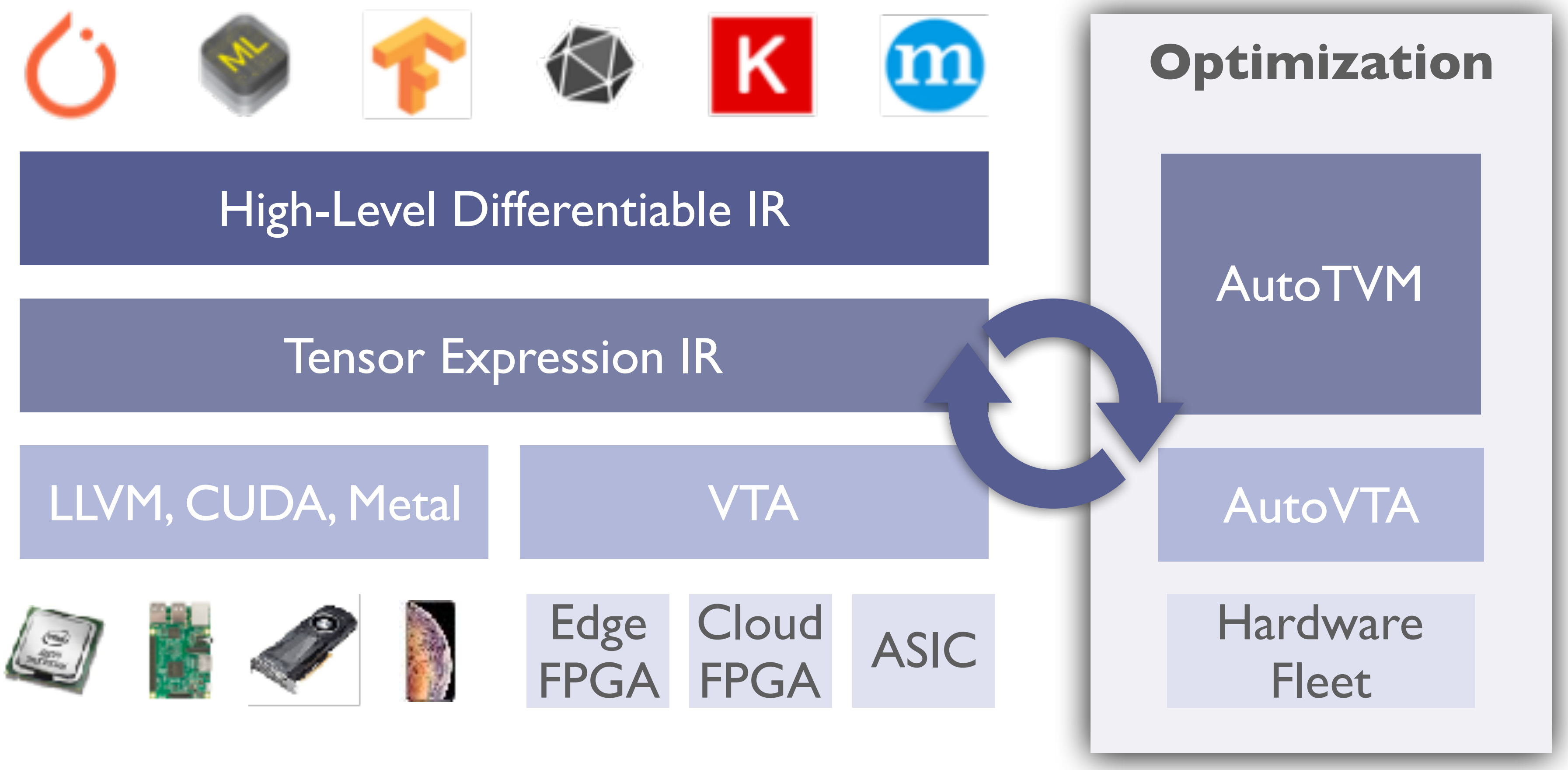
Scalable Distributed Training with Parameter Hub: a whirlwind tour

sampl

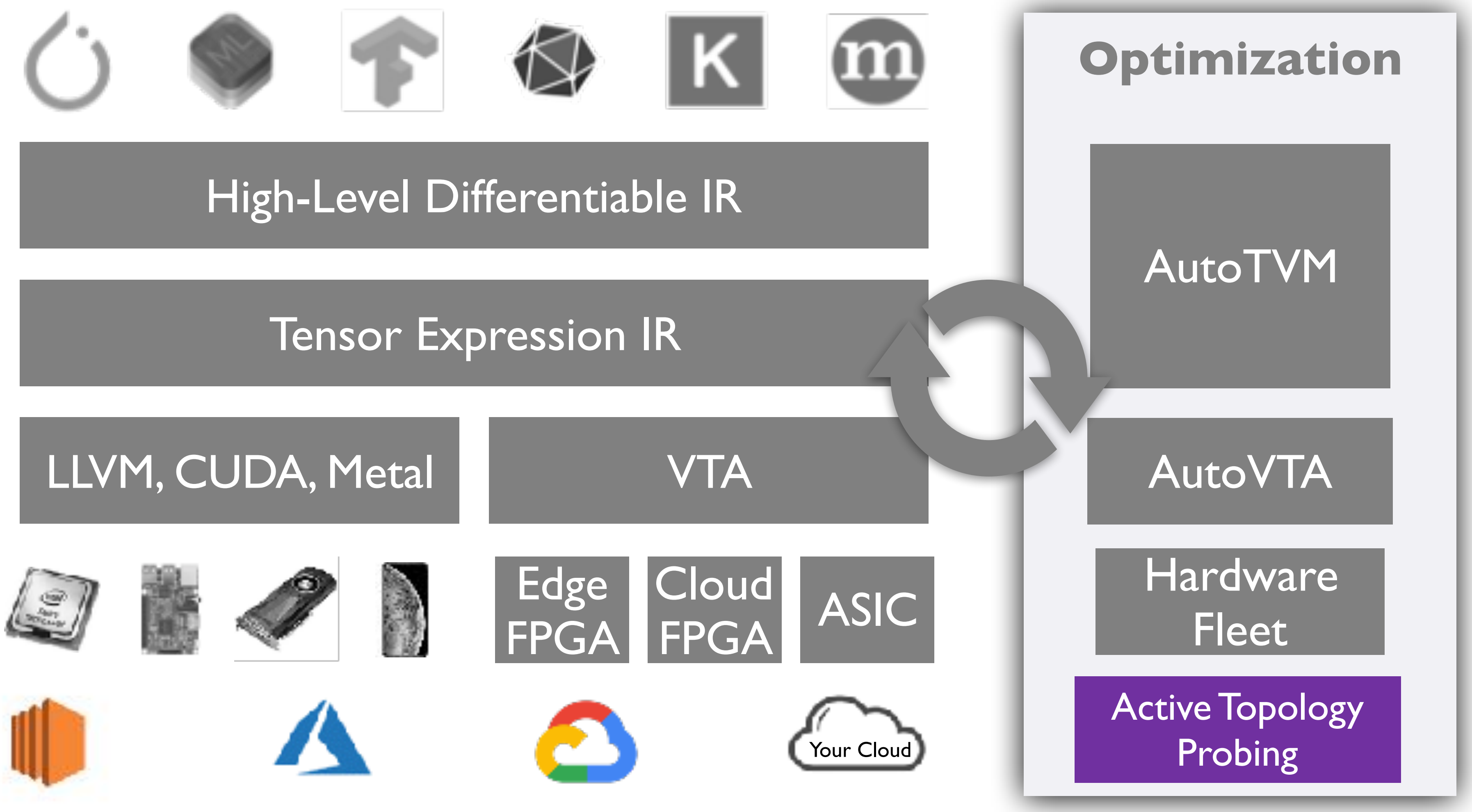
PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

tvm.ai

TVM Stack



High-Level Differentiable IR
Tensor Expression IR
LLVM, CUDA, Metal
VTA
Edge FPGA
Cloud FPGA
ASIC
Optimization
AutoTVM
AutoVTA
Hardware Fleet



Groundwork for bringing TVM to the distributed world for training and inference, on commercial cloud, or in your own cluster.

Parameter Hub

Optimized, topology-aware and dynamic mechanism for inter-machine communication

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

Parameter Hub

Optimized, topology-aware and dynamic mechanism for inter-machine communication

* In the cloud-based training context

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

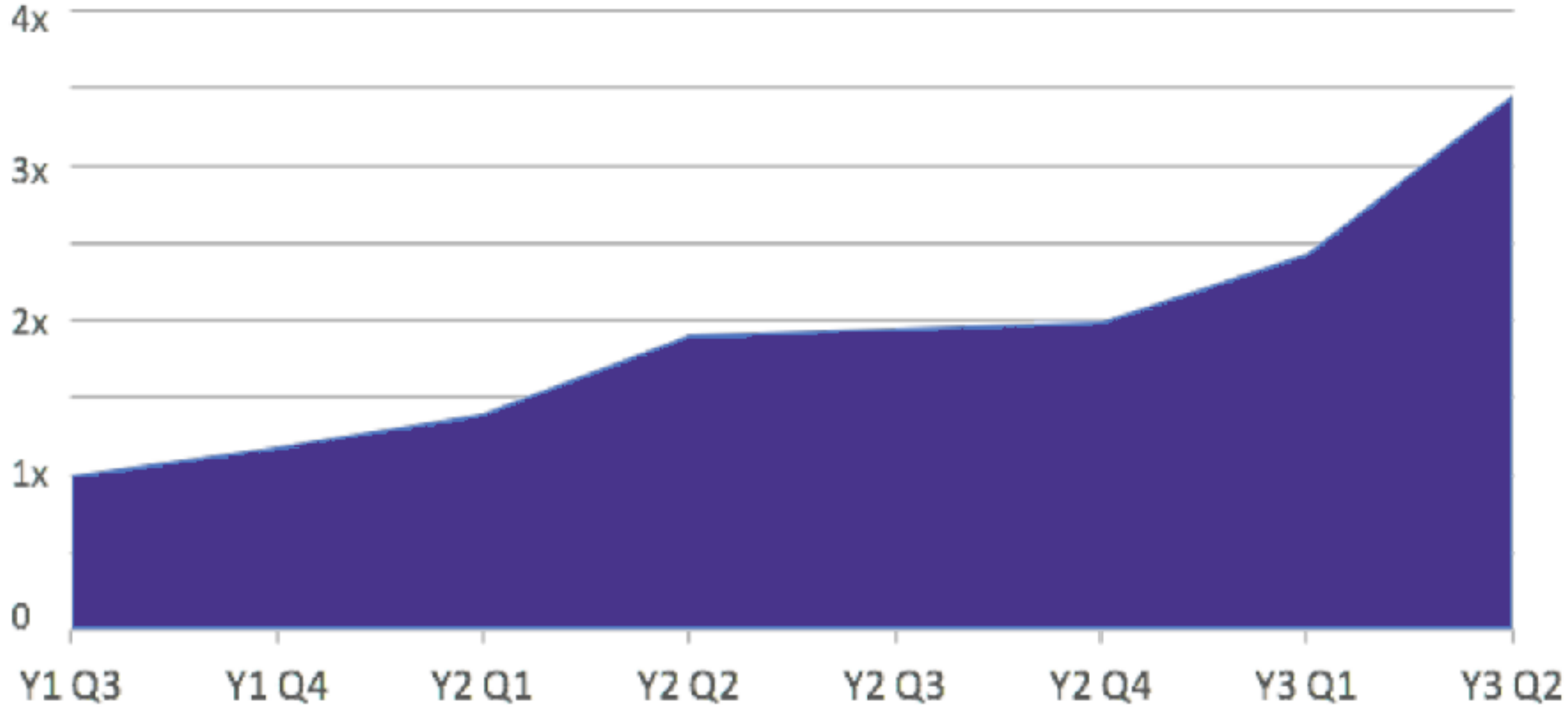
Deep Learning constitutes an important workload in cloud today.

Major cloud providers all have an ecosystem for cloud learning.

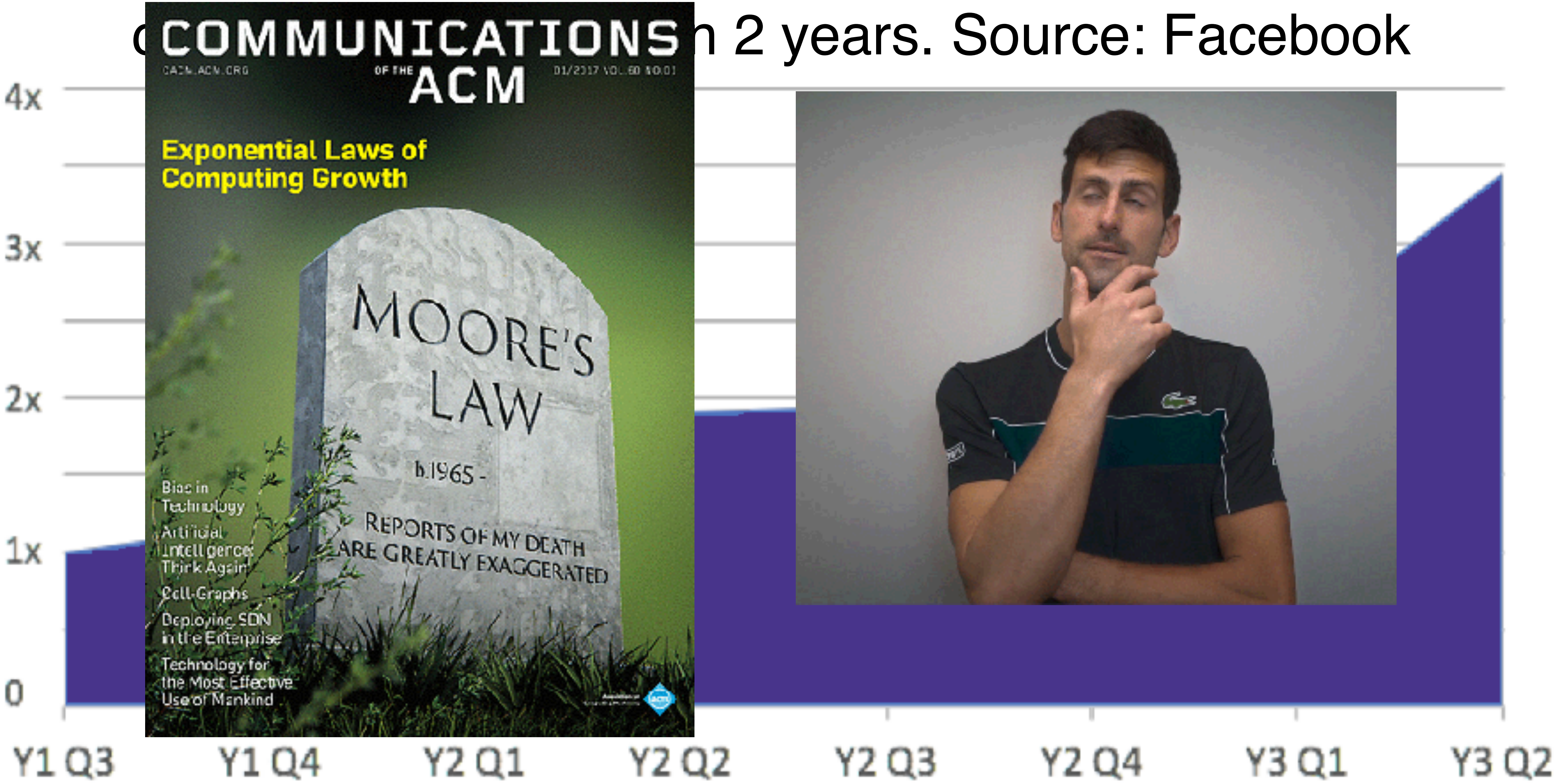
Deep Learning constitutes an important workload in cloud today.

Major cloud providers all have an ecosystem for cloud learning.

Server demand for DL inference across data centers nearly quadrupled in less than 2 years. Source: Facebook



Server demand for DL inference across data centers nearly c h 2 years. Source: Facebook



sir-g8389r8m	instance	g3.8xlarge	closed	i-084d42d56d09...	instance-terminated-no-c...	one-time	27 minutes ago	$2.28
sir-ffbga5fp	instance	g3.8xlarge	closed	i-0f040f9bd6a78...	instance-terminated-no-c...	one-time	27 minutes ago	$2.28
sir-7rerb21m	instance	g3.8xlarge	closed	i-0d844616a2e...	instance-terminated-capa...	one-time	an hour ago	$2.28
sir-pwvi36bq	instance	g3.8xlarge	closed	i-09dd933e7856...	instance-terminated-no-c...	one-time	an hour ago	$2.28
sir-rwqradim	instance	g3.8xlarge	closed	i-0f149c2c1585	instance-terminated-no-c	one-time	an hour ago	$2.28
sir-r1bi81mq	instance	g3.8xlarge	closed	i-087784d7ad8b...	instance-terminated-capa...	one-time	an hour ago	$2.28
sir-33xi8qxm	instance	g3.8xlarge	closed	i-0f50b38522a0...	instance-terminated-capa...	one-time	an hour ago	$2.28
sir-vfdradsn	instance	g3.8xlarge	closed	i-009d0d289b86	instance-terminated-capa	one-time	an hour ago	$2.28
sir-94er9wmn	instance	g3.8xlarge	closed	i-0208ba570400...	instance-terminated-capa...	one-time	an hour ago	$2.28
sir-bmc8brsq	instance	g3.8xlarge	closed	i-0c16a6cb261b...	instance-terminated-capa...	one-time	an hour ago	$2.28
sir-k3rgamdn	instance	g3.8xlarge	closed	i-0e565c2ccc68	instance-terminated-no-c	one-time	an hour ago	$2.28
sir-fgnl9wan	instance	g3.8xlarge	closed	i-07090471bd76...	instance-terminated-no-c...	one-time	an hour ago	$2.28
sir-wzvra82n	instance	g3.8xlarge	closed	i-0c57cac2c5c8...	instance-terminated-capa...	one-time	an hour ago	$2.28
sir-ex789i4p	instance	g3.8xlarge	closed	i-03b1e92f296d	instance-terminated-capa	one-time	an hour ago	$2.28
sir-vvhran2q	instance	g3.8xlarge	closed	i-087663582f2f2...	instance-terminated-capa...	one-time	an hour ago	$2.28
sir-irkr87nq	instance	g3.8xlarge	closed	i-0d29ad96ca42...	instance-terminated-no-c...	one-time	an hour ago	$2.28
sir-w2eg88pm	instance	g3.8xlarge	closed	i-079a78d77ea0	instance-terminated-no-c	one-time	an hour ago	$2.28
sir-mphgb1rq	instance	g3.8xlarge	closed	i-0d12485d32f3...	instance-terminated-capa...	one-time	an hour ago	$2.28

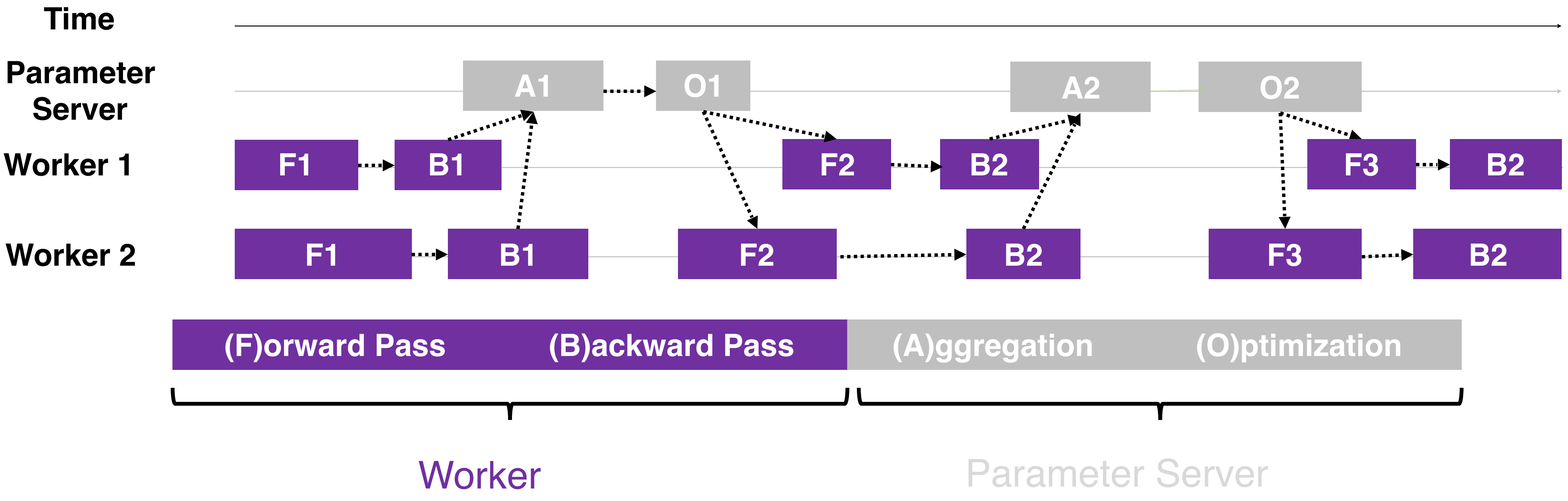
EC2 reclaims your GPU instances as they run out of capacity

sir-g8389r8m	instance	g3.8xlarge	closed	i-084d42d56d09...	instance-terminated-no-c...	one-time	27 minutes ago	$2.28
sir-ffbga5fp	instance	g3.8xlarge	closed	i-0f040f9bd6a78...	instance-terminated-no-c...	one-time	27 minutes ago	$2.28
sir-7rerb21m	instance	g3.8xlarge	closed	i-0d8444616a2e...	instance-terminated-capa...	one-time	an hour ago	$2.28
sir-pwvi36bq	instance						an hour ago	$2.28
sir-rwqradim	instance						an hour ago	$2.28
sir-r1bi81mq	instance						an hour ago	$2.28
sir-33xi8qxm	instance						an hour ago	$2.28
sir-vfdradsn	instance						an hour ago	$2.28
sir-94er9wmn	instance						an hour ago	$2.28
sir-bmc8brsq	instance						an hour ago	$2.28
sir-k3rgamdn	instance						an hour ago	$2.28
sir-fgnl9wan	instance						an hour ago	$2.28
sir-wzvra82n	instance						an hour ago	$2.28
sir-ex789i4p	instance						an hour ago	$2.28
sir-wvhran2q	instance						an hour ago	$2.28
sir-irkr87nq	instance						an hour ago	$2.28
sir-w2eg88pm	instance	g3.8xlarge	closed	i-079a78d77ea0	instance-terminated-no-c	one-time	an hour ago	$2.28
sir-mphgb1rq	instance	g3.8xlarge	closed	i-0d12485d32f3...	instance-terminated-capa...	one-time	an hour ago	$2.28

EC2 reclaims your GPU instances as they run out of capacity

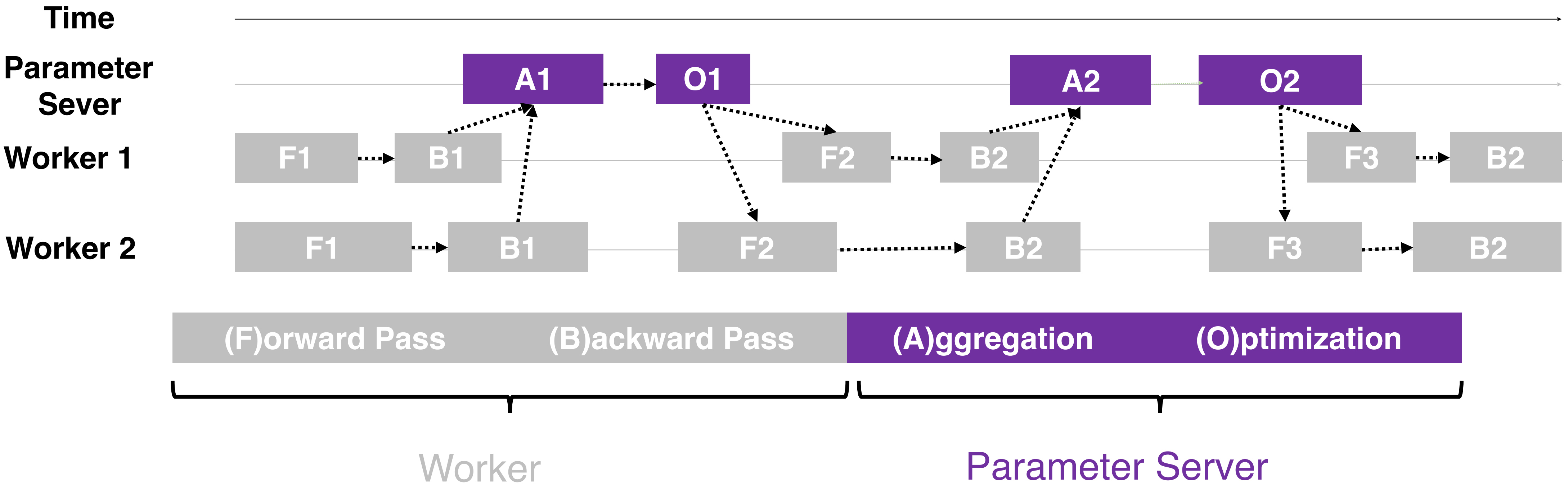
Distributed Training

INDEPENDENT FORWARD/BACKWARD PASSES + COORDINATED PARAMETER EXCHANGE



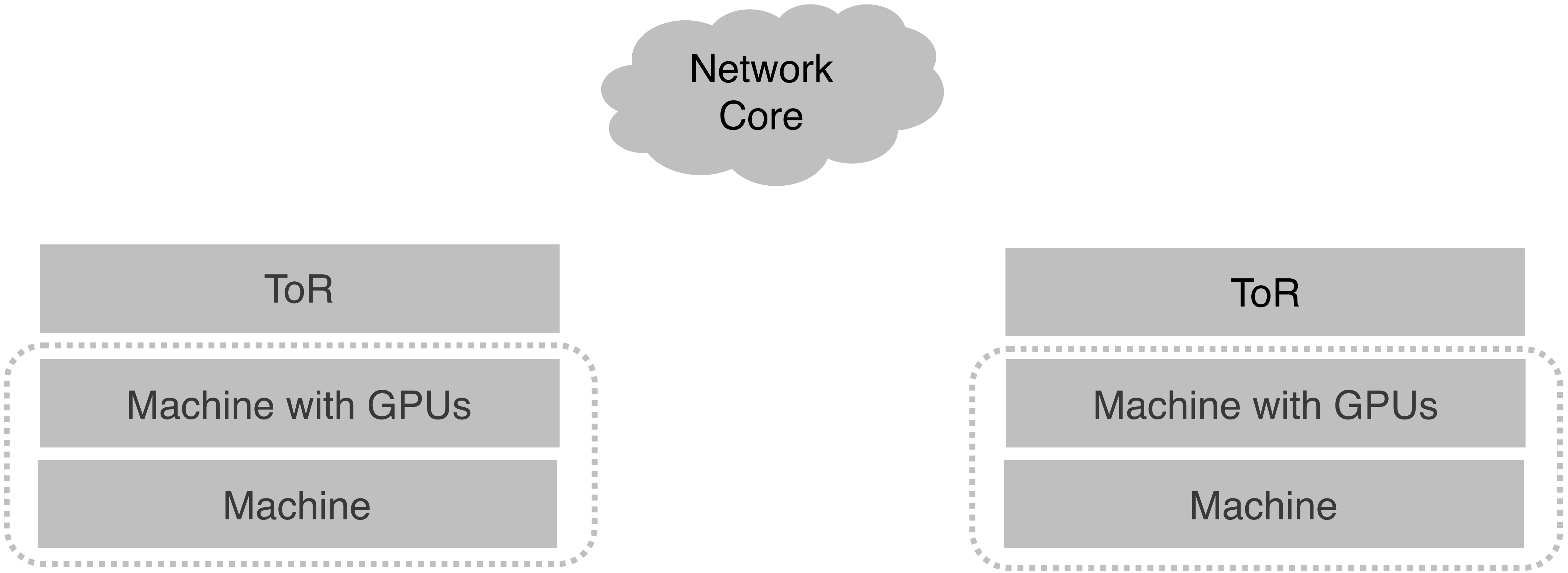
Distributed Training

INDEPENDENT FORWARD/BACKWARD PASSES + COORDINATED PARAMETER EXCHANGE



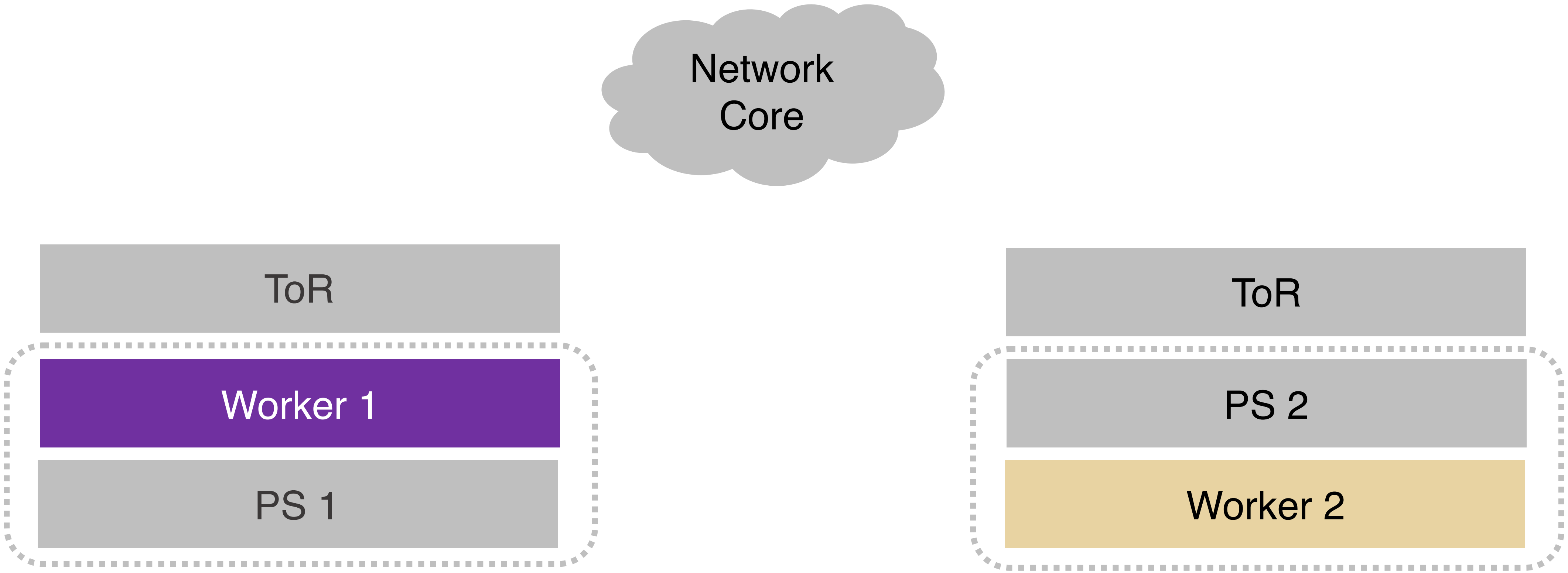
Distributed Training Today

IN THE CONTEXT OF THE CLOUD



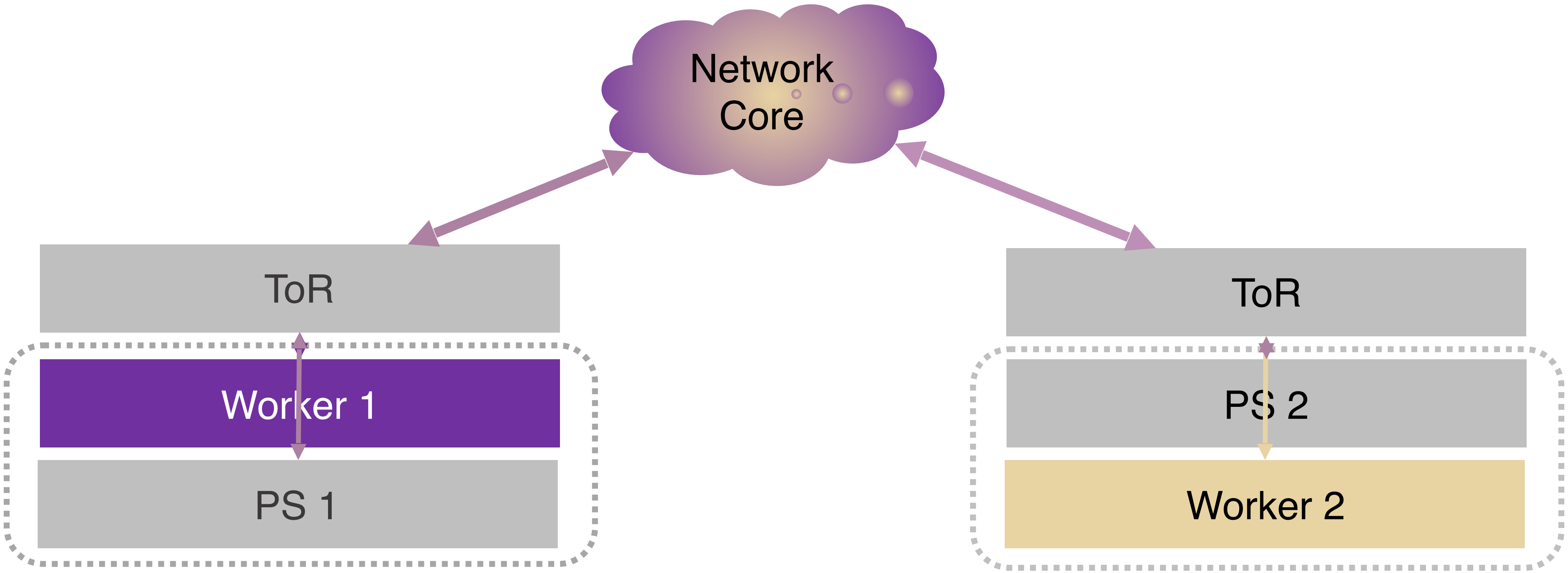
Distributed Training Today

FORWARD AND BACKWARD PASSES IN WORKER



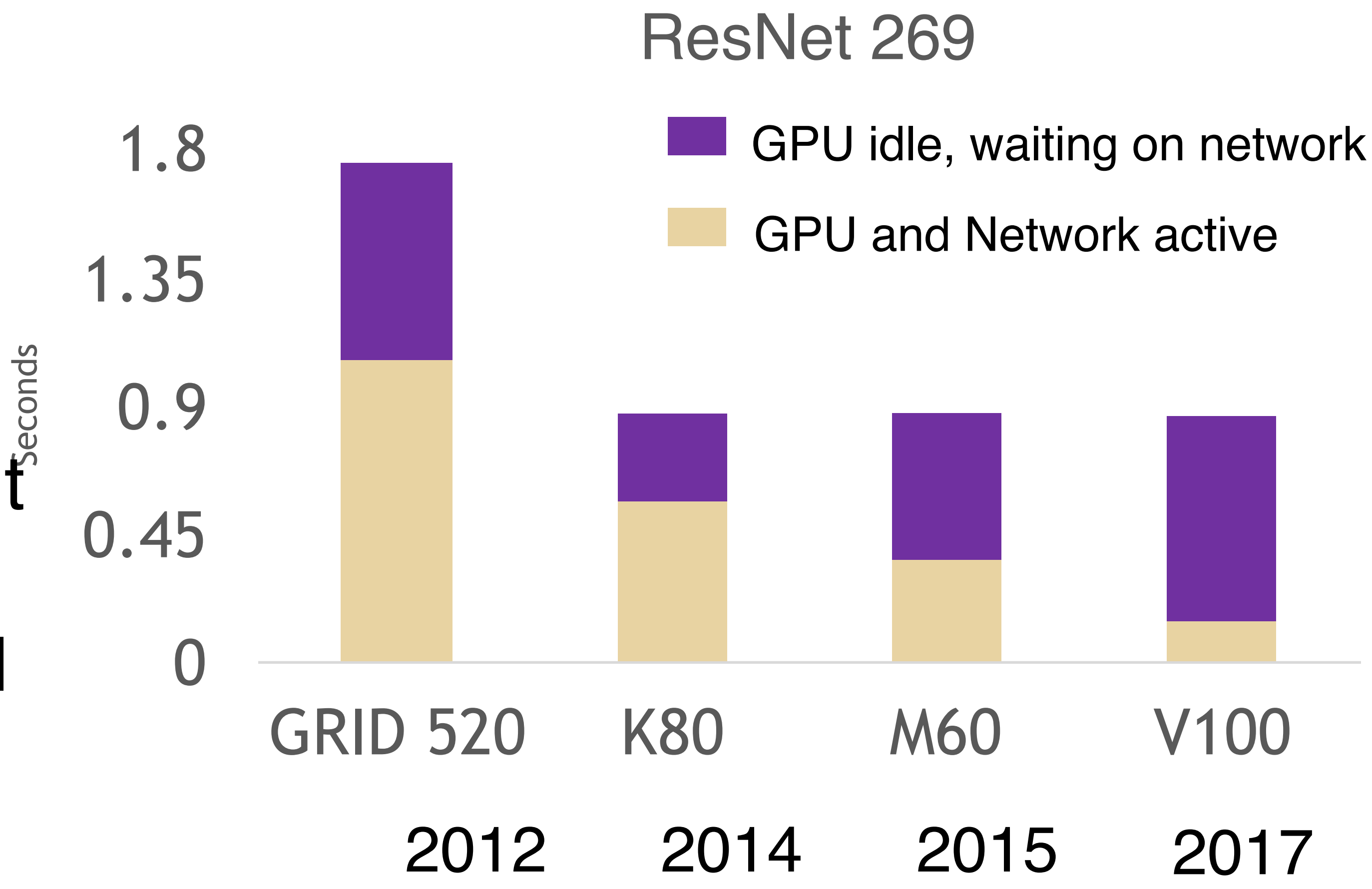
Distributed Training Today

AGGREGATION AND OPTIMIZATION IN PS

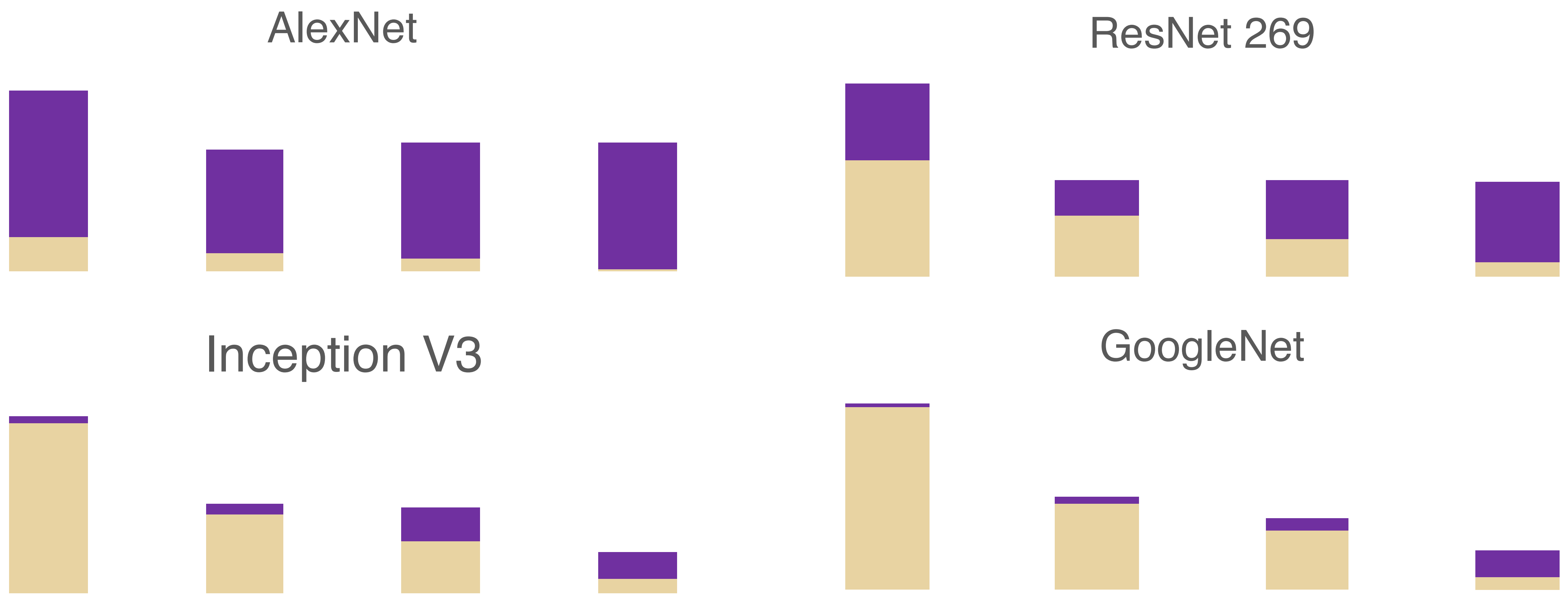


Distributed training is communication bound

- 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.

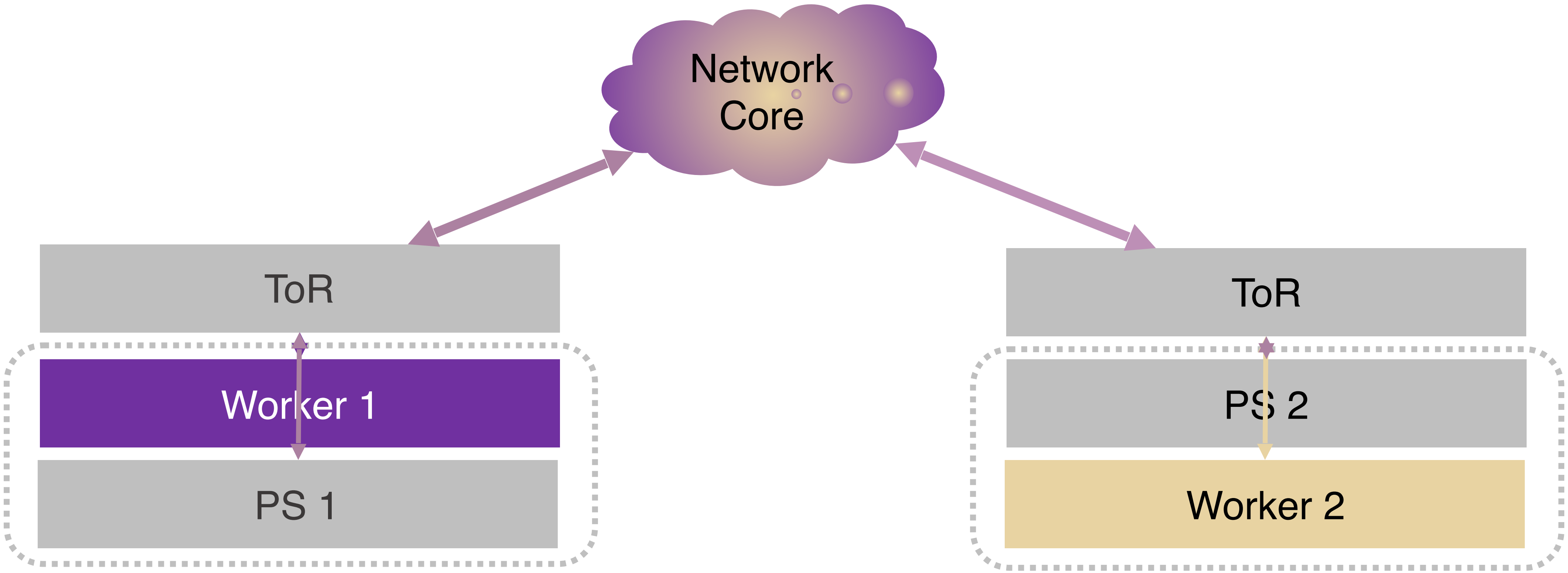


Distributed training is communication bound



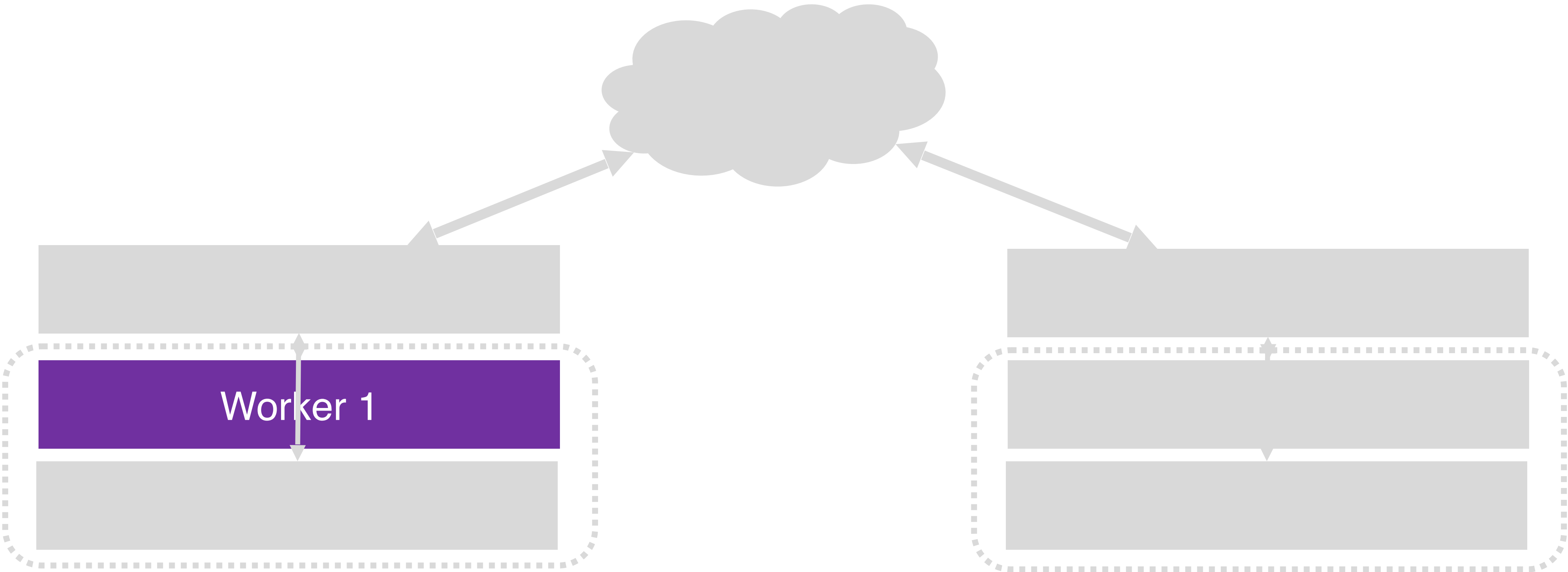
Bottlenecks in DDNN training

MAPPING OF TRAINING WORKLOAD TO THE CLOUD IS INEFFICIENT.



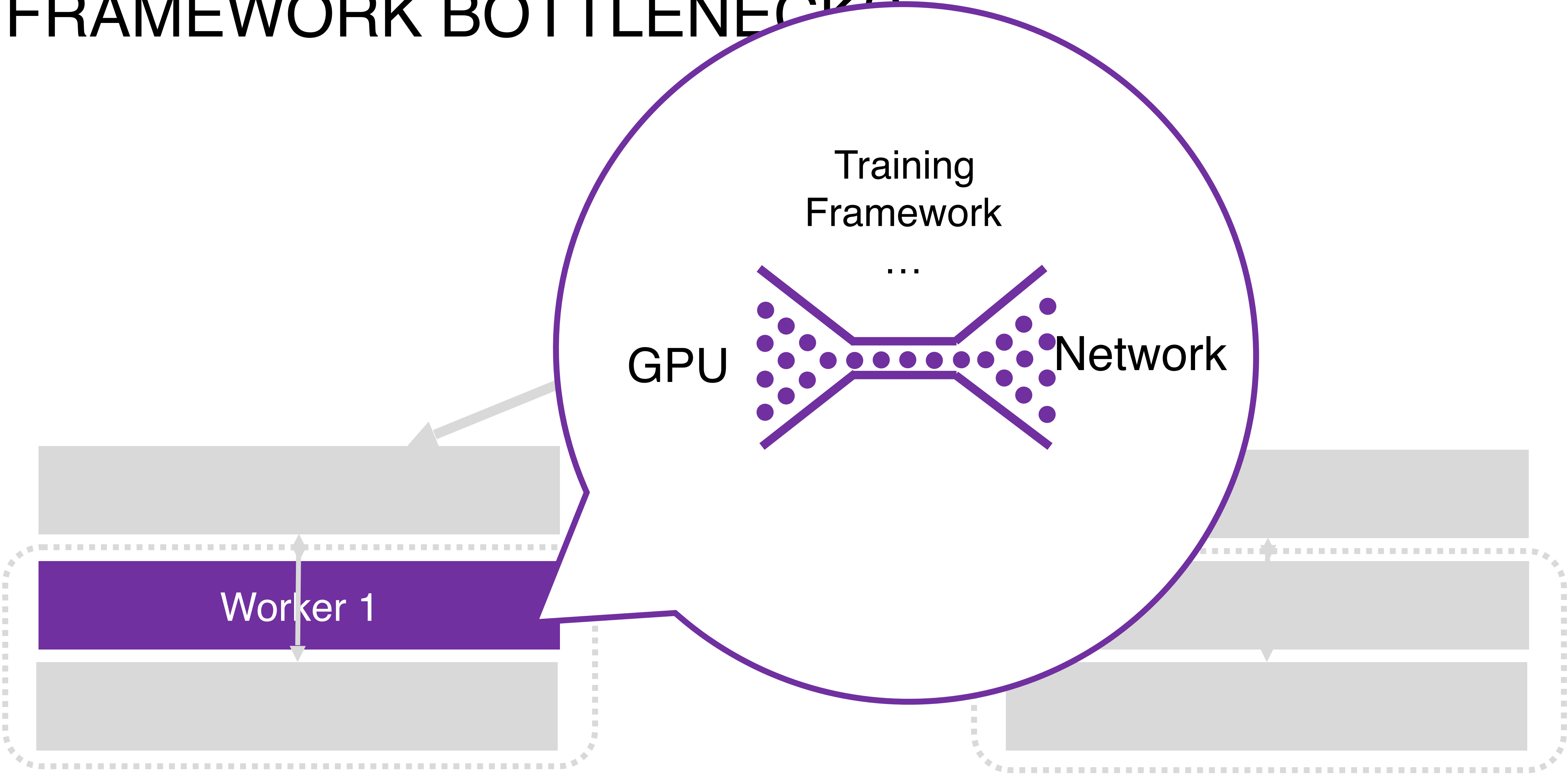
Bottlenecks in DDNN training

FRAMEWORK BOTTLENECKS



Bottlenecks in DDNN training

FRAMEWORK BOTTLENECKS

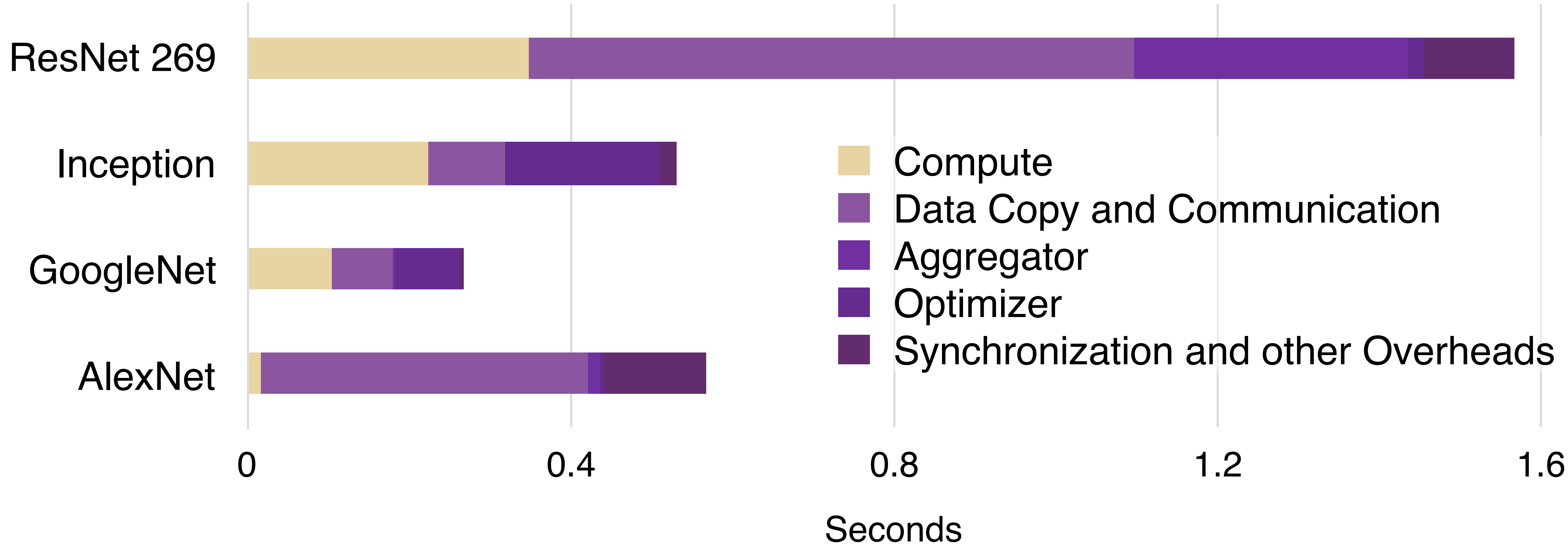


Bottlenecks in DDNN training

FRAMEWORK BOTTLENECKS

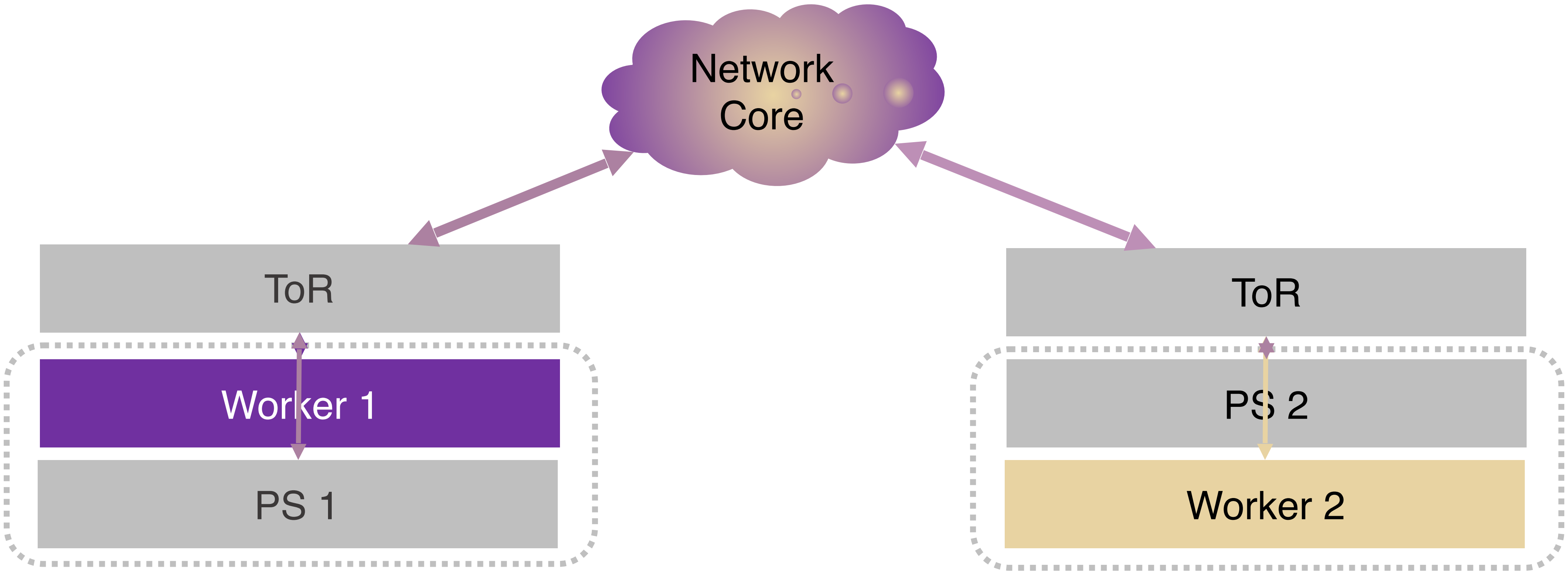
Bottlenecks in DDNN training

FRAMEWORK BOTTLENECKS



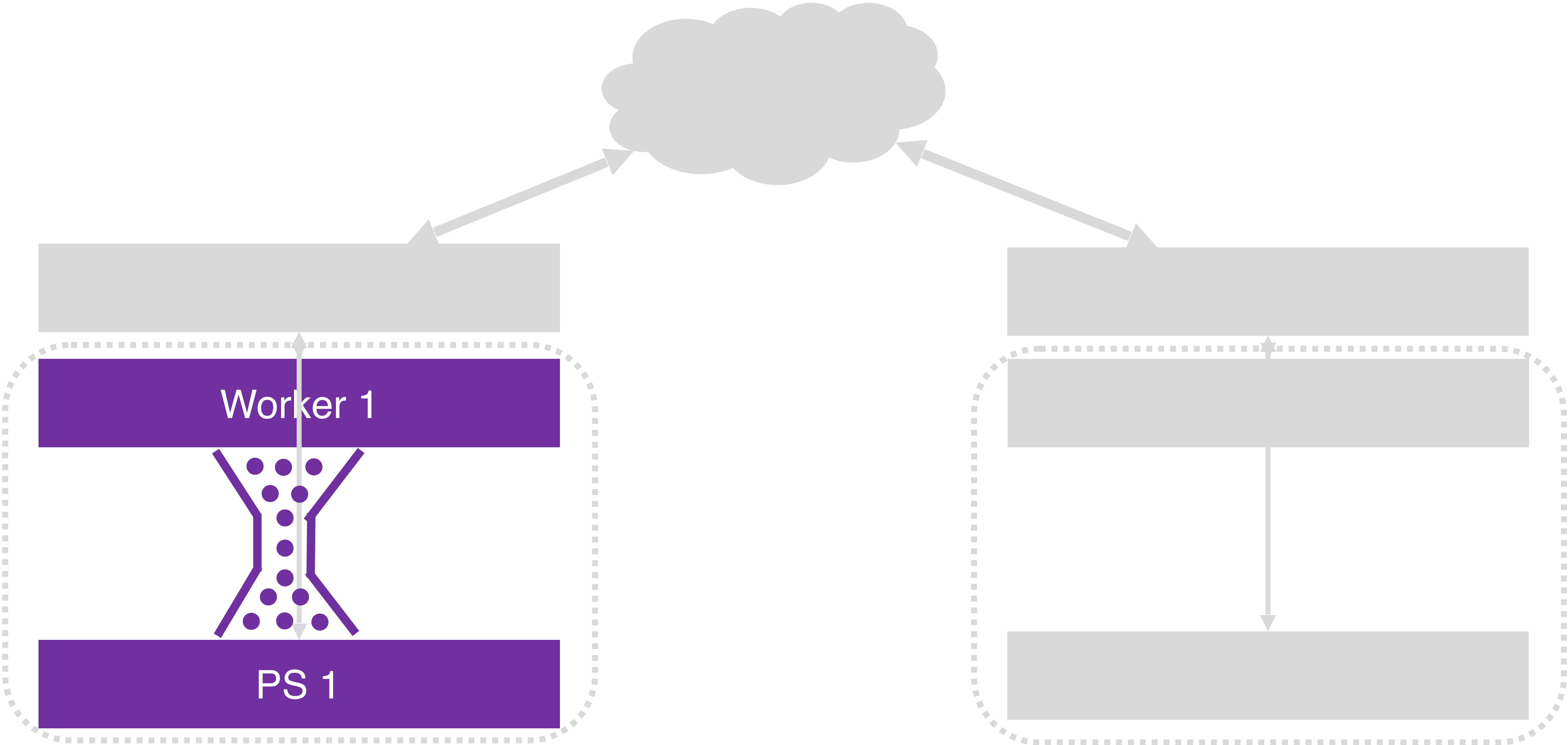
Bottlenecks in DDNN training

MAPPING OF TRAINING WORKLOAD TO THE CLOUD IS INEFFICIENT.



Bottlenecks in 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?

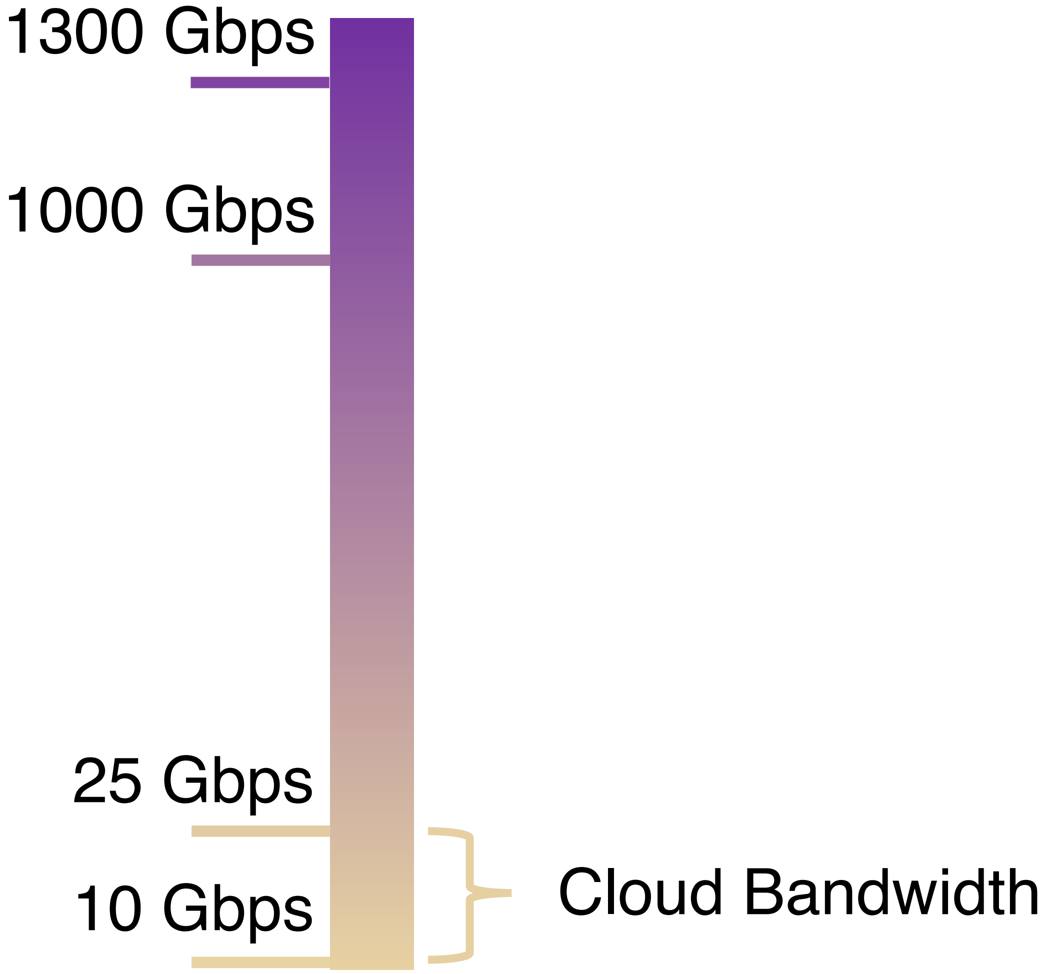
8 workers, GTX 1080 Ti, central parameter servers. MxNet

Bottlenecks in Cloud-based DDNN training

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

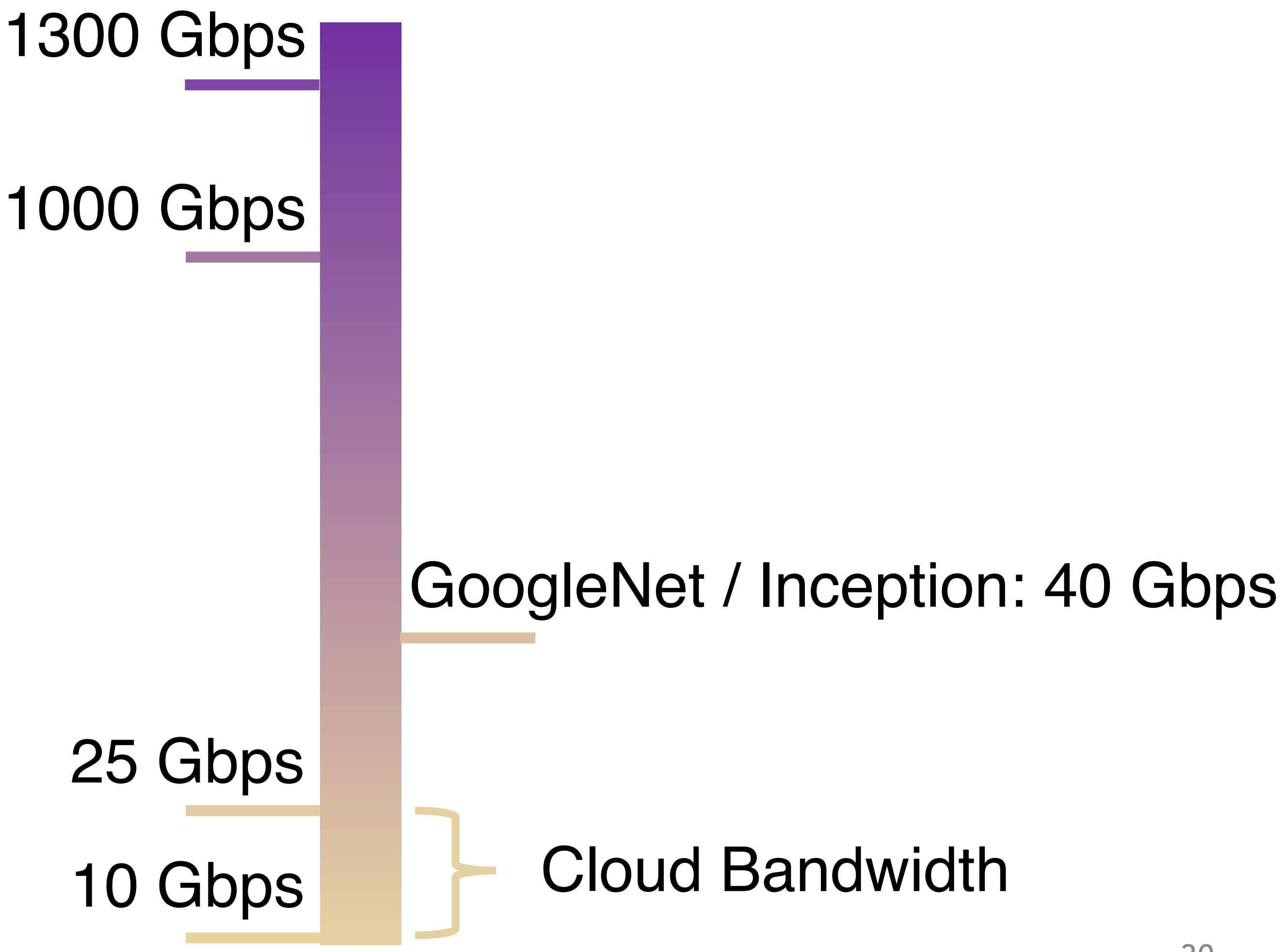


Bottlenecks in Cloud-based DDNN training

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

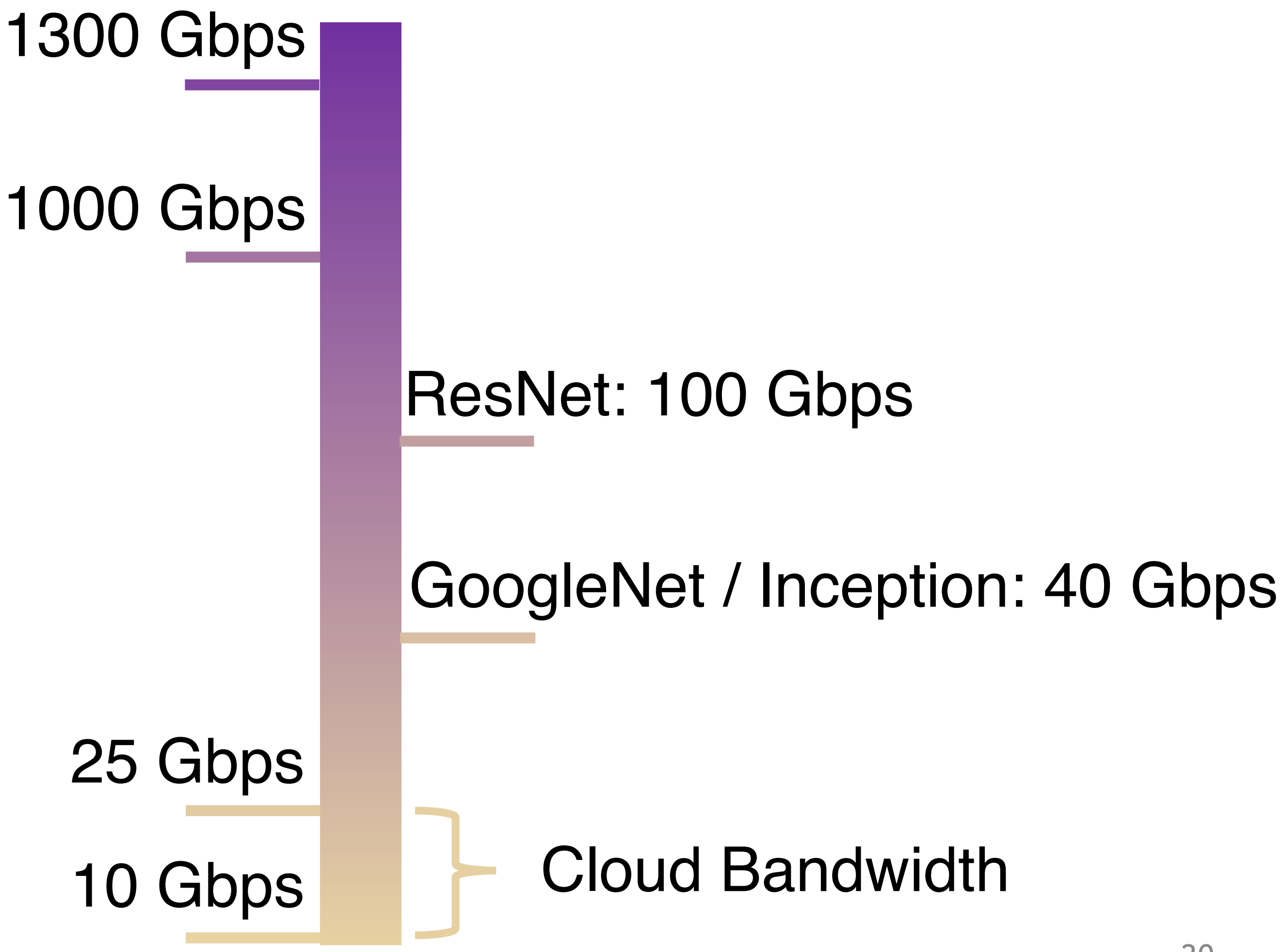


Bottlenecks in Cloud-based DDNN training

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

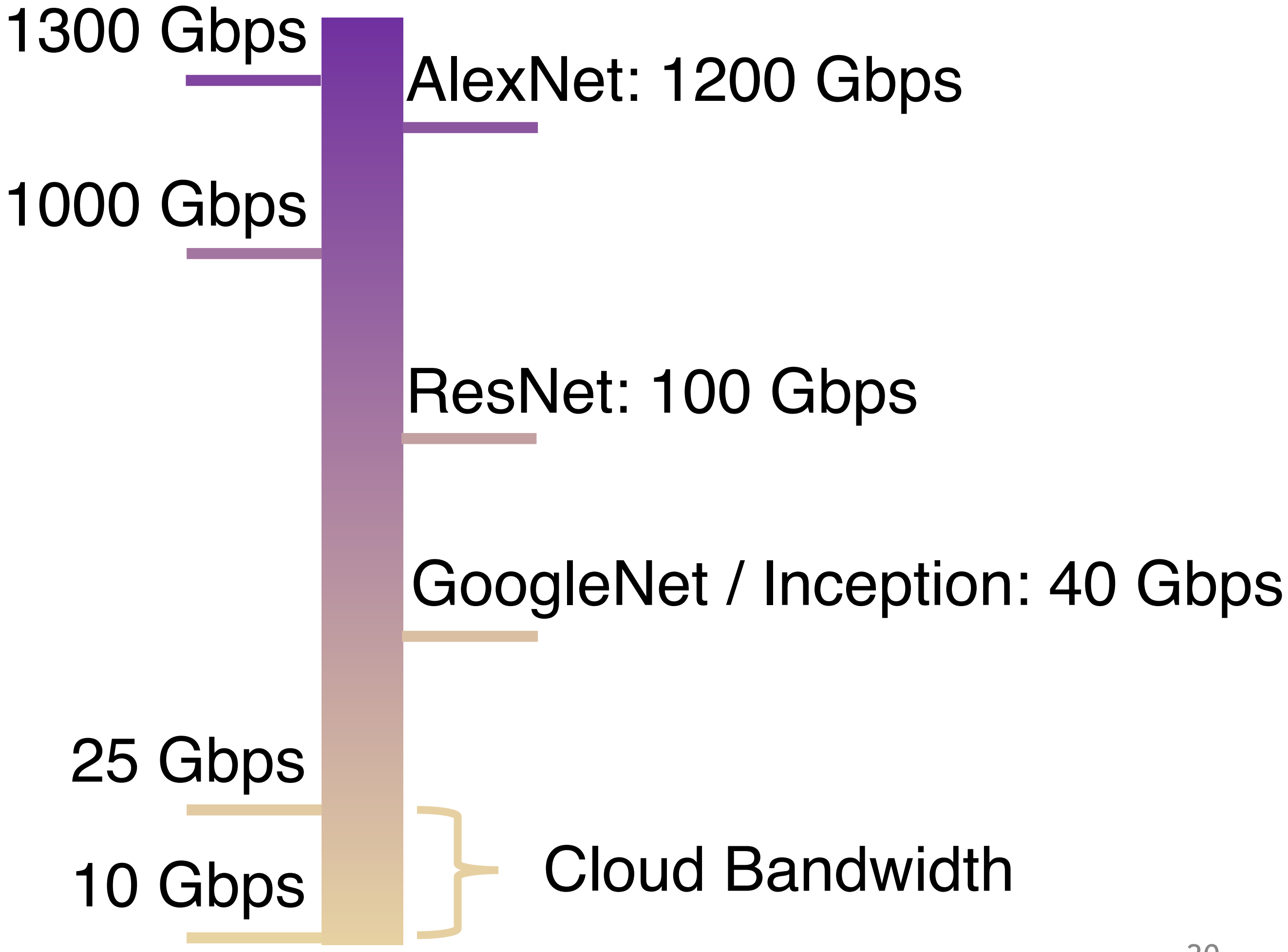


Bottlenecks in Cloud-based DDNN training

INSUFFICIENT BANDWIDTH

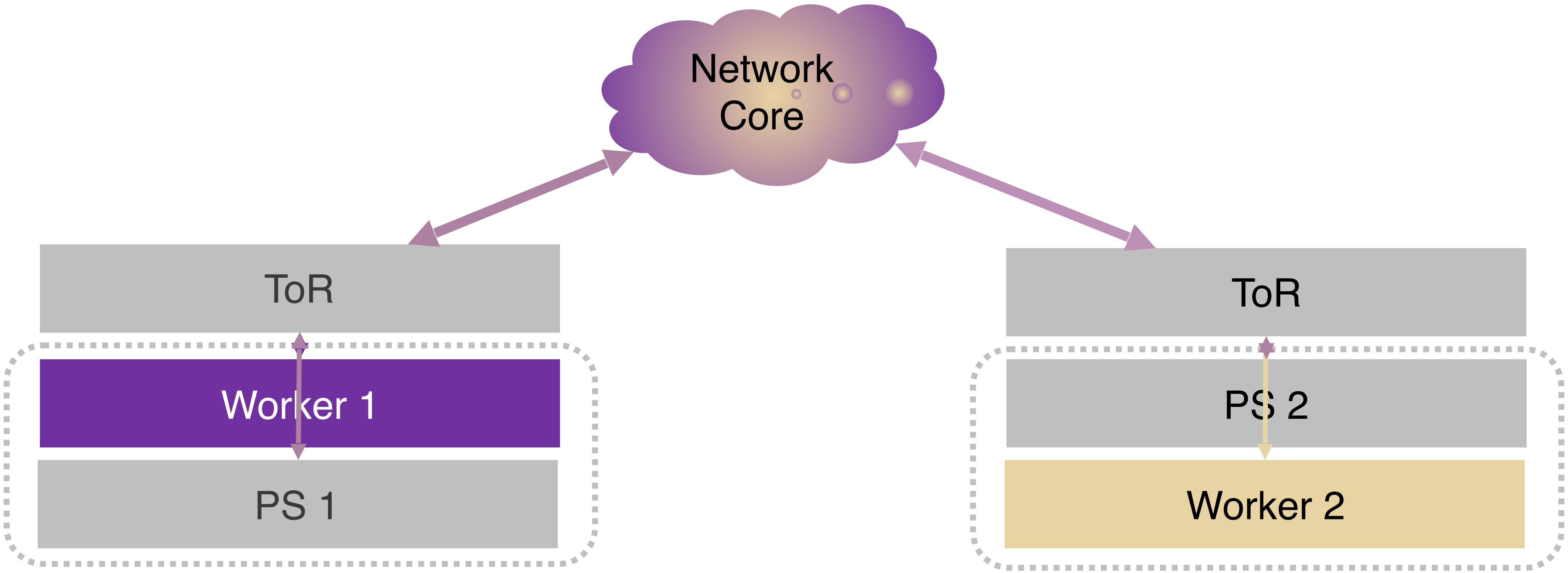
Minimum bandwidth required for each of the popular NNs for communication to not bottleneck computation?

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



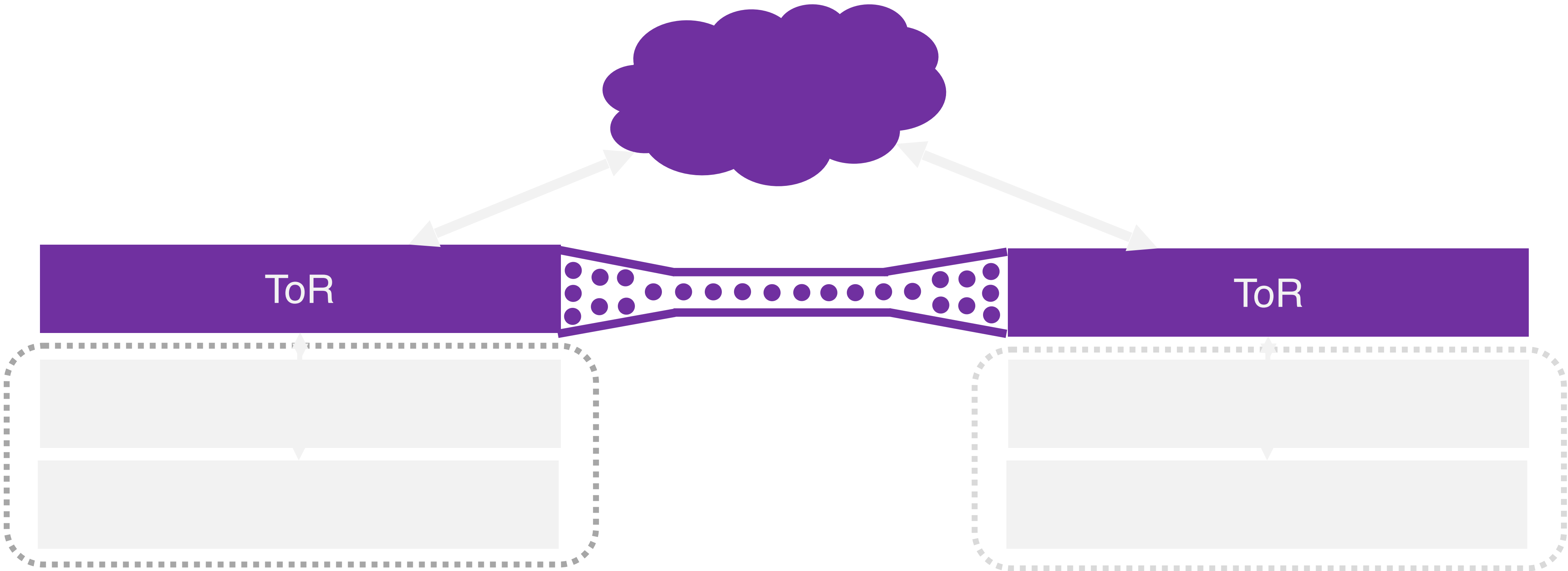
Bottlenecks in Cloud-based DDNN training

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.

Hosts

	1	2	3	4	5	6	7	8
1								
2	4.7							
3	8.9	4.7						
4	8.9	4.7	8.9					
5	8.9	4.7	8.9	8.9				
6	4.7	9.0	4.7	4.7	4.7			
7	8.	4.	9.	8.	9.	4.		

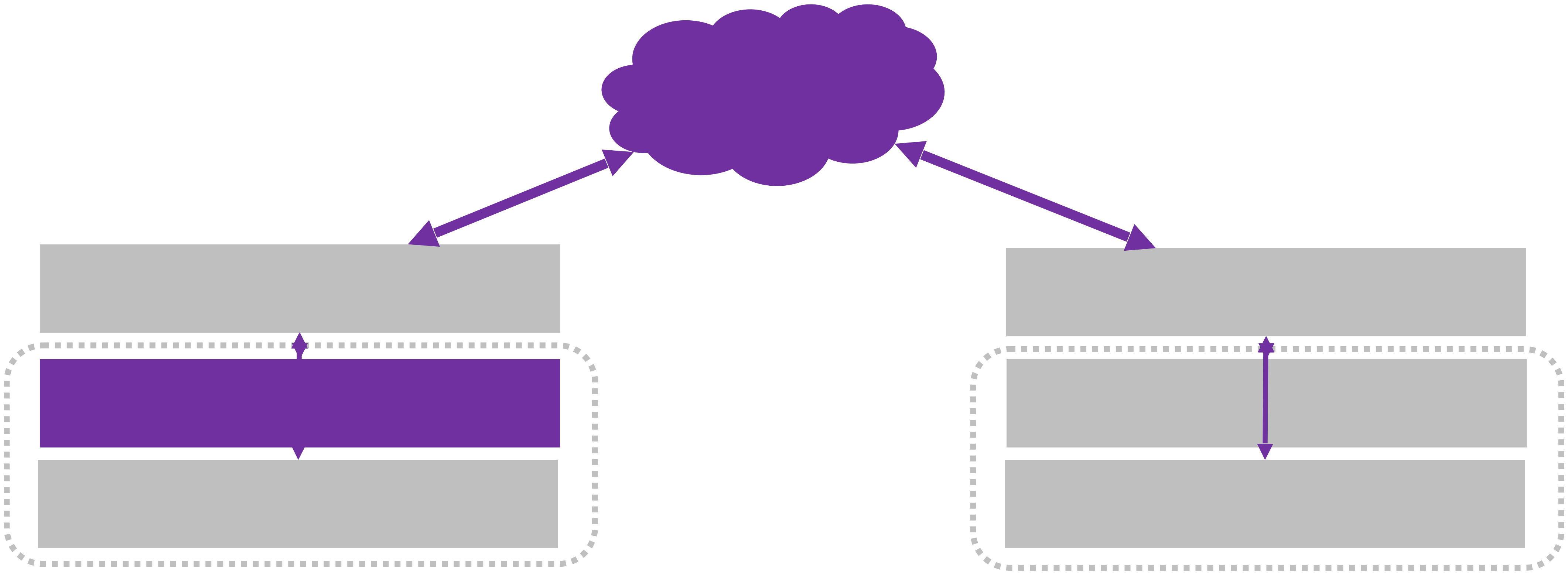
Cluster 1: 1 3 4 5 7
Cluster 2: 2 6 8

9 Gbps

4 Gbps

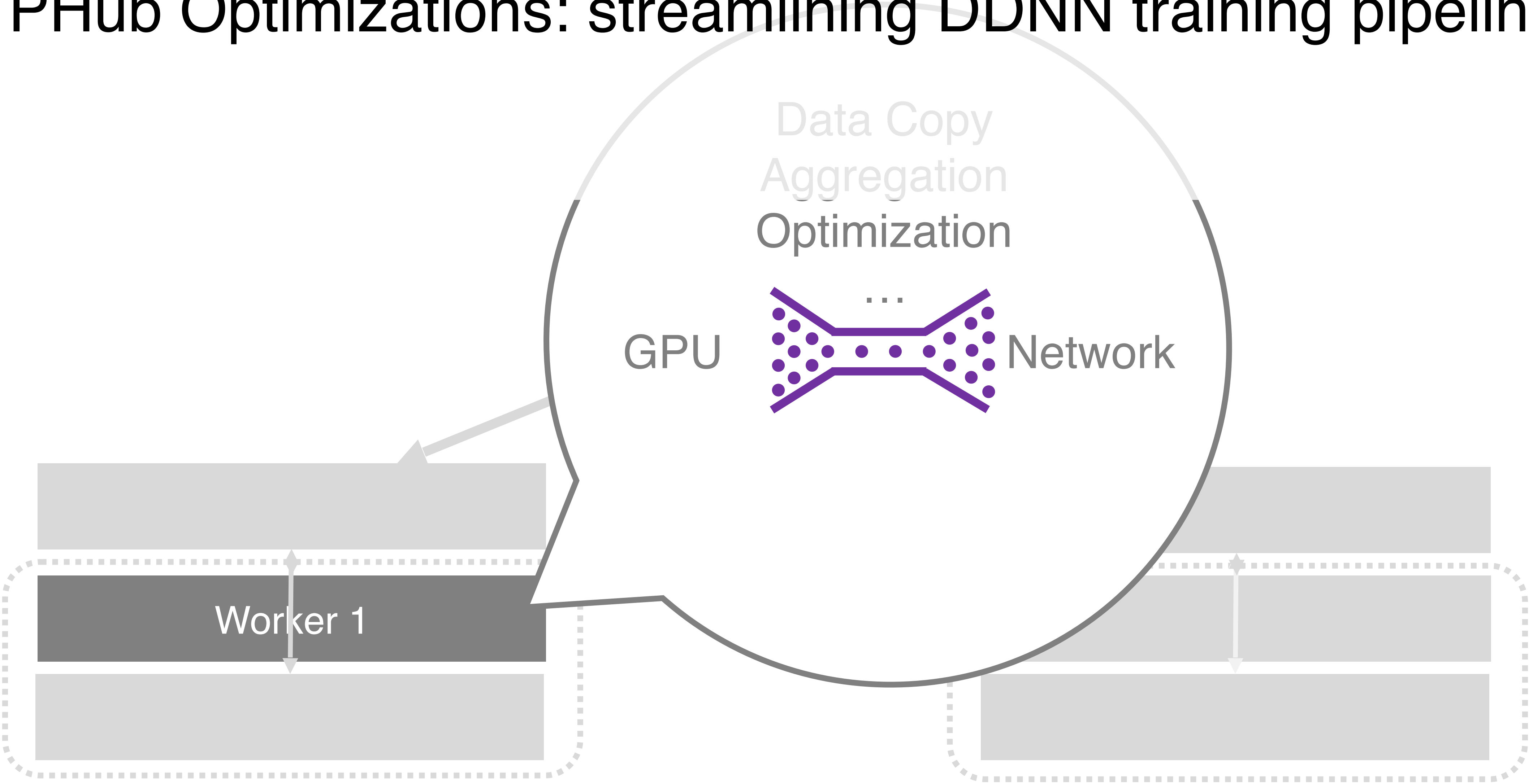
Parameter Hub Optimizations

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



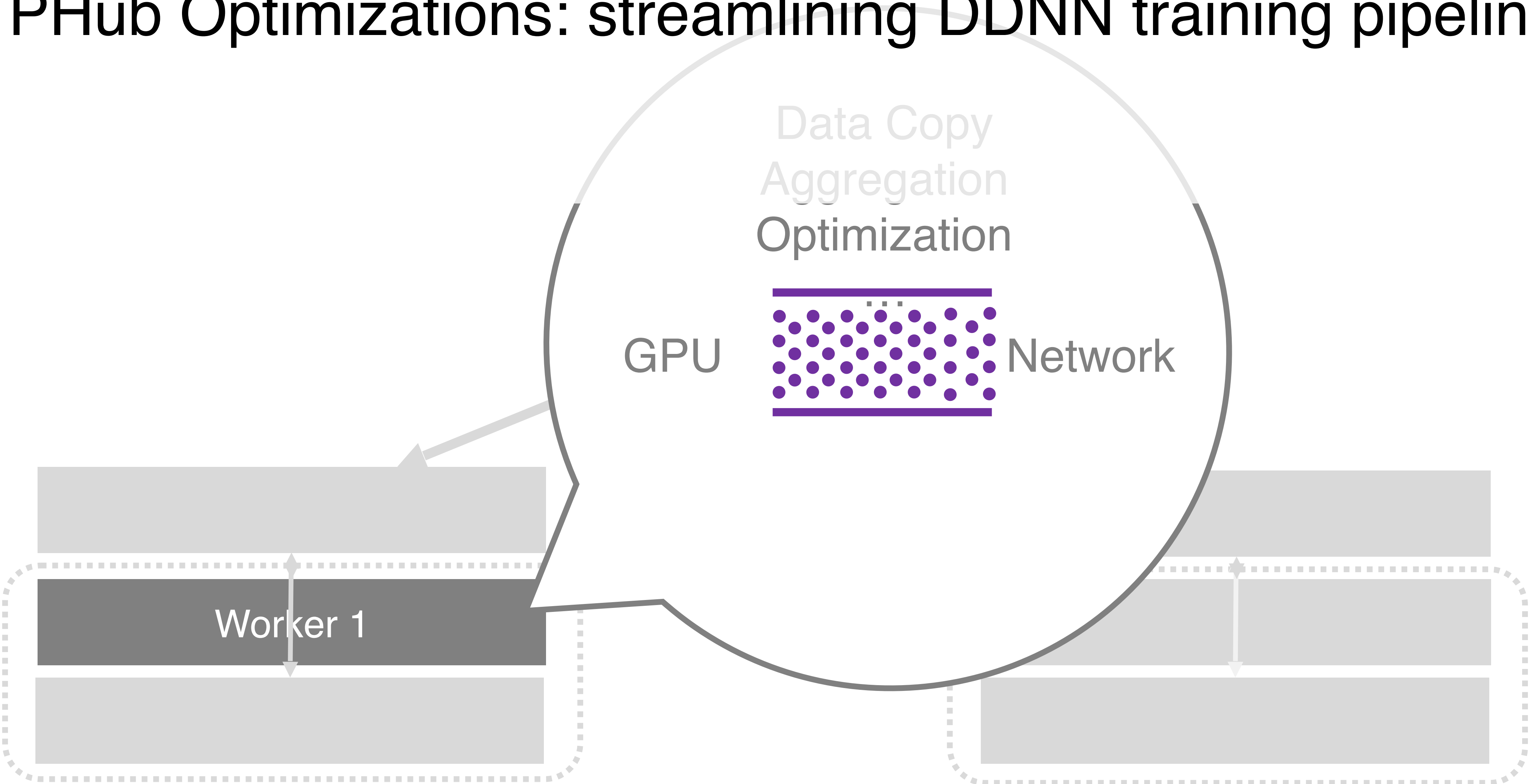
Eliminating framework bottlenecks:

PHub Optimizations: streamlining DDNN training pipeline

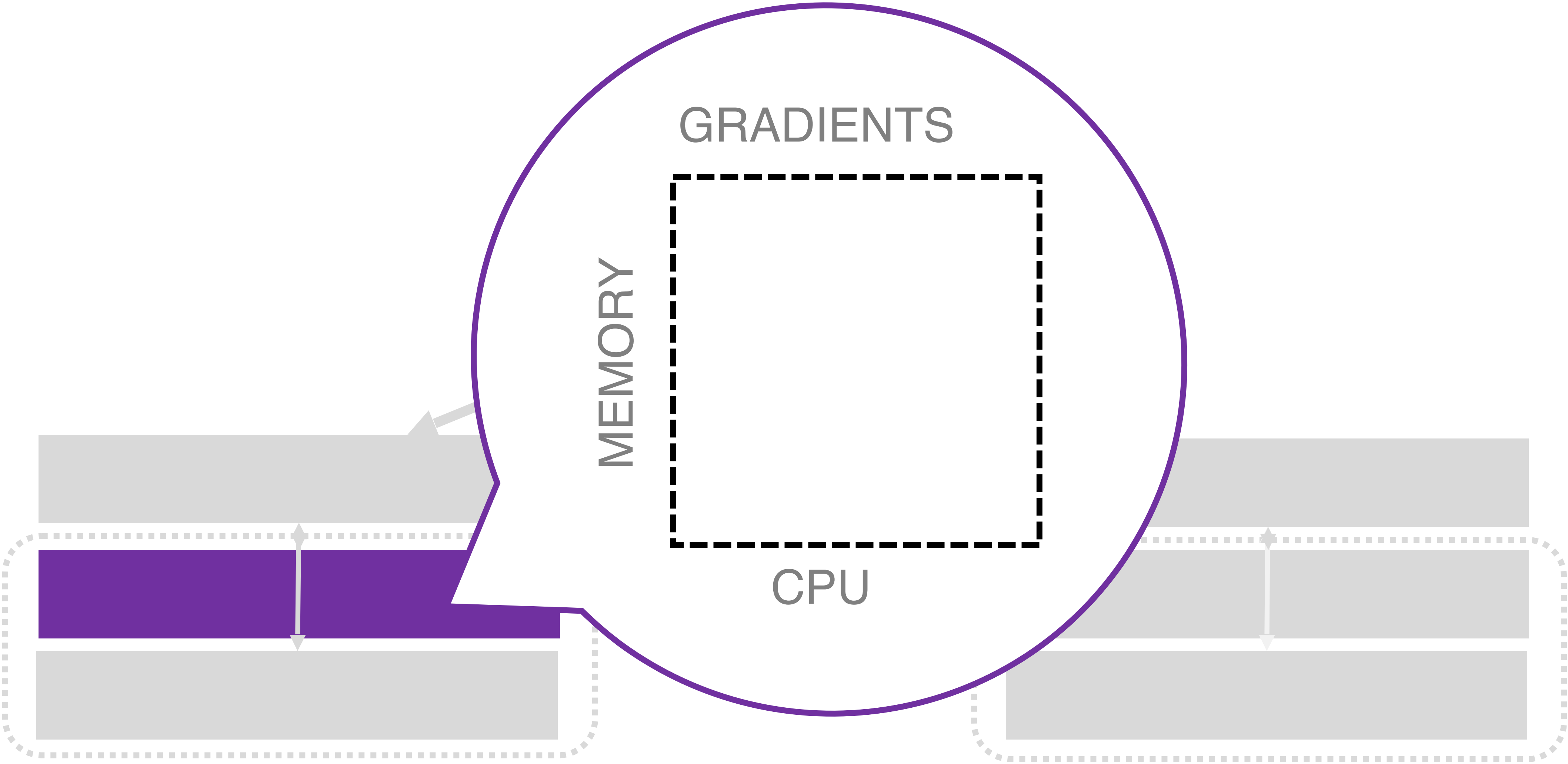


Eliminating framework bottlenecks:

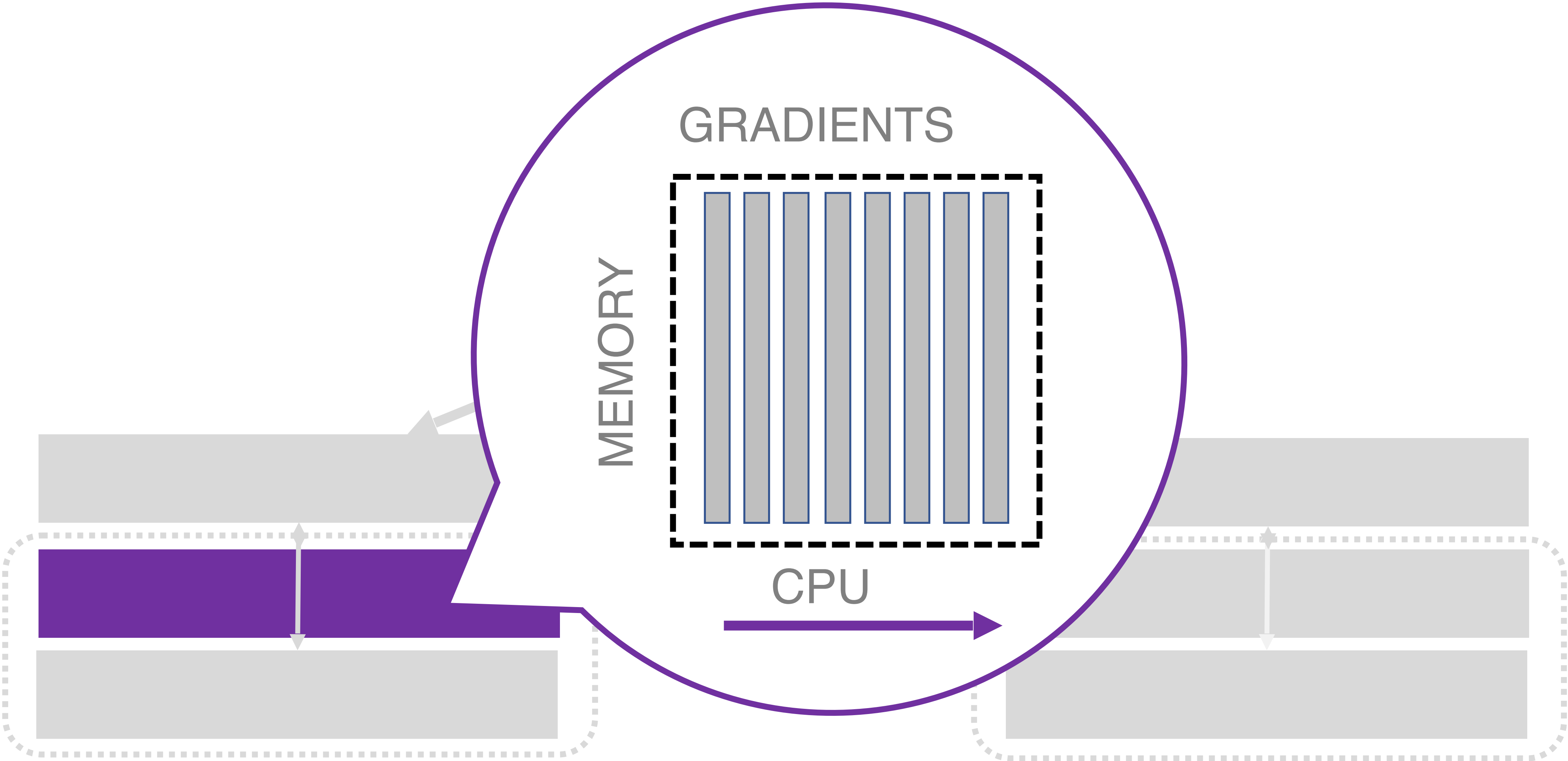
PHub Optimizations: streamlining DDNN training pipeline



Software Optimizations



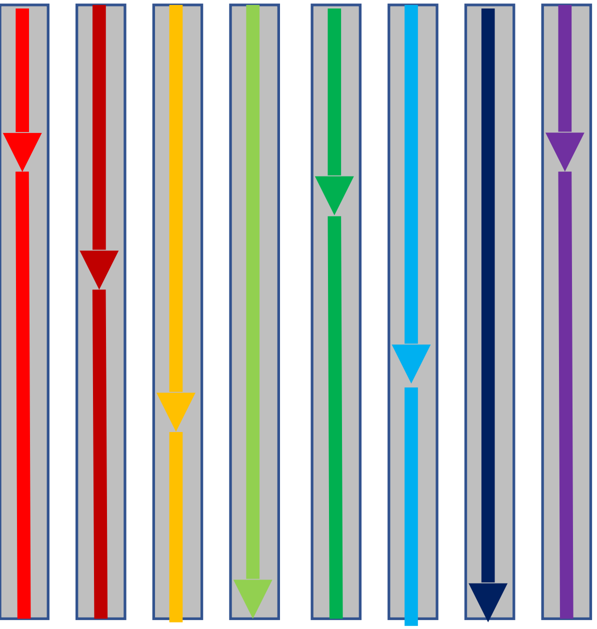
Software Optimizations



Software Optimizations

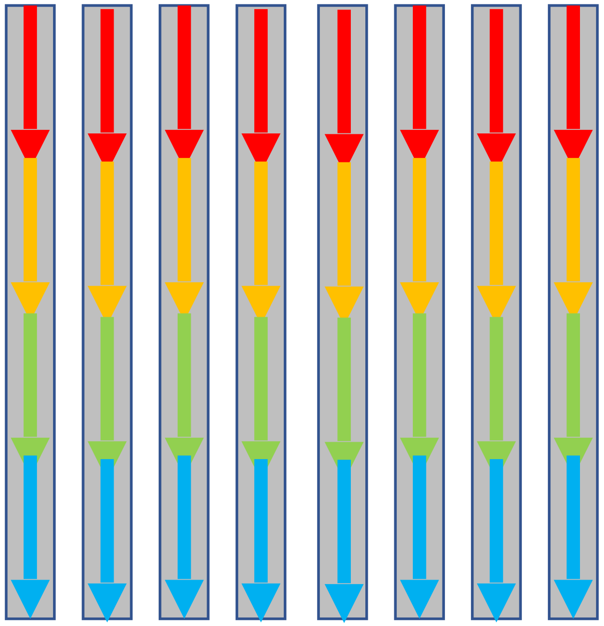
GRADIENT AGGREGATION AND OPTIMIZATION

Requires synchronization.



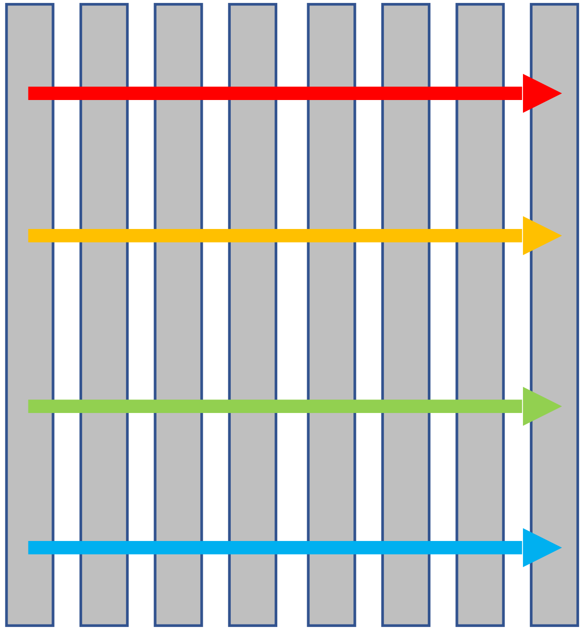
Each core reads the input Q from different workers and writes to different locations to the output queue

Great locality. No synchronization



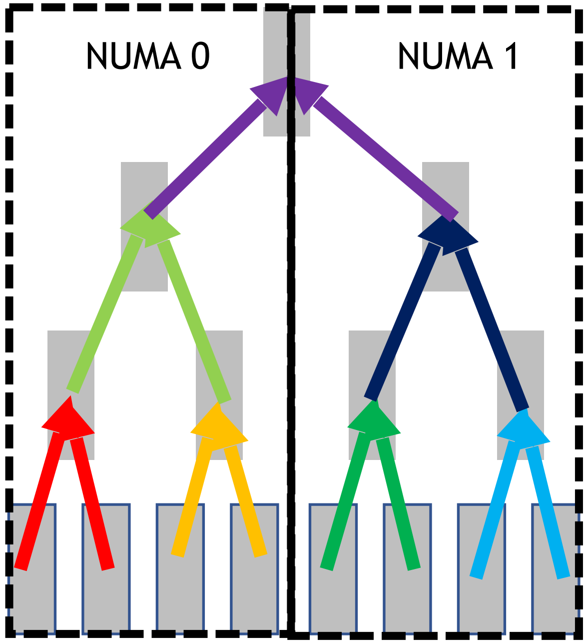
For each input Q, launch a series of threads for aggregation. This is used in MxNet. (Wide Aggregation)

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

GRADIENT AGGREGATION AND OPTIMIZATION

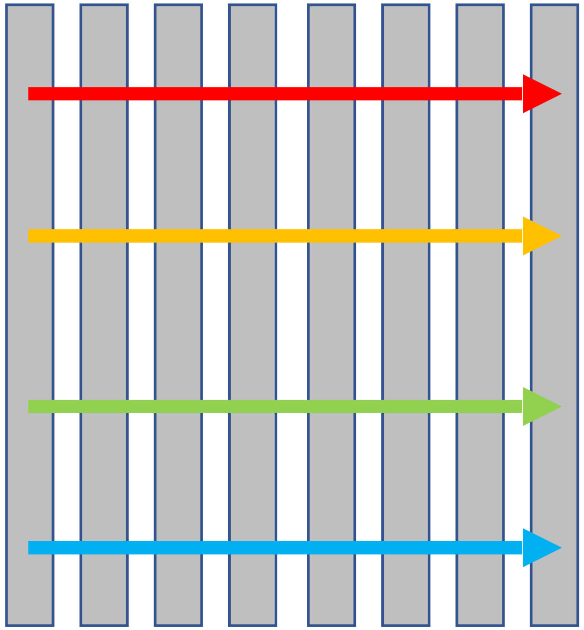
Requires synchronization.

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



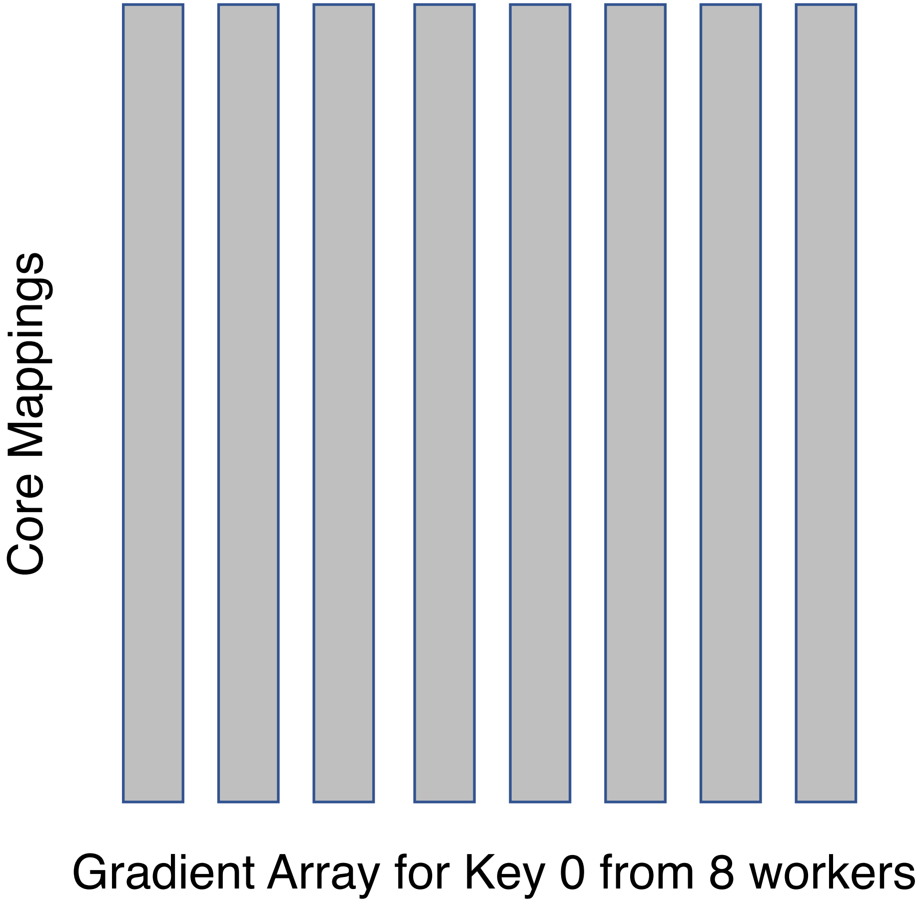
Sequentially aggregates the same portion of gradients within each queue. (Tall Aggregation)

Too much coherence and synchronization

Organize processors into hierarchy. Perform NUMA aware tree reduction.

Software Optimizations

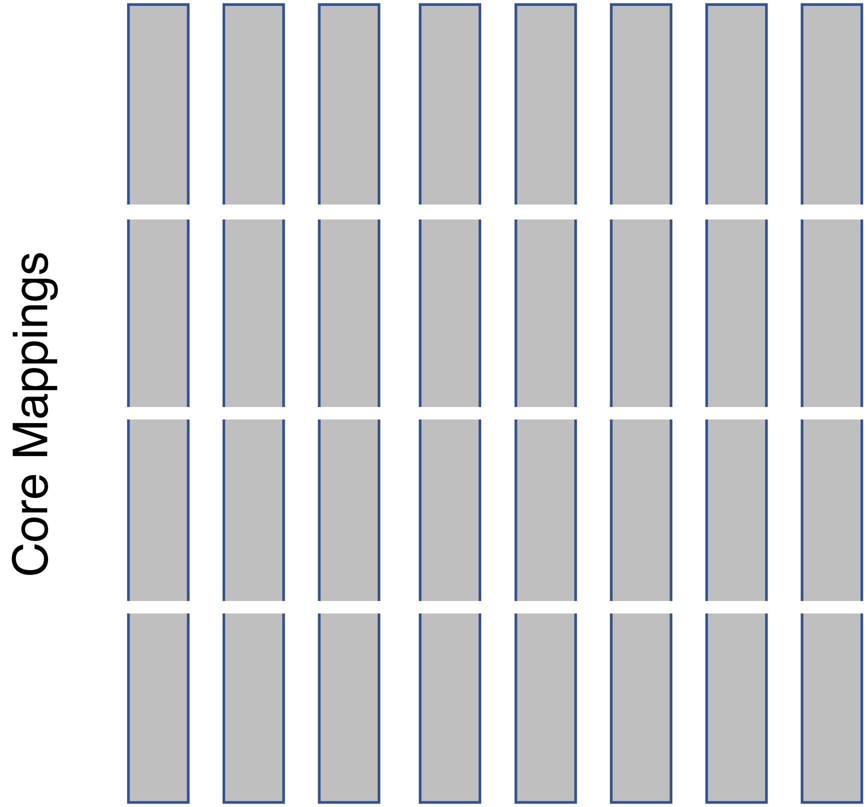
TALL AGGREGATION AND OPTIMIZATION



- Chunk a gradient into a series of virtual gradients deterministically.
- A virtual gradient is mapped to a particular core on the server.

Software Optimizations

TALL AGGREGATION AND OPTIMIZATION

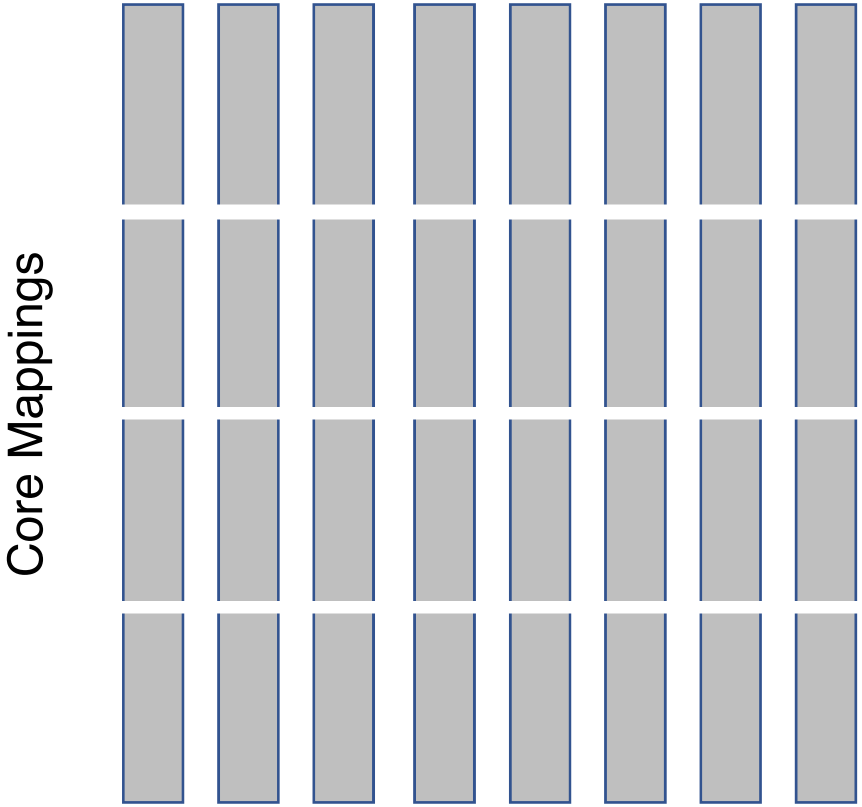


Gradient Array for Key 0 from 8 workers

- Chunk a gradient into a series of virtual gradients deterministically.
- A virtual gradient is mapped to a particular core on the server.

Software Optimizations

TALL AGGREGATION AND OPTIMIZATION

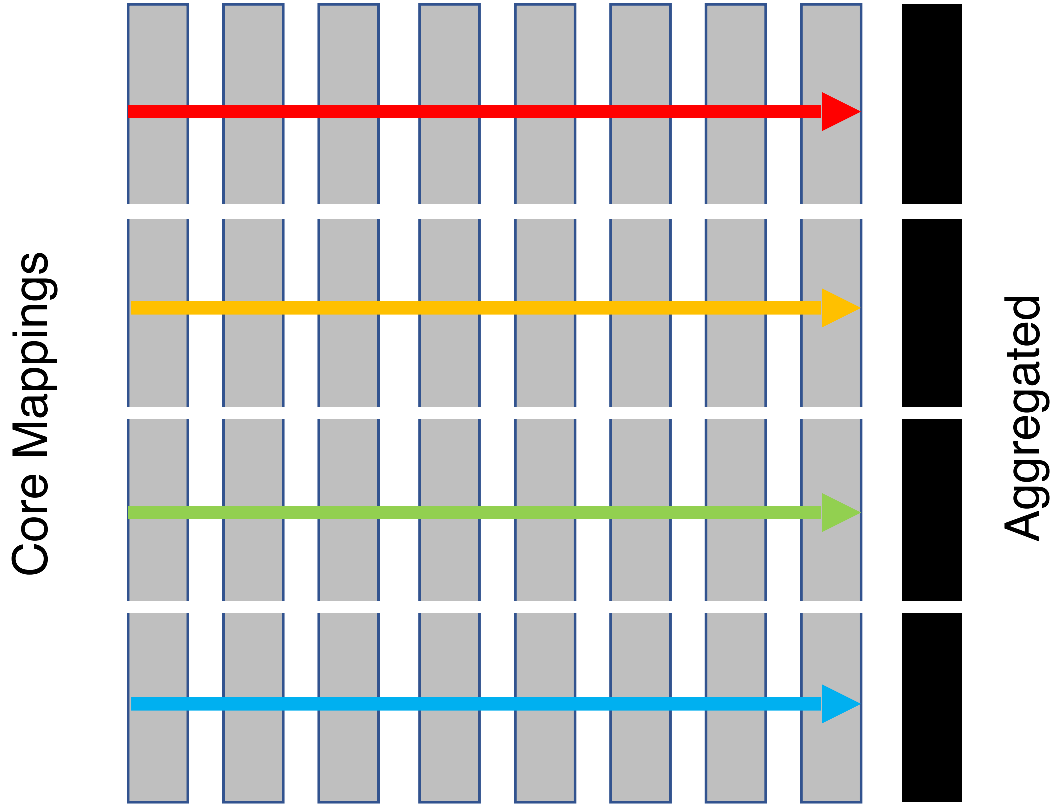


Gradient Array for Key 0 from 8 workers

- 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.

Software Optimizations

TALL AGGREGATION AND OPTIMIZATION

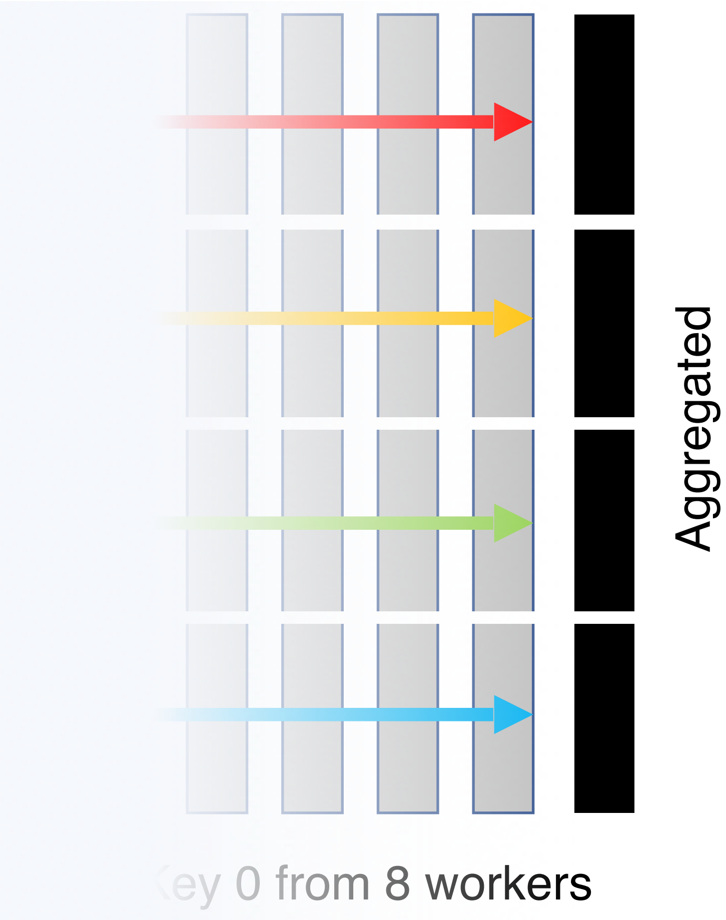


Gradient Array for Key 0 from 8 workers

- 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

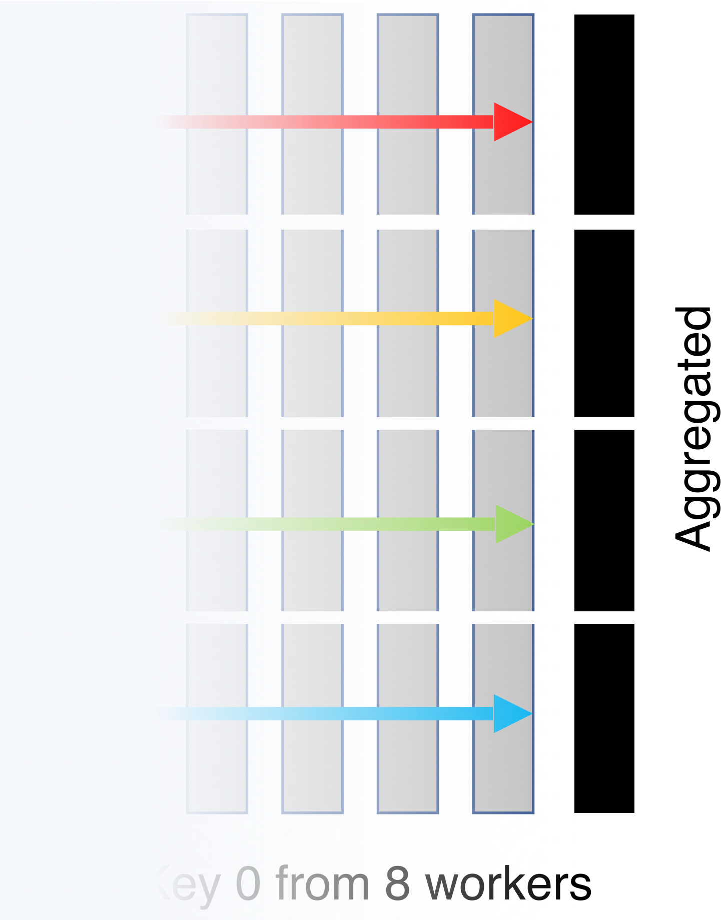


When Aggregation is done, PHub:

- 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.
- FP32-level streaming aggregation and optimization to hide communication latency.

Software Optimizations

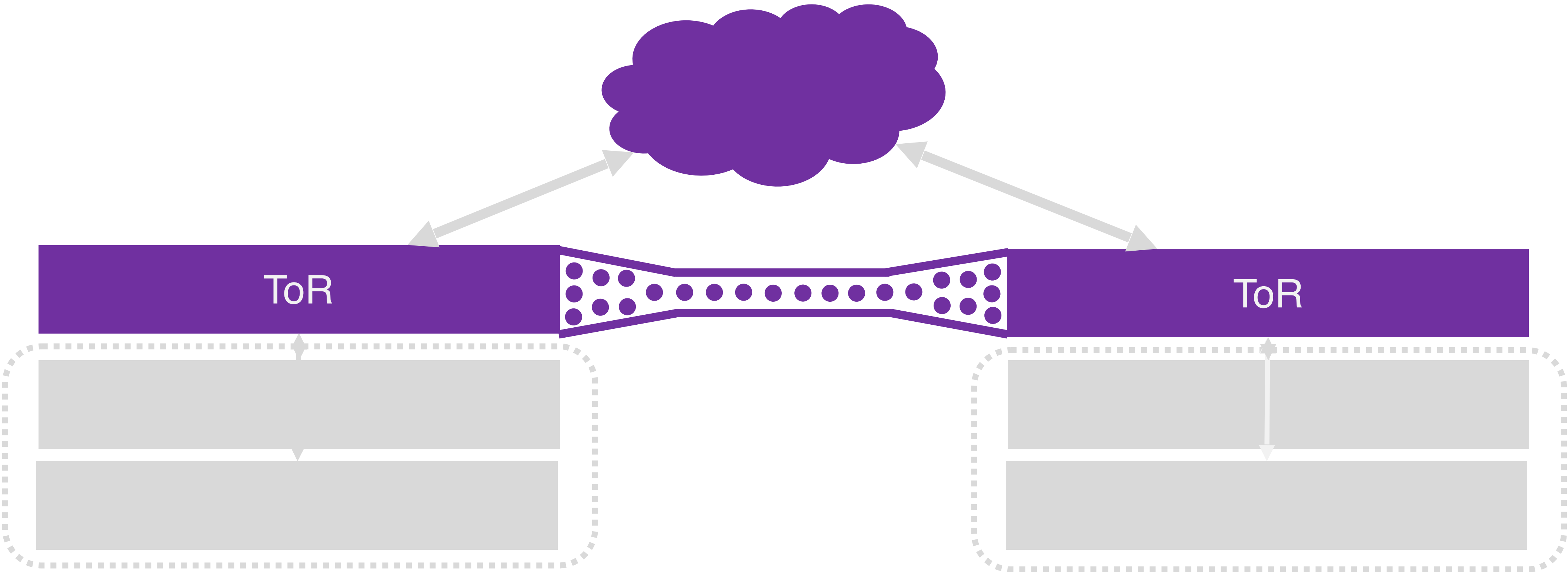
TALL AGGREGATION AND OPTIMIZATION

When Aggregation is done, PHub:

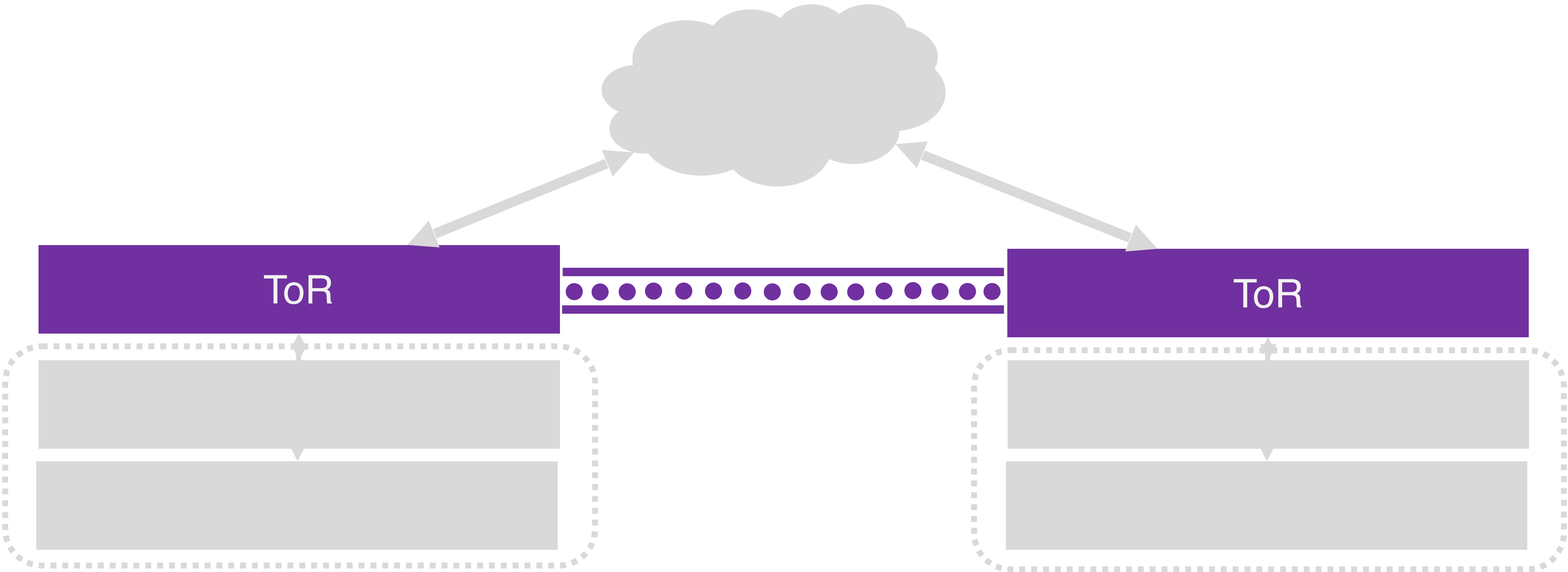
- PHub optimizes a chunk with the same core that aggregates that chunk.
- FP32-level streaming aggregation and optimization to hide communication latency.

Eliminating deployment bottlenecks:

PHub hierarchical reduction: reducing cross rack traffic

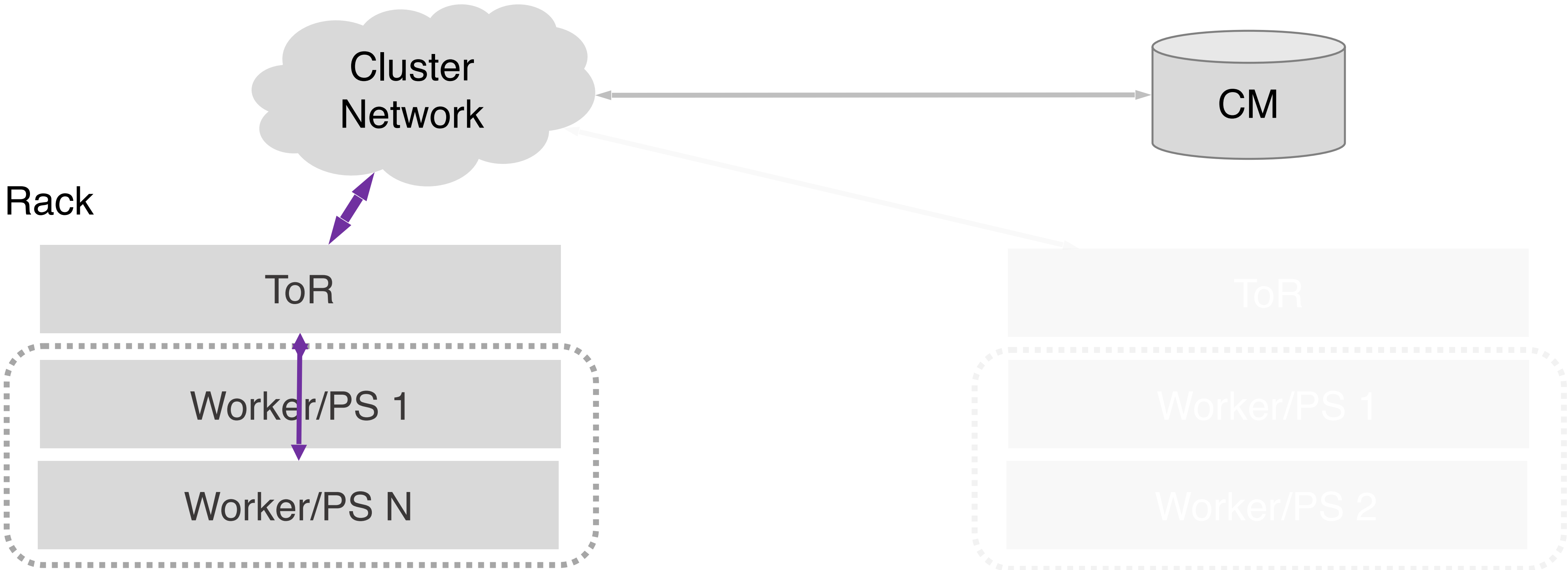


Eliminating deployment bottlenecks:
PHub hierarchical reduction: reducing cross rack traffic



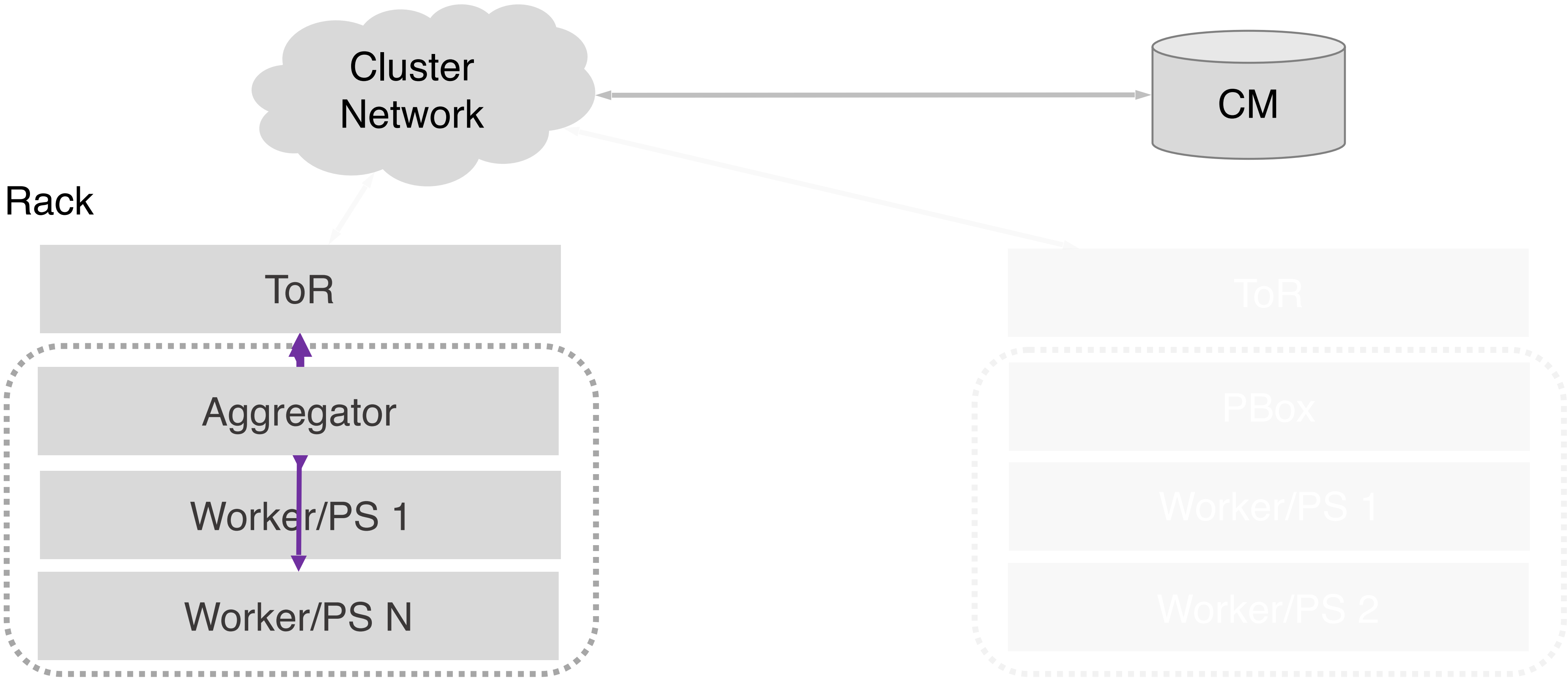
Two-Phase Hierarchical Aggregation

RACK SCALE PARAMETER SERVICE



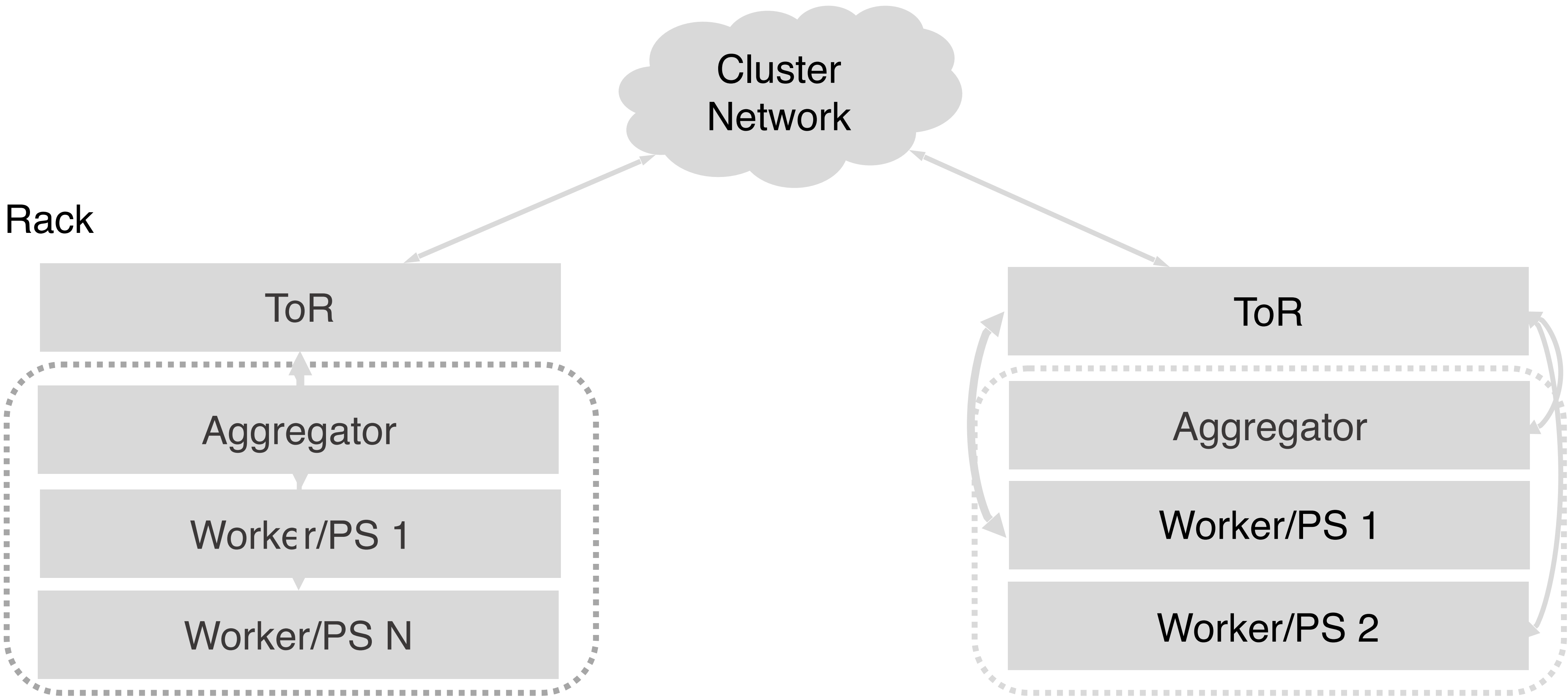
Two-Phase Hierarchical Aggregation

RACK SCALE PARAMETER SERVICE



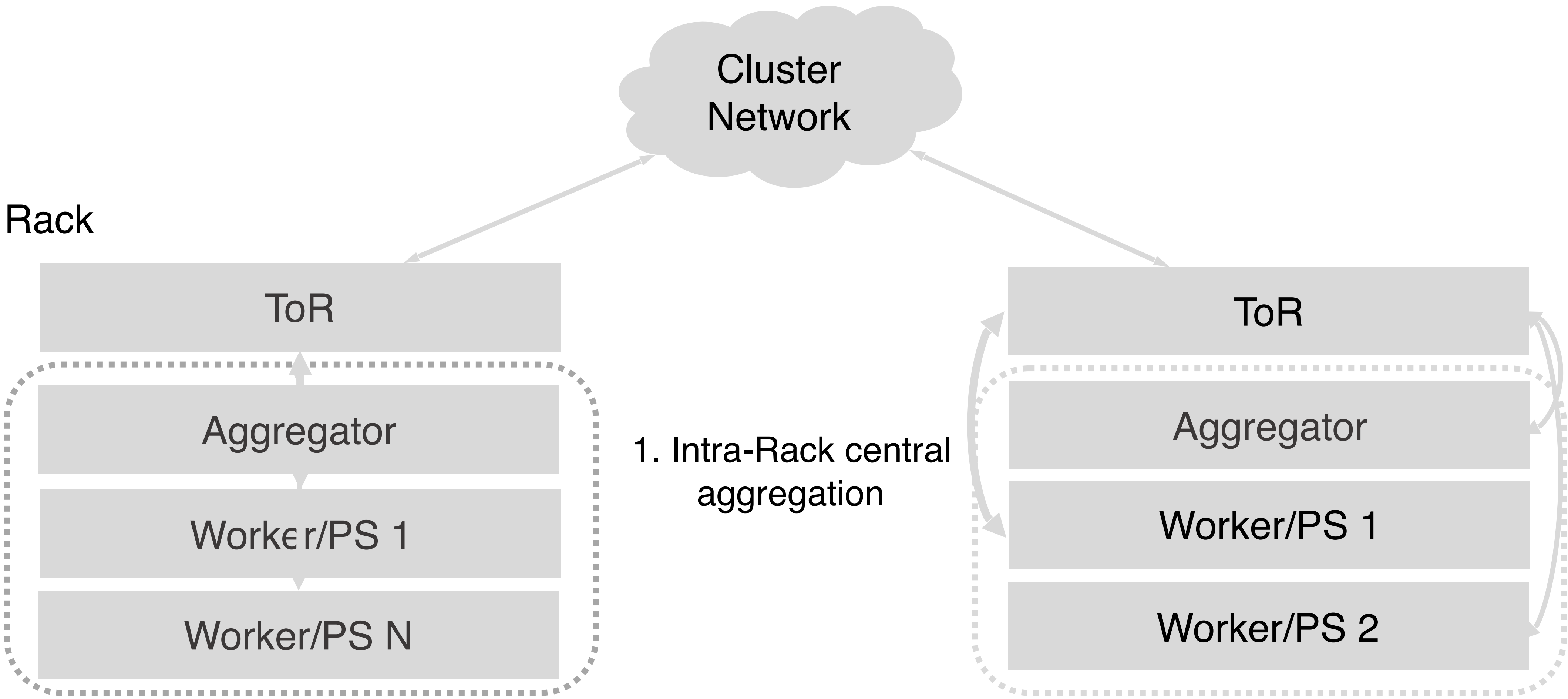
Two-Phase Hierarchical Aggregation

ADAPTING TO THE DATACENTER NETWORK TOPOLOGY



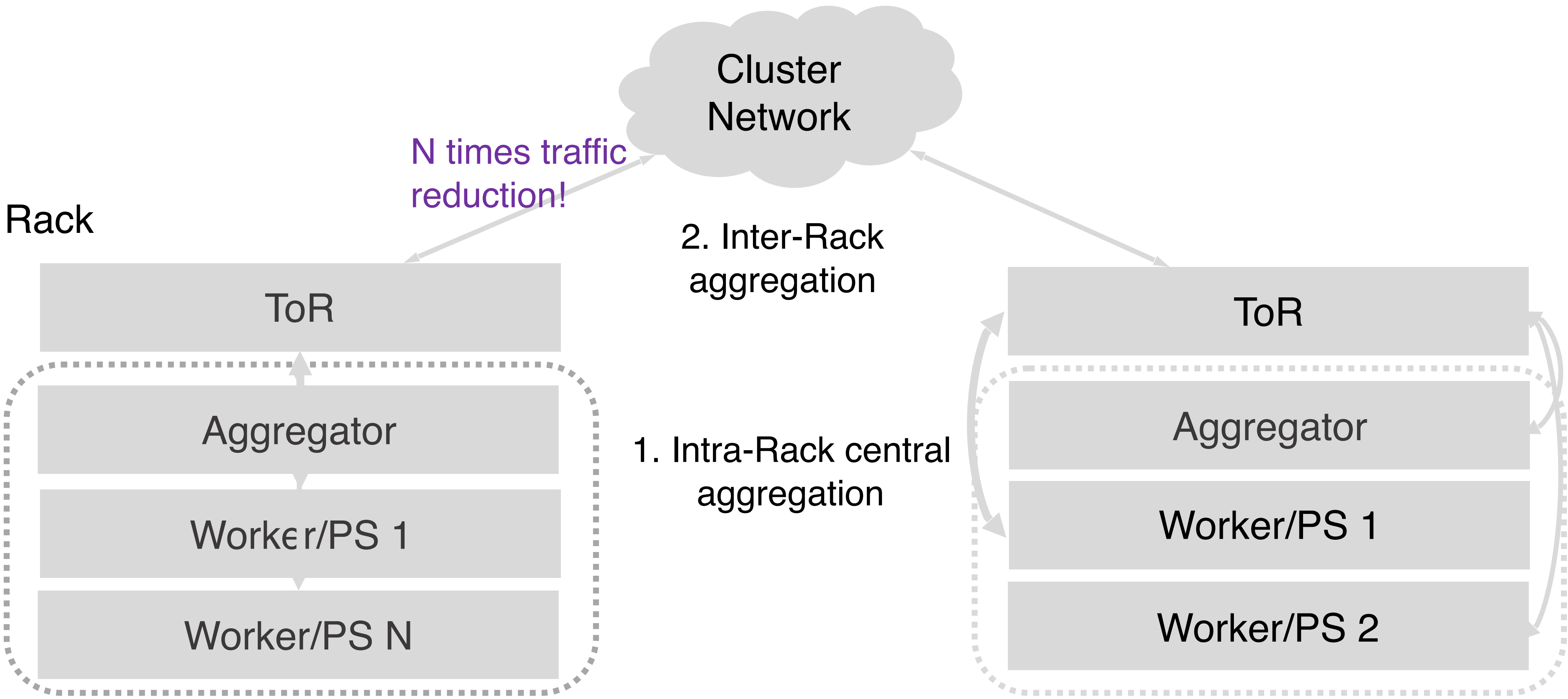
Two-Phase Hierarchical Aggregation

ADAPTING TO THE DATACENTER NETWORK TOPOLOGY



Two-Phase Hierarchical Aggregation

ADAPTING TO THE DATACENTER NETWORK TOPOLOGY



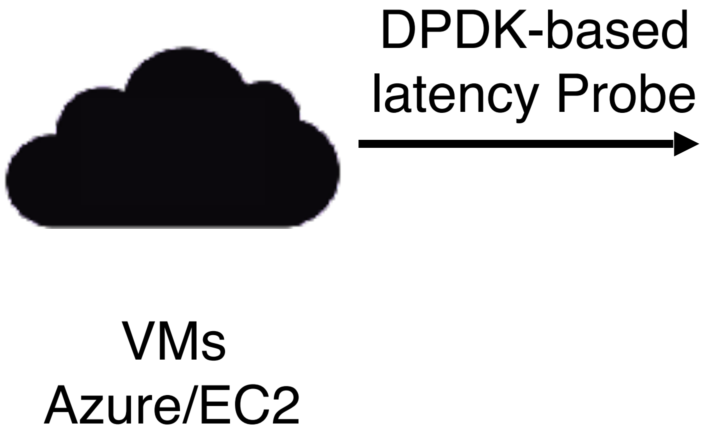
Efficient DDNN Training in Commercial Cloud

ACTIVE TOPOLOGY PROBING

VMs
Azure/EC2

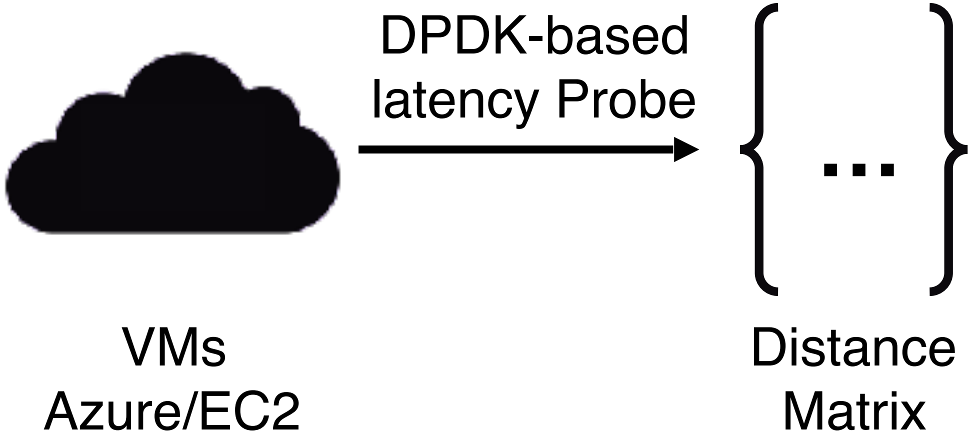
Efficient DDNN Training in Commercial Cloud

ACTIVE TOPOLOGY PROBING



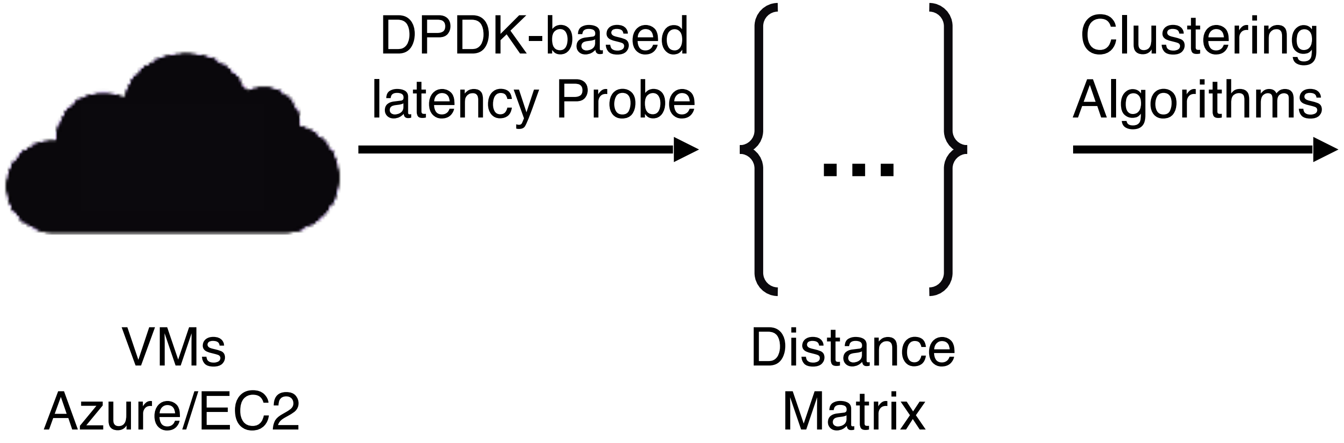
Efficient DDNN Training in Commercial Cloud

ACTIVE TOPOLOGY PROBING



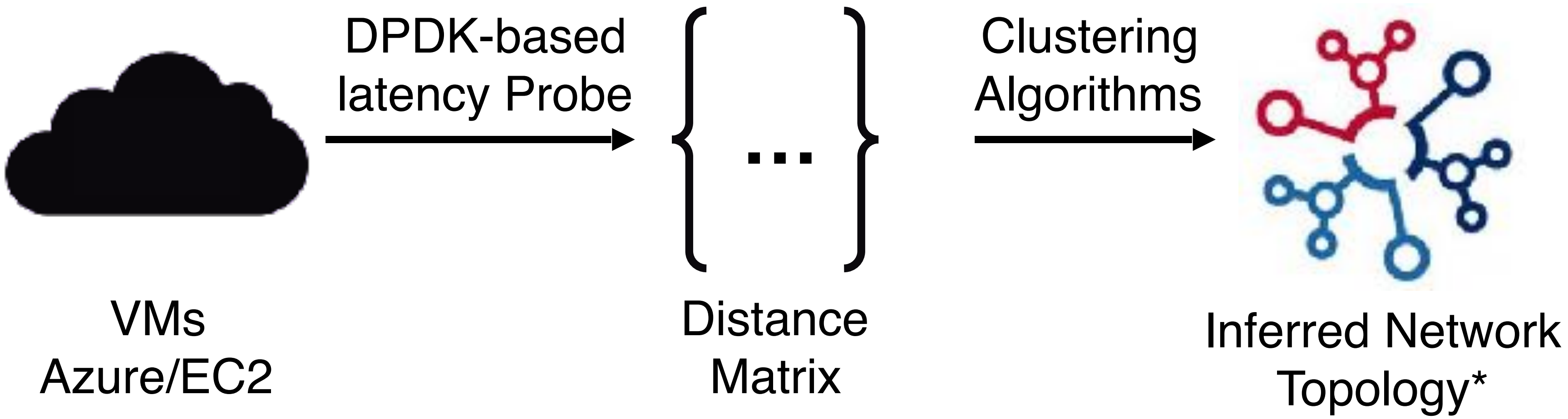
Efficient DDNN Training in Commercial Cloud

ACTIVE TOPOLOGY PROBING



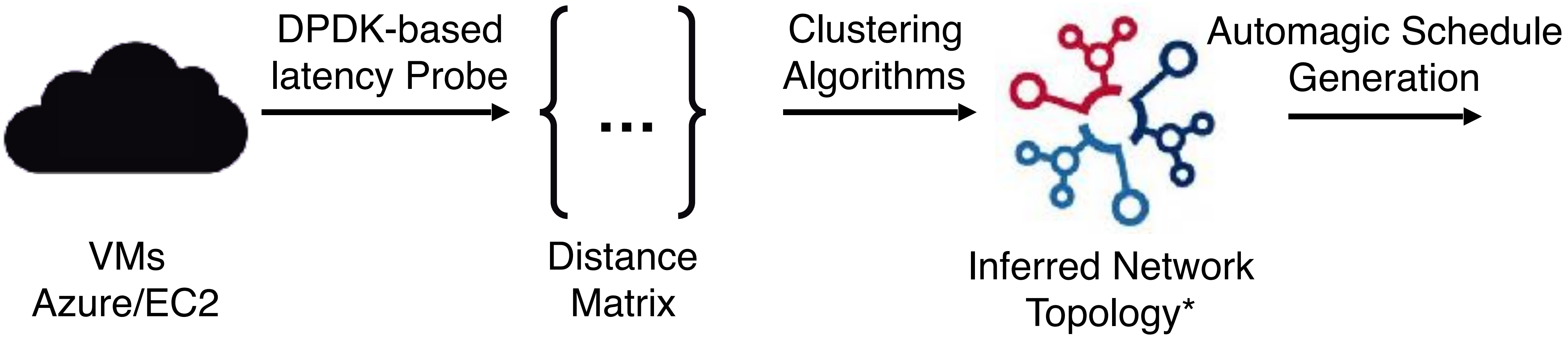
Efficient DDNN Training in Commercial Cloud

ACTIVE TOPOLOGY PROBING



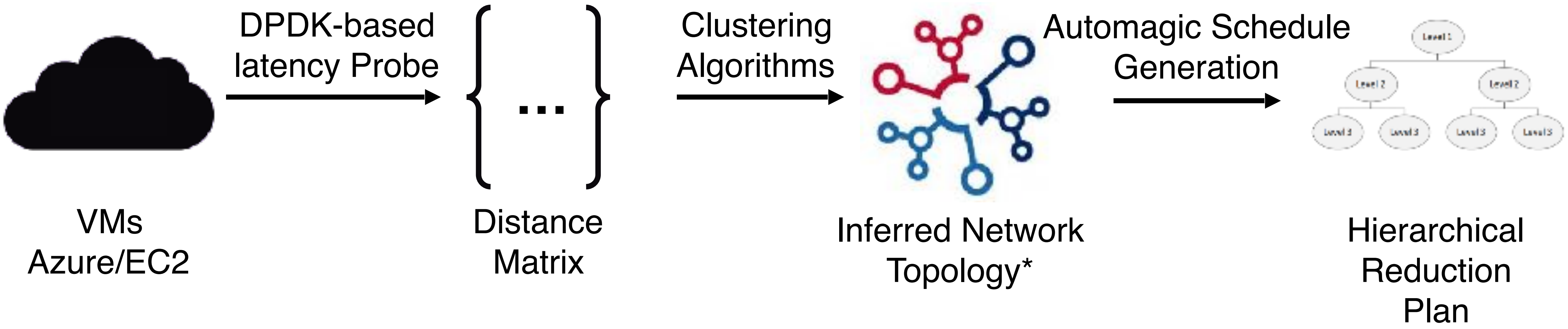
Efficient DDNN Training in Commercial Cloud

ACTIVE TOPOLOGY PROBING



Efficient DDNN Training in Commercial Cloud

ACTIVE TOPOLOGY PROBING



Efficient DDNN Training in Commercial Cloud

ACTIVE TOPOLOGY PROBING

VMs
Azure/EC2

DPDK-based
latency Probe

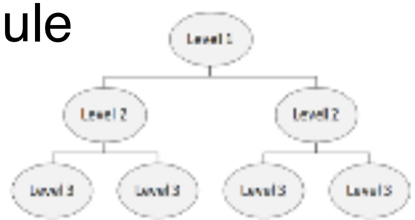
{ ... }

Distance
Matrix

Clustering
Algorithms

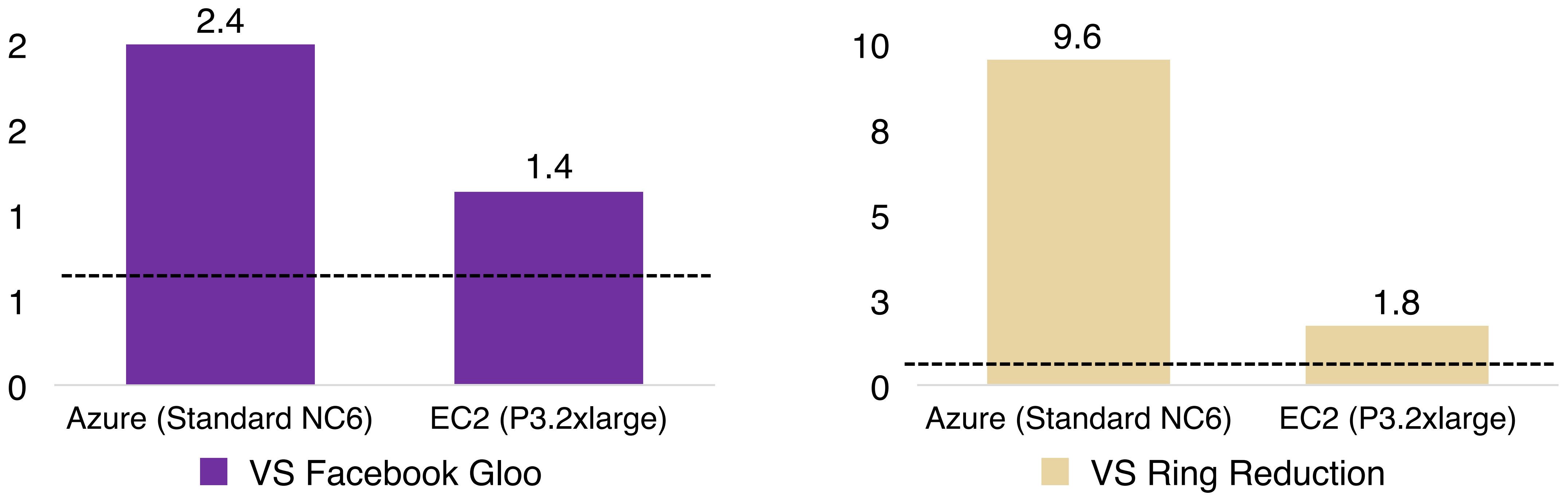
Inferred Network
Topology*

Automagic Schedule
Generation



Hierarchical
Reduction
Plan

Performance in commercial cloud with PHub

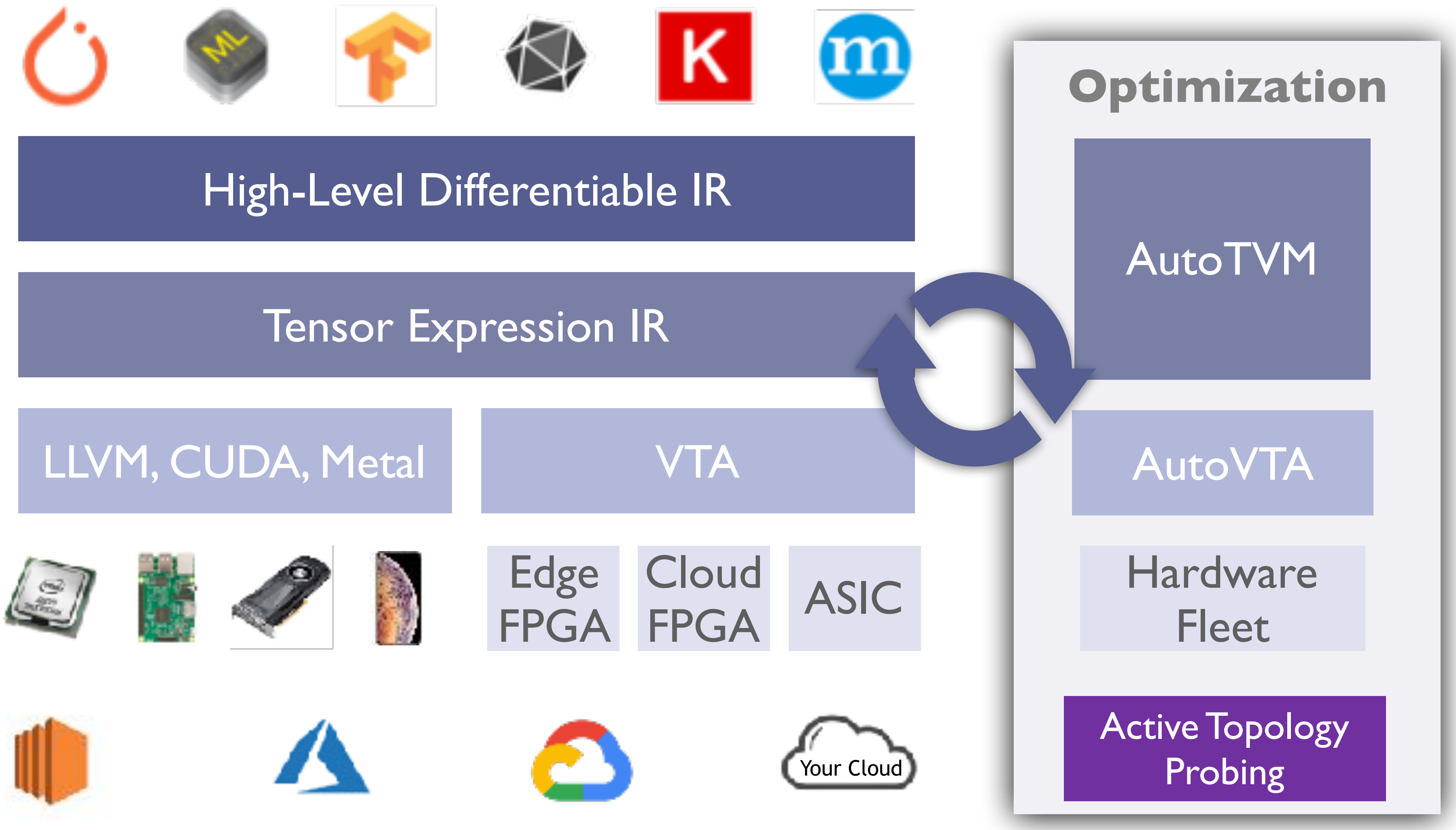


Windows Azure and Amazon EC2. 32 instances. Up to 10 Gbps. Standard_NC6: Nvidia K80. Batch Size = 512. P3.2xLarge: Nvidia V100. Batch Size = 512. Facebook Caffe2/ Pytorch. ResNet 50.

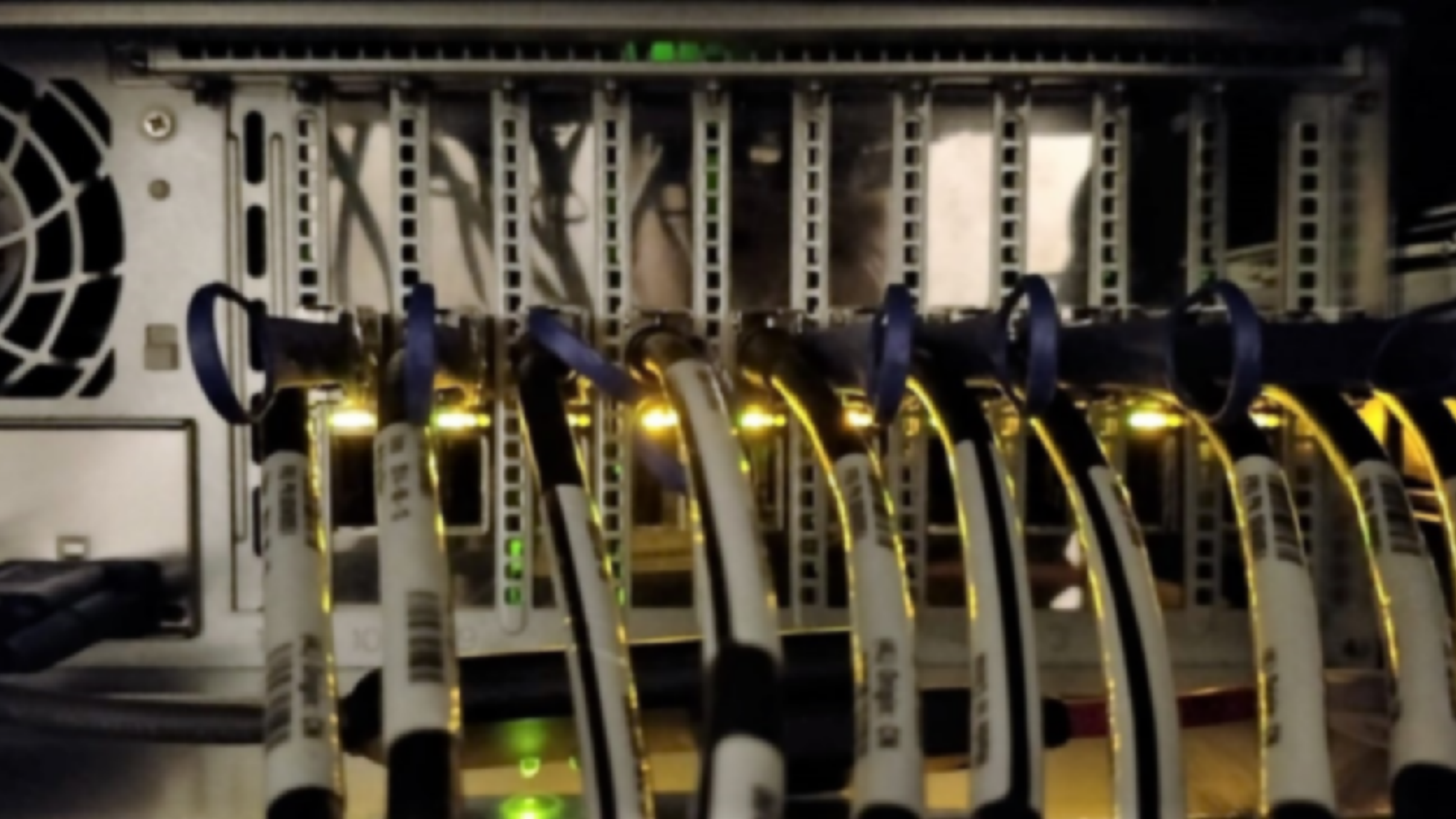
Framework Integration

Support for Mxnet/Pytorch/Caffe2.

```
var pHub = std::make_shared<PHub>(cfg.redisIp, nMap, keySize, appAddrs, cntr,
sizeof(float), cfg.rank, plp);
pHub->ToggleUseSchedule(pSchedule);
pHub->Reduce();
```



Groundwork for bringing TVM to the distributed world for training and inference, on commercial cloud, or in your own cluster.



Hardware Parameter Hub

Hardware Parameter Hub

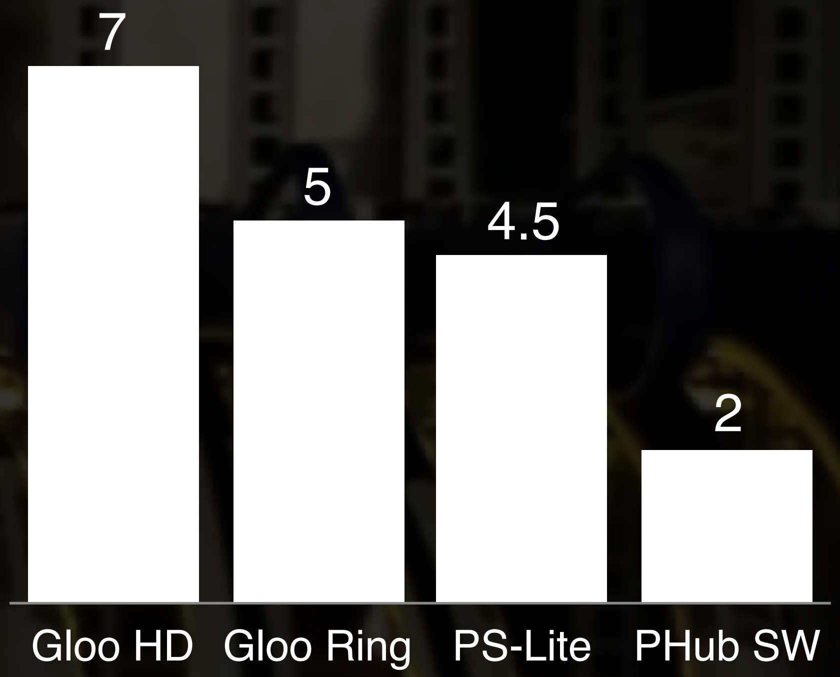
Hardware Parameter Hub

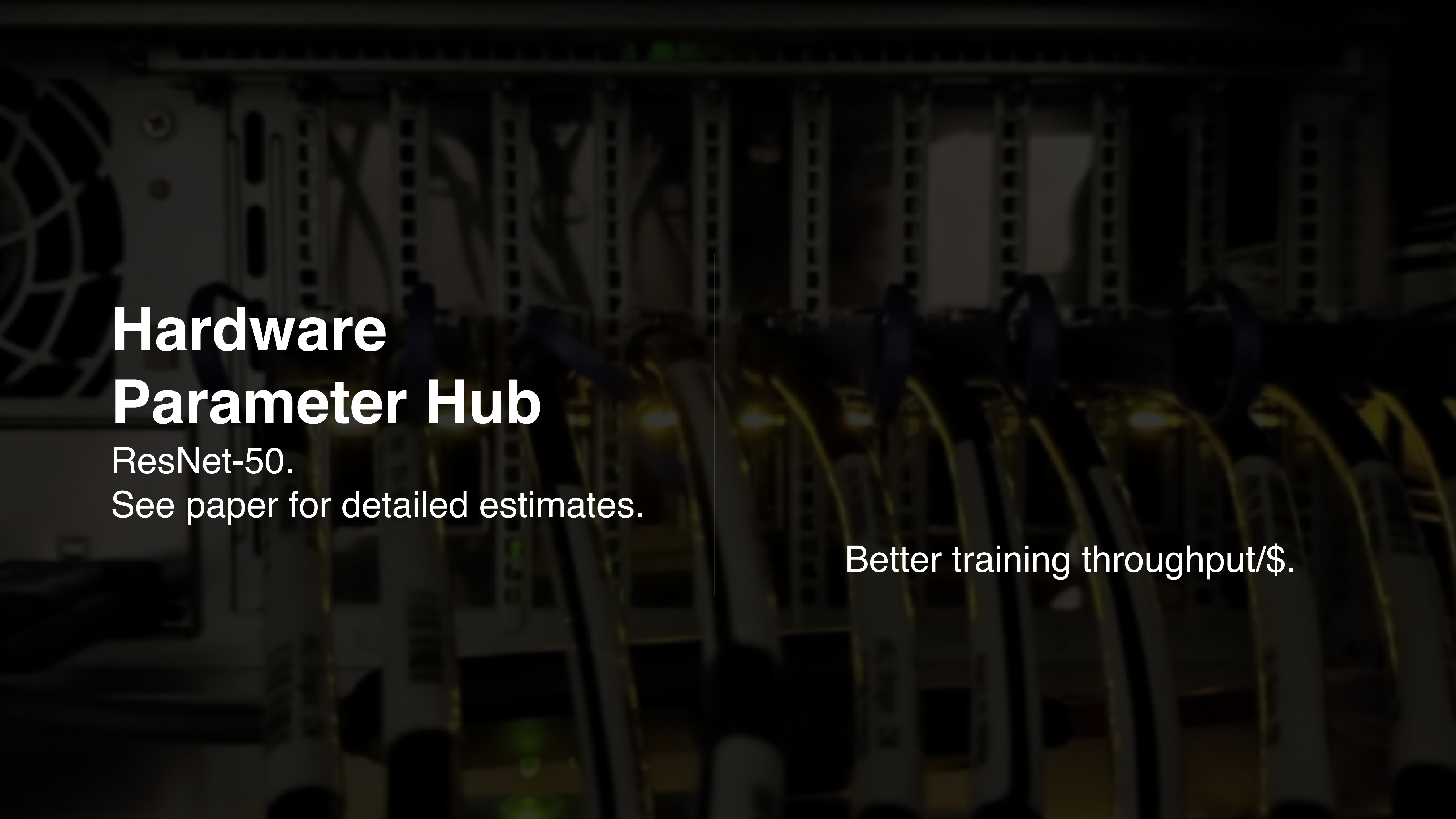
Balanced computation and communication resource.

- 10 ConnectX-3 Card
- 560+Gbps Network BW
- 800Gbps PCIe
- Fully supported by Software Parameter Hub

Hardware Parameter Hub

35GB/s aggregation throughput. Supports 100+ ResNet-50 training nodes with a single machine.





Hardware Parameter Hub

ResNet-50.
See paper for detailed estimates.

Better training throughput/$.

Hardware Parameter Hub

ResNet-50.
See paper for detailed estimates.

25%

Better training throughput/$.