Heterogeneous Acceleration Pipeline for Recommendation System Training

Muhammad Adnan
Yassaman Ebrahimzadeh Maboud, Divya Mahajan, Prashant J. Nair

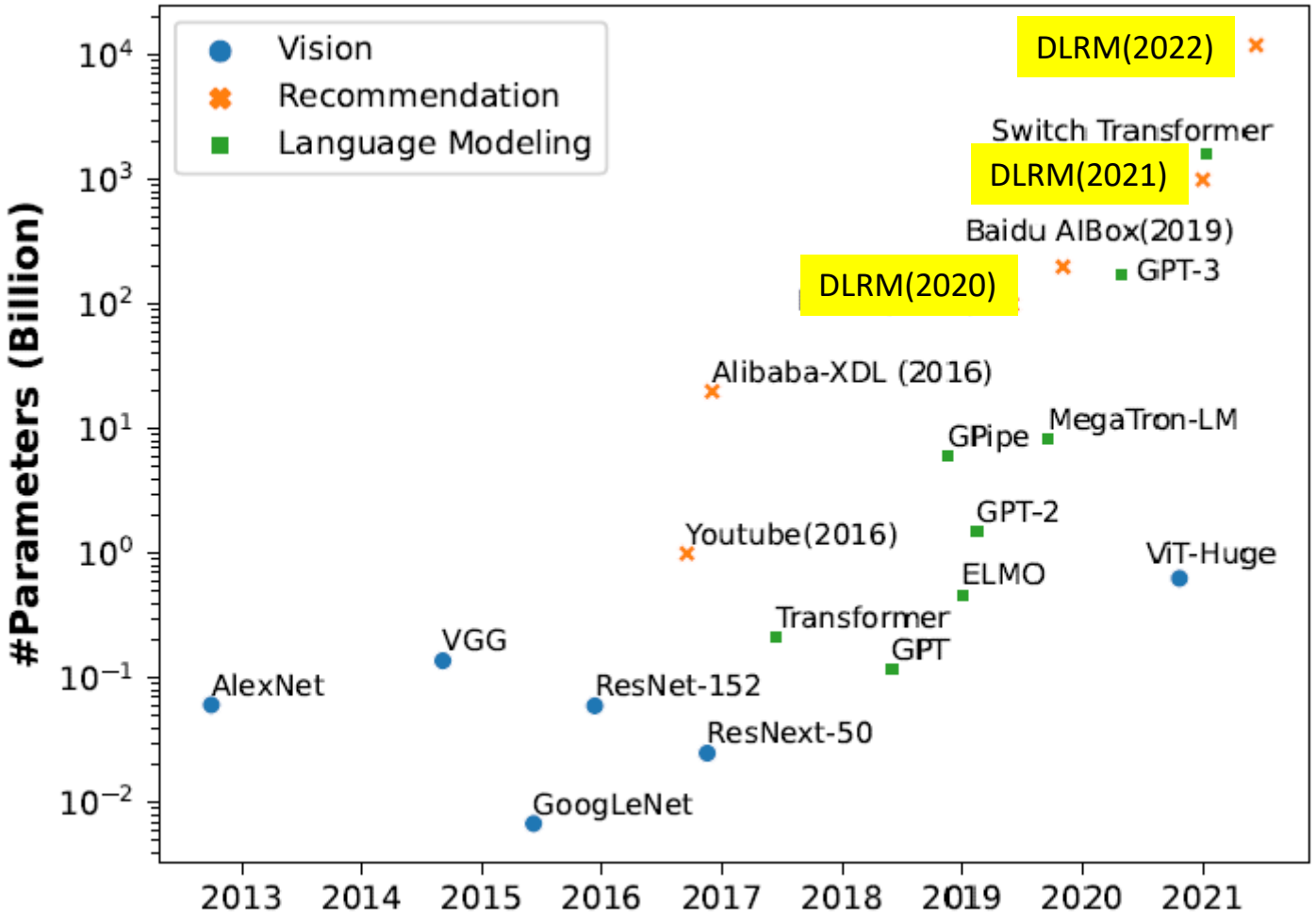
ISCA, 2024
Buenos Aires, Argentina

THE UNIVERSITY OF BRITISH COLUMBIA

Georgia Institute of Technology

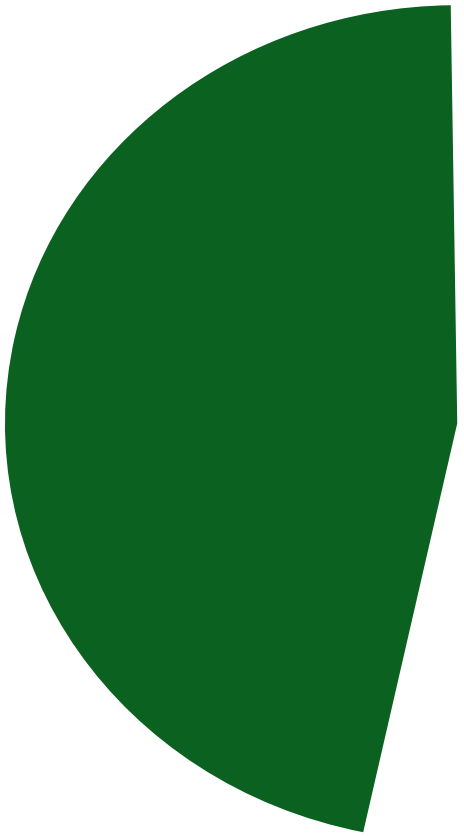
Background

Recommendation Models are getting larger

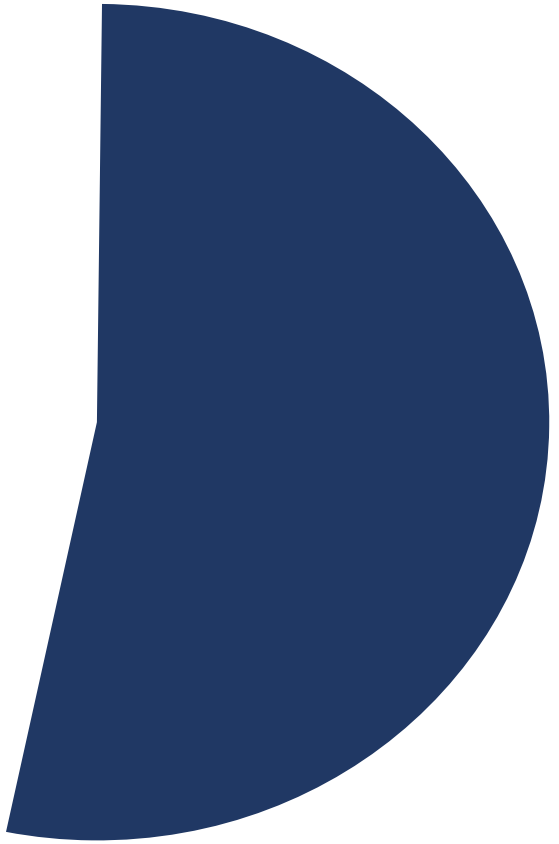


Ardalani et al. arXiv'22.

AI Infrastructure Share



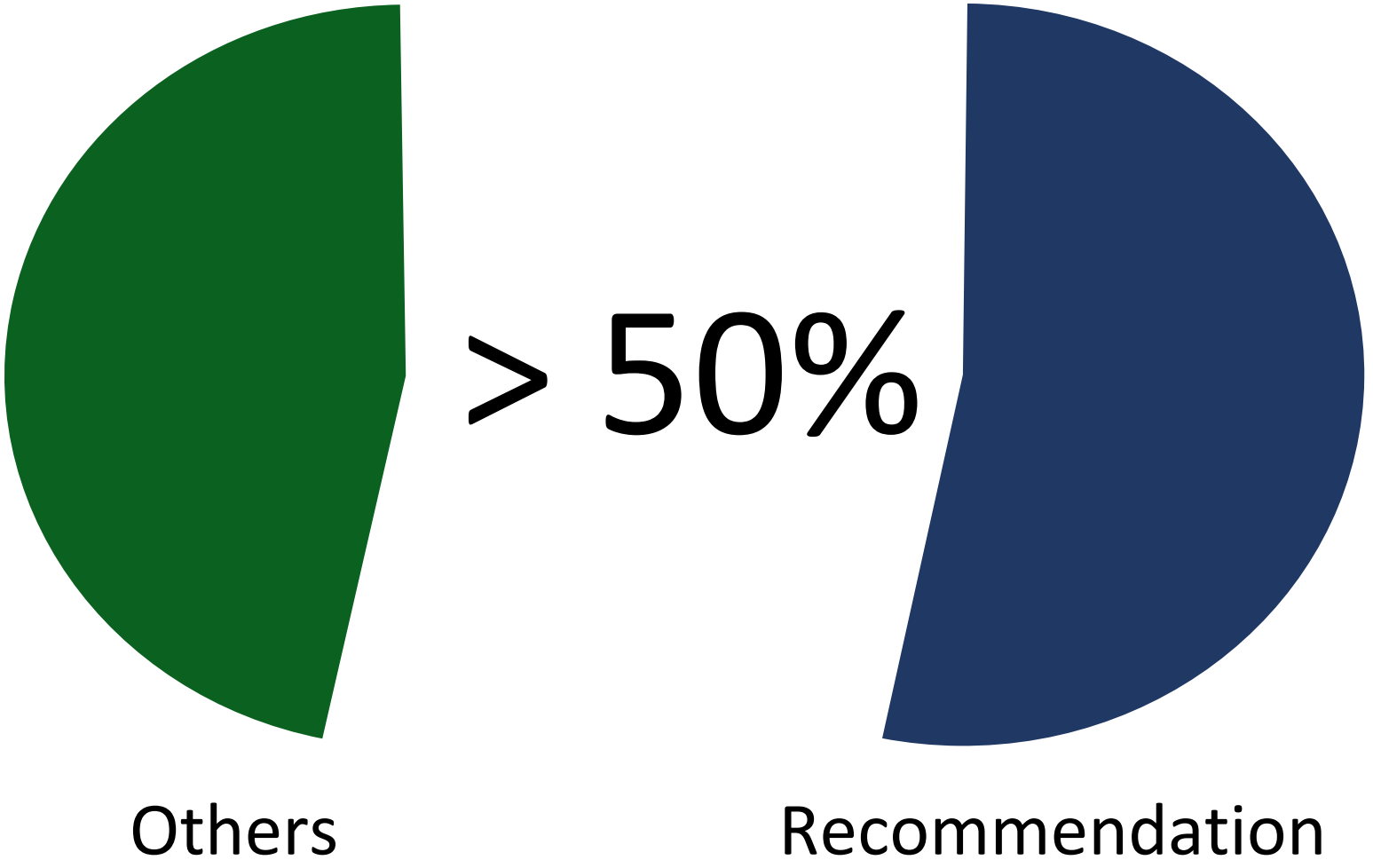
Others



Recommendation

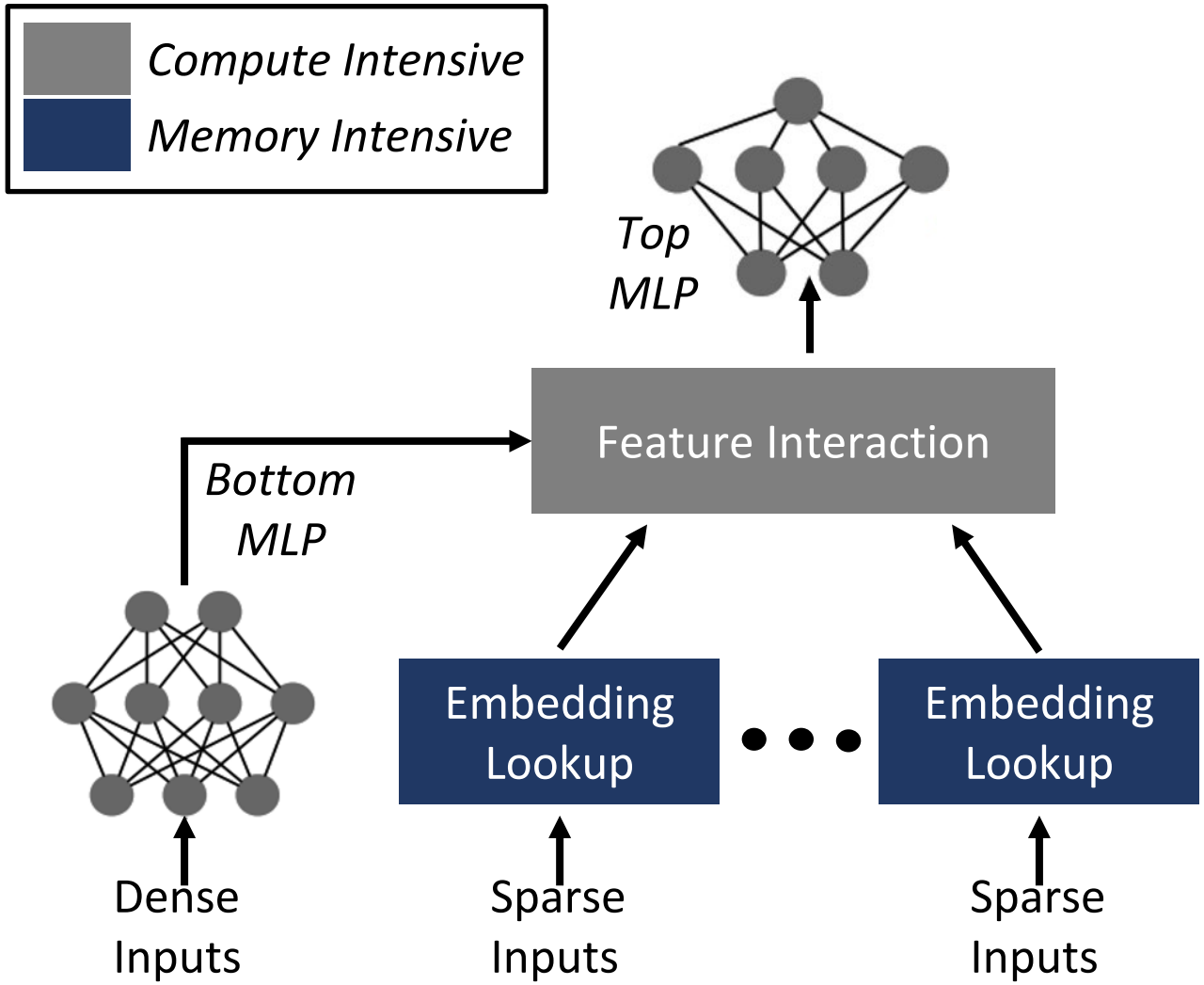
Naumov et al. arXiv'20.

AI Infrastructure Share



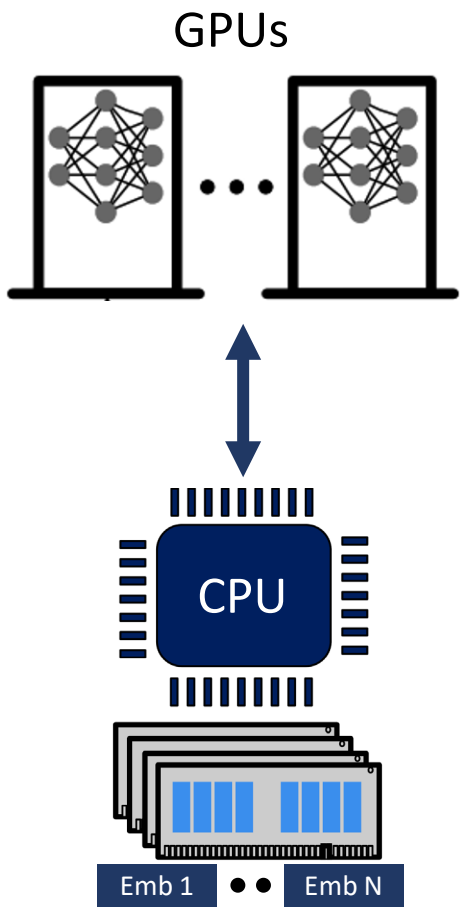
Naumov et al. arXiv'20.

DLRM – High Level Overview



DLRM – Training

Hybrid CPU-GPU



DLRM – Training

Compute Intensive

Memory Intensive

Top MLP

Feature Interaction

Bottom MLP

Embedding Lookup

Embedding Lookup

Dense Inputs

Sparse Inputs

Sparse Inputs

Hybrid CPU-GPU

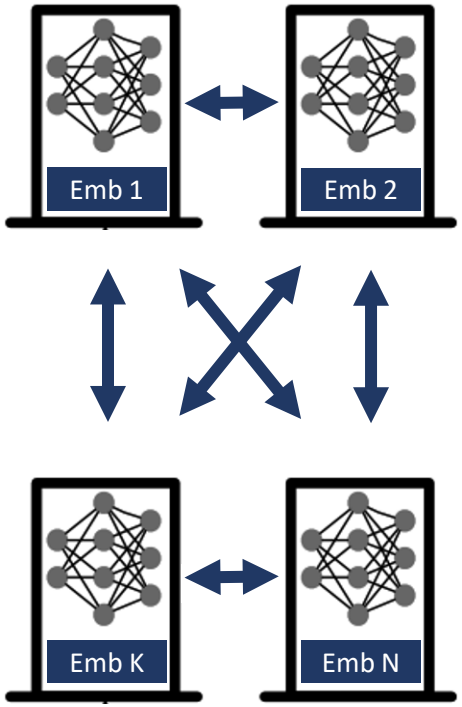
GPUs

CPU

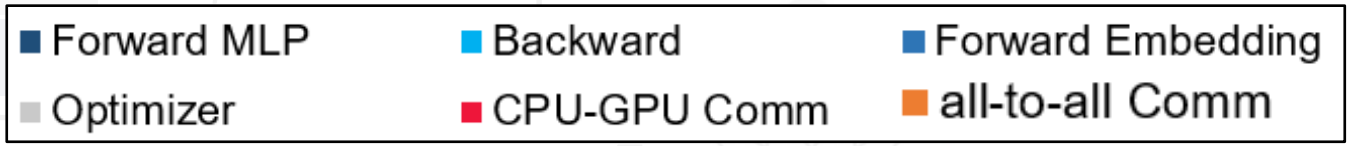
Emb 1

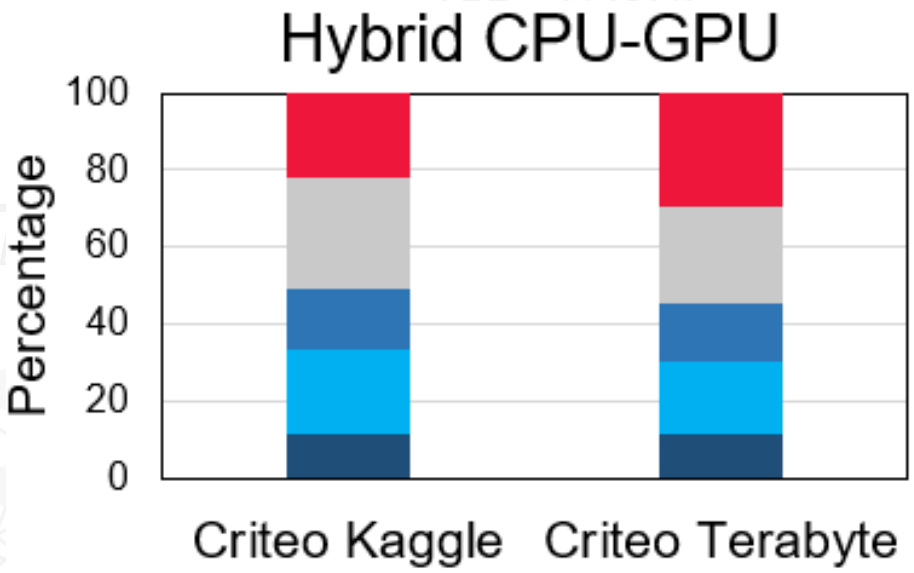
Emb N

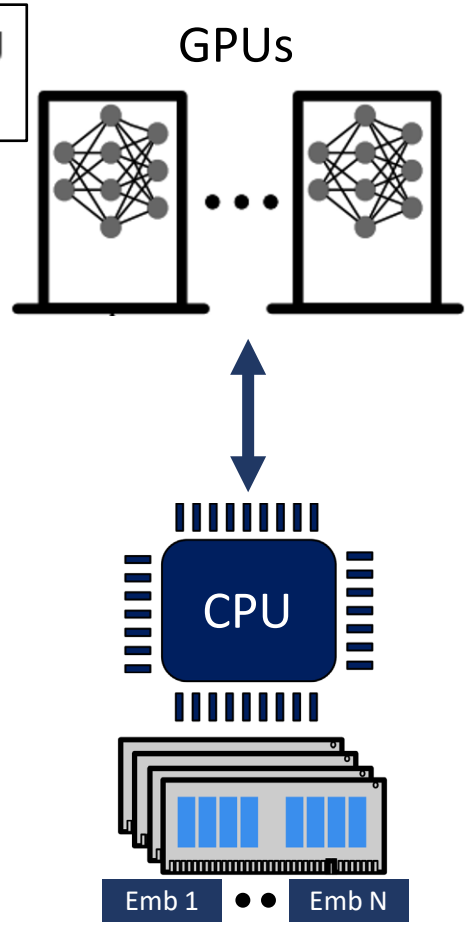
GPU-only



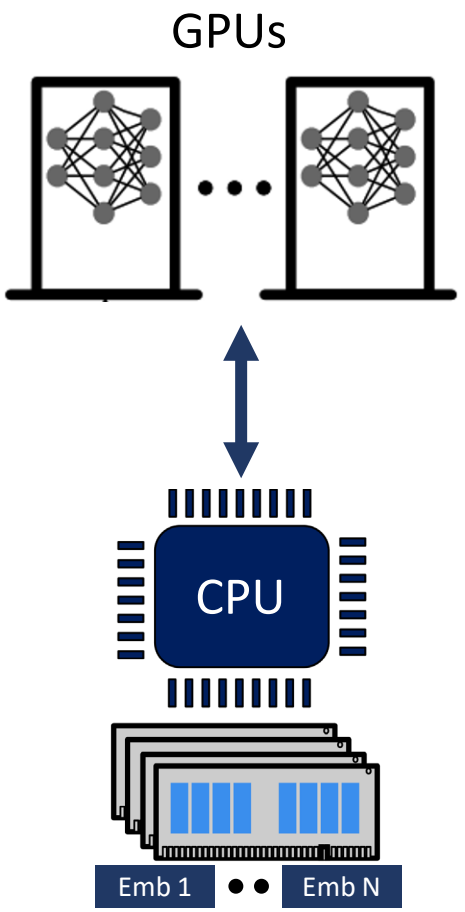
Hybrid CPU – GPU







Hybrid CPU – GPU



Embeddings' CPU-GPU Communication

30%

Embeddings' Operations

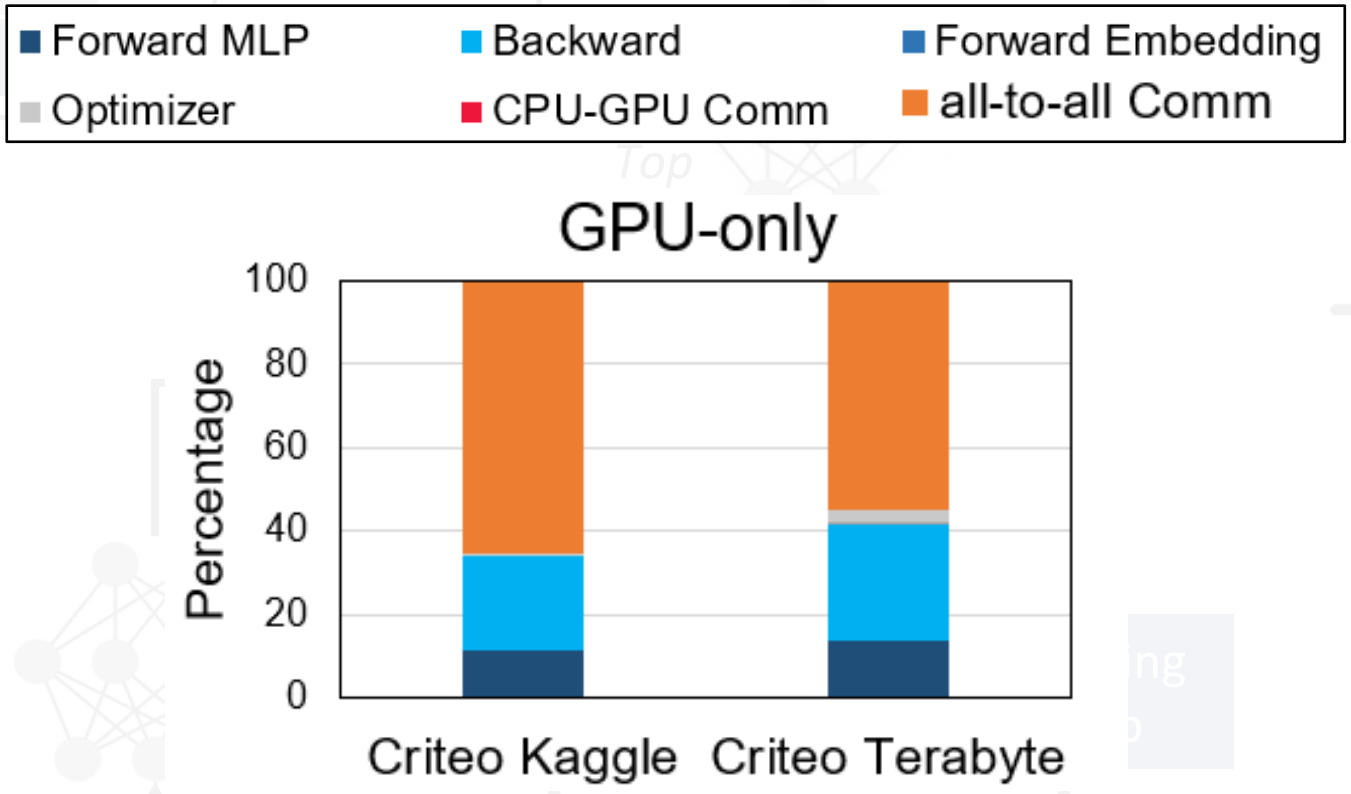
10x slower

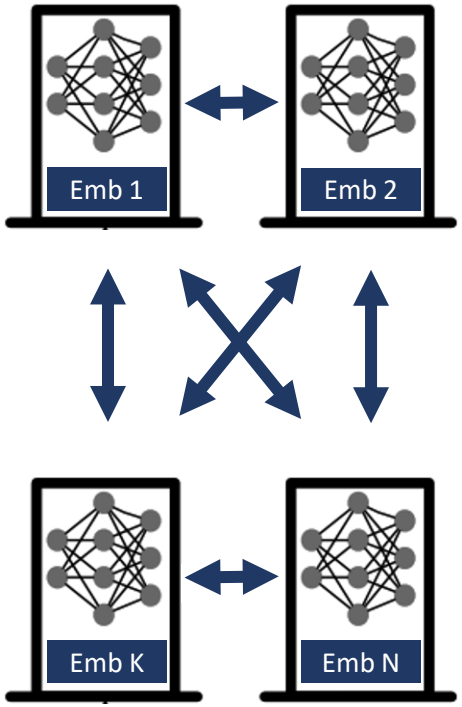
Hybrid CPU – GPU

CPU-GPU Embeddings' Communication: PCIe Bandwidth

Embeddings' Operations: CPU Main Memory Bandwidth

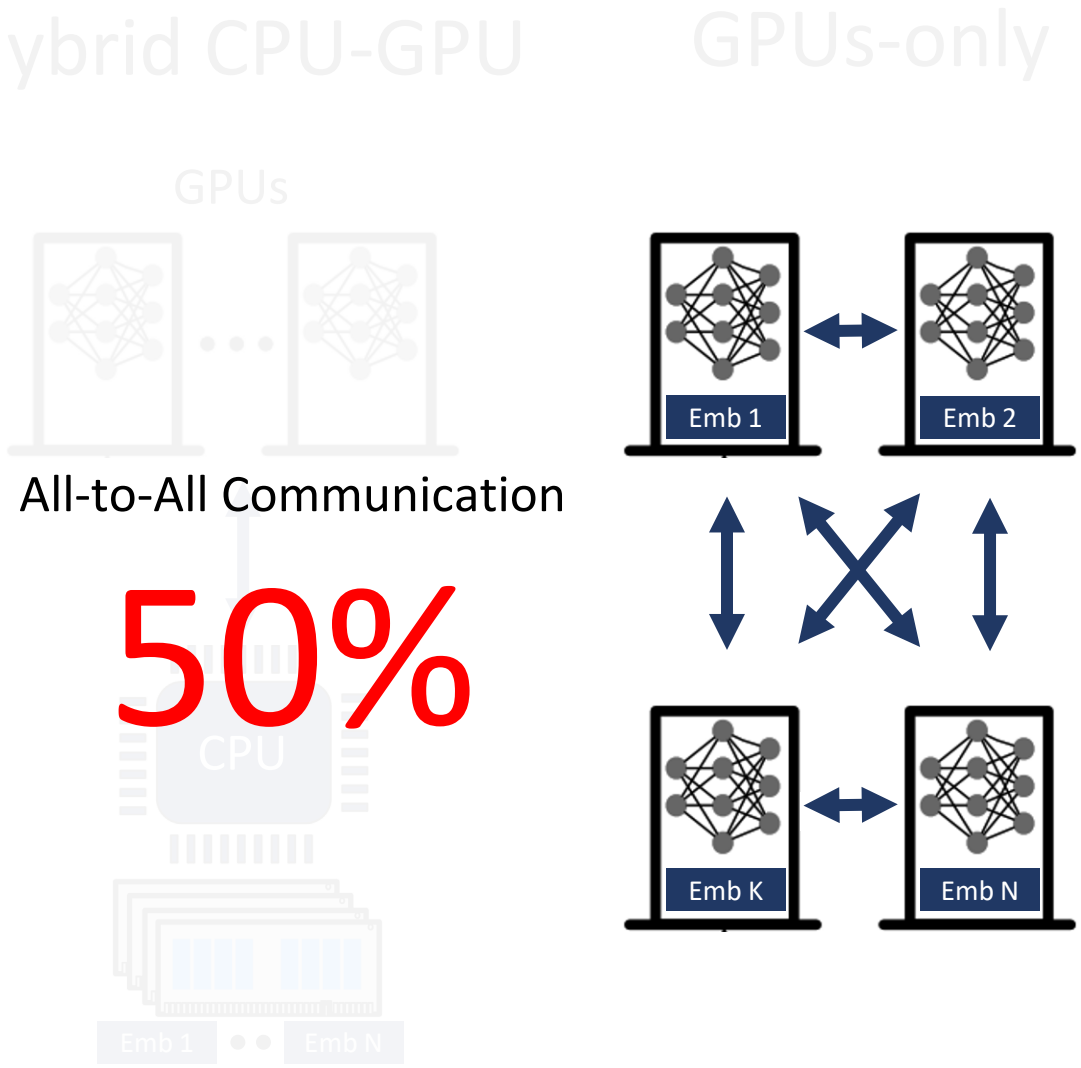
GPU – only


Forward MLP
Backward
Forward Embedding
Optimizer
CPU-GPU Comm
all-to-all Comm
GPU-only
Percentage
100
80
60
40
20
0
Criteo Kaggle
Criteo Terabyte


Emb 1
Emb 2
Emb K
Emb N

GPU – only

Compute Intensive
Memory Intensive
Top MLP
Feature Interaction
Bottom MLP
Embedding Lookup
Embedding Lookup
Dense Inputs
Sparse Inputs
Sparse Inputs


Hybrid CPU-GPU
GPUs-only
GPUs
CPU
Emb 1
Emb N
All-to-All Communication
50%
Emb 1
Emb 2
Emb K
Emb N

GPU – only

GPU Scaling to accommodate Embeddings

All to All Communication

Goals

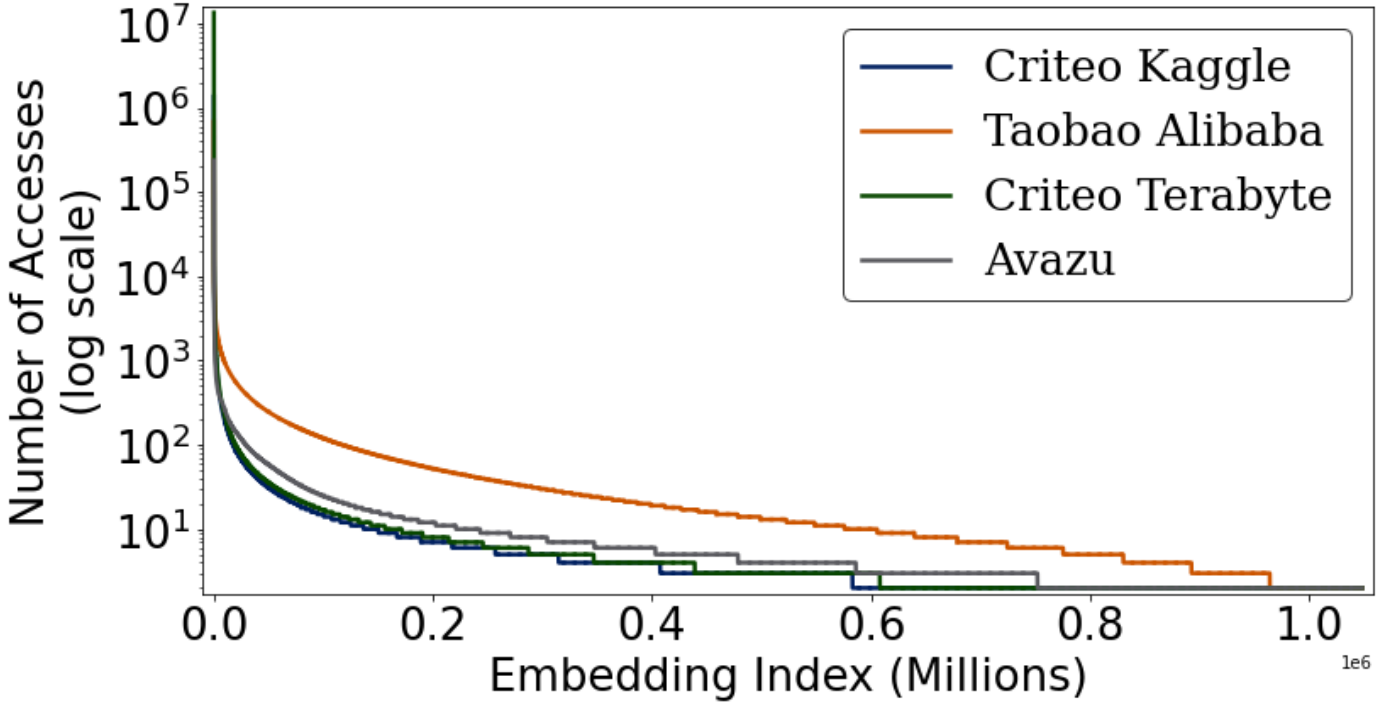
Embeddings → CPU DRAM Capacity

Embeddings → GPU HBM Performance

Minimize CPU-GPU Communication

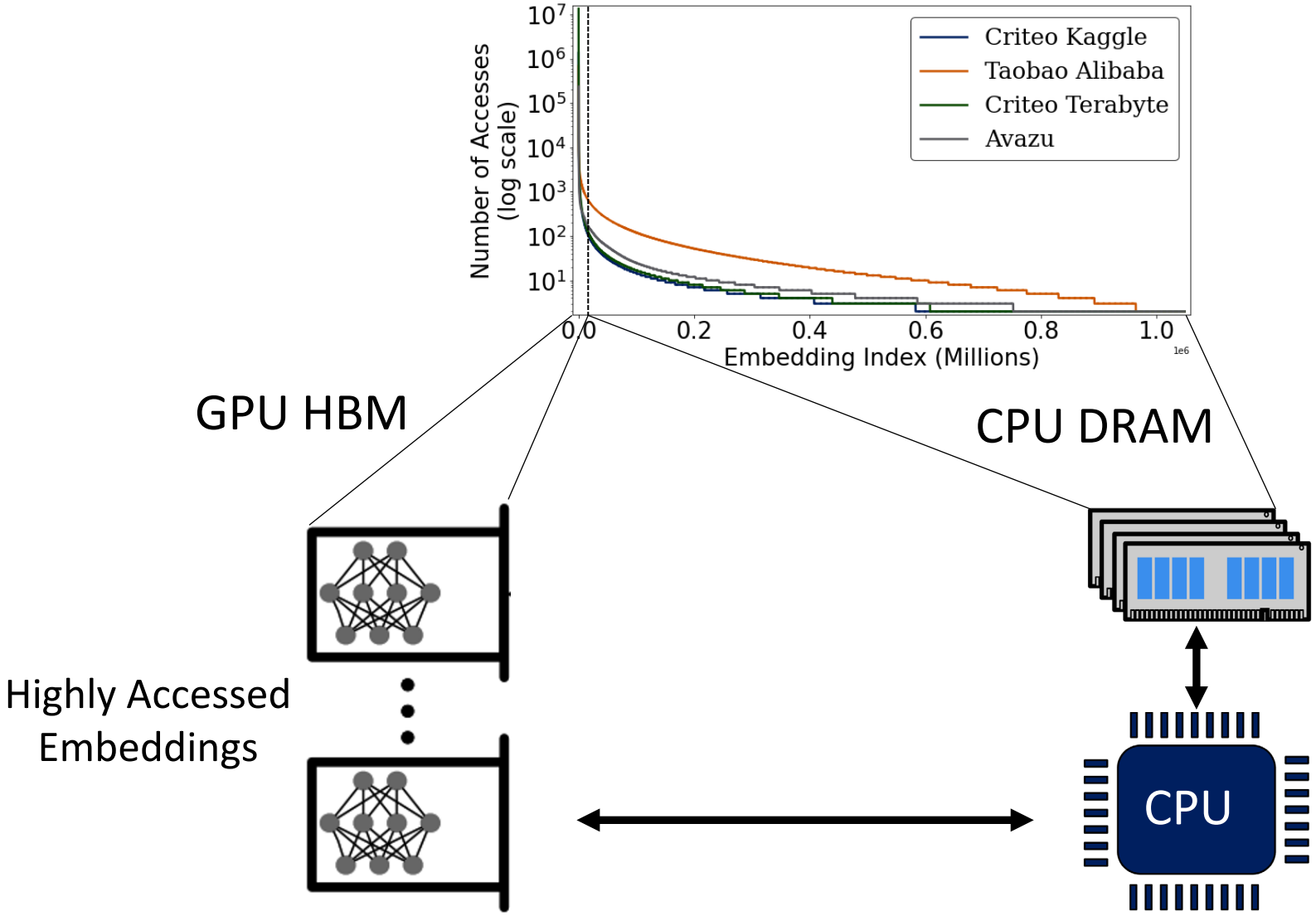
Minimize All to All Communication

Key Insight: Embedding Access Skew

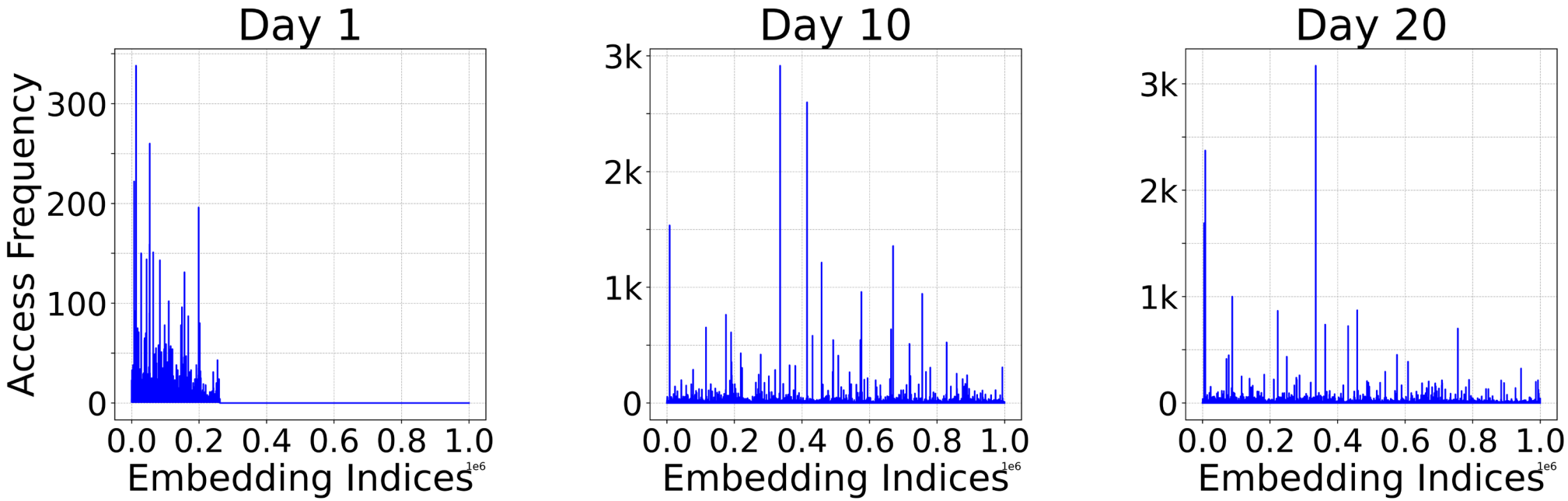


1- Adnan et al. VLDB'22
2- Sethi et al. ASPLOS'22

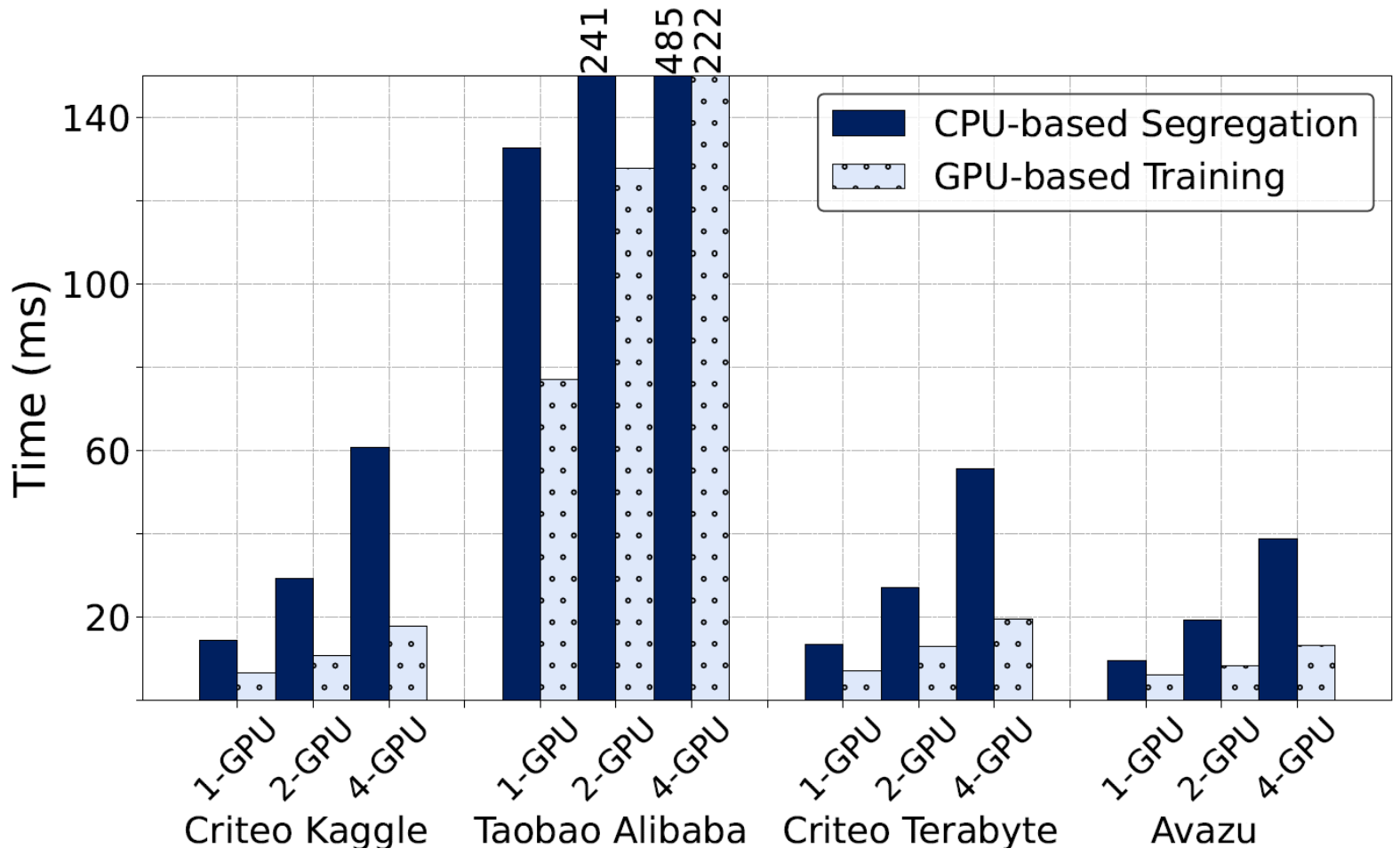
Embedding Layout across Memories



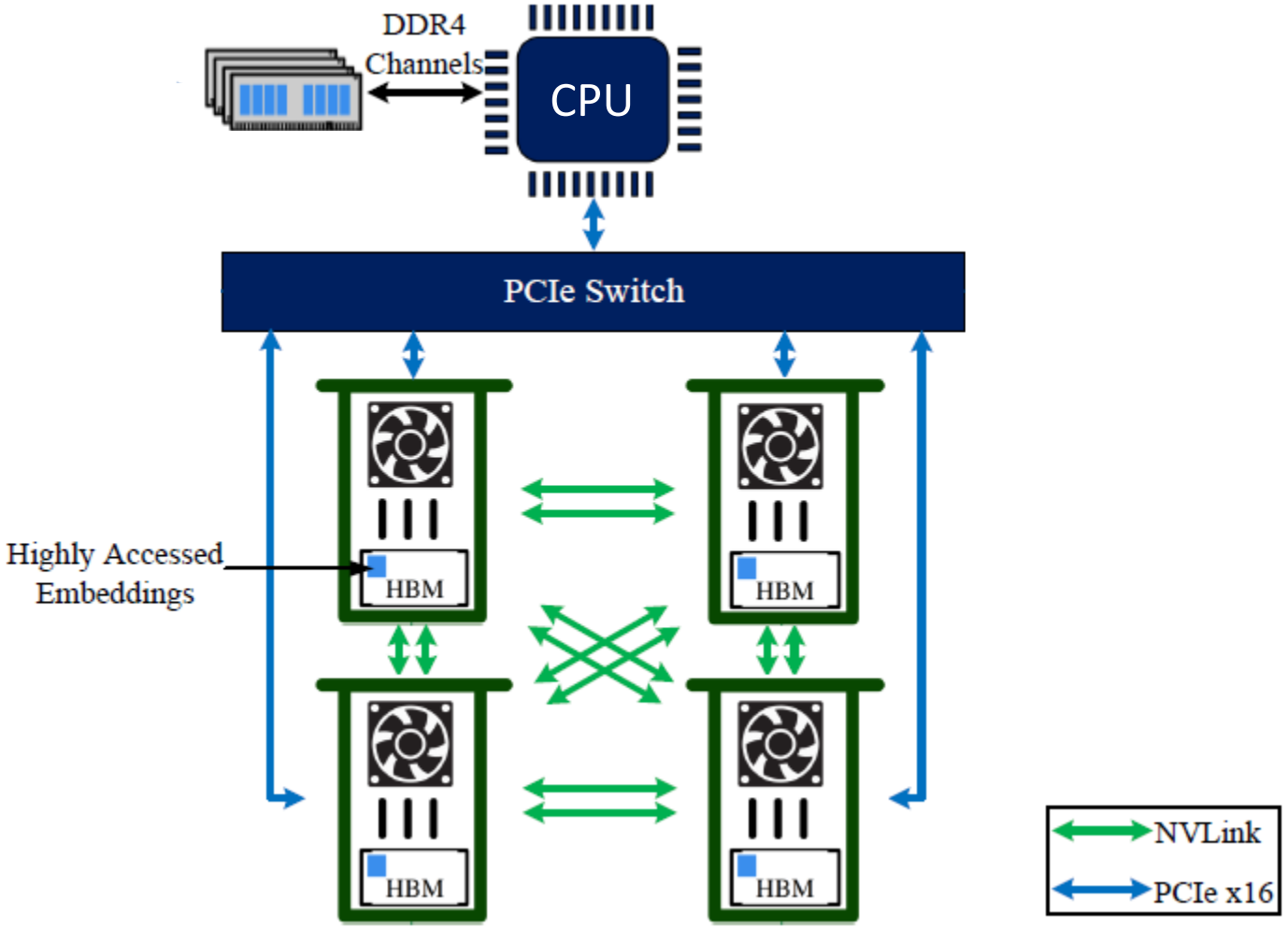
Challenge: Evolving Access Skew



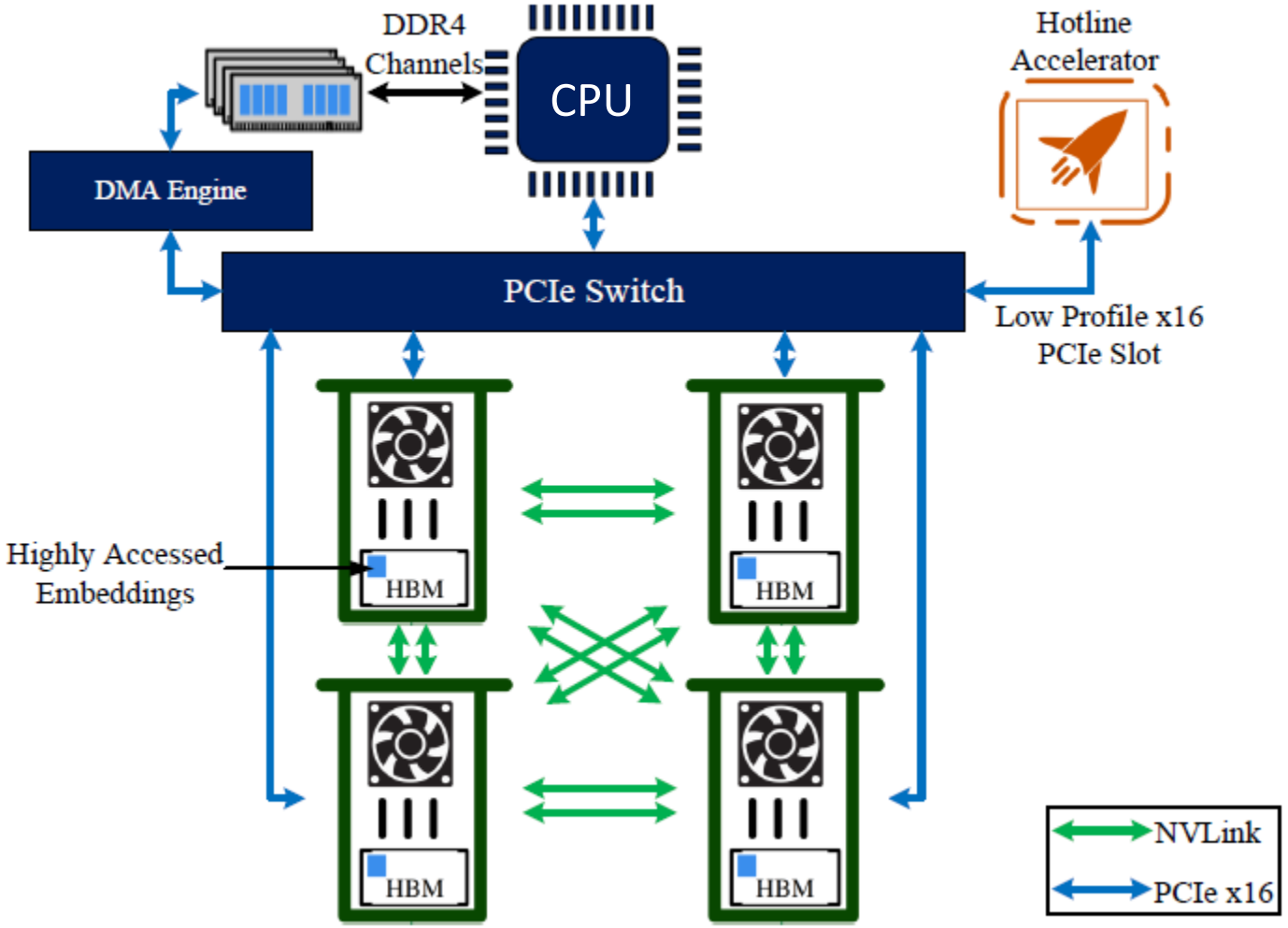
Limitation: CPU-based Pipeline Scheduler



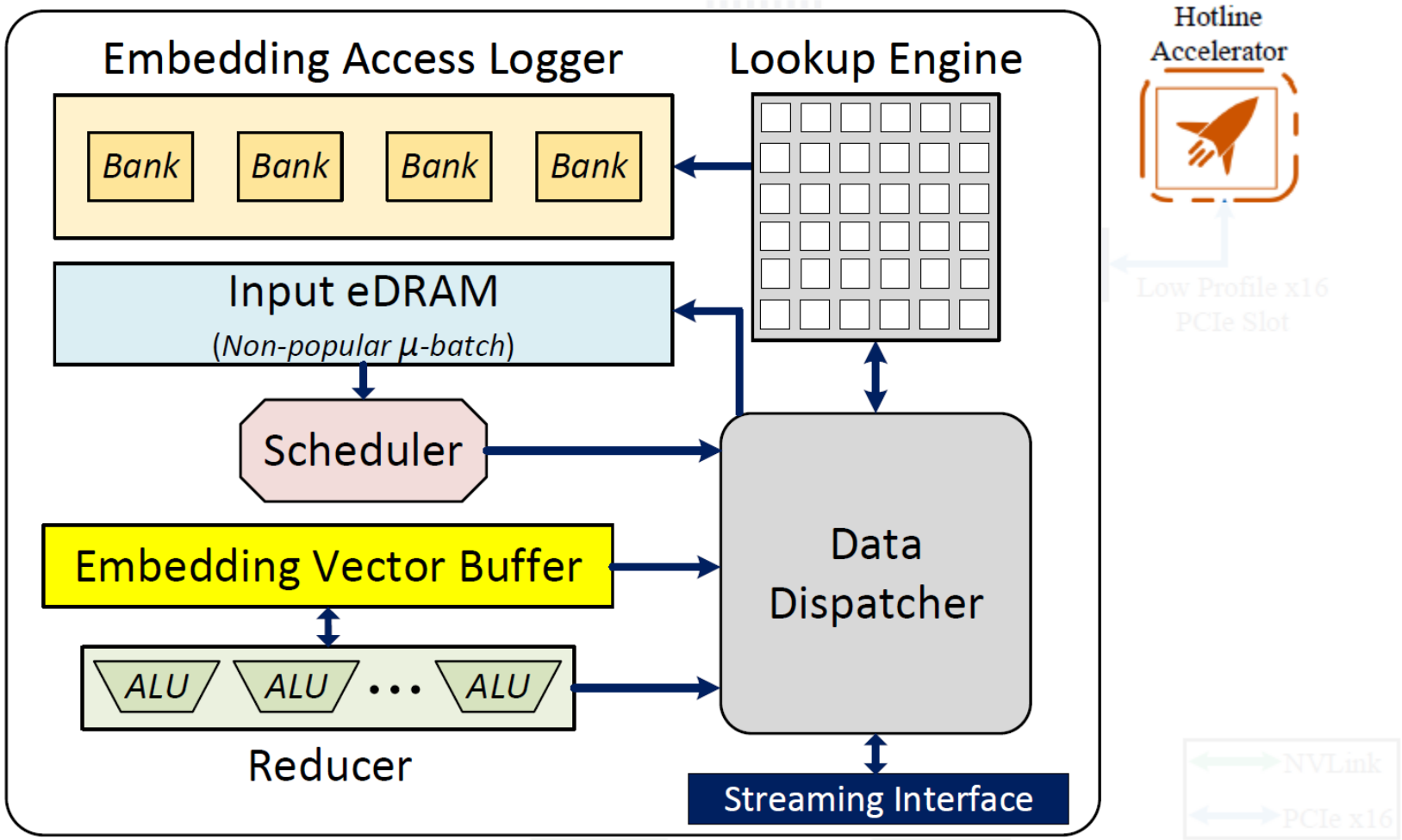
Heterogeneous Acceleration Pipeline



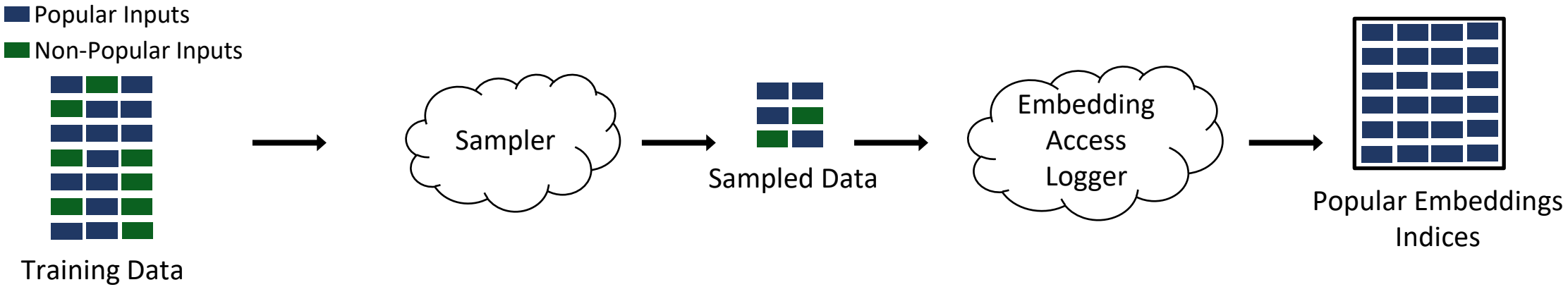
Heterogeneous Acceleration Pipeline



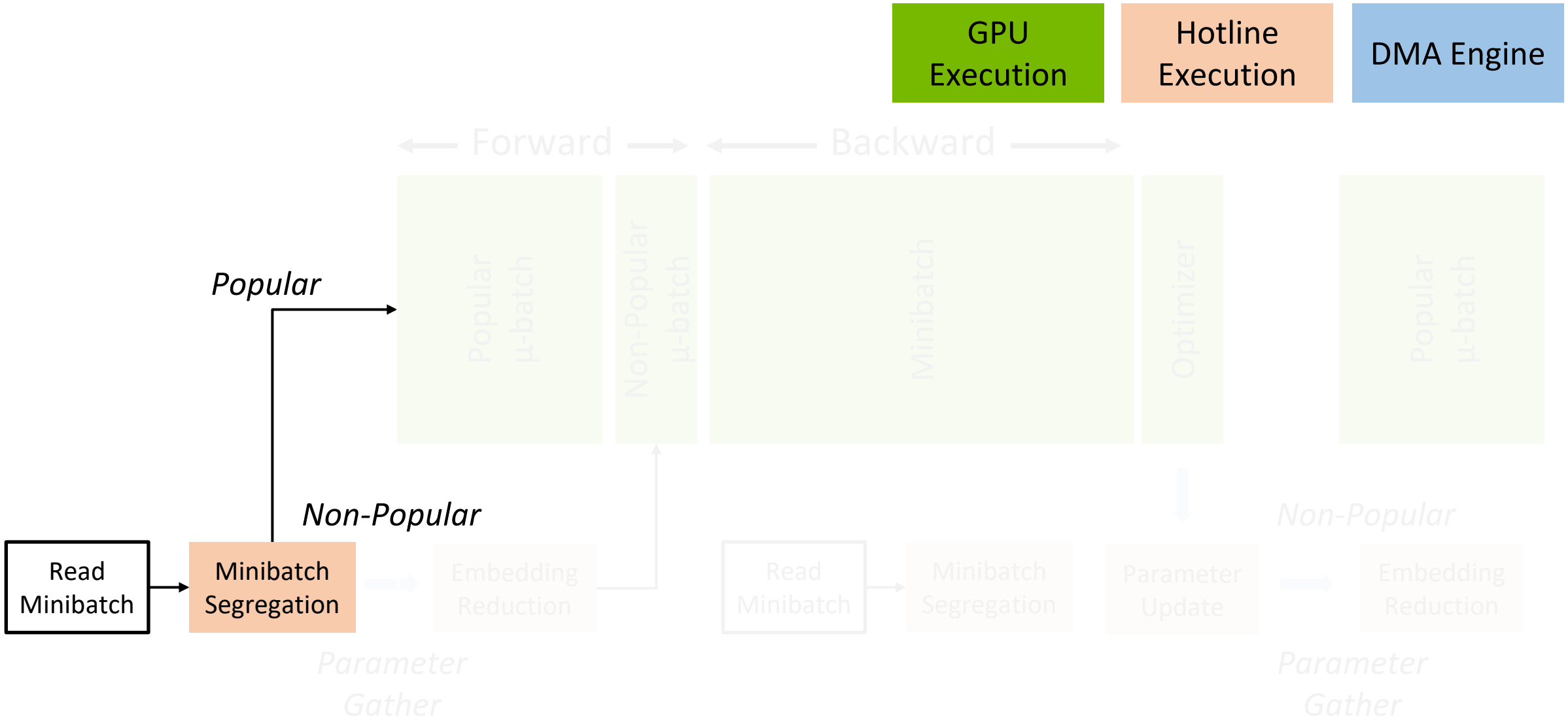
Heterogeneous Acceleration Pipeline



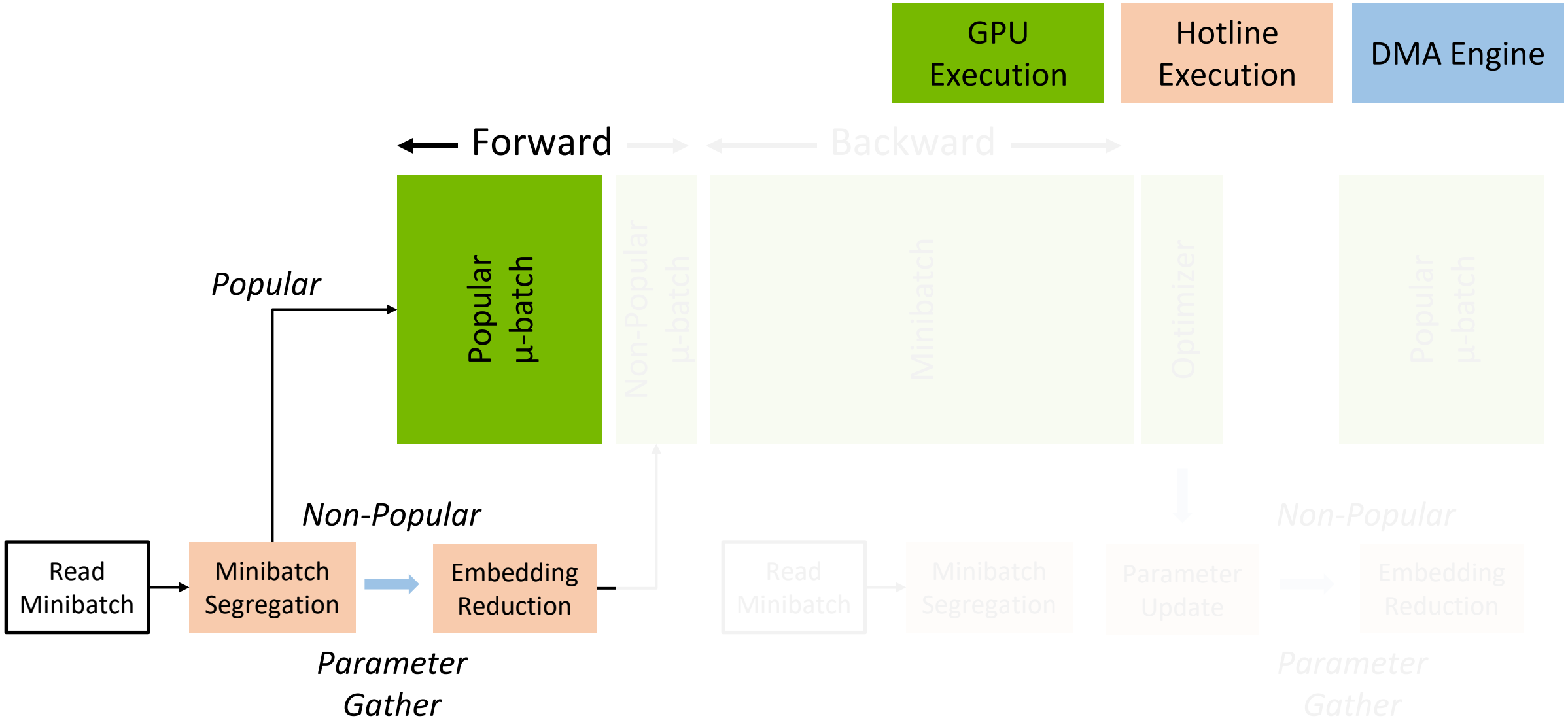
Hotline – Learning Phase



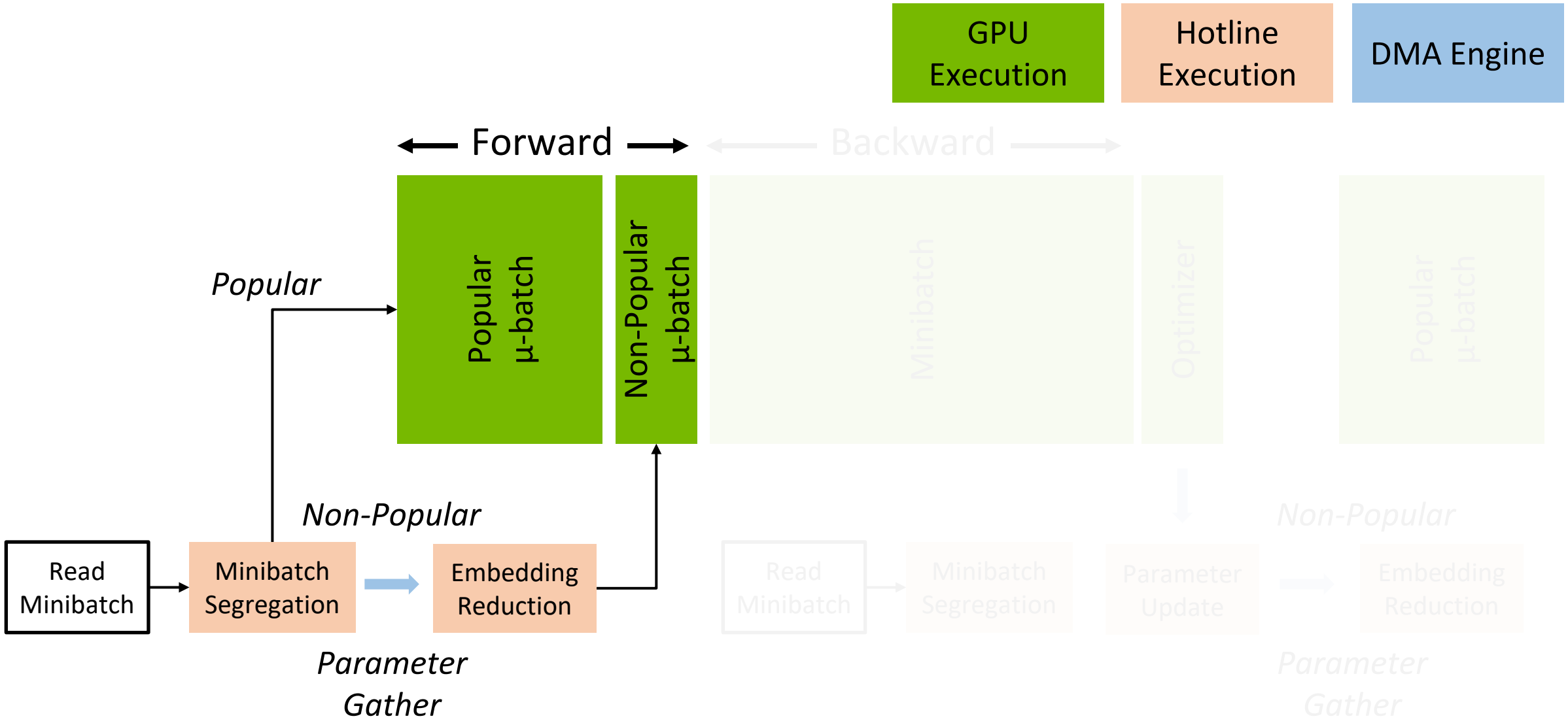
Hotline – Pipeline Scheduler



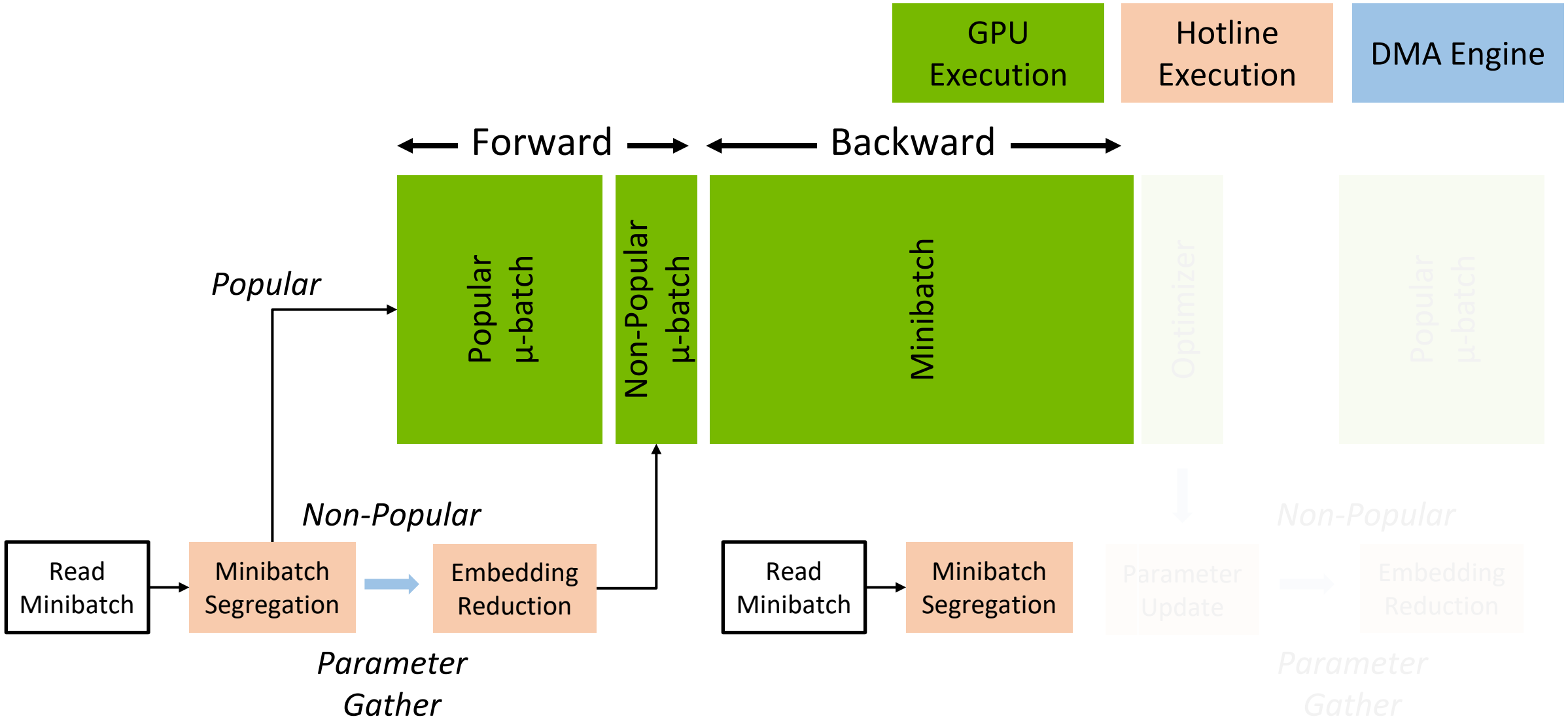
Hotline – Pipeline Scheduler



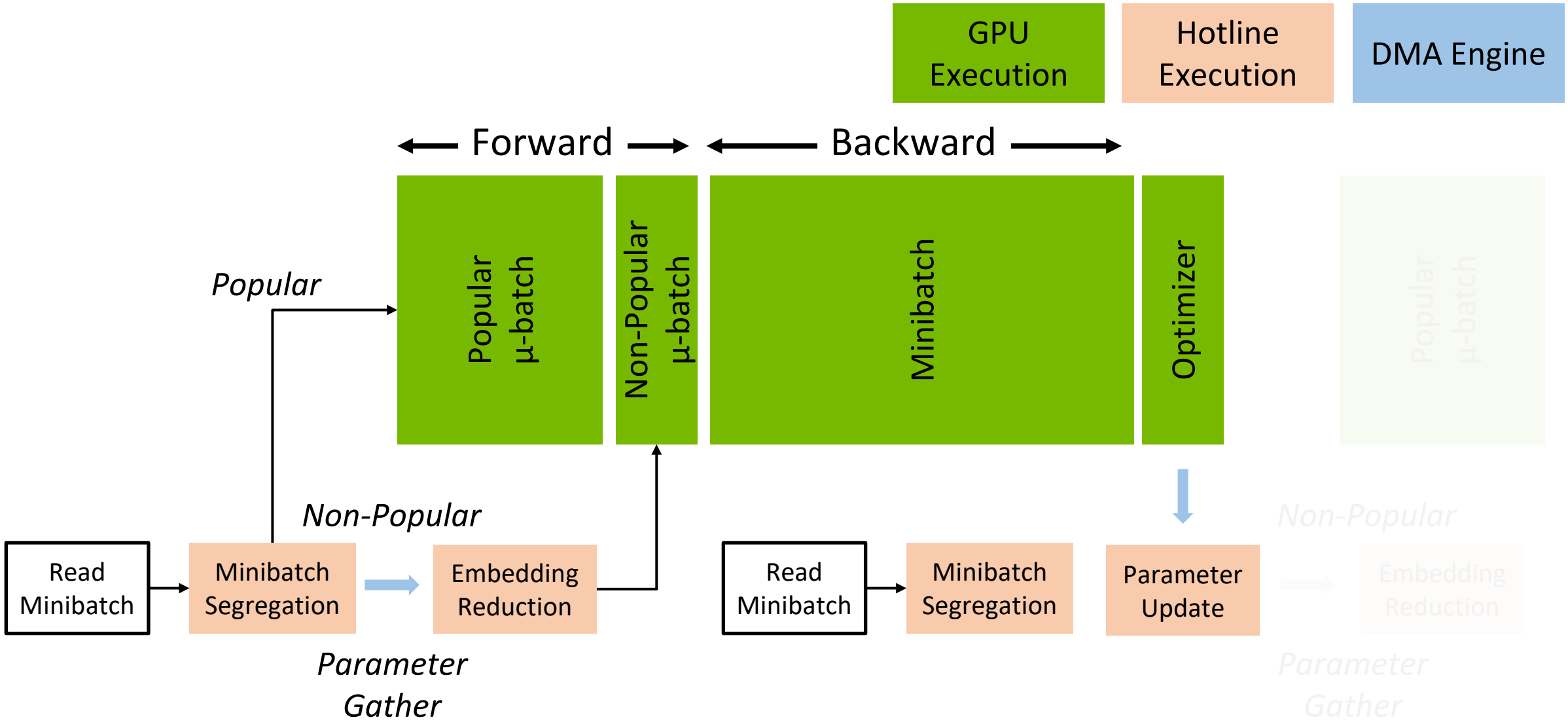
Hotline – Pipeline Scheduler



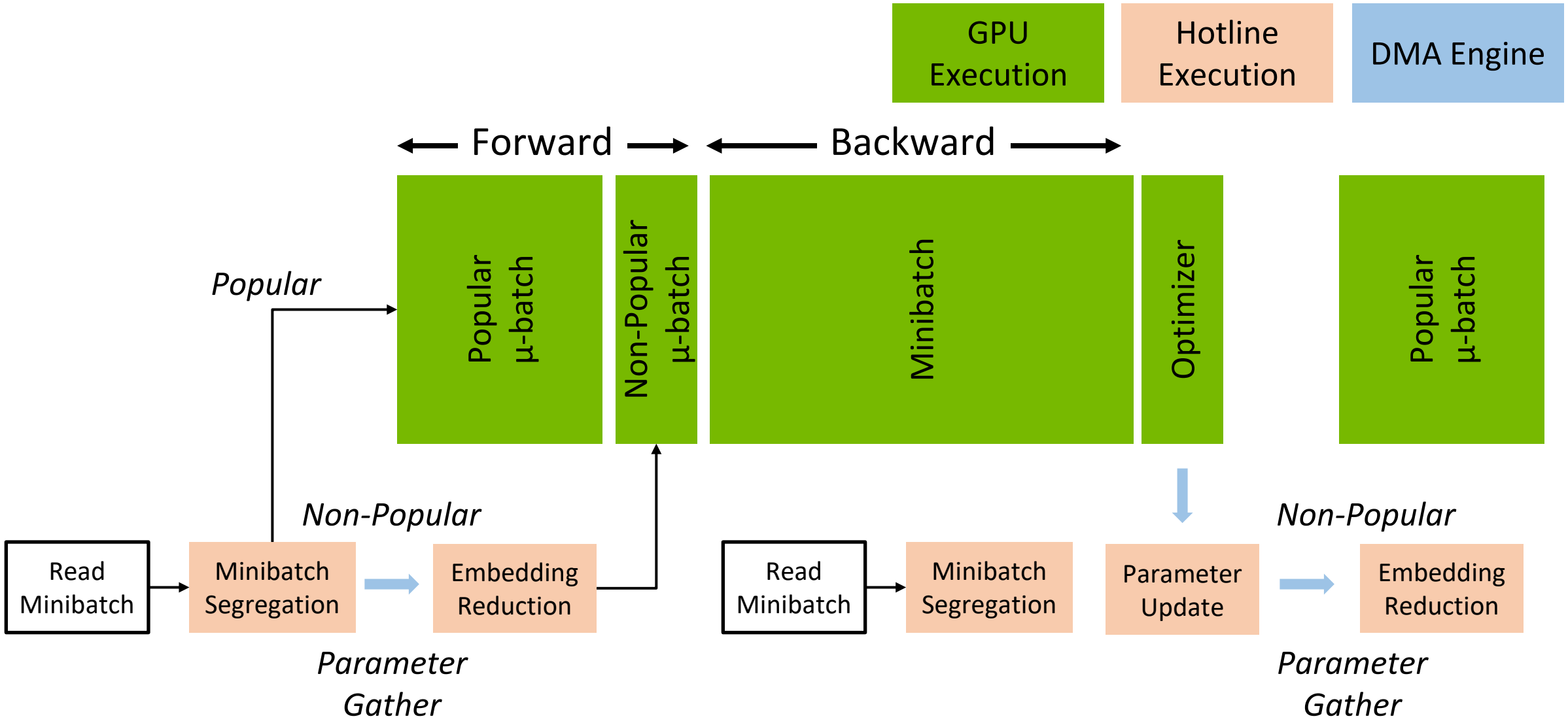
Hotline – Pipeline Scheduler



Hotline – Pipeline Scheduler



Hotline – Pipeline Scheduler



Evaluation

Baseline	XDL[1]	Open Source DLRM[2]	FAE[3]	HugeCTR[4]
Dataset	Criteo Terabyte	Criteo Kaggle	Taobao Alibaba	Avazu

1 - Jiang et al. DLP-KDD'19
2 - Naumov et al. arXiv'19
3 - Adnan et al. VLDB'22
4 - NVIDIA

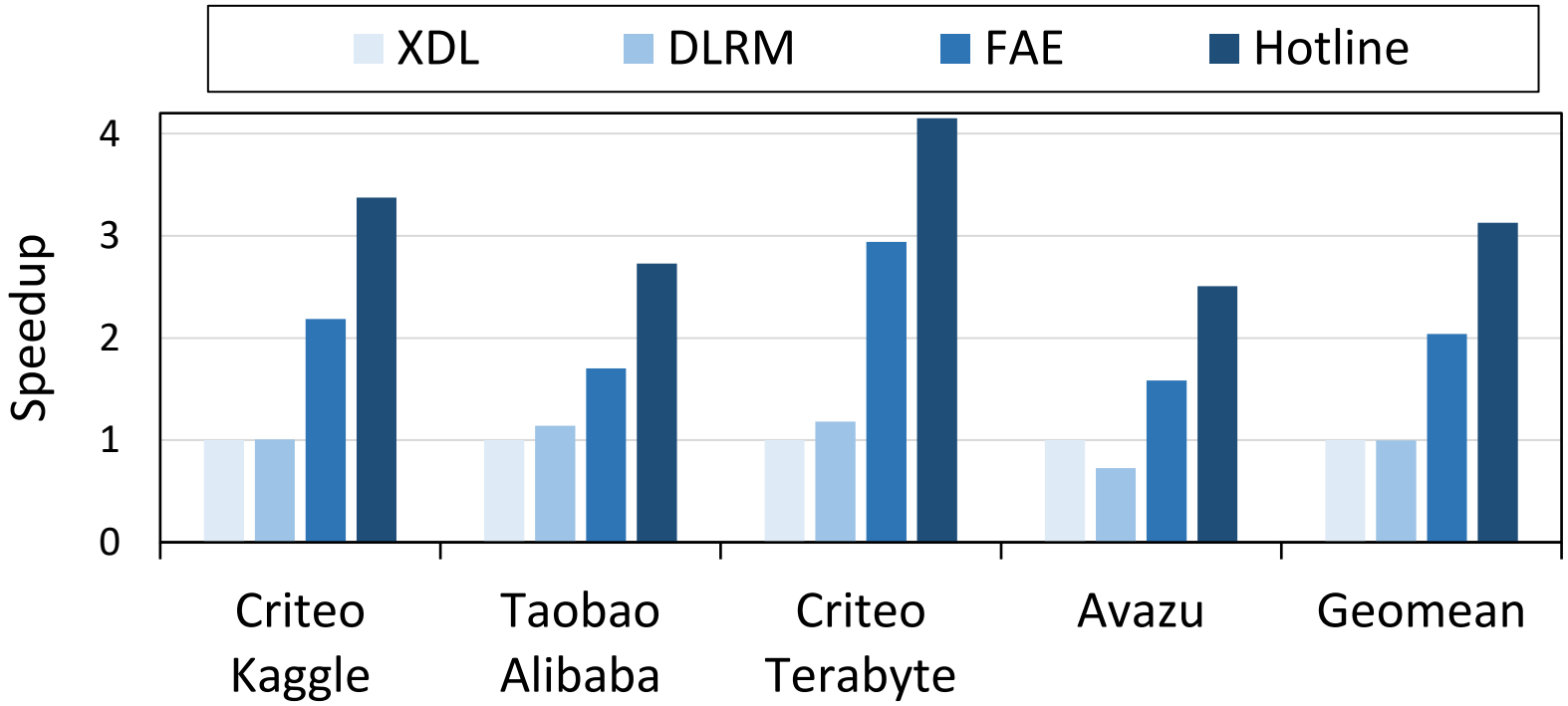
Evaluation

Baseline	XDL[1]	Open Source DLRM[2]	FAE[3]	HugeCTR[4]

Dataset	Criteo Terabyte	Criteo Kaggle	Taobao Alibaba	Avazu

1 - Jiang et al. DLP-KDD'19
2 - Naumov et al. arXiv'19
3 - Adnan et al. VLDB'22
4 - NVIDIA

Performance (Hybrid CPU-GPU)

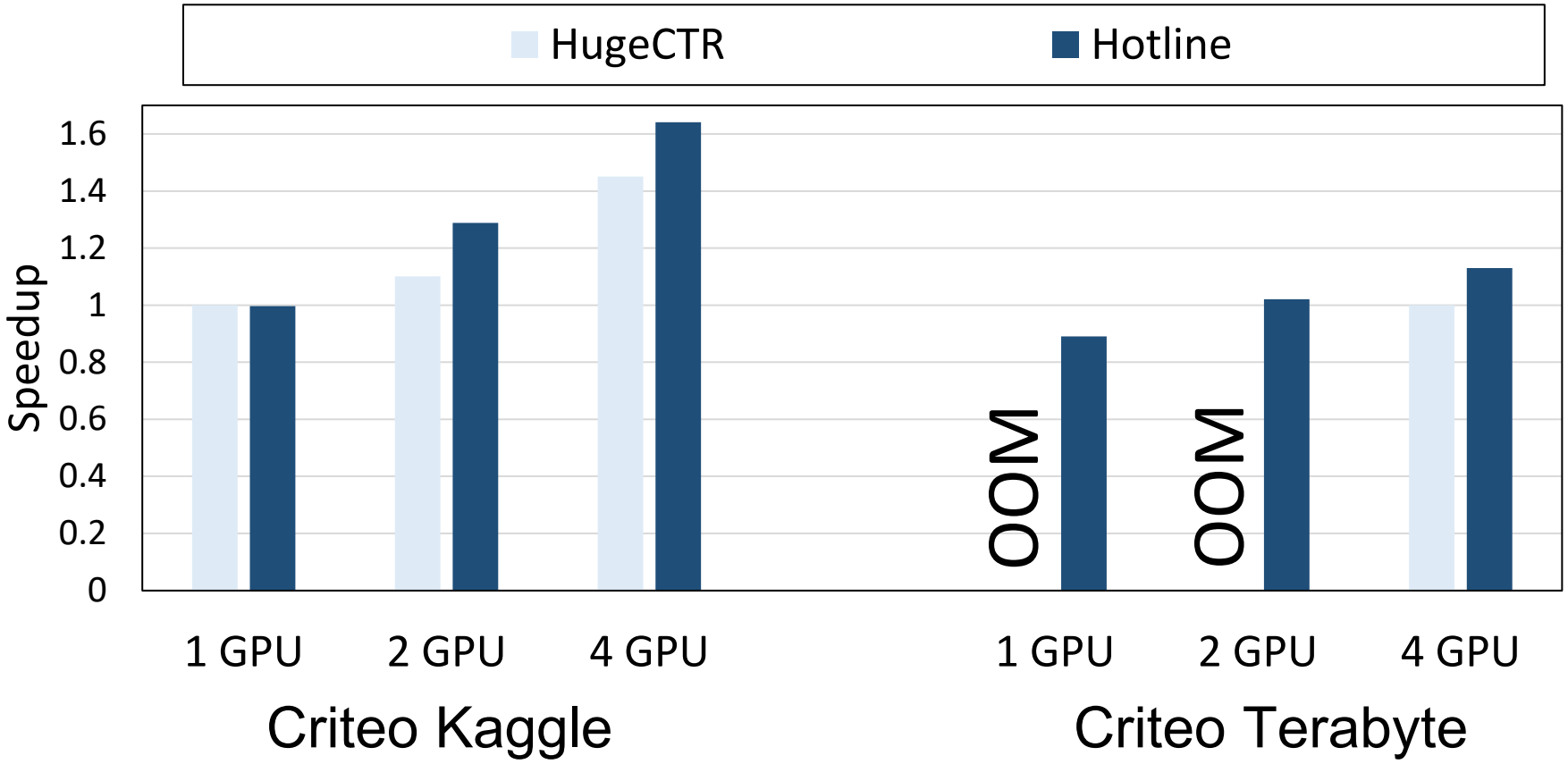


Hotline → 3.4x in comparison to XDL

Hotline → 2.2x in comparison to DLRM

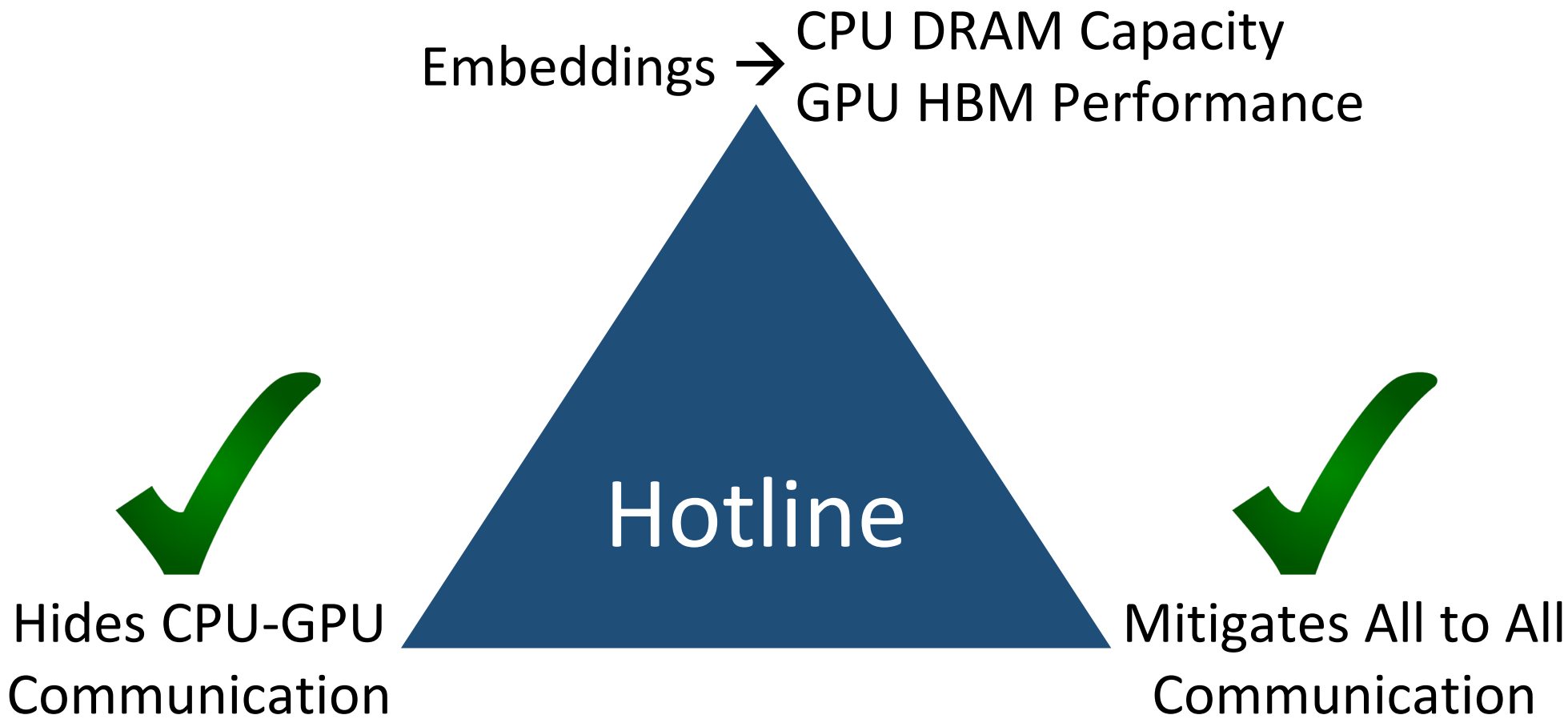
Hotline → 1.4x in comparison to FAE

Performance (GPU-only)



Hotline → 1.13x in comparison to HugeCTR

Key Takeaways



Conclusion

- Recommendation Models → Communication bottleneck
- Hybrid CPU-GPU → Embeddings CPU-GPU communication
- GPU-only → All to All communication (GPU Scaling for embeddings)
- Hotline → Data and Model-Aware Pipeline Scheduler

↓

Number of GPUs

↓2.2x

Training Time

↑2.6x

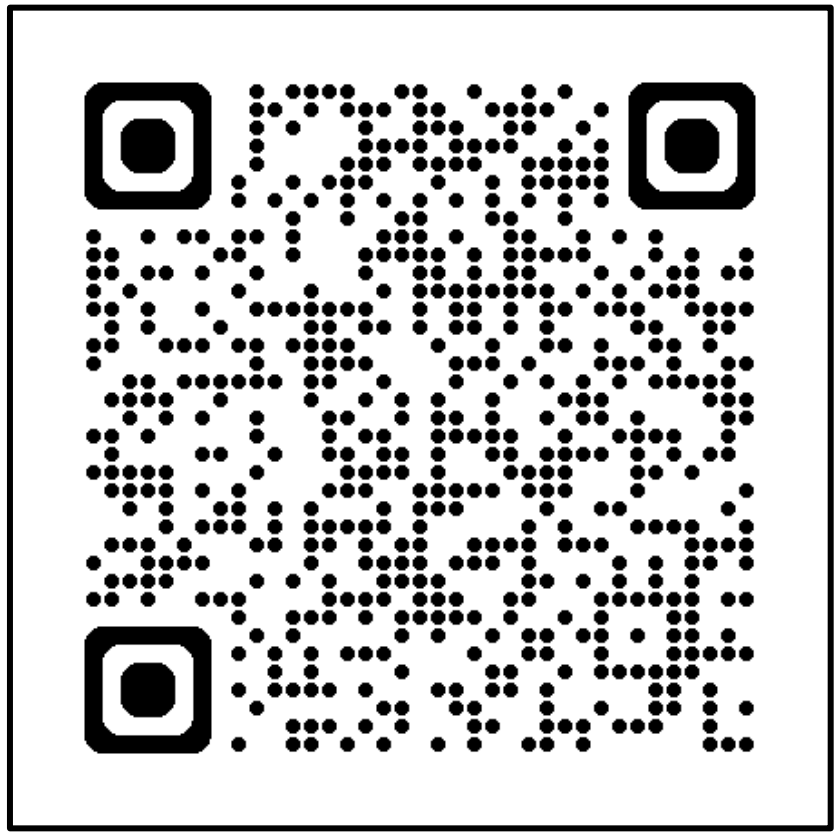
Training Throughput

Questions

adnan@ece.ubc.ca

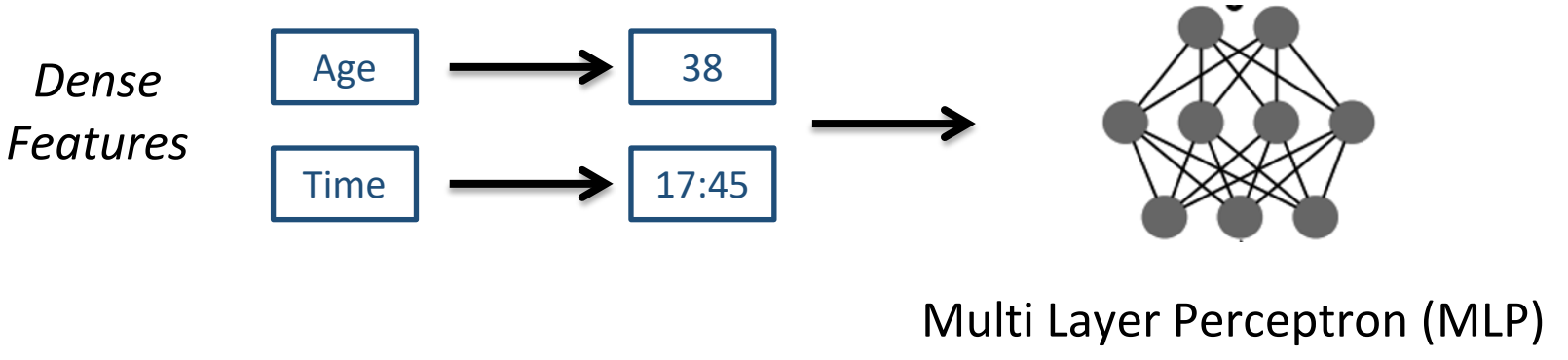
http://people.ece.ubc.ca/adnan/

adnan_tw33ts

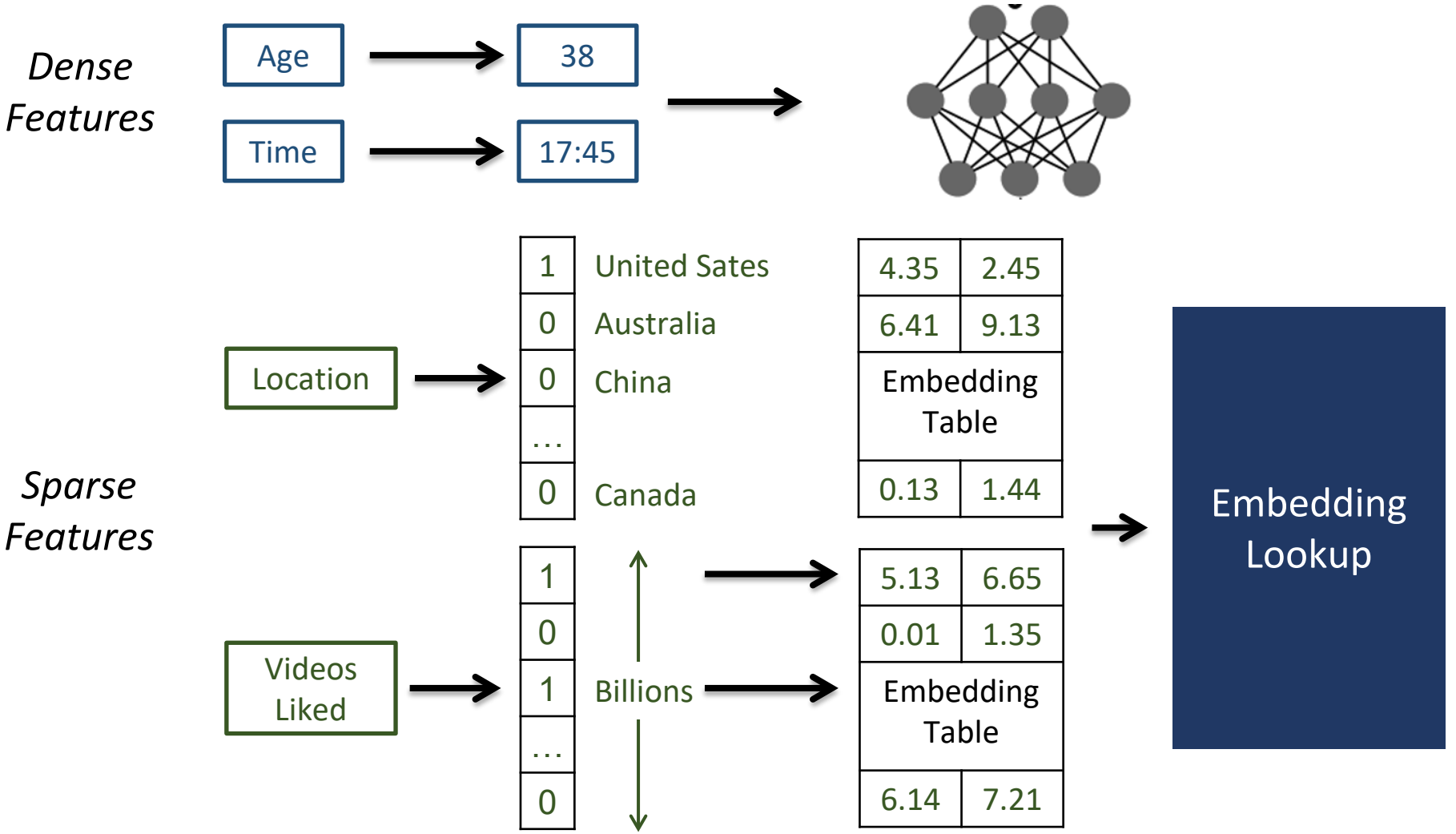


Backup Slides

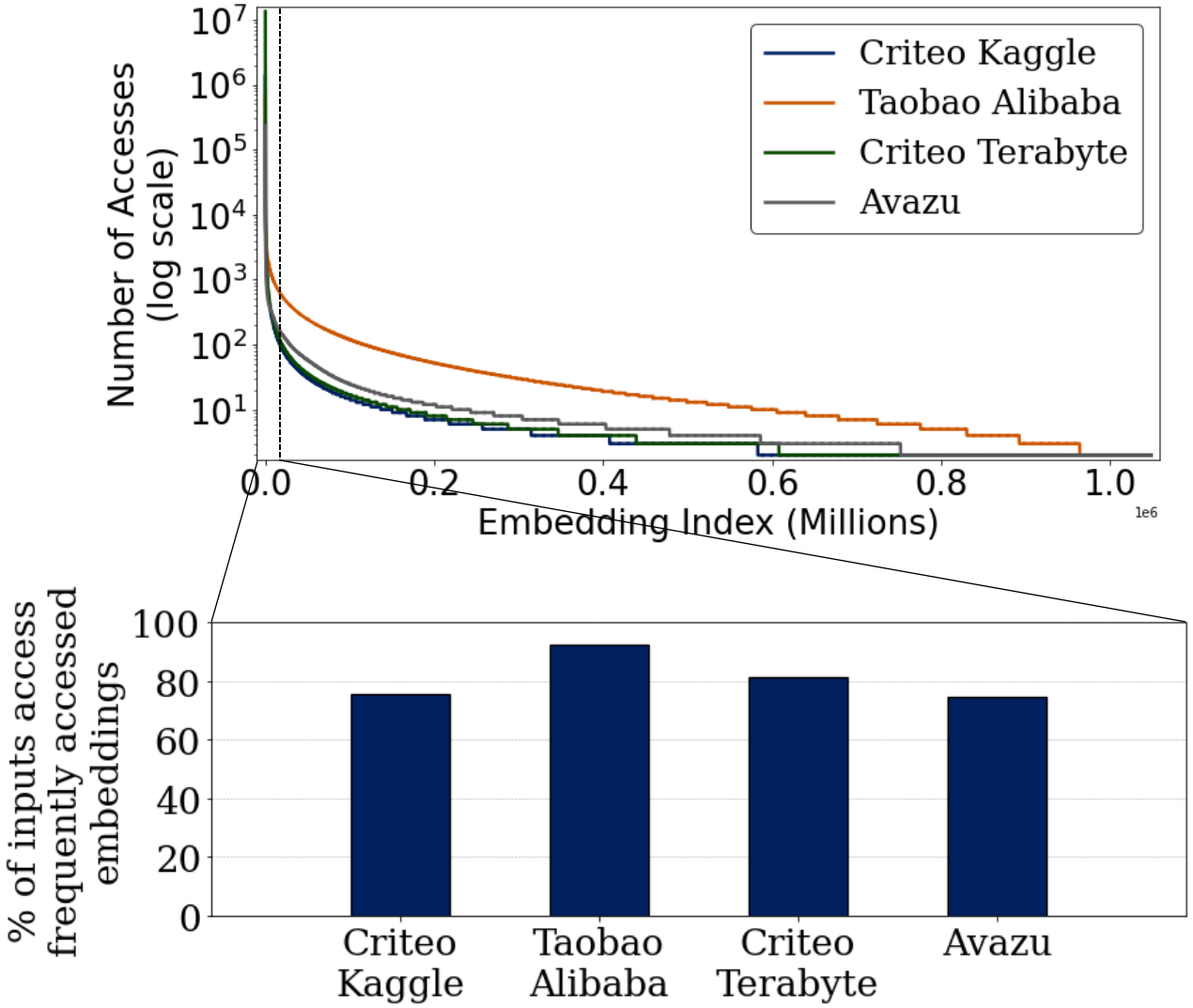
Dense Features



Sparse Features

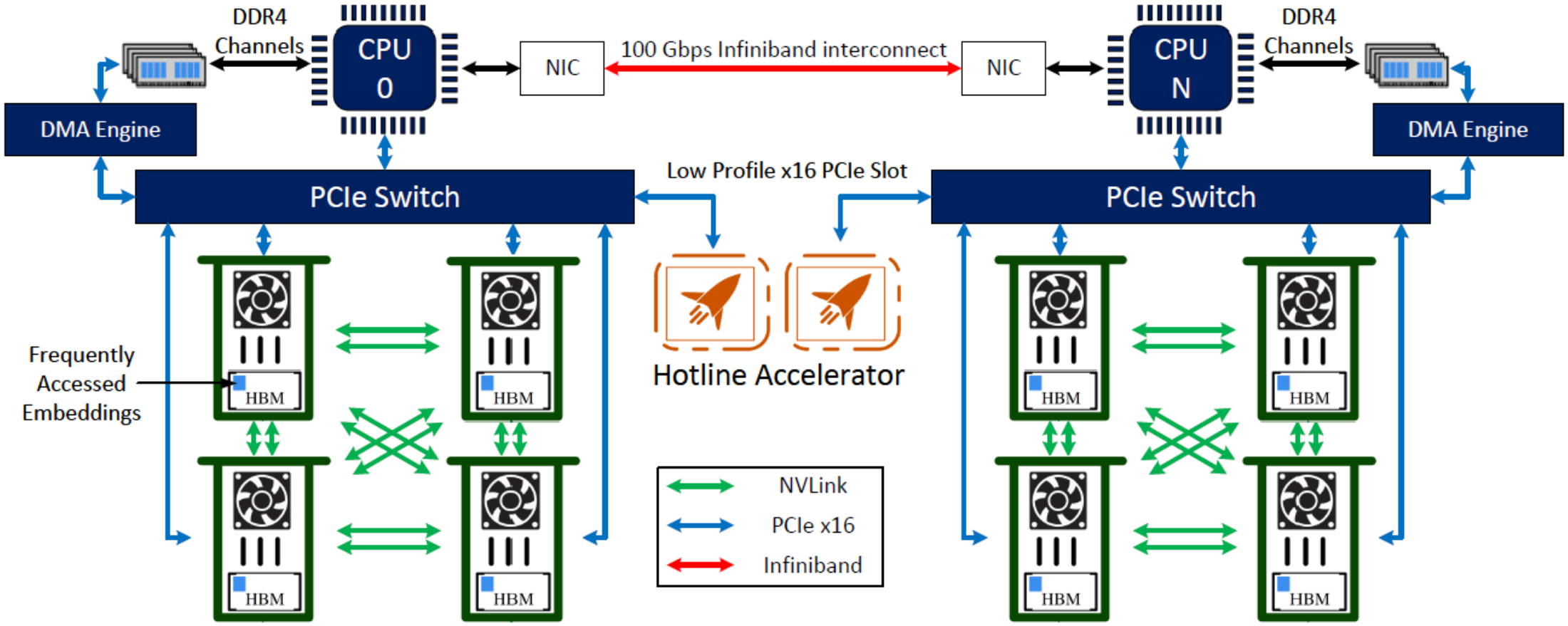


Key Insight: Embedding Access Skew

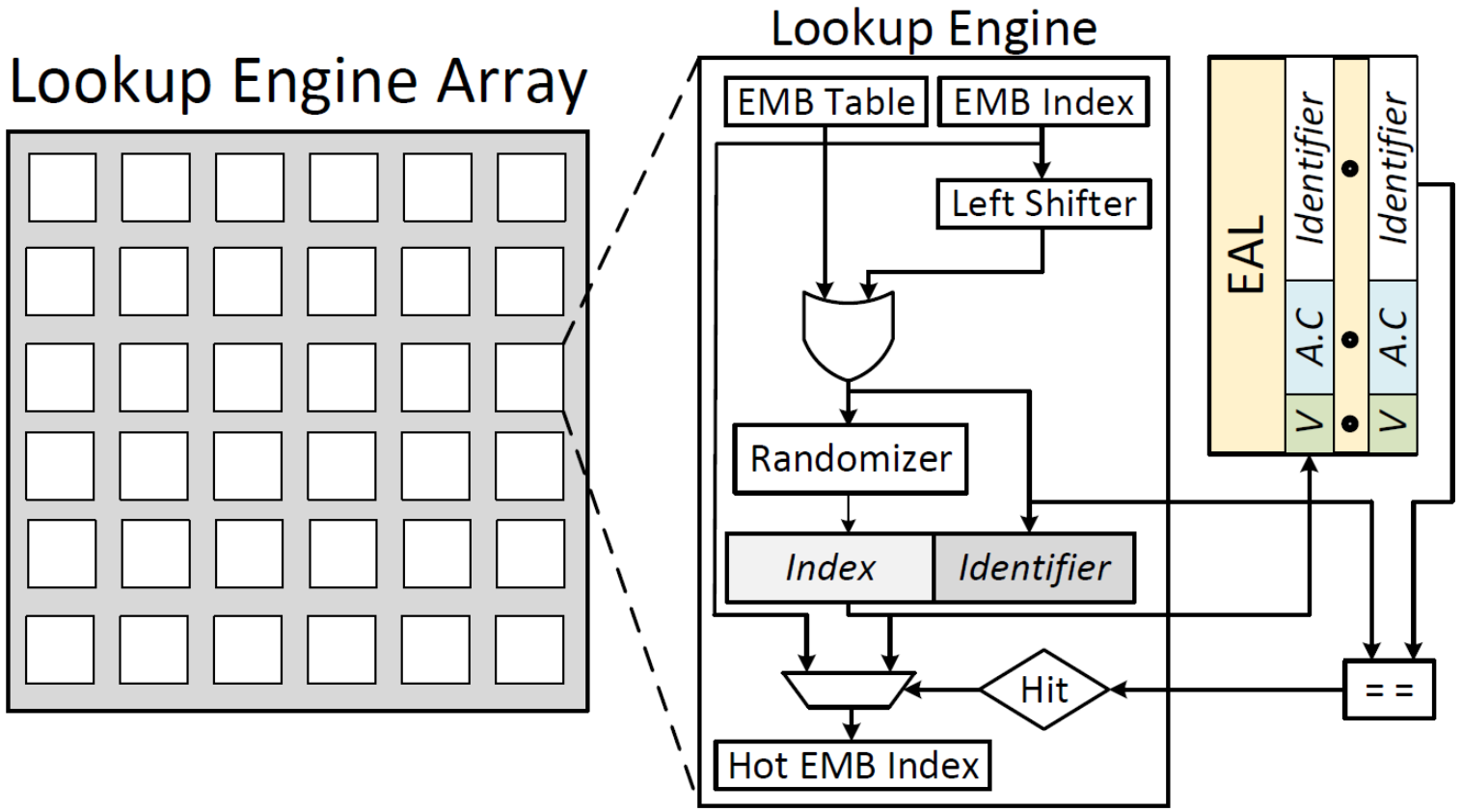


75%

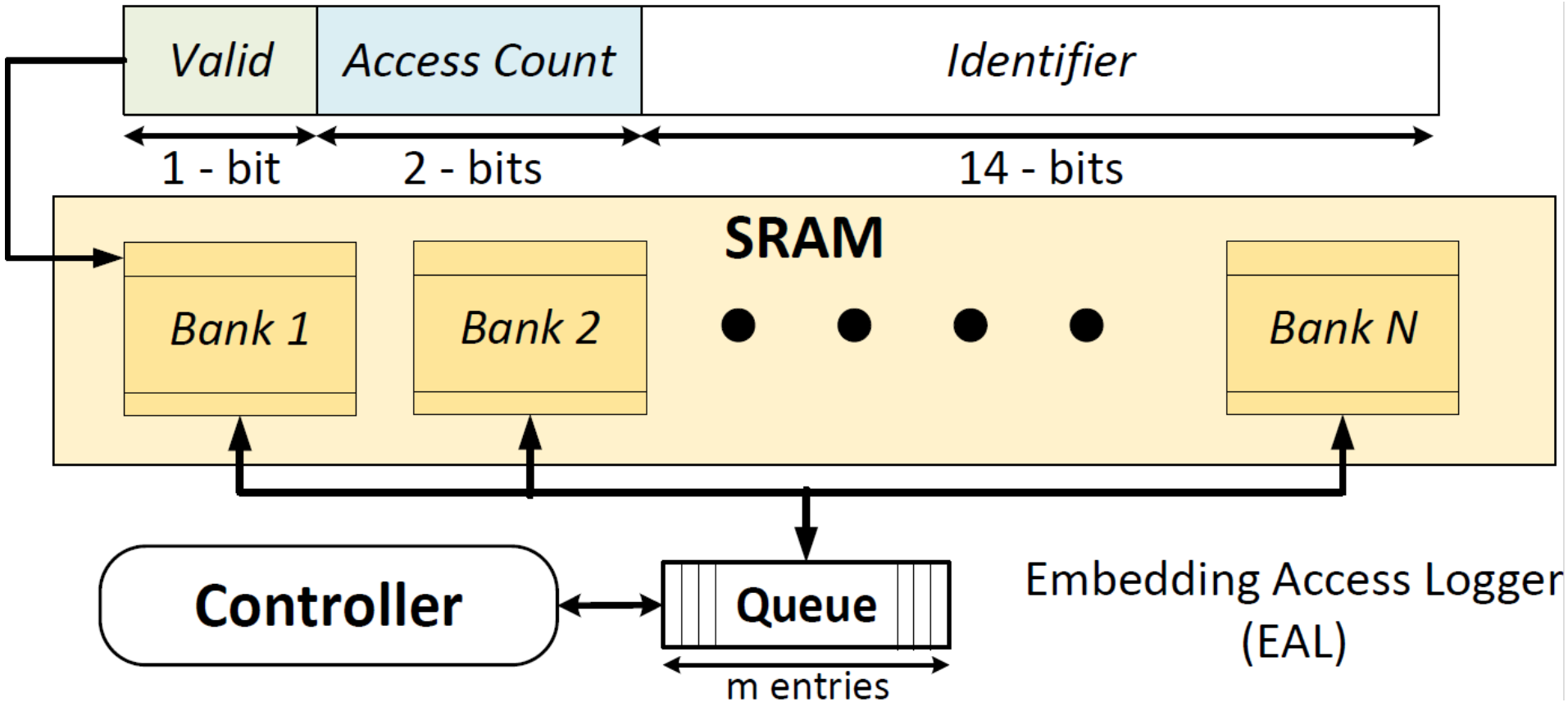
Heterogeneous Acceleration Pipeline



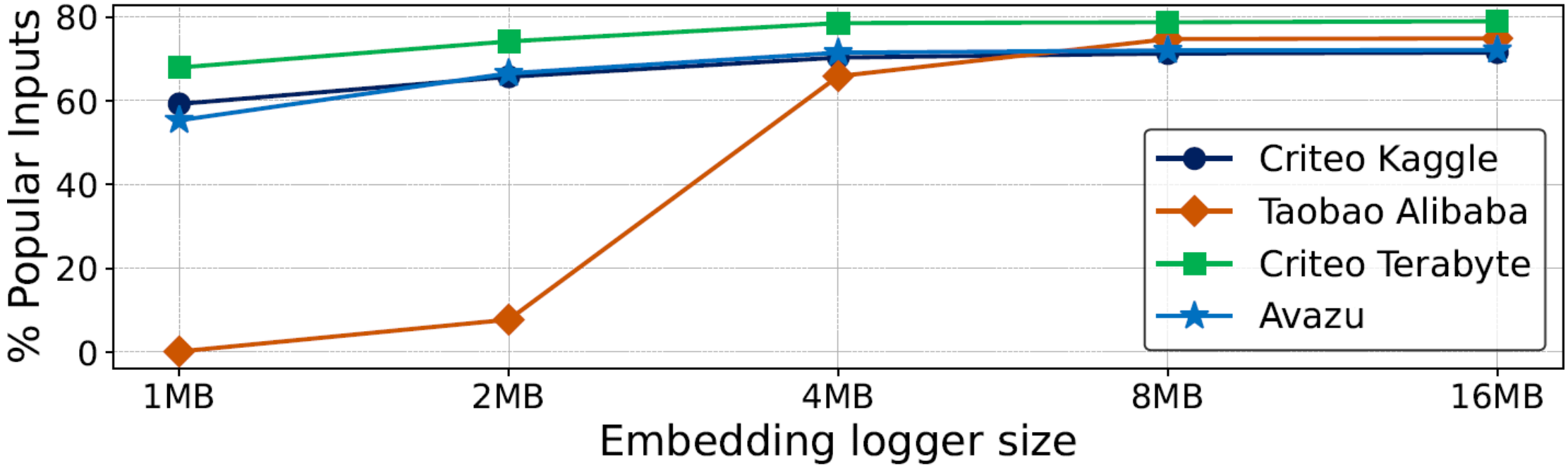
Lookup Engine Array



Embedding Access Logger (EAL)

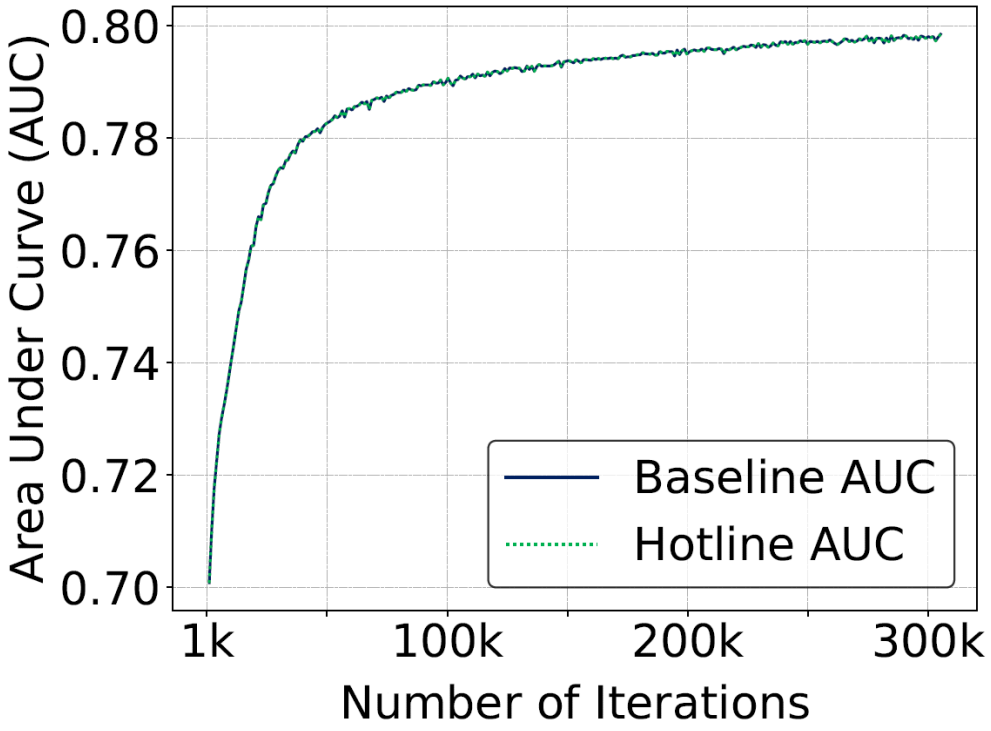


Embedding Access Logger (EAL) Size

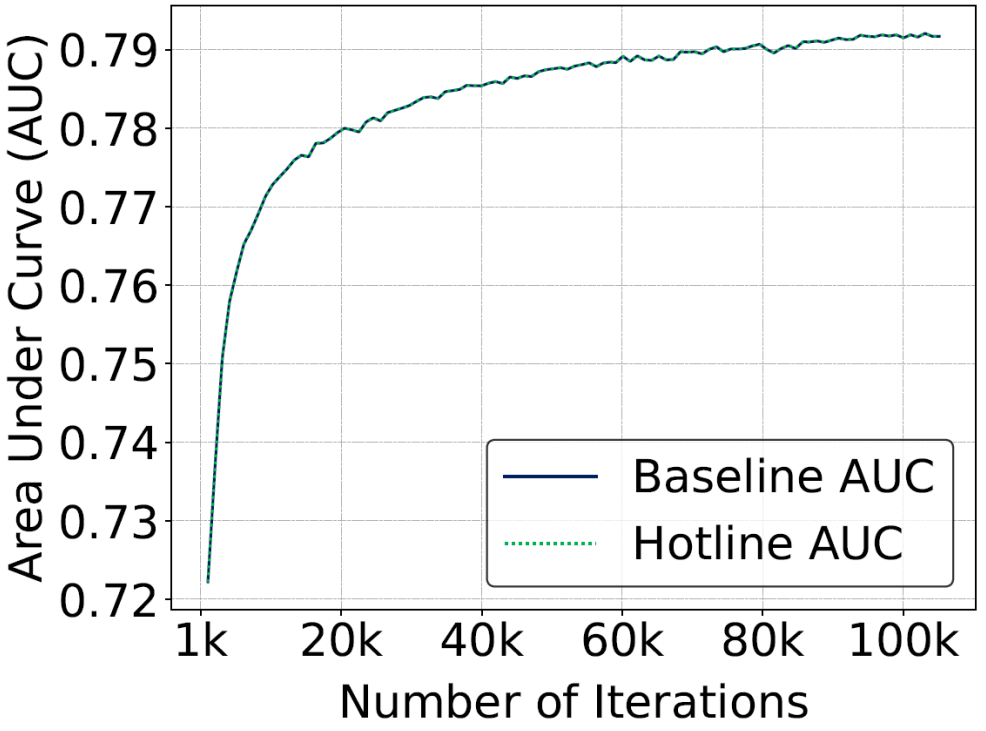


Accuracy

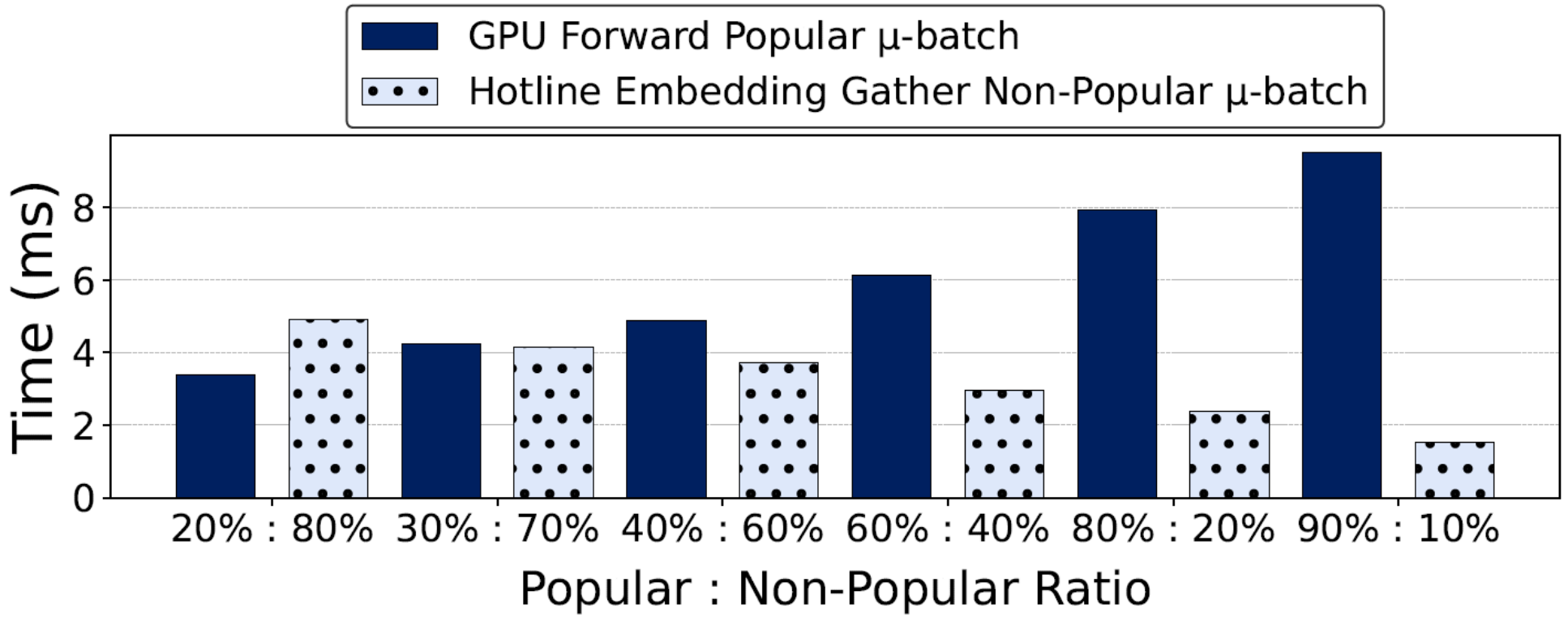
Criteo Kaggle



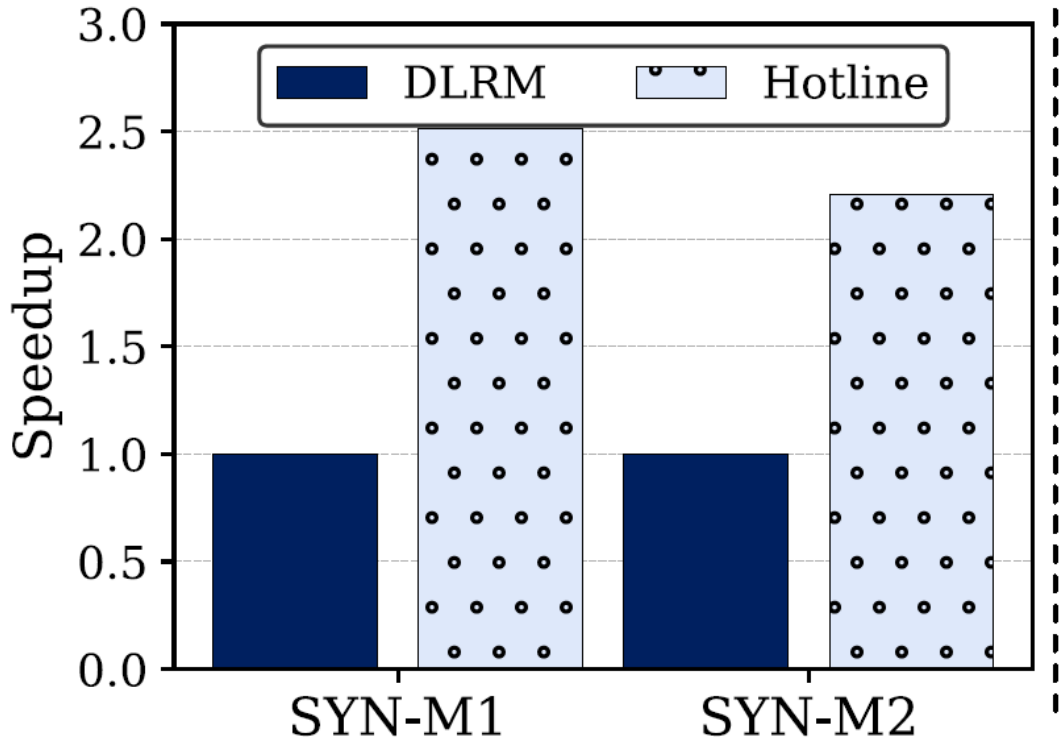
Criteo Terabyte



Popular vs Non-popular Ratio



Synthetic Models

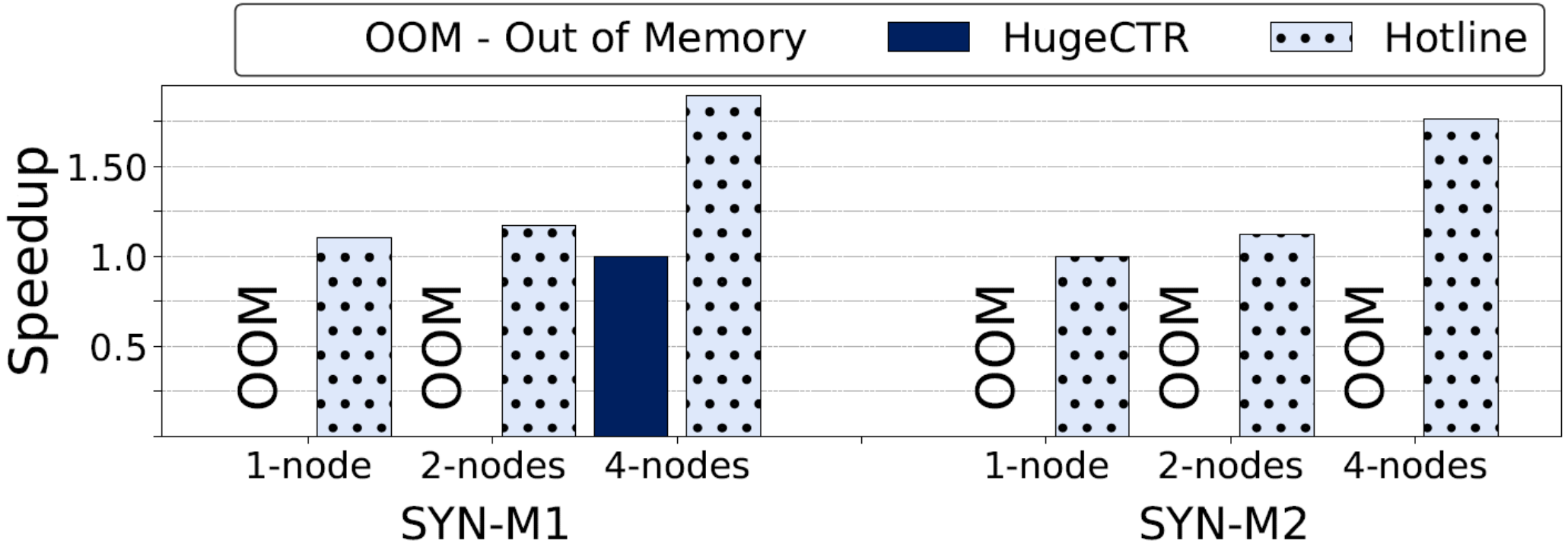


Model: SYN-M1	
Dense Features	54
Sparse Features	102
Size (GB)	196

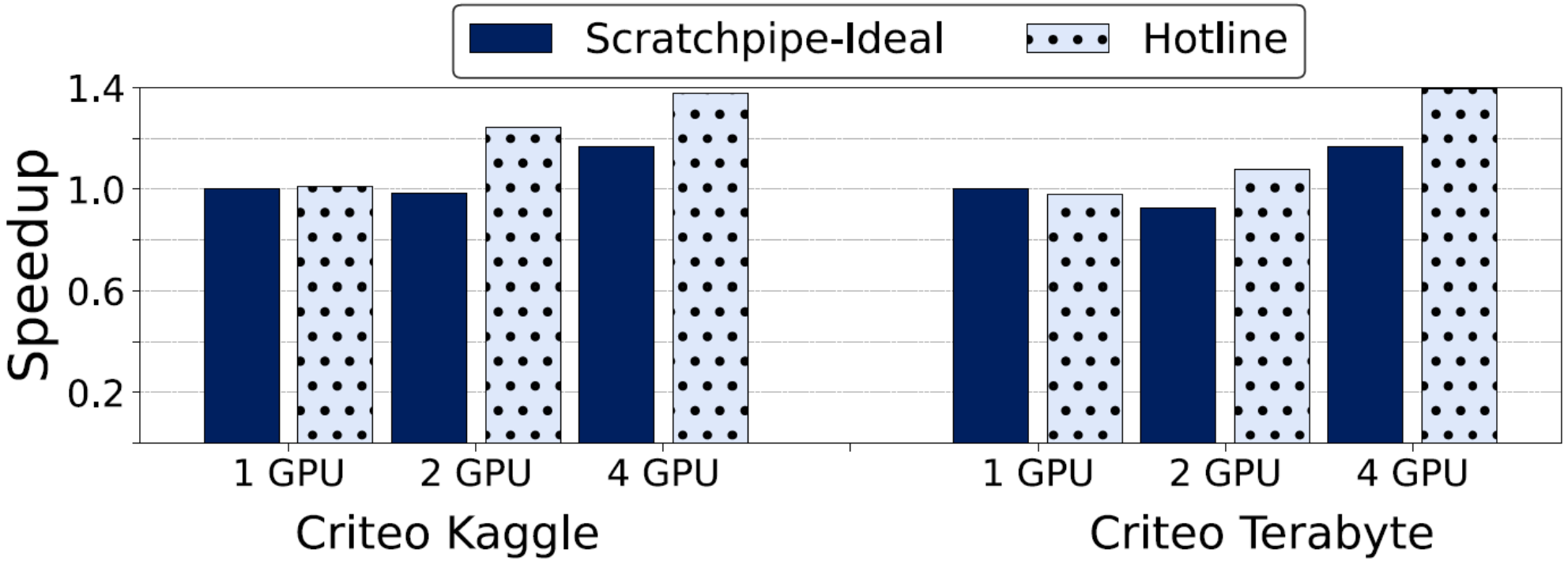
Model: SYN-M2	
Dense Features	102
Sparse Features	204
Size (GB)	390

Synthetic Model Configurations

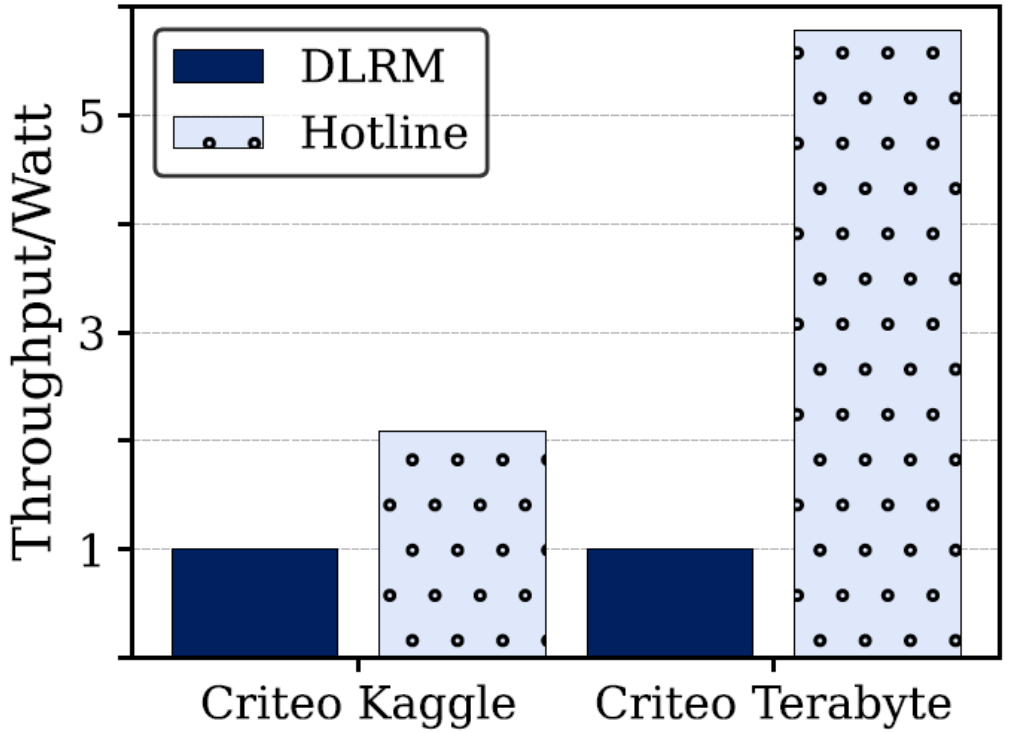
Multi-node Performance



Pipeline Comparison



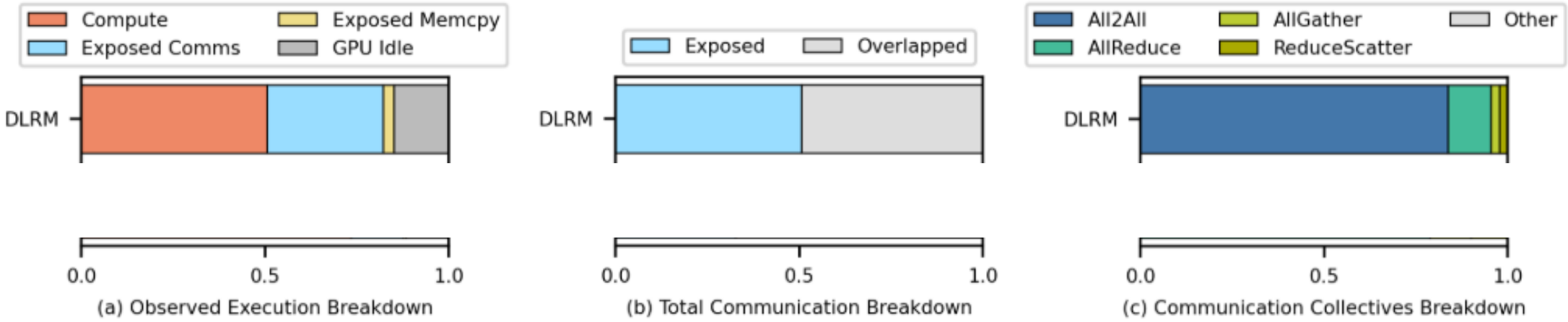
Performance/Watt



Component	Area	Power
EAL	65.72%	67.31%
Input eDRAM	23.44%	26.75%
Lookup Engine	5.80%	2.97%
Emb Vec Buffer	0.66%	0.99%
Others	4.38%	1.98%

Hotline Accelerator
Area and Power Breakdown

Breakdown



Hsia et al. ISCA'24.