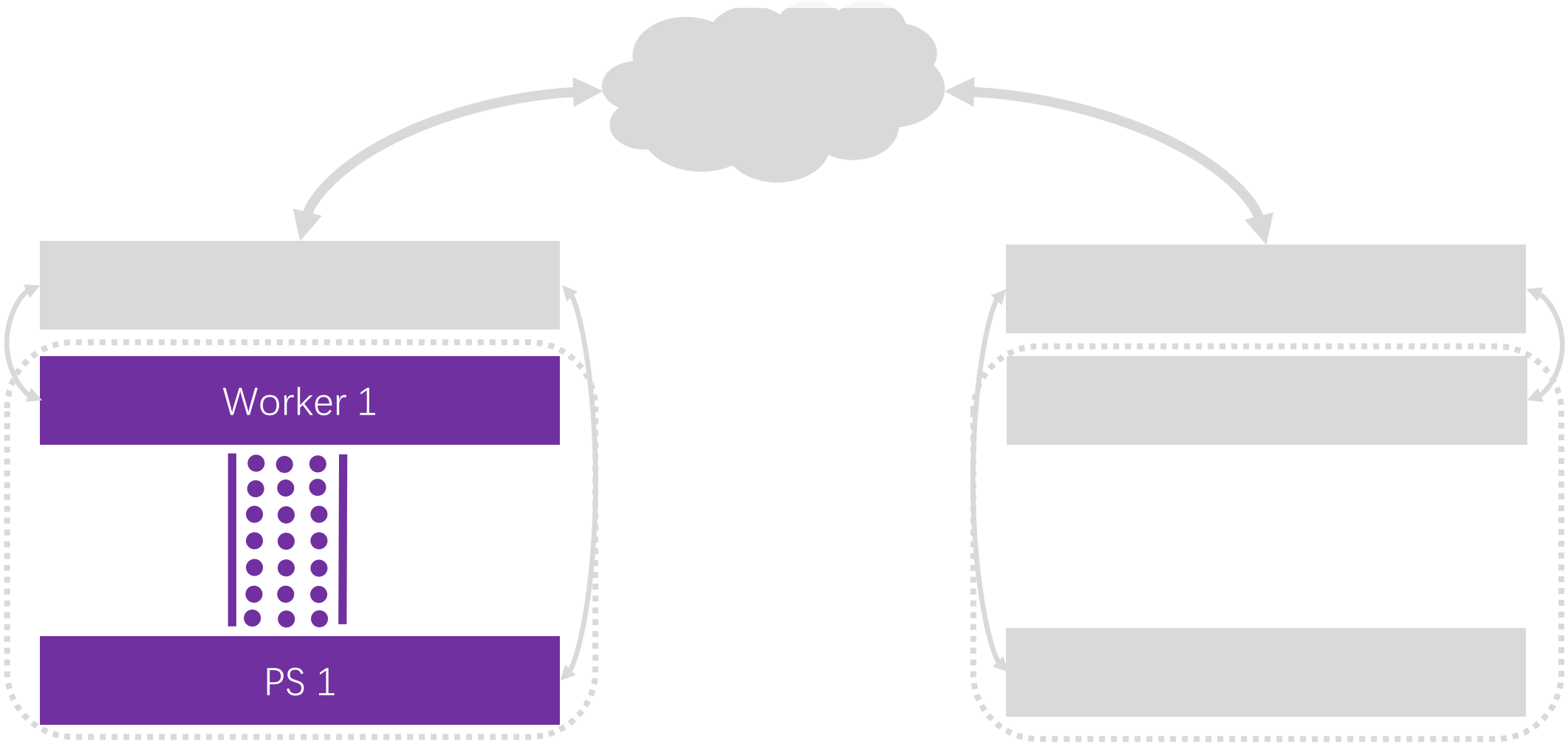
Eliminating bandwidth bottlenecks:

PBox hardware: balanced computation and communication resources.



Eliminating bandwidth bottlenecks:

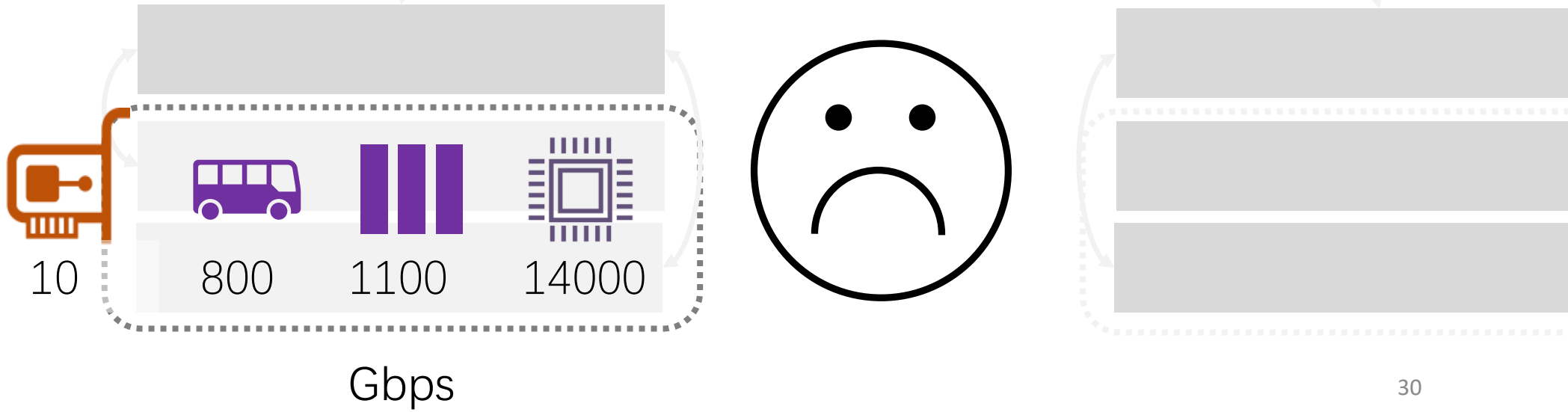
PBox hardware: balanced computation and communication resources.



Hardware Optimization

THE PBOX

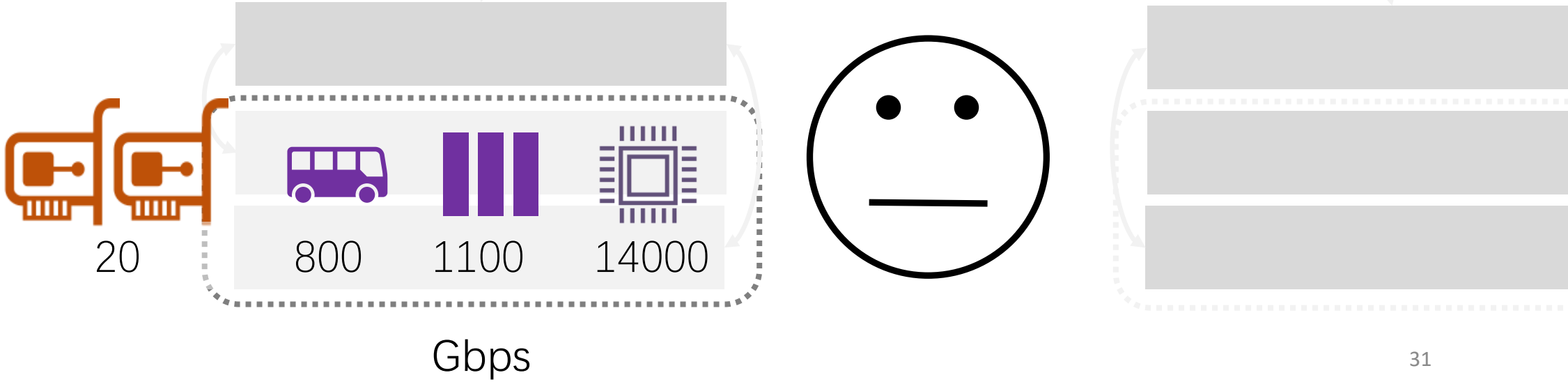
- Balanced computation and communication
- Extends the balance and locality notion across NUMA domains and NICs.



Hardware Optimization

THE PBOX

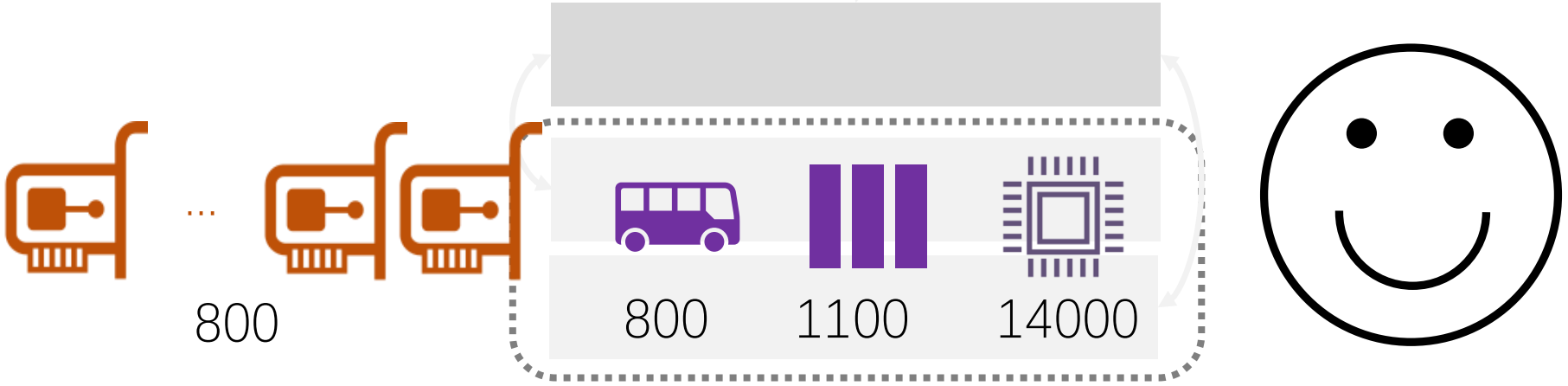
- Balanced computation and communication
- Extends the balance and locality notion across NUMA domains and NICs.



Hardware Optimization

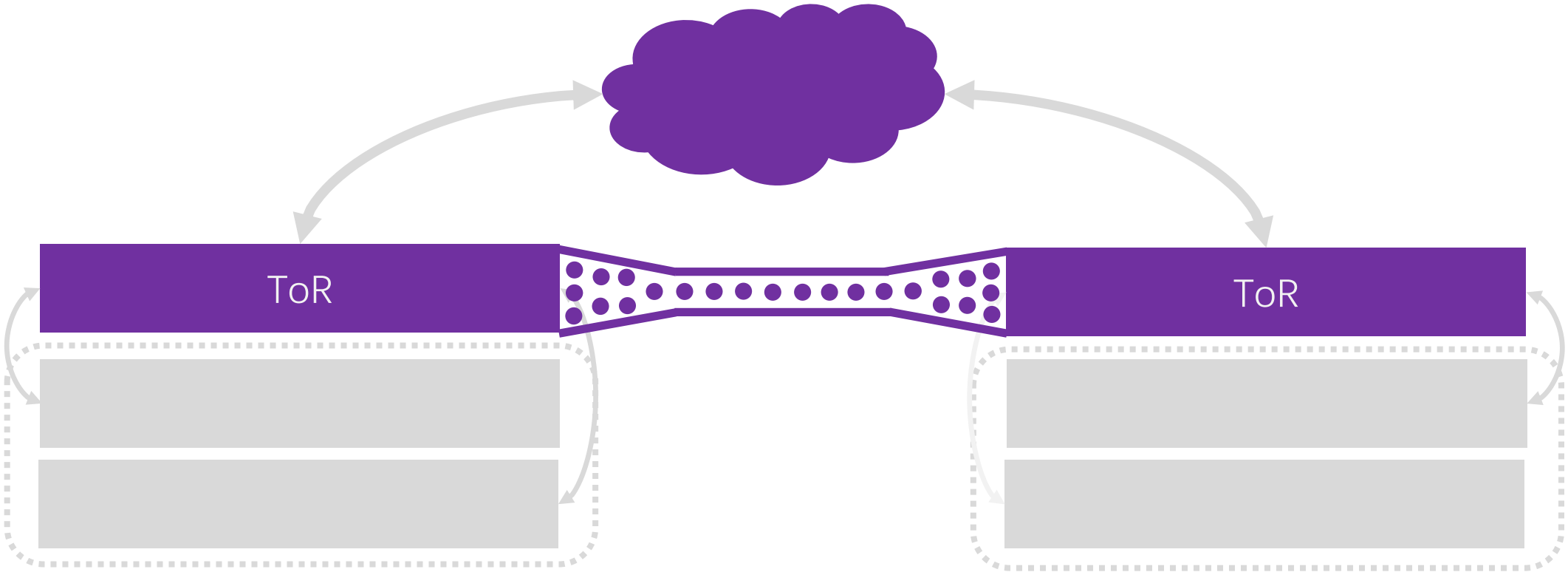
THE PBOX

- Balanced computation and communication
- Extends the balance and locality notion across NUMA domains and NICs.



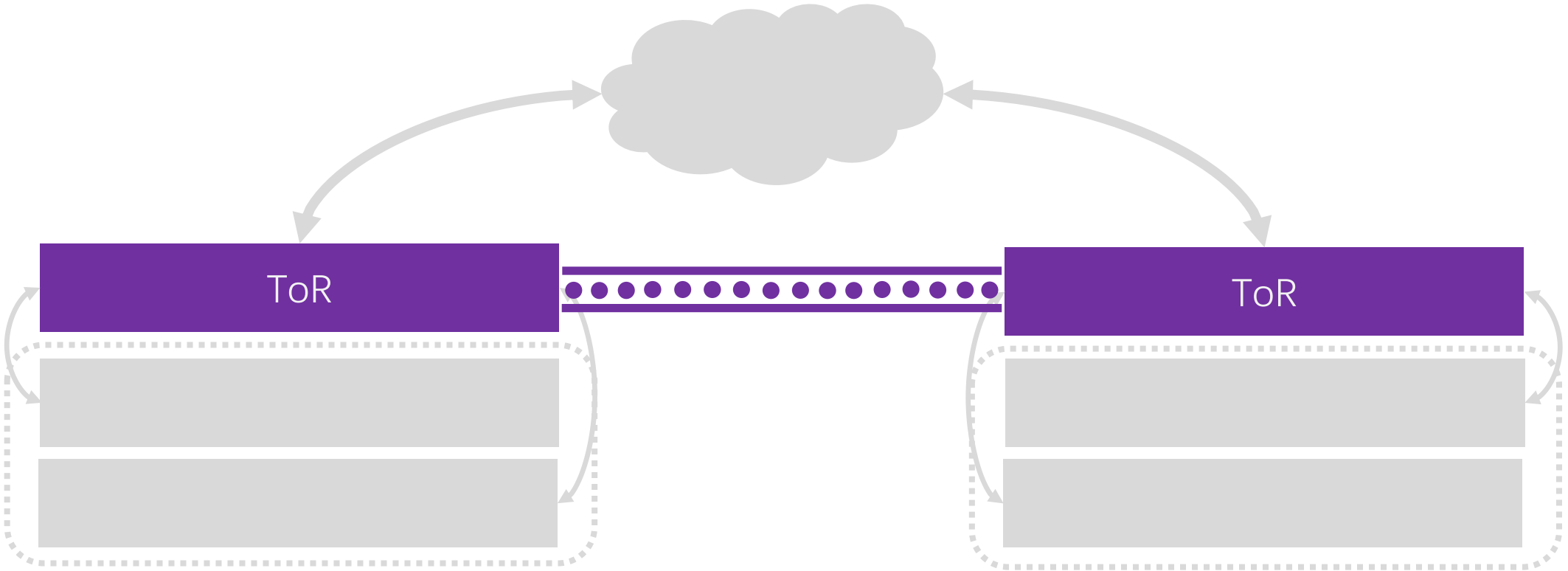
Eliminating deployment bottlenecks:

PHub hierarchical reduction: reducing cross rack traffic



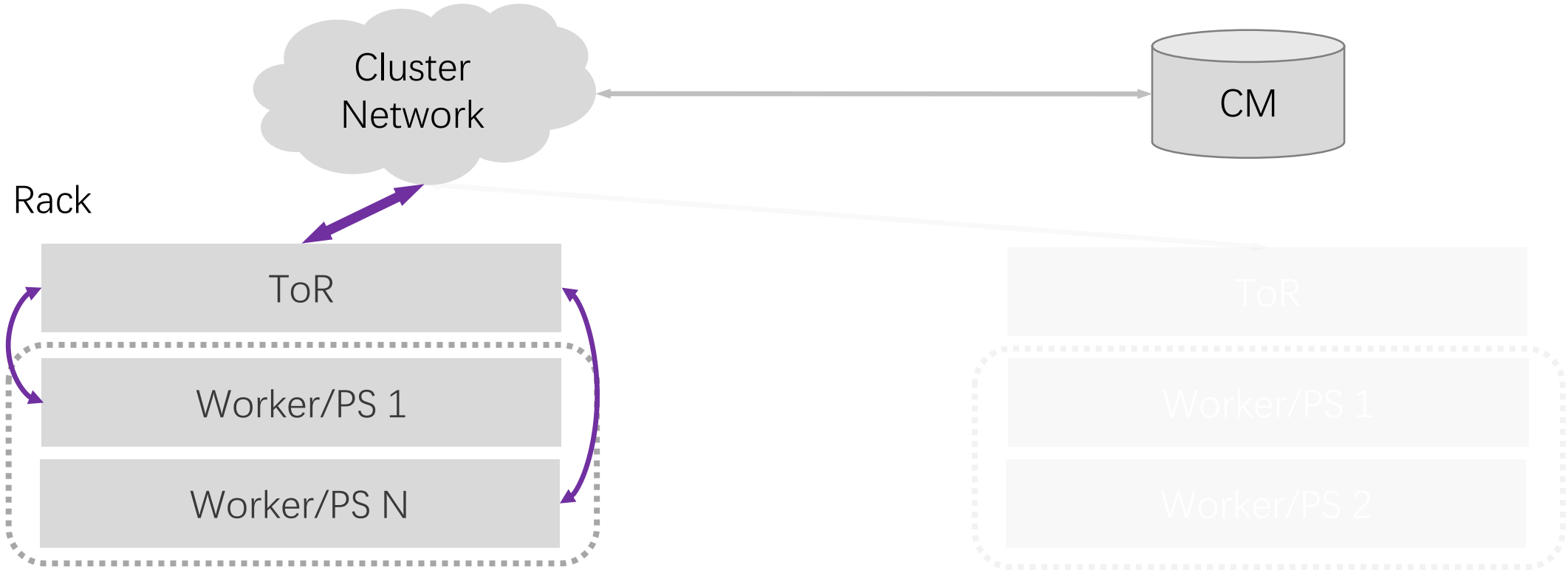
Eliminating deployment bottlenecks:

PHub hierarchical reduction: reducing cross rack traffic



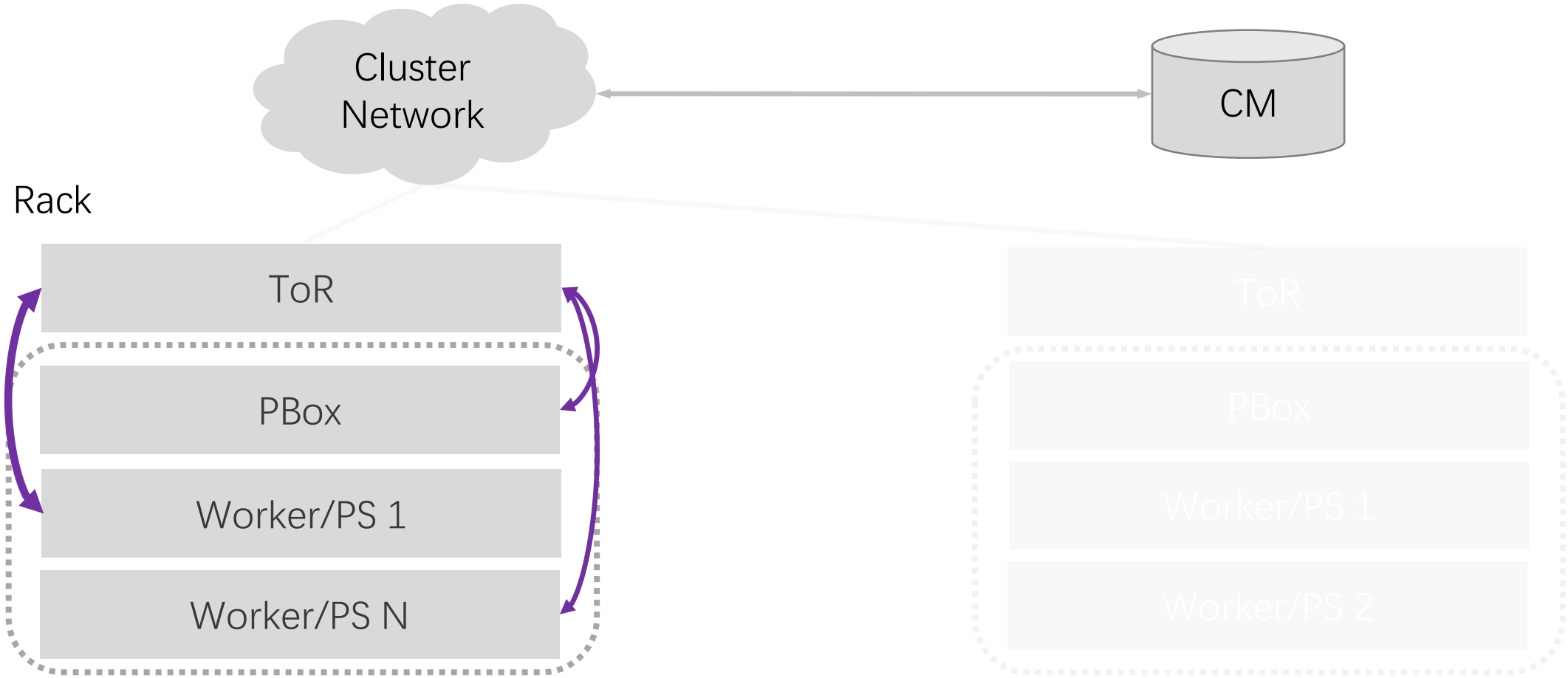
PBox Deployment

RACK SCALE PARAMETER SERVICE



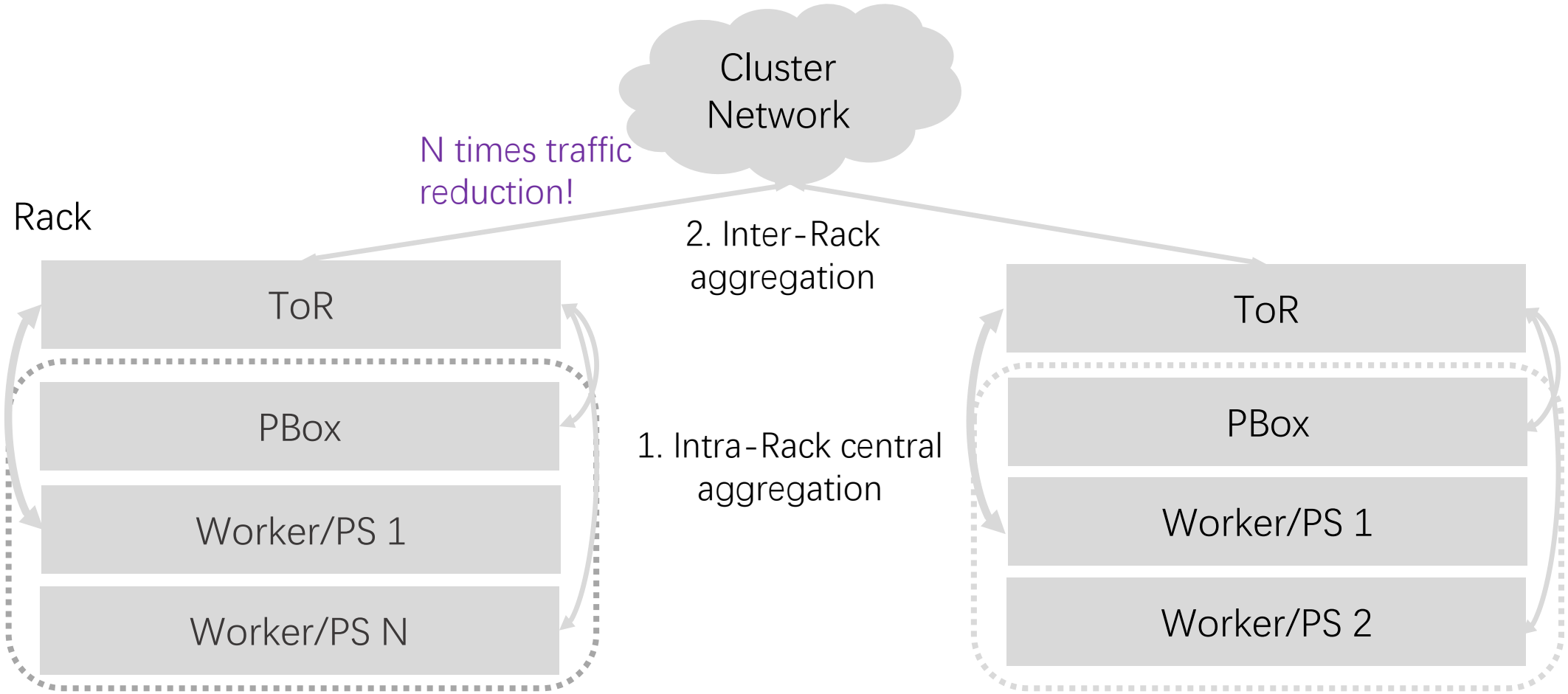
PBox Deployment

RACK SCALE PARAMETER SERVICE

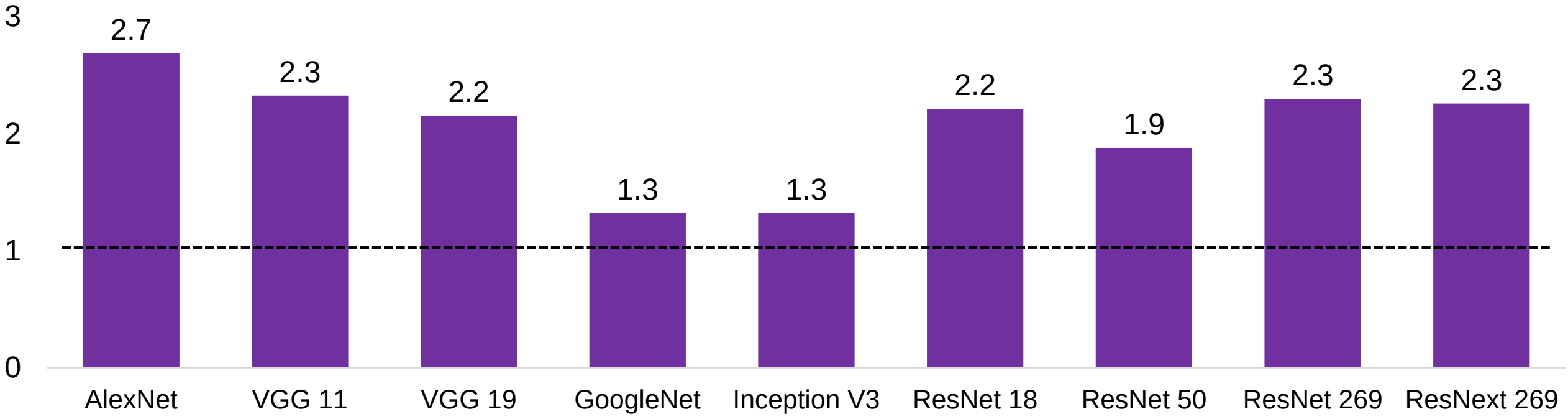


Two-Phase Hierarchical Aggregation

ADAPTING TO THE DATACENTER NETWORK TOPOLOGY



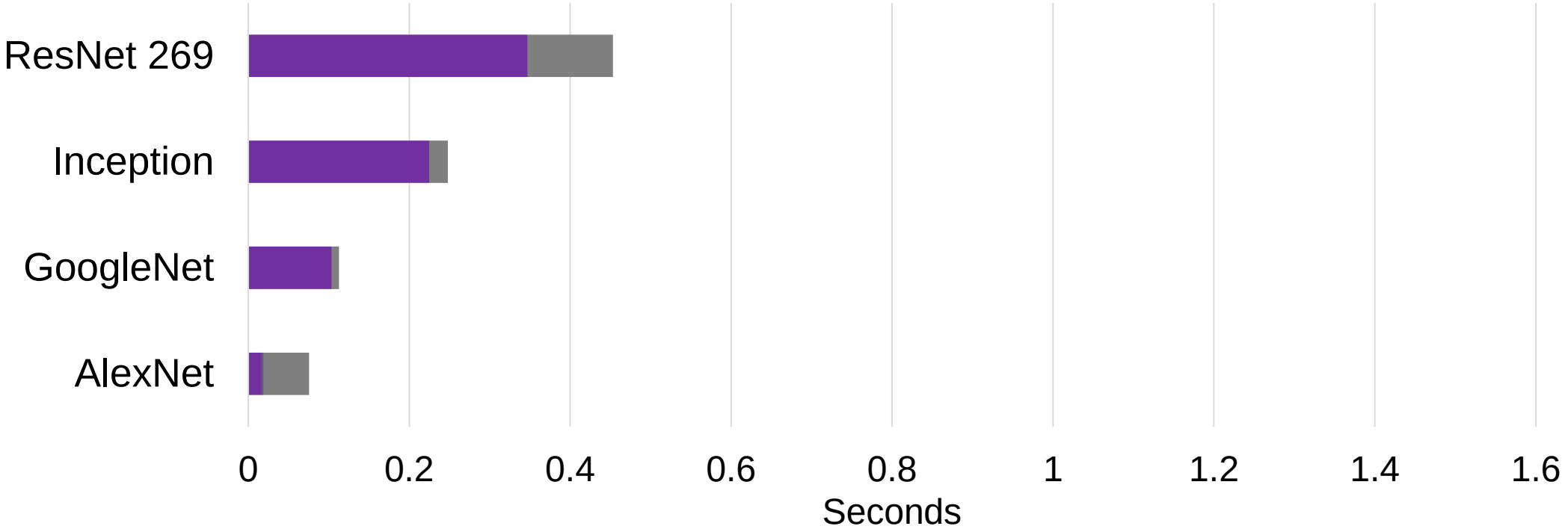
Up to 2.7x performance in 10Gbps cloud-like environment



8 Workers. GTX 1080 Ti. MxNet: InfiniBand-enhanced baseline. PBox. Batch Size 64 for ResNext, 128 for ResNet 269, 256 for all others.

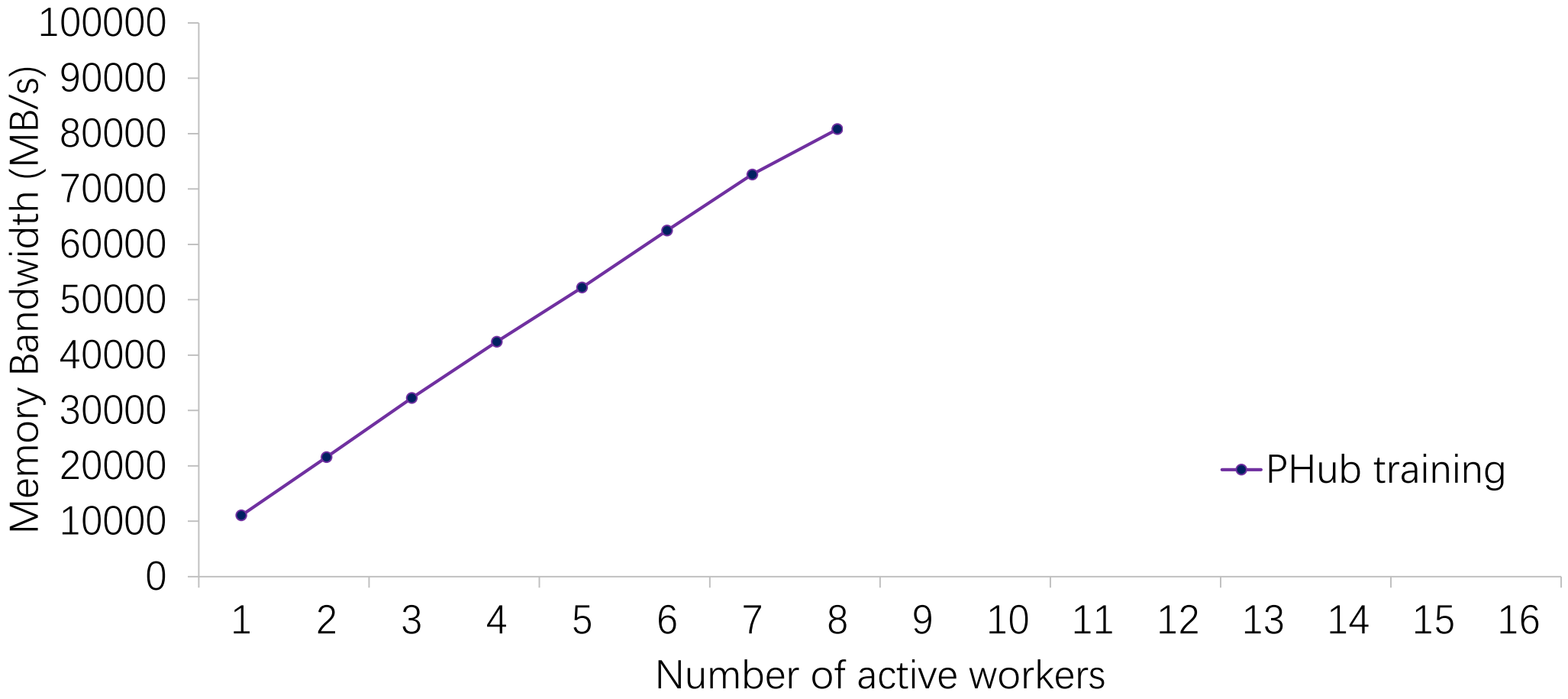
Framework Bottlenecks

- Data Copy
- Aggregation and Optimization
- Synchronization



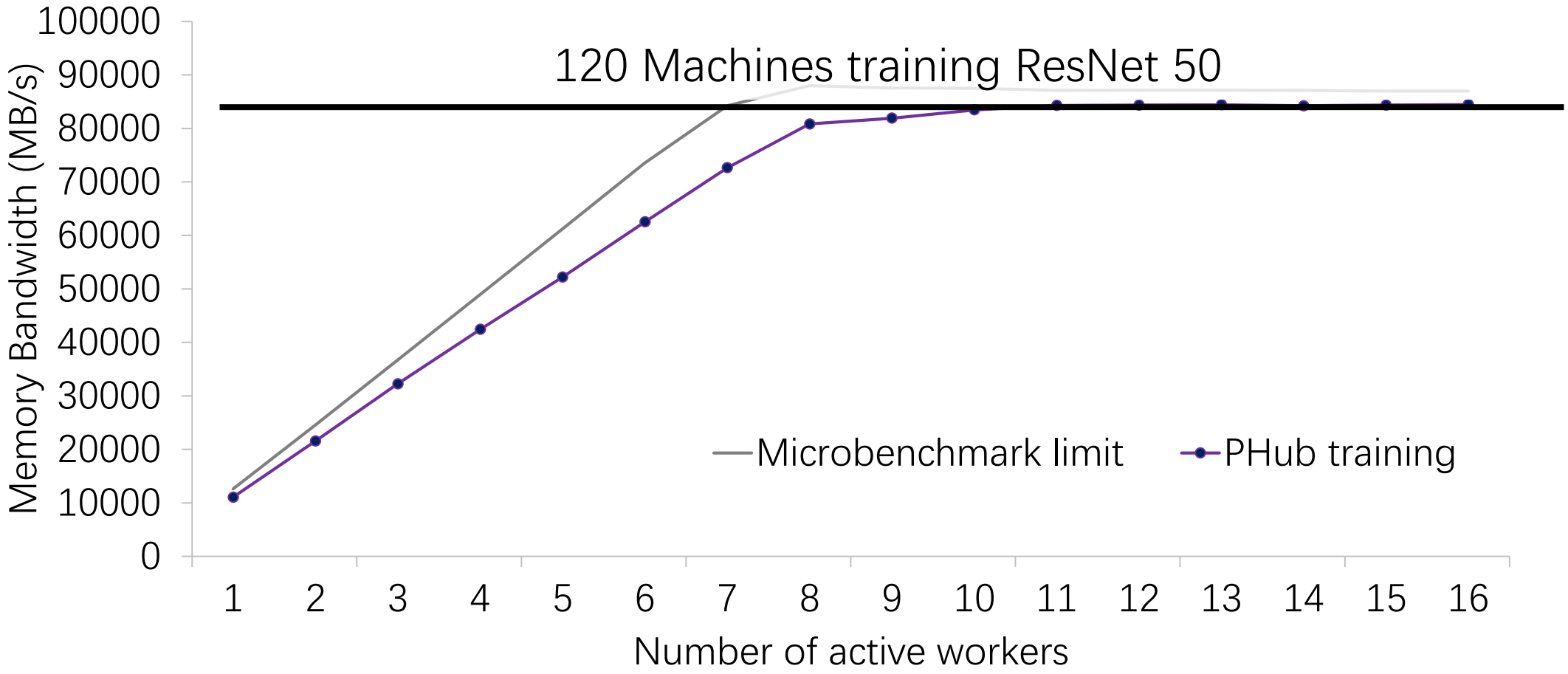
Scalability

LINEAR SCALING IN COMM. ONLY BENCHMARK



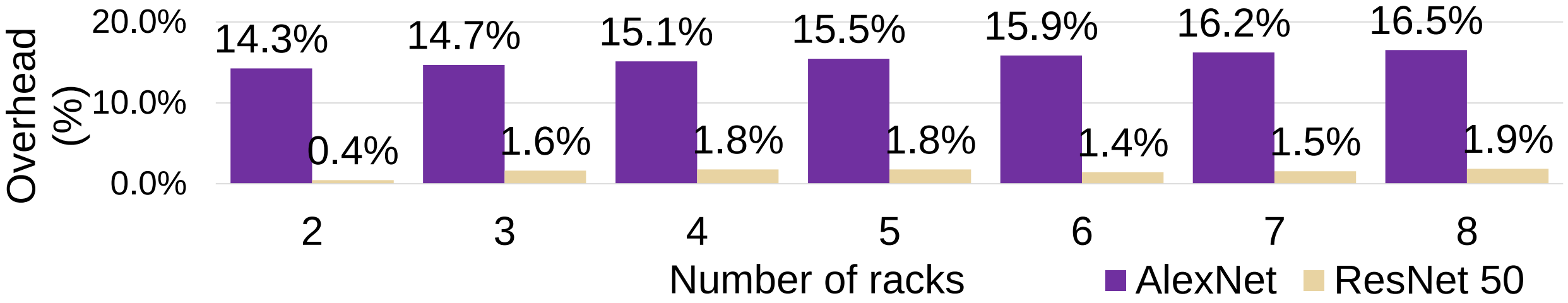
Scalability

PCI-E TO MEMORY SUBSYSTEM BRIDGE



Scalability Beyond a Single Rack

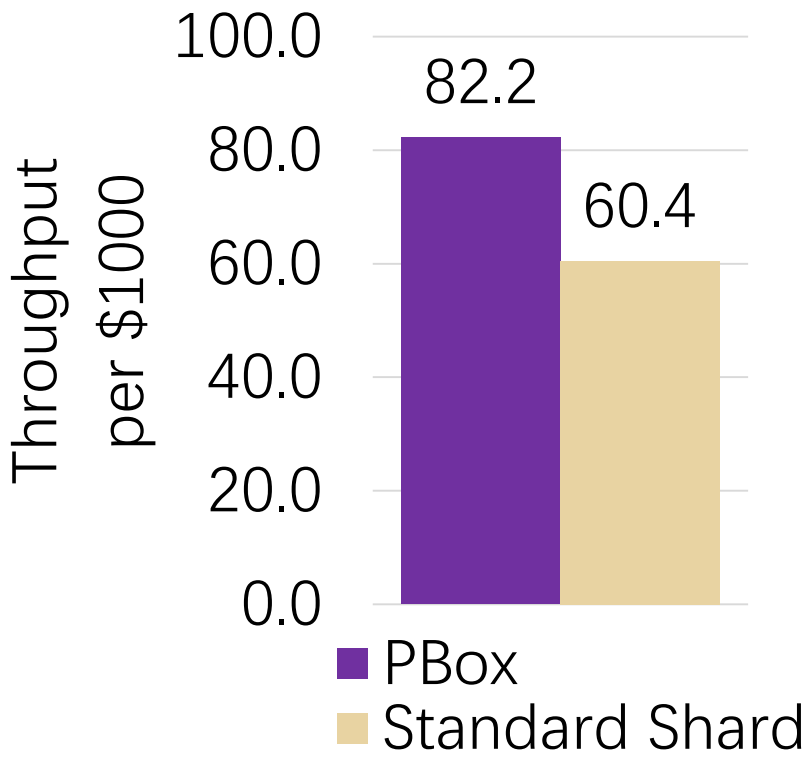
EMULATING HIERARCHICAL AGGREGATION



Overhead of Phub cross-rack synchronization

Cost Analysis – for infrastructure builders

25% BETTER THROUGHPUT/$



Accounting for network devices (switch ports, network adapters, network cables), GPU costs, and PBox's entire machine cost.

Core oversubscription 2:1

PAUL G. ALLEN SCHOOL

Parameter Hub

A software, hardware and cluster configuration codesign that target three major bottlenecks in the cloud for more efficient DDNN training