Deep Learning: It’s Not All About Recognizing Cats and Dogs

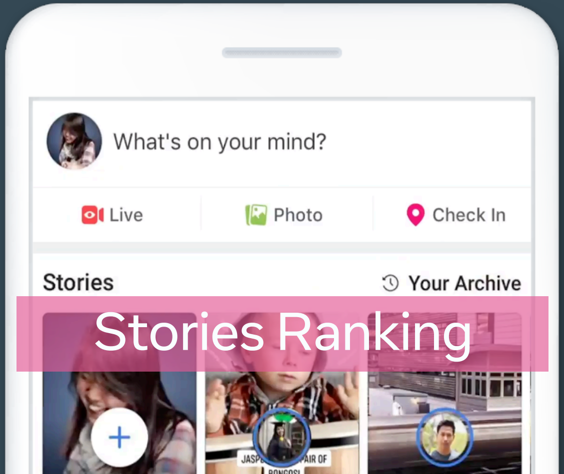
Carole-Jean Wu
Facebook AI Research

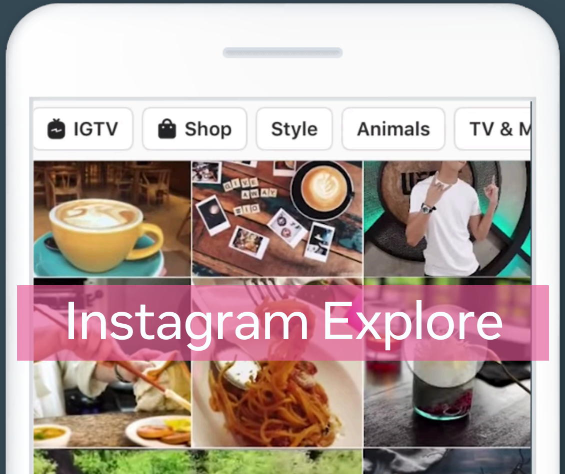
Computer Architecture Seminar
University of Wisconsin – Madison
Oct. 13, 2020

FACEBOOK AI

Recommendation Use Cases

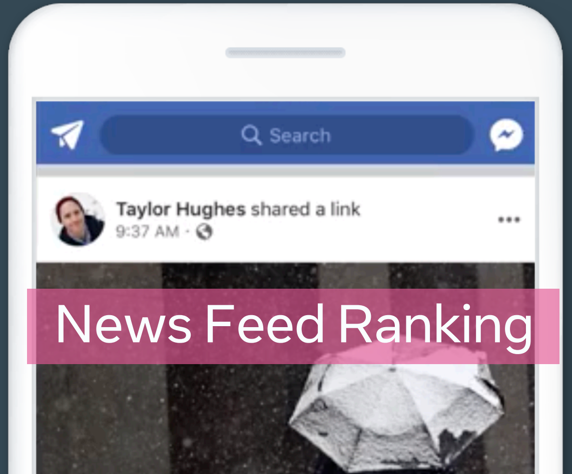
P_{CTR} = f(user features, item features, context features, …, model parameters)



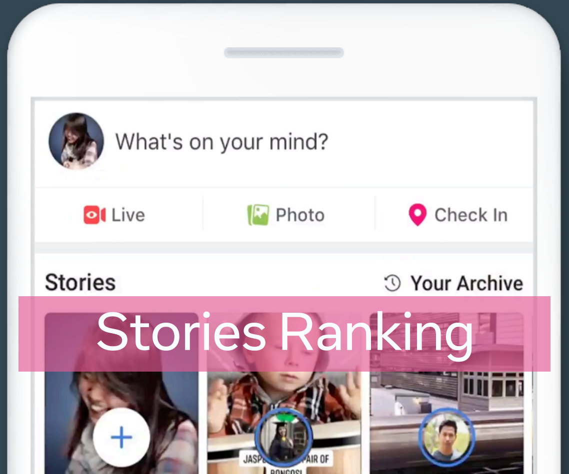


Recommendation Use Cases

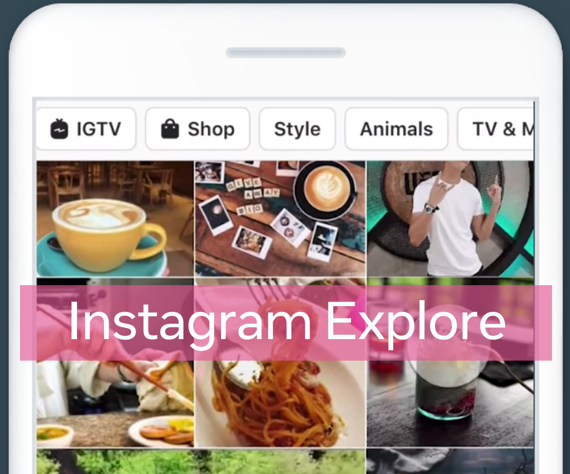
P_{CTR} = f(user features, item features, context features, …, model parameters)



News Feed Ranking



Stories Ranking

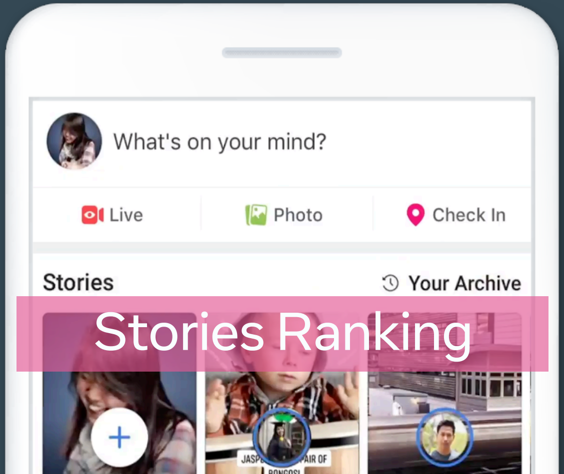


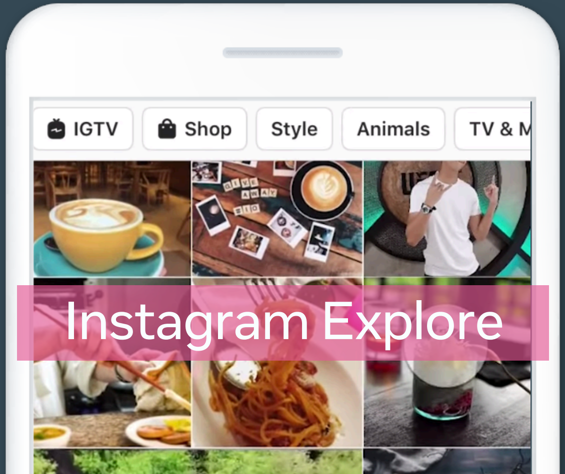
Instagram Explore

Recommendation Use Cases

P_{CTR} = f(user features, item features, context features, …, model parameters)

90%
87%
95%
99%
.
.
.
68%
94%





Evolution of Personalized Recommendation Systems and Algorithms

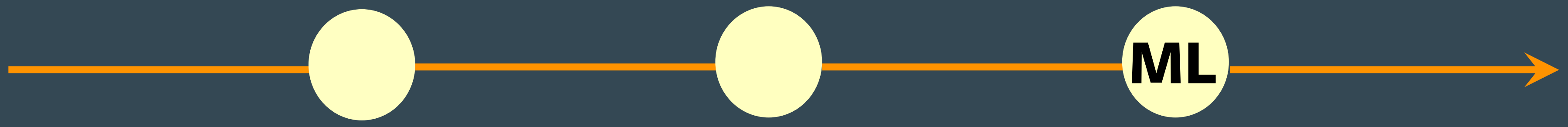
Content-based filter

Collaborative filter

Factorization-based Methods

LR; GBDT;

Pairwise Ranking



Circa 2006

Circa 2009

Evolution of Personalized Recommendation Systems and Algorithms

Content-based filter
Collaborative filter

Factorization-based Methods

LR; GBDT;
Pairwise Ranking

Wide & Deep
DeepFM/xDeepFM
DIN/DIEN
DLRM
DCN

Circa 2006

Circa 2009

Circa 2015

Compute Footprint of Recommendation

50%

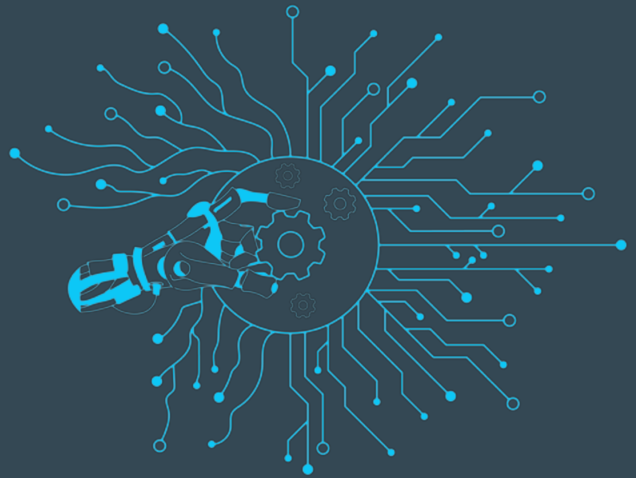
of all AI Training Cycles

Compute Footprint of Recommendation

50%

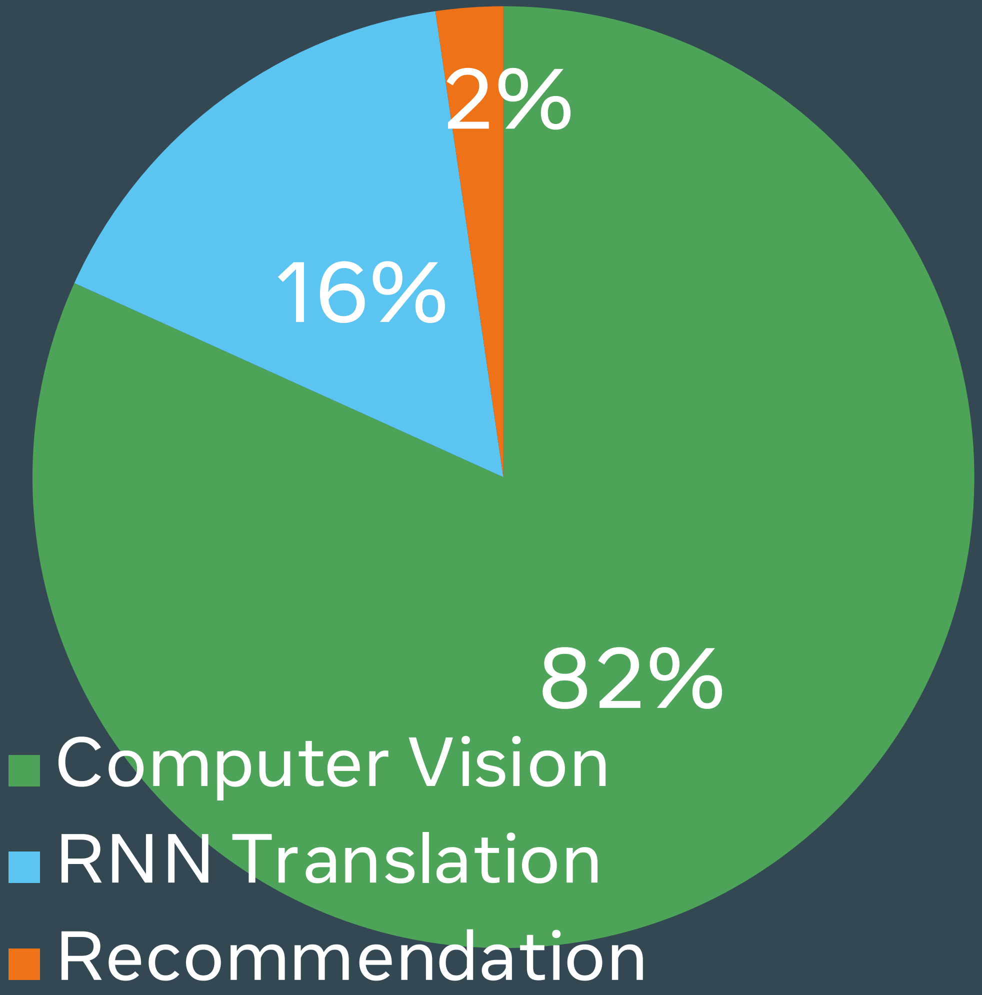
of all AI Training Cycles

80%



of all AI Inference Cycles

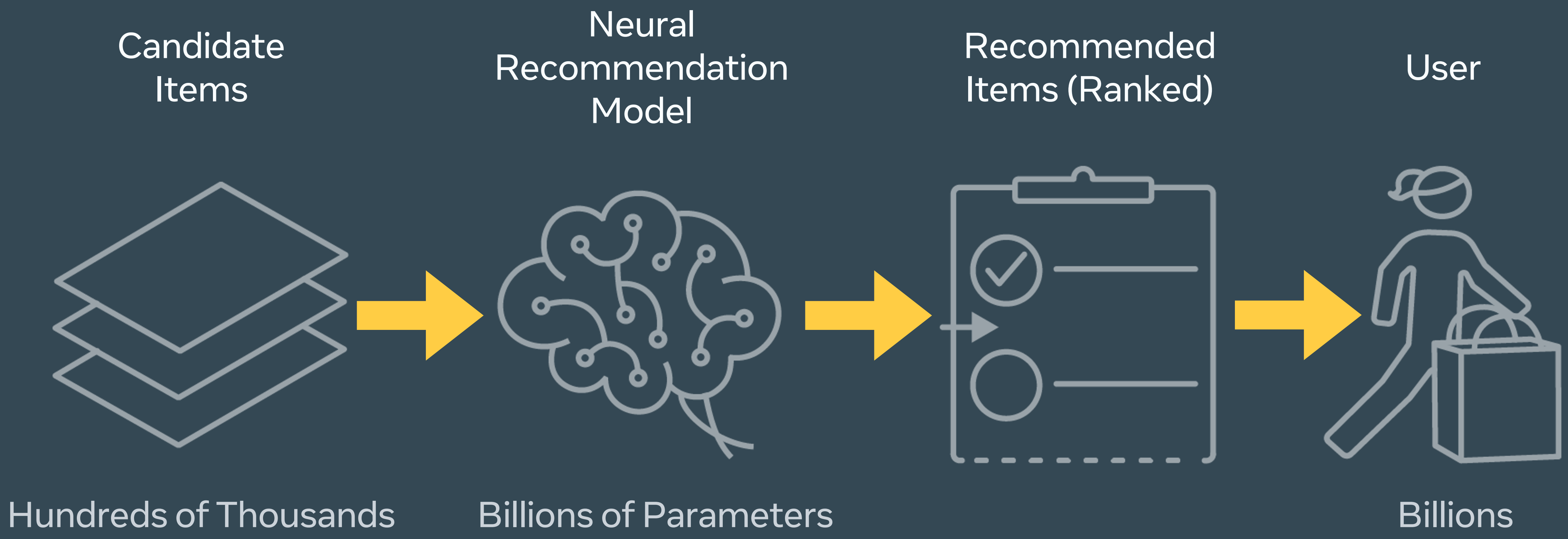
Publications - ML Systems Community



Why The Disconnect?

Personalized Recommendation

At Data Center Scale



How Do Recommender Systems Work?

How Do Recommender Systems Work?

Continuous (dense) features

Age

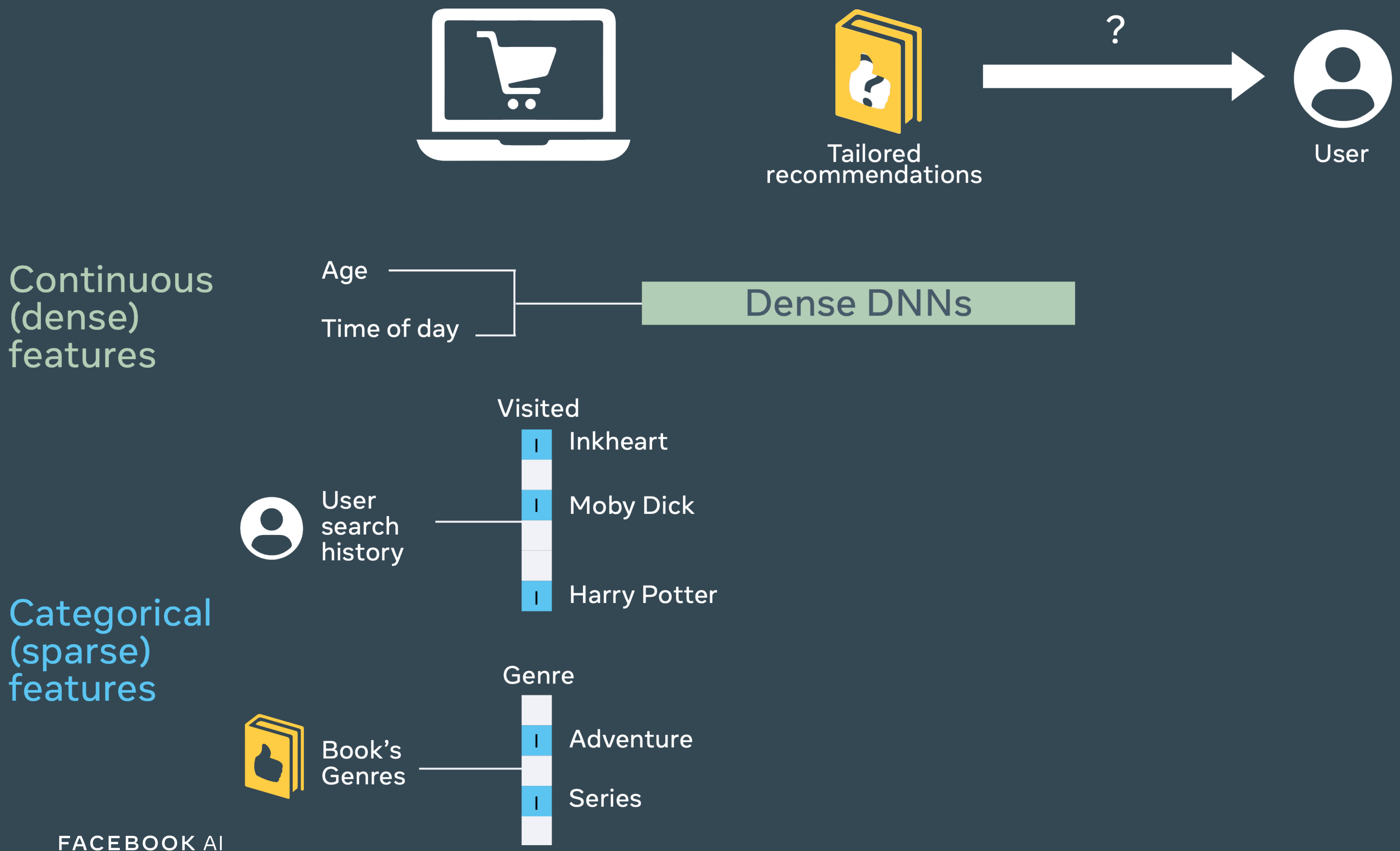
Time of day

User search history

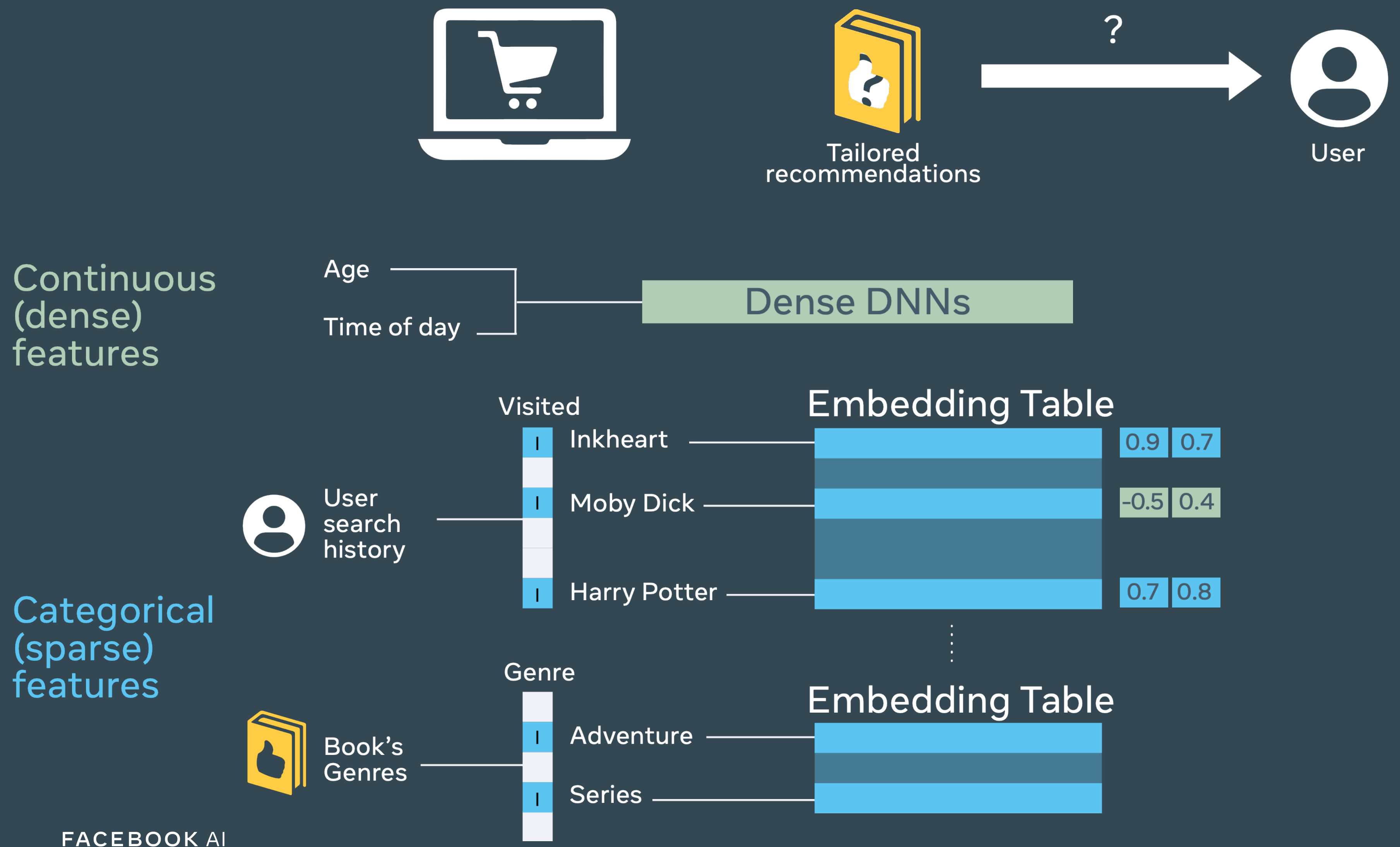
Categorical (sparse) features

Book's Genres

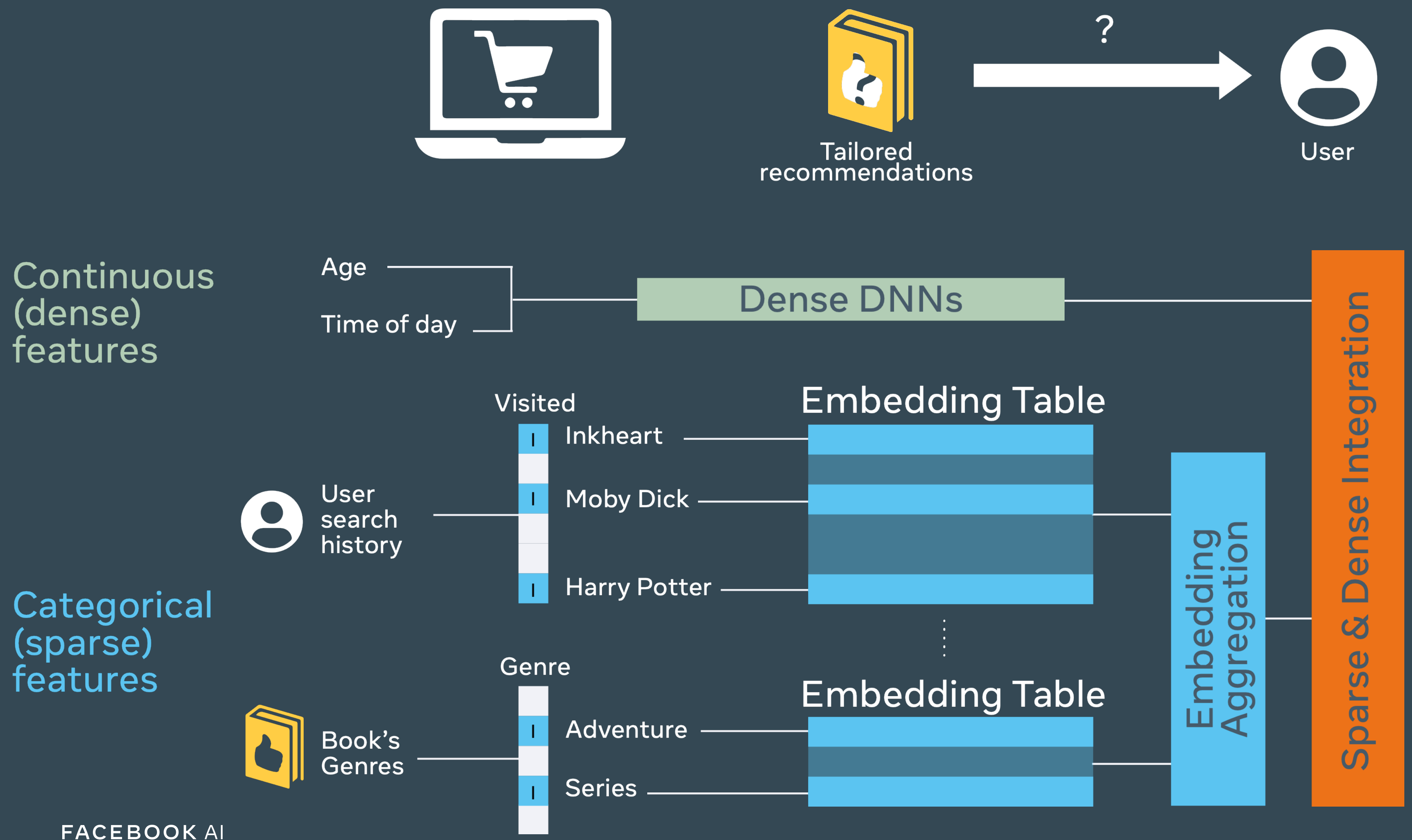
How Do Recommender Systems Work?



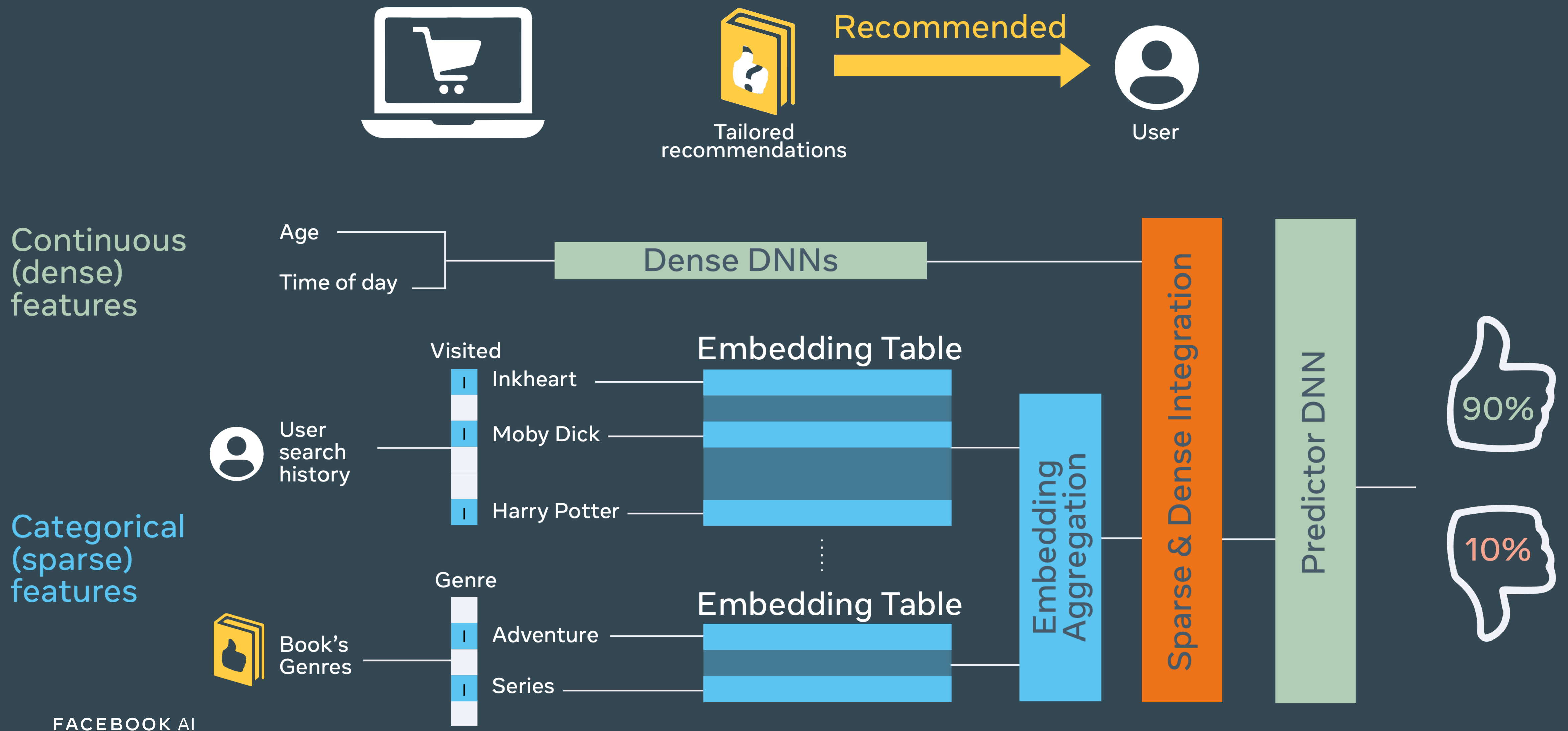
How Do Recommender Systems Work?



How Do Recommender Systems Work?



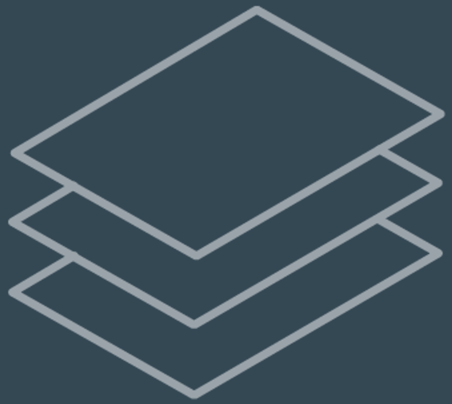
How Do Recommender Systems Work?



FACEBOOK AI

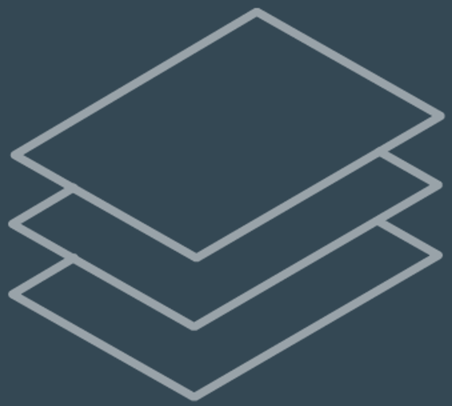
Ranking More Items Leads to Better Recommendations

High
Throughput



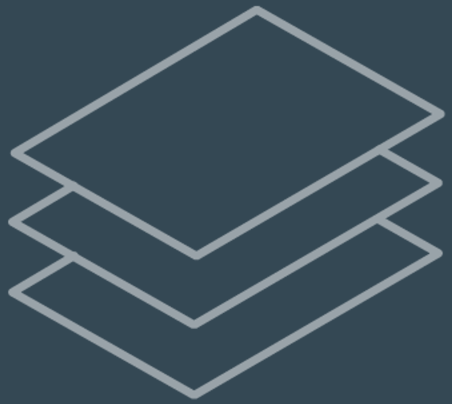
Ranking More Items Leads to Better Recommendations

High Throughput + Low Latency



Ranking More Items Leads to Better Recommendations

High Throughput + Low Latency = Latency-Bounded Throughput



FACEBOOK AI

Agenda

Motivation

Understanding the Unique System Challenges

Characterizing Performance Acceleration with GPUs

Optimizing Neural Recommendation Inference At-Scale

Conclusion and Future Work

Unique System Challenges

Embedding Tables

Unique System Challenges

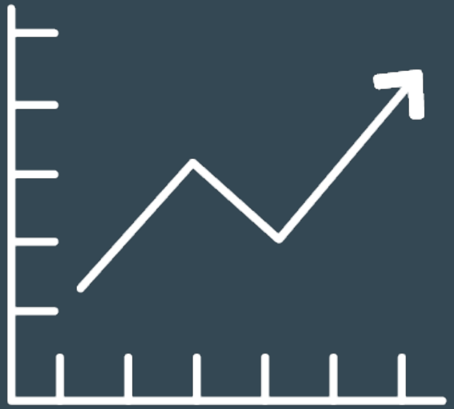
Embedding Tables

Model Heterogeneity

Unique System Challenges

Embedding Tables

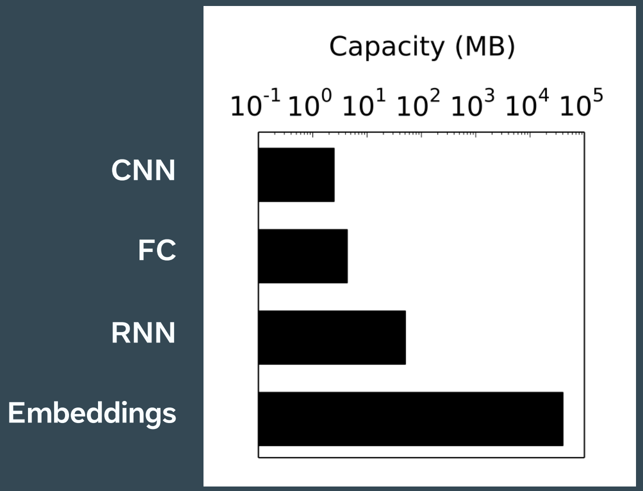
Model Heterogeneity



Performance Variance

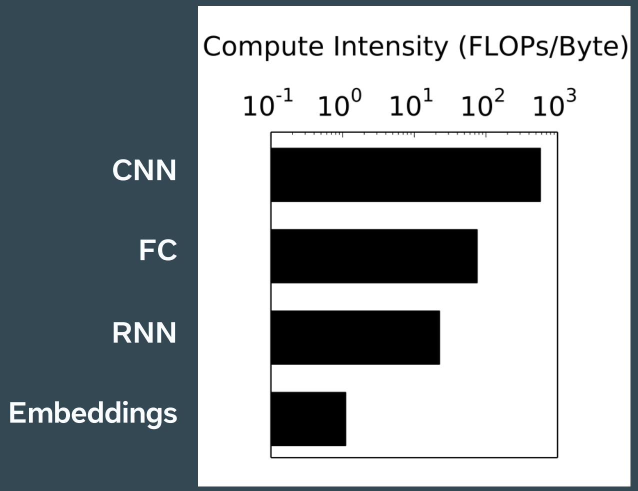
Challenge: Embedding Tables

Storage Capacity



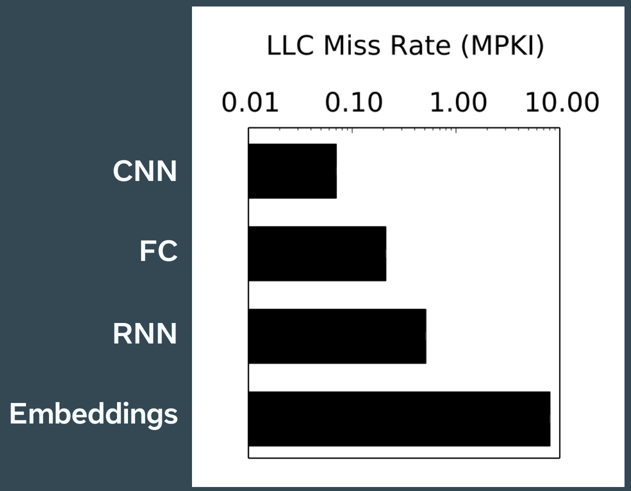
Orders of Magnitude
Larger

Compute Intensity



Orders of Magnitude
Fewer FLOPS/Byte

Memory Irregularity



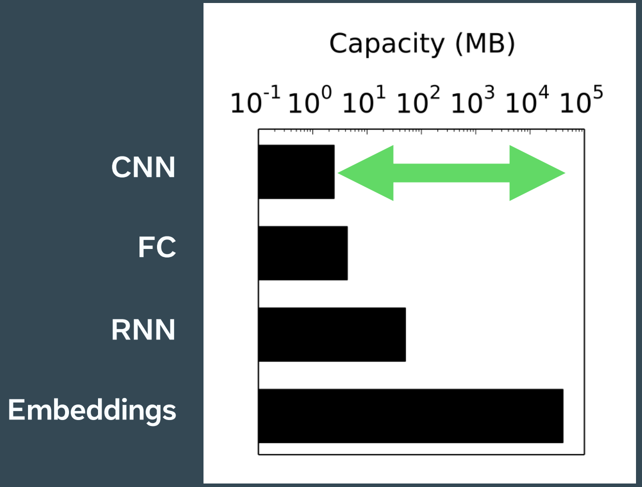
Sparse, Irregular
Memory Accesses

[HPCA 2020] **The Architectural Implications of Facebook's DNN-based Personalized Recommendation.** U. Gupta, C.-J. Wu, X. Wang, M. Naumov, B. Reagen, D. Brooks, B. Cottel, K. Hazelwood, M. Hempstead, B. Jia, H.-H. Lee, A. Malevich, D. Mudigere, M. Smelyanskiy, L. Xiong, X. Zhang.

[ISCA 2020] **RecNMP: Accelerating Personalized Recommendation with Near-Memory Processing.** L. Ke, U. Gupta, B. Cho, D. Brooks, V. Chandra, U. Diril, A. Firoozshahian, K. Hazelwood, B. Jia, H.-S. Lee, M. Li, B. Maher, D. Mudigere, M. Naumov, M. Schatz, M. Smelyanskiy, X. Wang, B. Reagen, C.-J. Wu, M. Hempstead, X. Zhang.

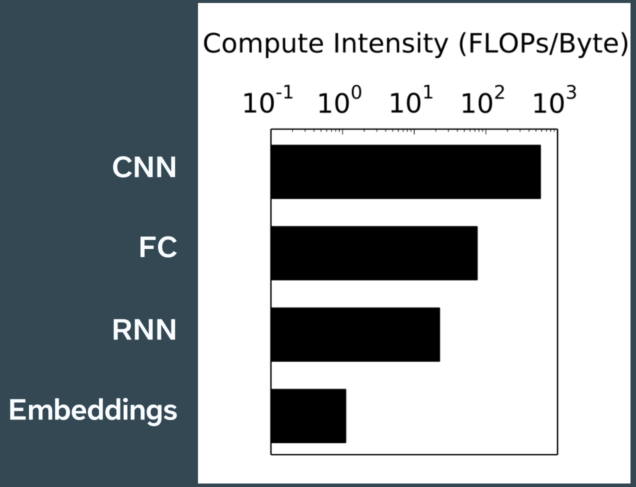
Challenge: Embedding Tables

Storage Capacity



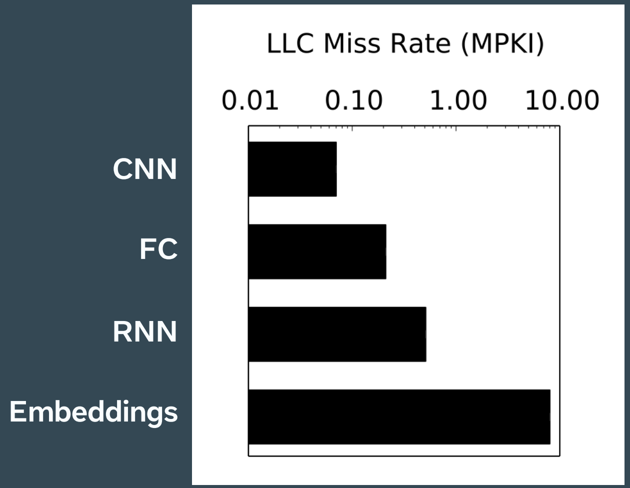
Orders of Magnitude
Larger

Compute Intensity



Orders of Magnitude
Fewer FLOPS/Byte

Memory Irregularity



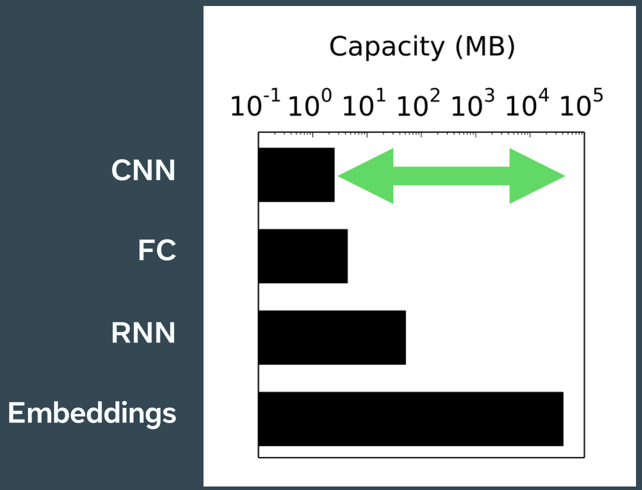
Sparse, Irregular
Memory Accesses

[HPCA 2020] **The Architectural Implications of Facebook's DNN-based Personalized Recommendation.** U. Gupta, C.-J. Wu, X. Wang, M. Naumov, B. Reagen, D. Brooks, B. Cottel, K. Hazelwood, M. Hempstead, B. Jia, H.-H. Lee, A. Malevich, D. Mudigere, M. Smelyanskiy, L. Xiong, X. Zhang.

[ISCA 2020] **RecNMP: Accelerating Personalized Recommendation with Near-Memory Processing.** L. Ke, U. Gupta, B. Cho, D. Brooks, V. Chandra, U. Diril, A. Firoozshahian, K. Hazelwood, B. Jia, H.-S. Lee, M. Li, B. Maher, D. Mudigere, M. Naumov, M. Schatz, M. Smelyanskiy, X. Wang, B. Reagen, C.-J. Wu, M. Hempstead, X. Zhang.

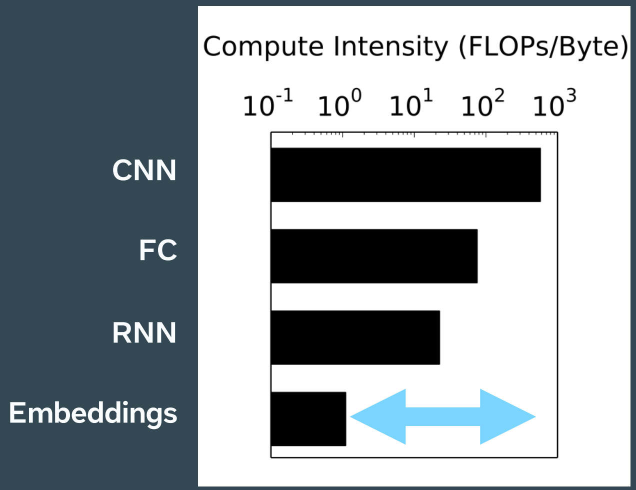
Challenge: Embedding Tables

Storage Capacity



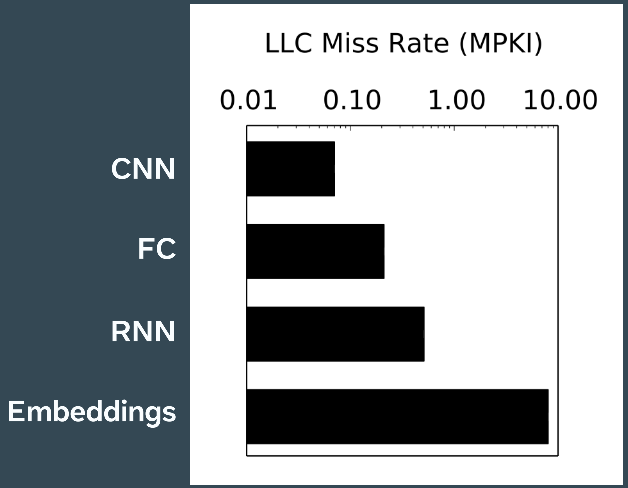
Orders of Magnitude
Larger

Compute Intensity



Orders of Magnitude
Fewer FLOPS/Byte

Memory Irregularity



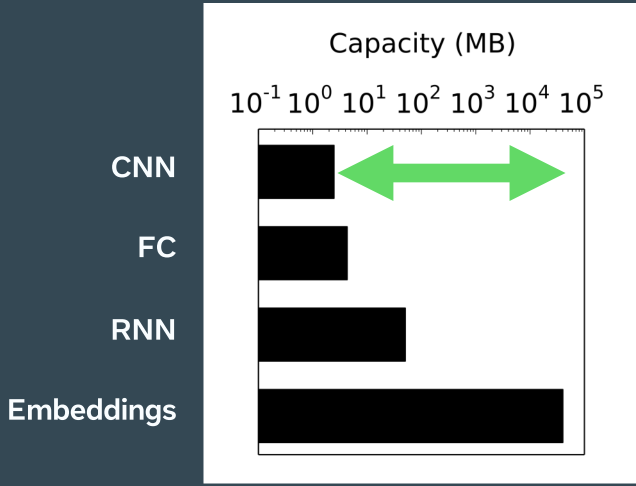
Sparse, Irregular
Memory Accesses

[HPCA 2020] **The Architectural Implications of Facebook's DNN-based Personalized Recommendation.** U. Gupta, C.-J. Wu, X. Wang, M. Naumov, B. Reagen, D. Brooks, B. Cottel, K. Hazelwood, M. Hempstead, B. Jia, H.-H. Lee, A. Malevich, D. Mudigere, M. Smelyanskiy, L. Xiong, X. Zhang.

[ISCA 2020] **RecNMP: Accelerating Personalized Recommendation with Near-Memory Processing.** L. Ke, U. Gupta, B. Cho, D. Brooks, V. Chandra, U. Diril, A. Firoozshahian, K. Hazelwood, B. Jia, H.-S. Lee, M. Li, B. Maher, D. Mudigere, M. Naumov, M. Schatz, M. Smelyanskiy, X. Wang, B. Reagen, C.-J. Wu, M. Hempstead, X. Zhang.

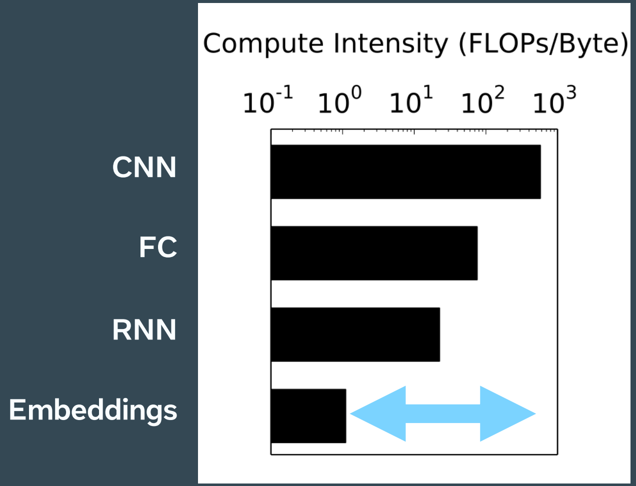
Challenge: Embedding Tables

Storage Capacity



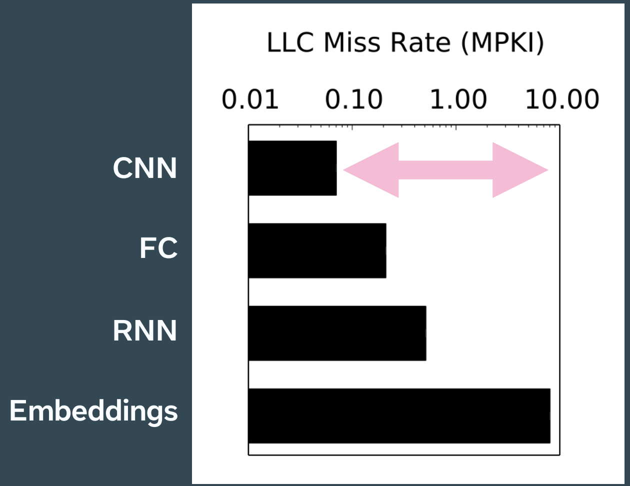
Orders of Magnitude
Larger

Compute Intensity



Orders of Magnitude
Fewer FLOPS/Byte

Memory Irregularity



Sparse, Irregular
Memory Accesses

[HPCA 2020] **The Architectural Implications of Facebook's DNN-based Personalized Recommendation.** U. Gupta, C.-J. Wu, X. Wang, M. Naumov, B. Reagen, D. Brooks, B. Cottel, K. Hazelwood, M. Hempstead, B. Jia, H.-H. Lee, A. Malevich, D. Mudigere, M. Smelyanskiy, L. Xiong, X. Zhang.

[ISCA 2020] **RecNMP: Accelerating Personalized Recommendation with Near-Memory Processing.** L. Ke, U. Gupta, B. Cho, D. Brooks, V. Chandra, U. Diril, A. Firoozshahian, K. Hazelwood, B. Jia, H.-S. Lee, M. Li, B. Maher, D. Mudigere, M. Naumov, M. Schatz, M. Smelyanskiy, X. Wang, B. Reagen, C.-J. Wu, M. Hempstead, X. Zhang.

Challenge: Model Heterogeneity

Three Facebook Recommendation Models

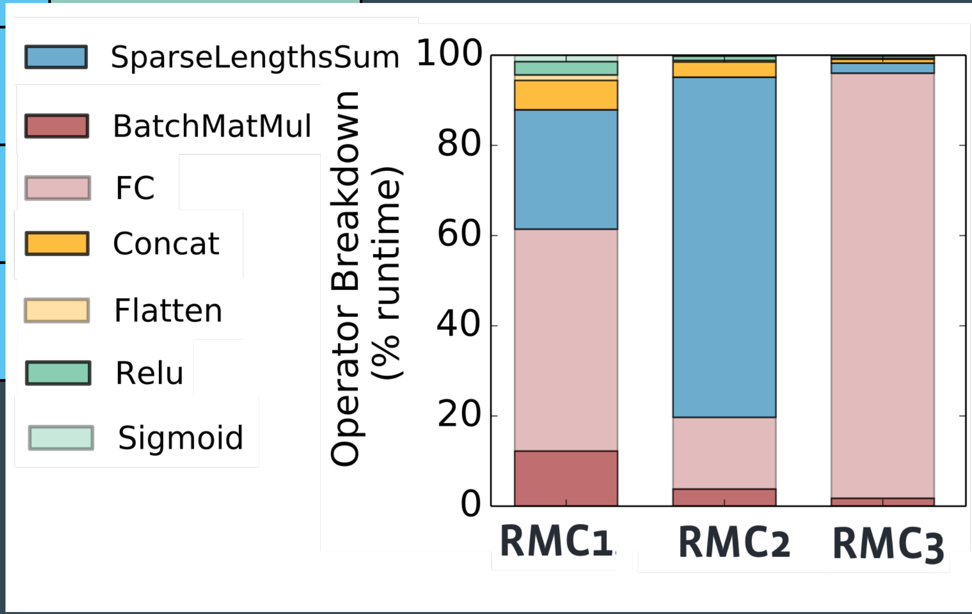
	RM1	RM2	RM3
FC sizes	Small	Medium	Large
Number of embedding tables	O(10)	O(50)	O(10)
Size of embeddings	Small	Medium	Large
Number of lookups per table	O(100)	O(100)	O(10)

[HPCA 2020] **The Architectural Implications of Facebook's DNN-based Personalized Recommendation.** U. Gupta, C.-J. Wu, X. Wang, M. Naumov, B. Reagen, D. Brooks, B. Cottel, K. Hazelwood, M. Hempstead, B. Jia, H.-H. Lee, A. Malevich, D. Mudigere, M. Smelyanskiy, L. Xiong, X. Zhang.

Challenge: Model Heterogeneity

Three Facebook Recommendation Models

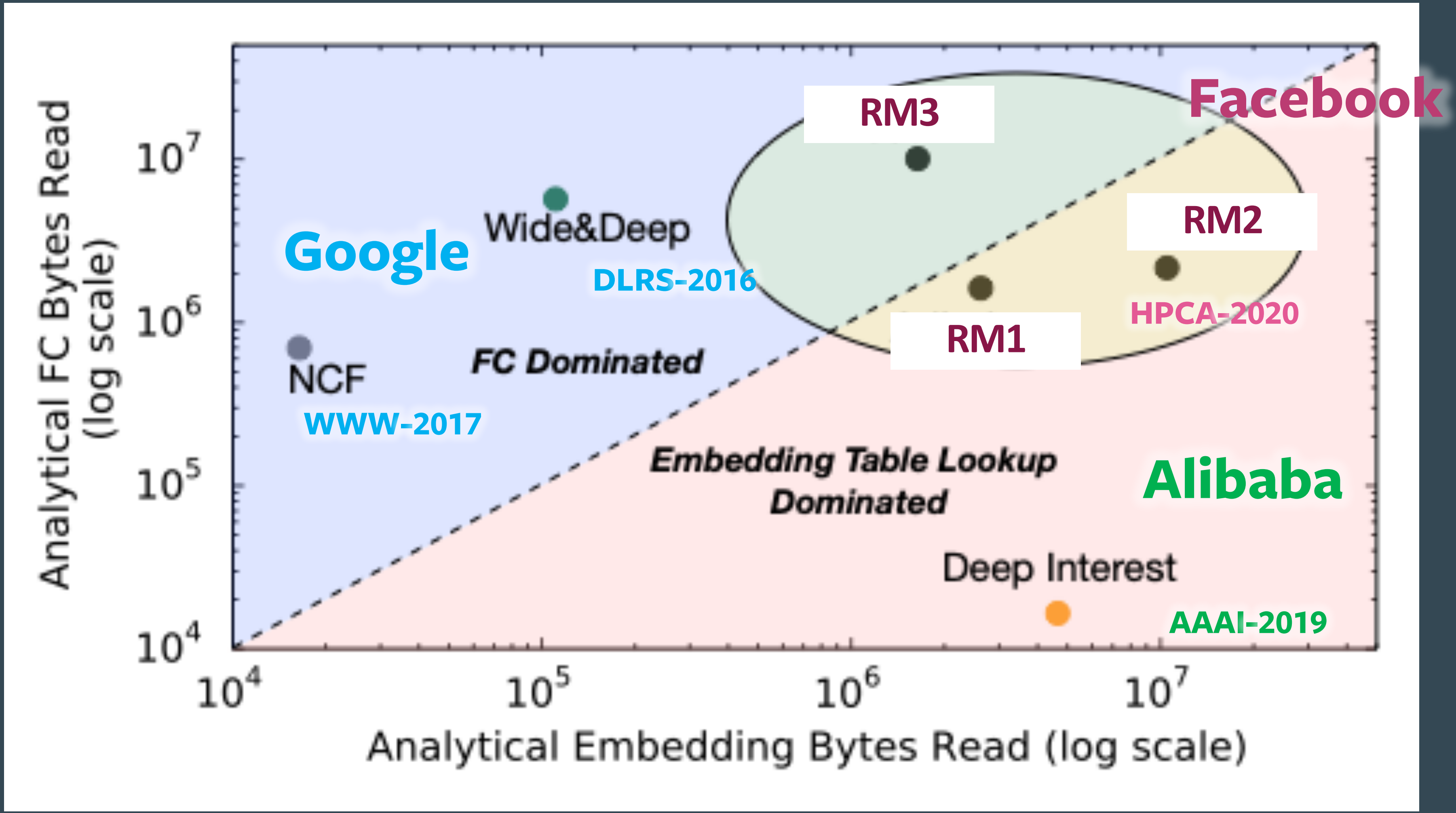
	RM1	RM2	RM3
FC sizes	Small	Medium	Large
Number of embedding tables	O(10)	O(50)	
Size of embeddings	Small	Medium	
Number of lookups per table	O(100)	O(100)	



[HPCA 2020] **The Architectural Implications of Facebook's DNN-based Personalized Recommendation.** U. Gupta, C.-J. Wu, X. Wang, M. Naumov, B. Reagen, D. Brooks, B. Cottel, K. Hazelwood, M. Hempstead, B. Jia, H.-H. Lee, A. Malevich, D. Mudigere, M. Smelyanskiy, L. Xiong, X. Zhang.

Challenge: Model Heterogeneity

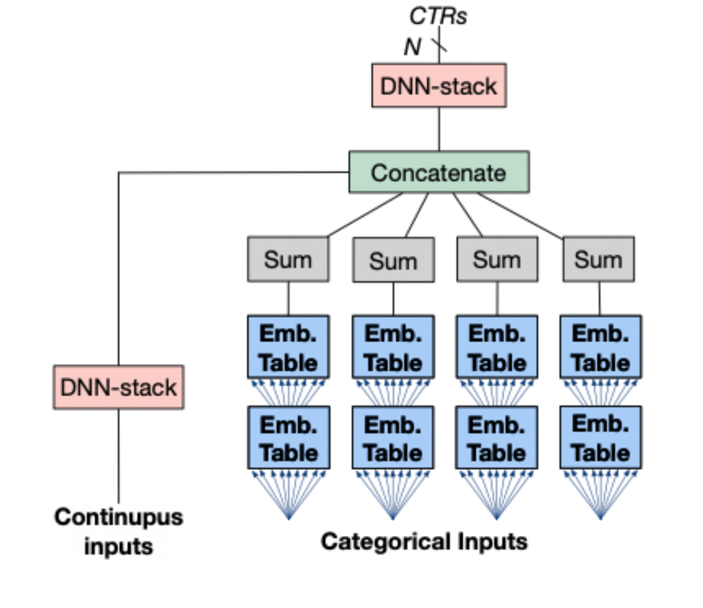
The Landscape of Modern Recommendation Models



Challenge: Model Heterogeneity

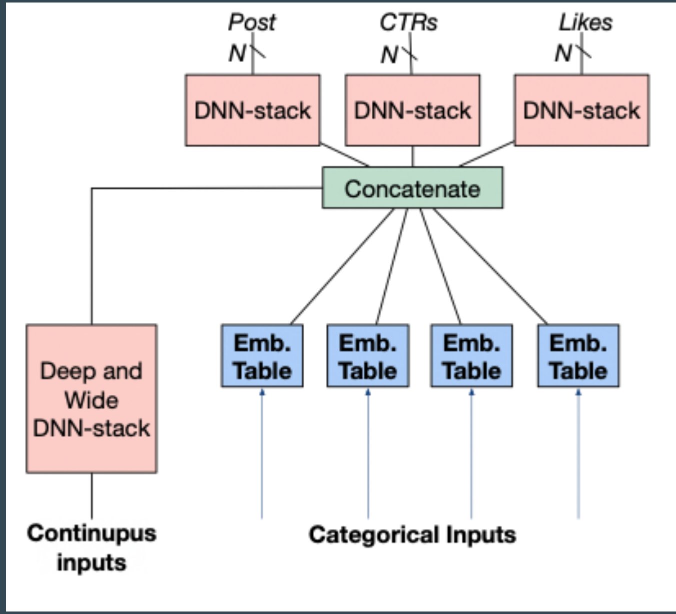
Unique Categories of Recommendation Model Architecture

Embedding-dominated



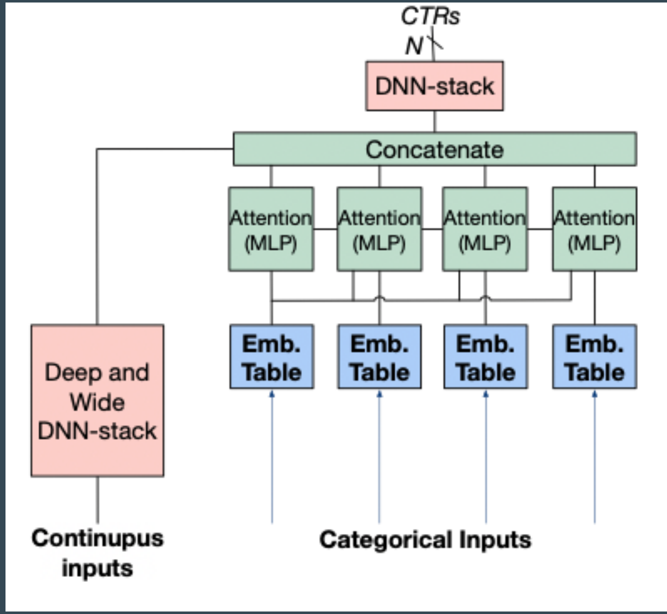
Many embedding tables
Tens to hundreds of lookups

MLP-dominated



Deep, wide MLP layers
Many output DNN stacks

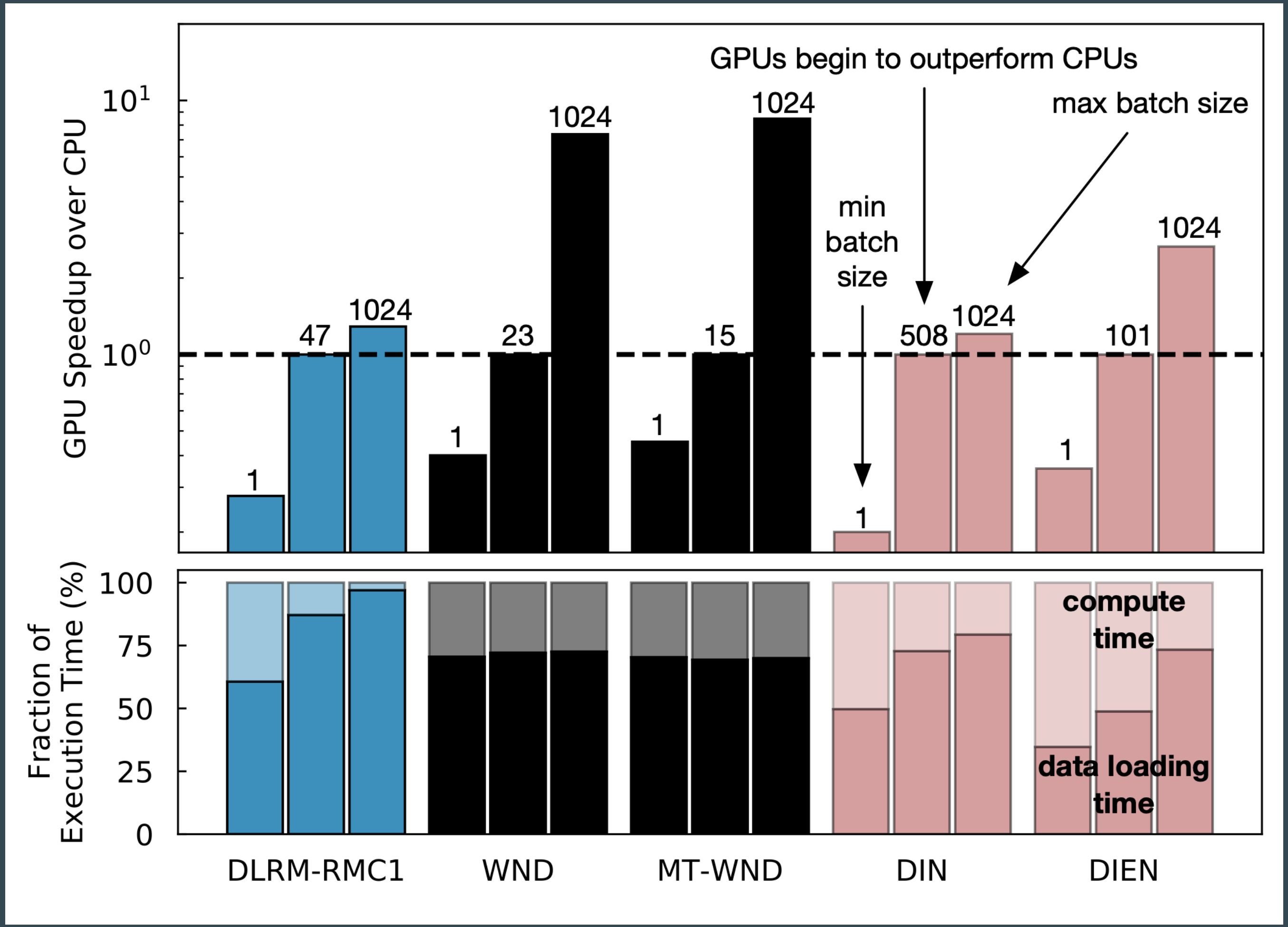
Attention-dominated



Complex *attention* and *sequential modeling* for feature interaction

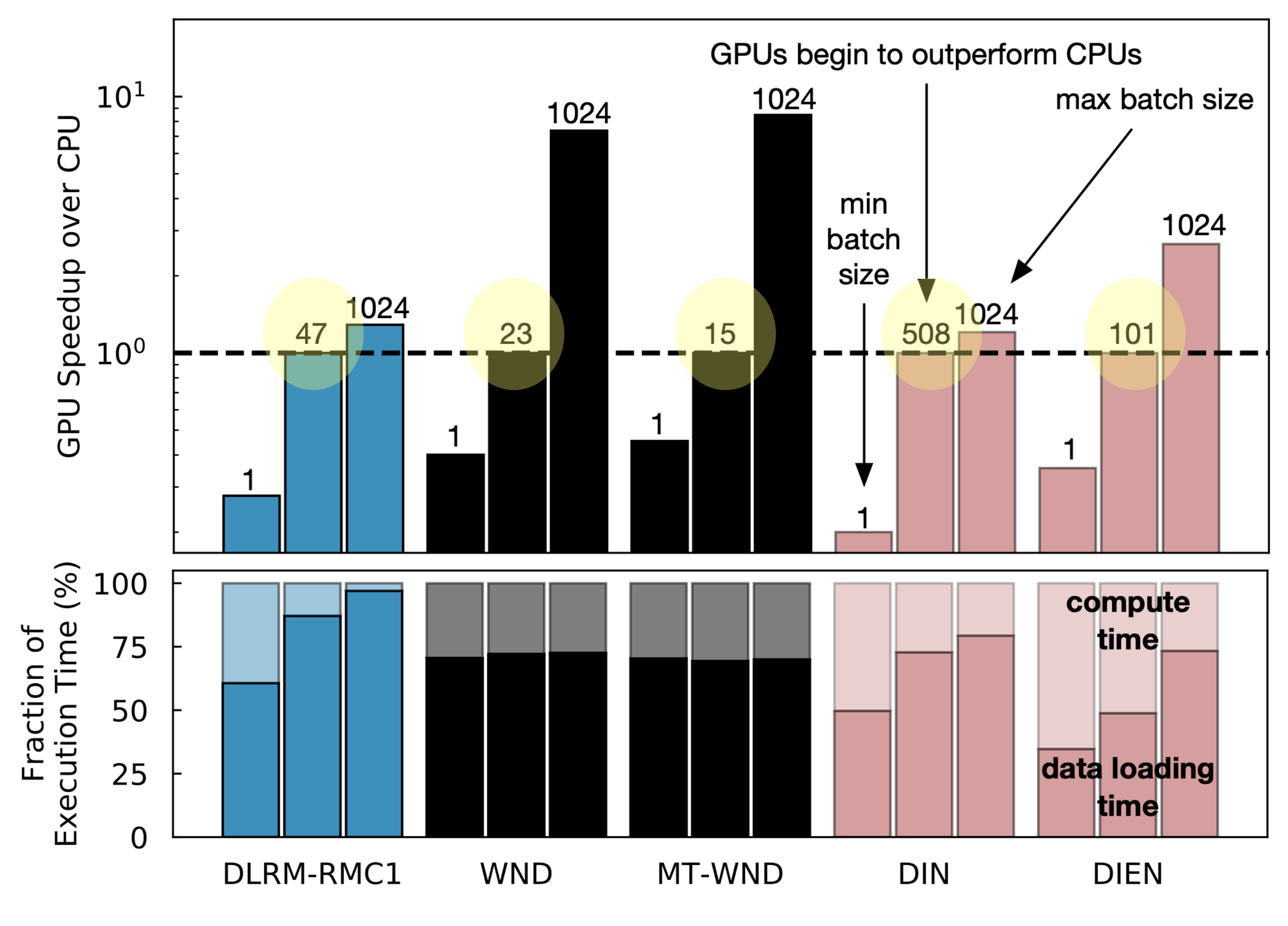
Challenge: Optimal System Config Varies

Batch Sizes, Compute Platforms



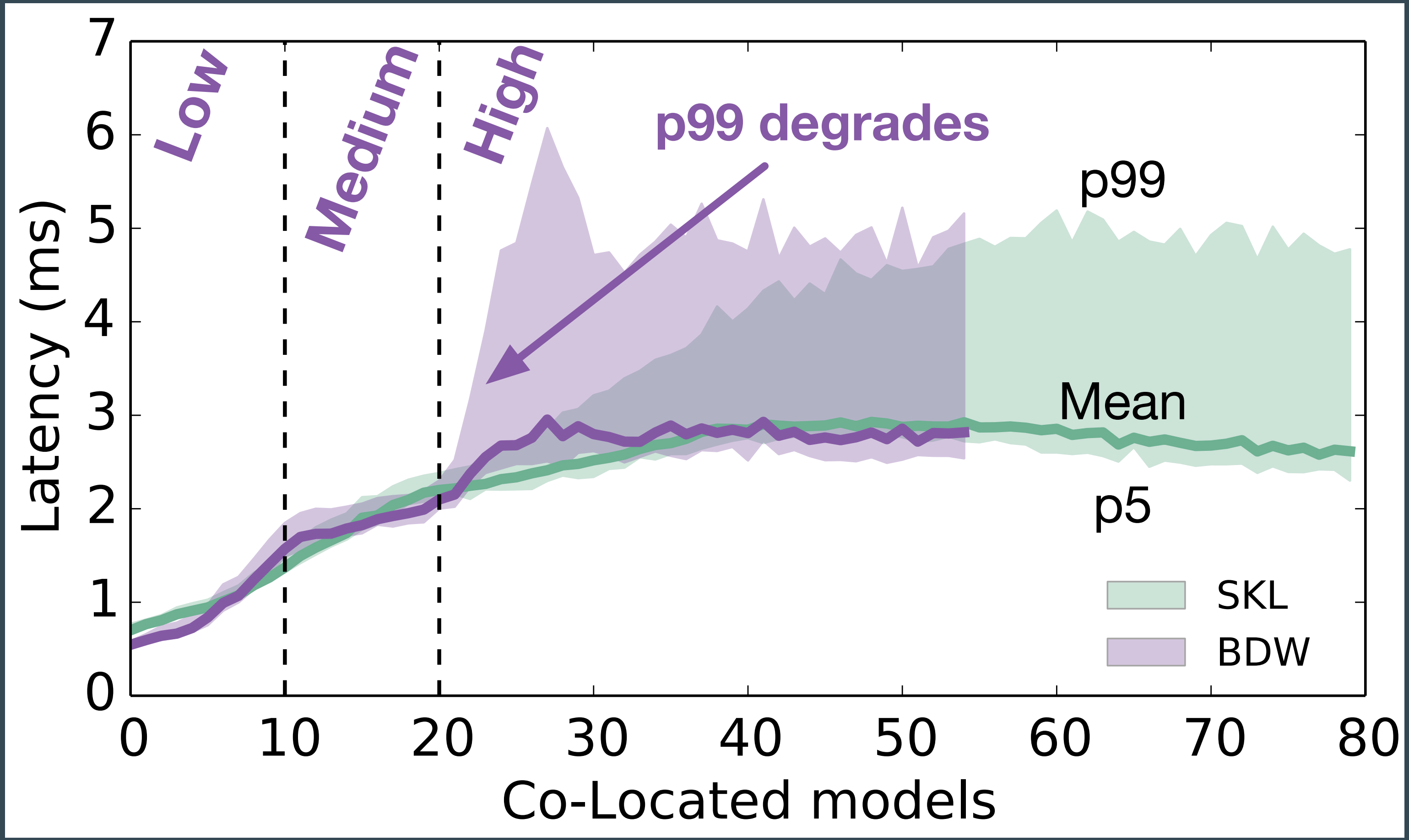
Challenge: Optimal System Config Varies

Batch Sizes, Compute Platforms



Challenge: Performance Variance

Co-Location Across Different Compute Platforms



Agenda

Motivation

Understanding the Unique Systems Challenges

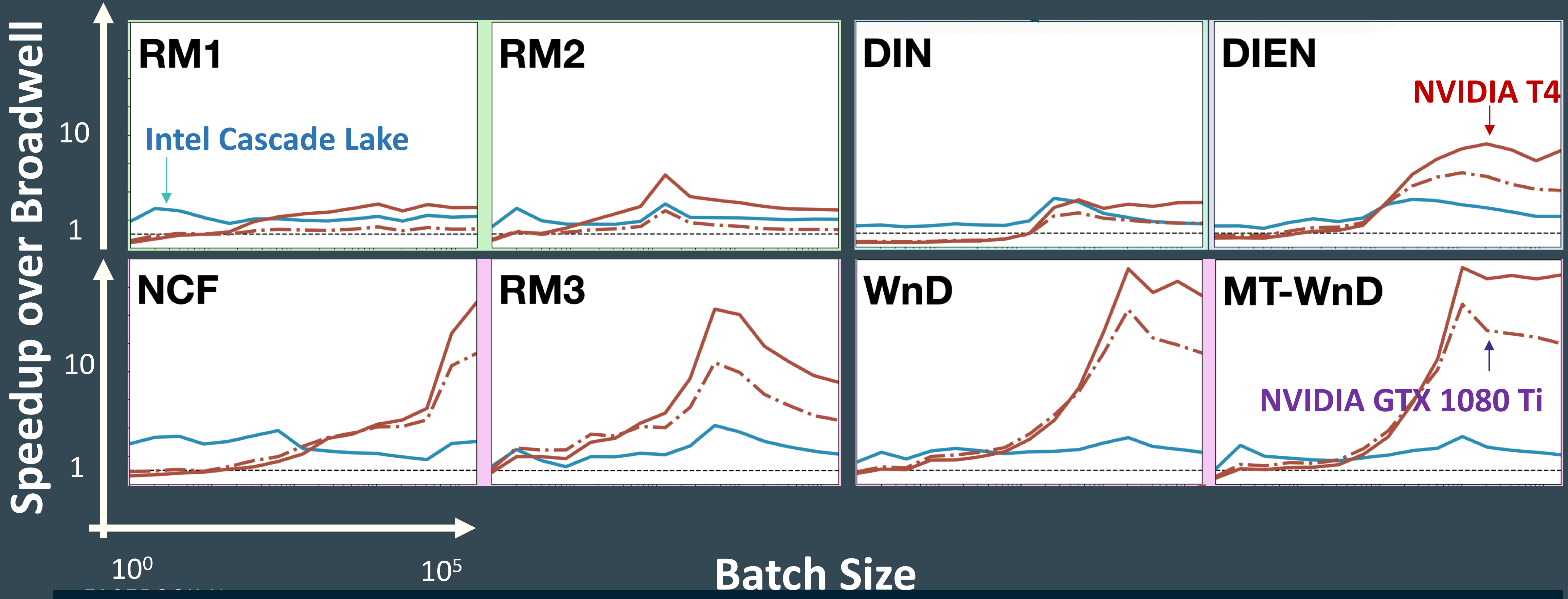
Characterizing Performance Acceleration with GPUs

Optimizing Neural Recommendation Inference At-Scale

Conclusion and Future Work

System Implications of Model Heterogeneity

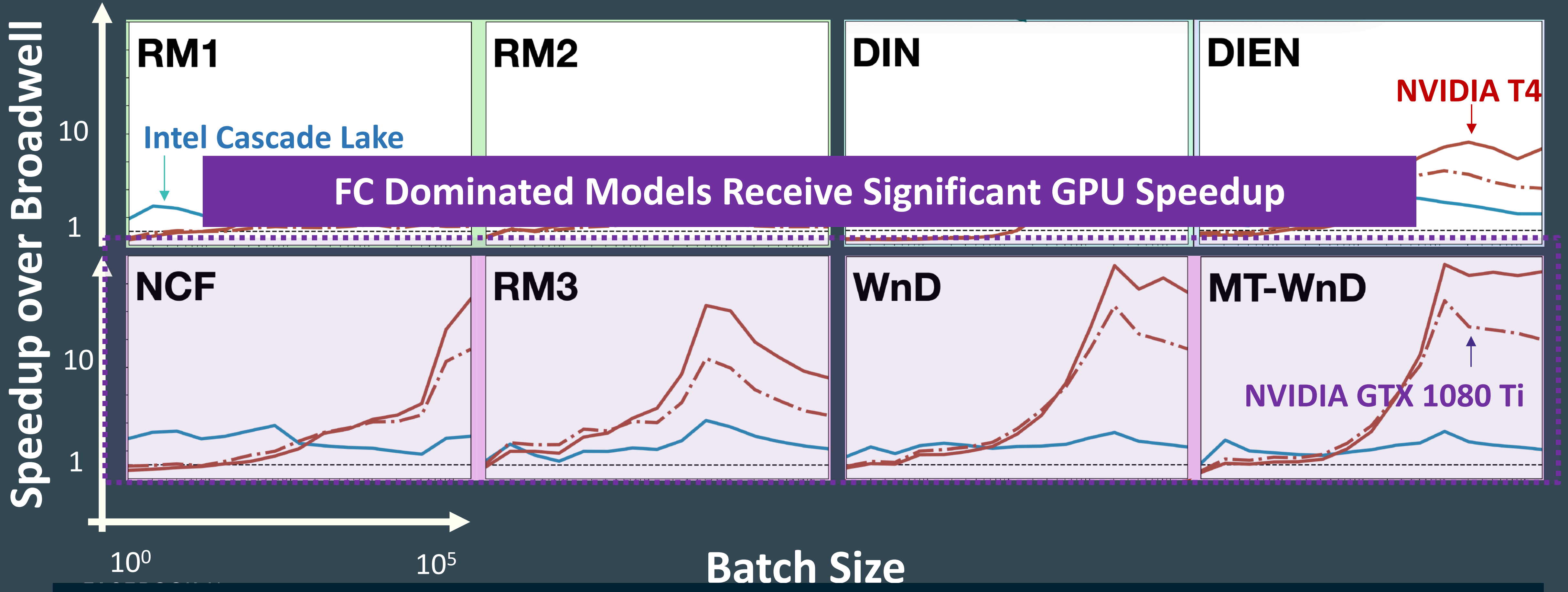
Model Architectures Play a Significant Role in Recommendation Inference Acceleration



[IISWC 2020] **Cross-Stack Workload Characterization of Deep Recommendation Systems.** S. Hsia, U. Gupta, M. Wilkening, C.-J. Wu, G.-Y. Wei, D. Brooks.

System Implications of Model Heterogeneity

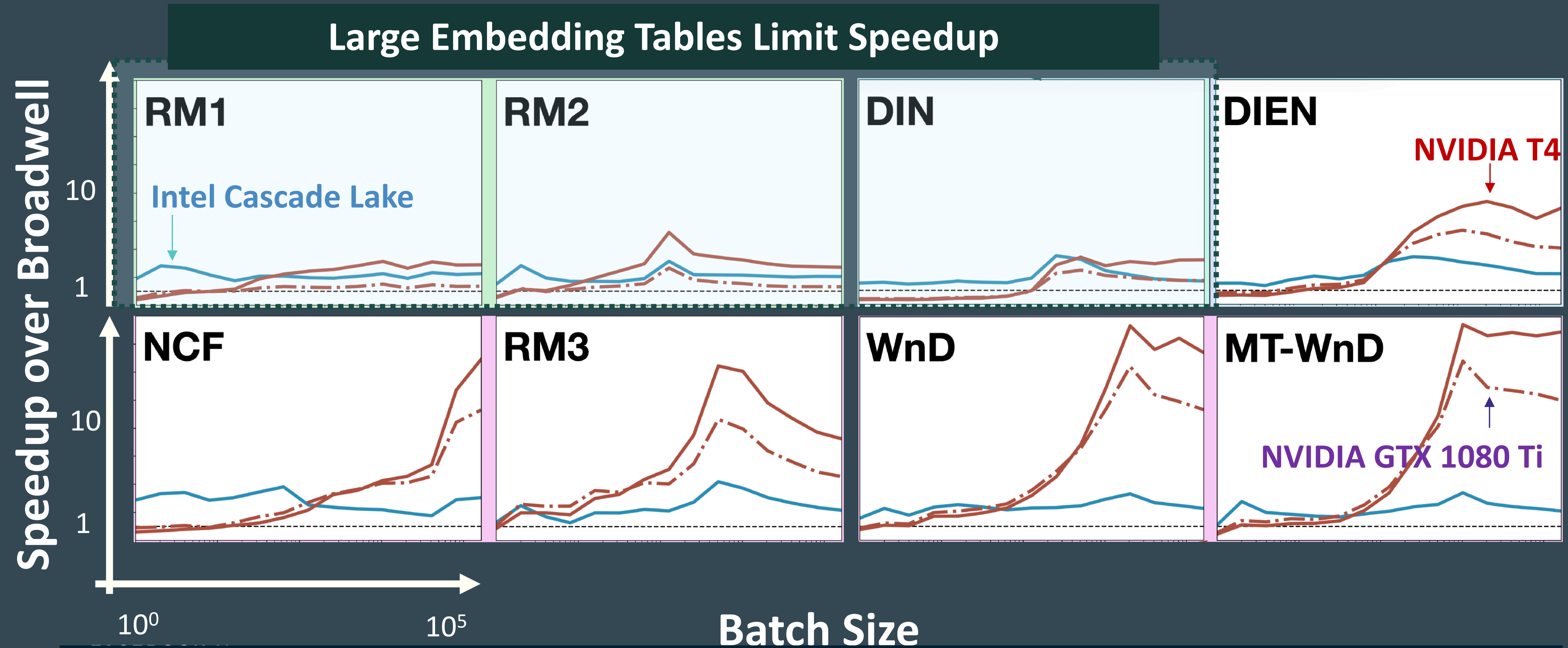
Model Architectures Play a Significant Role in Recommendation Inference Acceleration



[IISWC 2020] **Cross-Stack Workload Characterization of Deep Recommendation Systems.** S. Hsia, U. Gupta, M. Wilkening, C.-J. Wu, G.-Y. Wei, D. Brooks.

System Implications of Model Heterogeneity

Model Architectures Play a Significant Role in Recommendation Inference Acceleration

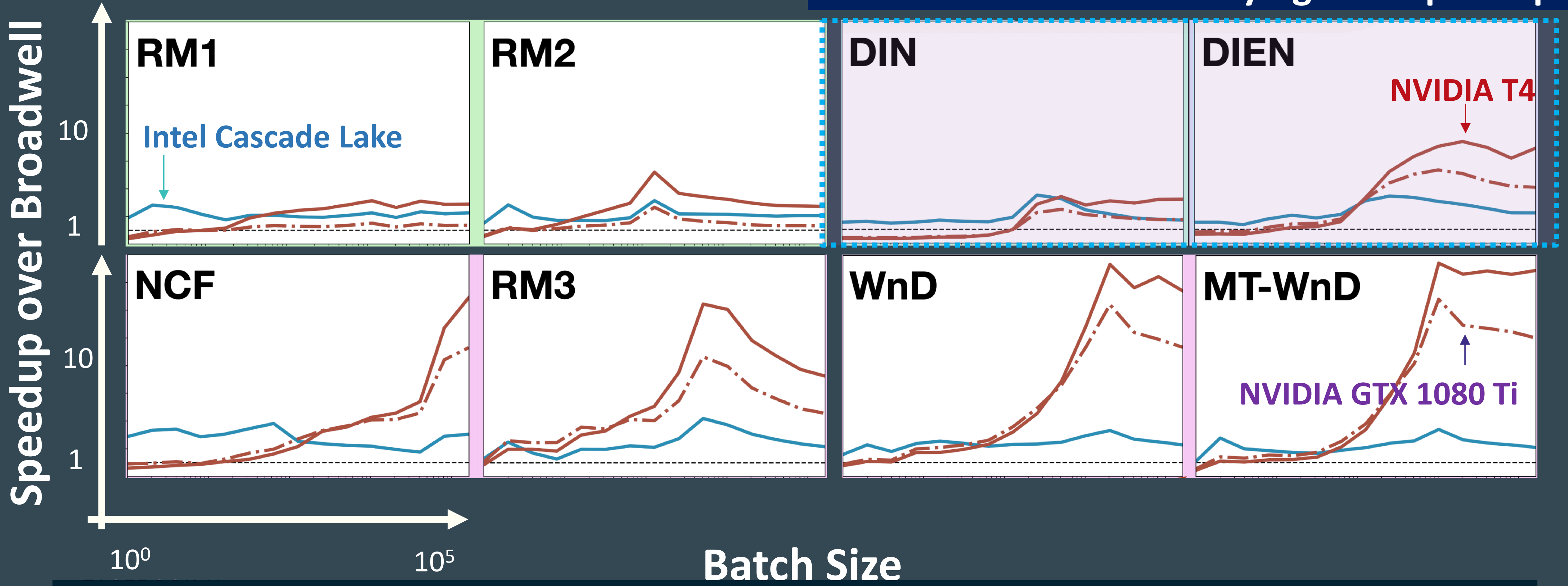


[IISWC 2020] **Cross-Stack Workload Characterization of Deep Recommendation Systems.** S. Hsia, U. Gupta, M. Wilkening, C.-J. Wu, G.-Y. Wei, D. Brooks.

System Implications of Model Heterogeneity

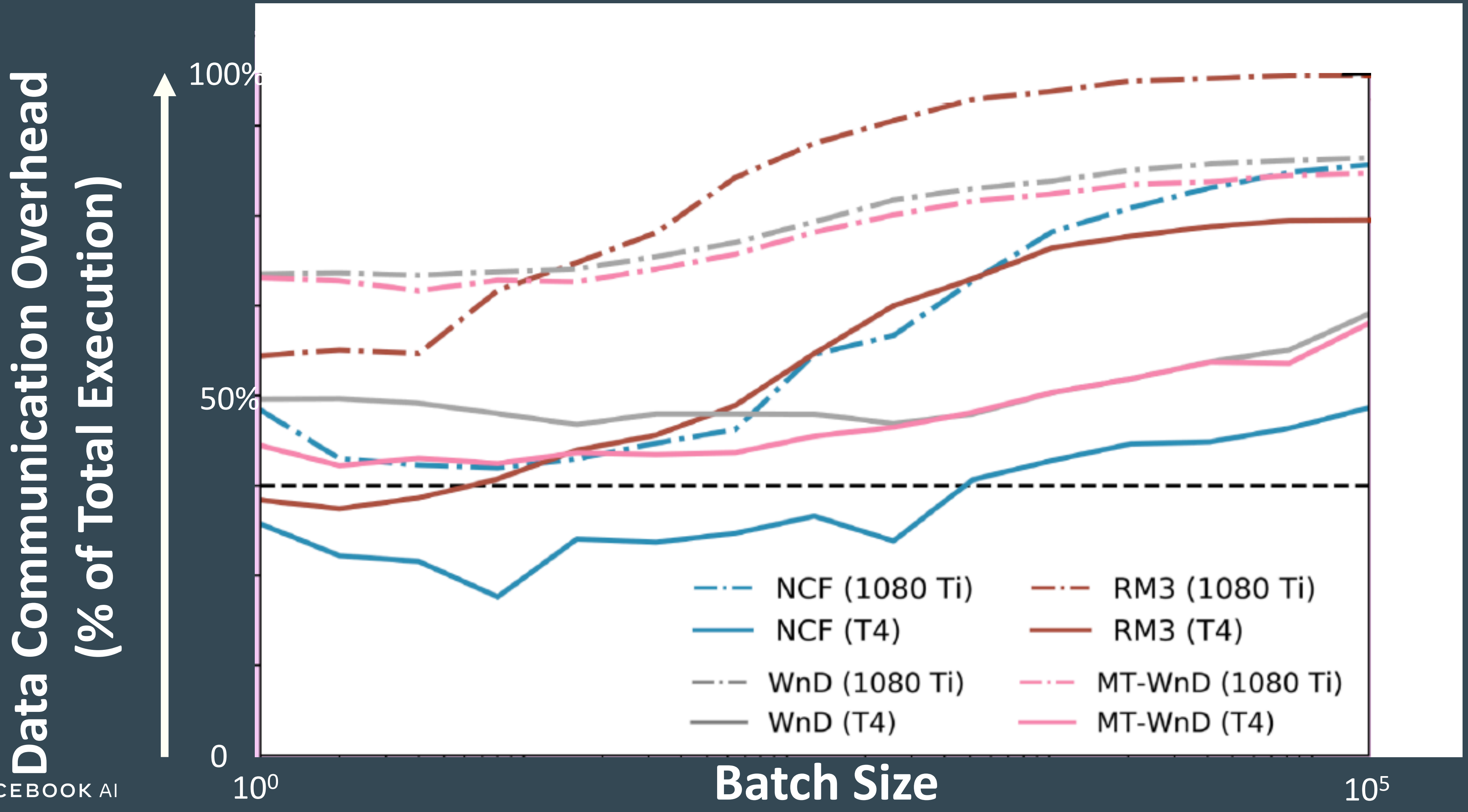
Model Architectures Play a Significant Role in Recommendation Inference Acceleration

Different Implementations of Attention Mechanisms Lead to Varying GPU Speedup



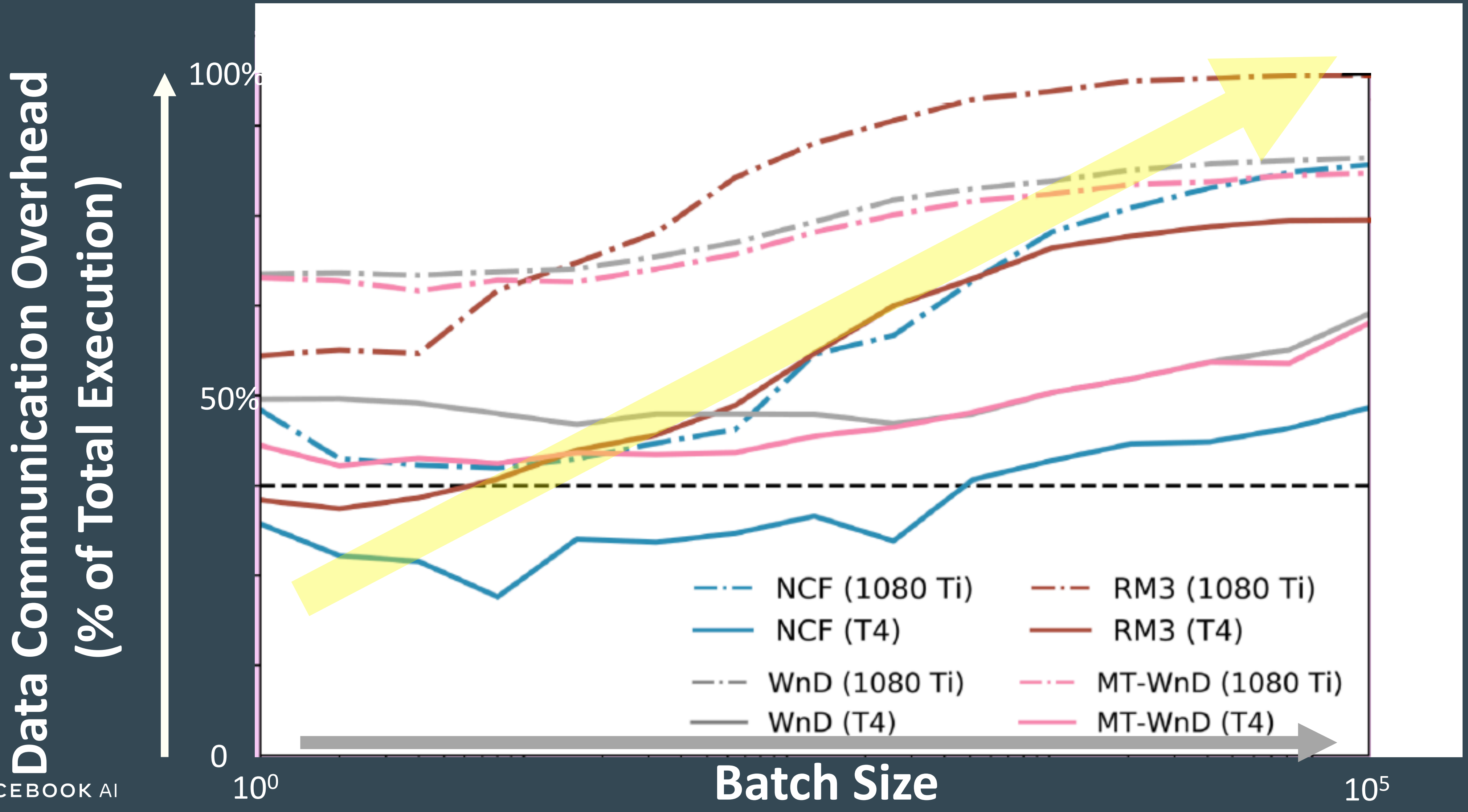
[IISWC 2020] **Cross-Stack Workload Characterization of Deep Recommendation Systems.** S. Hsia, U. Gupta, M. Wilkening, C.-J. Wu, G.-Y. Wei, D. Brooks.

Speedup Limited by Data Communication

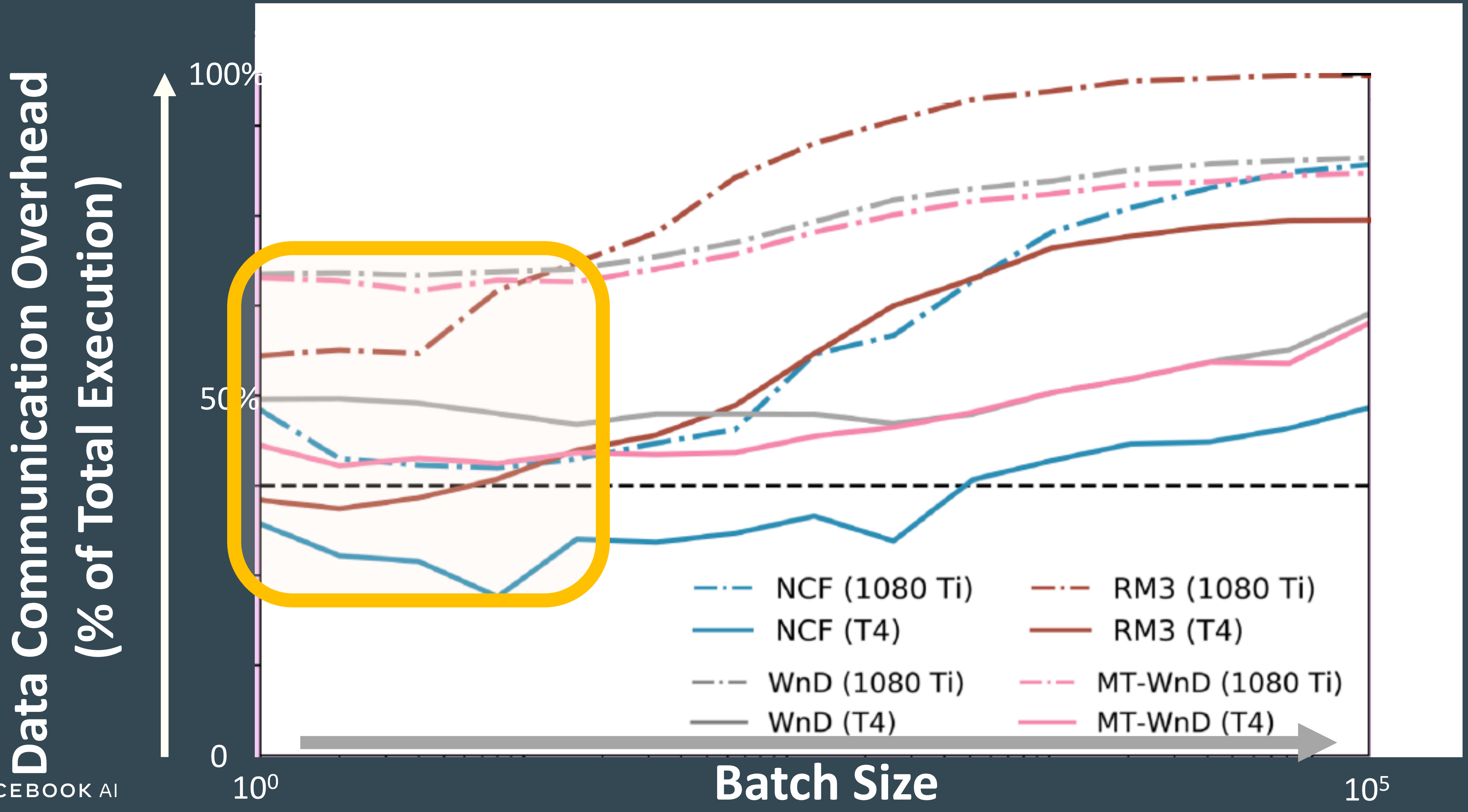


FACEBOOK AI

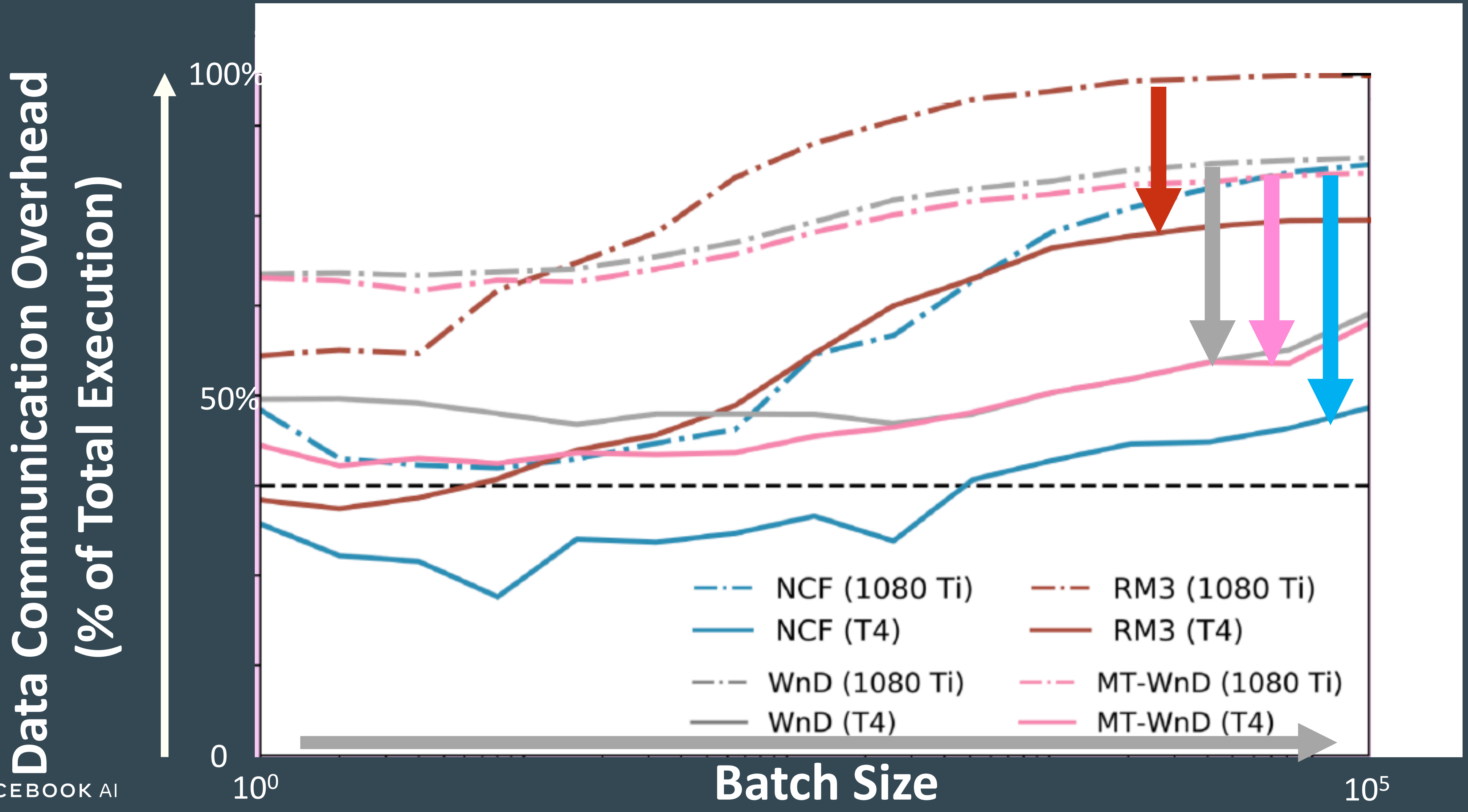
Speedup Limited by Data Communication



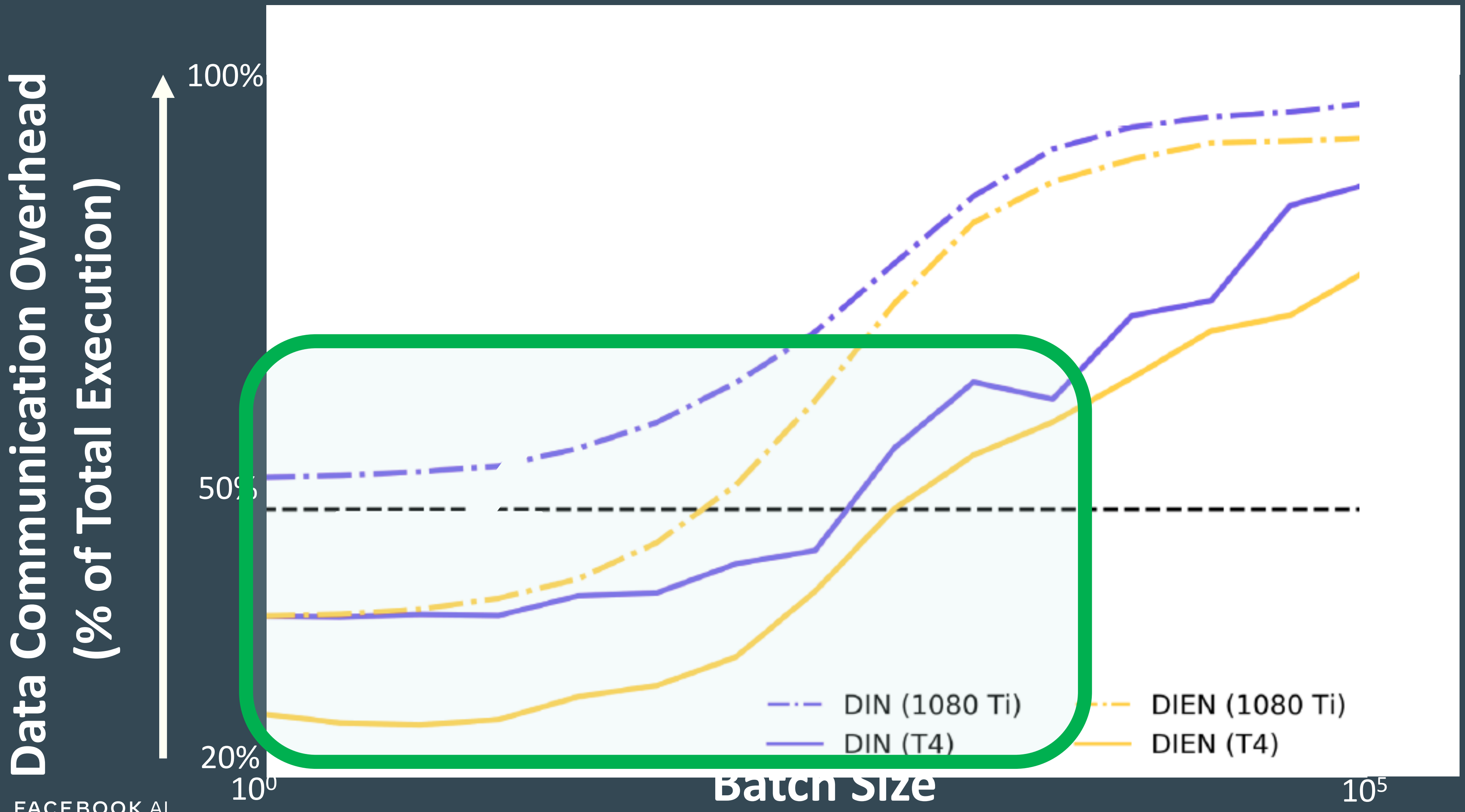
Speedup Limited by Data Communication



Speedup Limited by Data Communication

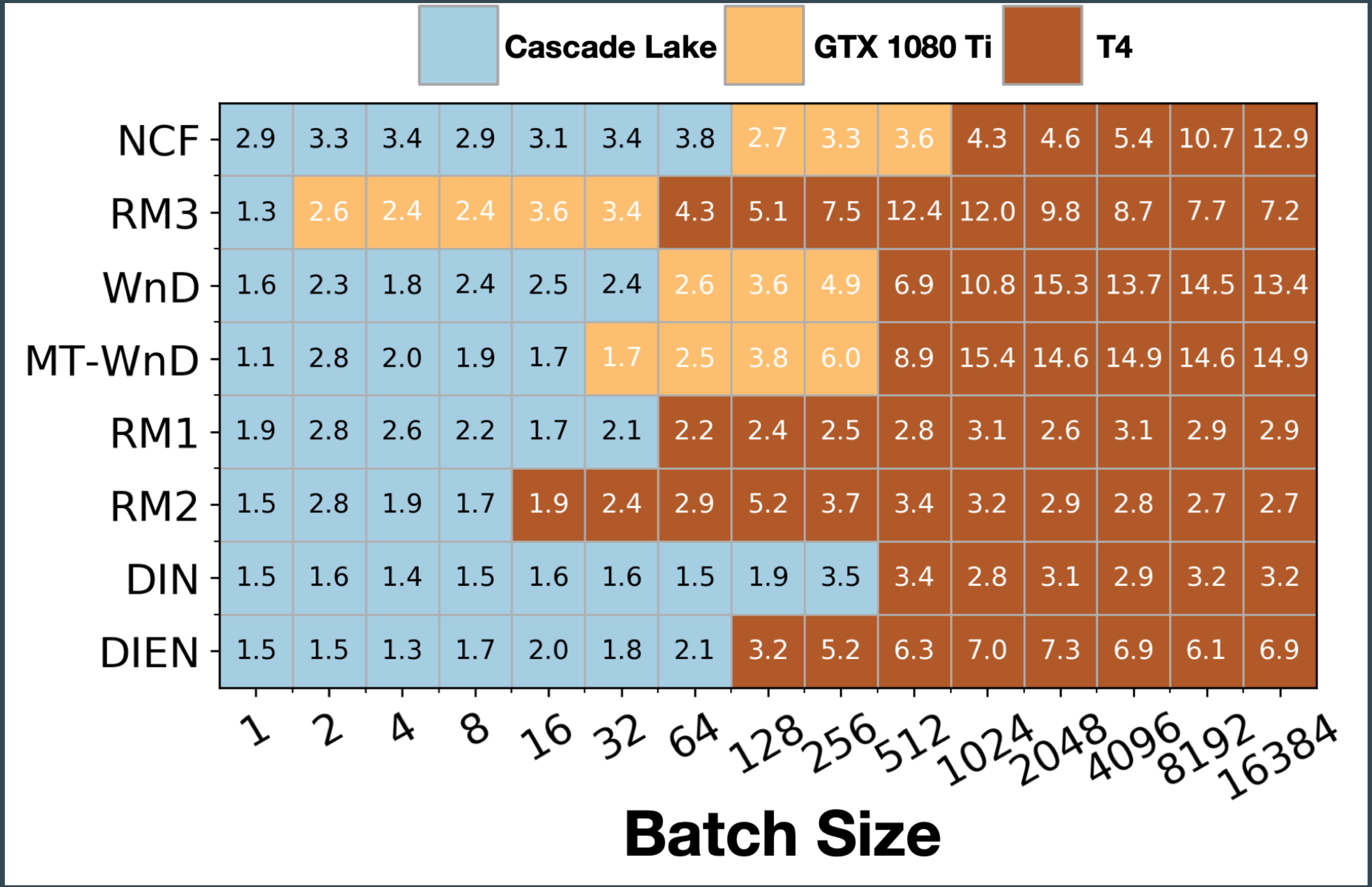


Speedup Limited by Data Communication



Optimal Hardware Varies

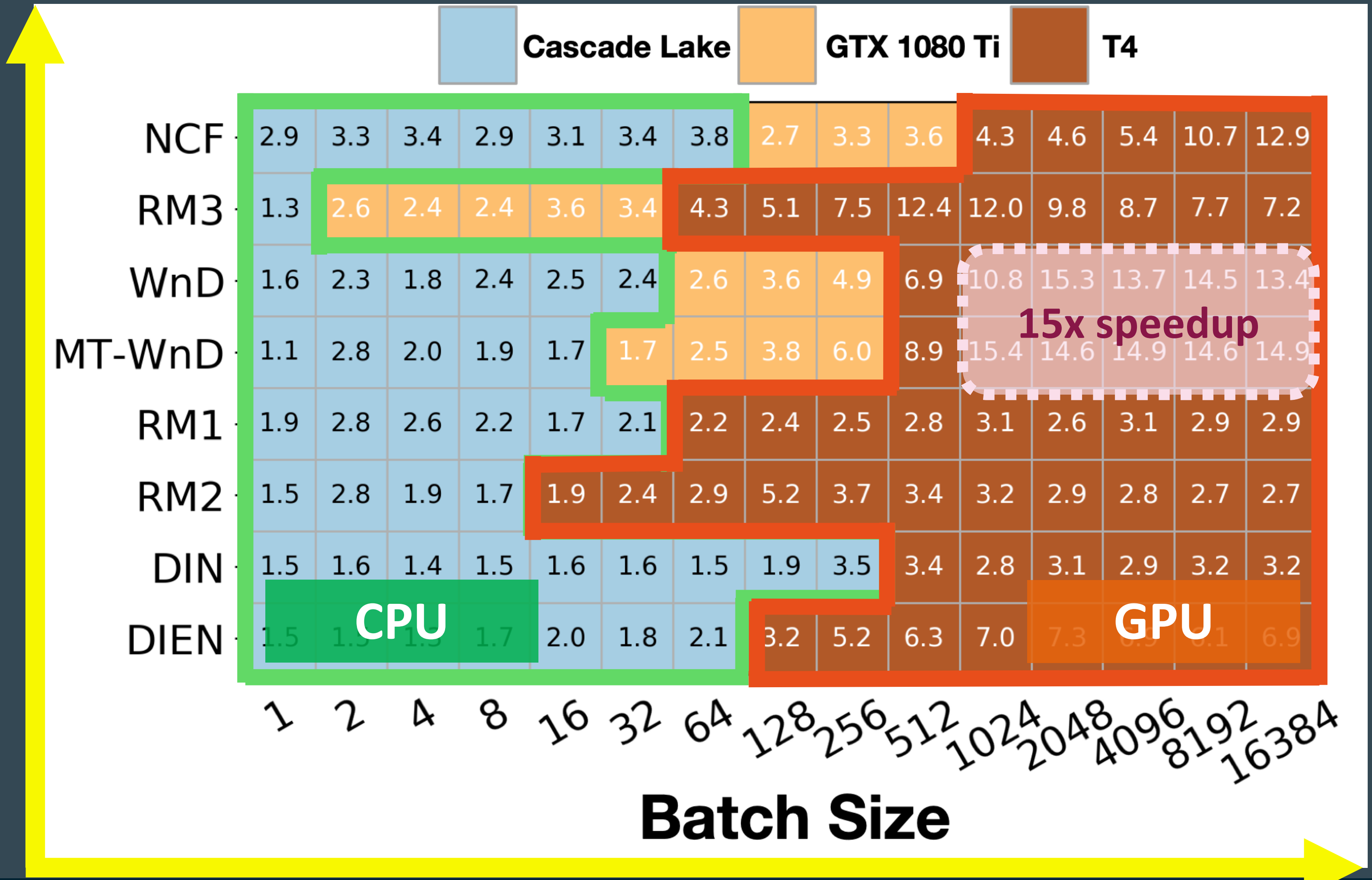
Across Model Architectures and Input Batch Size



[IISWC 2020] **Cross-Stack Workload Characterization of Deep Recommendation Systems.** S. Hsia, U. Gupta, M. Wilkening, C.-J. Wu, G.-Y. Wei, D. Brooks.

Optimal Hardware Varies

Across Model Architectures and Input Batch Size



[IISWC 2020] **Cross-Stack Workload Characterization of Deep Recommendation Systems.** S. Hsia, U. Gupta, M. Wilkening, C.-J. Wu, G.-Y. Wei, D. Brooks.

Agenda

Motivation

Understanding the Unique Systems Challenges

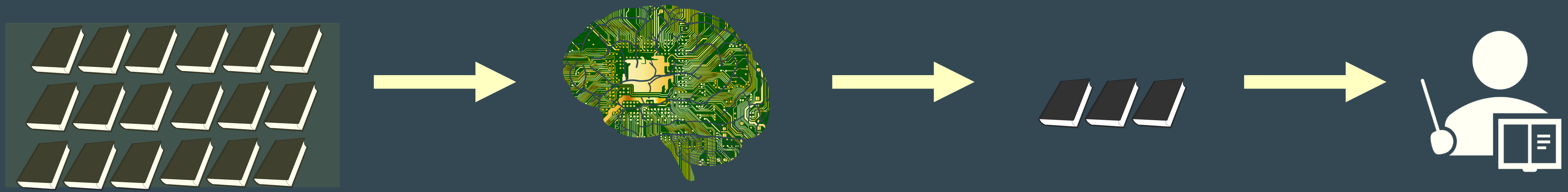
Characterizing Performance Acceleration with GPUs

Optimizing Neural Recommendation Inference At-Scale

Conclusion and Future Work

Let’s Consider Runtime Effects

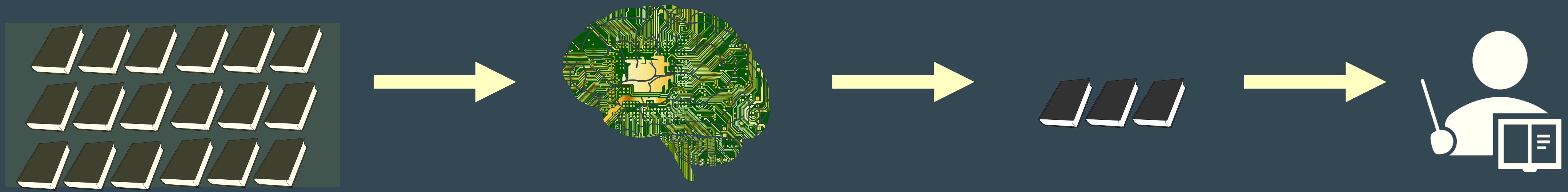
Number of Items to Rank Varies across Queries



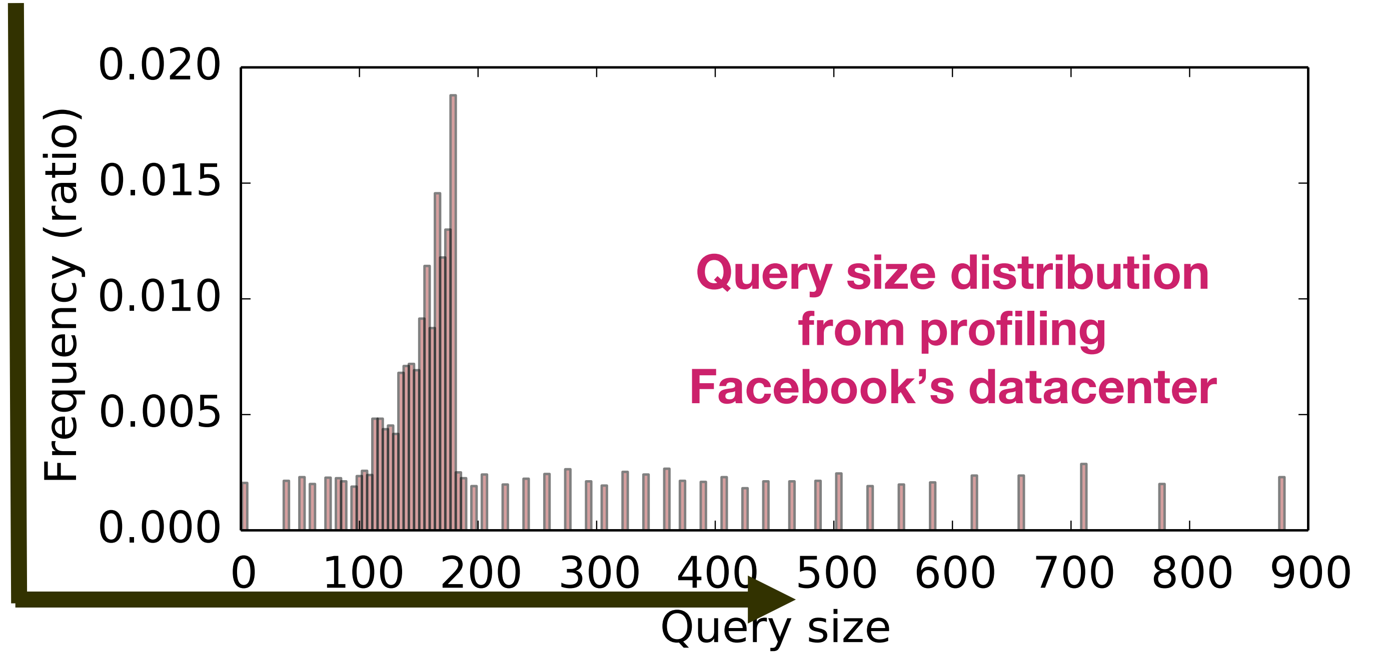
Number of candidate items

Let’s Consider Runtime Effects

Number of Items to Rank Varies across Queries



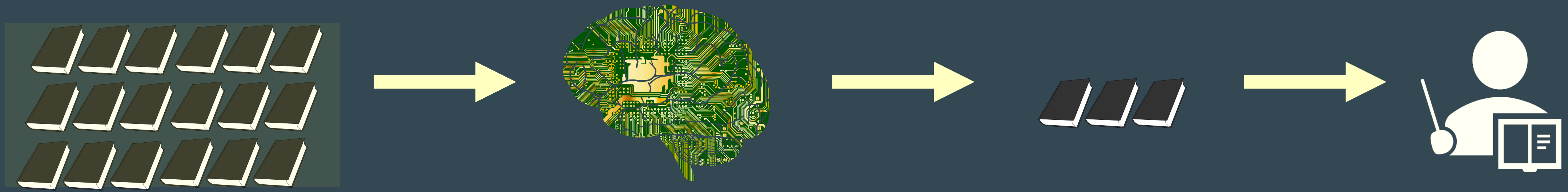
Number of candidate items



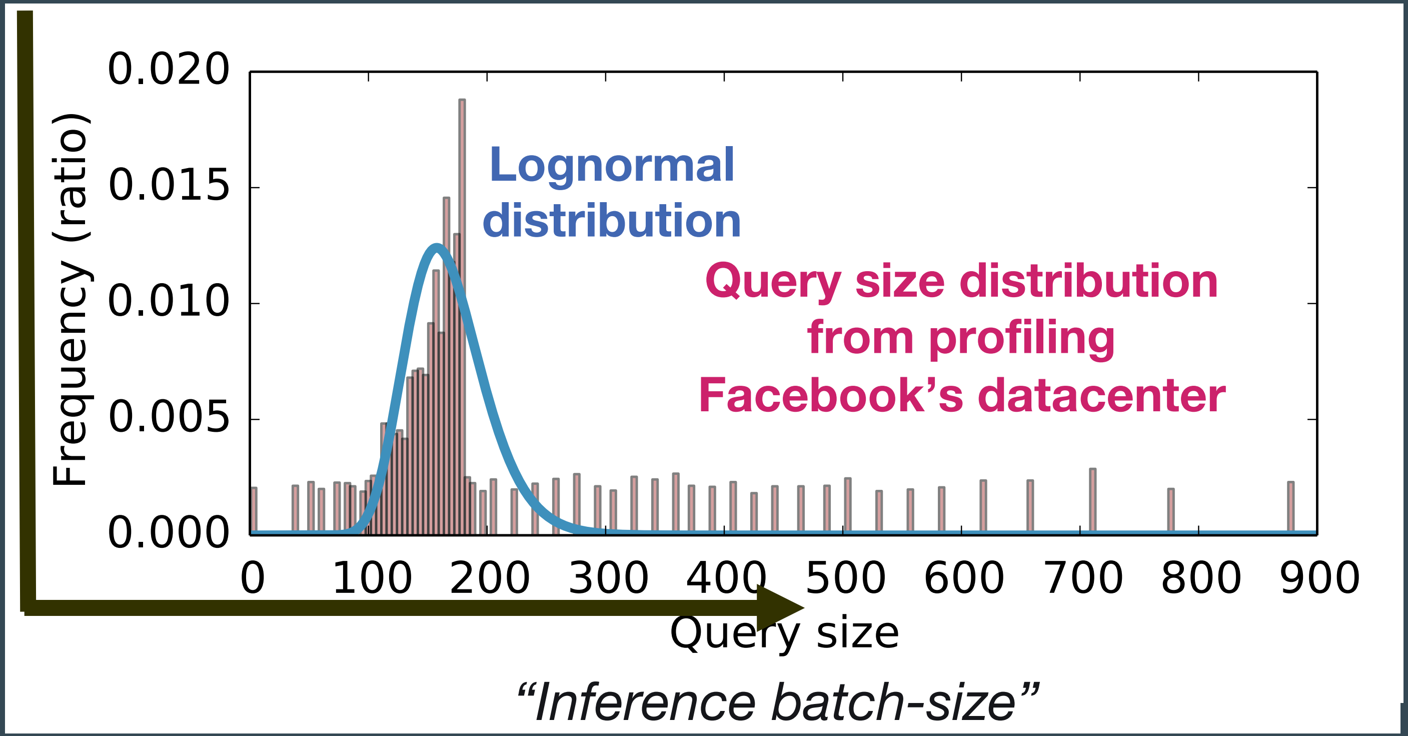
“Inference batch-size”

Let’s Consider Runtime Effects

Number of Items to Rank Varies across Queries

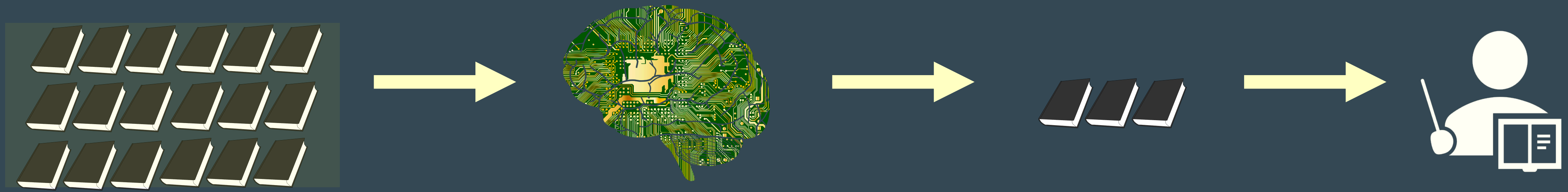


Number of candidate items

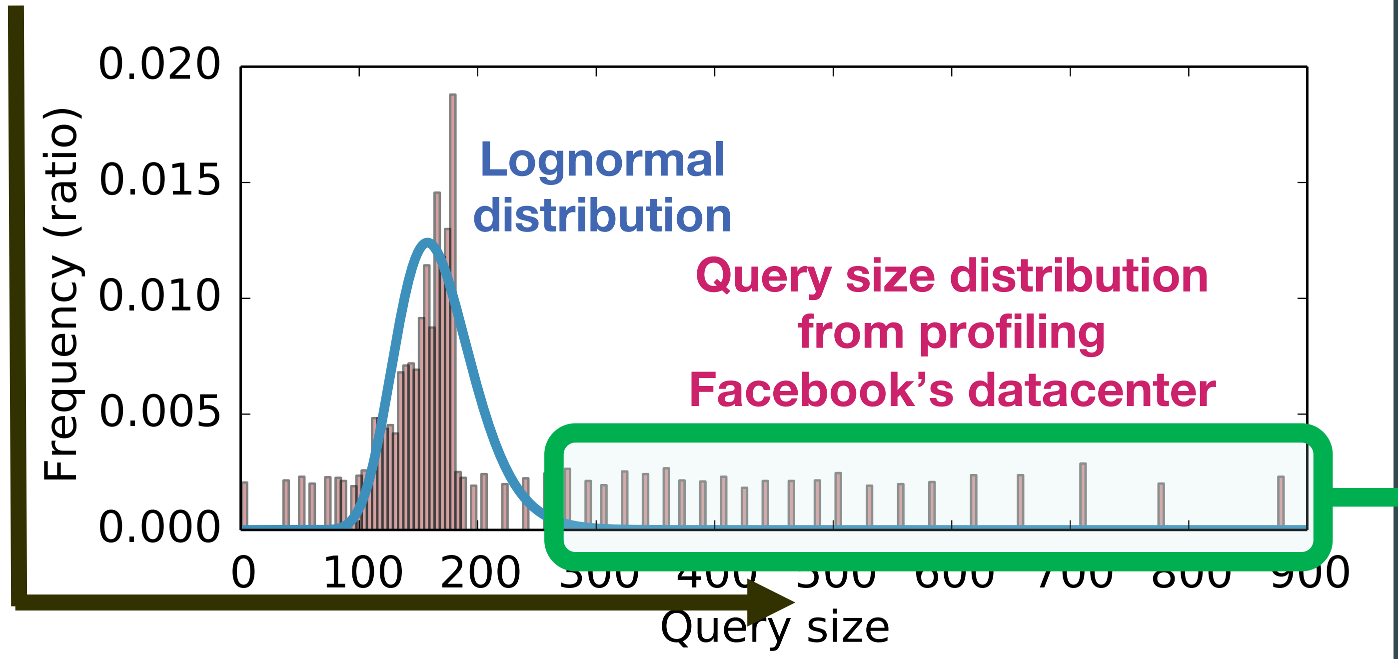


Let’s Consider Runtime Effects

Number of Items to Rank Varies across Queries



Number of candidate items



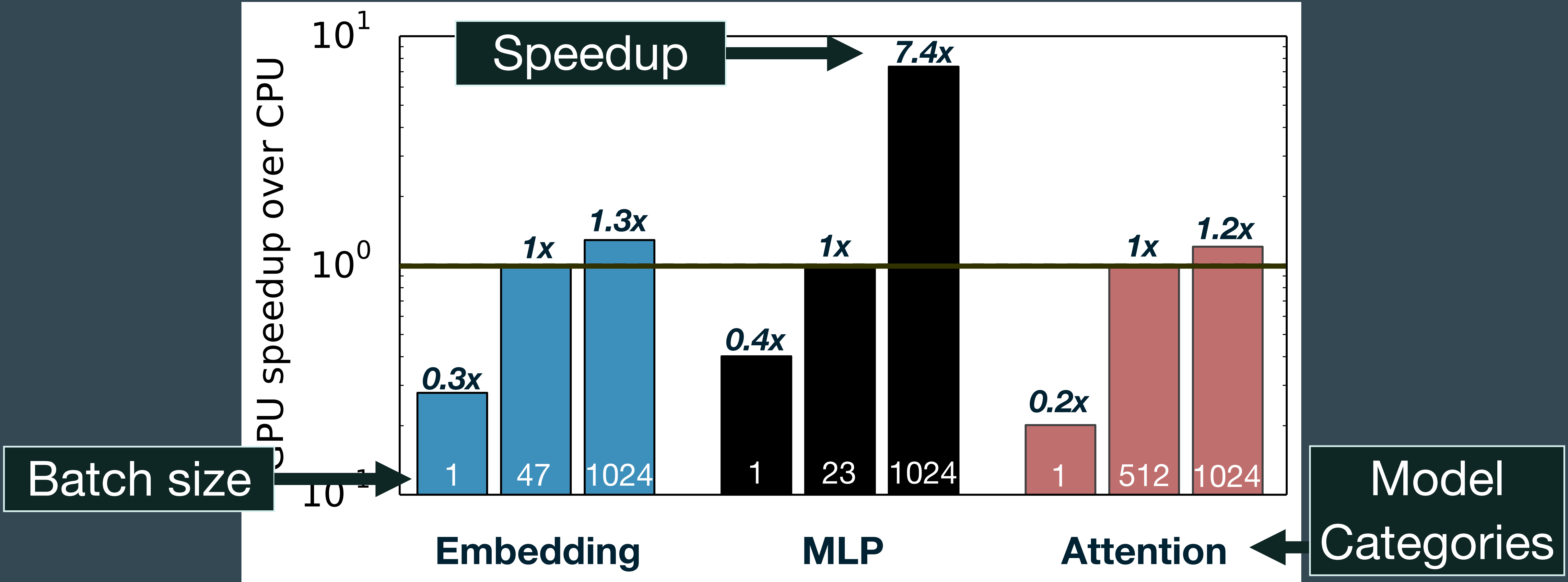
“Inference batch-size”

Acceleration opportunity

System Heterogeneity At-Scale

Optimal Hardware and Batch Sizes and Vary

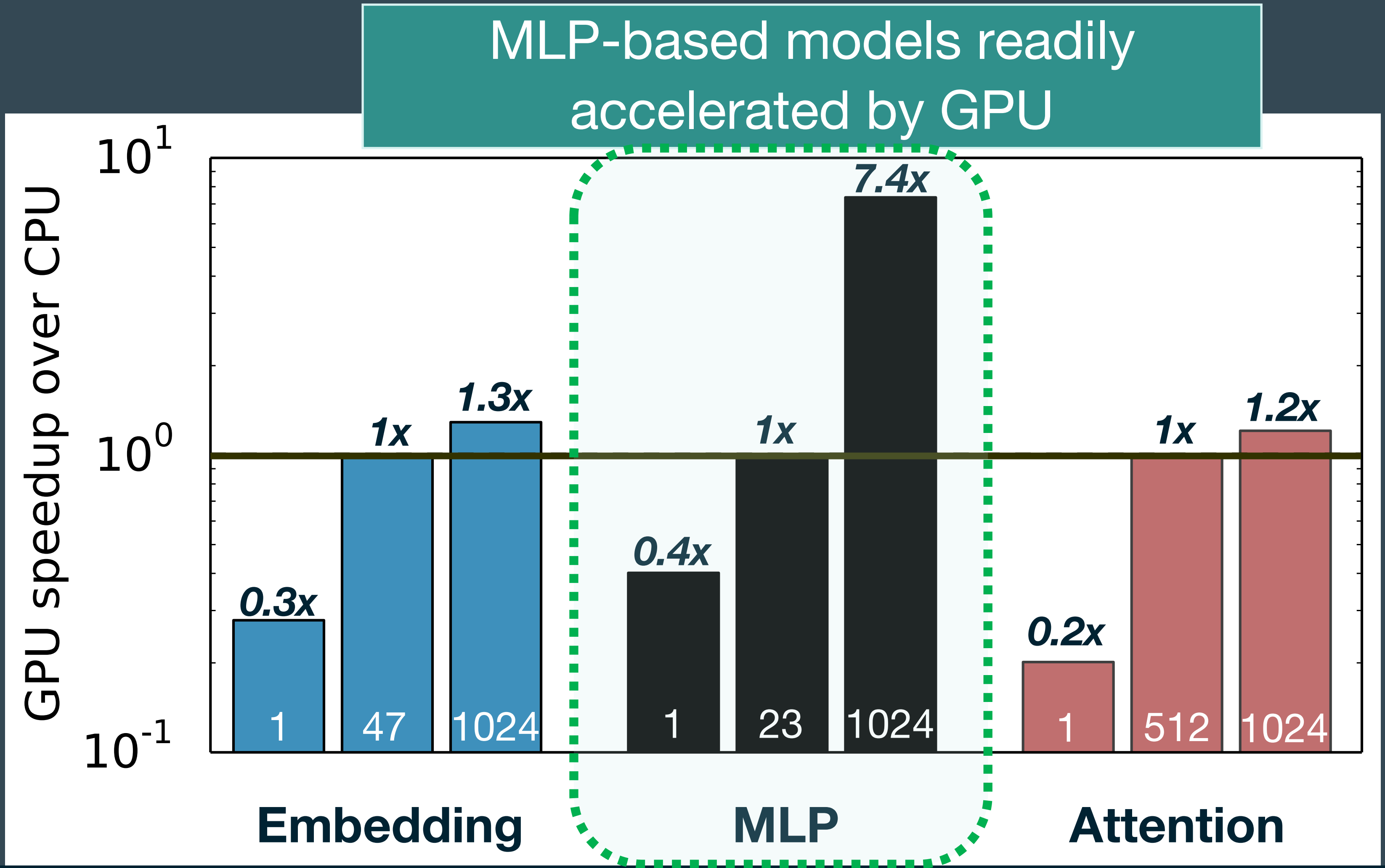
When considering runtime effects



[ISCA 2020] **DeepRecSys: A System for Optimizing End-to-end At-scale Neural Recommendation Inference.** U. Gupta, S. Hsia, V. Saraph, X. Wang, B. Reagen, G.-Y. Wei, H.-S. Lee, D. Brooks, and C.-J. Wu.

Optimal Hardware and Batch Sizes and Vary

When considering runtime effects



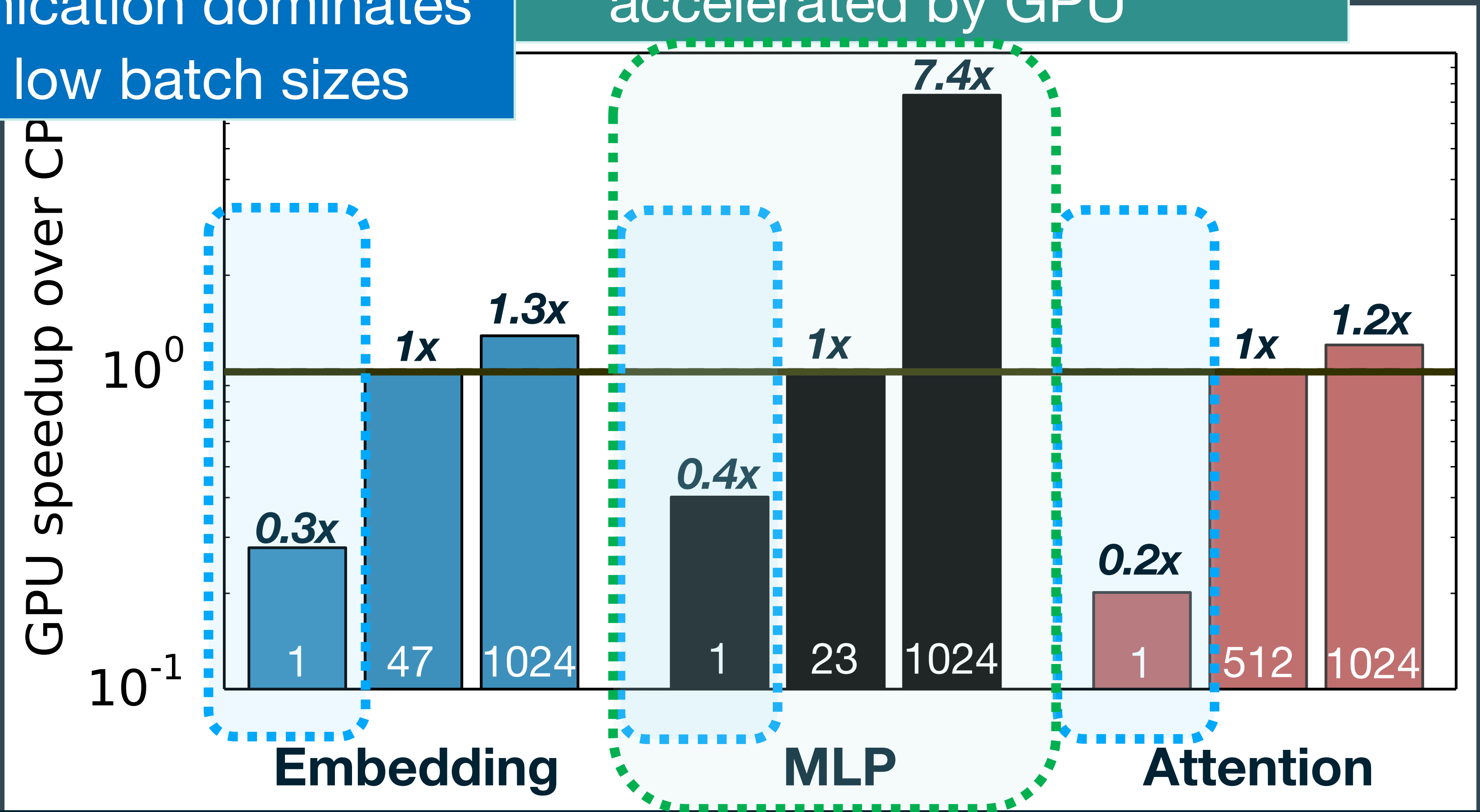
[ISCA 2020] **DeepRecSys: A System for Optimizing End-to-end At-scale Neural Recommendation Inference.** U. Gupta, S. Hsia, V. Saraph, X. Wang, B. Reagen, G.-Y. Wei, H.-S. Lee, D. Brooks, and C.-J. Wu.

Optimal Hardware and Batch Sizes and Vary

When considering runtime effects

Data communication dominates run-time at low batch sizes

MLP-based models readily accelerated by GPU



[ISCA 2020] **DeepRecSys: A System for Optimizing End-to-end At-scale Neural Recommendation Inference.** U. Gupta, S. Hsia, V. Saraph, X. Wang, B. Reagen, G.-Y. Wei, H.-S. Lee, D. Brooks, and C.-J. Wu.

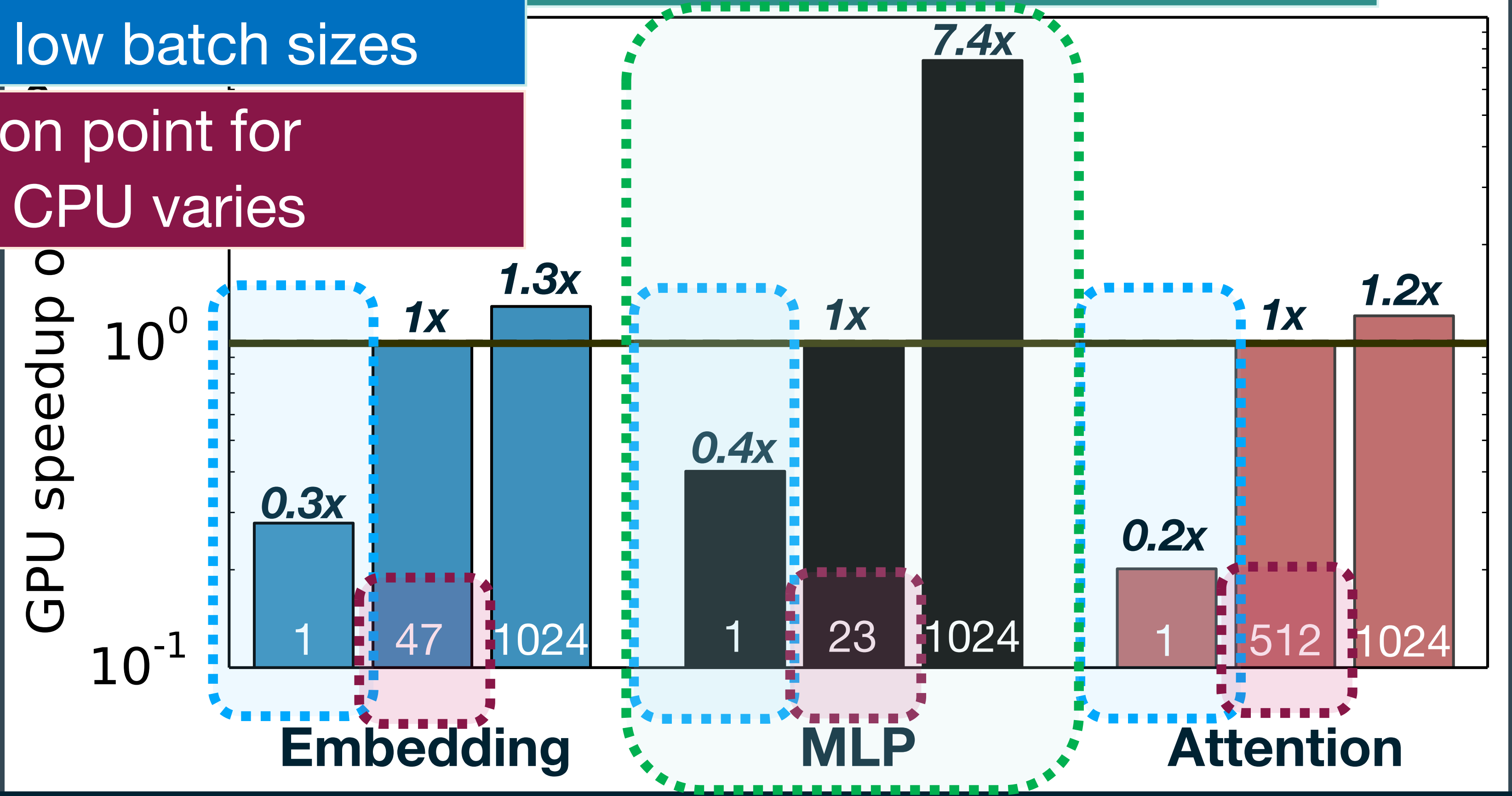
Optimal Hardware and Batch Sizes and Vary

When considering runtime effects

MLP-based models readily accelerated by GPU

Data communication dominates run-time at low batch sizes

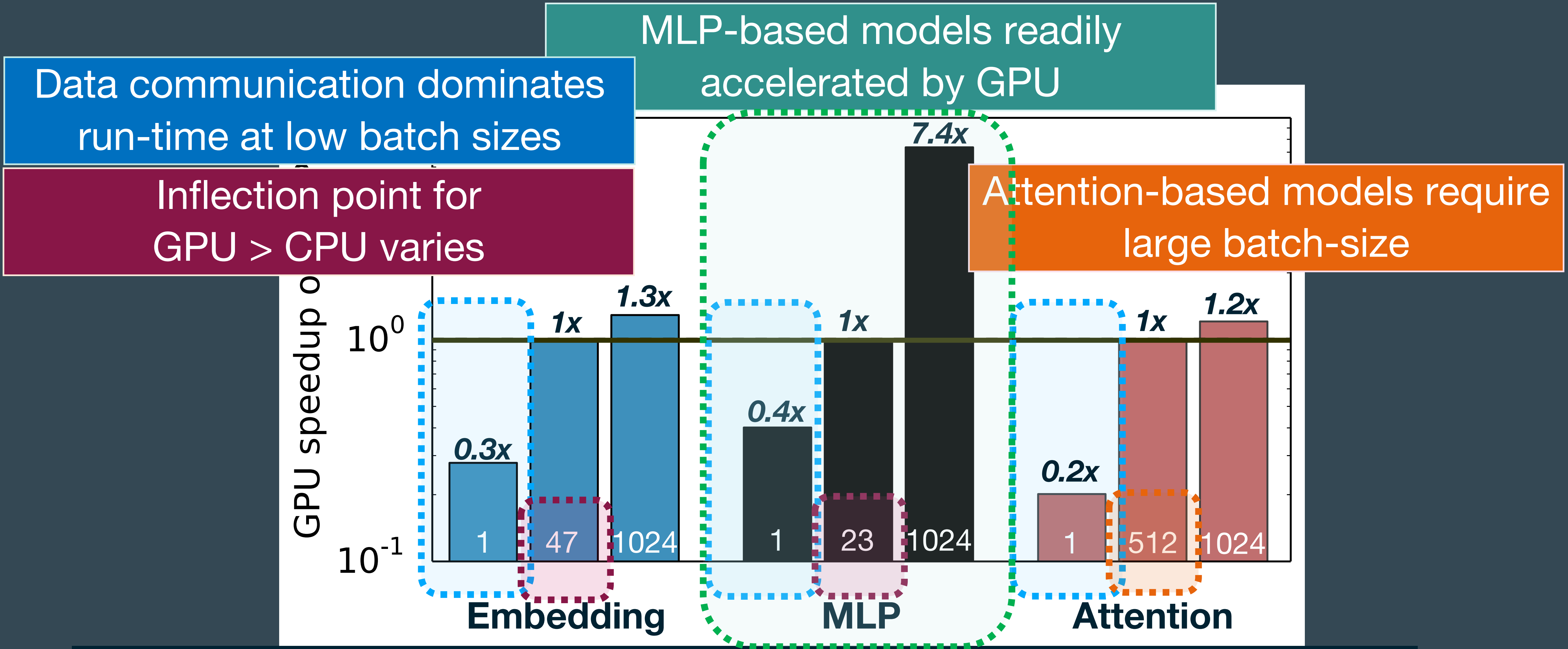
Inflection point for GPU > CPU varies



[ISCA 2020] **DeepRecSys: A System for Optimizing End-to-end At-scale Neural Recommendation Inference.** U. Gupta, S. Hsia, V. Saraph, X. Wang, B. Reagen, G.-Y. Wei, H.-S. Lee, D. Brooks, and C.-J. Wu.

Optimal Hardware and Batch Sizes and Vary

When considering runtime effects

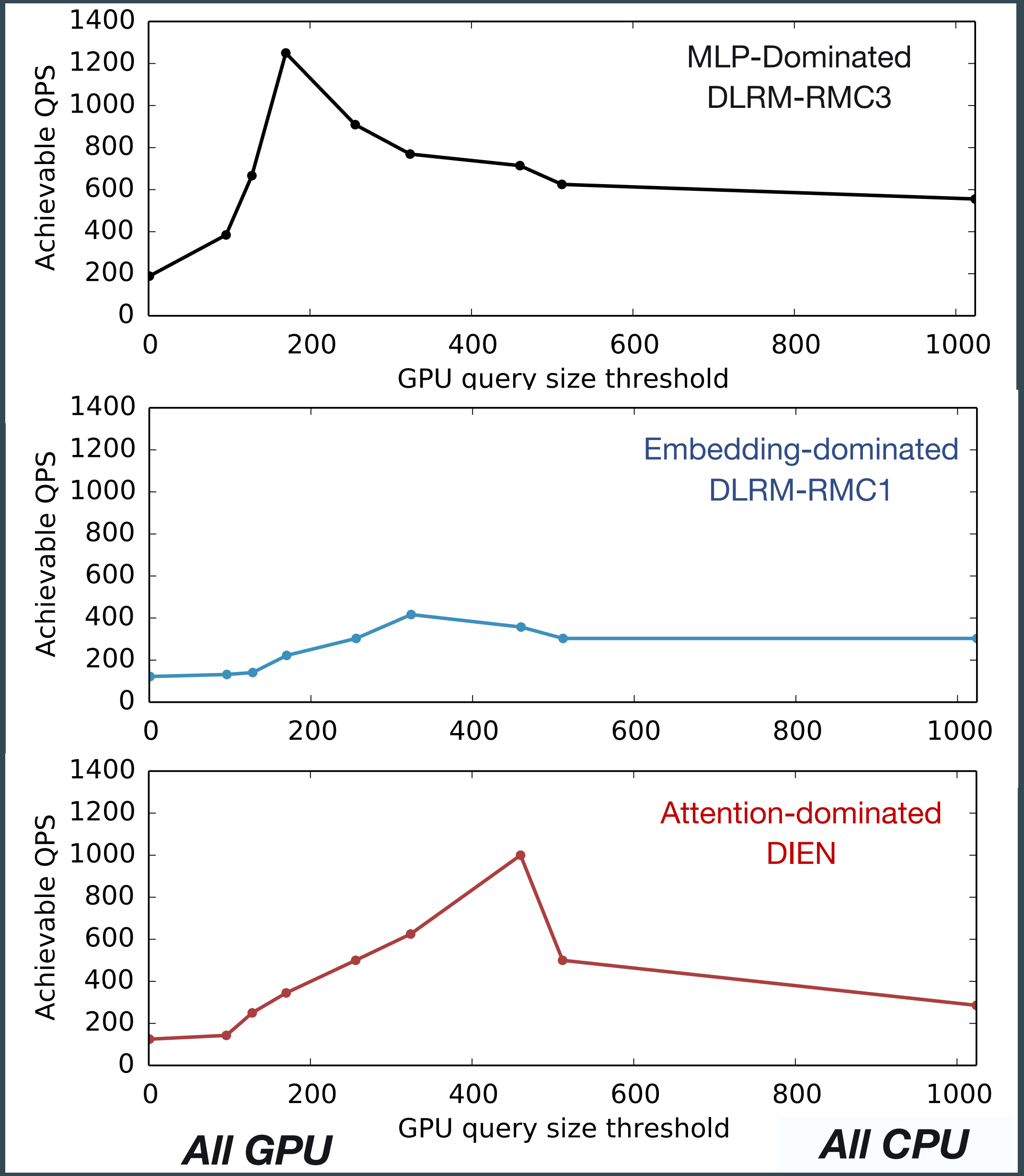


[ISCA 2020] **DeepRecSys: A System for Optimizing End-to-end At-scale Neural Recommendation Inference.** U. Gupta, S. Hsia, V. Saraph, X. Wang, B. Reagen, G.-Y. Wei, H.-S. Lee, D. Brooks, and C.-J. Wu.

Latency-bound QPS Optimization

Optimal execution depends on

- Recommendation models
- AI system architectures
  - CPUs vs. AI accelerators
- Runtime characteristics
  - Query arrival and working set sizes
- Application SLA requirement

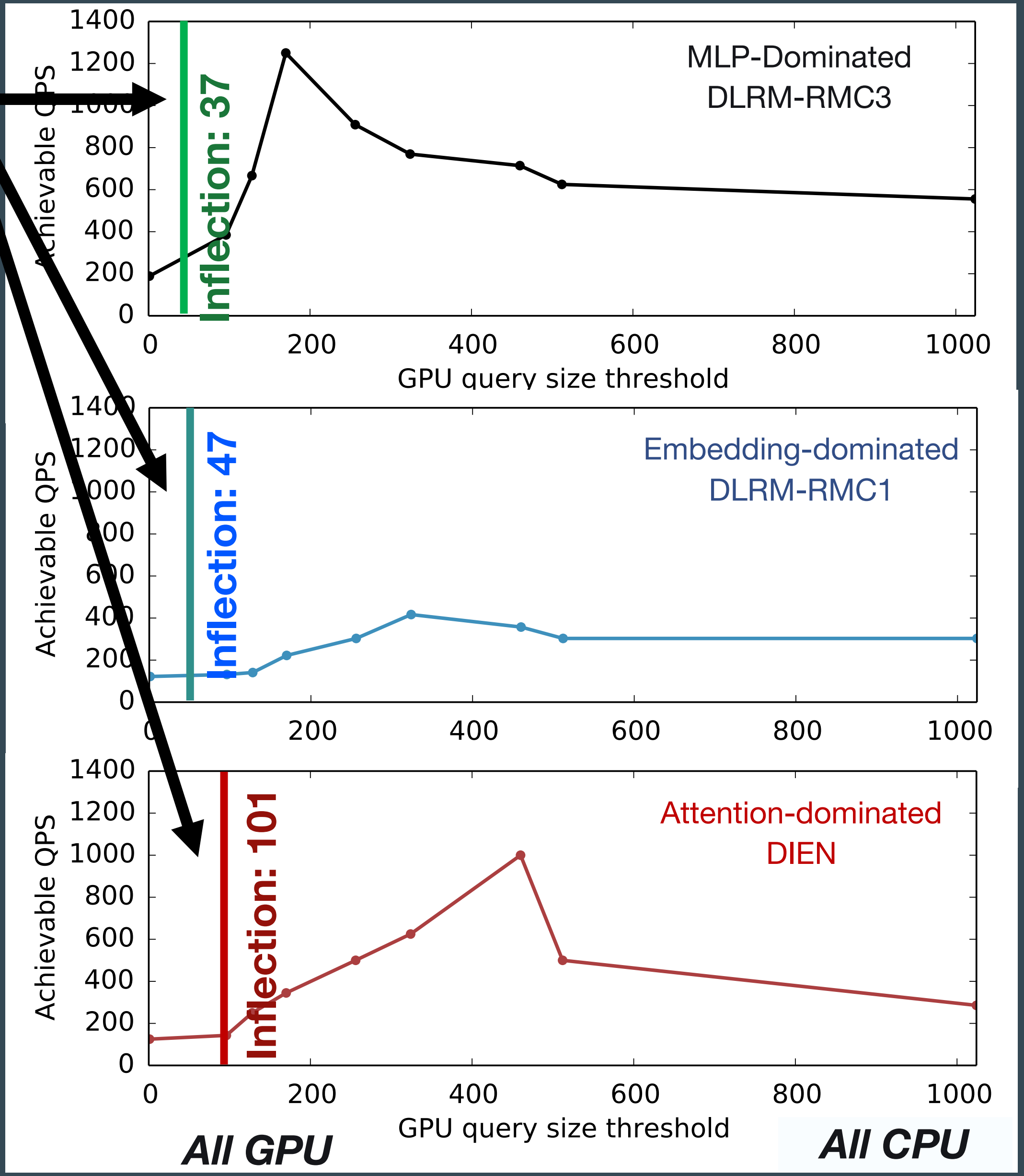


Latency-bound QPS Optimization

Inflection point for GPU > CPU

Optimal execution depends on

- Recommendation models
- AI system architectures
  - CPUs vs. AI accelerators
- Runtime characteristics
  - Query arrival and working set sizes
- Application SLA requirement



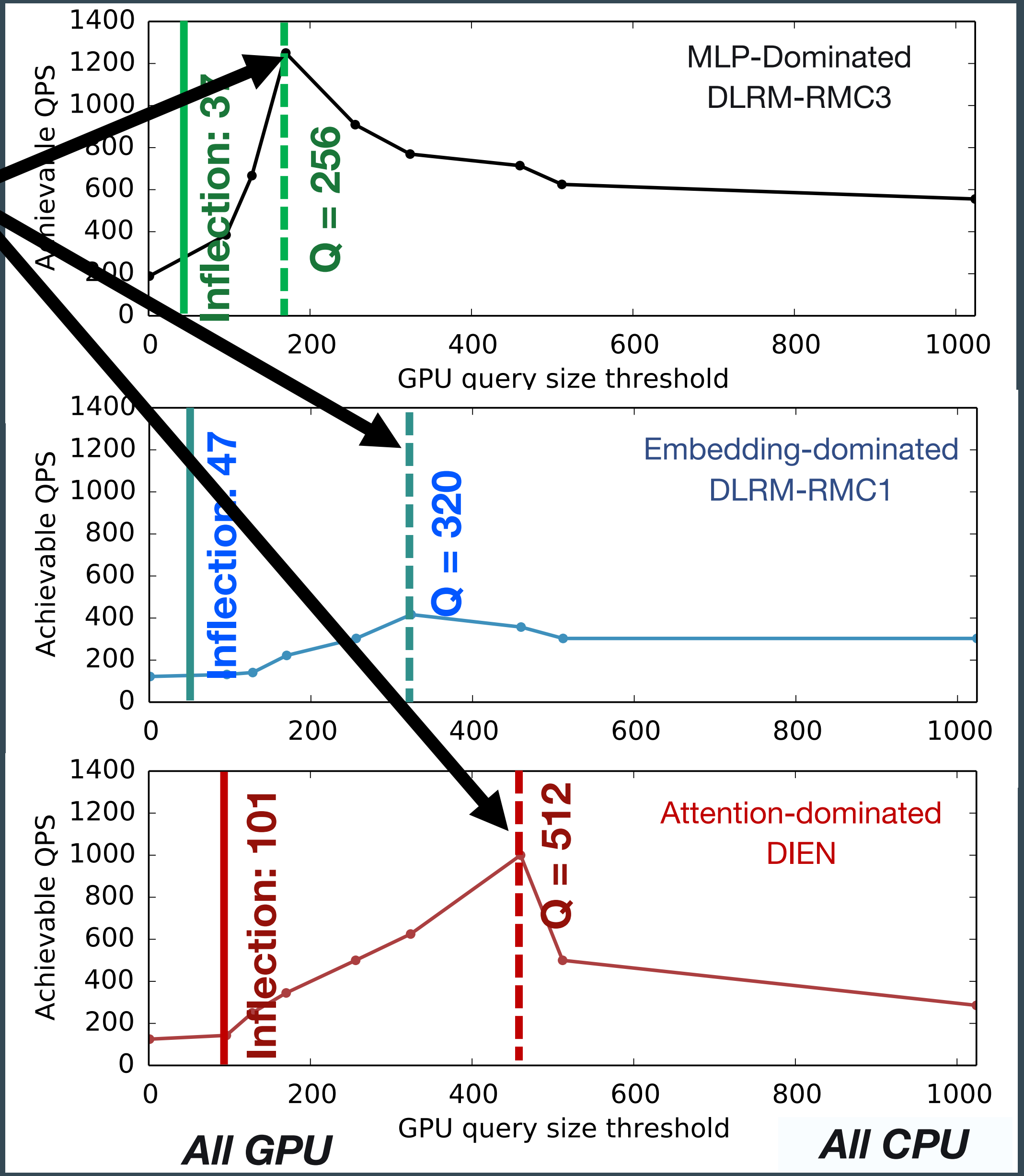
Latency-bound QPS Optimization

Inflection point for GPU > CPU

Optimal Query Size Threshold (Q)
Parallelism on CPUs

Optimal execution depends on

- Recommendation models
- AI system architectures
    - CPUs vs. AI accelerators
- Runtime characteristics
    - Query arrival and working set sizes
- Application SLA requirement



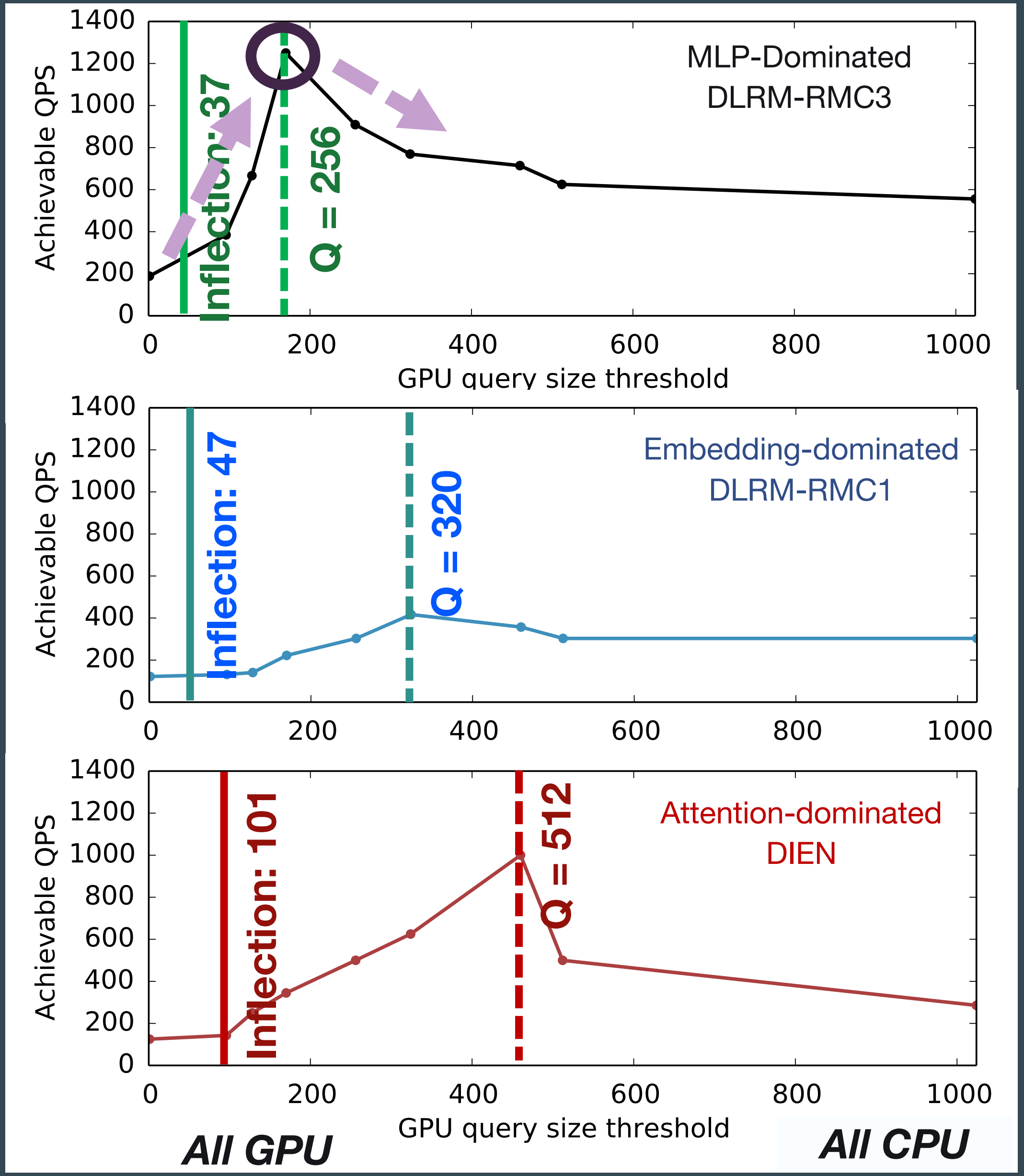
Latency-bound QPS Optimization

Inflection point for GPU > CPU

Optimal Query Size Threshold (Q)
Parallelism on CPUs

Optimal execution depends on

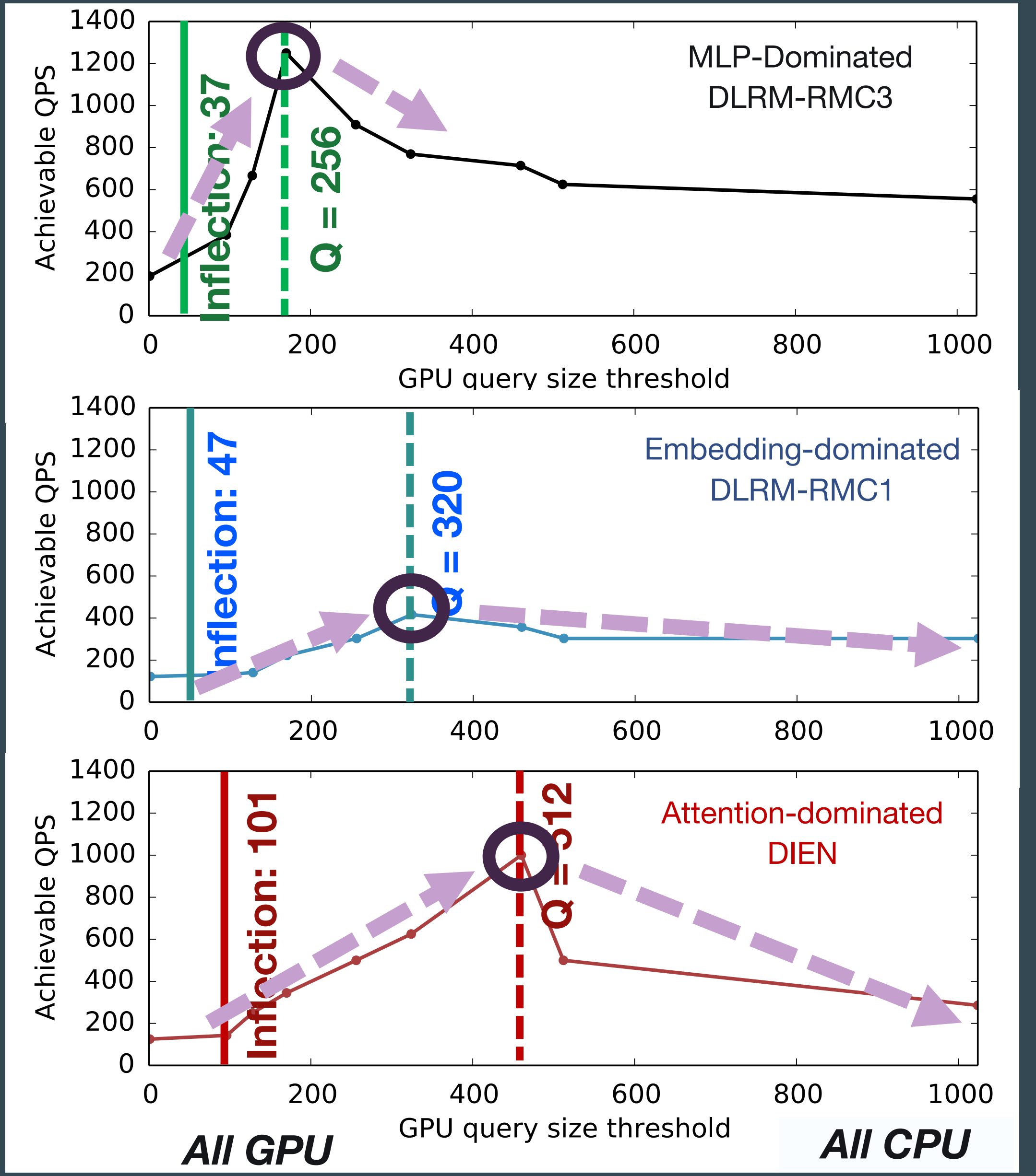
- Recommendation models
- AI system architectures
  - CPUs vs. AI accelerators
- Runtime characteristics
  - Query arrival and working set sizes
- Application SLA requirement



Latency-bound QPS Optimization

DeepRecSched uses simple hill-climbing search for
- optimal offloading threshold (Q), and
- batch size

- Recommendation models
- AI system architectures
  - CPUs vs. AI accelerators
- Runtime characteristics
  - Query arrival and working set sizes
- Application SLA requirement



Experimental Setup

More Detail in the Paper

DeepRecSys

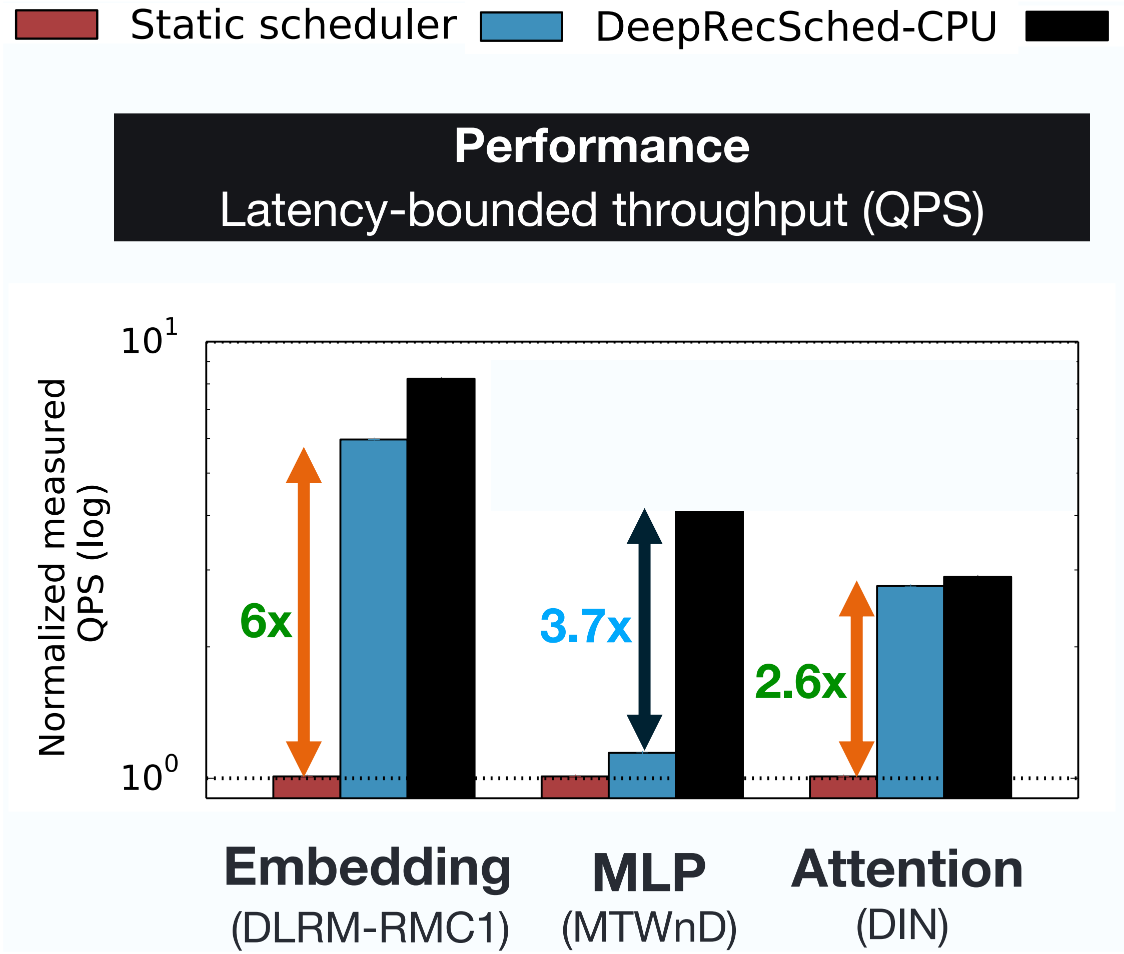
- Runtime recommendation query patterns (Poisson arrival & production working set size)
- 8 Industry-Representative Deep Learning Recommendation Model Architectures: DLRM-RM-1; DLRM-RM-2; DLRM-RM3; NCF; WND; MTWND; DIN; DIEN

Experimental systems

- Intel dual-socket Broadwell/Skylake CPUs; Intel MKL
- NVIDIA GTX 1080Ti; CUDA/cuDNN 10.1

Evaluation Results

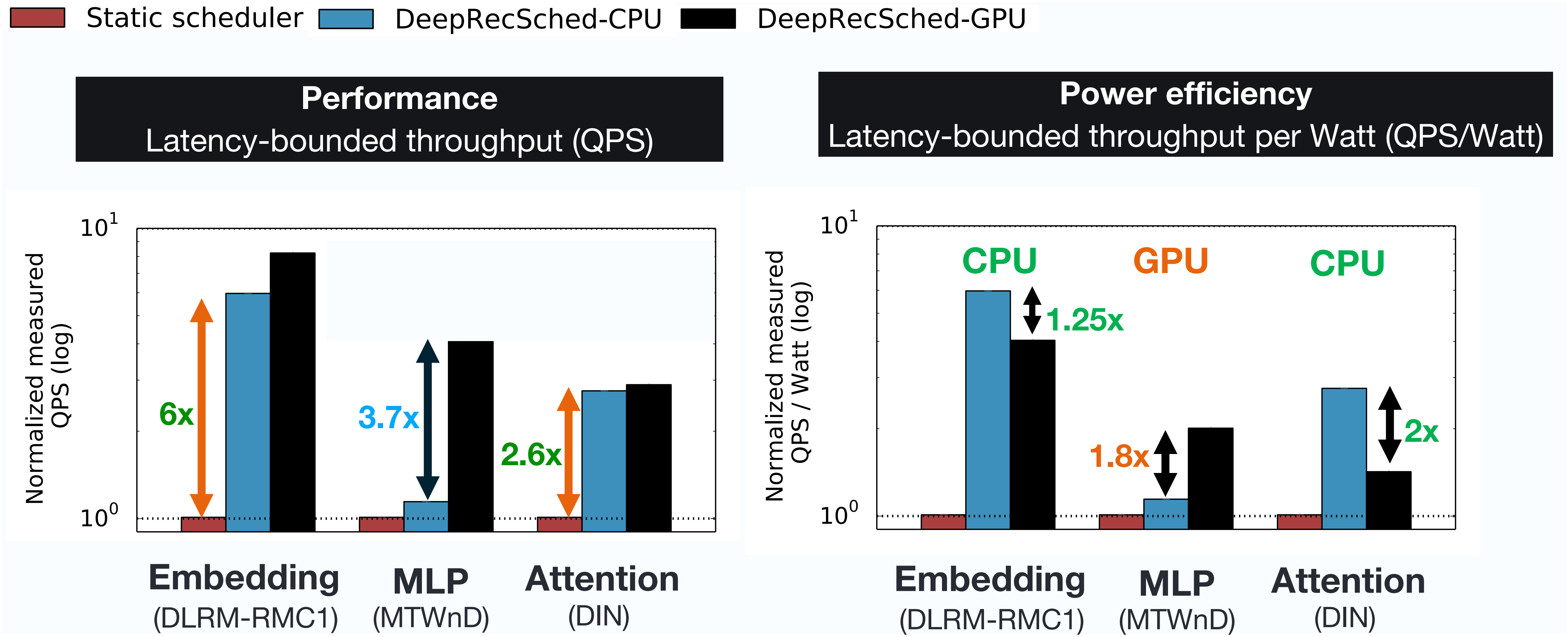
Performance and Power Efficiency Advantages



[ISCA 2020] **DeepRecSys: A System for Optimizing End-to-end At-scale Neural Recommendation Inference.** U. Gupta, S. Hsia, V. Saraph, X. Wang, B. Reagen, G.-Y. Wei, H.-S. Lee, D. Brooks, and C.-J. Wu.

Evaluation Results

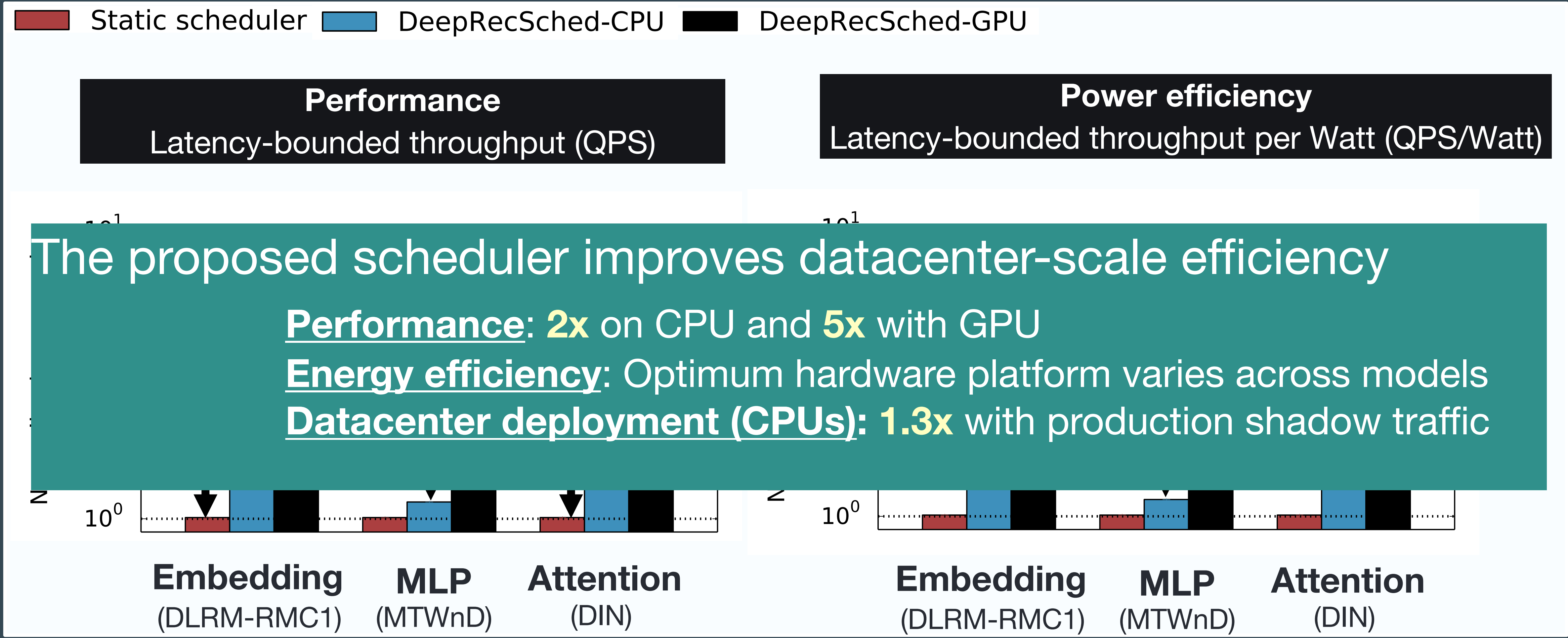
Performance and Power Efficiency Advantages



[ISCA 2020] **DeepRecSys: A System for Optimizing End-to-end At-scale Neural Recommendation Inference.** U. Gupta, S. Hsia, V. Saraph, X. Wang, B. Reagen, G.-Y. Wei, H.-S. Lee, D. Brooks, and C.-J. Wu.

Evaluation Results

Performance and Power Efficiency Advantages



The proposed scheduler improves datacenter-scale efficiency

Performance: **2x** on CPU and **5x** with GPU

Energy efficiency: Optimum hardware platform varies across models

Datacenter deployment (CPUs): 1.3x with production shadow traffic

[ISCA 2020] **DeepRecSys: A System for Optimizing End-to-end At-scale Neural Recommendation Inference.** U. Gupta, S. Hsia, V. Saraph, X. Wang, B. Reagen, G.-Y. Wei, H.-S. Lee, D. Brooks, and C.-J. Wu.

Agenda

Motivation

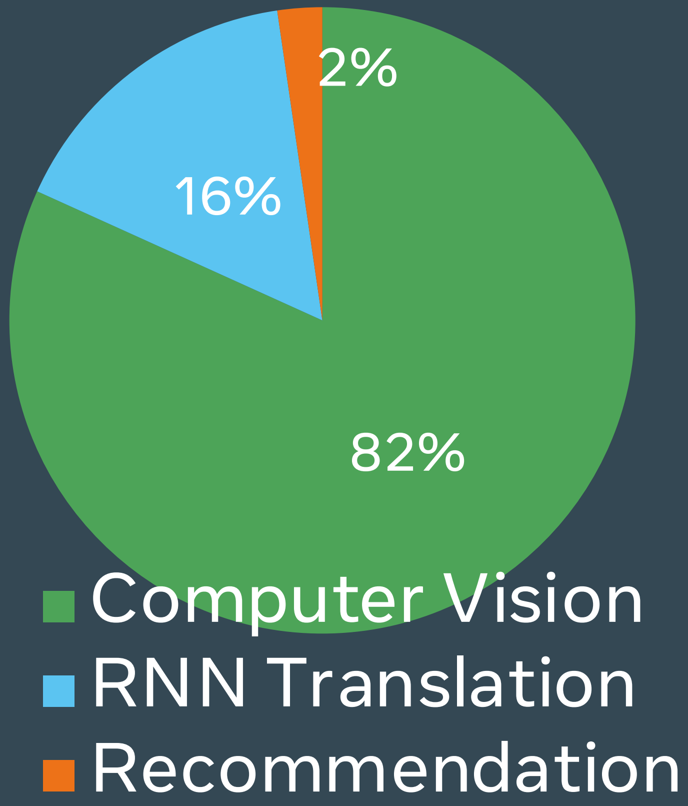
Understanding the Unique Systems Challenges

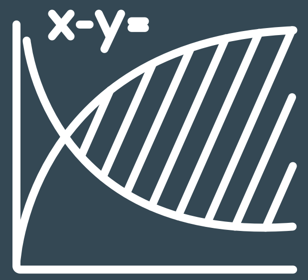
Characterizing Performance Acceleration with GPUs

Optimizing Neural Recommendation Inference At-Scale

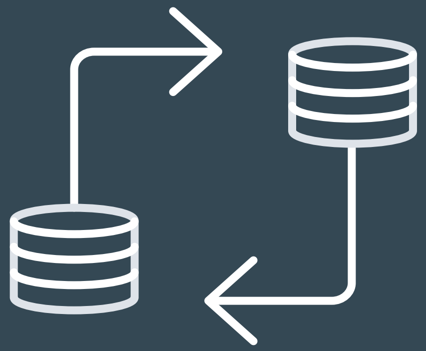
Conclusion and Future Work

Resolving the Underinvestment





Representative Benchmarks



Datasets

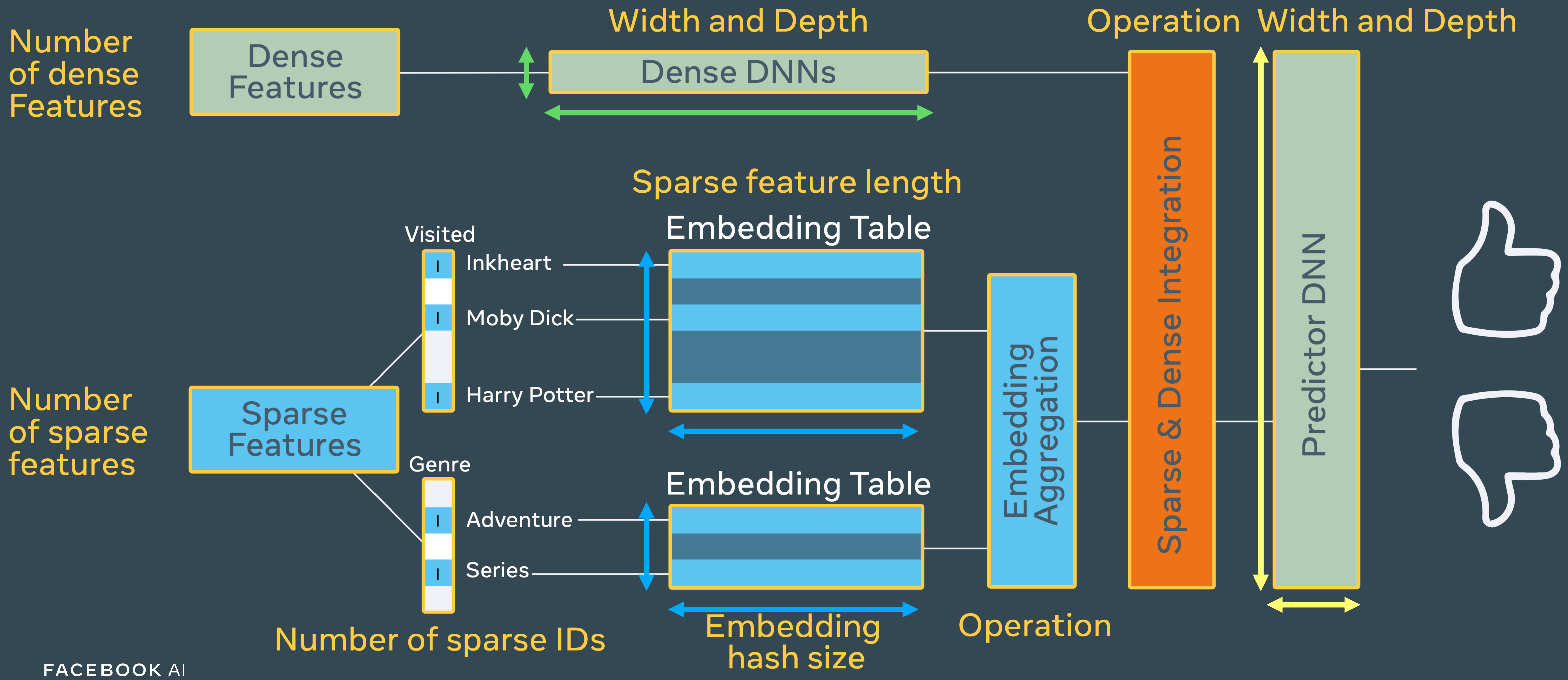
CV: **ImageNet**

NLP: **LibriSpeech**

Recommendation?

DLRM: Deep Learning Recommendation Model

A Configurable Benchmark for E2E Models



FACEBOOK AI

https://github.com/facebookresearch/dlrm

DeepRecSys: Industry-Representative Neural Recommendation Models

DeepRecSys: A System for Optimizing End-To-End At-scale Neural Recommendation Inference

https://github.com/harvard-acc/DeepRecSys

DeepRecSys provides an end-to-end infrastructure to study and optimize at-scale neural recommendation inference. The infrastructure is configurable across three main dimensions that represent different recommendation use cases: the load generator (query arrival patterns and size distributions), neural recommendation models, and underlying hardware platforms.

Neural recommendation models

This repository supports 8-industry representative neural recommendation models based on open-source publications from various Internet services in Caffe2:

1. Deep Learning Recommendation Models (DLRM-RMC1, DLRM-RMC2, DLRM-RMC3); link
2. Neural Collaborative Filtering (NCF); link
3. Wide and Deep (WnD); link

MLPerf includes DLRM + Criteo Ads Dataset

MLPerf

A machine learning performance benchmark suite with broad industry and academic support

MLPerf Includes DLRM + Criteo Ads Dataset

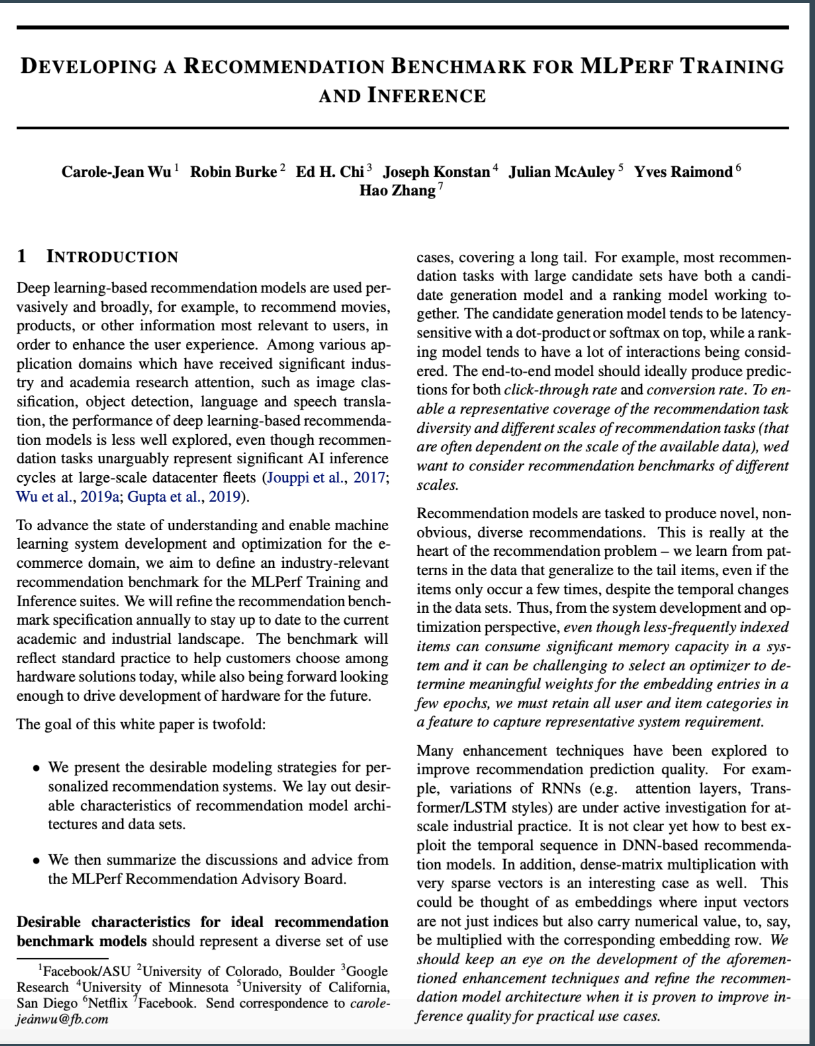
Recommendation Benchmark Advisory Board

Recommendation Model

- Cover a diverse set of use cases with the goal to optimize for both *click-through-rate* and *conversion-rate*, as well as to improve *long-term values*

Recommendation Datasets

- *Capture the degree of sparsity found in industry-scale problems*
- *Cover user- and item-features as well as user-item interactions*



DEVELOPING A RECOMMENDATION BENCHMARK FOR MLPERF TRAINING AND INFERENCE

Carole-Jean Wu [1] Robin Burke [2] Ed H. Chi [3] Joseph Konstan [4] Julian McAuley [5] Yves Raimond [6] Hao Zhang [7]

1 INTRODUCTION

Deep learning-based recommendation models are used pervasively and broadly, for example, to recommend movies, products, or other information most relevant to users, in order to enhance the user experience. Among various application domains which have received significant industry and academia research attention, such as image classification, object detection, language and speech translation, the performance of deep learning-based recommendation models is less well explored, even though recommendation tasks unarguably represent significant AI inference cycles at large-scale datacenter fleets (Jouppi et al., 2017; Wu et al., 2019a; Gupta et al., 2019).

To advance the state of understanding and enable machine learning system development and optimization for the e-commerce domain, we aim to define an industry-relevant recommendation benchmark for the MLPerf Training and Inference suites. We will refine the recommendation benchmark specification annually to stay up to date to the current academic and industrial landscape. The benchmark will reflect standard practice to help customers choose among hardware solutions today, while also being forward looking enough to drive development of hardware for the future.

The goal of this white paper is twofold:

- We present the desirable modeling strategies for personalized recommendation systems. We lay out desirable characteristics of recommendation model architectures and data sets.
- We then summarize the discussions and advice from the MLPerf Recommendation Advisory Board.

Desirable characteristics for ideal recommendation benchmark models should represent a diverse set of use cases, covering a long tail. For example, most recommendation tasks with large candidate sets have both a candidate generation model and a ranking model working together. The candidate generation model tends to be latency-sensitive with a dot-product or softmax on top, while a ranking model tends to have a lot of interactions being considered. The end-to-end model should ideally produce predictions for both *click-through rate* and *conversion rate*. *To enable a representative coverage of the recommendation task diversity and different scales of recommendation tasks (that are often dependent on the scale of the available data), wed want to consider recommendation benchmarks of different scales.*

Recommendation models are tasked to produce novel, non-obvious, diverse recommendations. This is really at the heart of the recommendation problem – we learn from patterns in the data that generalize to the tail items, even if the items only occur a few times, despite the temporal changes in the data sets. Thus, from the system development and optimization perspective, *even though less-frequently indexed items can consume significant memory capacity in a system and it can be challenging to select an optimizer to determine meaningful weights for the embedding entries in a few epochs, we must retain all user and item categories in a feature to capture representative system requirement.*

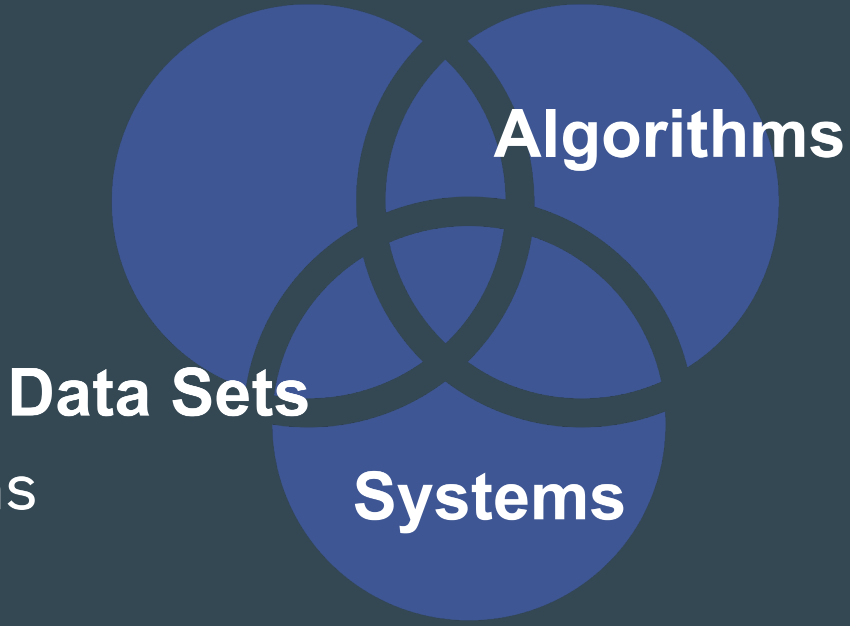
Many enhancement techniques have been explored to improve recommendation prediction quality. For example, variations of RNNs (e.g. attention layers, Transformer/LSTM styles) are under active investigation for at-scale industrial practice. It is not clear yet how to best exploit the temporal sequence in DNN-based recommendation models. In addition, dense-matrix multiplication with very sparse vectors is an interesting case as well. This could be thought of as embeddings where input vectors are not just indices but also carry numerical value, to, say, be multiplied with the corresponding embedding row. *We should keep an eye on the development of the aforementioned enhancement techniques and refine the recommendation model architecture when it is proven to improve inference quality for practical use cases.*

[1]Facebook/ASU [2]University of Colorado, Boulder [3]Google Research [4]University of Minnesota [5]University of California, San Diego [6]Netflix [7]Facebook. Send correspondence to *carole-jeanwu@fb.com*

[ArXiv 2020] **Developing a Recommendation Benchmark for MLPerf Training and Inference.** C.-J. Wu, R. Burke, E. Chi, J. Konstan, J. McAuley, Y. Raimond, H. Zhang.

Tutorials on Personalized Recommendation Systems and Algorithms

https://personal-tutorial.com/



- Understand the evolution of recommendation systems
- Discuss challenges of recommendation systems
- Provide a hands-on tutorial on open-source benchmarks and datasets (training and inference)
- Brainstorm novel solutions for efficient personalized recommendation

With ASPLOS-2020 & ISCA-2020

Recommendation Systems ...

1 Are Important

2 Are Underinvested

3 Have Unique Systems Challenges

4 Building Systems for Deep Learning Recommendation

5 New Benchmarks and Datasets Are NOW Available

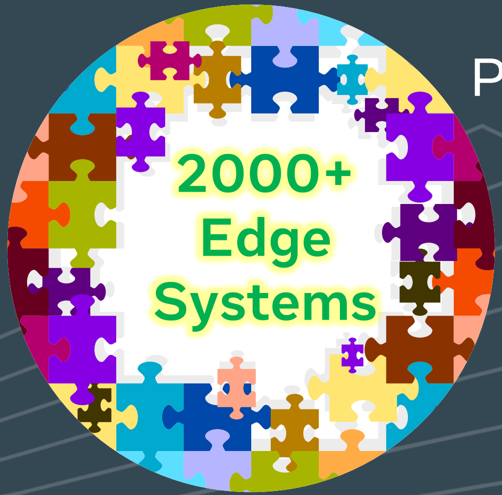
Responsible AI

Interpretability

Privacy

Responsible AI

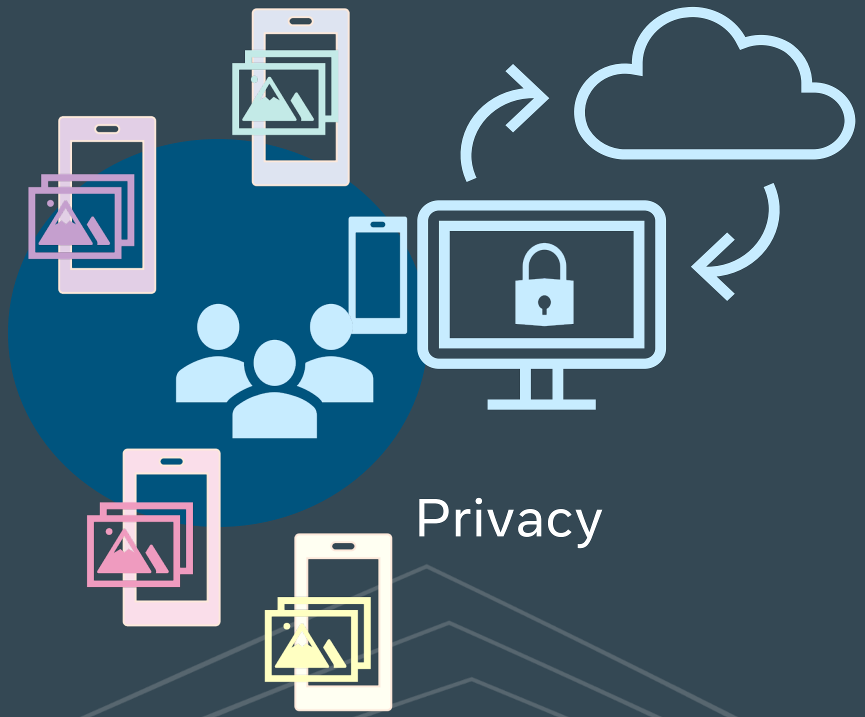
Interpretability



Privacy

Responsible AI

Interpretability



Privacy

Responsible AI

Interpretability

Privacy

Sustainability

AutoScale: Energy Efficiency Optimization
for Stochastic Edge Inference Using RL
(MICRO-2020)

Responsible AI

Interpretability

Privacy

Sustainability

5x 

References

[HPCA 2019] **Machine Learning at Facebook: Understanding Inference at the Edge**
C.-J. Wu, D. Brooks, K. Chen, D. Chen, S. Choudhury, M. Dukhan, K. Hazelwood, E. Isaac, Y. Jia, B. Jia, T. Leyvand, H. Lu, Y. Lu, L. Qiao, B. Reagen, J. Spisak, F. Sun, A. Tulloch, P. Vajda, X. Wang, Y. Wang, B. Wasti, Y. Wu, R. Xian, S. Yoo, P. Zhang. In *Proceedings of the 25th IEEE International Symposium on High Performance Computer Architecture (HPCA)*, Washington DC, USA, 2019.

[MICRO 2020] **AutoScale: Energy Efficiency Optimization for Stochastic Edge Inference Using Reinforcement Learning**
Y. G. Kim and C.-J. Wu. To Appear in *Proceedings of the IEEE International Symposium on Microarchitecture (MICRO)*, Athens, Greece, October 2020.

[HPCA 2020] **The Architectural Implications of Facebook's DNN-based Personalized Recommendation**
U. Gupta, C.-J. Wu, X. Wang, M. Naumov, B. Reagen, D. Brooks, B. Cottel, K. Hazelwood, M. Hempstead, B. Jia, H.-H. Lee, A. Malevich, D. Mudigere, M. Smelyanskiy, L. Xiong, X. Zhang. In *Proceedings of the IEEE International Symposium on High Performance Computer Architecture (HPCA)*, San Diego CA, 2020.

[ISCA 2020] **DeepRecSys: A System for Optimizing End-to-end At-scale Neural Recommendation Inference**
U. Gupta, S. Hsia, V. Saraph, X. Wang, B. Reagen, G.-Y. Wei, H.-S. Lee, D. Brooks, and C.-J. Wu. In *Proceedings of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, Virtual, 2020.

[ISCA 2020] **RecNMP: Accelerating Personalized Recommendation with Near-Memory Processing**
L. Ke, U. Gupta, B. Cho, D. Brooks, V. Chandra, U. Diril, A. Firoozshahian, K. Hazelwood, B. Jia, H.-S. Lee, M. Li, B. Maher, D. Mudigere, M. Naumov, M. Schatz, M. Smelyanskiy, X. Wang, B. Reagen, C.-J. Wu, M. Hempstead, X. Zhang. In *Proceedings of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, Virtual, 2020.

[IISWC 2020] **Cross-Stack Workload Characterization of Deep Recommendation Systems**
S. Hsia, U. Gupta, M. Wilkening, C.-J. Wu, G.-Y. Wei, D. Brooks. In *Proceedings of the IEEE International Symposium on Workload Characterization (IISWC),* Virtual, 2020.

References

[ArXiv 2019] **Deep Learning Recommendation Model for Personalization and Recommendation Systems**
M. Naumov, D. Mudigere, H.-J. Shi, J. Huang, N. Sundaraman, J. Park, X. Wang, U. Gupta, C.-J. Wu, A. Azzolini, D. Dzhulgakov, A. Mallevich, I. Cherniavskii, Y. Lu, R. Krishnamoorthi, A. Yu, V. Kondratenko, S. Pereira, X. Chen, W. Chen, V. Rao, B. Jia, L. Xiong, M. Smelyanskiy. In CoRR abs/1906.00091

[ArXiv 2020] **Developing a Recommendation Benchmark for MLPerf Training and Inference**
C.-J. Wu, R. Burke, E. Chi, J. Konstan, J. McAuley, Y. Raimond, H. Zhang. In *CoRR abs/2003.07336*.

[IEEE Micro 2020] **MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance**
P. Mattson, V. Reddi, C. Cheng, C. Coleman, G. Diamos, D. Kanter, P. Micikevicius, D. Patterson, G. Schmuelling, H. Tang, G.-Y. Wei, C.-J. Wu. In Proceedings of the *IEEE Micro*, 2020.

[MLSys 2020] **MLPerf Training Benchmark**
P. Mattson, C. Cheng, C. Coleman, G. Diamos, P. Micikevicius, D. Patterson, H. Tang, G.-Y. Wei, P. Ballis, V. Bittorf, D. Brooks, D. Chen, D. Dutta, U. Gupta, K. Hazelwood, A. Hock, X. Huang, B. Jia, D. Kang, N. Kumar, J. Liao, G. Ma, D. Narayanan, T. Oguntebi, G. Pekhimenko, L. Pentecost, V. Reddi, T. Robie, T. St. John, C.-J. Wu, L. Xu, C. Young, M. Zaharia. In *Proceedings of the Conference on Machine Learning and Systems (MLSys)*, Austin TX, 2020.

[ISCA 2020] **MLPerf Inference Benchmark**
V. Reddi, C. Cheng, D. Kanter, P. Mattson, G. Schmuelling, C.-J. Wu, B. Anderson, M. Breughe, M. Charlebois, W. Chou, R. Chukka, C. Coleman, S. Davis, P. Deng, G. Diamos, J. Duke, D. Fick, J. Gardner, I. Hubara, S. Idgunji, T. Jablin, J. Jiao, T. St. John, P. Kanwar, D. Lee, J. Liao, A. Lokhmotov, F. Massa, P. Meng, P. Micikevicius, C. Osborne, G. Pekhimenko, A. Rajan, D. Sequeira, A. Sirasao, F. Sun, H. Tang, M. Thomson, F. Wei, E. Wu, L. Xu, K. Yamada, B. Yu, G. Yuan, A. Zhong, P. Zhang, Y. Zhou. In *Proceedings of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, Virtual, 2020.

Thank you!

TO LEARN MORE, VISIT

research.fb.com
github.com/facebookresearch/dlrm
github.com/harvard-acc/DeepRecSys
personal-tutorial.com
mlperf.org

FACEBOOK AI