# GeePS: Scalable deep learning on distributed GPUs with a GPU-specialized parameter server

Henggang Cui

Hao Zhang, Gregory R. Ganger, Phillip B. Gibbons, and Eric P. Xing

PARALLEL DATA LABORATORY

Carnegie Mellon University

Carnegie Mellon
Parallel Data Laboratory

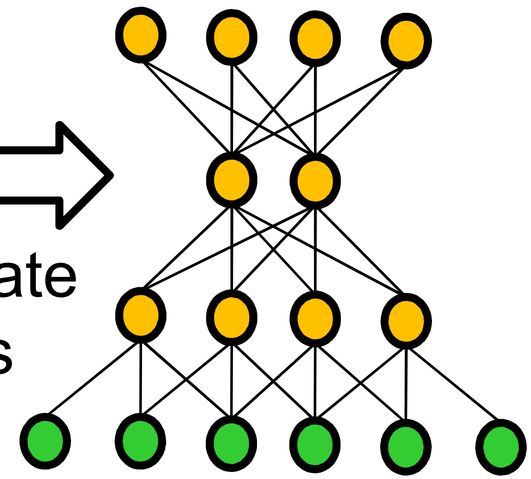
# Image classification w/ deep learning


Training data:
images w/ labels

Machine learning
program

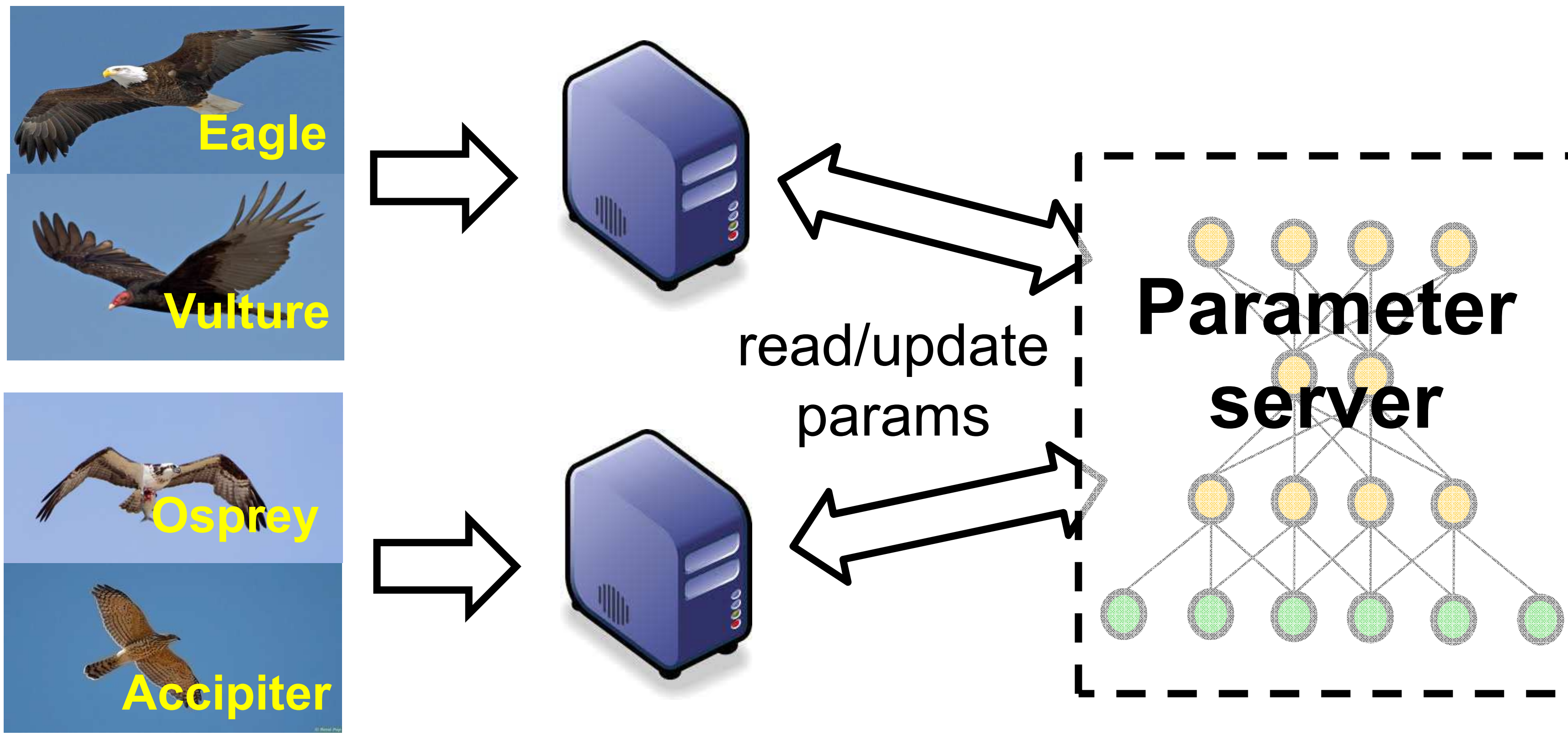
read/update
params

Deep neural network:
interconnected neurons



Model parameters:
connection weights
(solution)

# Distributed deep learning

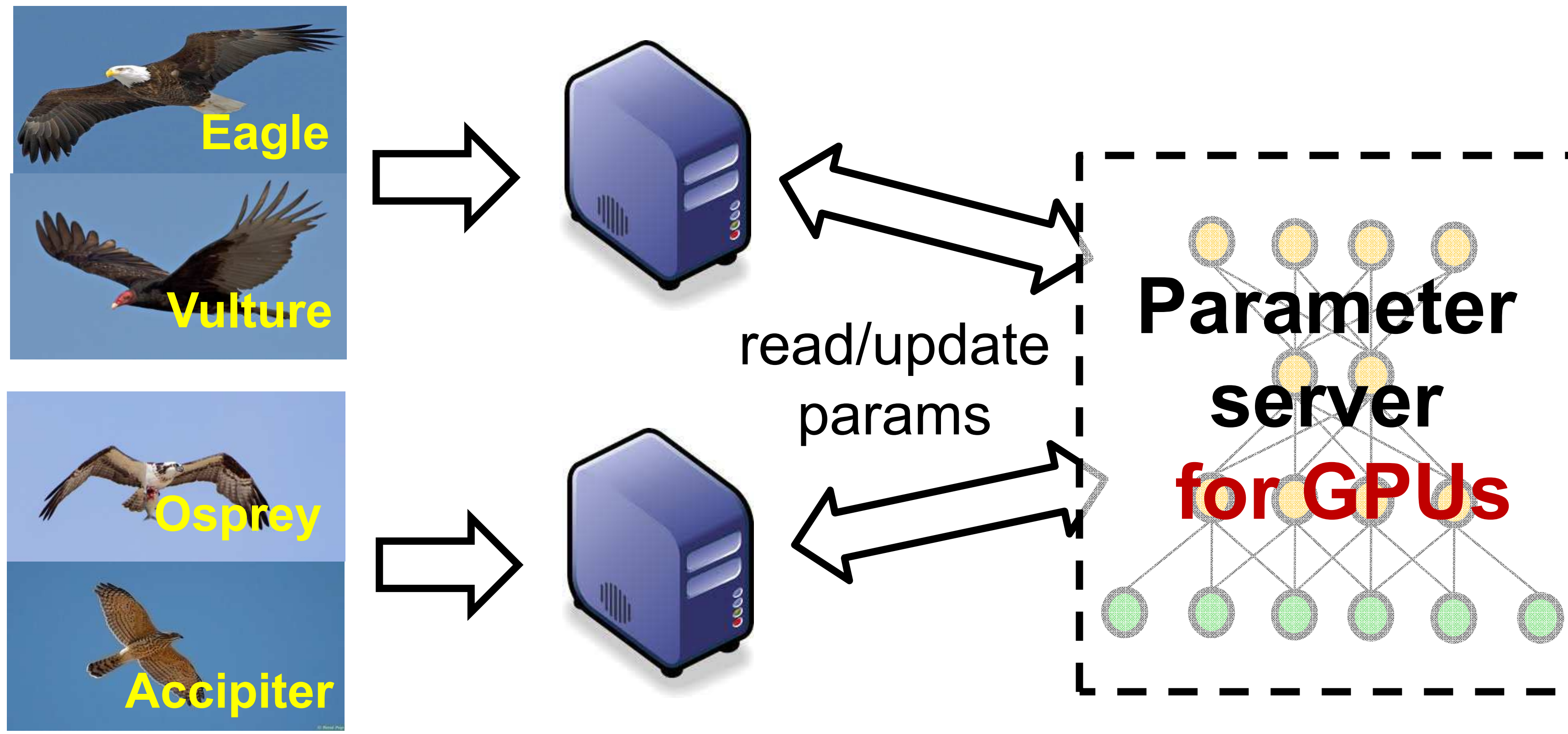



**Partitioned**
Training data

**Distributed**
ML workers

**Shared**
model parameters

# Distributed deep learning




**Partitioned**
Training data

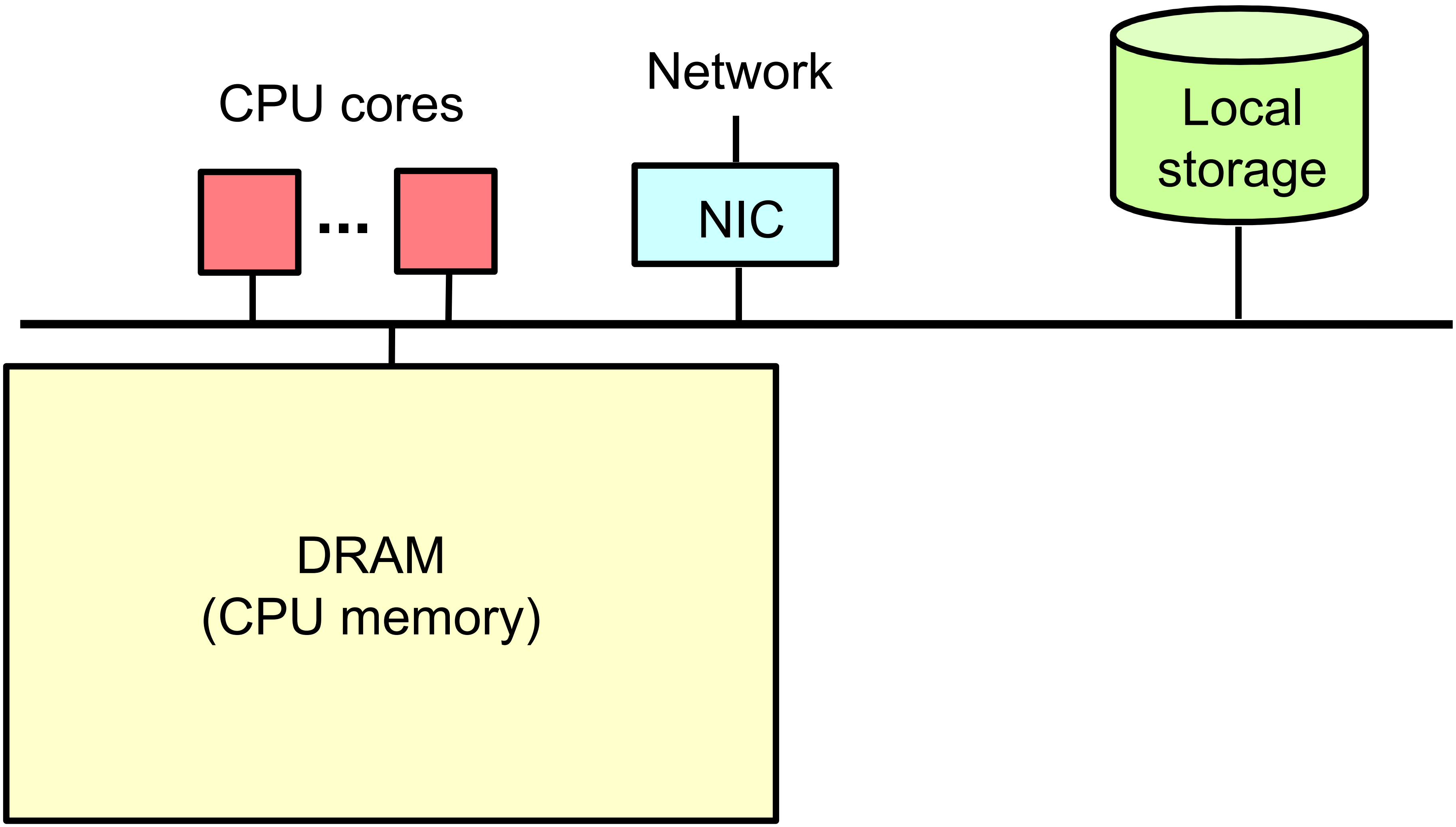
**Distributed**
**GPU** ML workers

**Shared**
model parameters

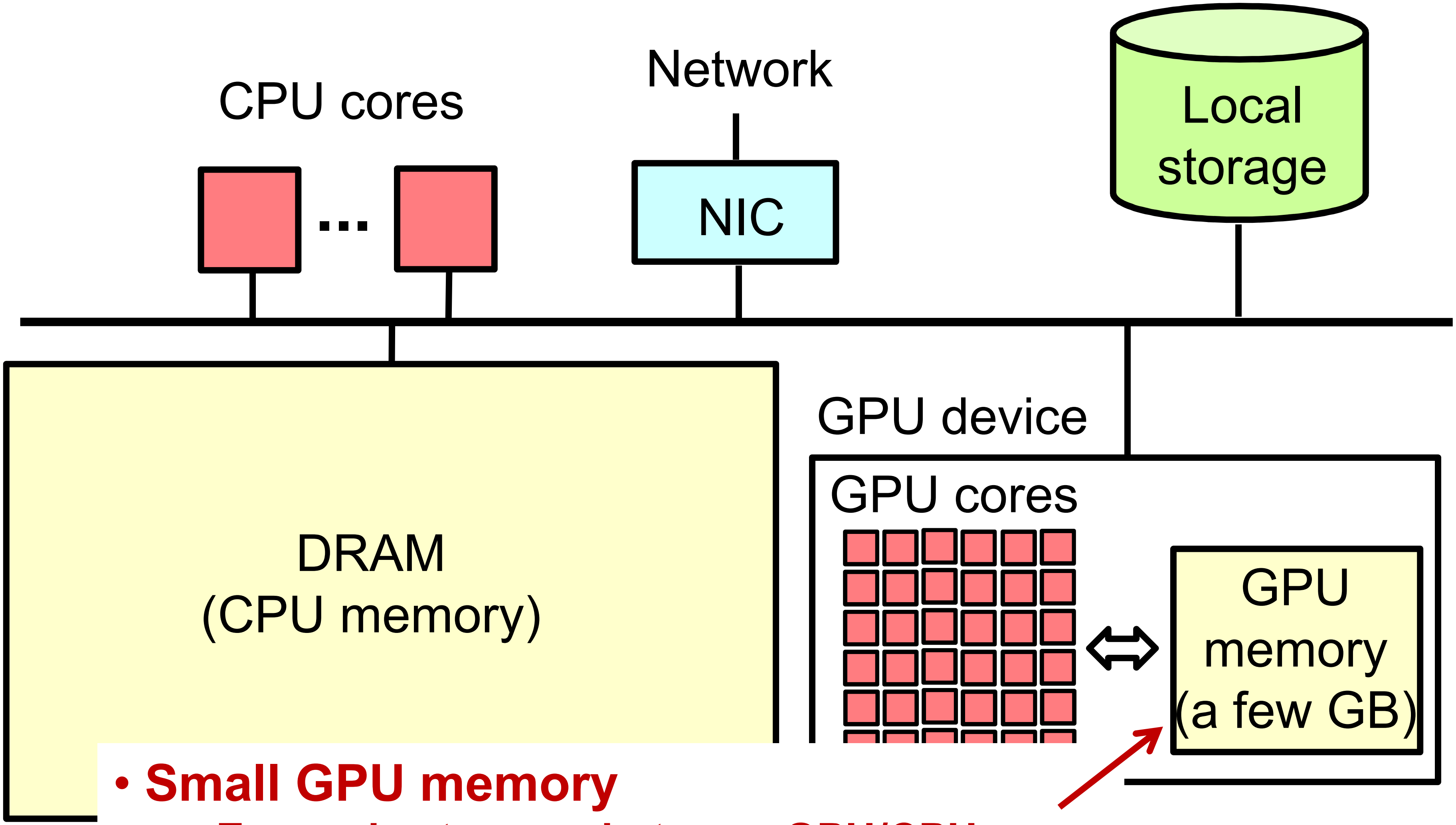
# Outline

- Background
  - Deep learning with GPUs
  - Parallel ML using parameter servers
- GeePS: GPU-specialized parameter server
- Experiment results

# A machine with no GPU

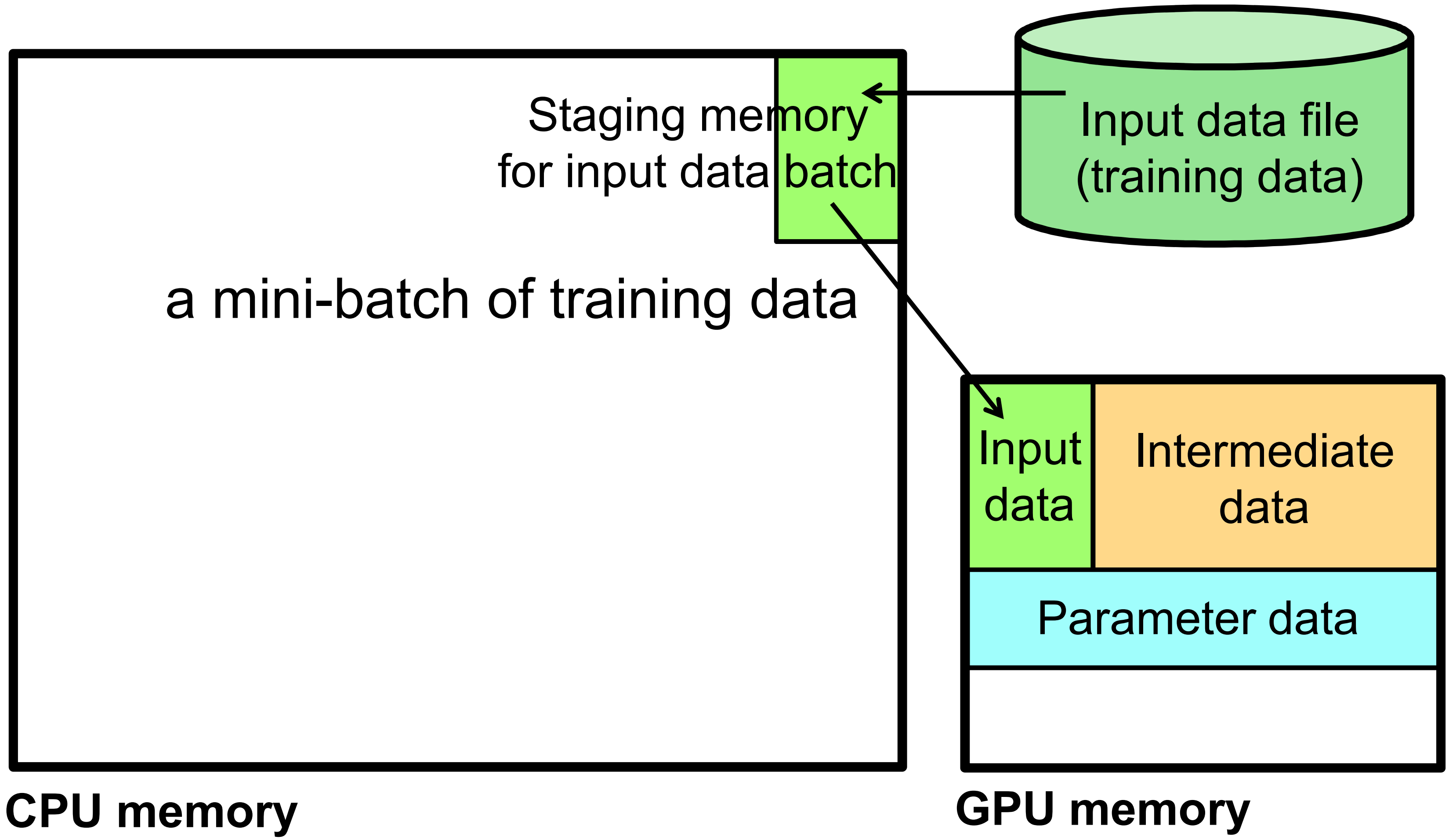


# A machine with a GPU device

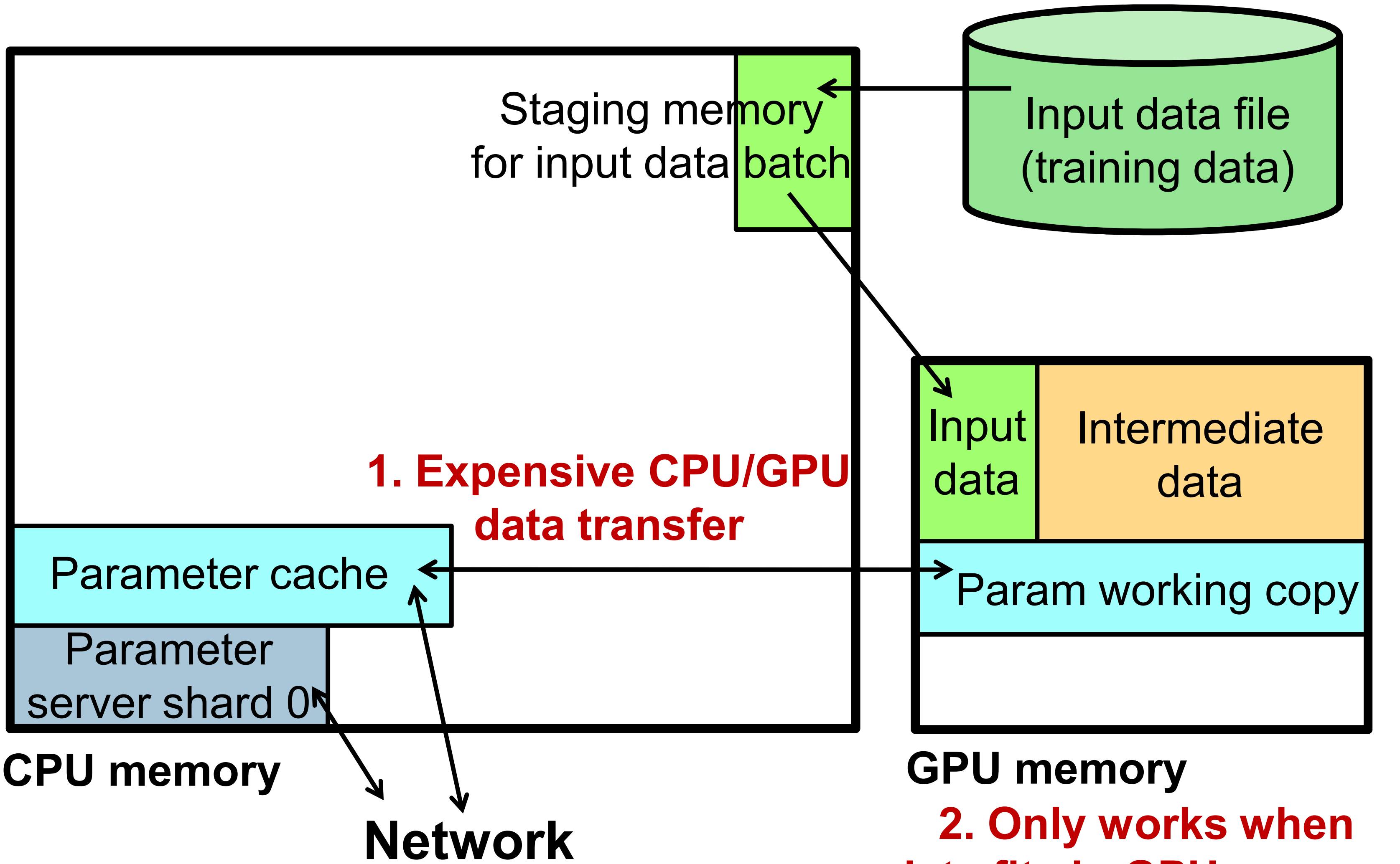



- **Small GPU memory**
  - **Expensive to copy between GPU/CPU mem**

# Single GPU machine learning



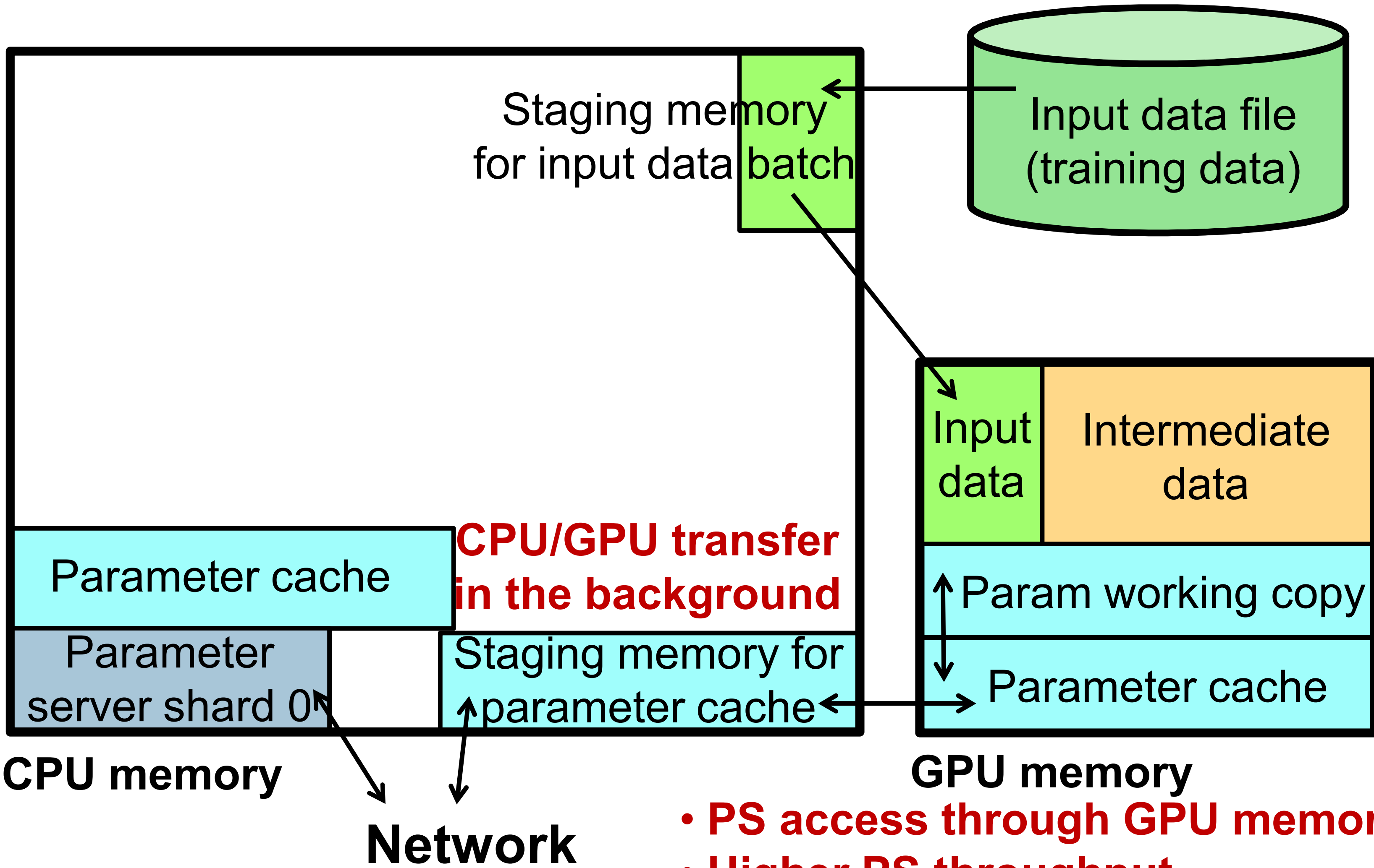
# Multi-GPU ML via CPU param. serv.



# Outline

- Background
  - Deep learning with GPUs
  - Parallel ML using parameter servers
- GeePS: GPU-specialized parameter server
  - Maintaining the parameter cache in GPU memory
  - Batch access with GPU cores for higher throughput
  - Managing limited GPU device memory
- Experiment results

# Multi-GPU ML via GeePS

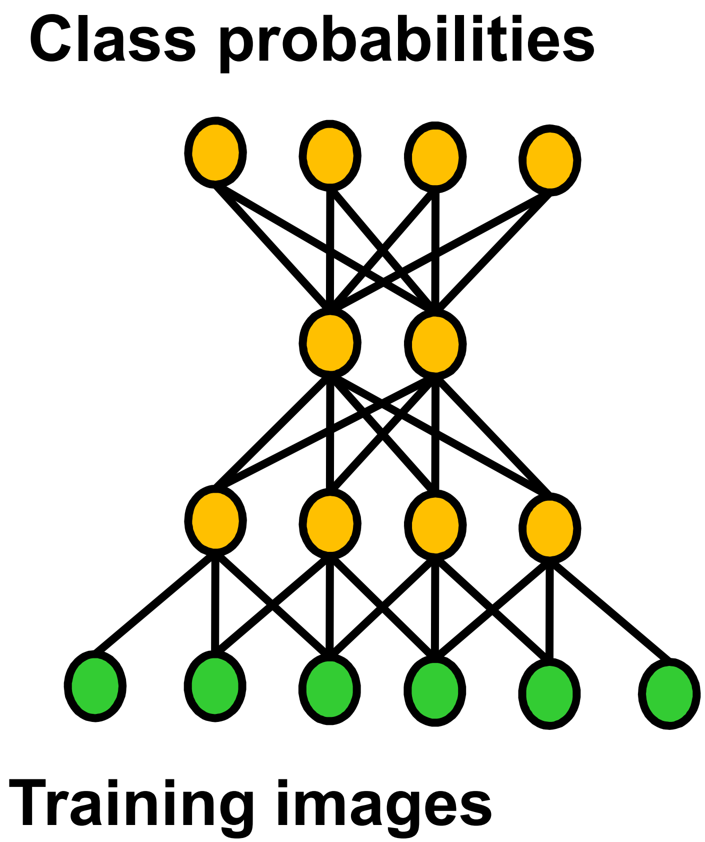



- **PS access through GPU memory**
- **Higher PS throughput**

# Outline

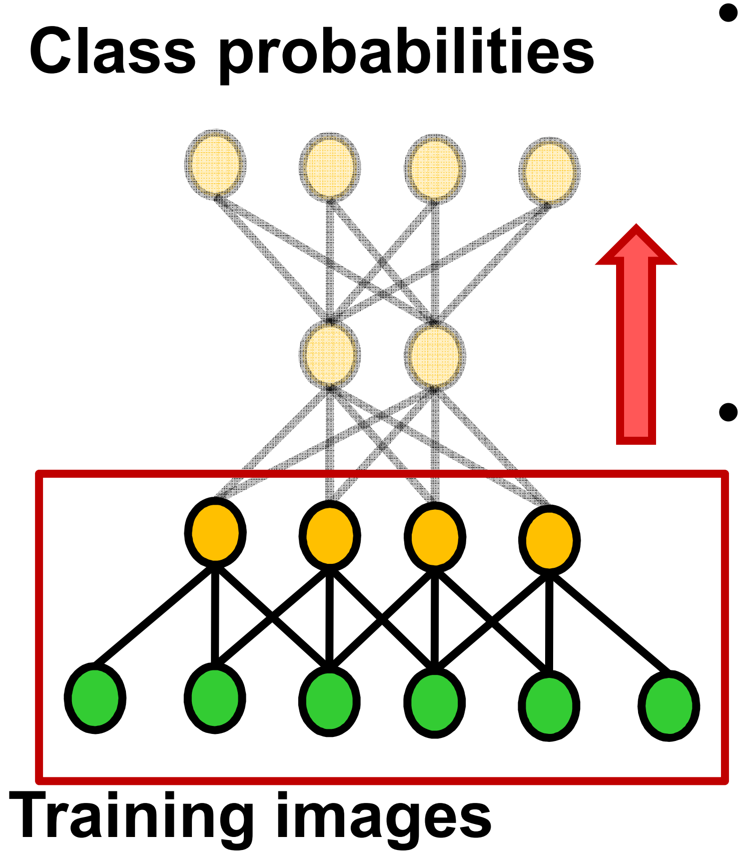
- Background
- GeePS: GPU-specialized parameter server
  - Maintaining the parameter cache in GPU memory
  - Batch access with GPU cores for higher throughput
  - Managing limited GPU device memory
- Experiment results

# Layer-by-layer computation for DNN



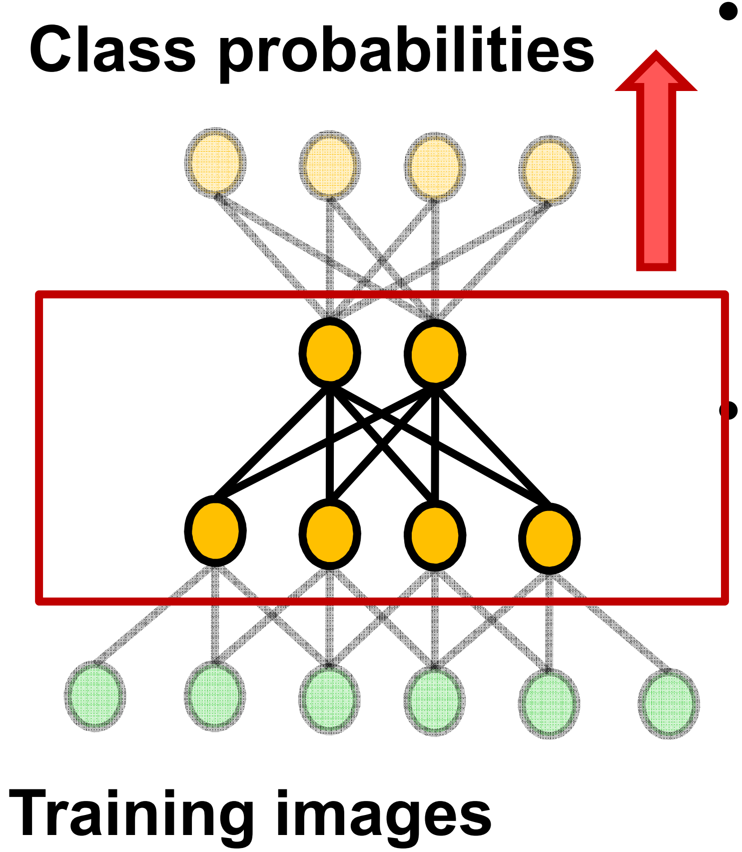

- For each iteration (mini-batch)
  - A forward pass
  - Then a backward pass
- Each time only data of two layers are used

# Layer-by-layer computation for DNN



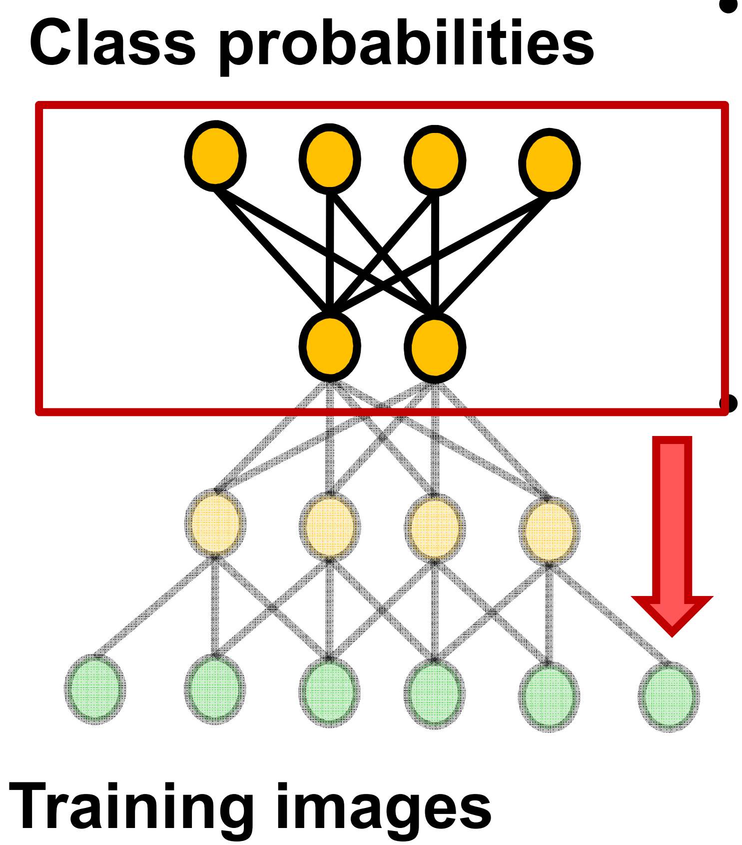

- For each iteration (mini-batch)
  - A forward pass
  - Then a backward pass
- Each time only data of two layers are used

# Layer-by-layer computation for DNN



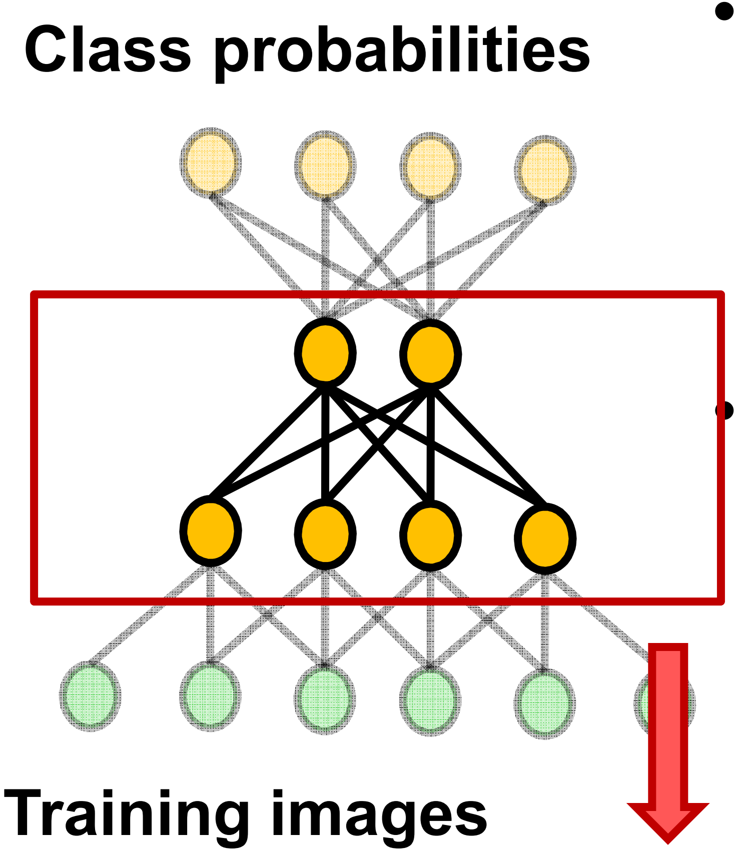

- For each iteration (mini-batch)
  - A forward pass
  - Then a backward pass
- Each time only data of two layers are used

# Layer-by-layer computation for DNN



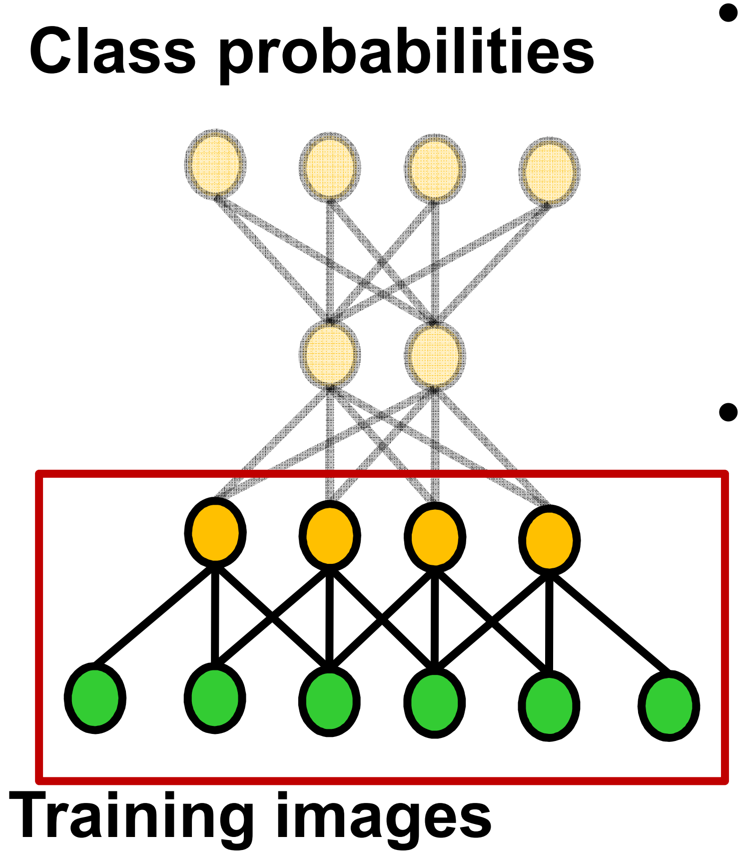

- For each iteration (mini-batch)
  - A forward pass
  - Then a backward pass
- Each time only data of two layers are used

# Layer-by-layer computation for DNN




- For each iteration (mini-batch)
  - A forward pass
  - Then a backward pass
- Each time only data of two layers are used

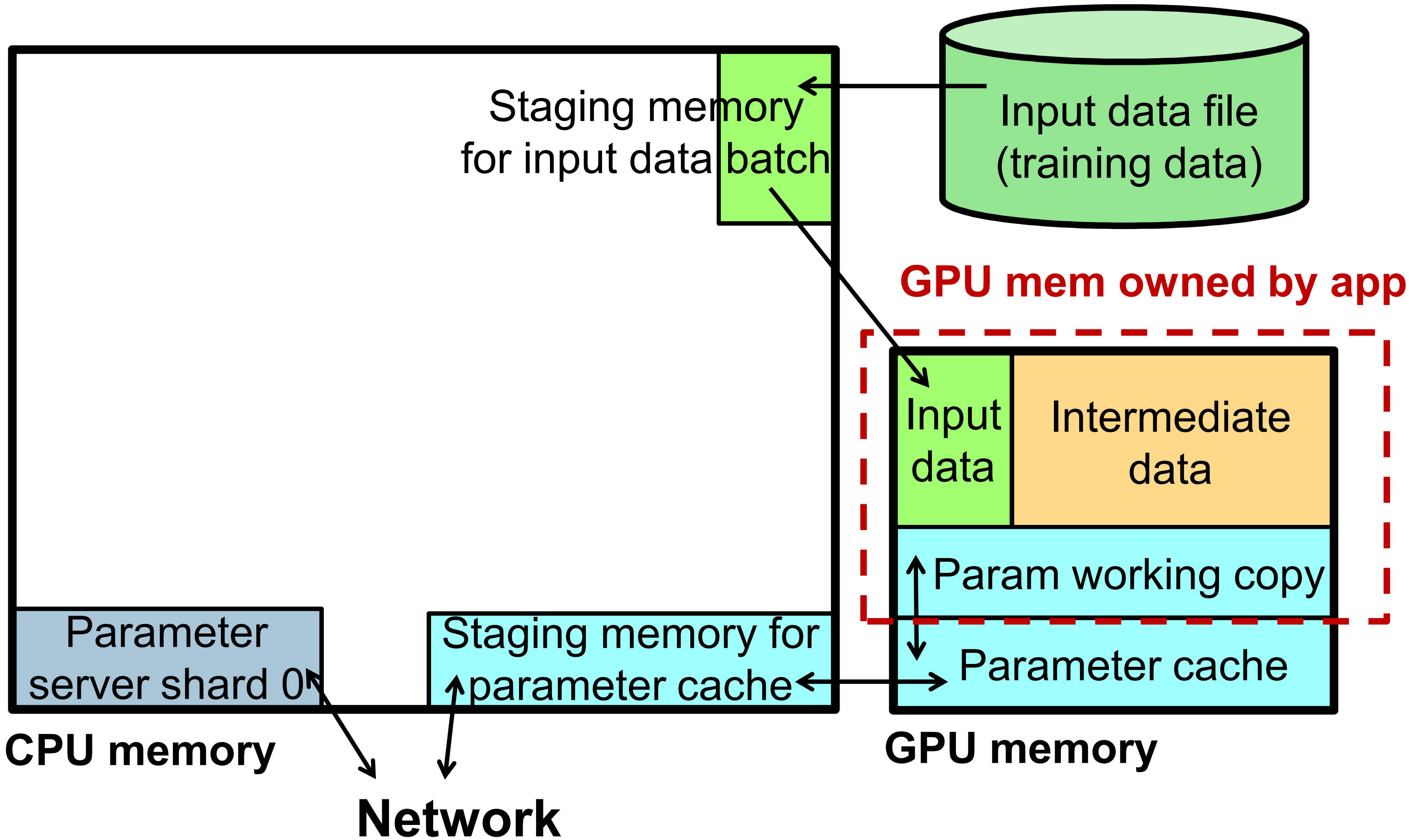
# Layer-by-layer computation for DNN



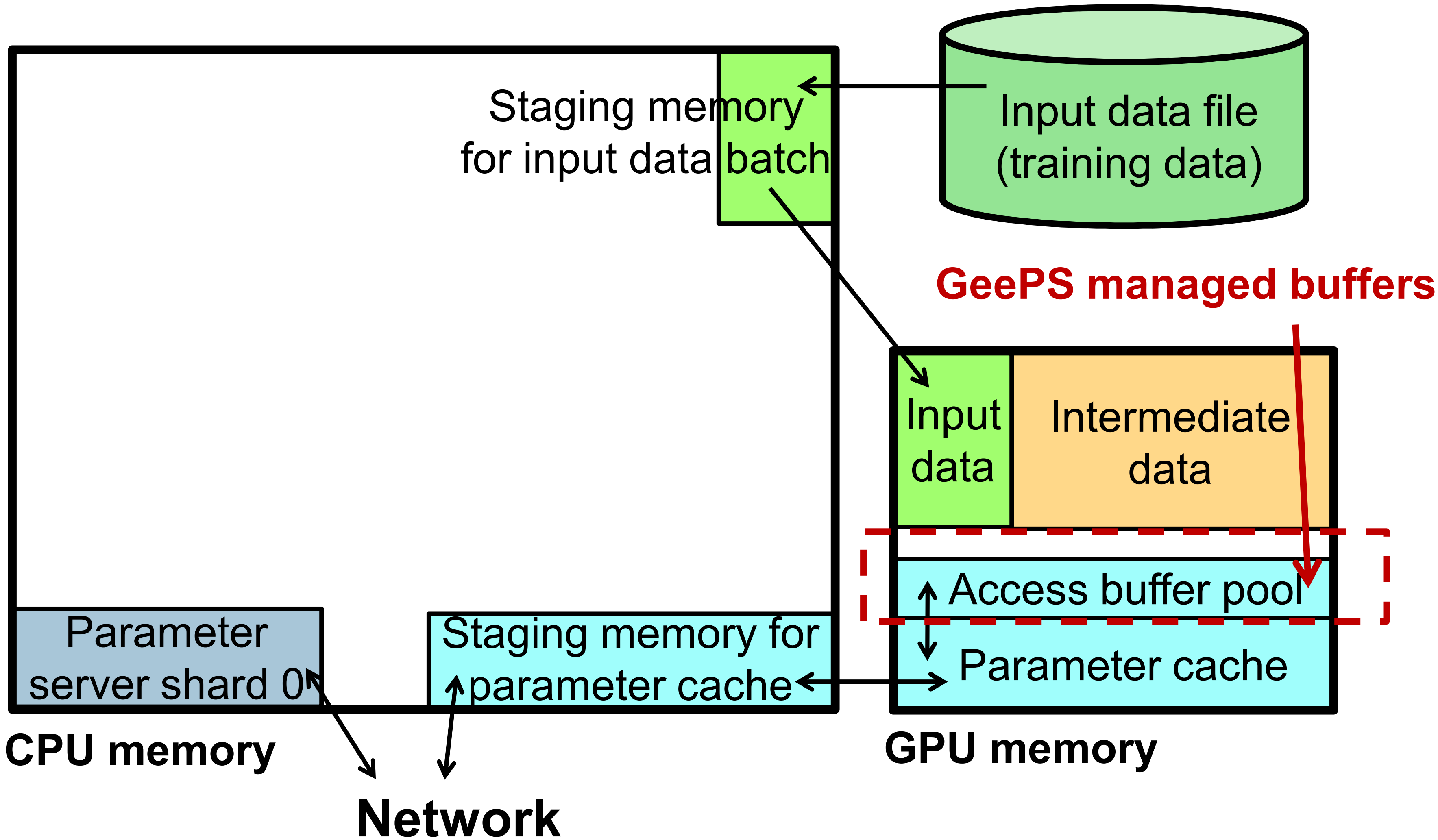

- For each iteration (mini-batch)
  - A forward pass
  - Then a backward pass
- Each time only data of two layers are used

- **Use GPU mem as a cache to keep actively used data**
- **Store the remaining in CPU mem**

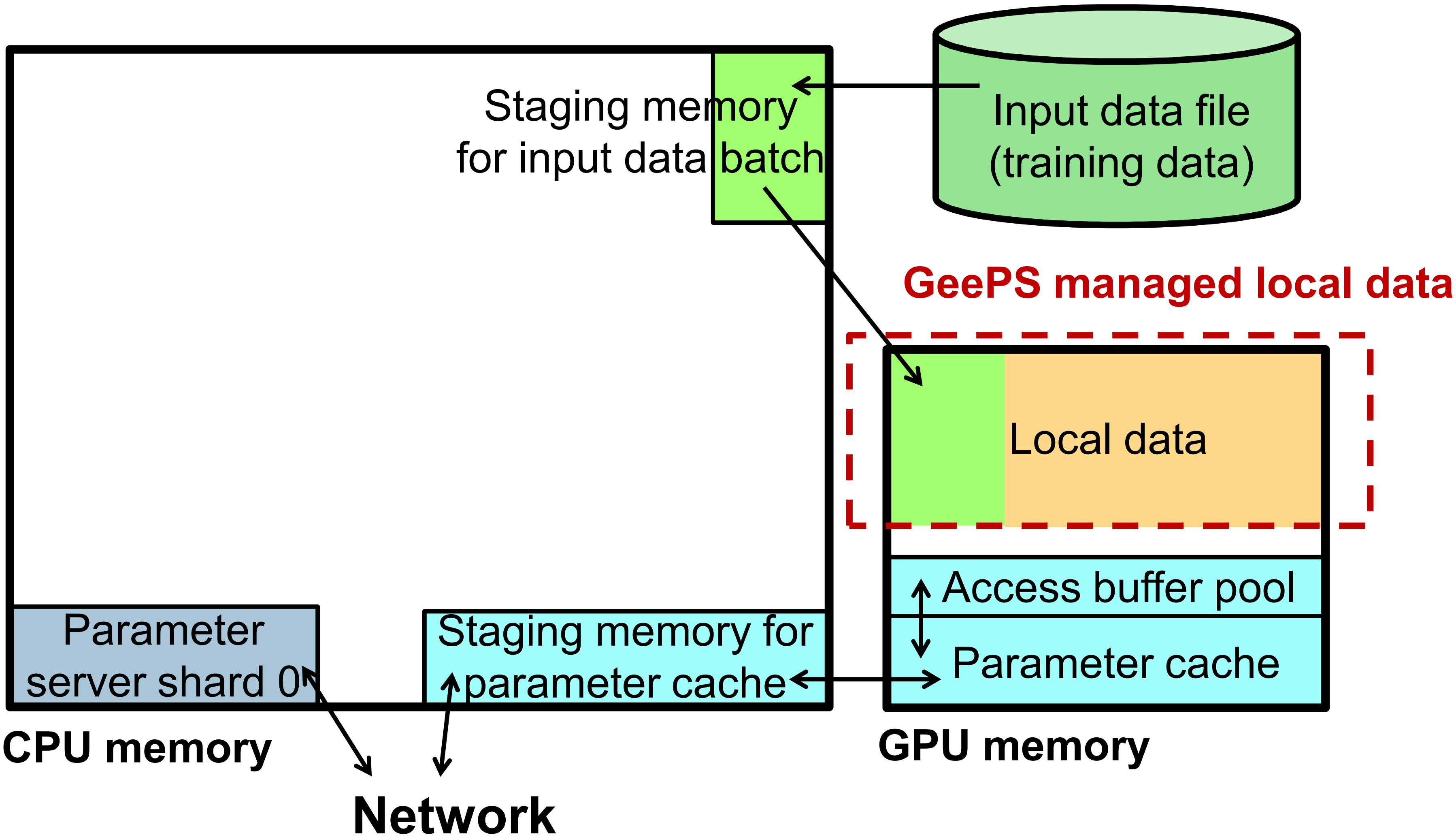
# GPU memory management



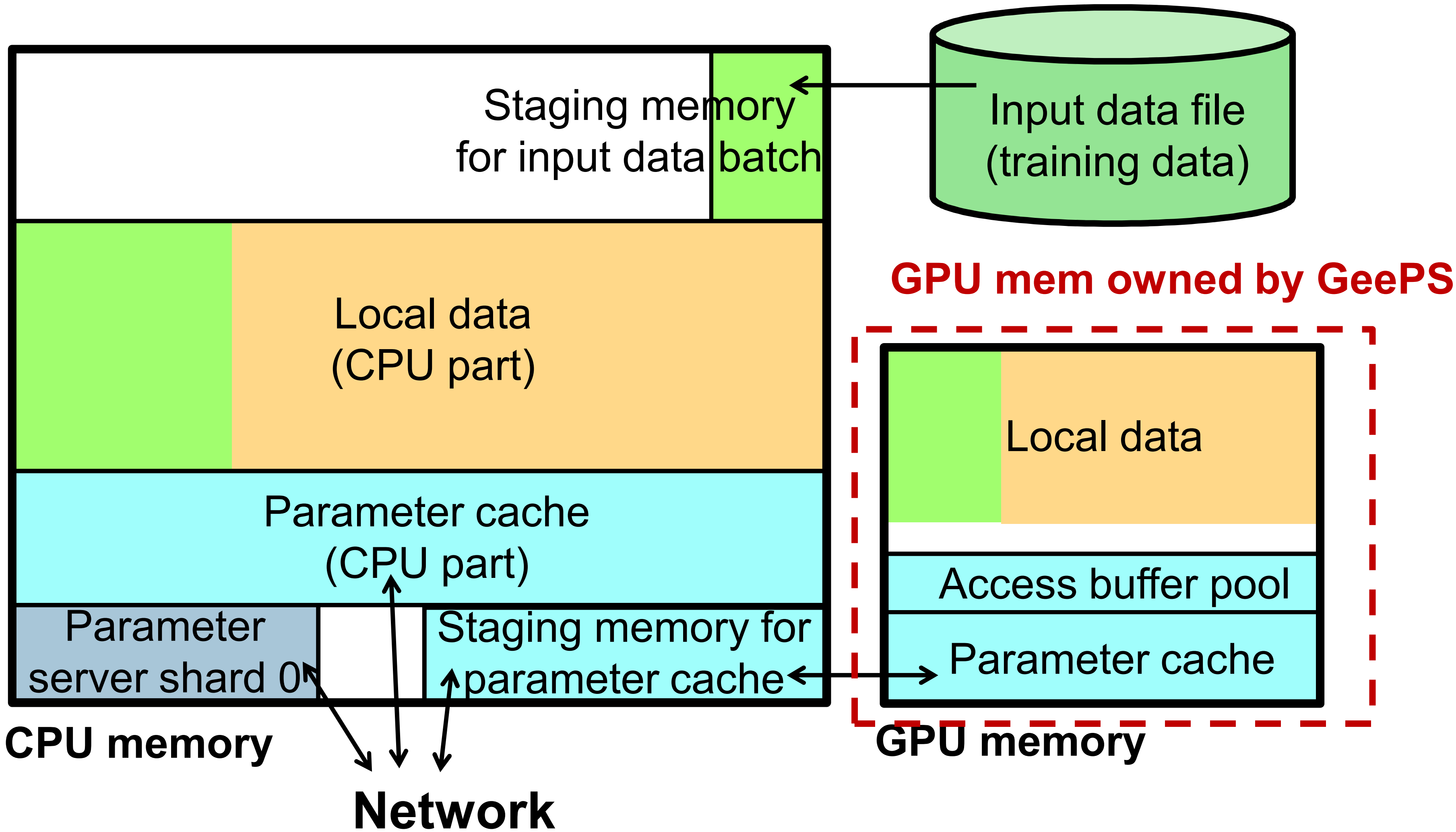
# GeePS-managed buffers



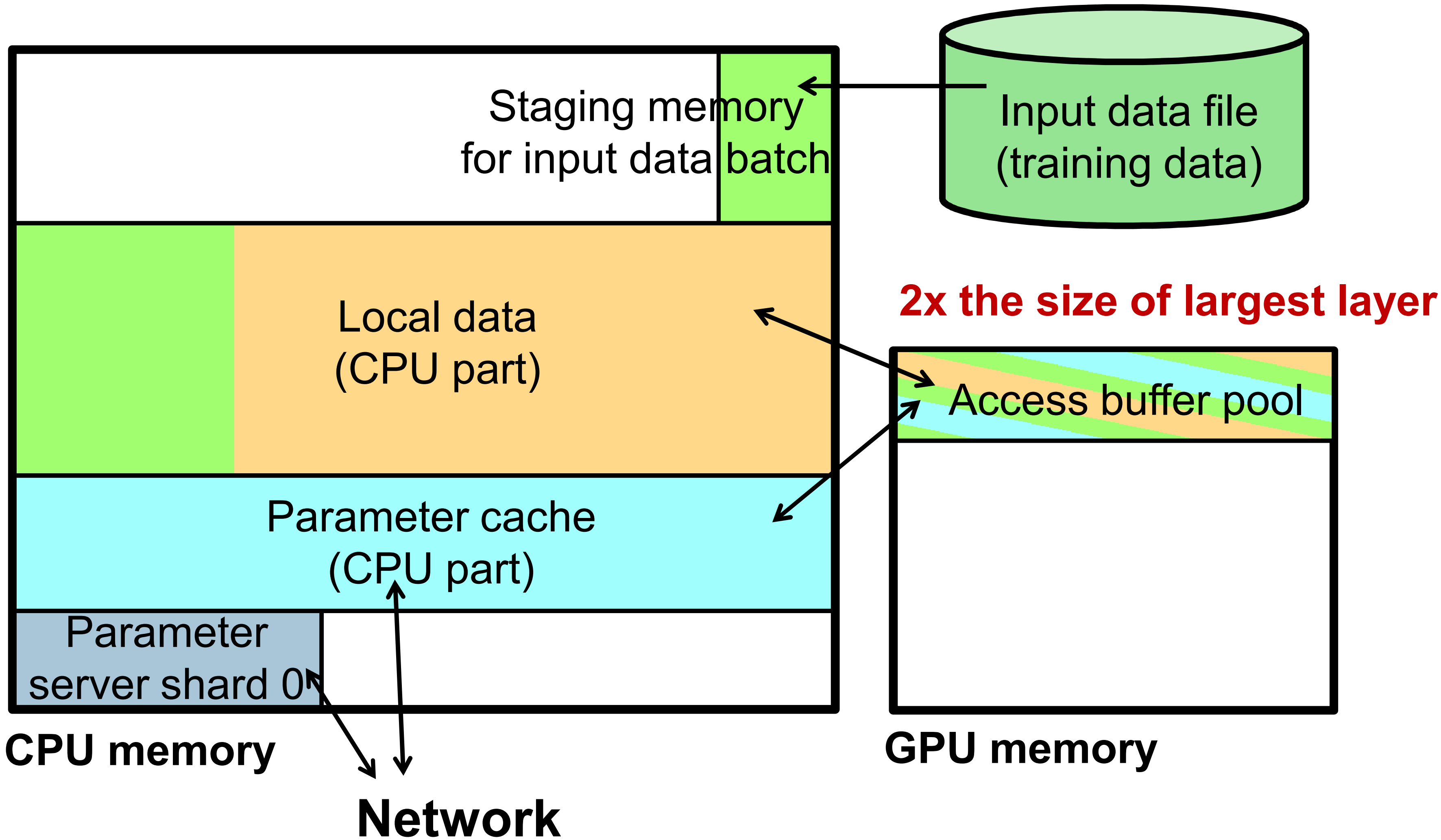
# GeePS manages local data also



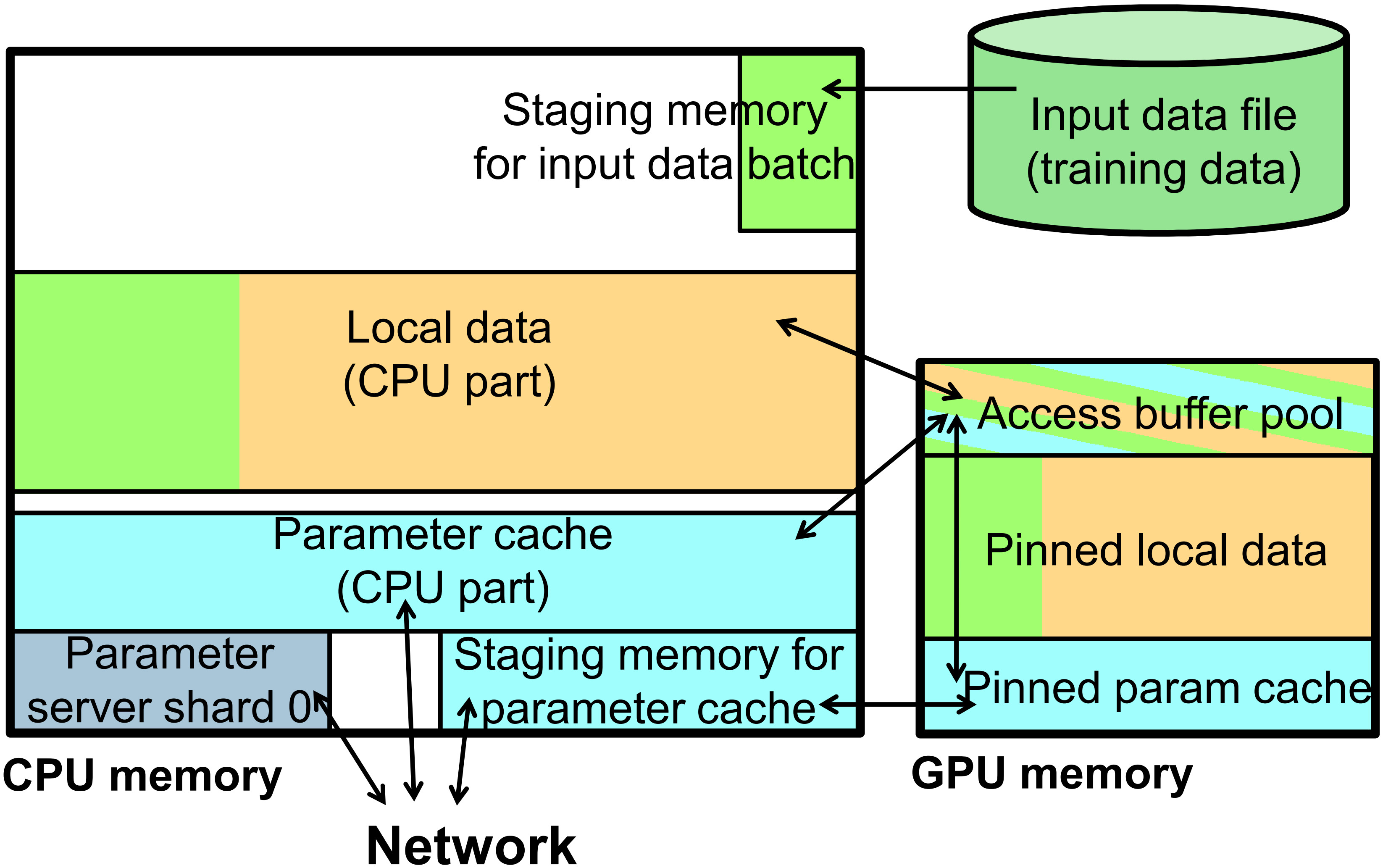
# Use CPU memory when not fit



# Use CPU memory when not fit



# Use CPU memory when not fit



# Outline

- Background
- GeePS: GPU-specialized parameter server
  - Maintaining the parameter cache in GPU memory
  - Batch access with GPU cores for higher throughput
  - Managing limited GPU device memory
- Experiment results

# Experimental setups

- Cluster information
  - Tesla K20C GPUs with 5 GB GPU memory

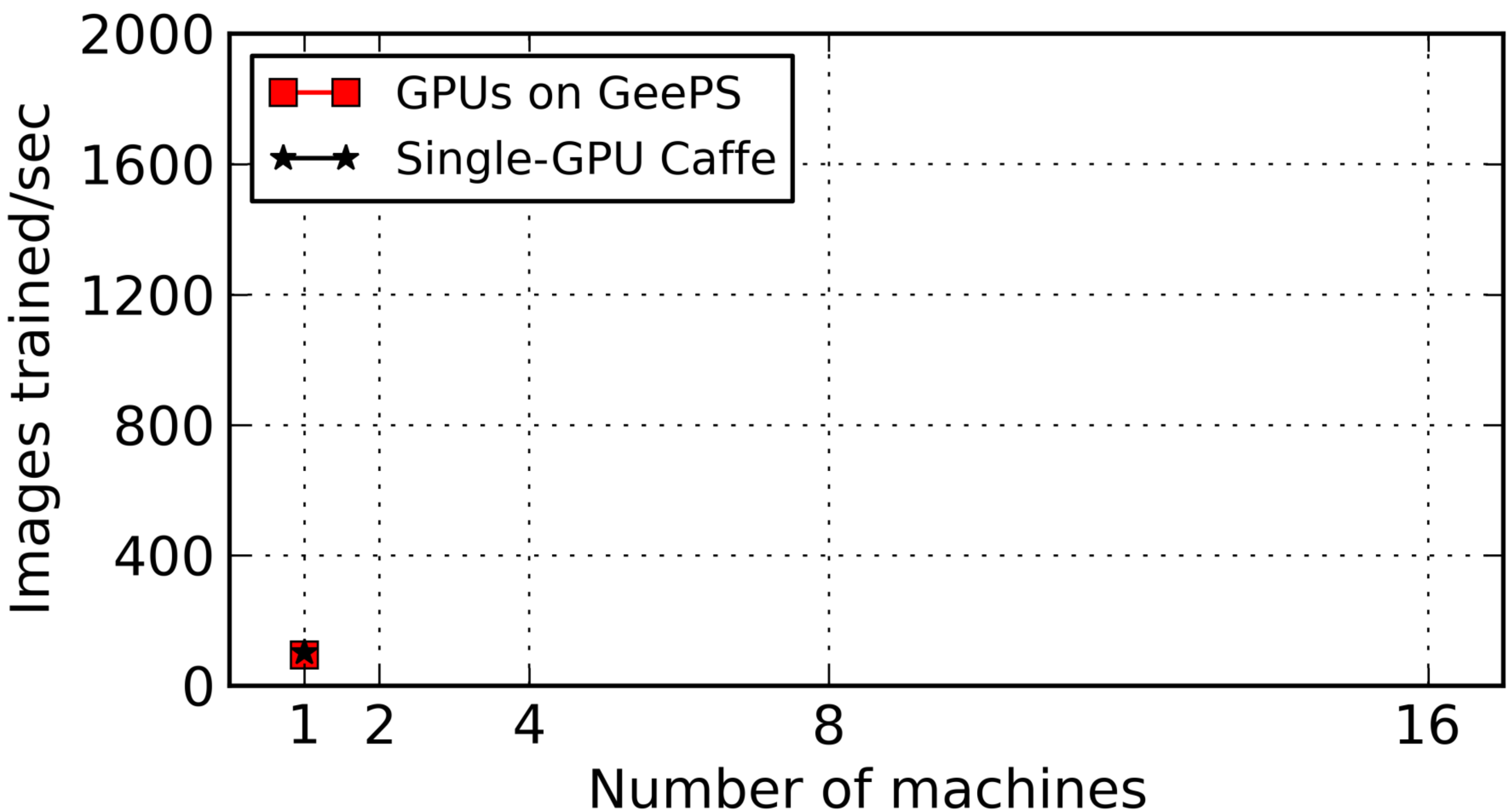
- Dataset and model
  - ImageNet: 7 million training images in 22,000 classes
  - Model: AlexNet
    - 25 layers, 2.4 billion conns
    - total memory consumption 4.5 GB

# System setups

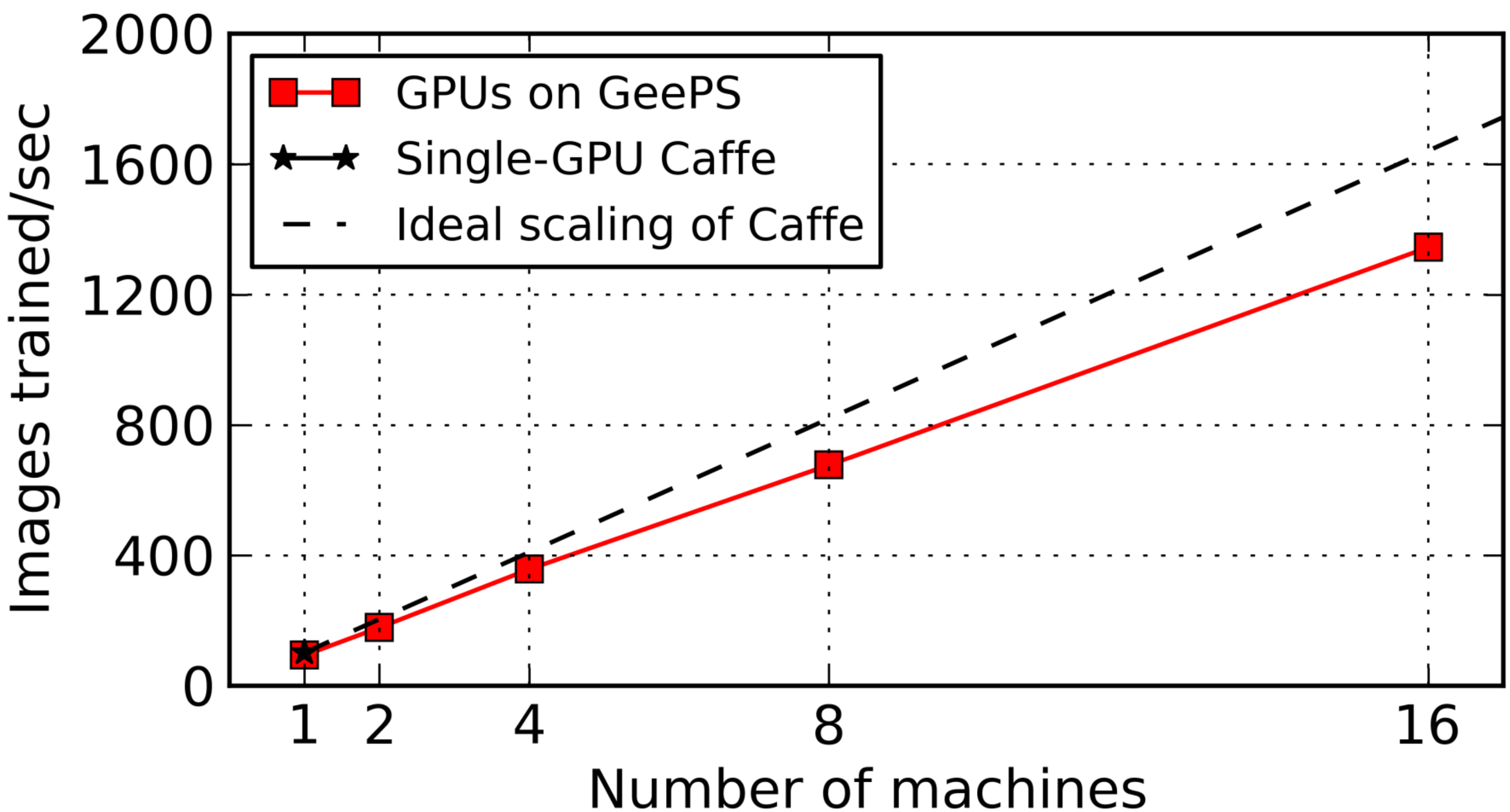
- GeePS-Caffe setups
  - Caffe: single-machine GPU deep learning system
  - GeePS-Caffe: Caffe linked with GeePS

- Baselines
  - The original unmodified Caffe
  - Caffe linked with CPU-based PS (IterStore [Cui SoCC'14])

# Training throughput

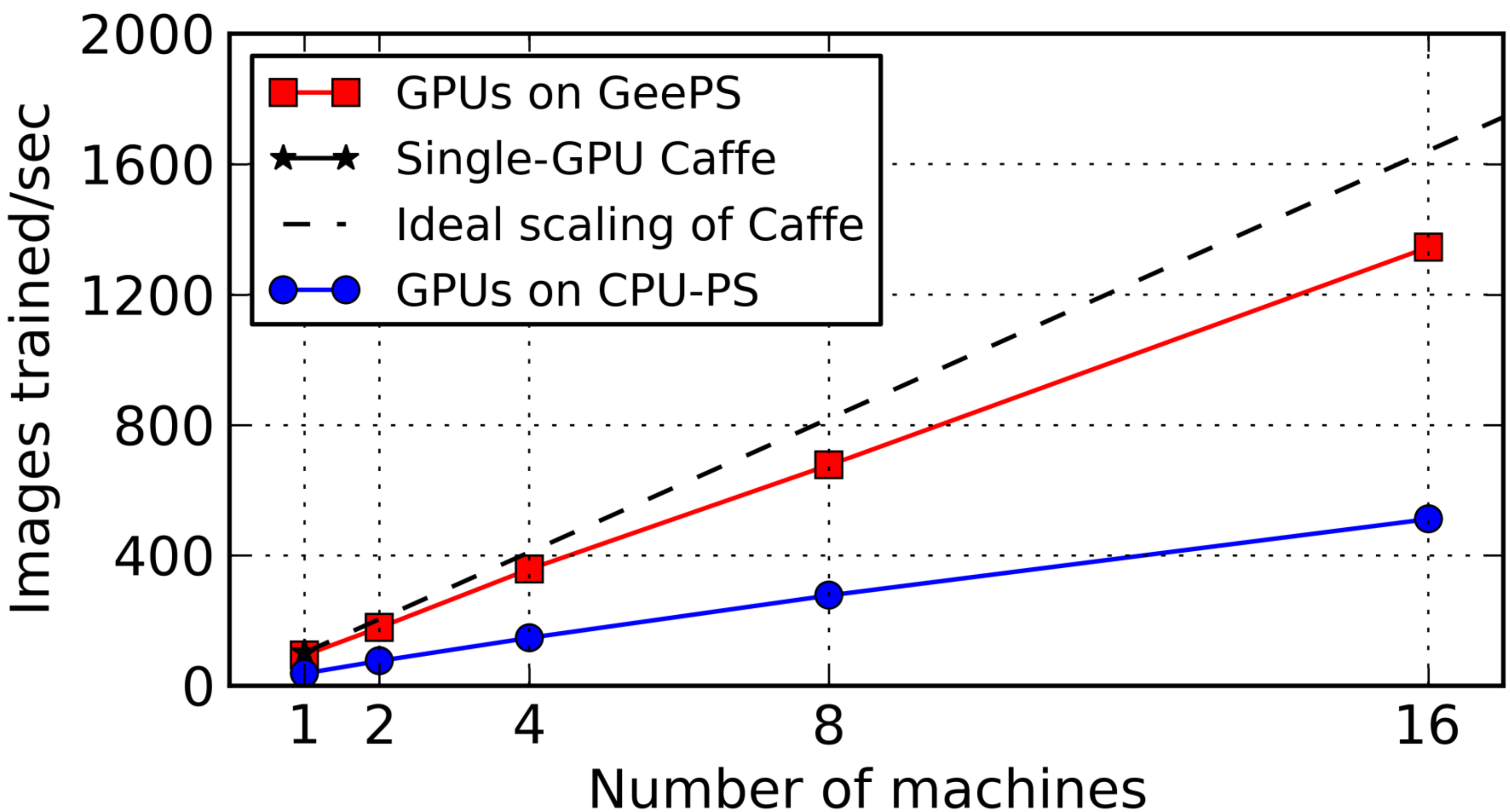


# Training throughput




- **GeePS scales close to linear with more machines**
  - **with 16 machines, it runs 13x faster than Caffe**
  - **only 8% GPU stall time**

# Training throughput




- **GeePS is much faster than CPU-based PS**
  - **2.6x higher throughput**
  - **reduces GPU stall time from 65% to 8%**

# More results in the paper

- Good scalability and convergence speed for
  - GoogLeNet network
  - RNN network for video classification

- Handle problems larger than GPU memory
  - Only 27% reduction in throughput with 35% memory
    - 3x bigger problems with little overhead
  - Handle models as large as 20 GB
  - Support 4x longer videos for video classification

# Conclusion

- GPU-specialized parameter server for GPU ML
  - 13x throughput speedup using 16 machines
  - 2x faster compared to CPU-based PS

- Managing limited GPU memory
  - By managing GPU memory inside GeePS as a cache
  - Efficiently handle problems larger than GPU memory

→ Enable use of data-parallel PS model

# References


- **[IterStore]** H. Cui, A. Tumanov, J. Wei, L. Xu, W. Dai, J. Haber-Kucharsky, Q. Ho, G. R. Ganger, P. B. Gibbons, G. A. Gibson, and E. P. Xing. Exploiting iterative-ness for parallel ML computations. In ACM SoCC, 2014.
- **[Caffe]** Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014.
- **[ImageNet]** J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE CVPR, 2009.
- **[ProjectAdam]** T. Chilimbi, Y. Suzue, J. Apacible, and K. Kalyanaraman. Project Adam: Building an efficient and scalable deep learning training system. In USENIX OSDI, 2014.

# Additional related work


- T. Chen, et al. MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274, 2015.
- H. Zhang, et al. Poseidon: A system architecture for efficient GPU-based deep learning on multiple machines. arXiv preprint arXiv:1512.06216, 2015.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.
- J. Dean, et al. Large scale distributed deep networks. In NIPS, 2012.
- C. Szegedy, et al. Going deeper with convolutions. arXiv preprint arXiv:1409.4842, 2014.
- R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun. Deep image: Scaling up image recognition. arXiv preprint arXiv:1501.02876, 2015.
- A. Coates, B. Huval, T.Wang, D.Wu, B. Catanzaro, and N. Andrew. Deep learning with COTS HPC systems. In ICML, 2013.