The Architectural Implications of Facebook's DNN-based Personalized Recommendation

Udit Gupta, Carole-Jean Wu, Xiaodong Wang, Maxim Naumov, Brandon Reagen

David Brooks, Bradford Cottel, Kim Hazelwood, Mark Hempstead, Bill Jia, Hsien-Hsin S. Lee, Andrey Malevich, Dheevatsa Mudigere, Mikhail Smelyanskiy, Liang Xiong, Xuan Zhang

IEEE Intl. Symposium on High-Performance Computer Architecture
HPCA 2020, Session 7a

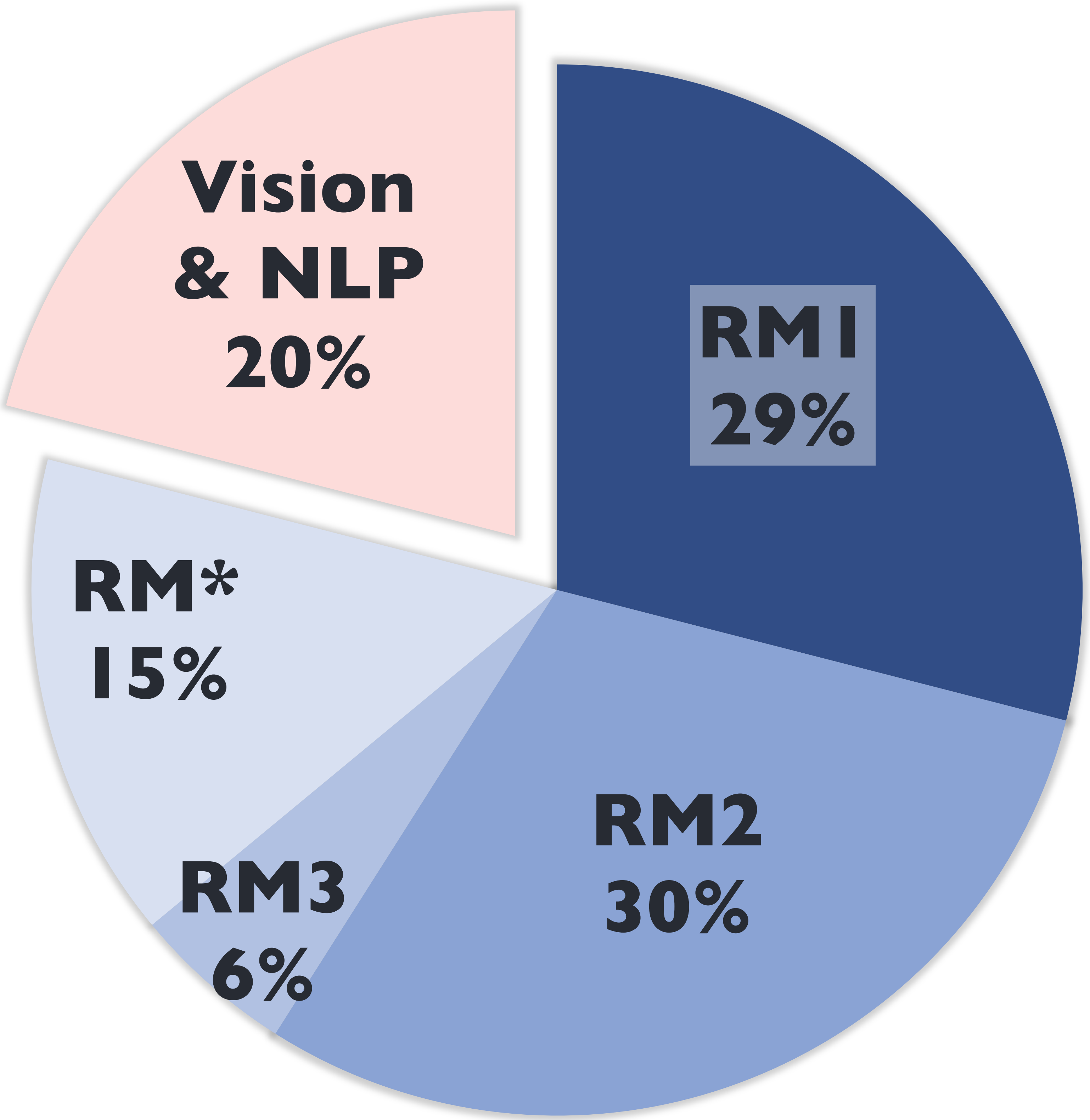
Personalized recommendation is everywhere

Personalized recommendation is everywhere

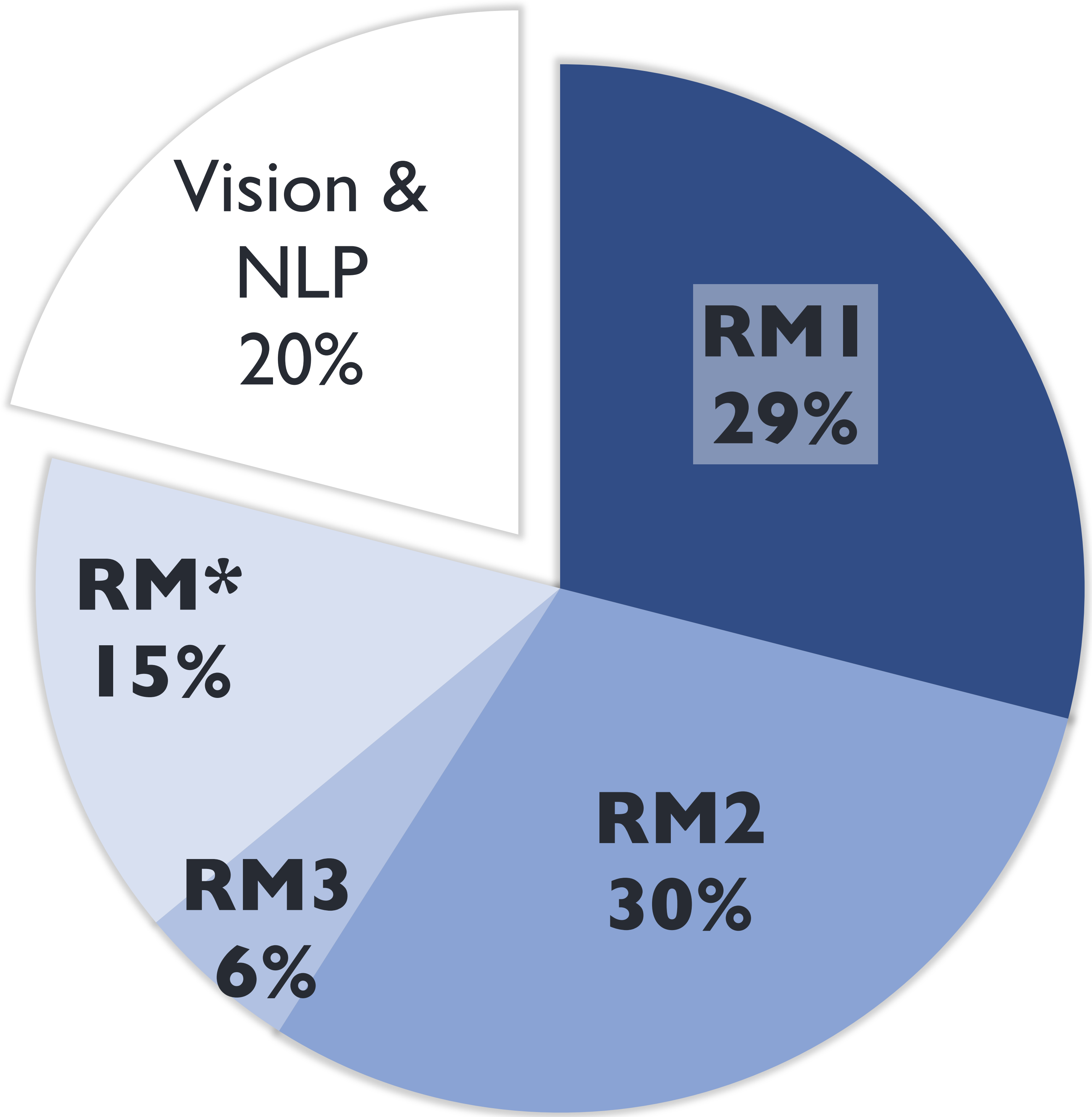
Personalized recommendation is everywhere

"35% of purchases on Amazon and 75% of videos on Netflix are powered by recommendation algorithms"
McKinsey & Co

AI inference cycles in Facebook's datacenter

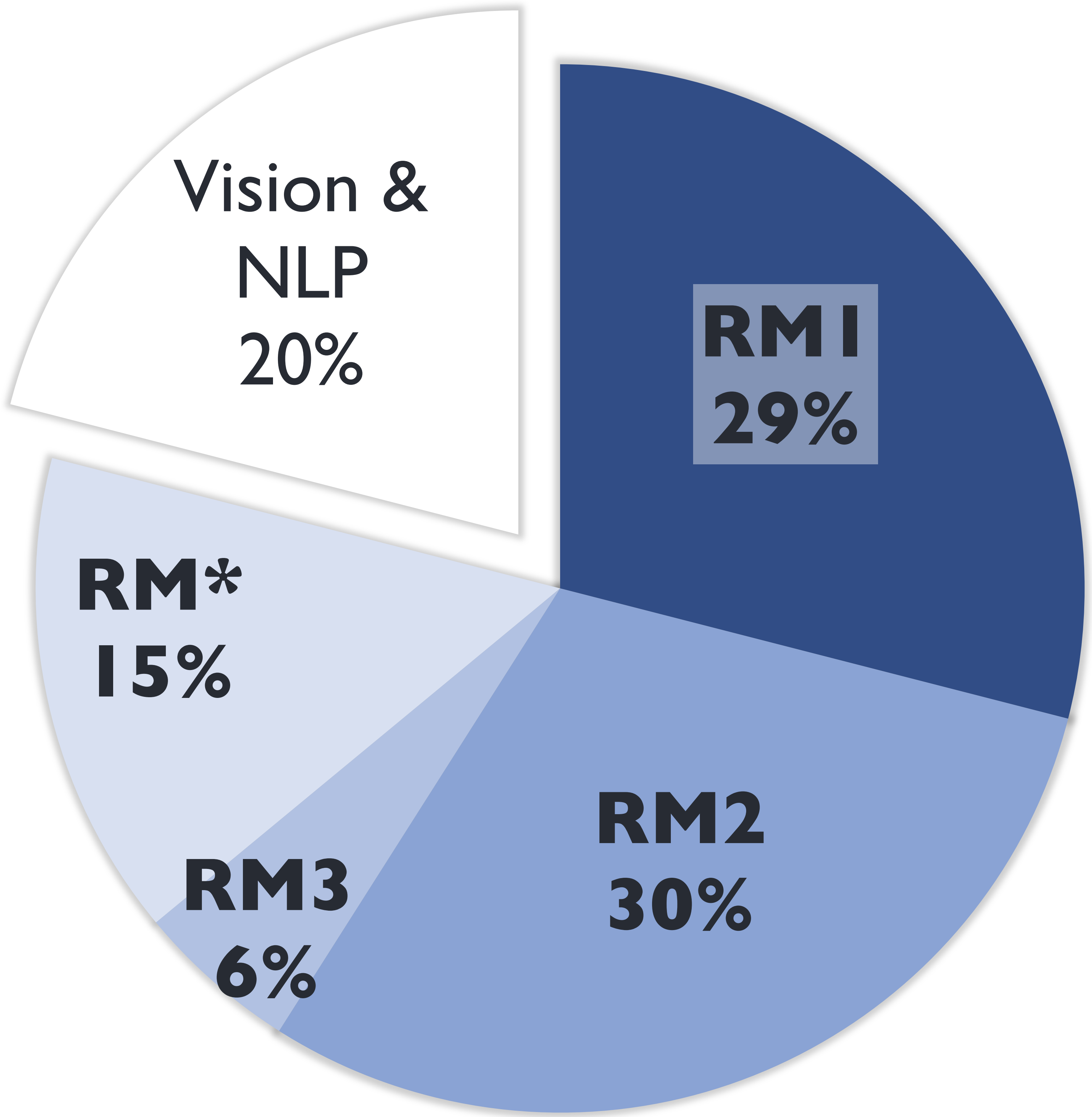


AI inference cycles in Facebook's datacenter



Recommendation uses cases account for over 80% of all AI inference cycles in Facebook's datacenter.

AI inference cycles in Facebook's datacenter



Recommendation uses cases account for over 80% of all AI inference cycles in Facebook's datacenter.

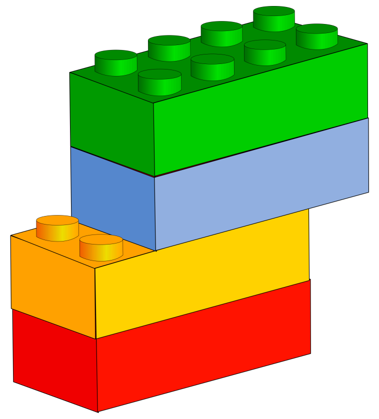
Given Facebook's datacenters perform 200+ trillion inferences every day, optimizing DNN-based recommendation is key.

https://engineering.fb.com/data-center-engineering/accelerating-infrastructure/

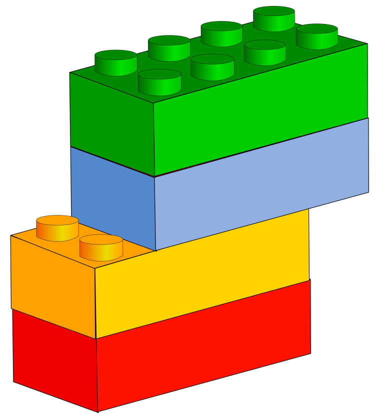
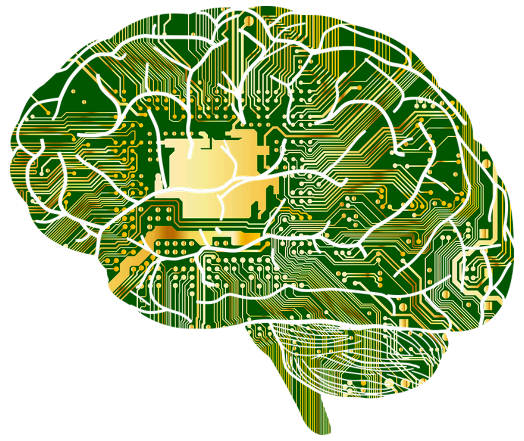
Hardware insights of recommendation

Algorithm	Hardware insights and opportunities

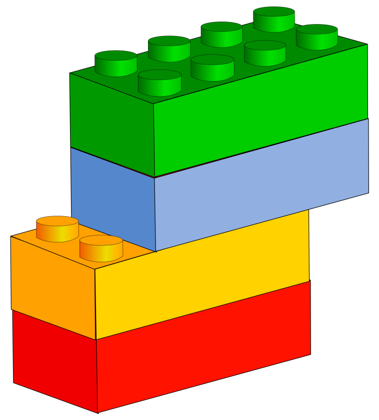
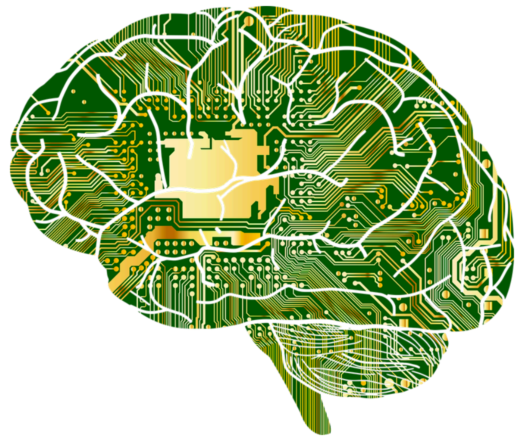
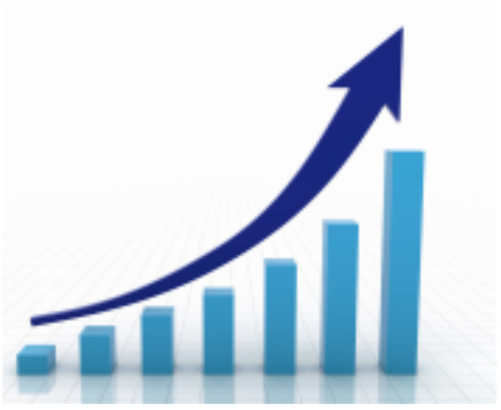
Hardware insights of recommendation

Algorithm	Hardware insights and opportunities
 General model structure	Optimize operators with new storage, compute, and memory access patterns

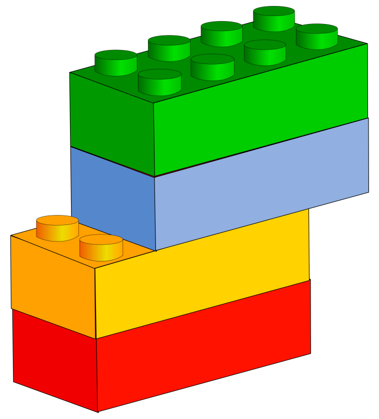
Hardware insights of recommendation

Algorithm	Hardware insights and opportunities
 General model structure	Optimize operators with new storage, compute, and memory access patterns
 Diverse networks architectures	Accelerate recommendation with flexible and diverse system solutions

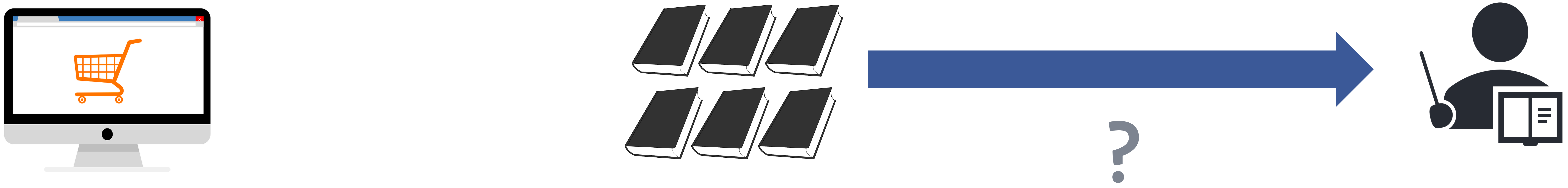
Hardware insights of recommendation

Algorithm		Hardware insights and opportunities
	General model structure	Optimize operators with new storage, compute, and memory access patterns
	Diverse networks architectures	Accelerate recommendation with flexible and diverse system solutions
	At-scale inference	Exploit hardware heterogeneity and parallelism to optimize latency-bounded throughput

Hardware insights of recommendation

Algorithm	Hardware insights and opportunities
 General model structure	Optimize operators with new storage, compute, and memory access patterns
Diverse networks architectures	Accelerate recommendation with flexible and diverse system solutions
At-scale inference	Exploit hardware heterogeneity and parallelism to optimize latency-bounded throughput

DNNs for recommendation



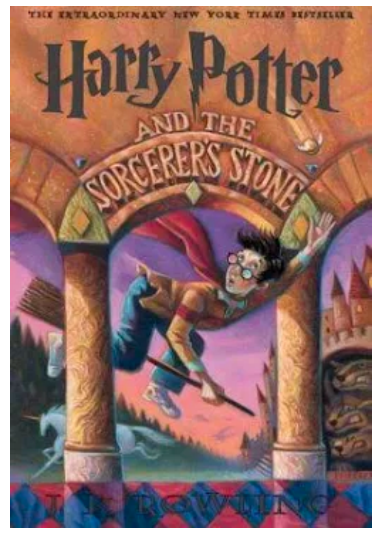
DNNs for recommendation

DNNs for recommendation

Continuous (dense) features

Categorical (sparse) features

DNNs for recommendation



?

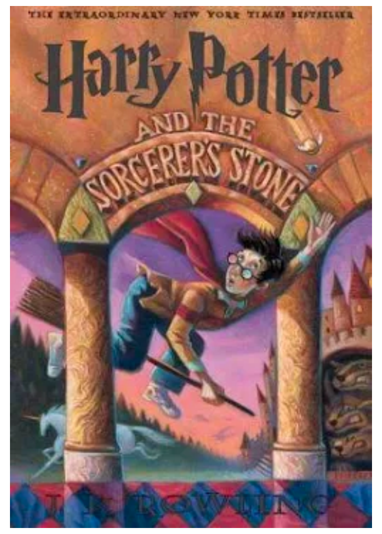
Continuous (dense) features

Age
Time of day

Dense DNNs

Categorical (sparse) features

DNNs for recommendation



?

Continuous (dense) features

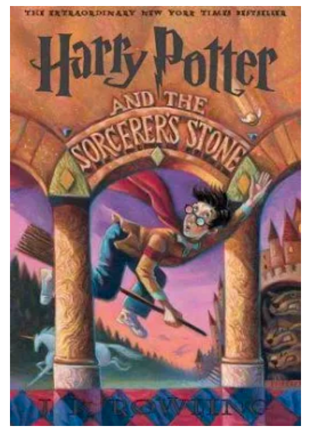
Age
Time of day

Dense DNNs

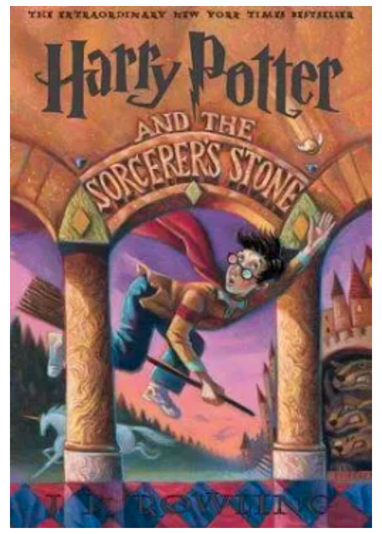
User purchase history

Categorical (sparse) features

Book's genre



DNNs for recommendation



?

Continuous (dense) features

Age
Time of day

Dense DNNs

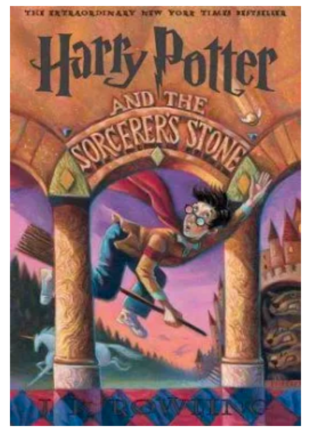
Categorical (sparse) features

User

User purchase history

Visited
- 1 Inkheart
- 1 Moby Dick
- 1 Hunger Games

Book's genre

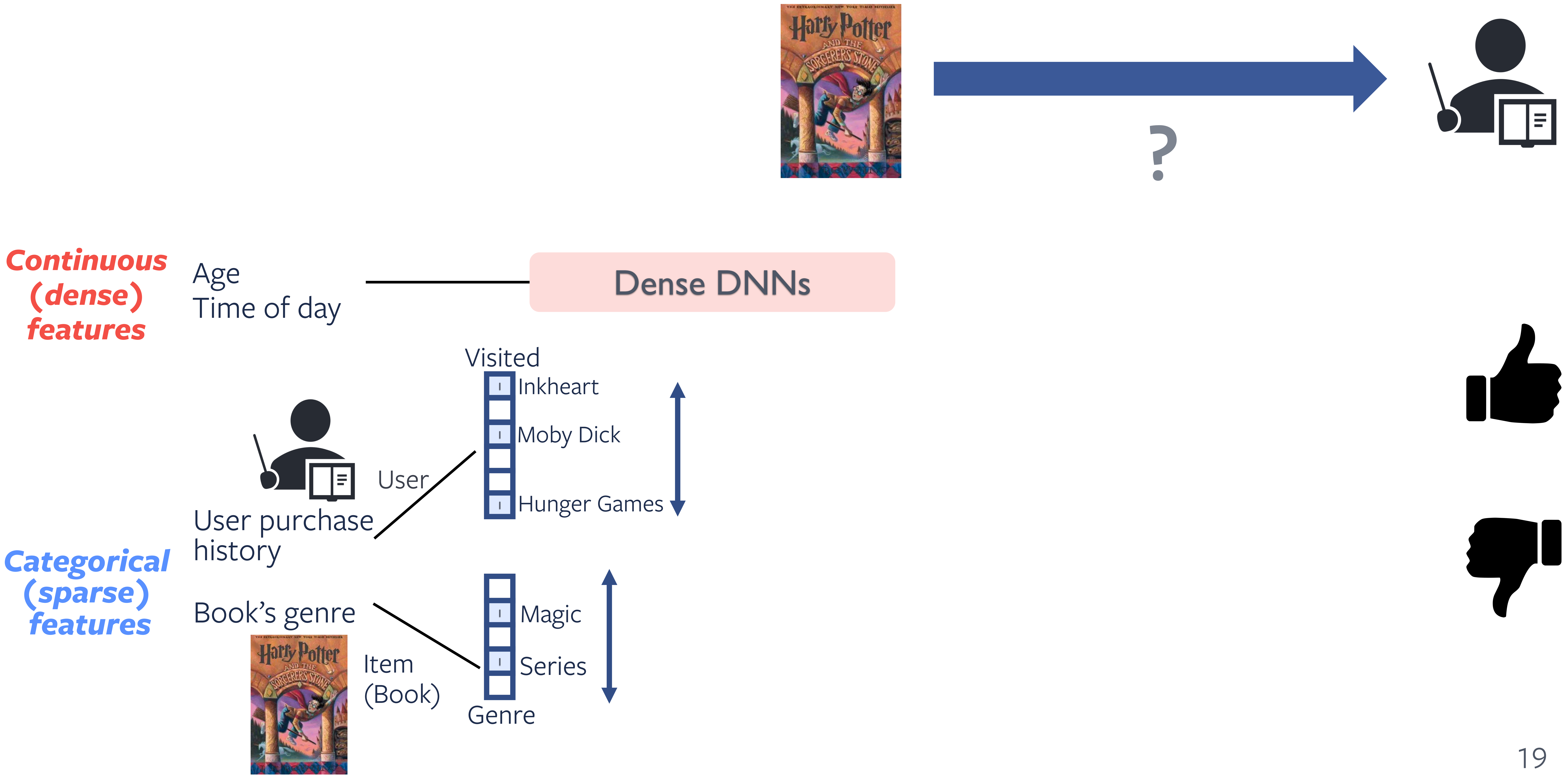


Item (Book)

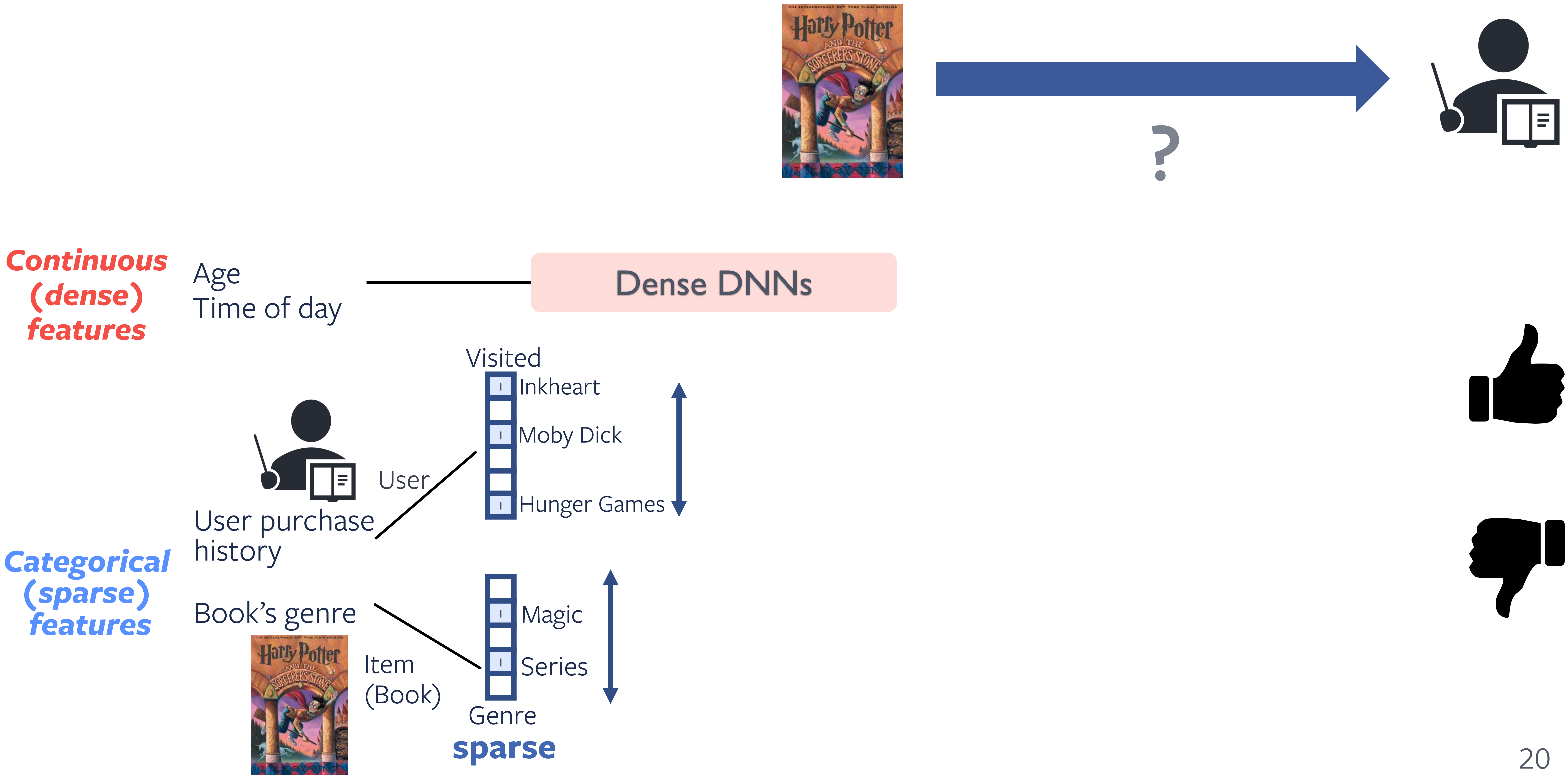
- 1 Magic
- 1 Series

Genre

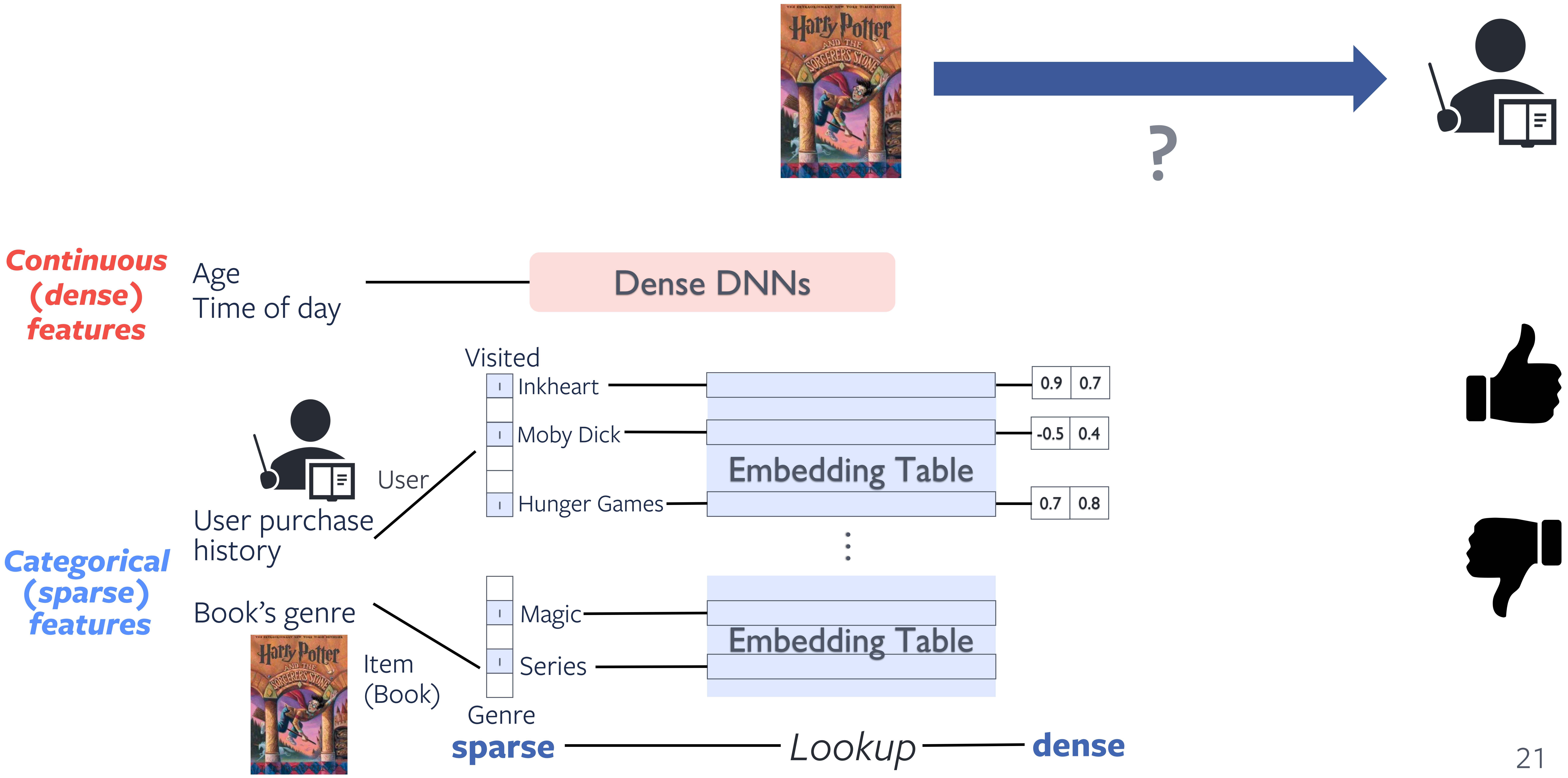
DNNs for recommendation



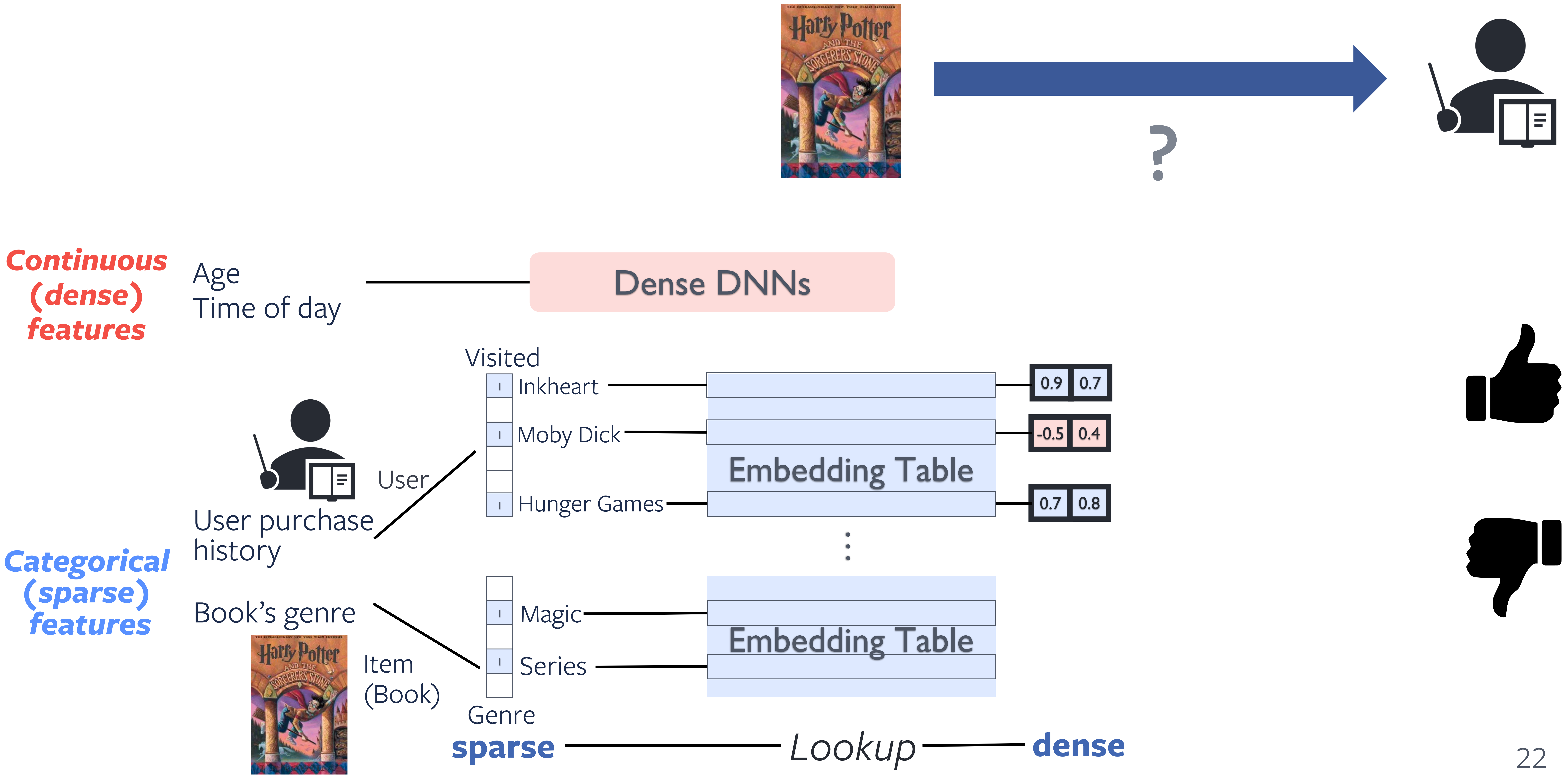
DNNs for recommendation



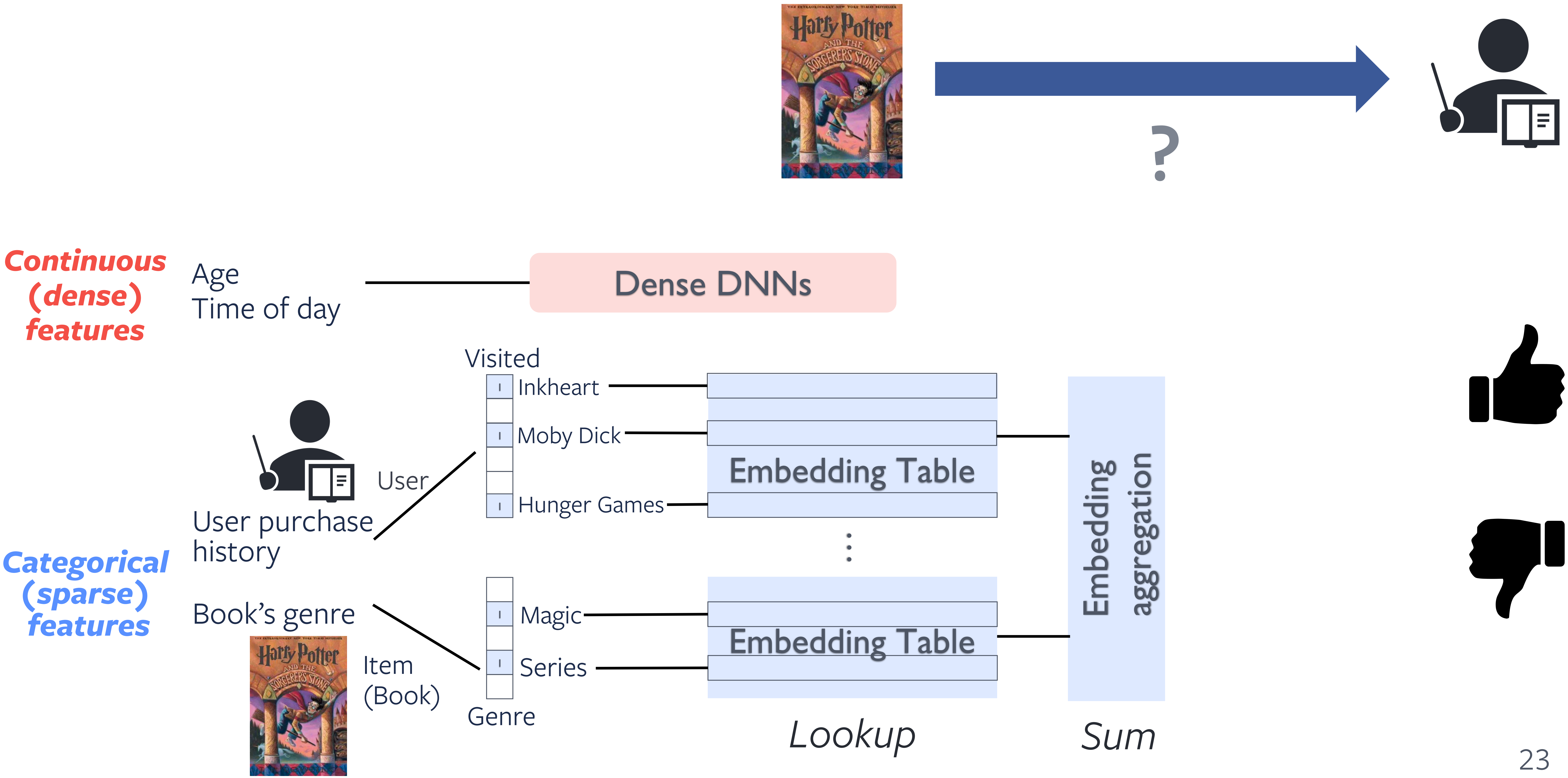
DNNs for recommendation



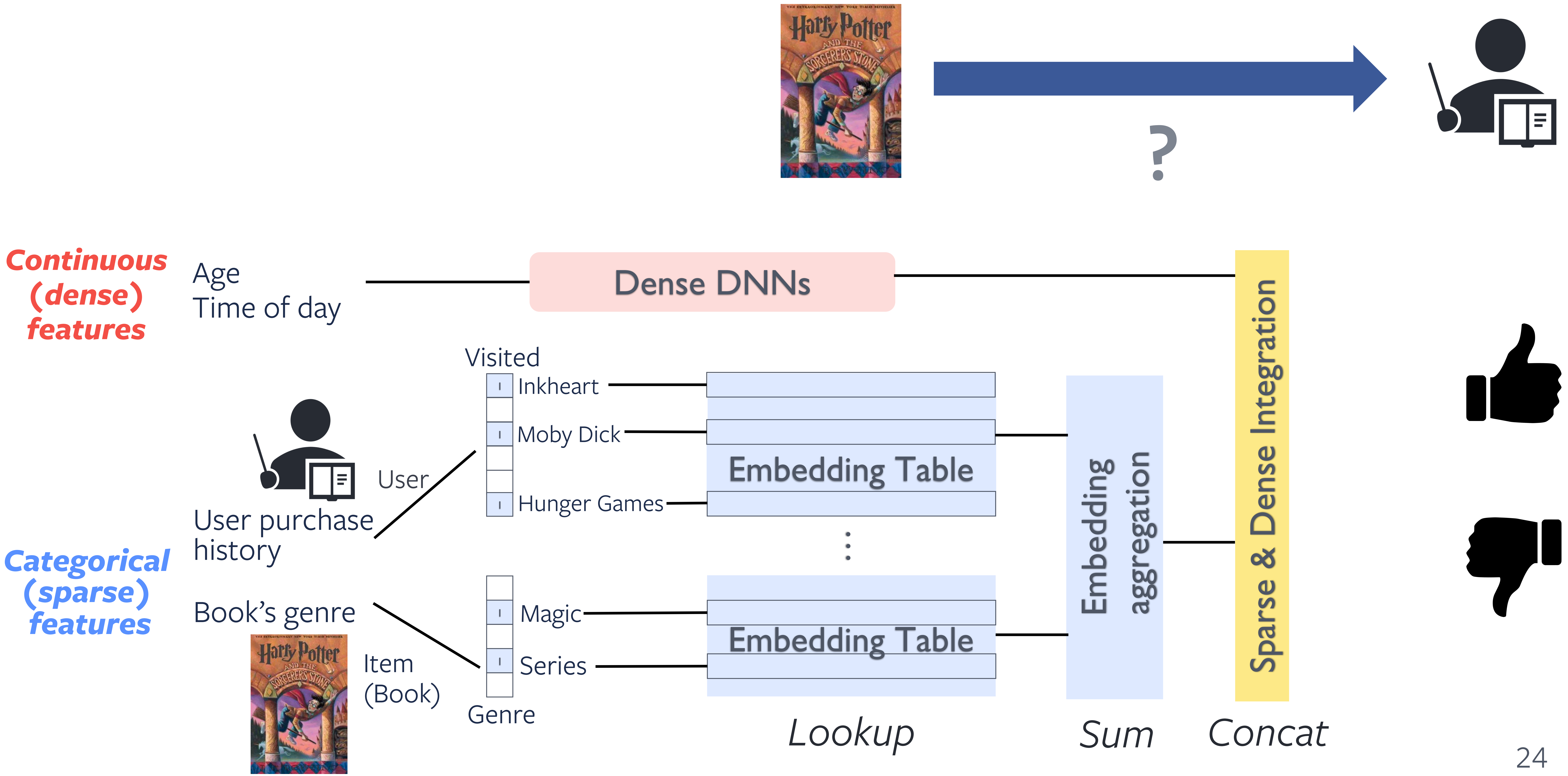
DNNs for recommendation



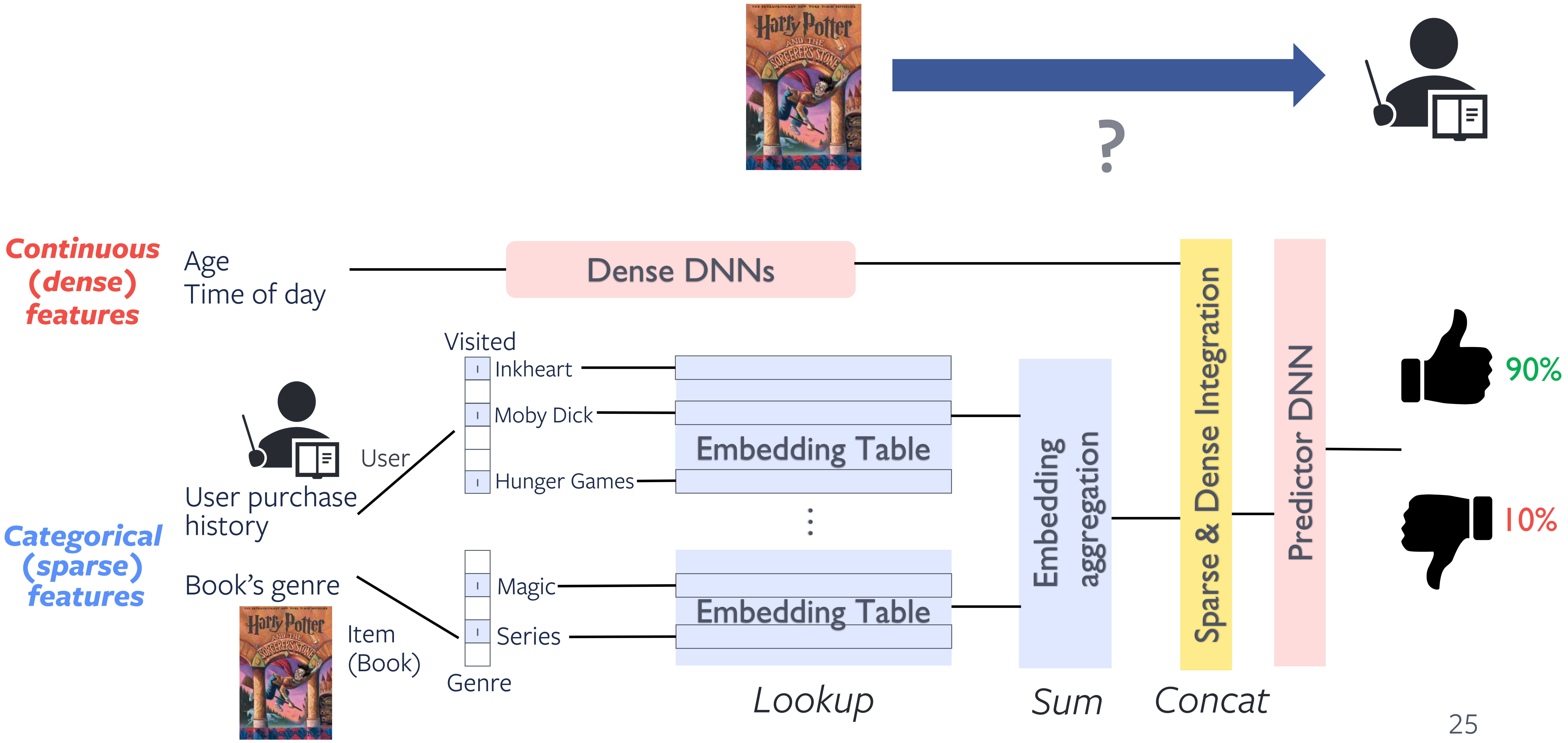
DNNs for recommendation



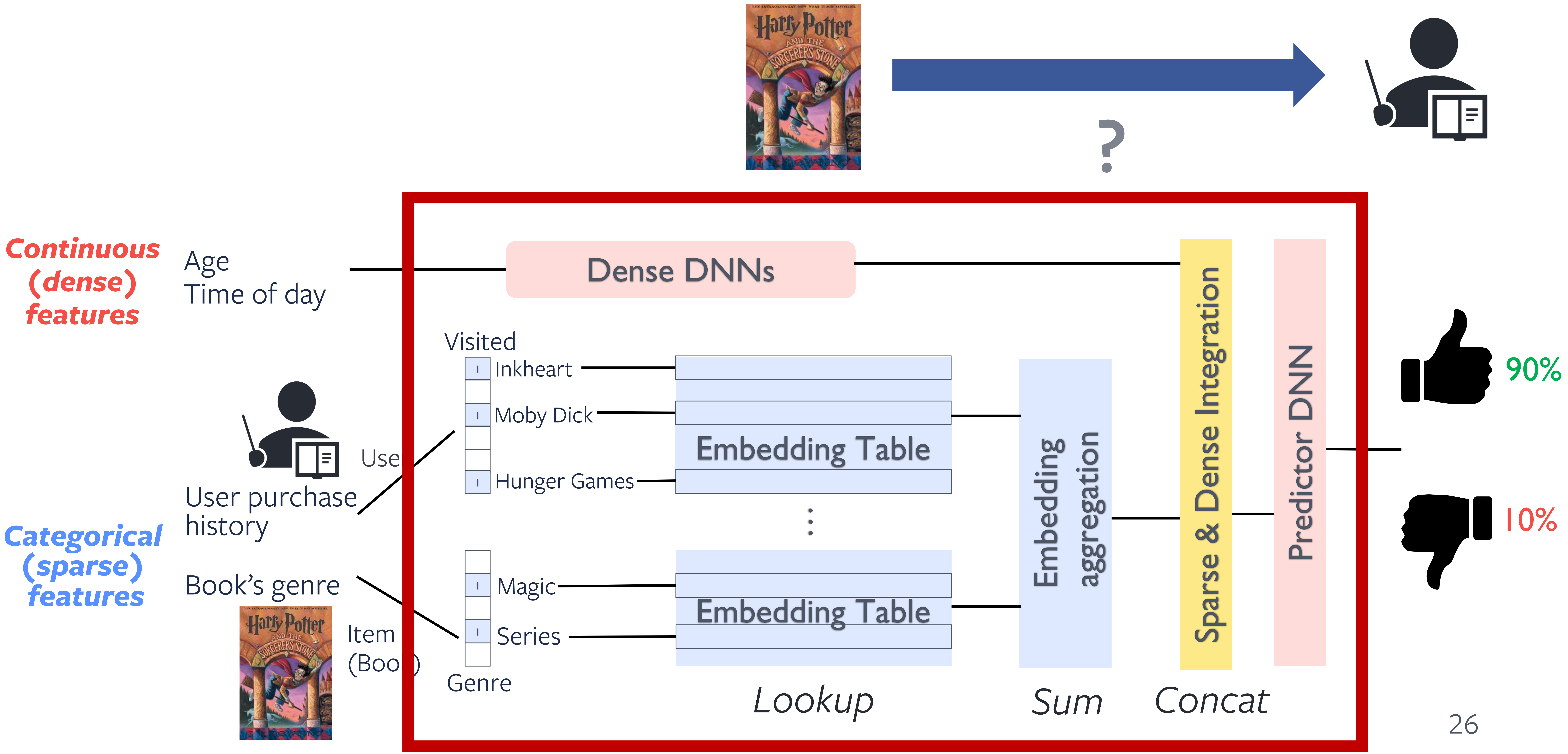
DNNs for recommendation



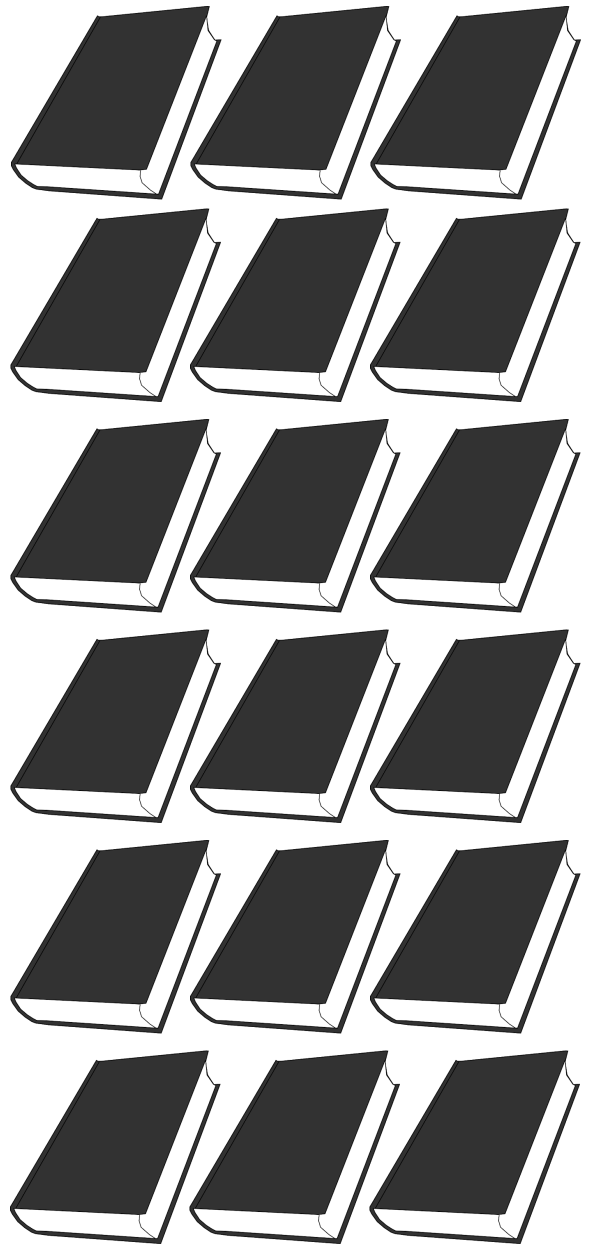
DNNs for recommendation



DNNs for recommendation

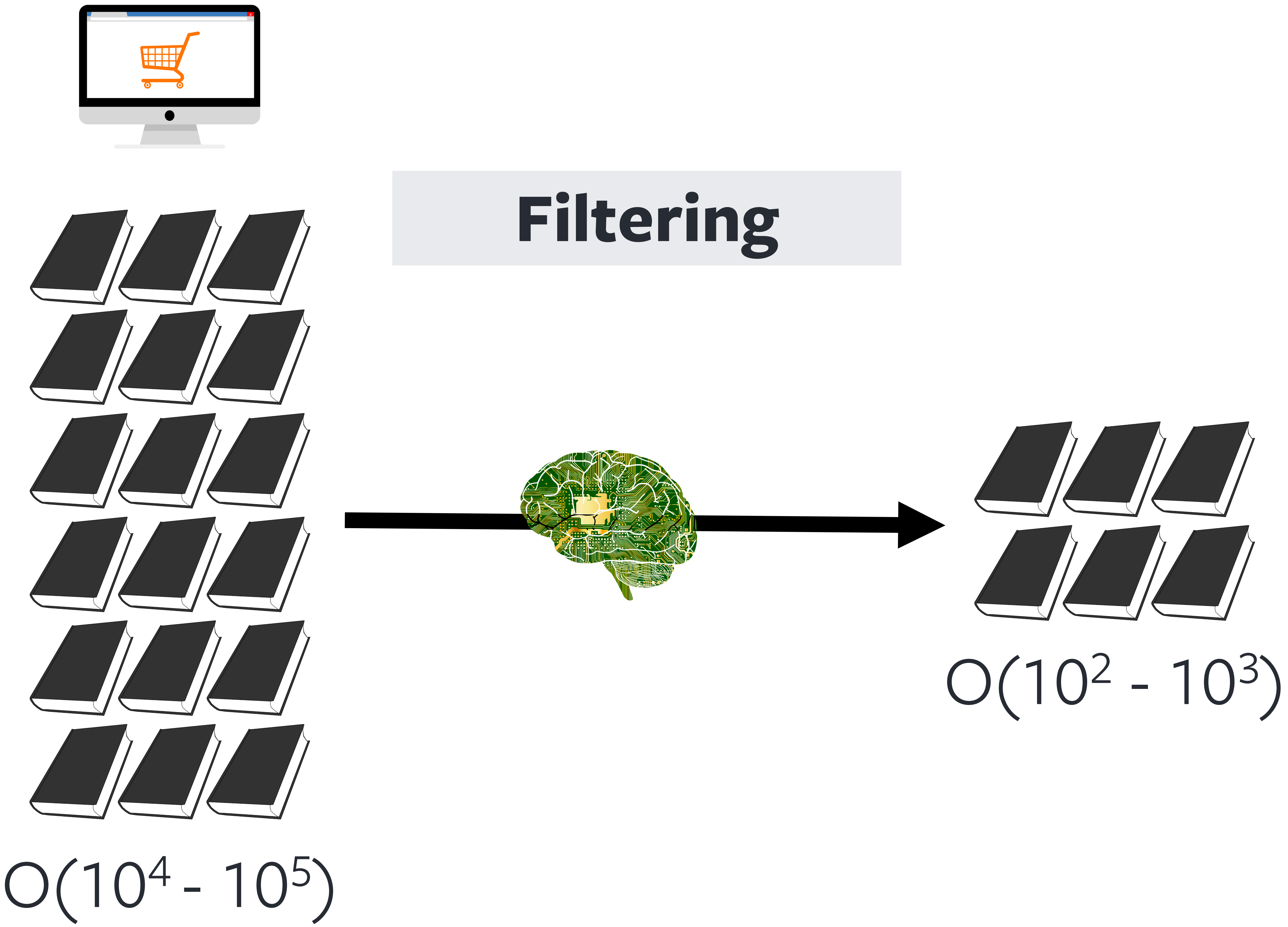


Ranking thousands of items at-scale

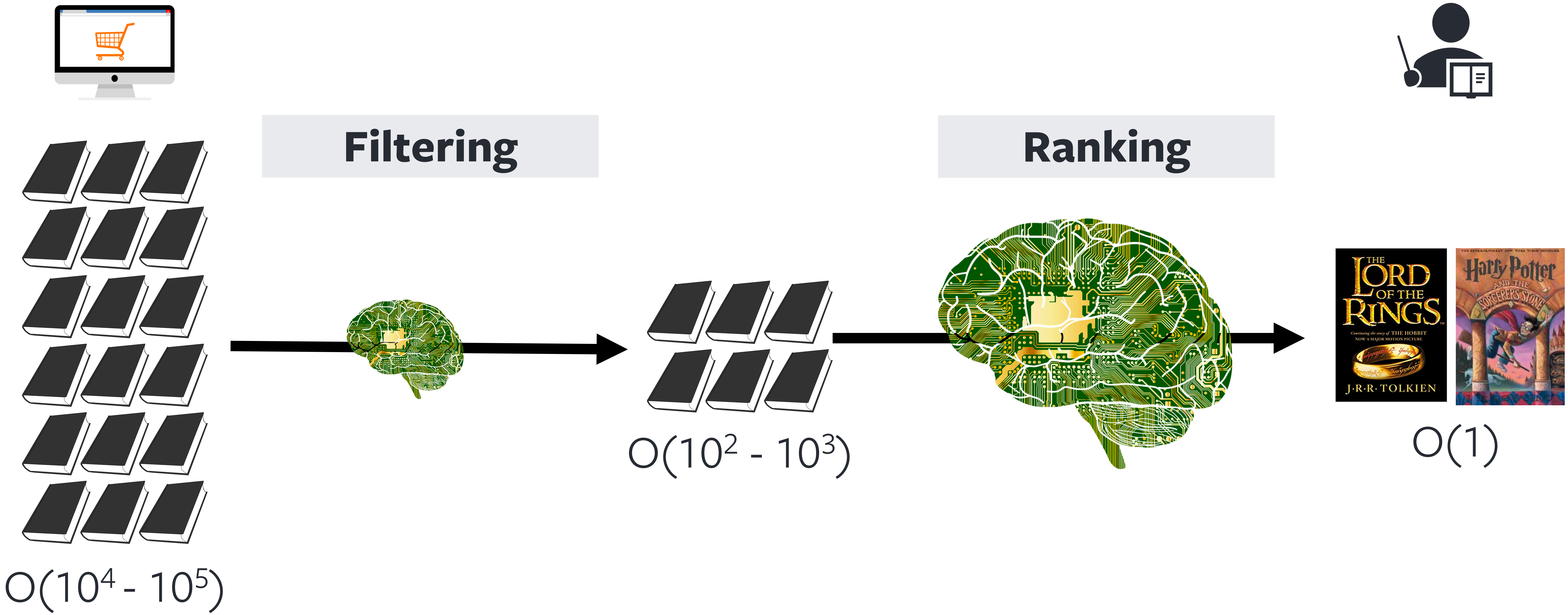


$O(10^4 - 10^5)$

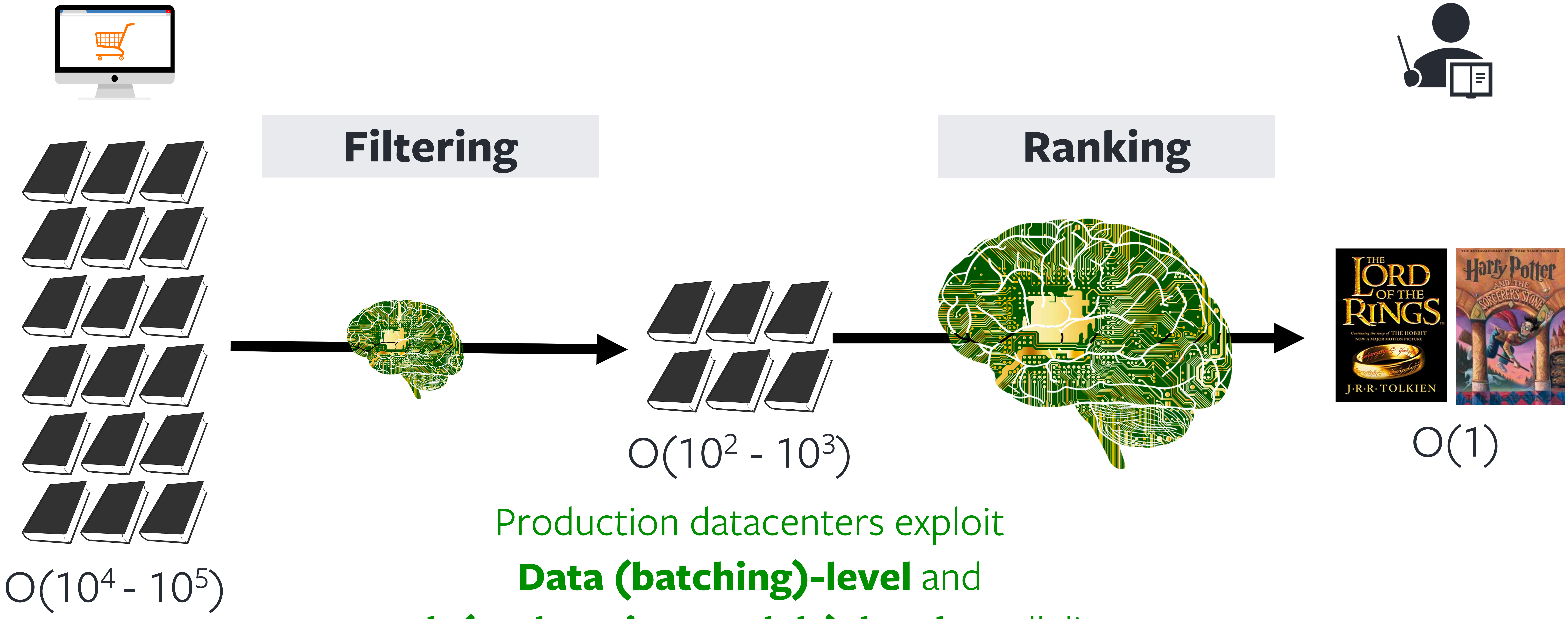
Ranking thousands of items at-scale



Ranking thousands of items at-scale



Ranking thousands of items at-scale



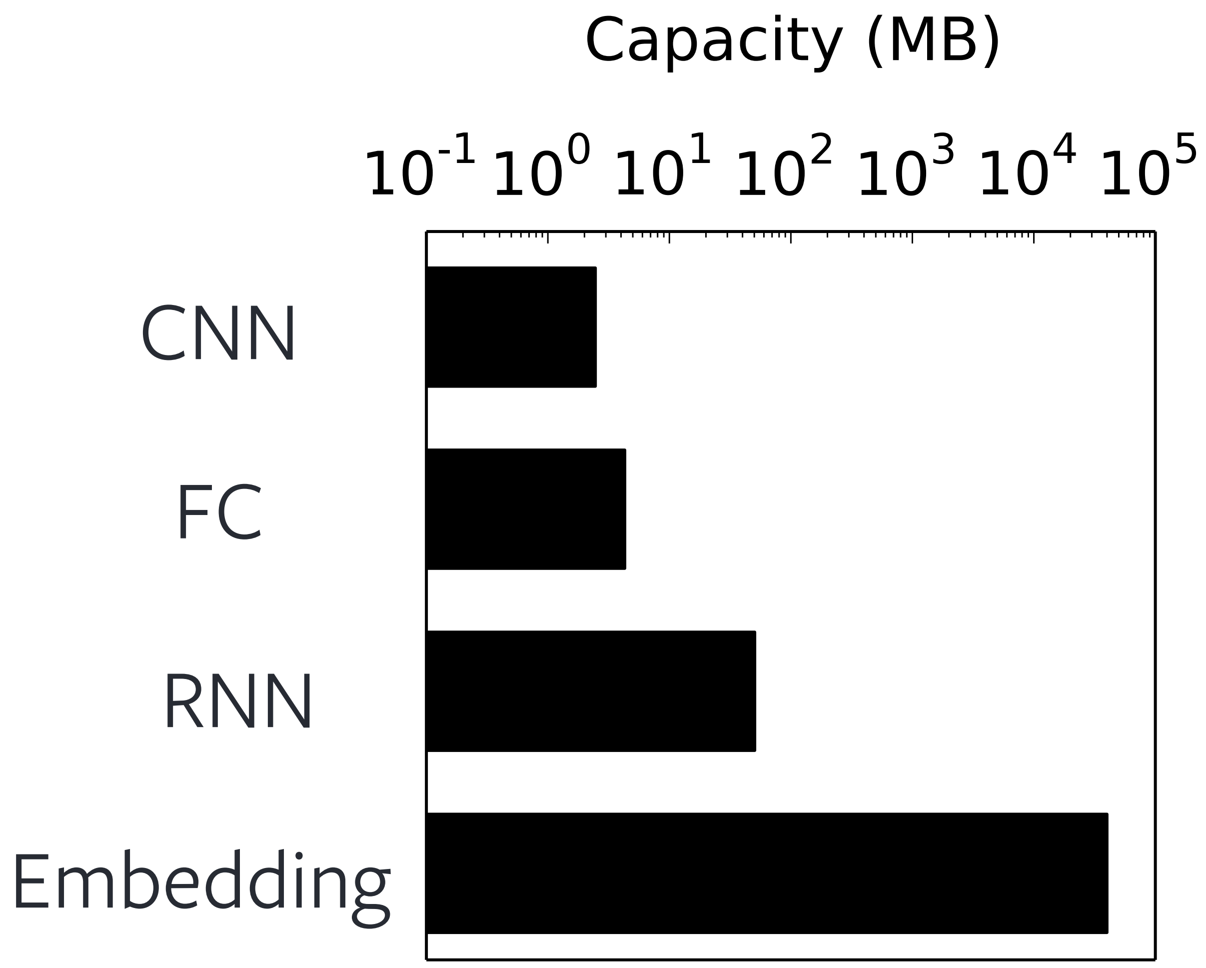
Production datacenters exploit
Data (batching)-level and
task (co-locating models)-level parallelism
for serving recommendations at-scale.

Embedding tables pose new challenges

Log Scale!

Embedding tables pose new challenges

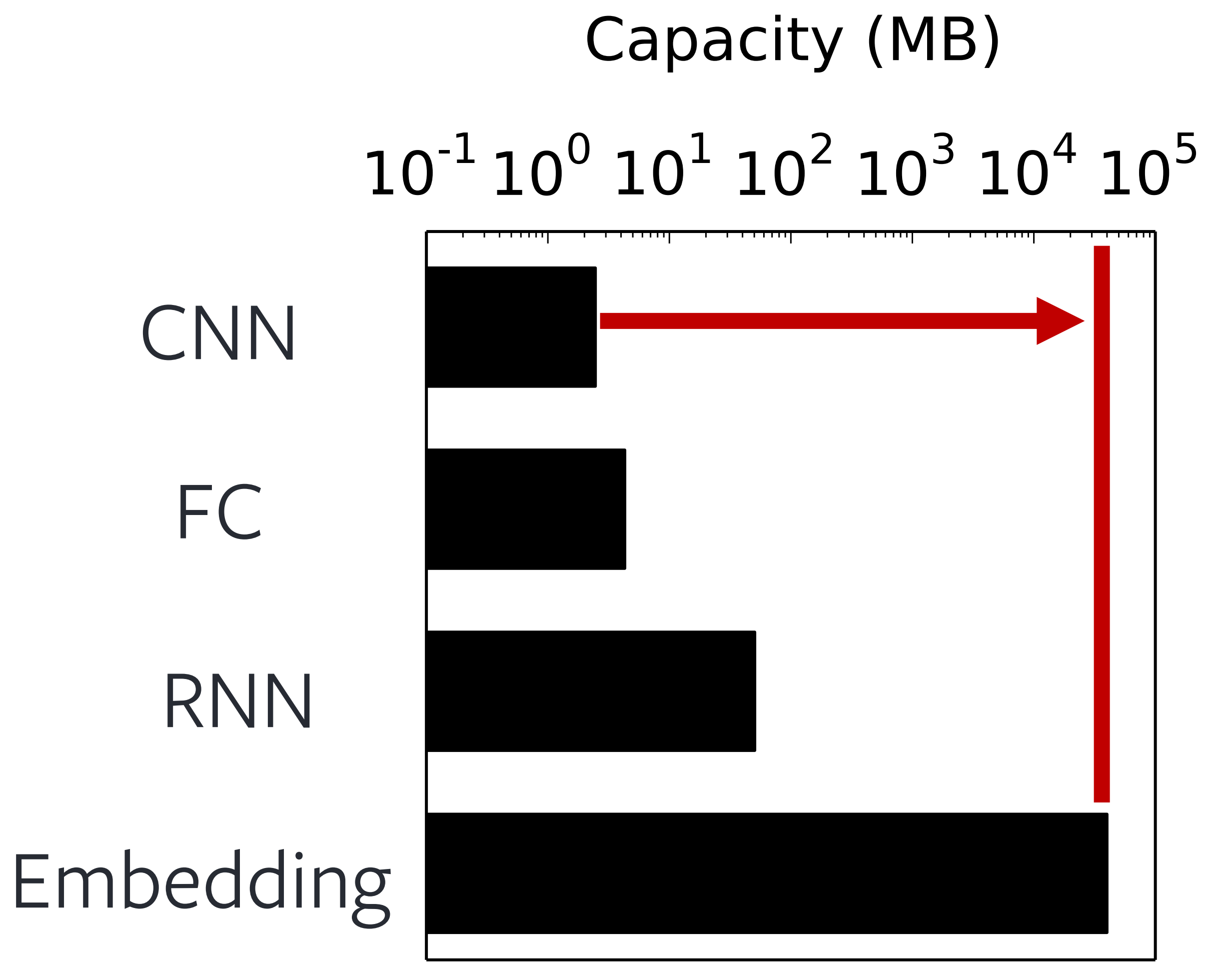
Log Scale!



Storage capacity

Embedding tables pose new challenges

Log Scale!

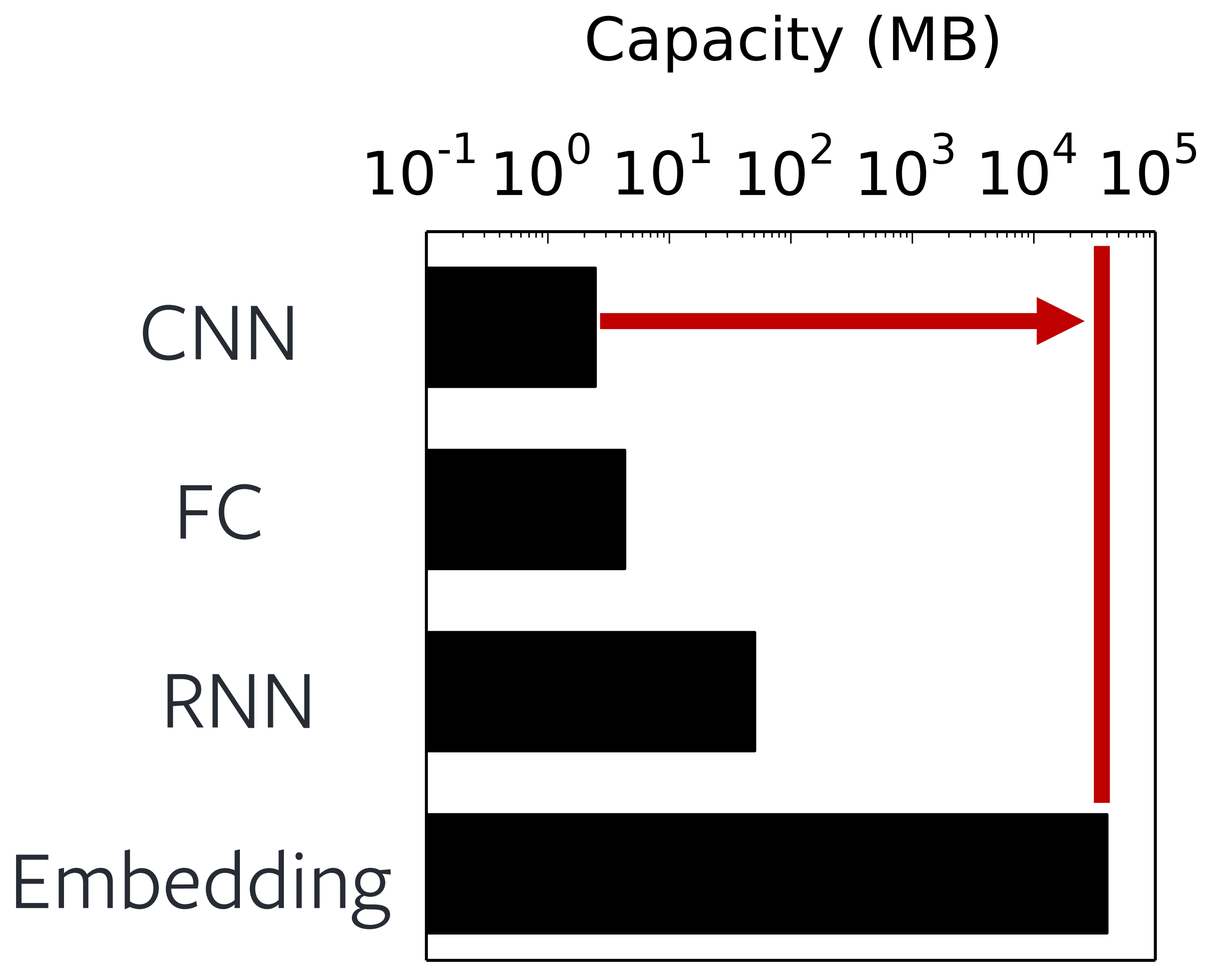


Storage capacity

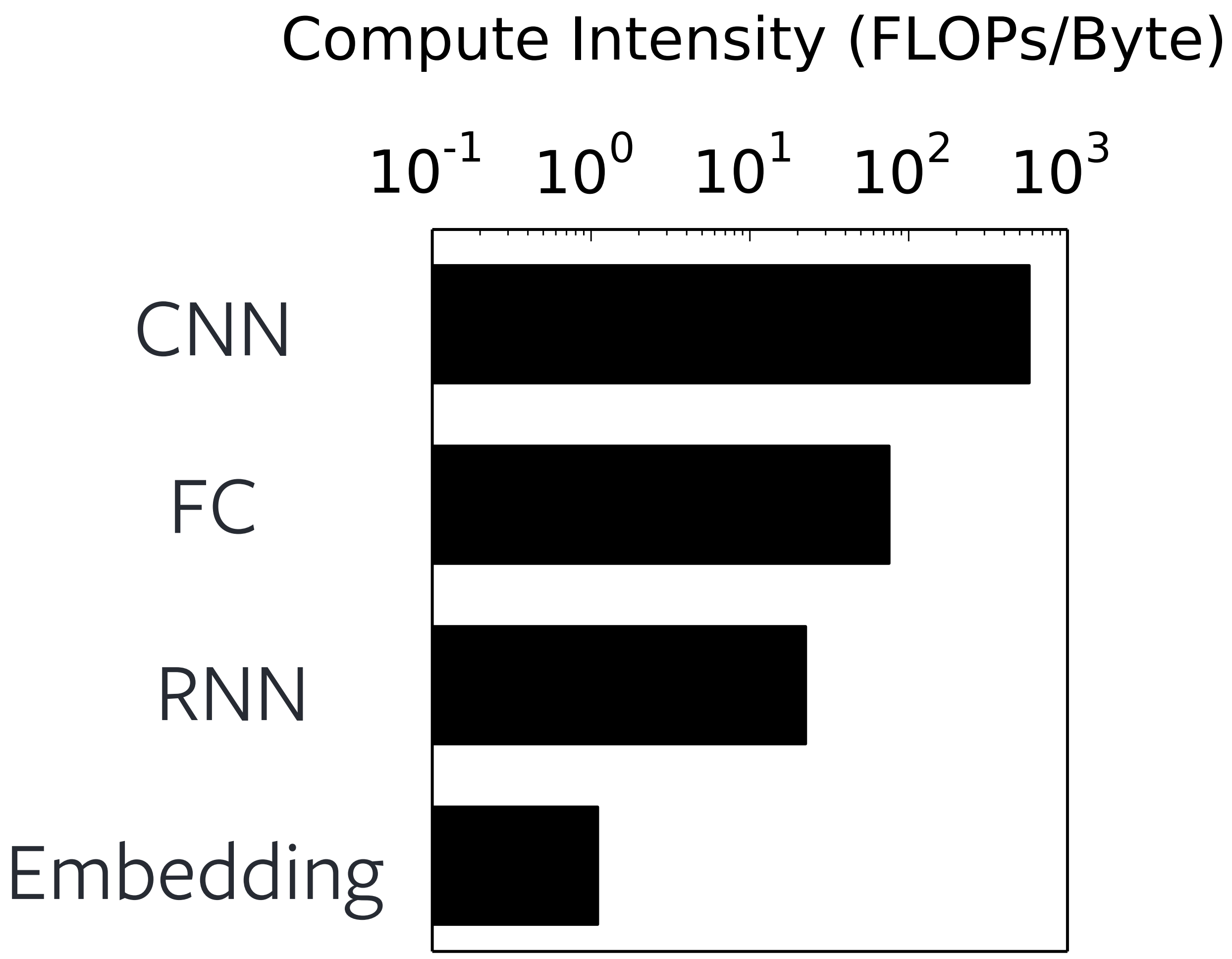
Up to tens of GBs

Embedding tables pose new challenges

Log Scale!



Storage capacity

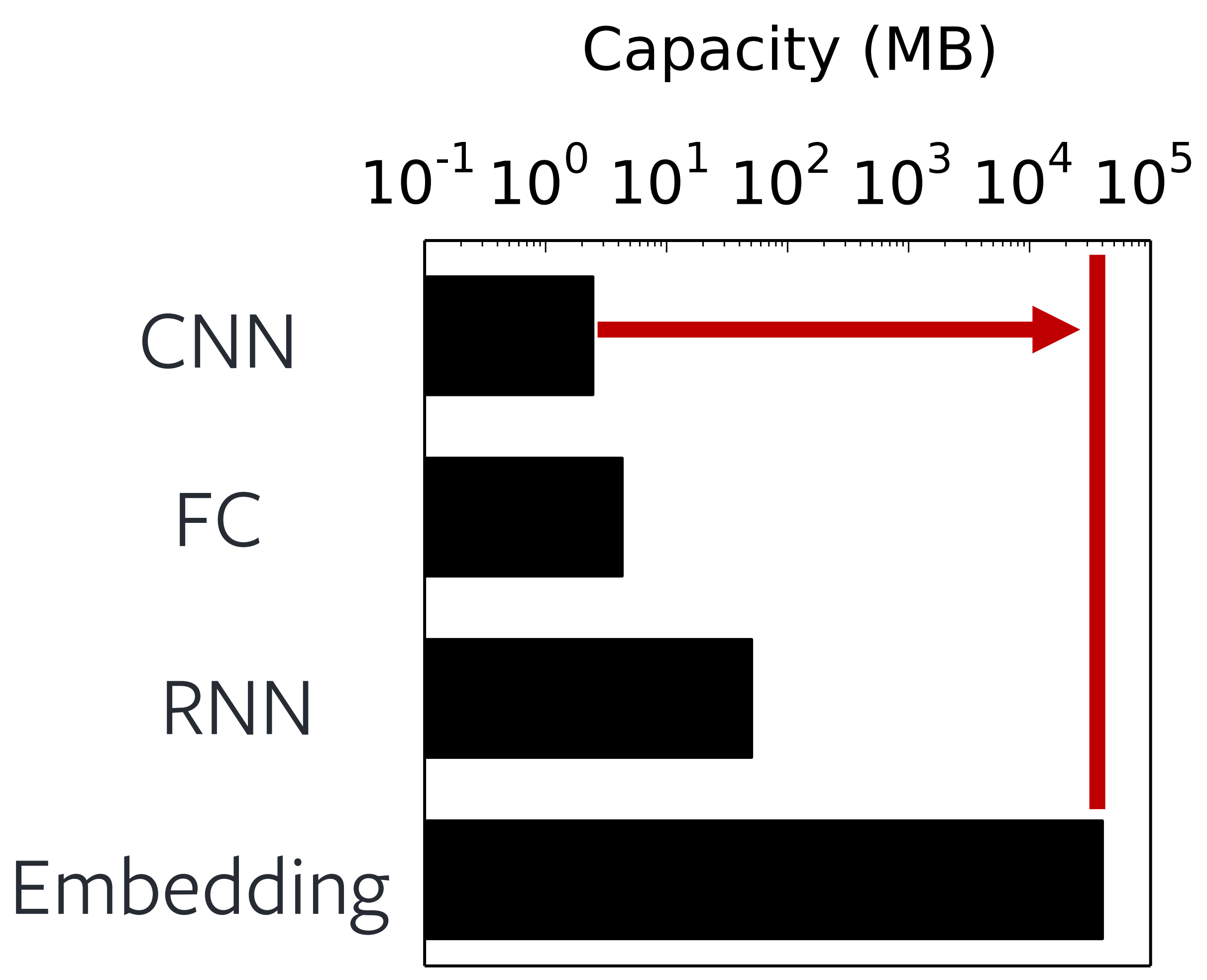


Compute intensity

Up to tens of GBs

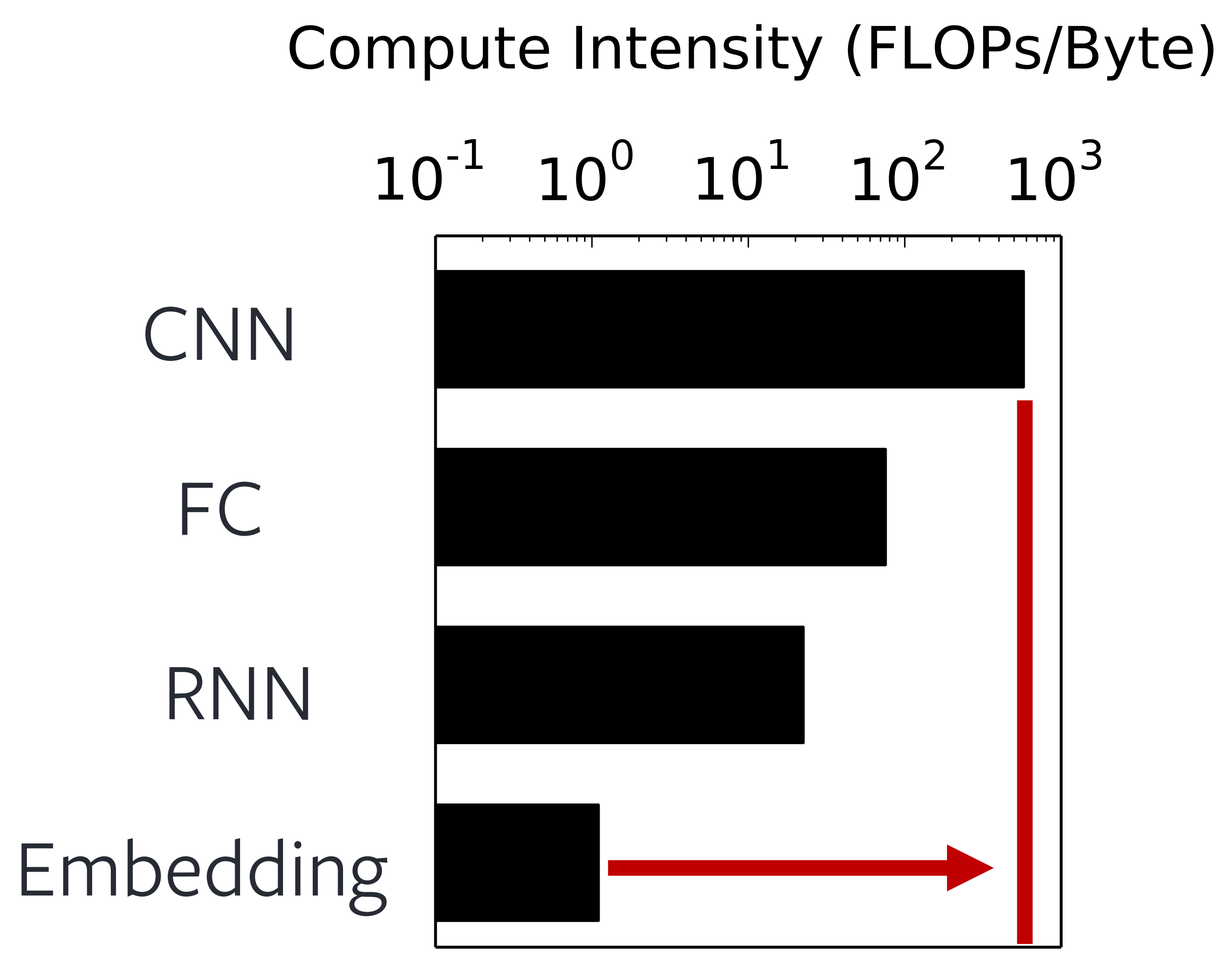
Embedding tables pose new challenges

Log Scale!



Storage capacity

Up to tens of GBs

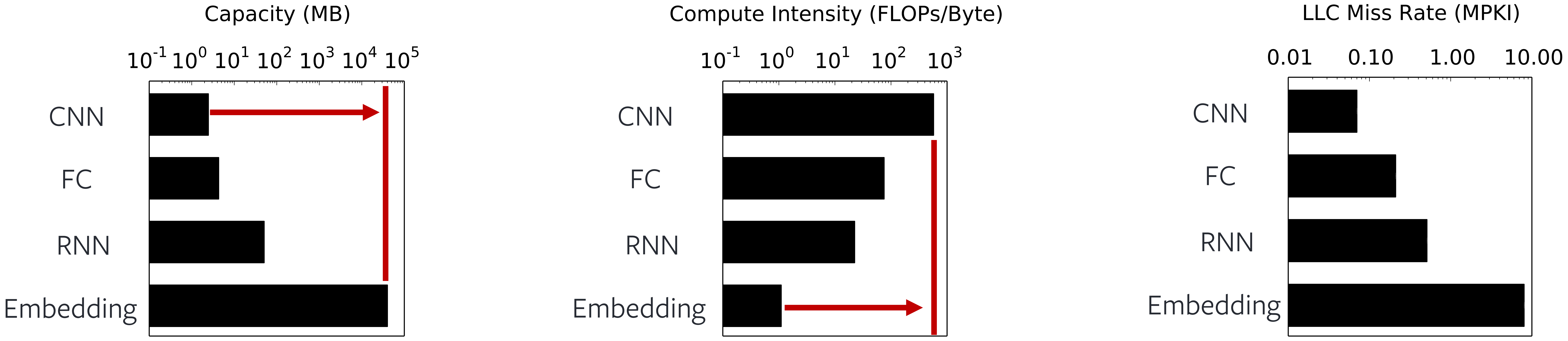


Compute intensity

Orders of magnitude
lower FLOPs/Byte

Embedding tables pose new challenges

Log Scale!



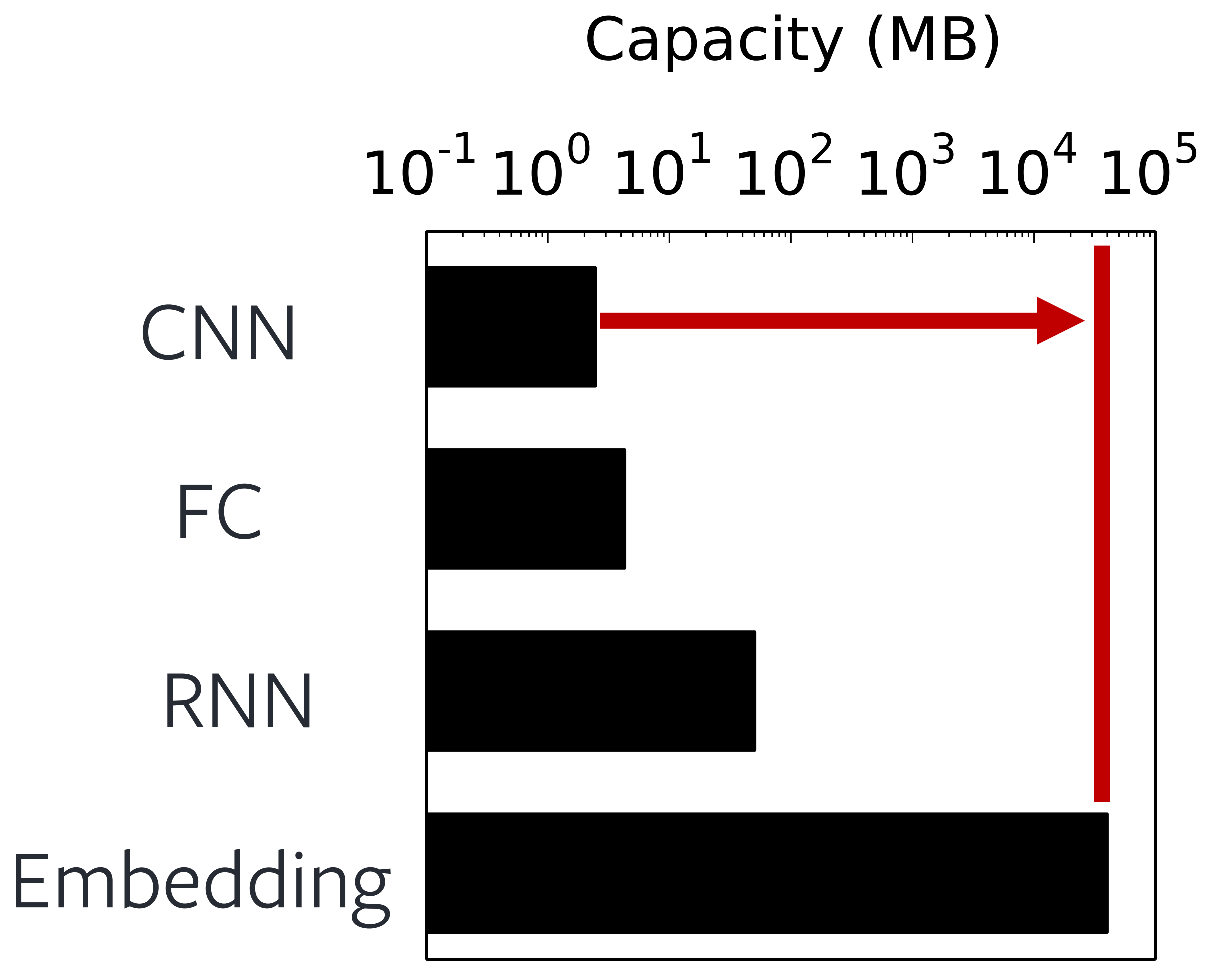
Storage capacity | **Compute intensity** | **Memory access pattern**

Up to tens of GBs

Orders of magnitude lower FLOPs/Byte

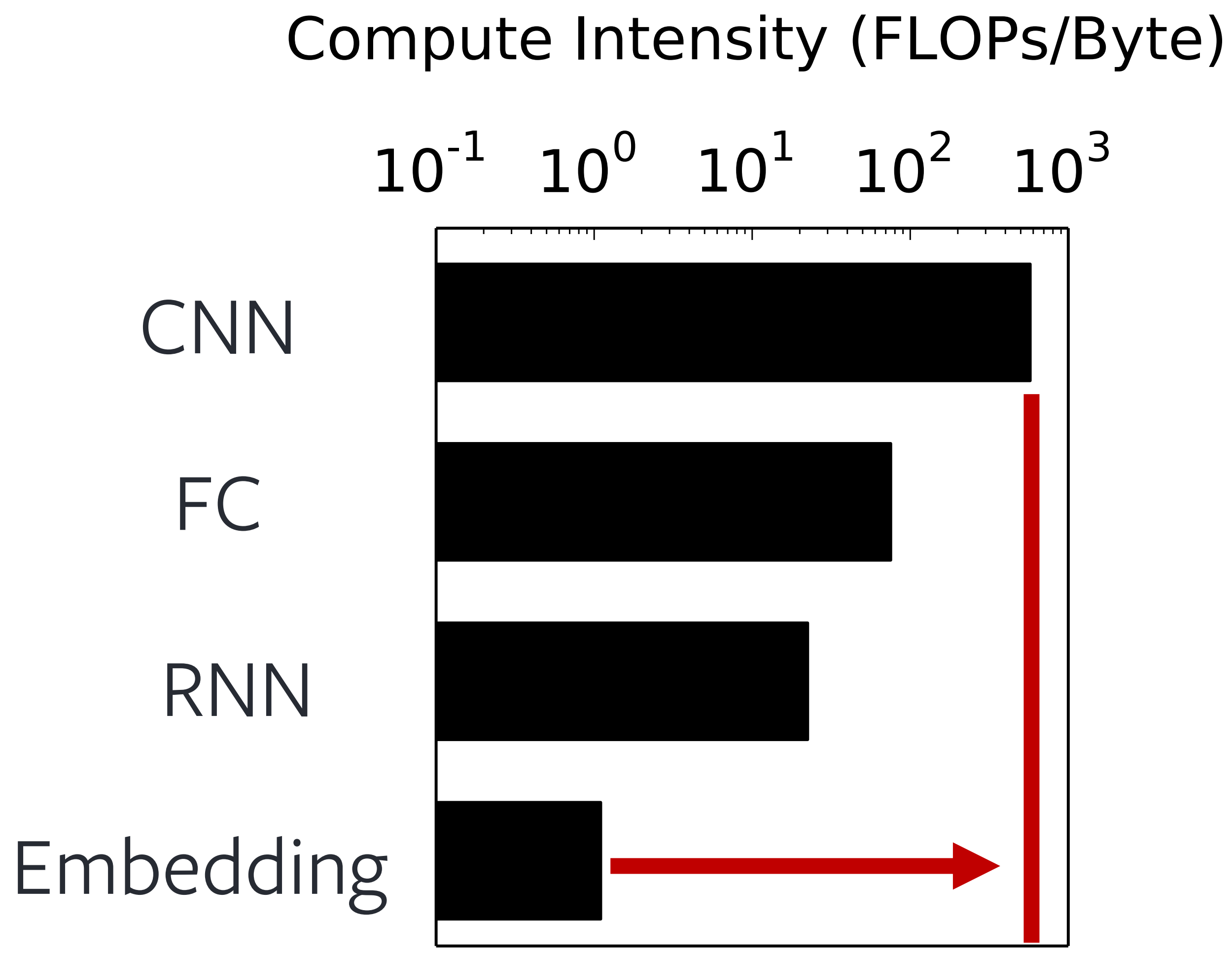
Embedding tables pose new challenges

Log Scale!



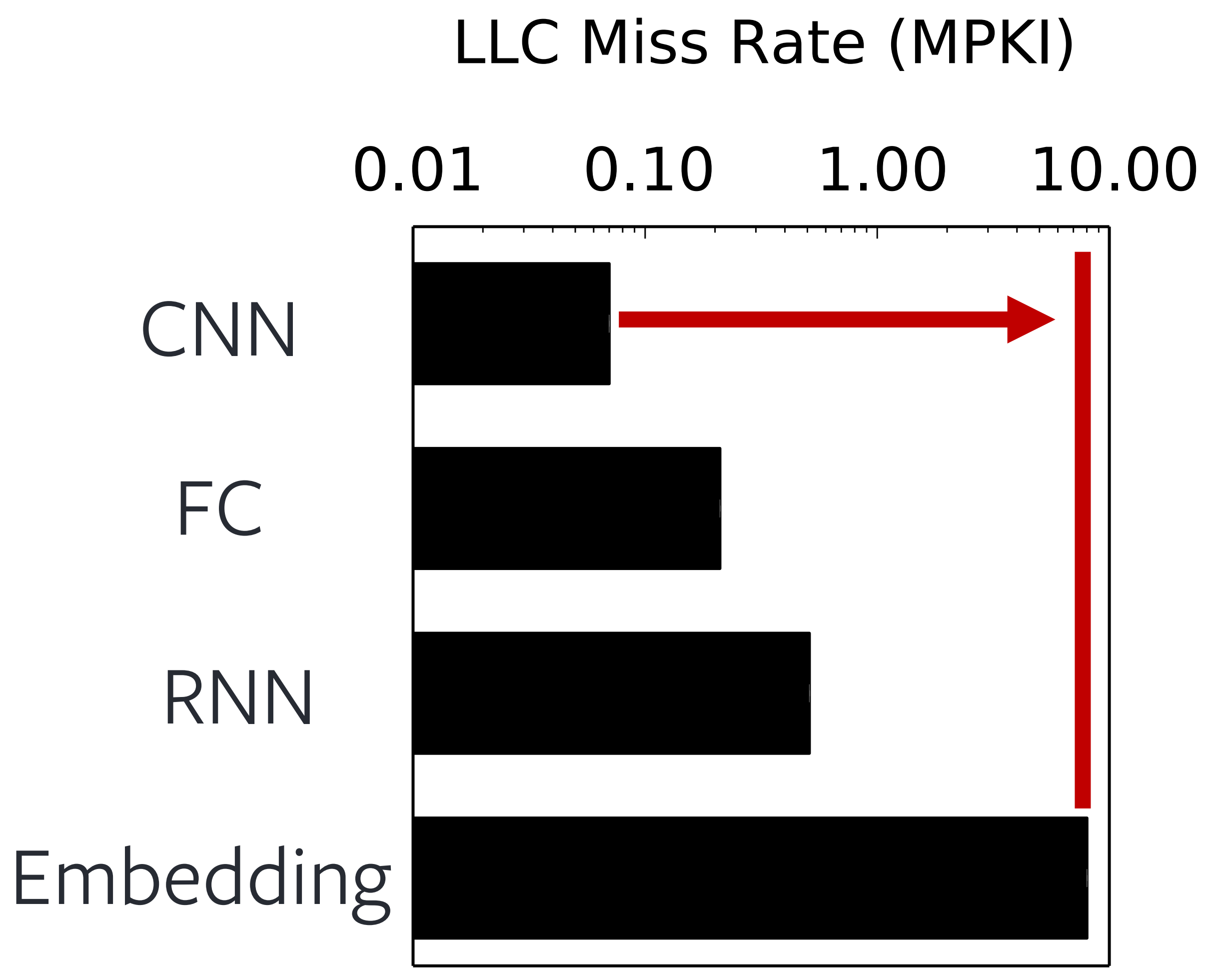
Storage capacity

Up to tens of GBs



Compute intensity

Orders of magnitude lower FLOPs/Byte

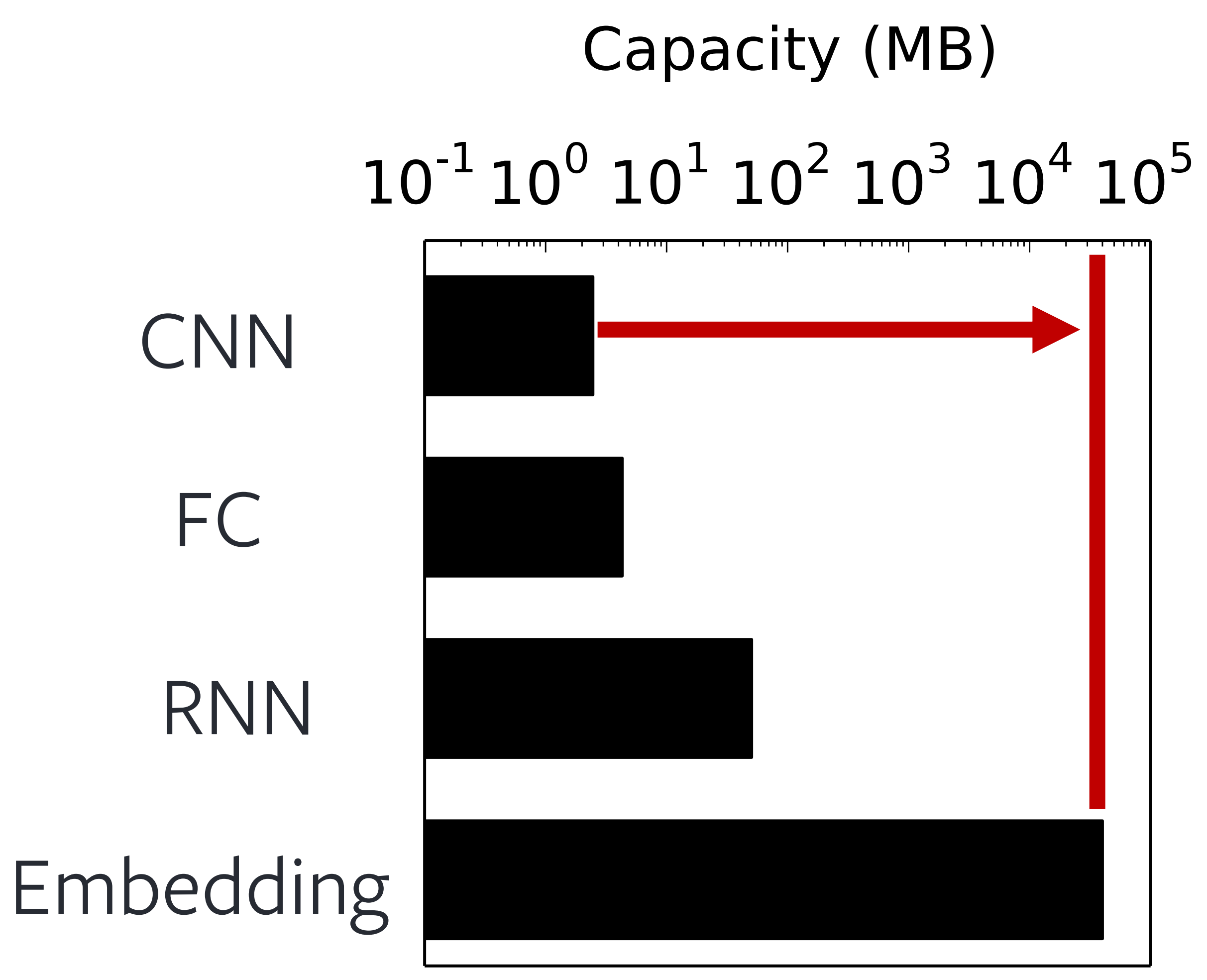


Memory access pattern

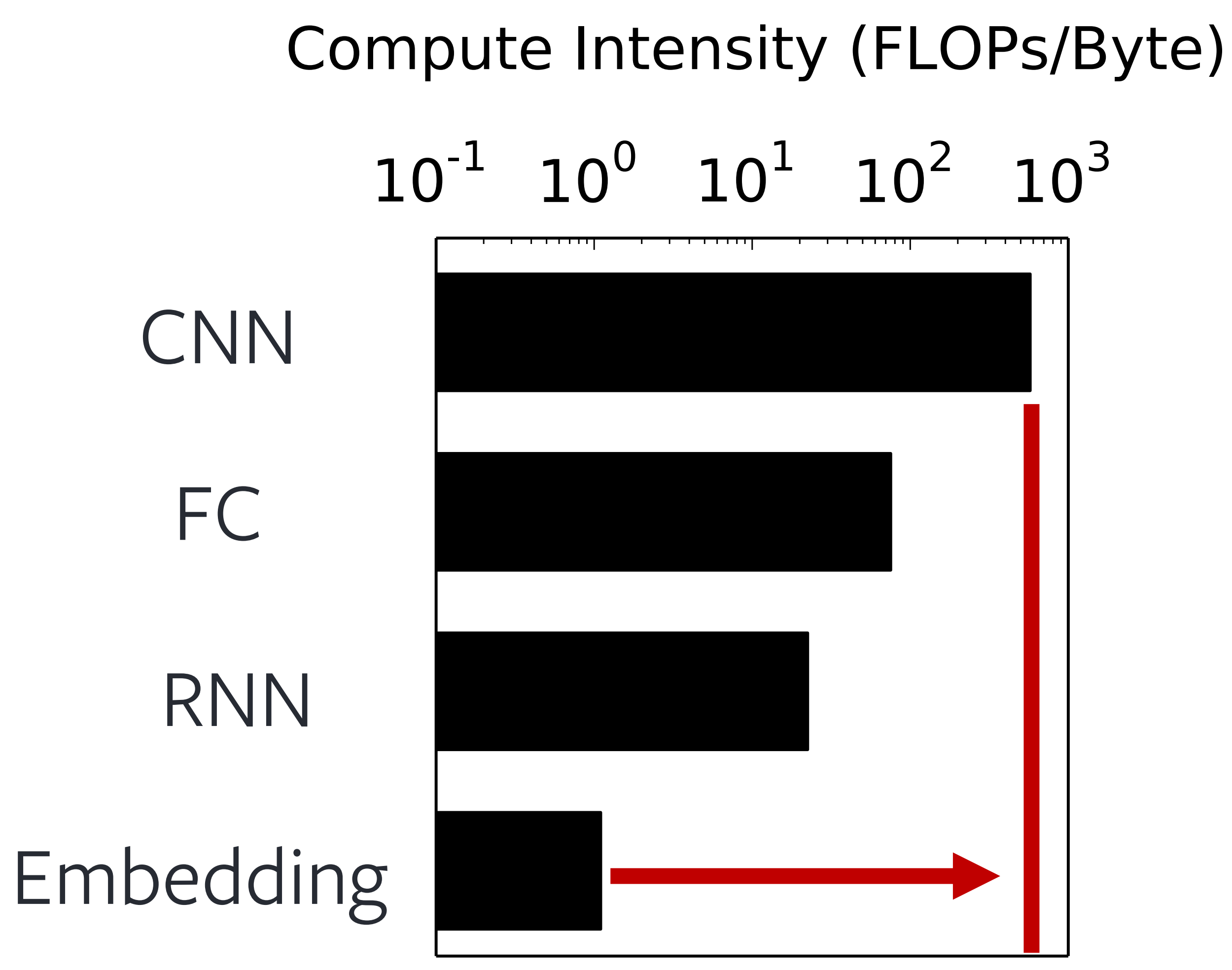
Sparse, irregular memory accesses

Embedding tables pose new challenges

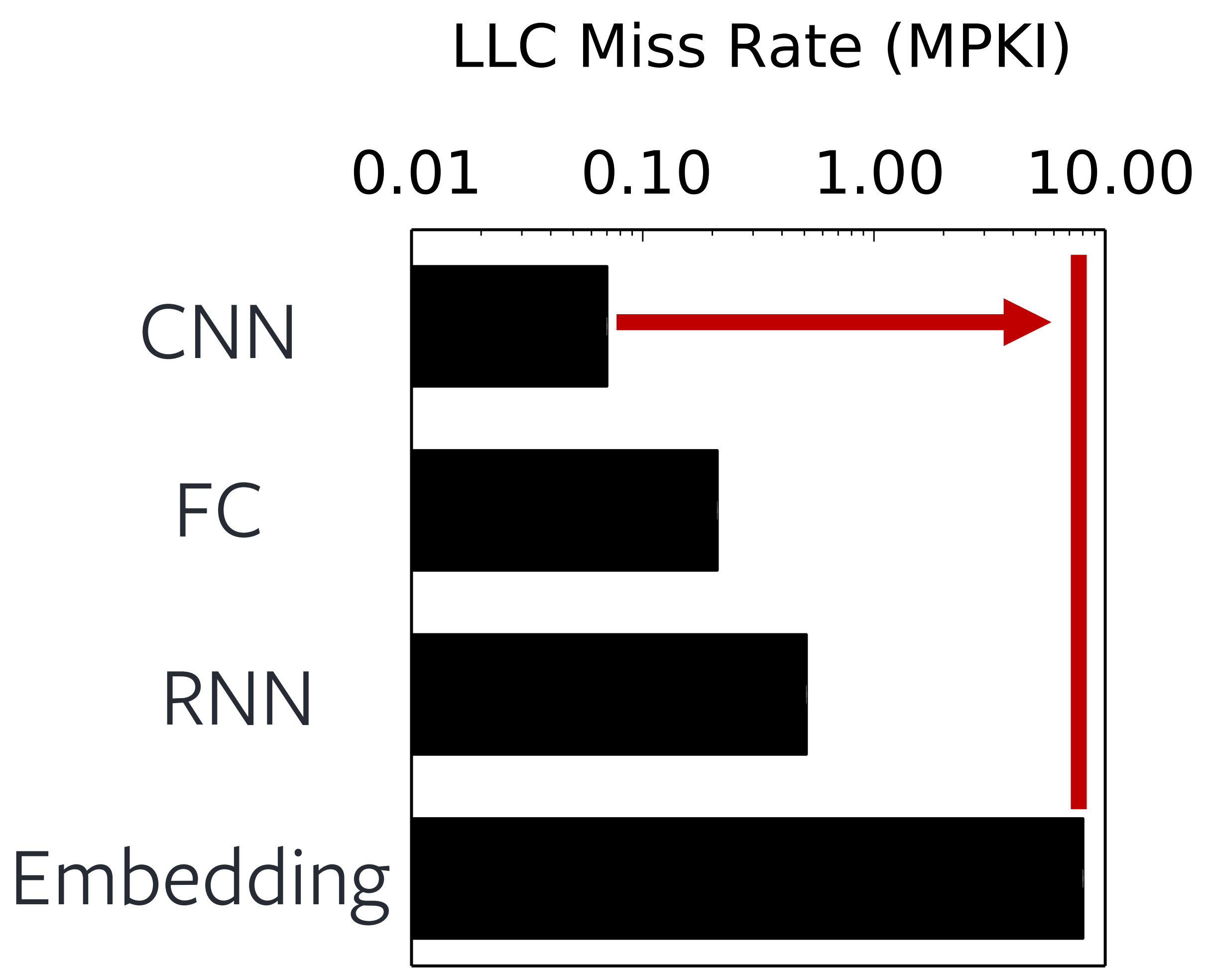
Log Scale!



Storage capacity



Compute intensity



Memory access pattern

Up to tens of GBs

Orders of magnitude lower FLOPs/Byte

Sparse, irregular memory accesses

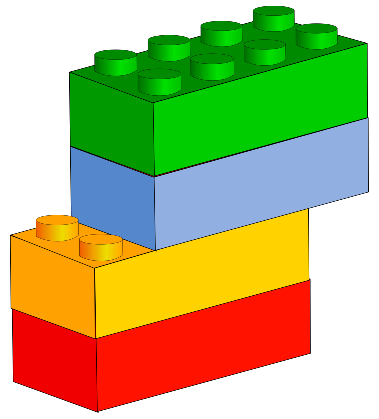
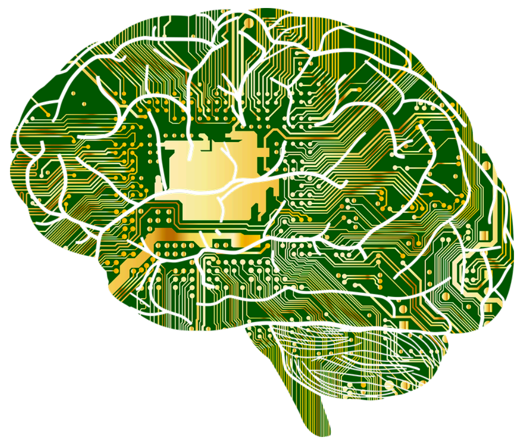
Systems and hardware opportunities

Off-chip memory (DRAM, NVM)

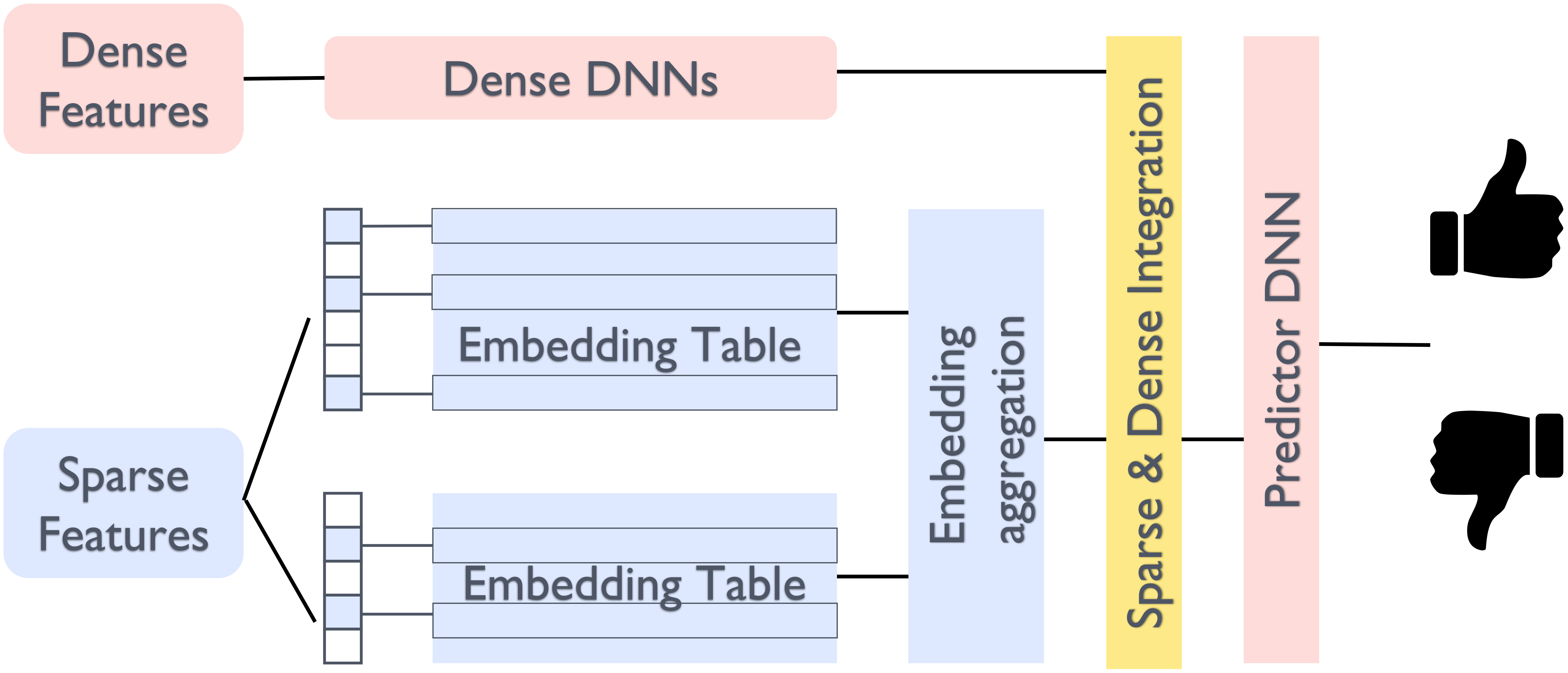
New accelerator designs (Near memory computing)

Specialized caching and pre-fetching capabilities

Hardware insights of recommendation

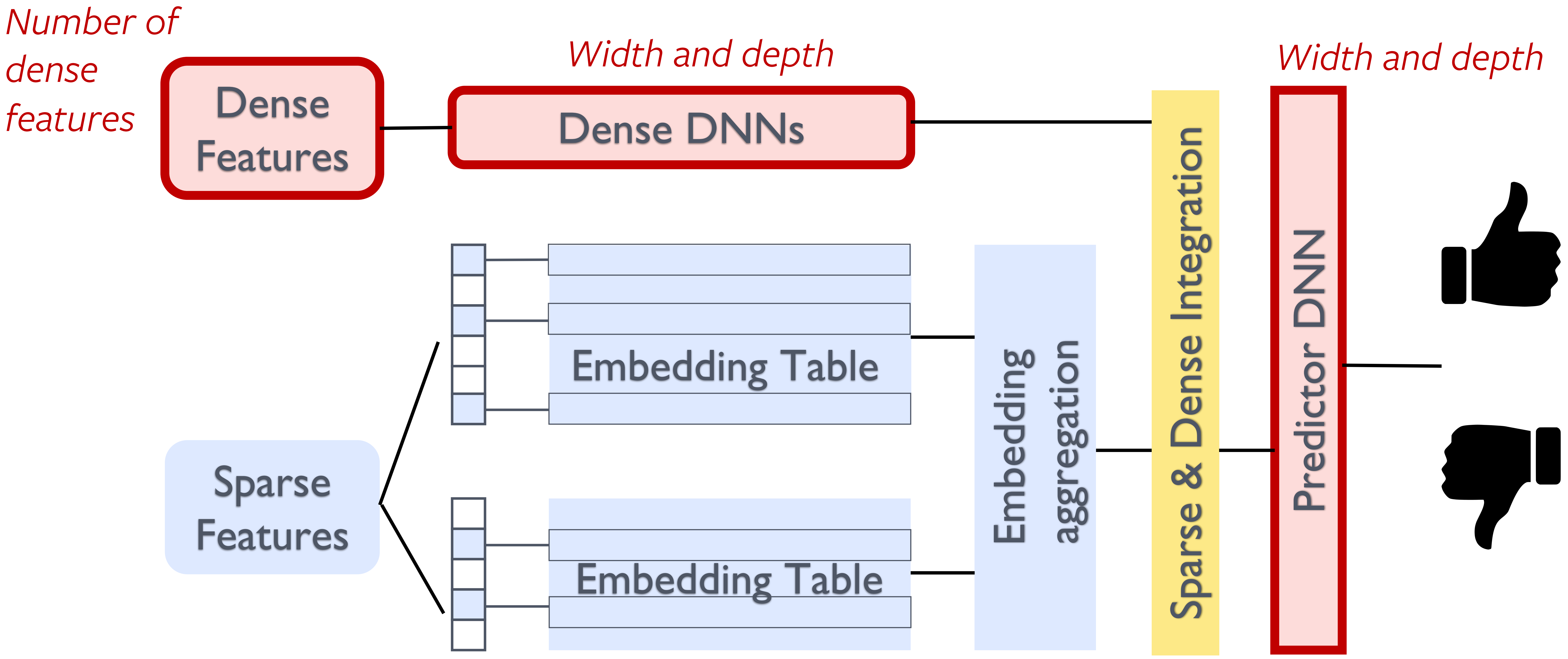
Algorithm	Hardware insights and opportunities
 General model structure	Optimize operators with new storage, compute, and memory access patterns
 Diverse networks architectures	Accelerate recommendation with flexible and diverse system solutions
At-scale inference	Exploit hardware heterogeneity and parallelism to optimize latency-bounded throughput

Facebook's DLRM: Configurable benchmark for end to end models



"Deep Learning Recommendation Model for Personalization and Recommendation Systems" Naumov, et. al.

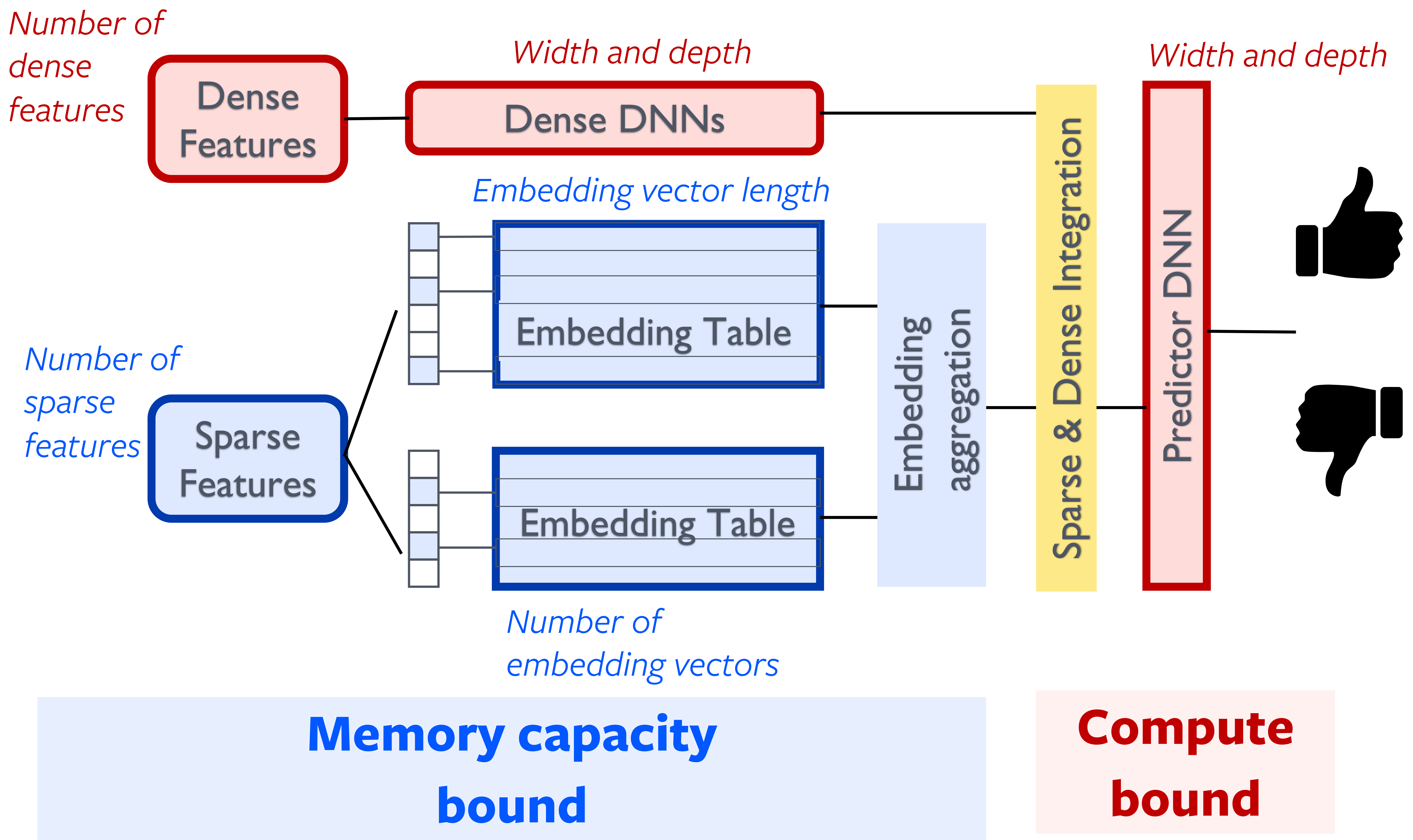
Facebook's DLRM: Configurable benchmark for end to end models



Compute bound

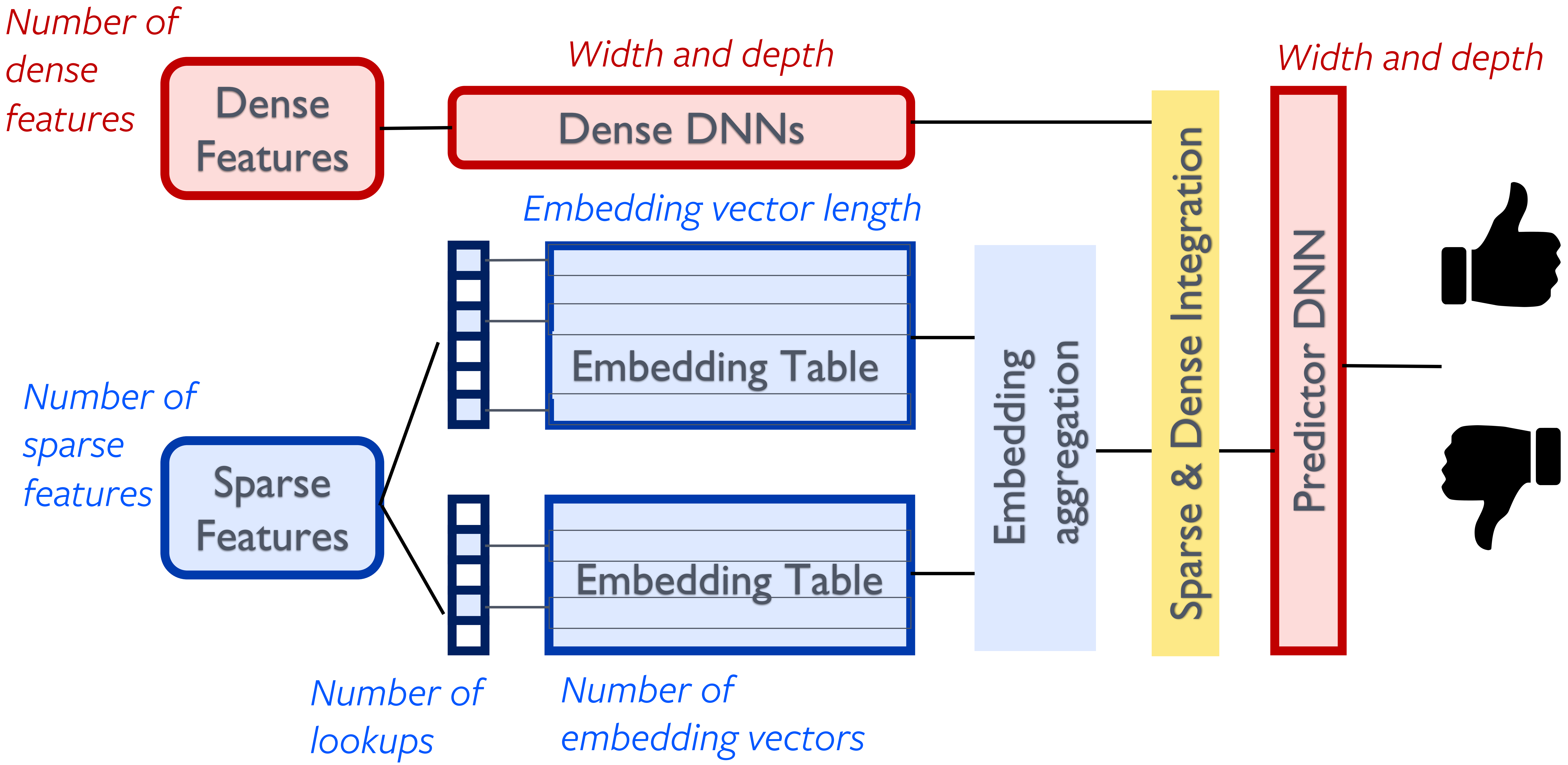
"Deep Learning Recommendation Model for Personalization and Recommendation Systems" Naumov, et. al.

Facebook's DLRM: Configurable benchmark for end to end models



"Deep Learning Recommendation Model for Personalization and Recommendation Systems" Naumov, et. al.

Facebook's DLRM: Configurable benchmark for end to end models



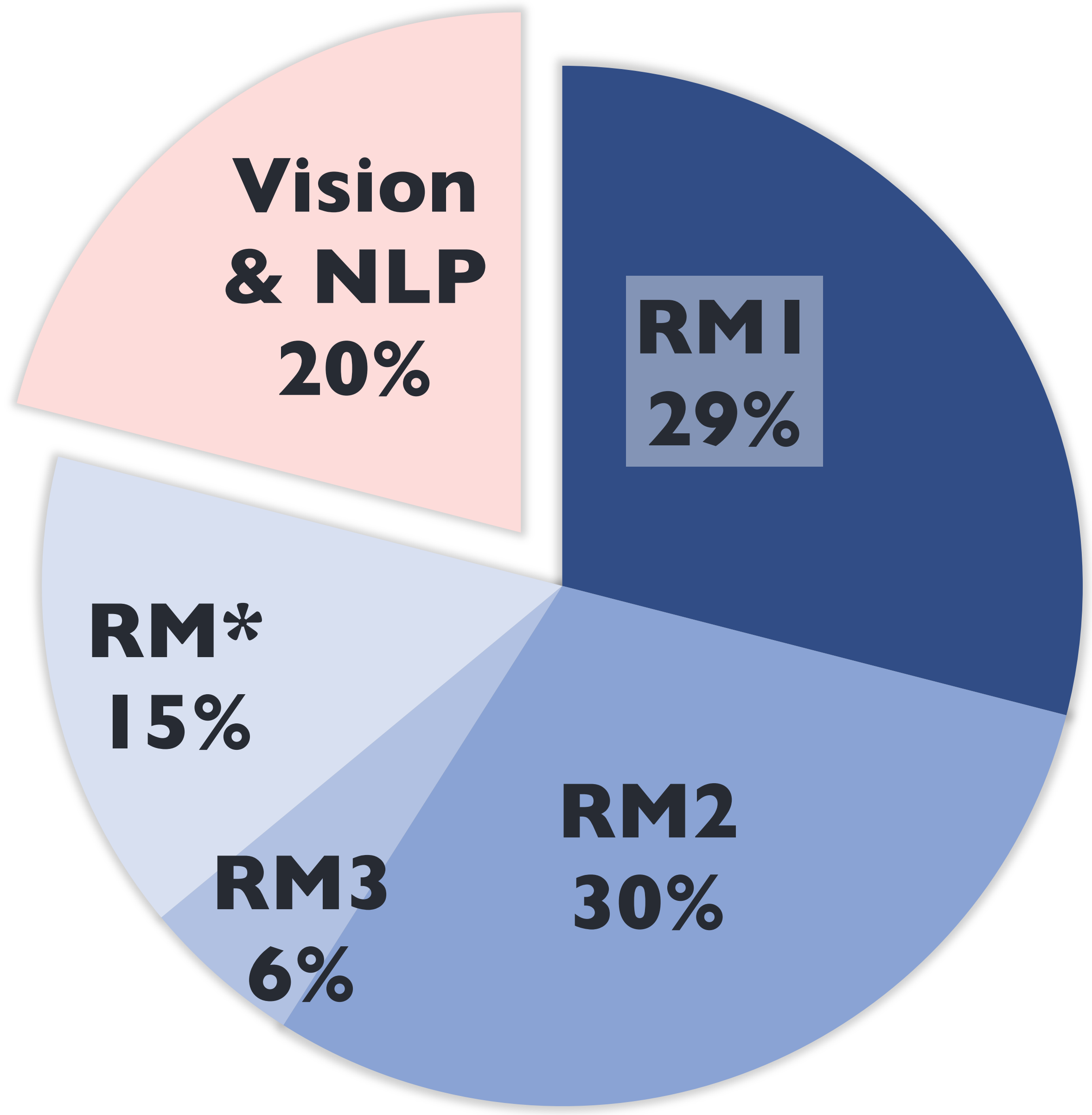
Memory capacity & bandwidth bound

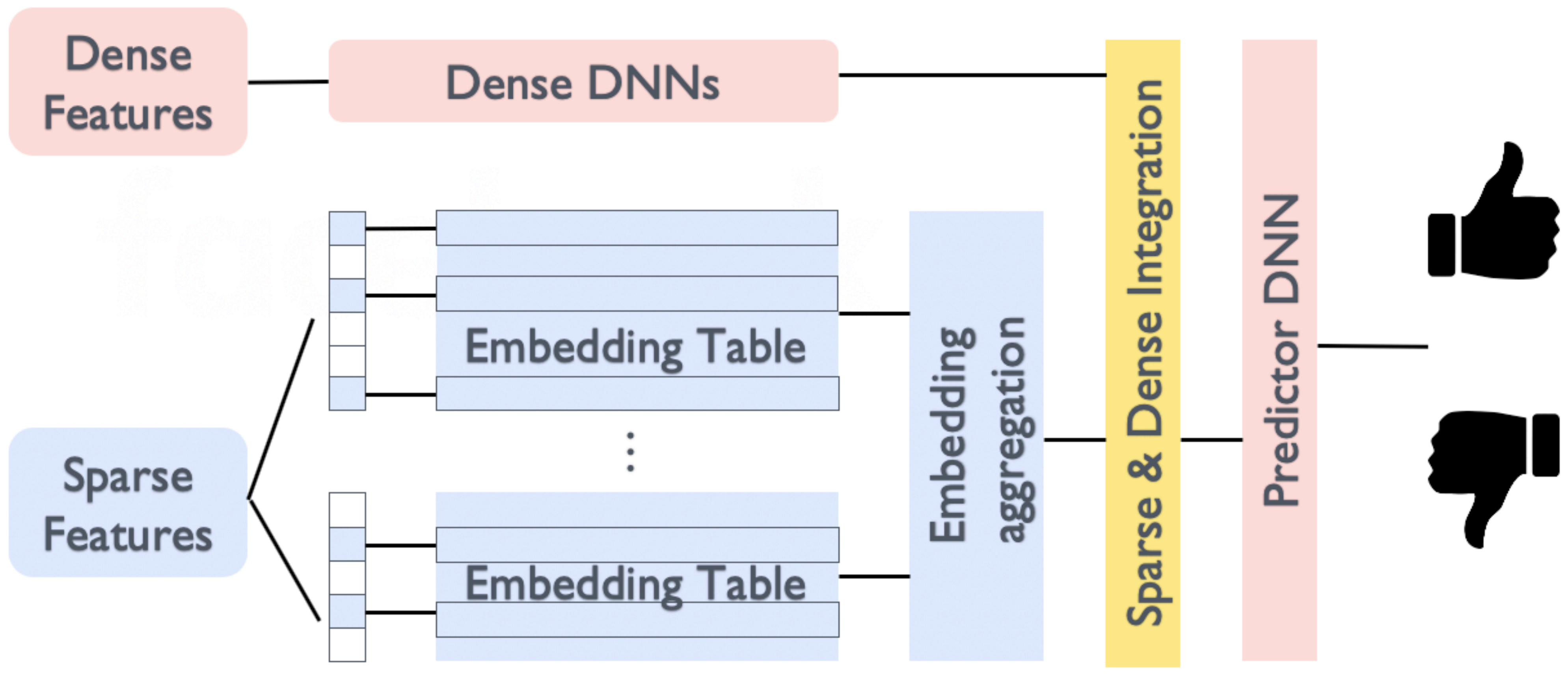
Compute bound

"Deep Learning Recommendation Model for Personalization and Recommendation Systems" Naumov, et. al.

Benchmarks represent key models in Facebook's datacenter

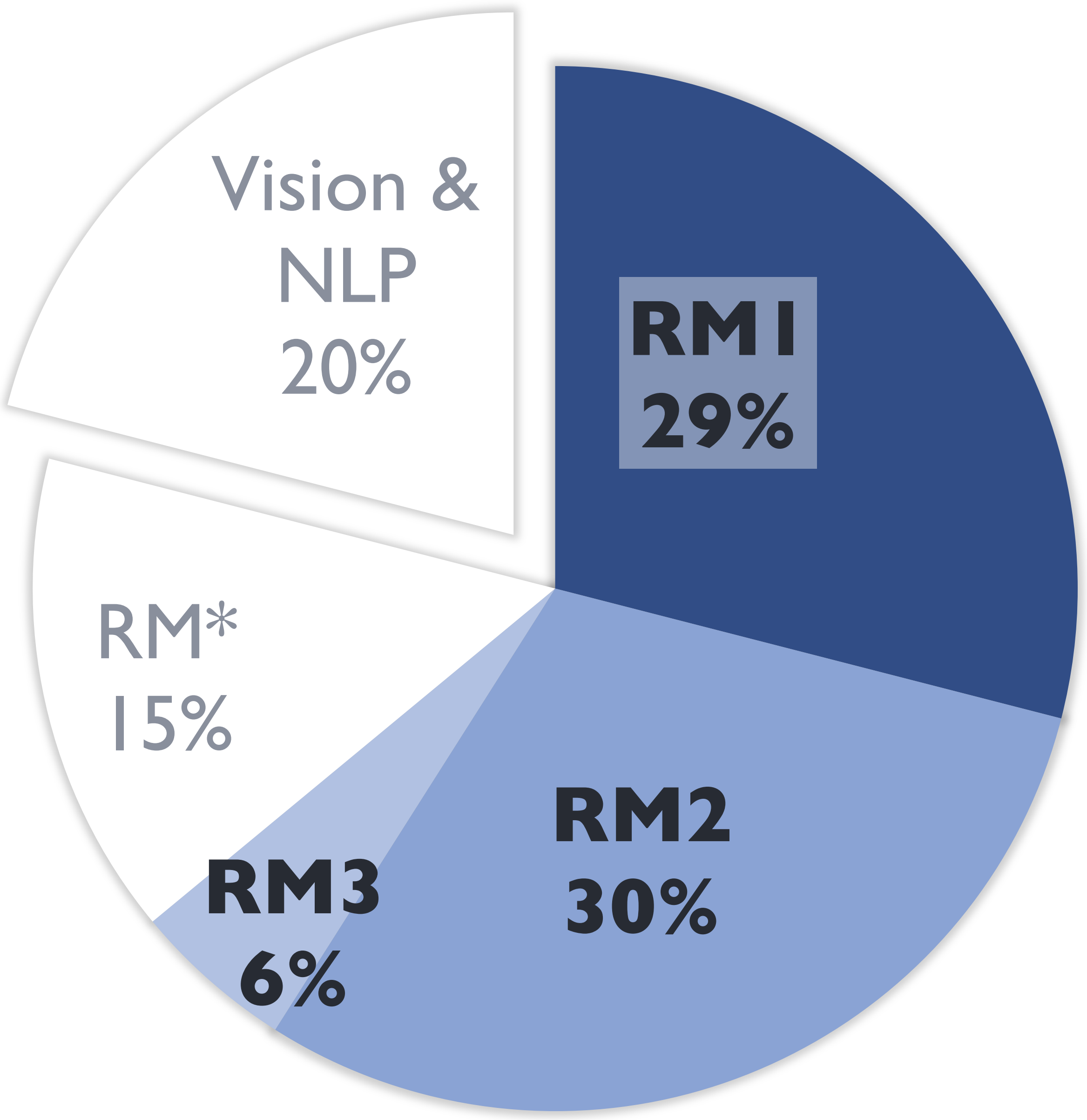
AI inference cycles in Facebook's datacenter

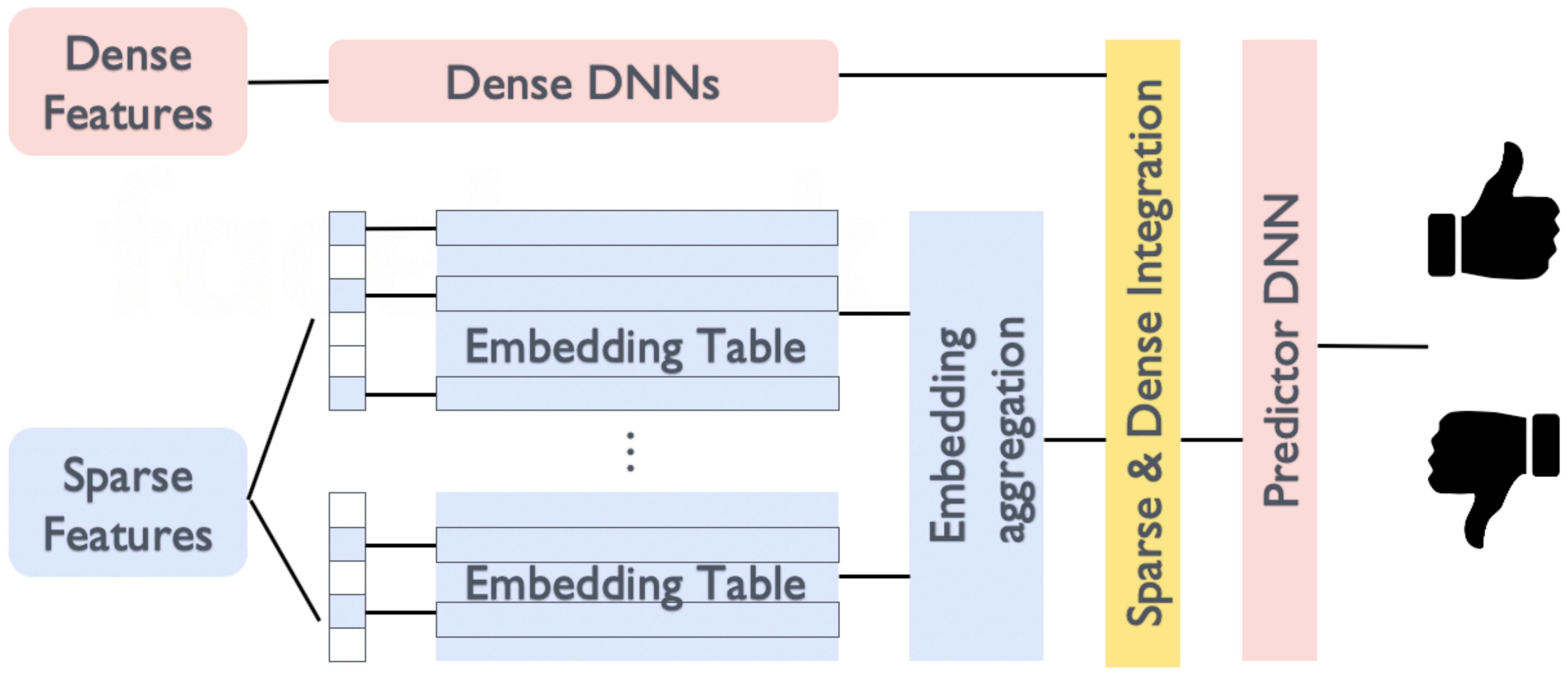




Benchmarks represent key models in Facebook's datacenter

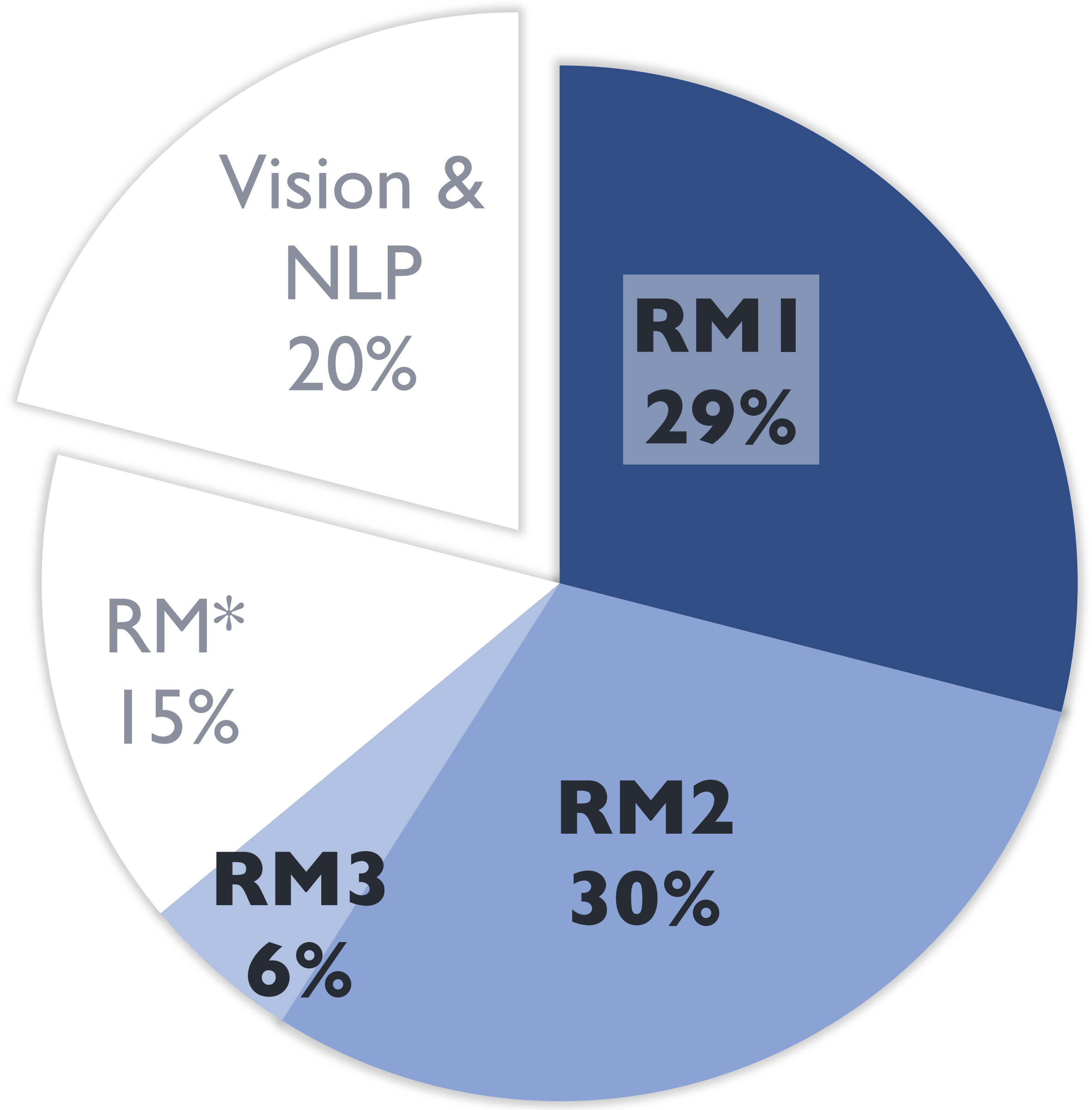
AI inference cycles in Facebook's datacenter

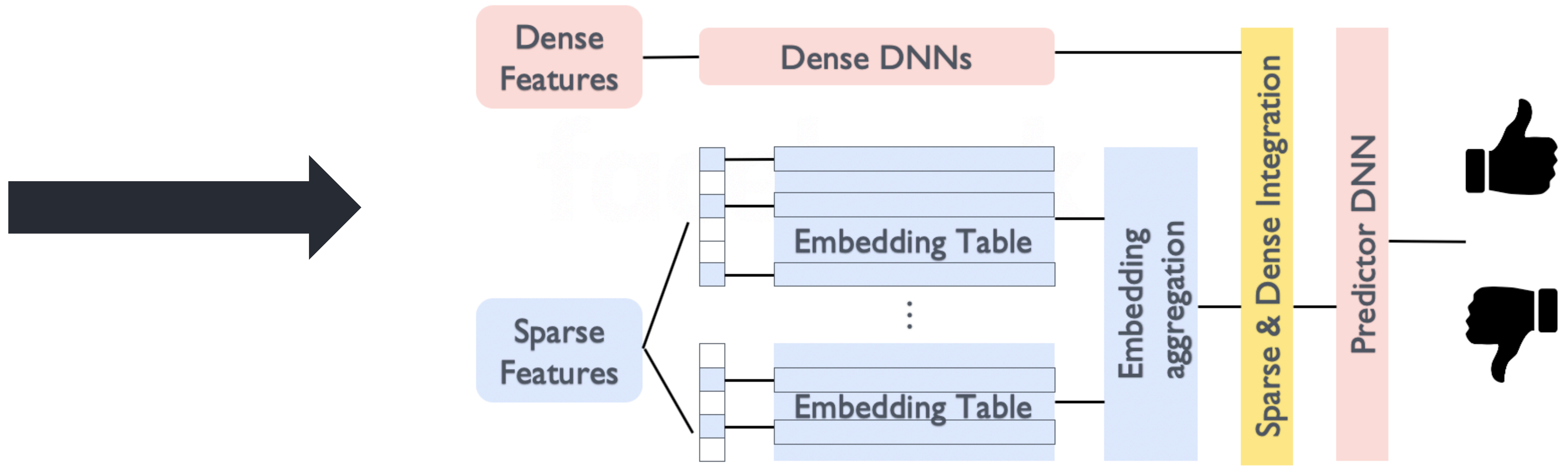




Benchmarks represent key models in Facebook's datacenter

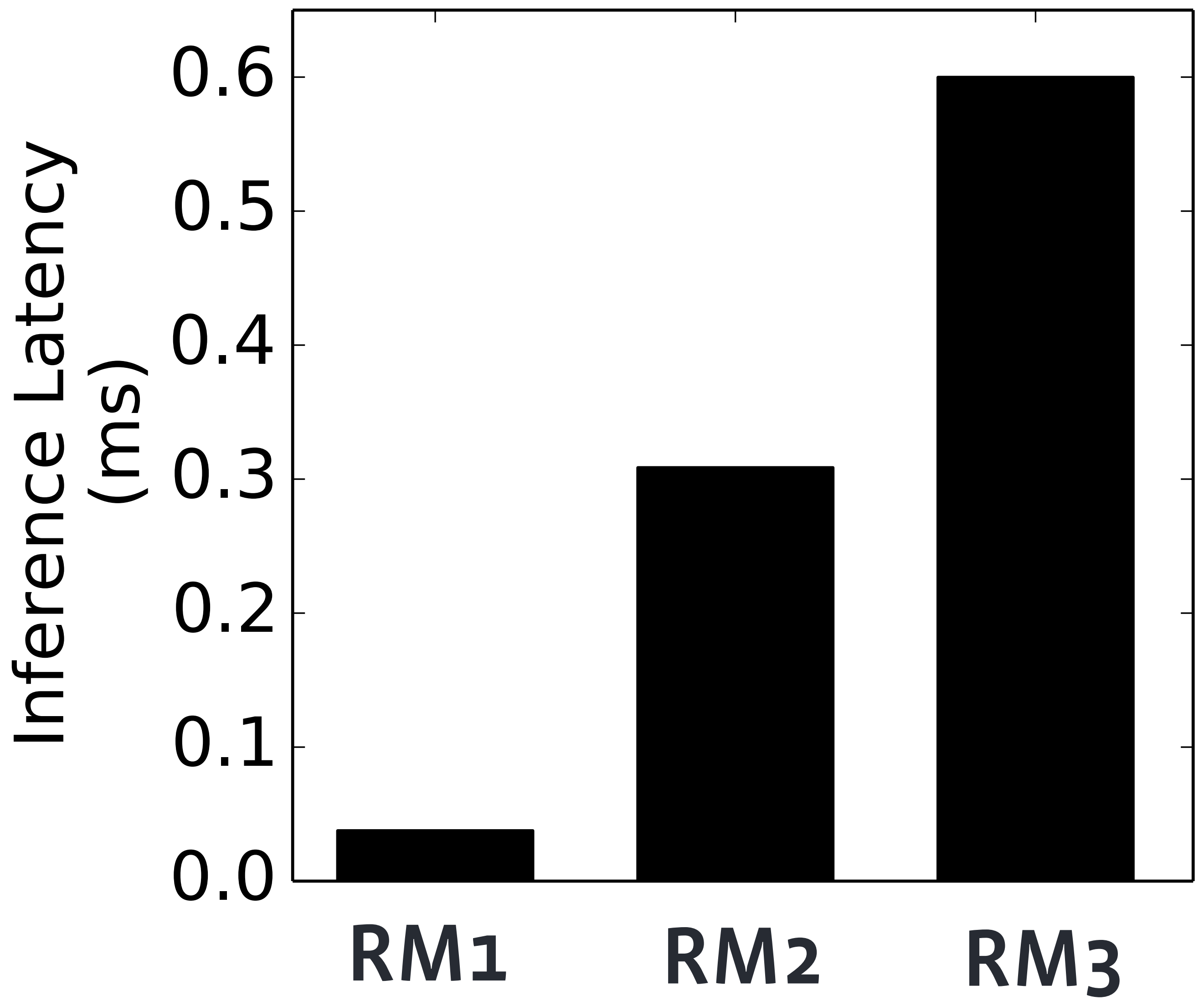
AI inference cycles in Facebook's datacenter



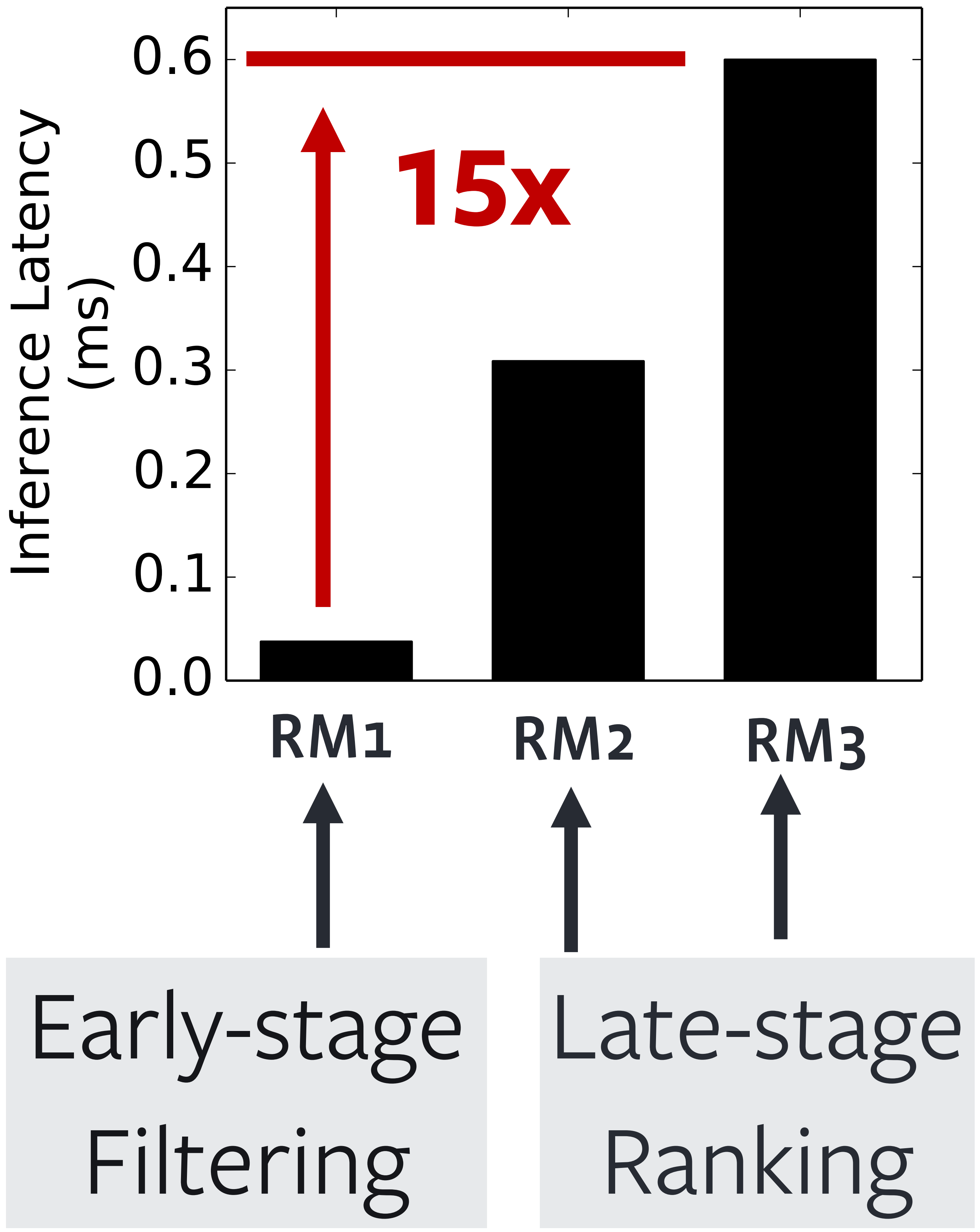


	RM1	RM2	RM3
Stage	Filtering	Ranking	Ranking
FC sizes	Small	Medium	Large
Number of embedding table	Few	Many	Few
Size of embeddings	Small	Medium	Large
Number of lookups per table	Hundreds	Hundreds	Tens

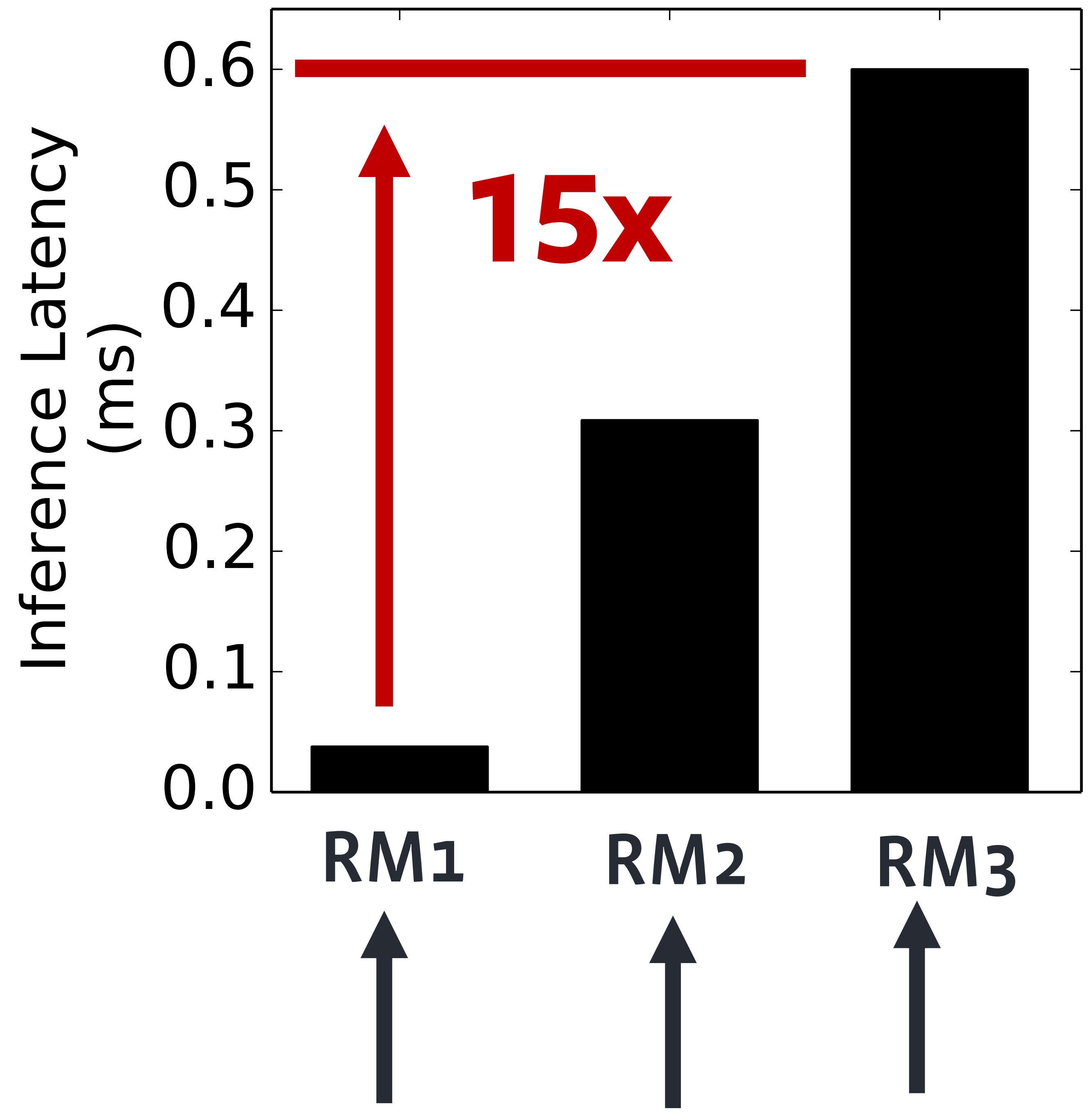
Performance characteristics of end-to-end models varies



Performance characteristics of end-to-end models varies

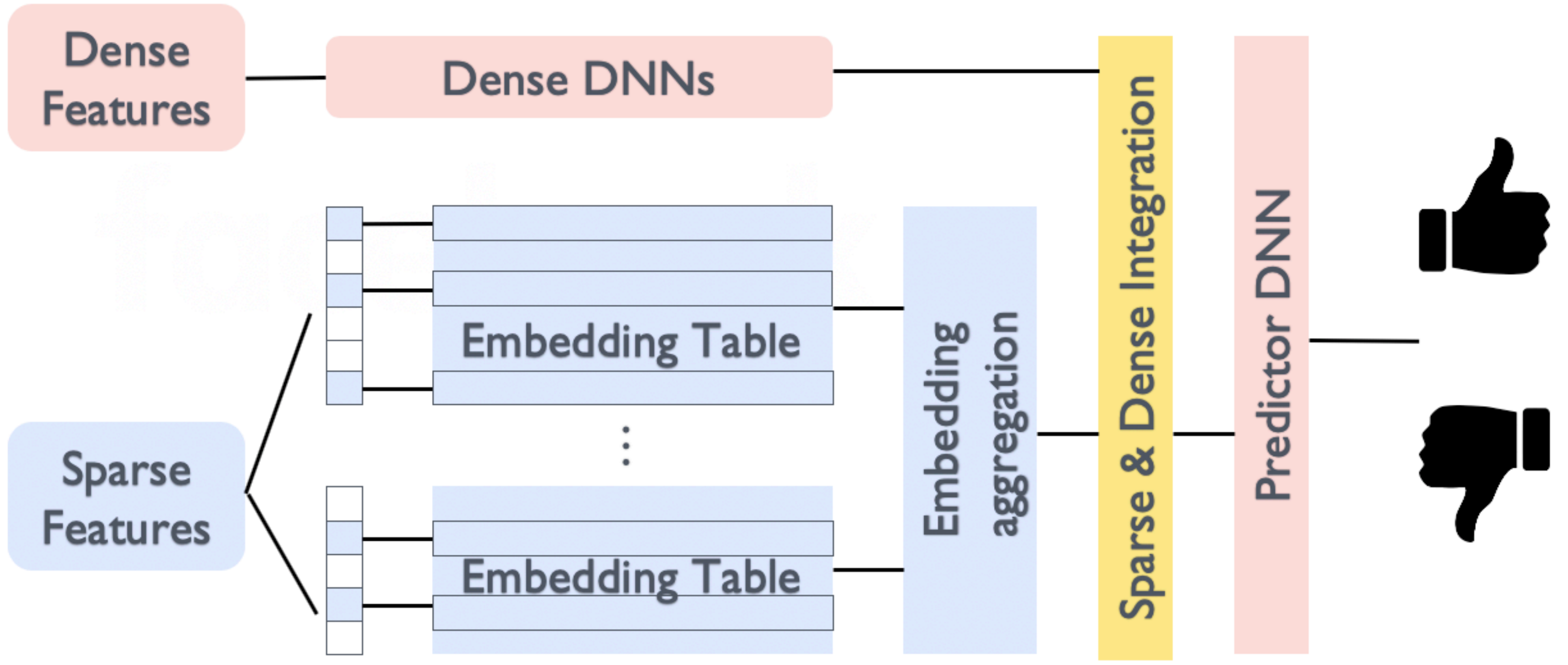


Performance characteristics of end-to-end models varies

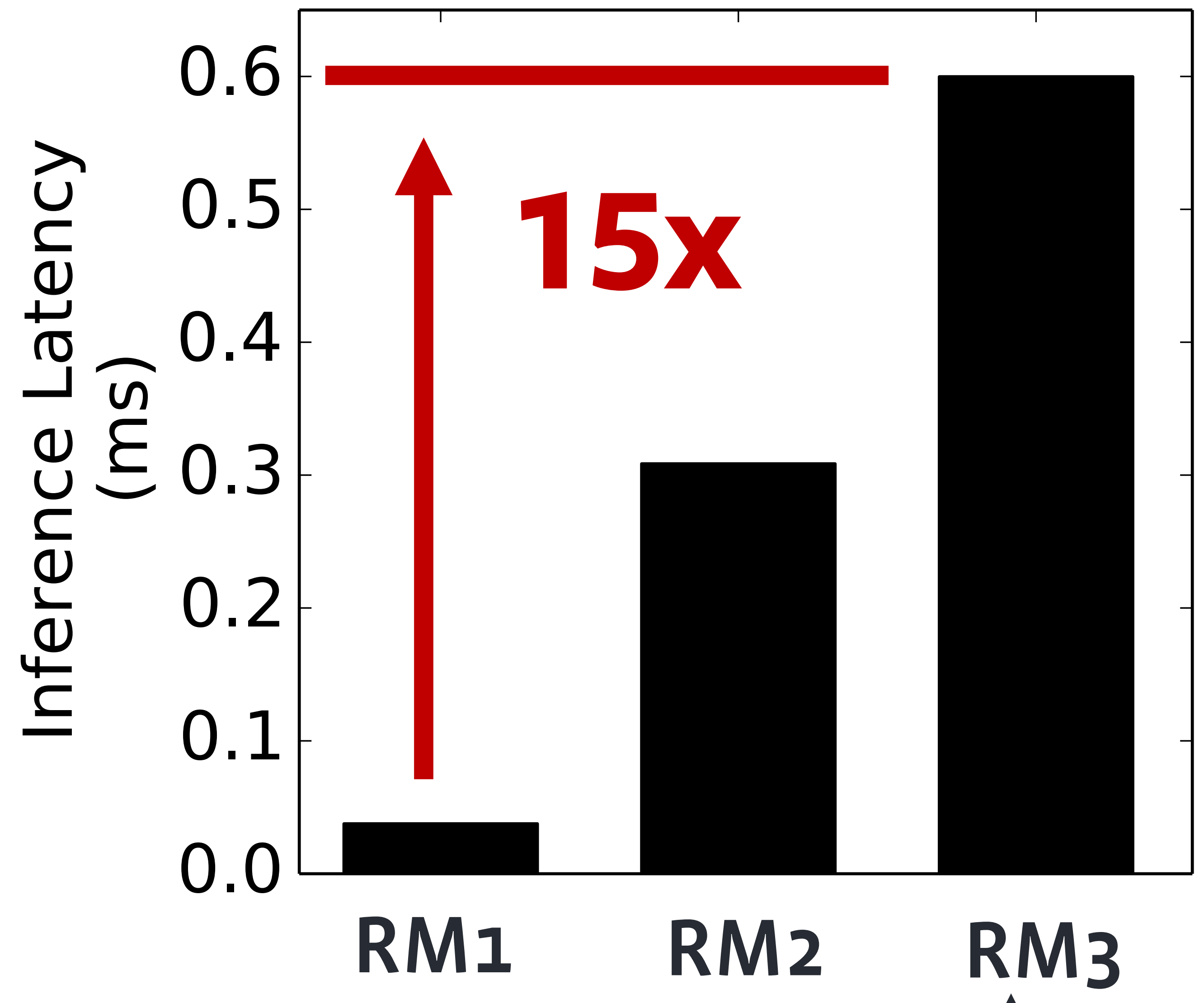


Early-stage Filtering

Late-stage Ranking

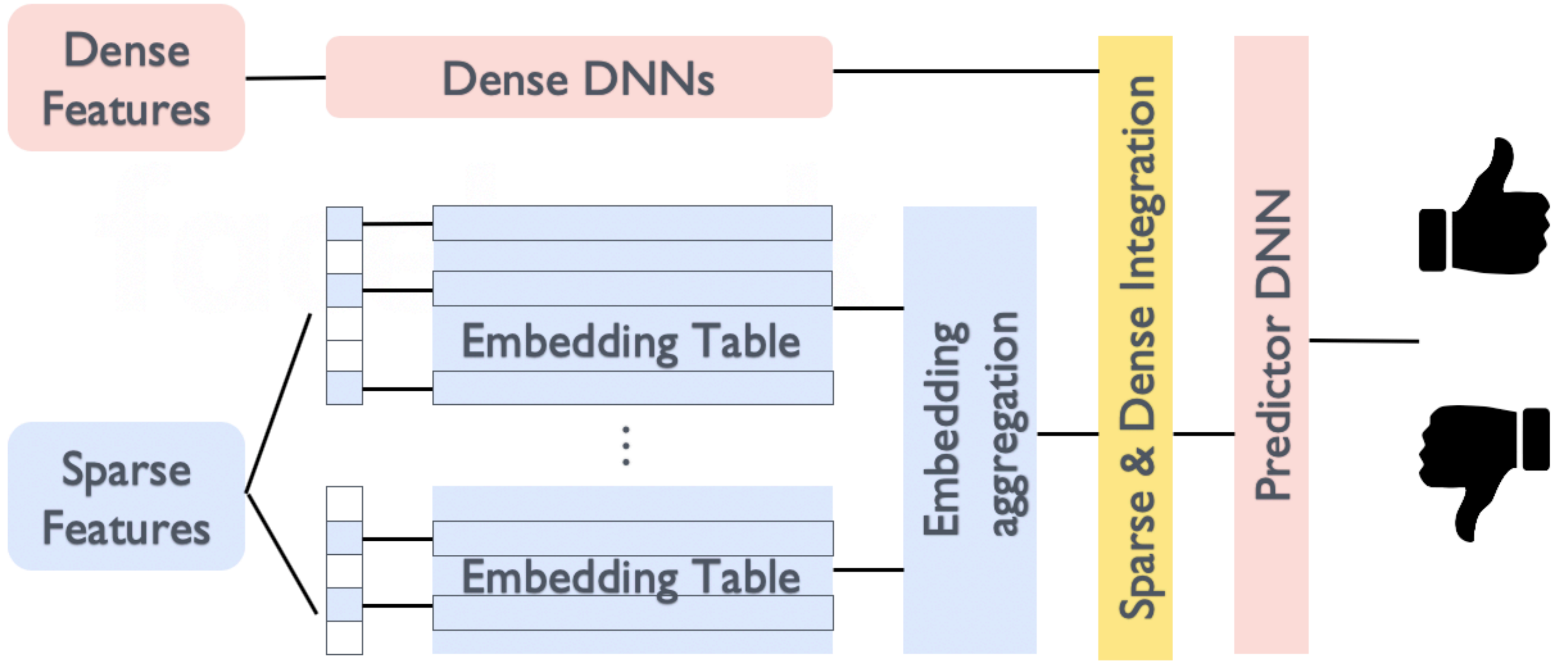


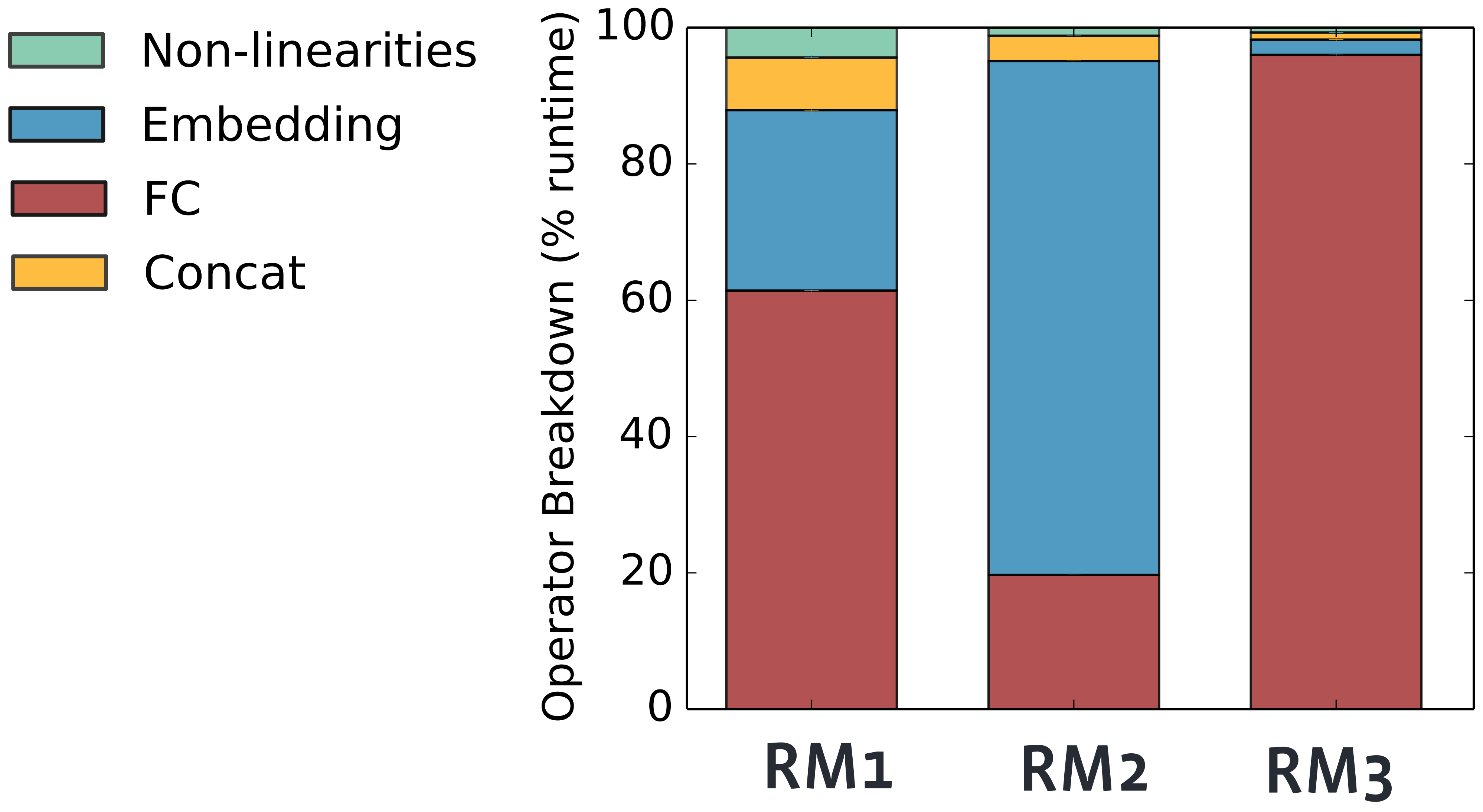
Performance characteristics of end-to-end models varies



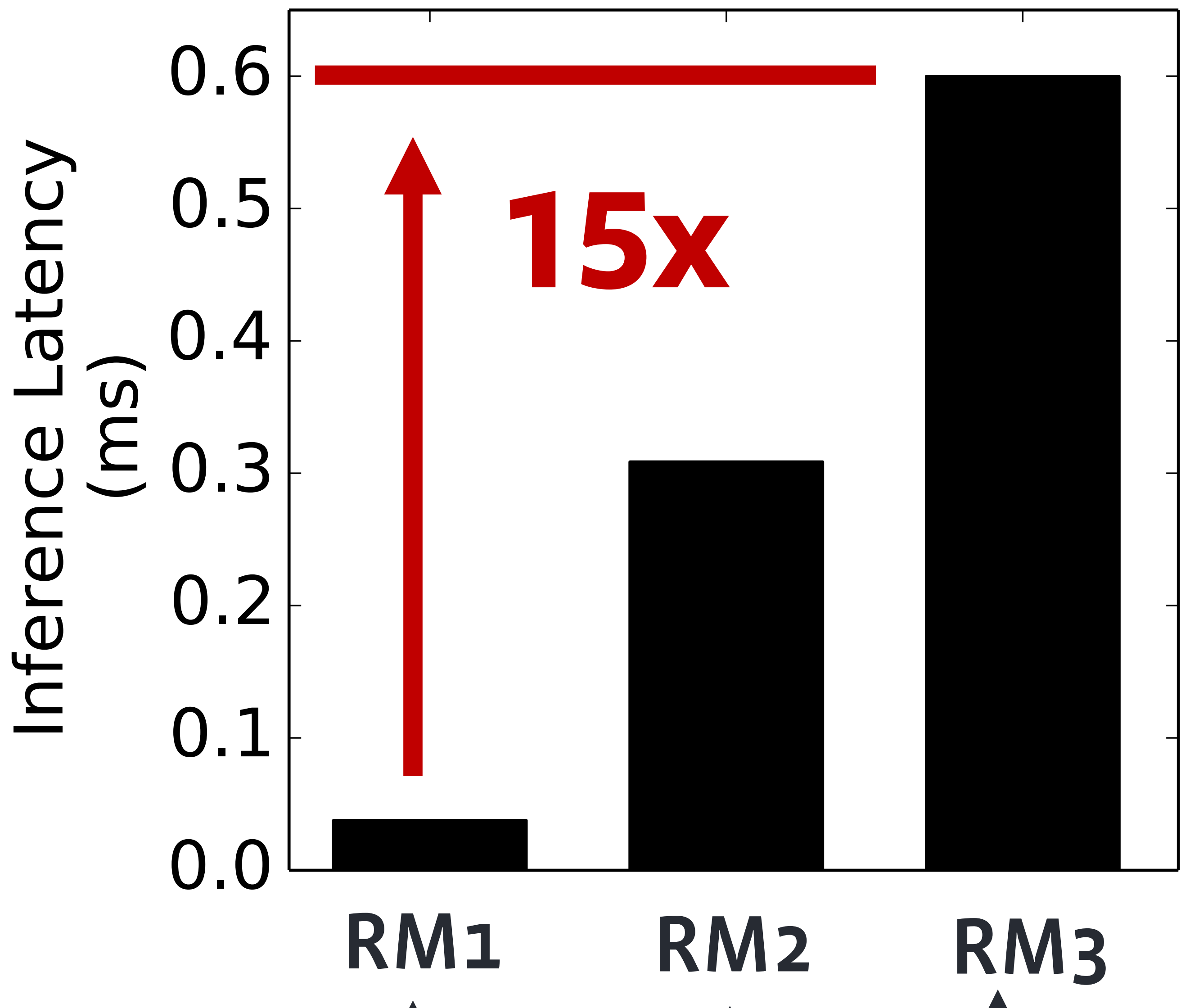
Early-stage Filtering

Late-stage Ranking



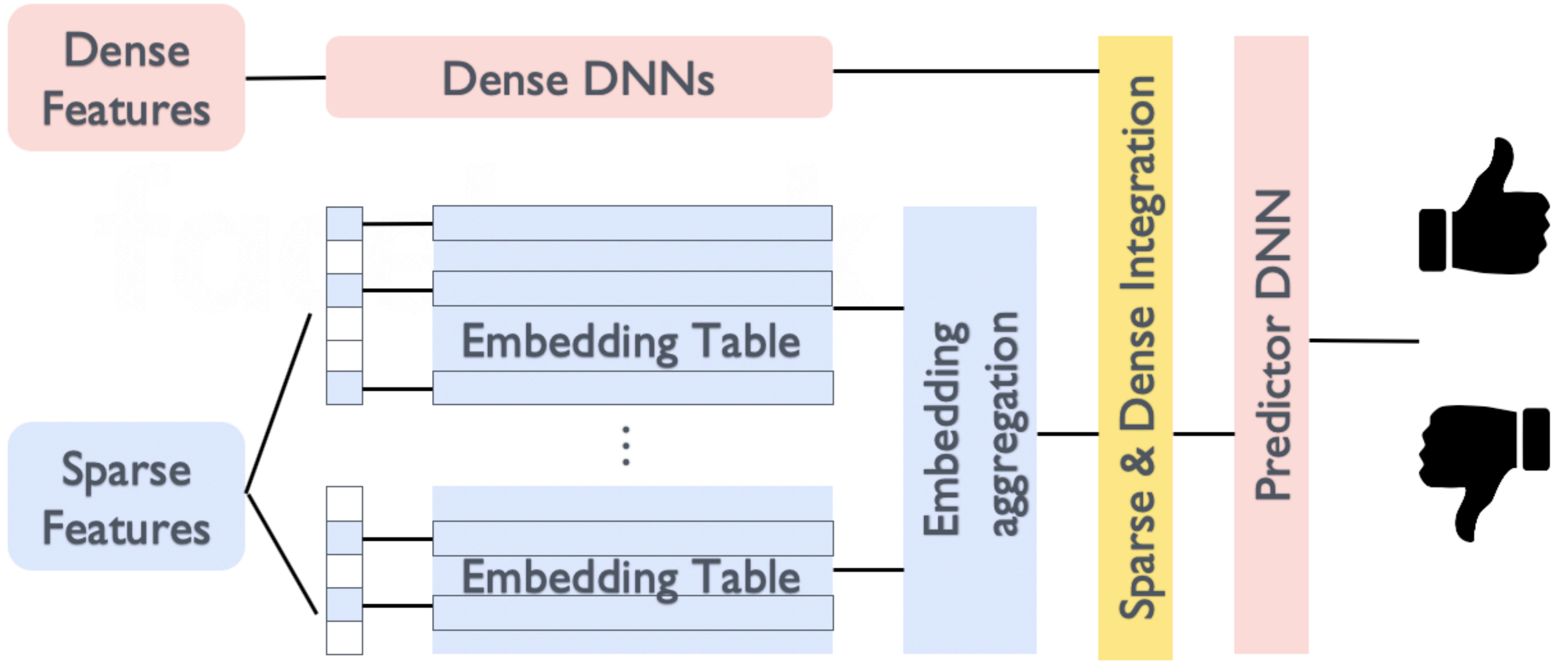


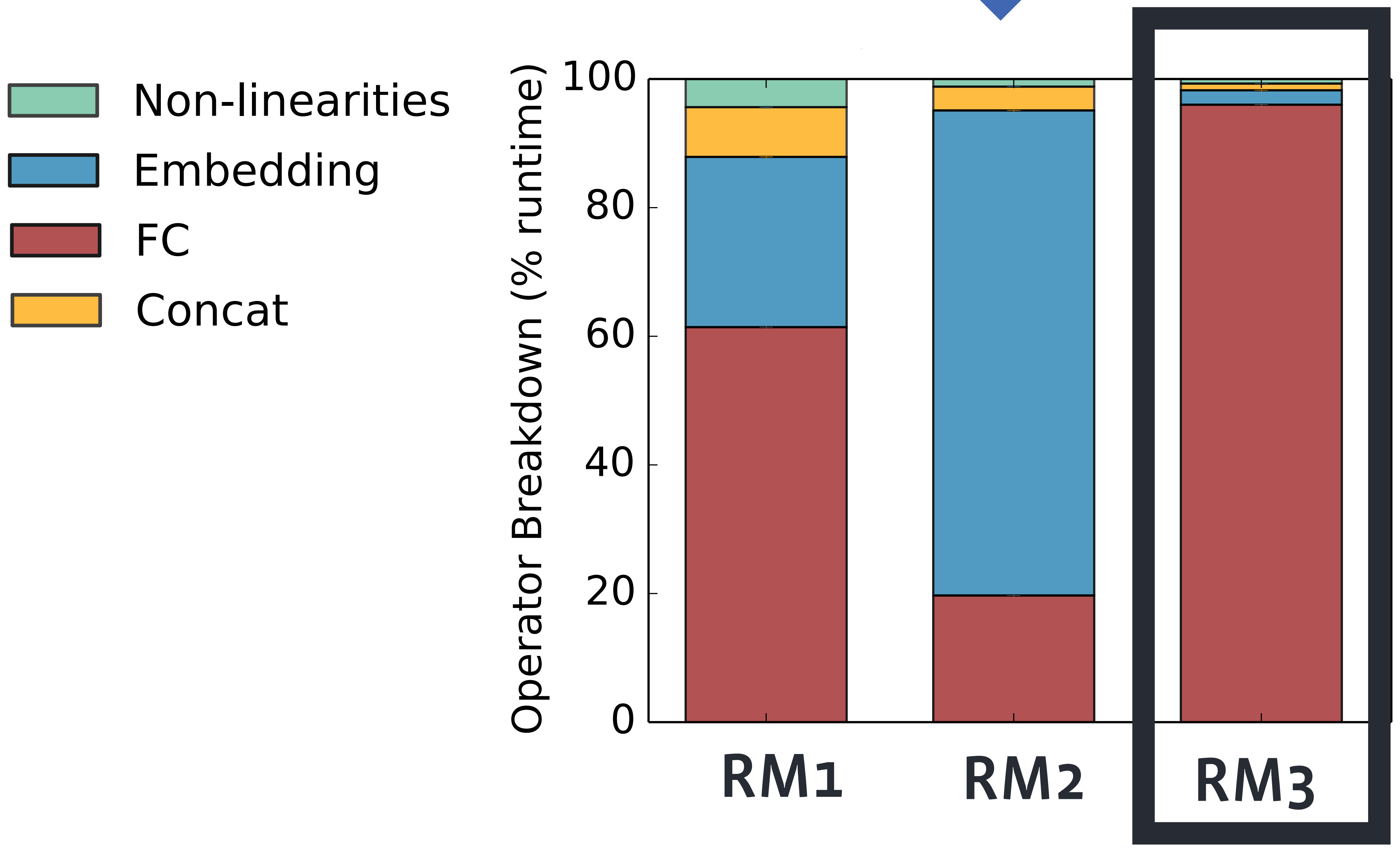
Performance characteristics of end-to-end models varies



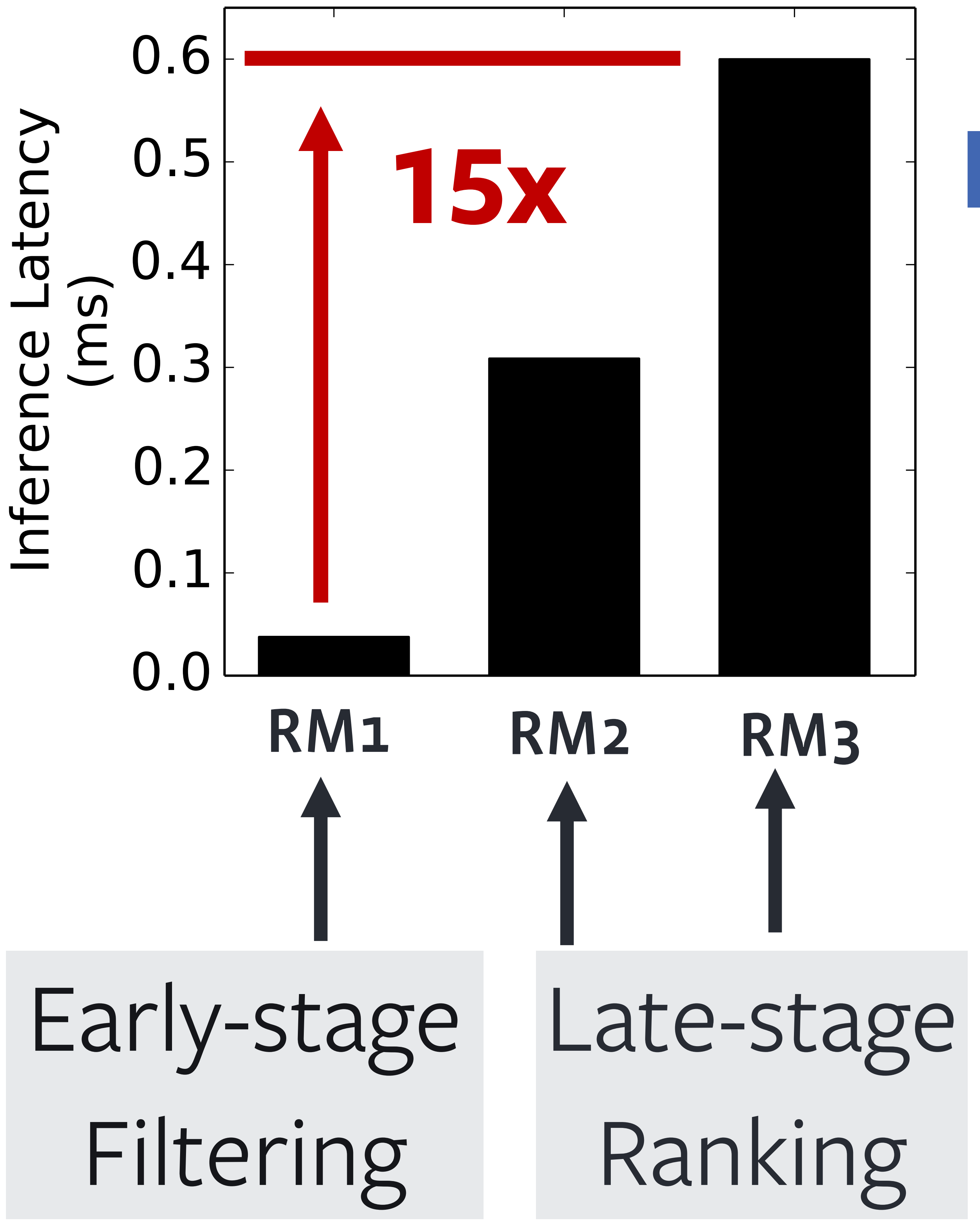
Early-stage Filtering

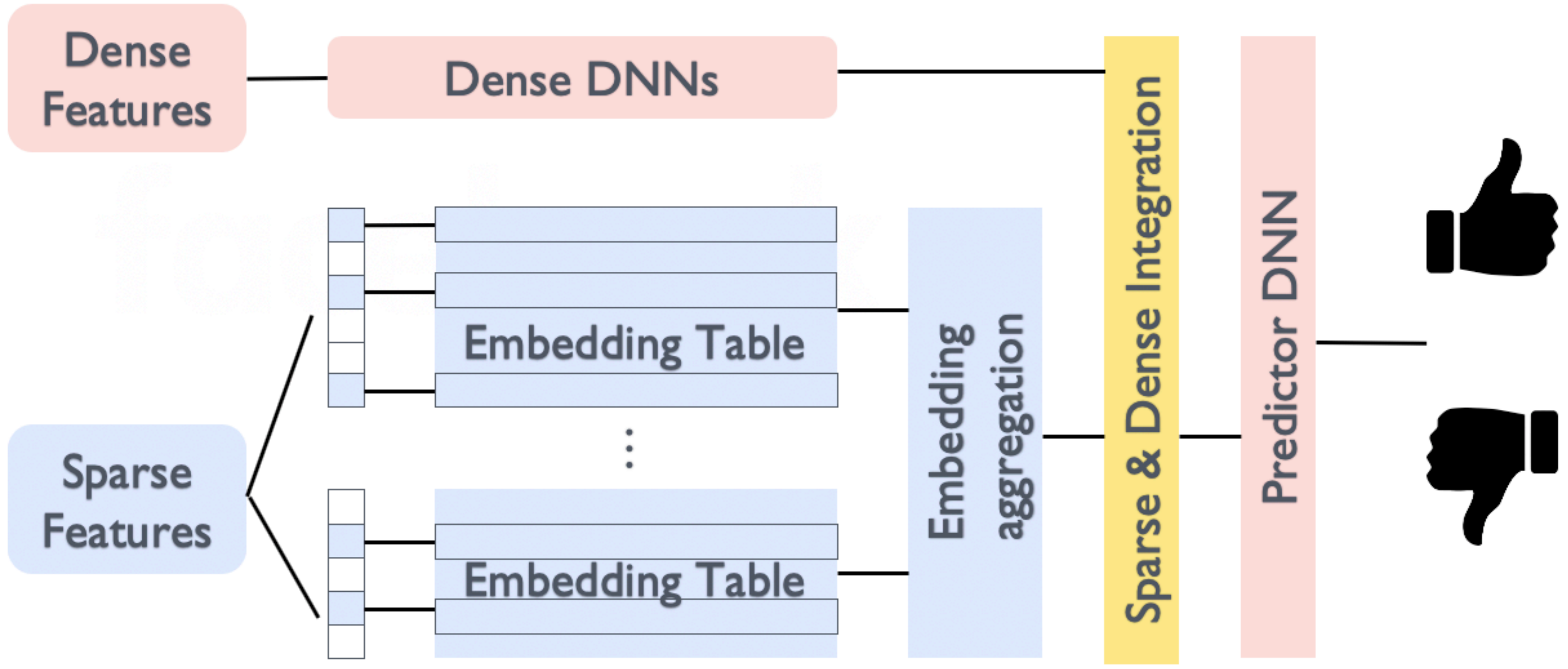
Late-stage Ranking

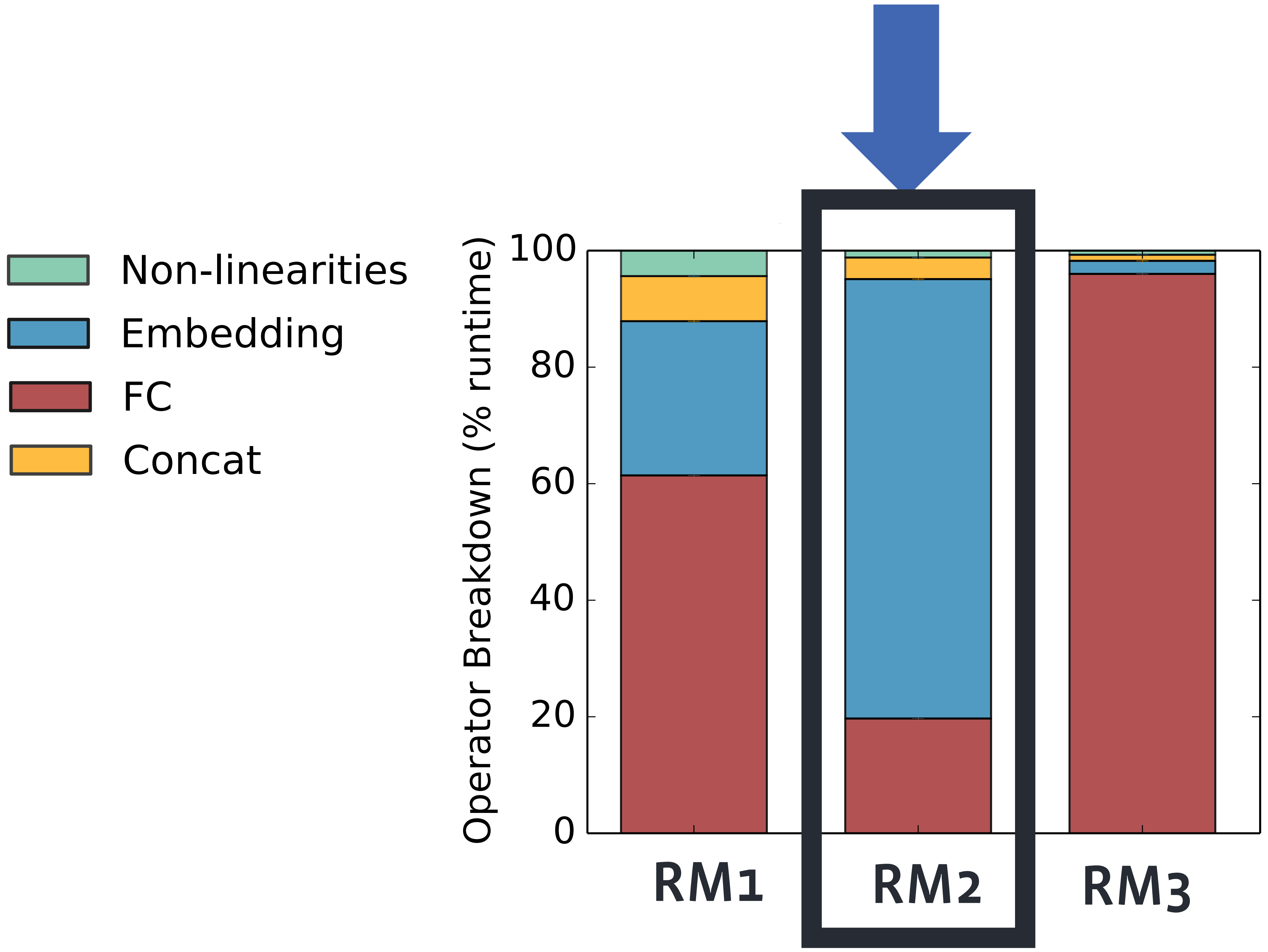




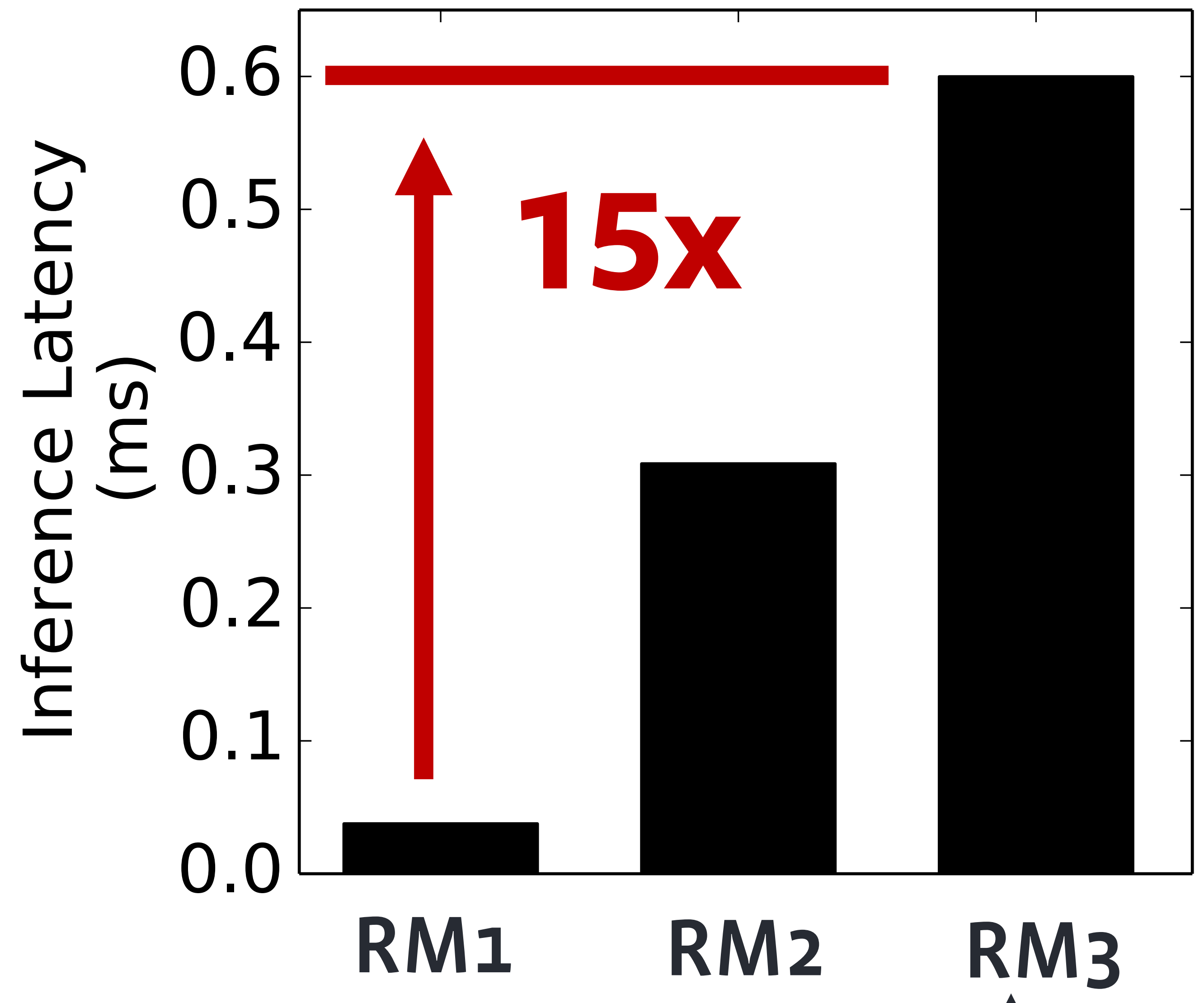
Performance characteristics of end-to-end models varies





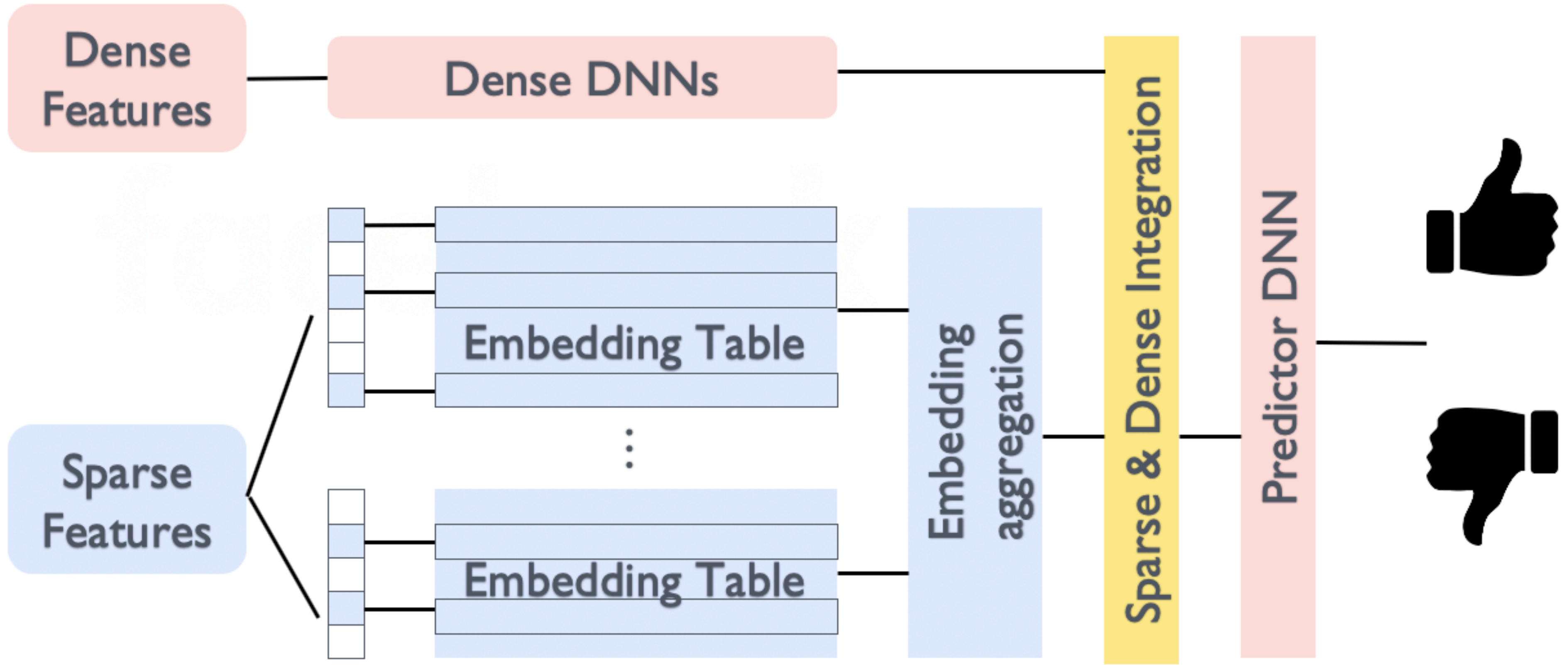


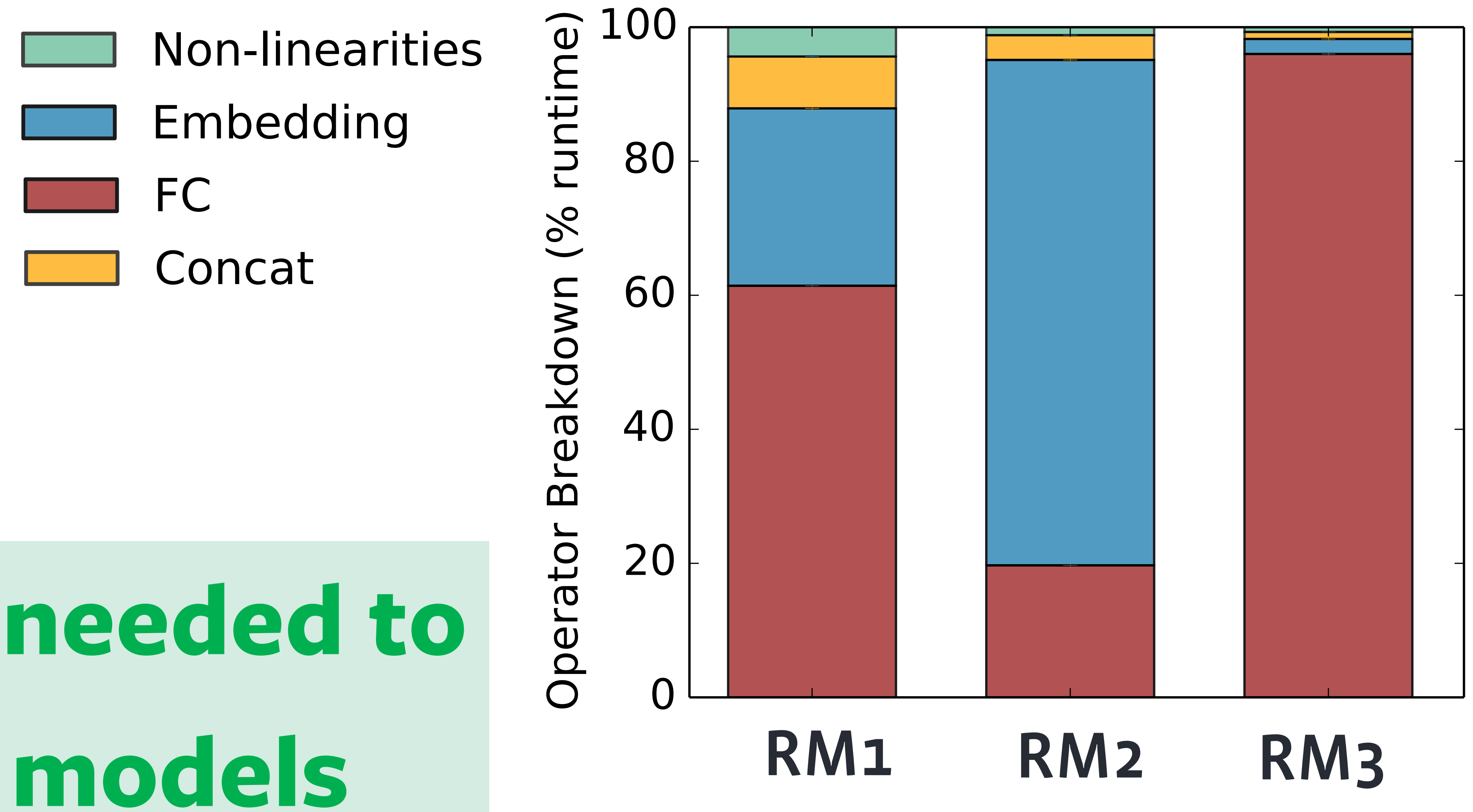
Performance characteristics of end-to-end models varies



Early-stage Filtering

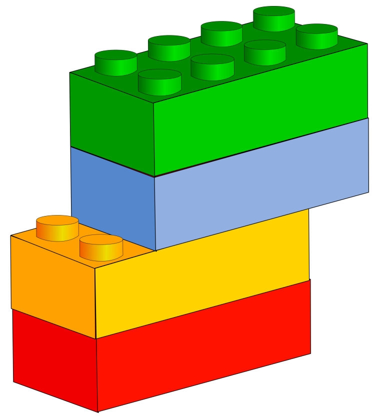
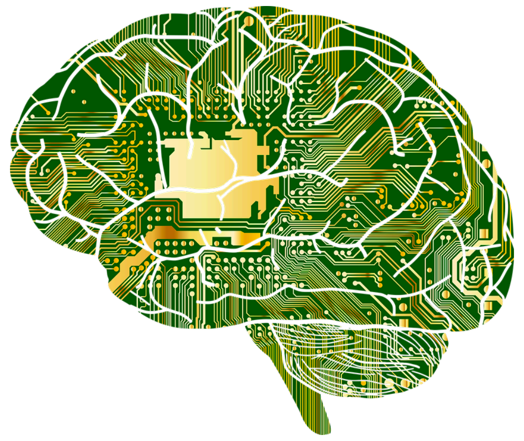
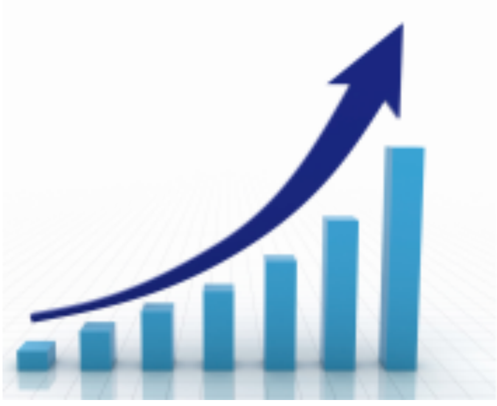
Late-stage Ranking





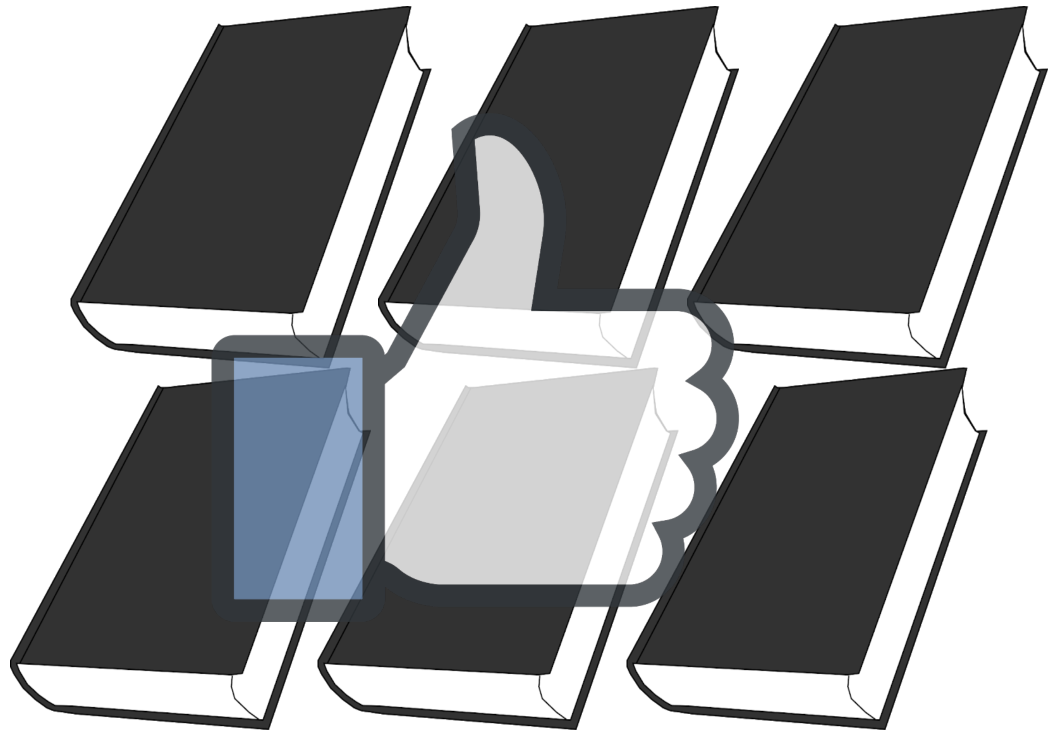
Diverse system solutions are needed to optimize recommendation models

Hardware insights of recommendation

Algorithm	Hardware insights and opportunities
 General model structure	Optimize operators with new storage, compute, and memory access patterns
 Diverse networks architectures	Accelerate recommendation with flexible and diverse system solutions
 At-scale inference	Exploit hardware heterogeneity and parallelism to optimize latency-bounded throughput

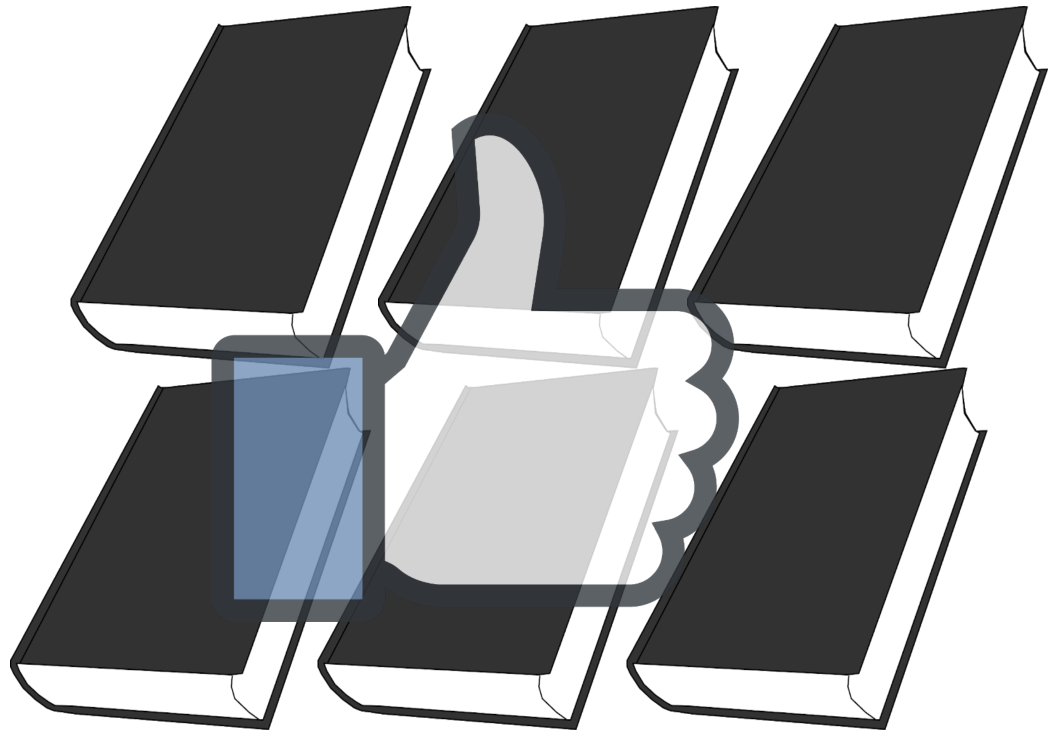
Ranking more items improves recommendation quality

High throughput!



Ranking more items improves recommendation quality

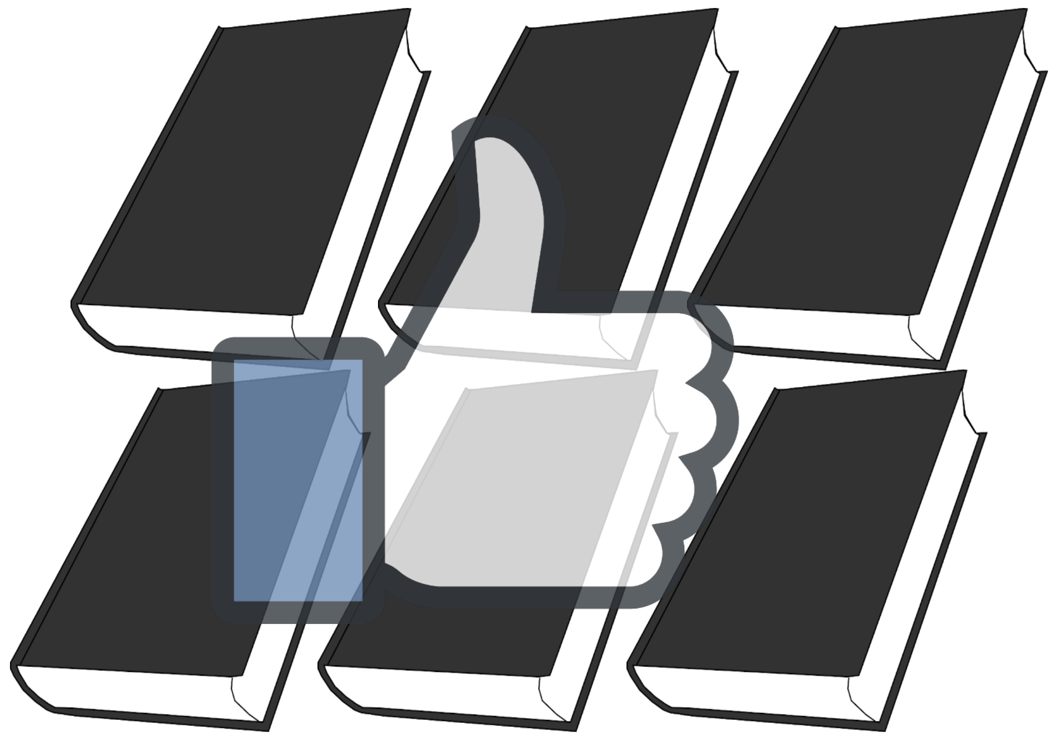
High throughput!



Low latency!

Ranking more items improves recommendation quality

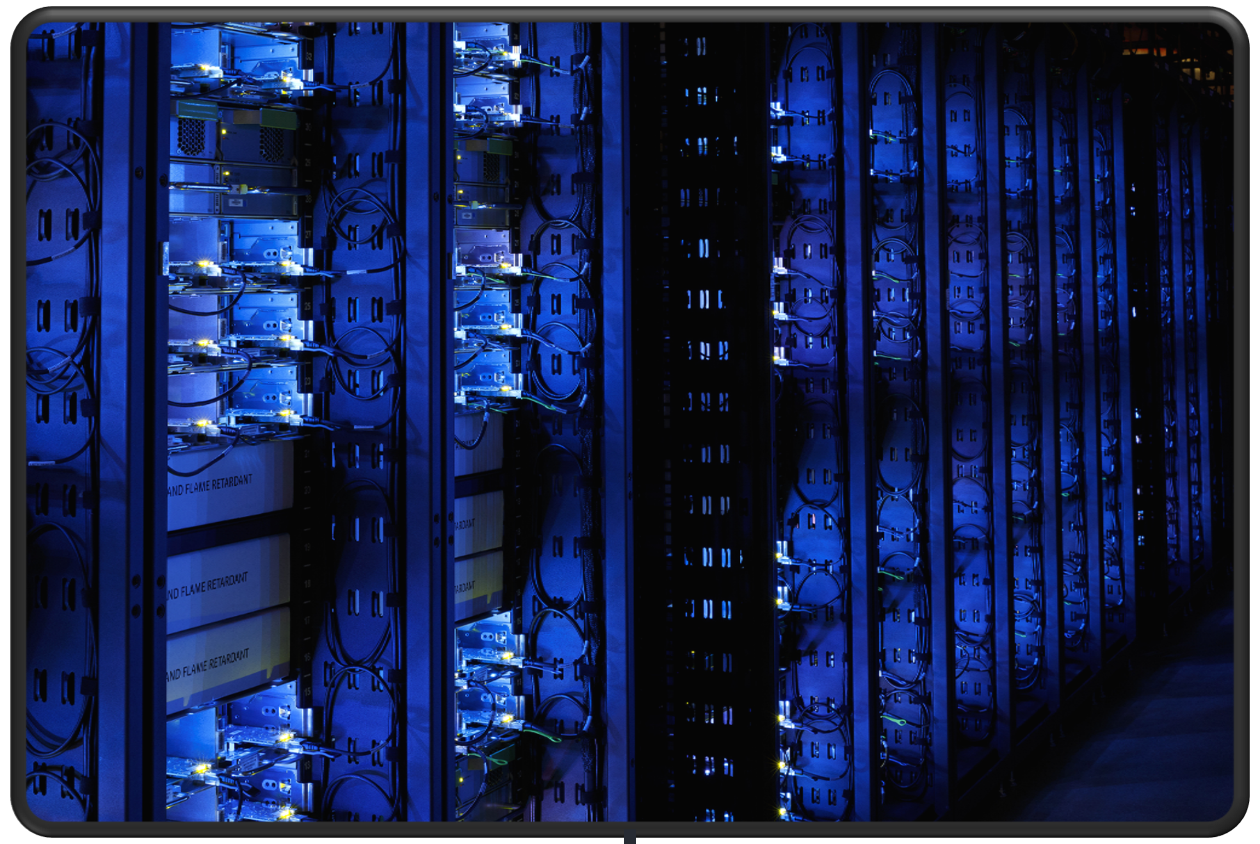
High throughput!

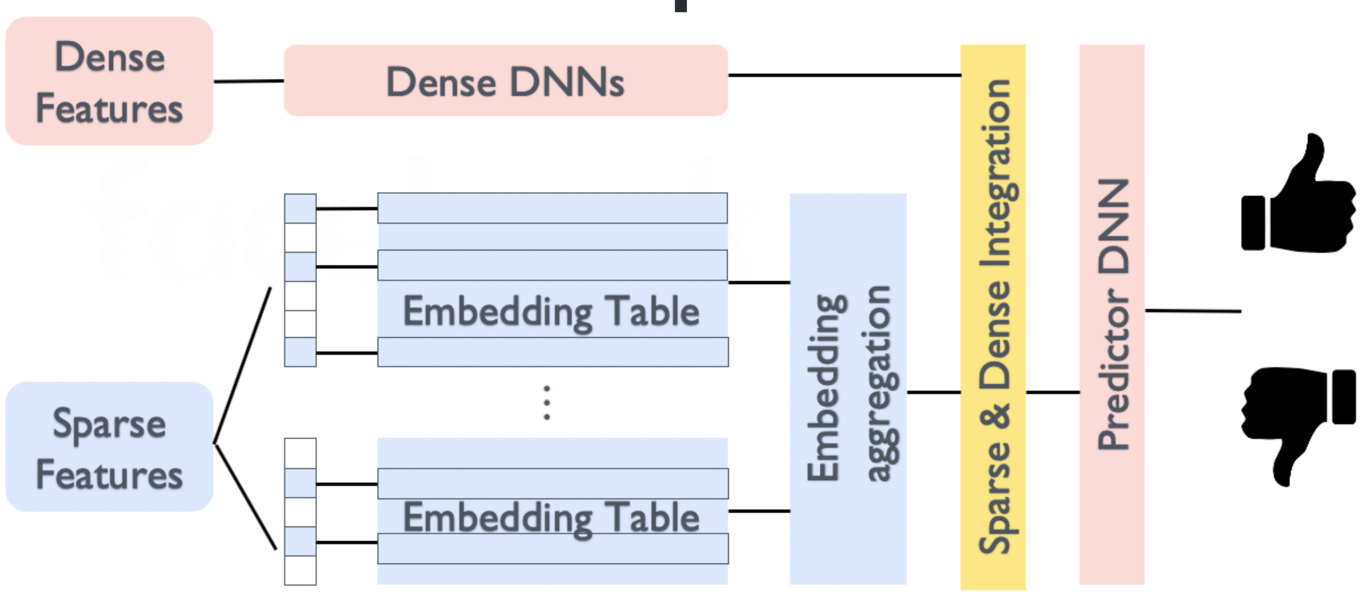


Low latency!

Optimize latency-bounded throughput

Characterizing latency bounded throughput design space





Models

Characterizing latency bounded throughput design space

2.5 GHz
AVX-2
Inclusive L2/L3 caches
DDR3

Haswell

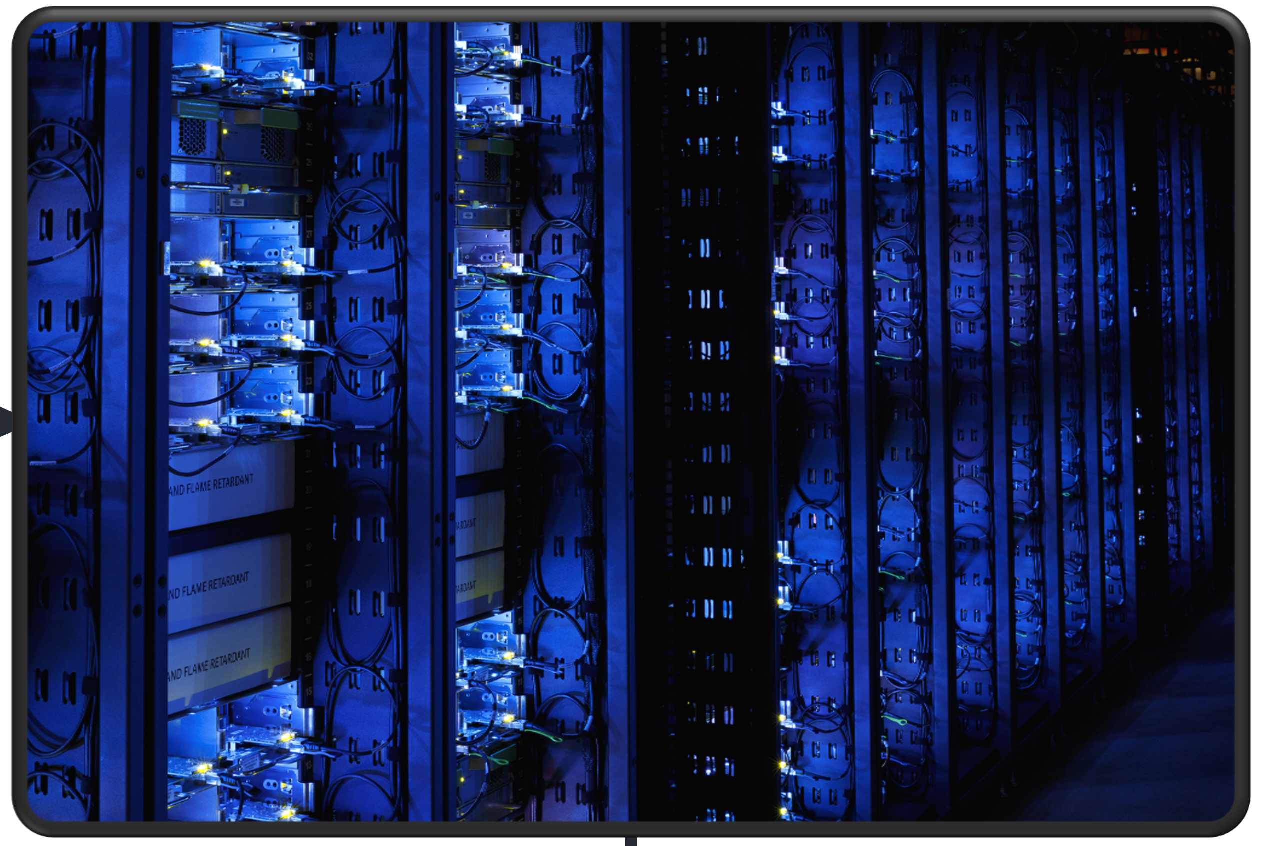
2.4 GHz
AVX-2
Inclusive L2/L3 caches
DDR4

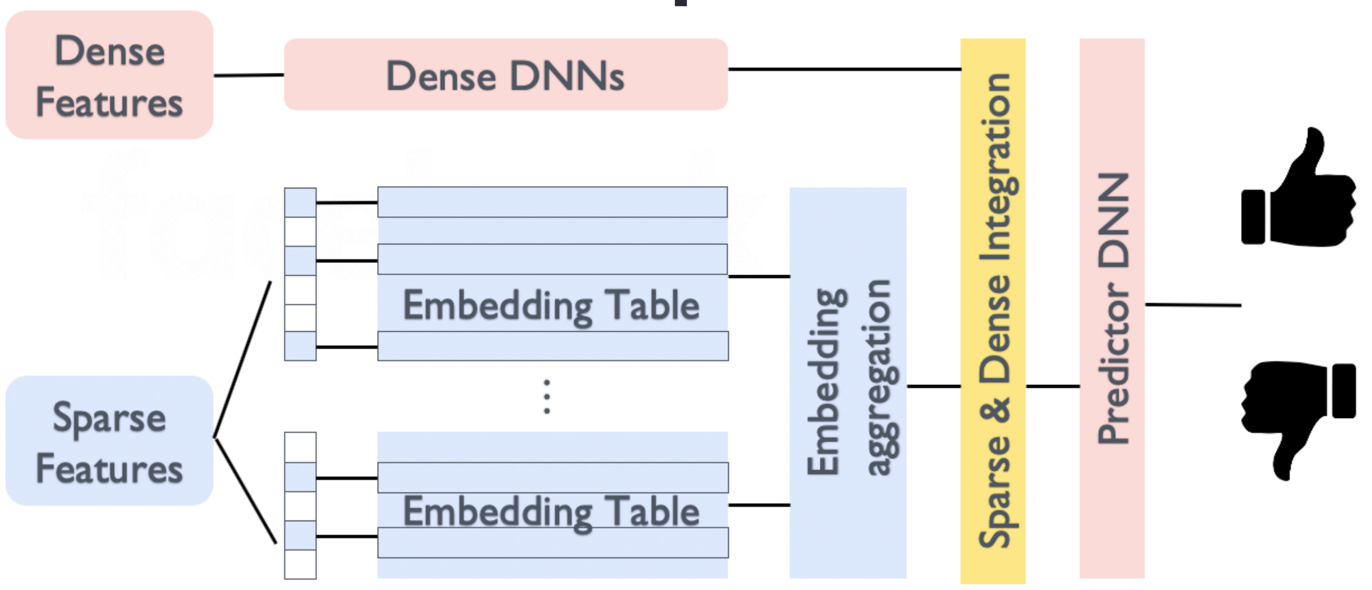
Broadwell

2.0 GHz
AVX-512
Exclusive L2/L3 caches
DDR4

Skylake

Hardware





Models

Characterizing latency bounded throughput design space

2.5 GHz
AVX-2
Inclusive L2/L3 caches
DDR3

Haswell

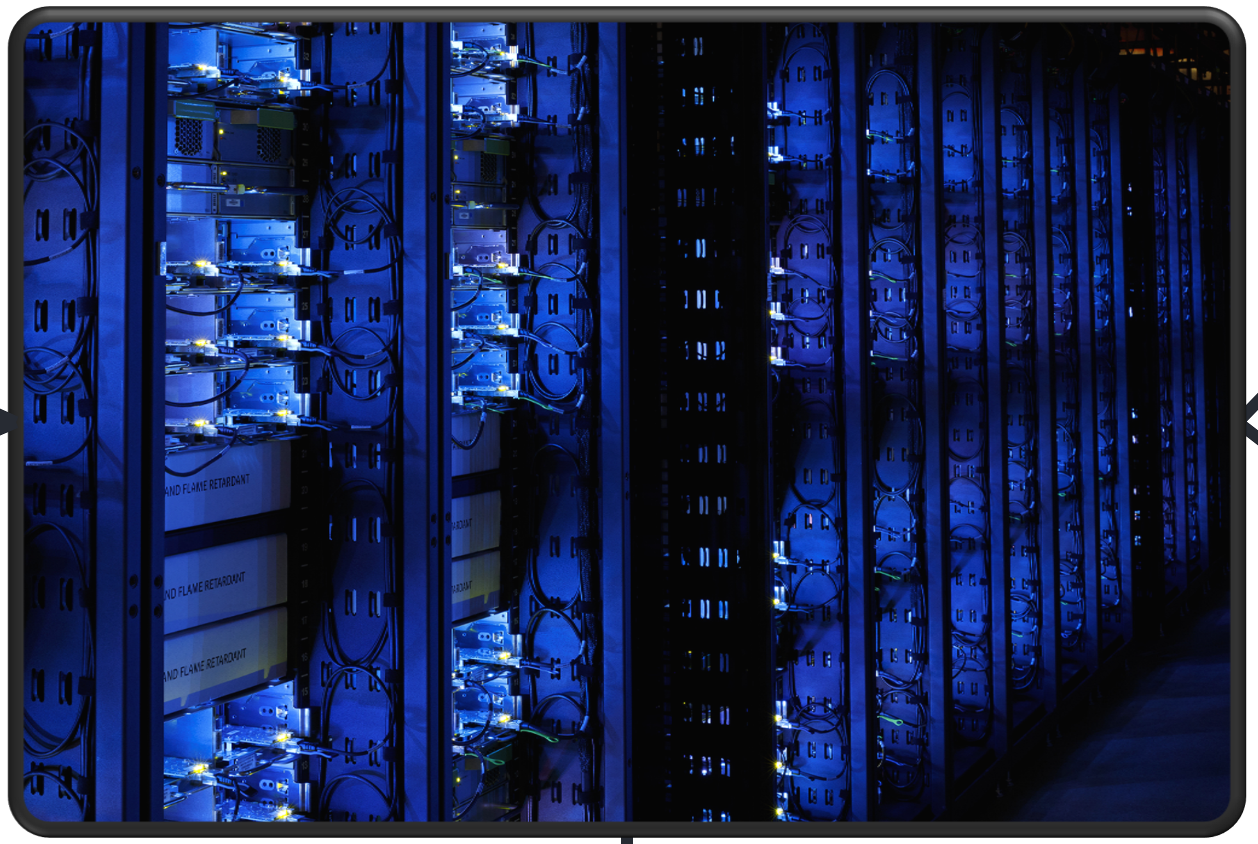
2.4 GHz
AVX-2
Inclusive L2/L3 caches
DDR4

Broadwell

2.0 GHz
AVX-512
Exclusive L2/L3 caches
DDR4

Skylake

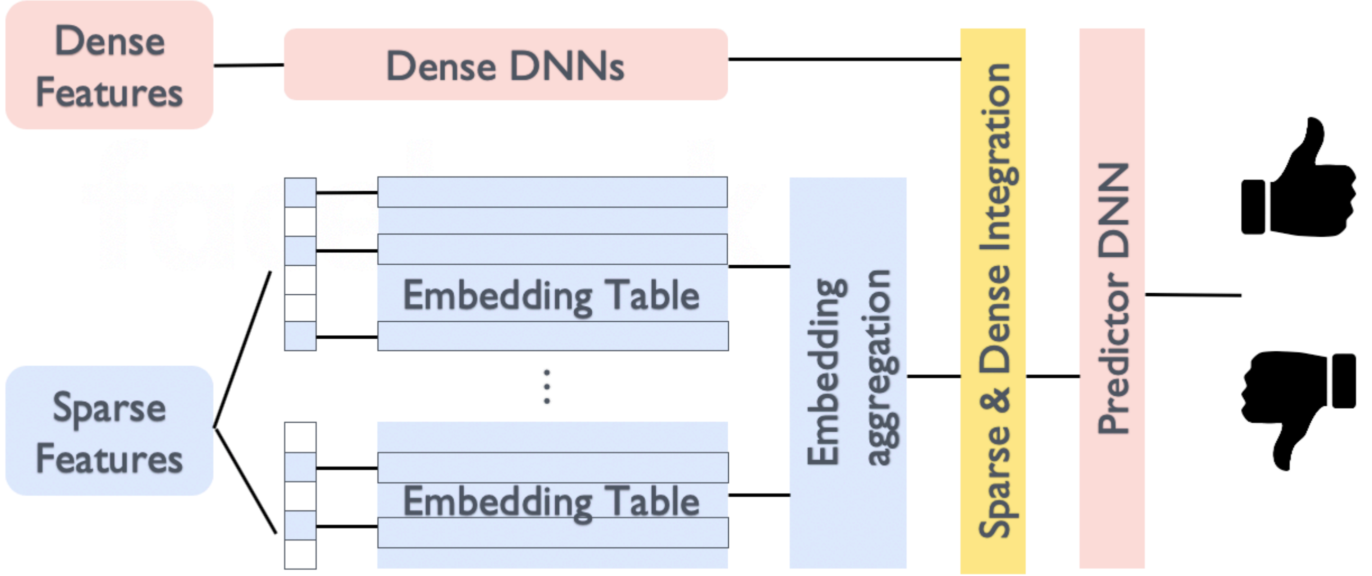
Hardware



Data level parallelism
(i.e., batch-size)

Task level parallelism
(i.e., co-locating models)

Parallelization



Models

Characterizing latency bounded throughput design space

2.5 GHz
AVX-2
Inclusive L2/L3 caches
DDR3

Haswell

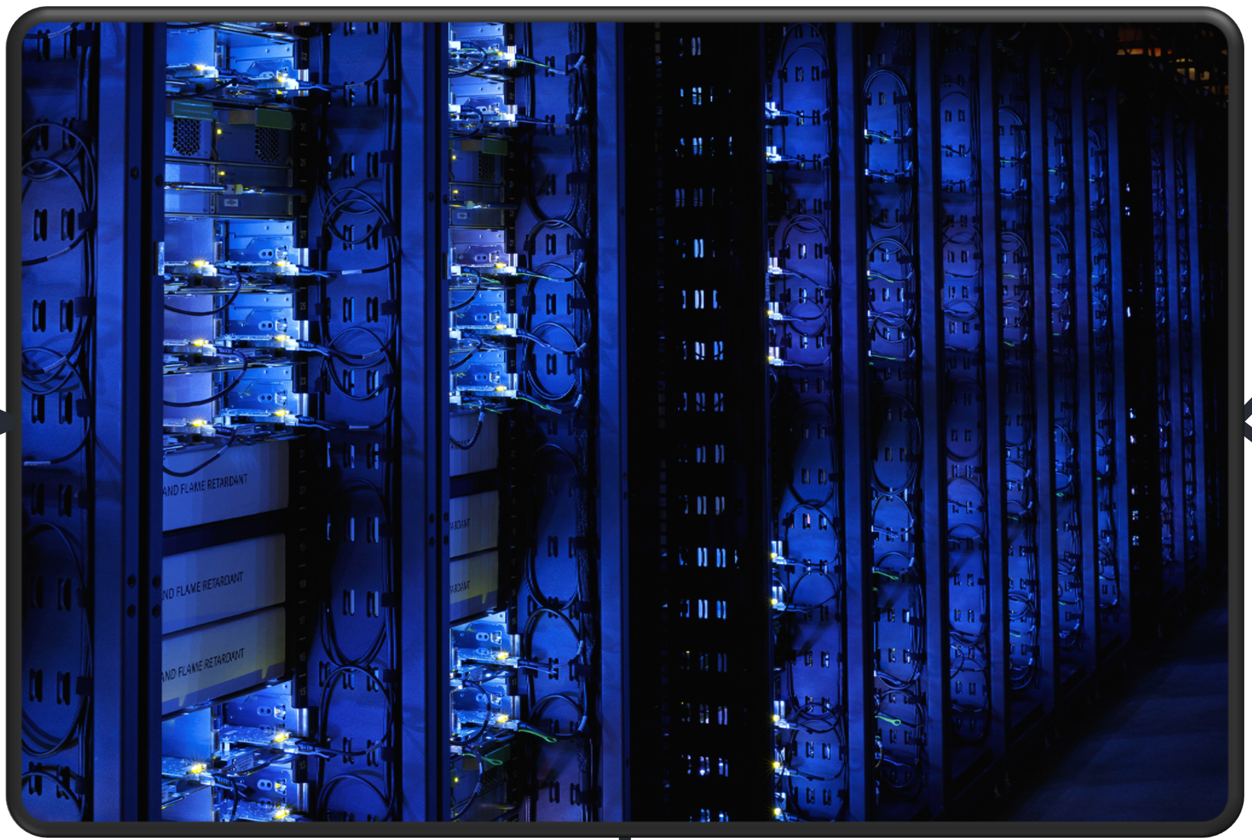
2.4 GHz
AVX-2
Inclusive L2/L3 caches
DDR4

Broadwell

2.0 GHz
AVX-512
Exclusive L2/L3 caches
DDR4

Skylake

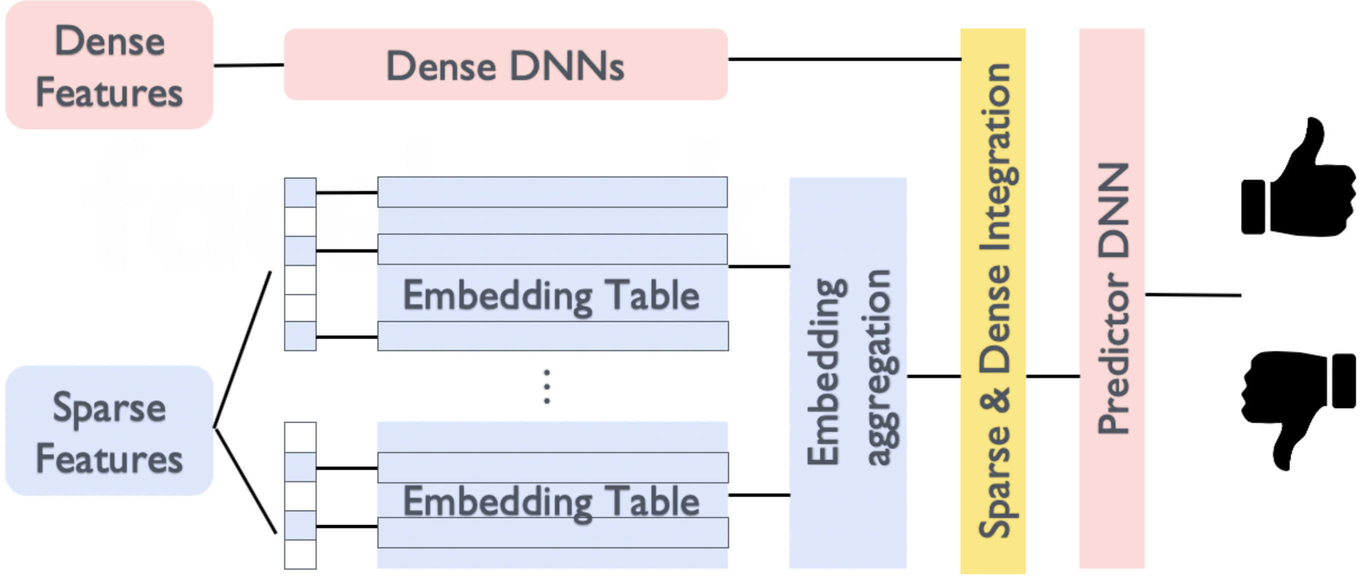
Hardware



Data level parallelism
(i.e., batch-size)

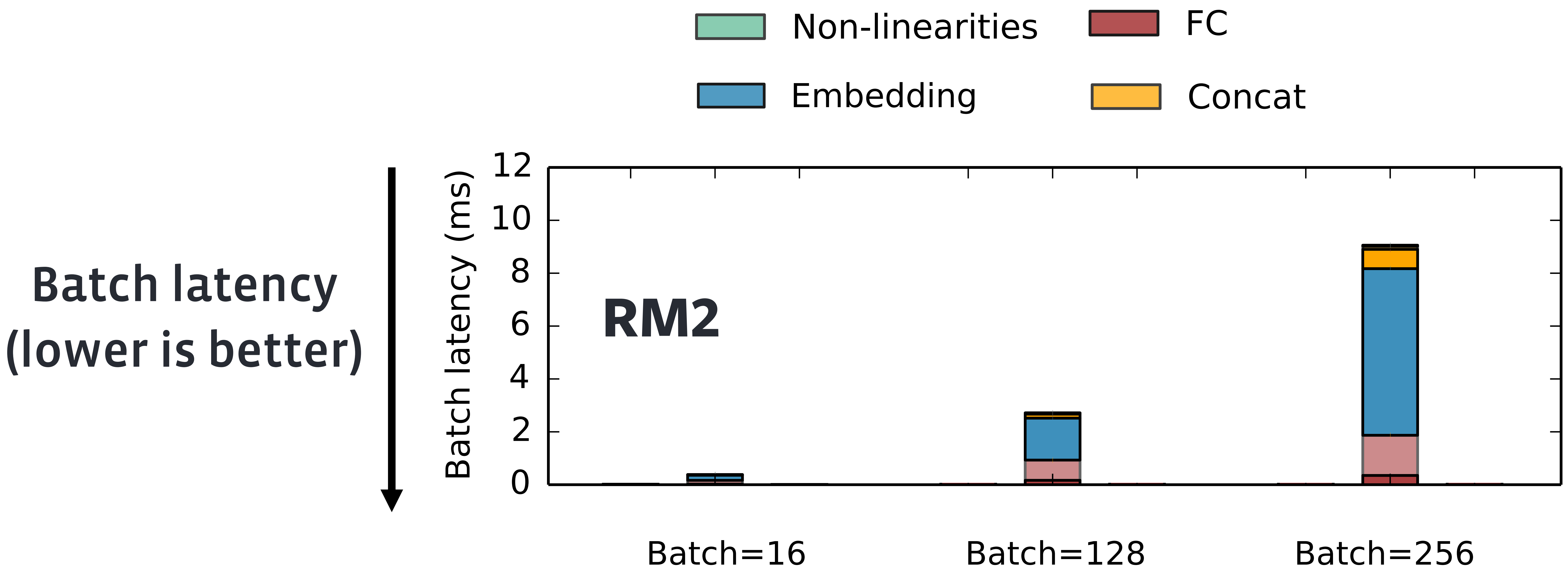
Task level parallelism
(i.e., co-locating models)

Parallelization



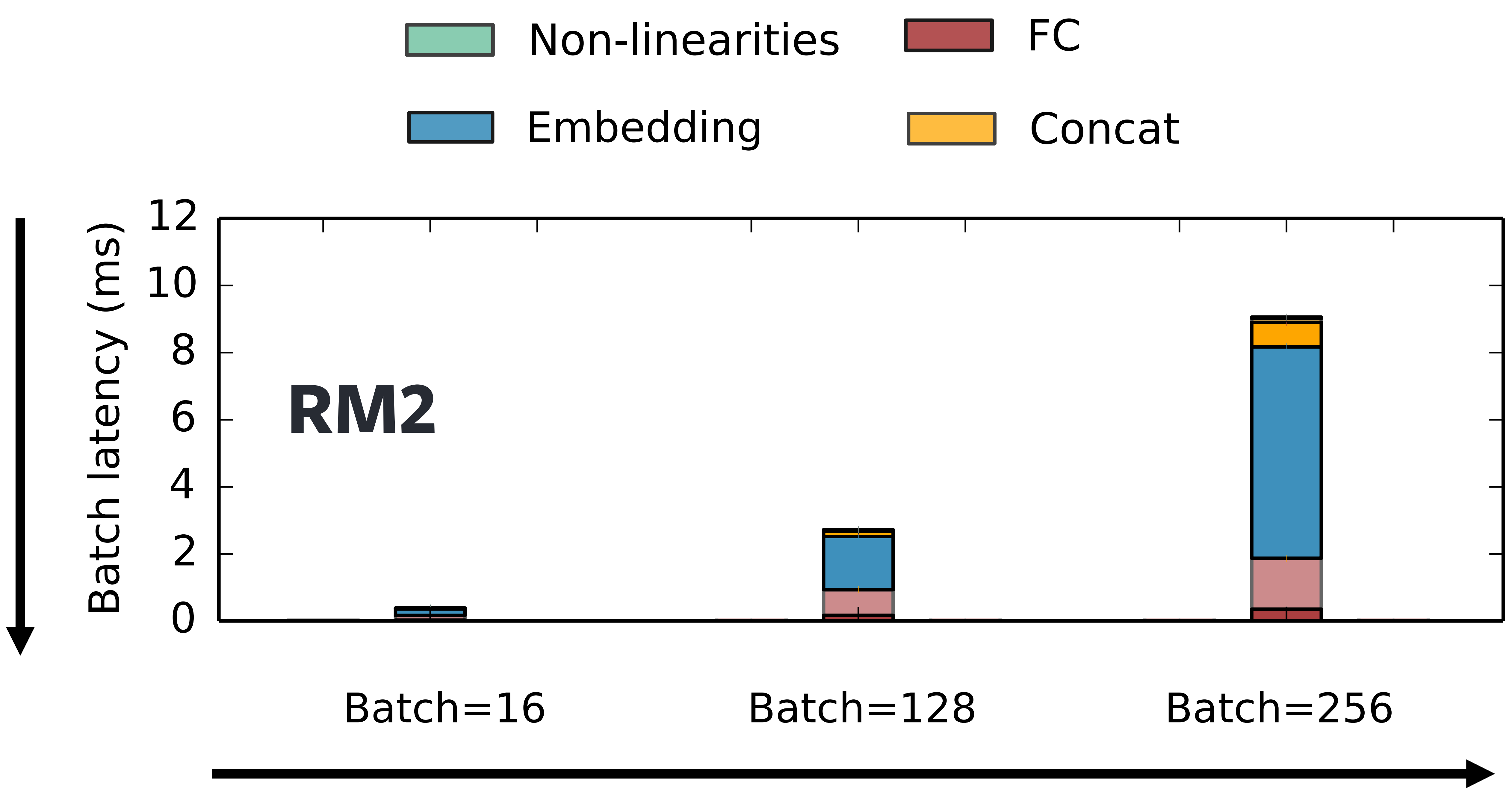
Models

Data parallelism: Characterizing latency bounded throughput design space



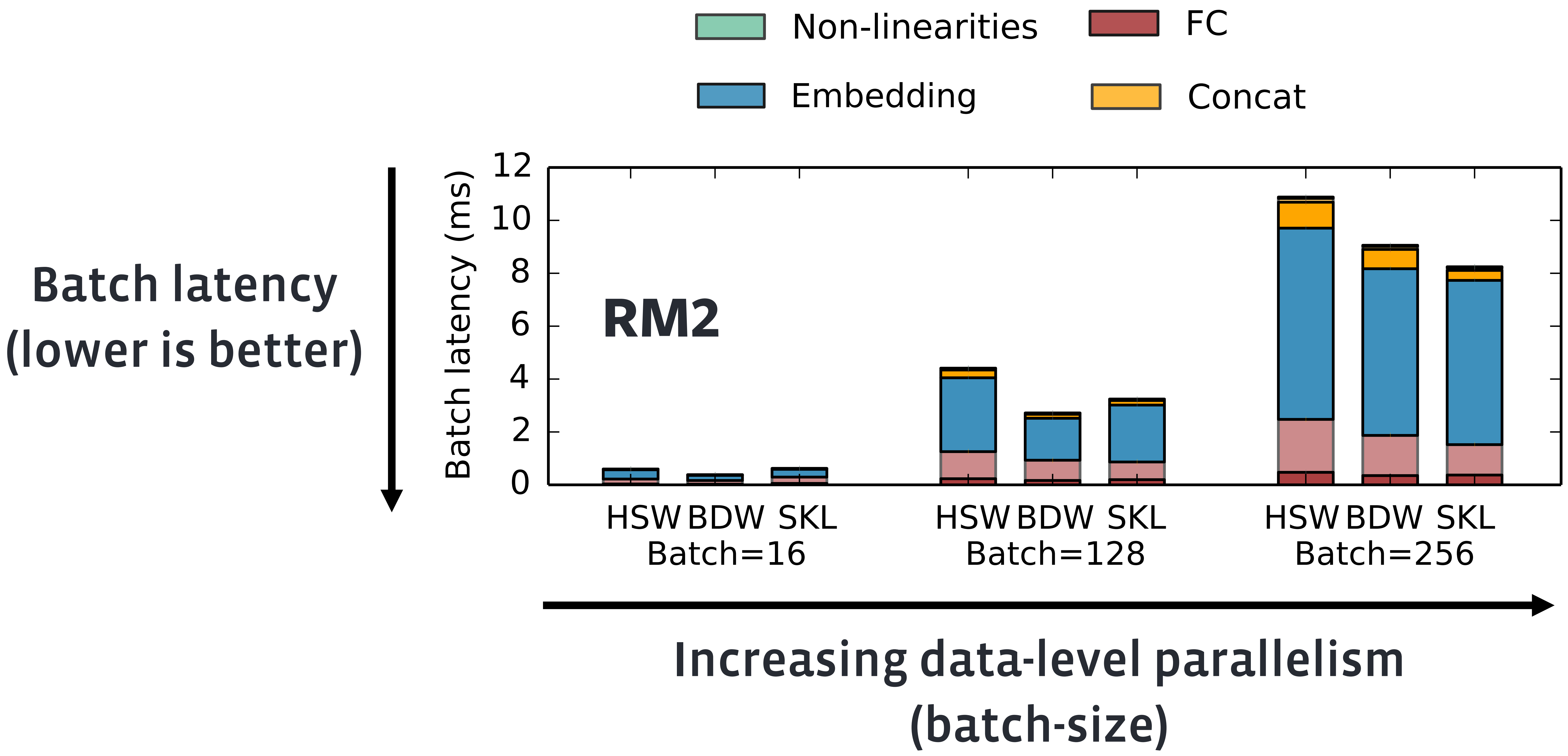
Data parallelism: Characterizing latency bounded throughput design space

Batch latency
(lower is better)

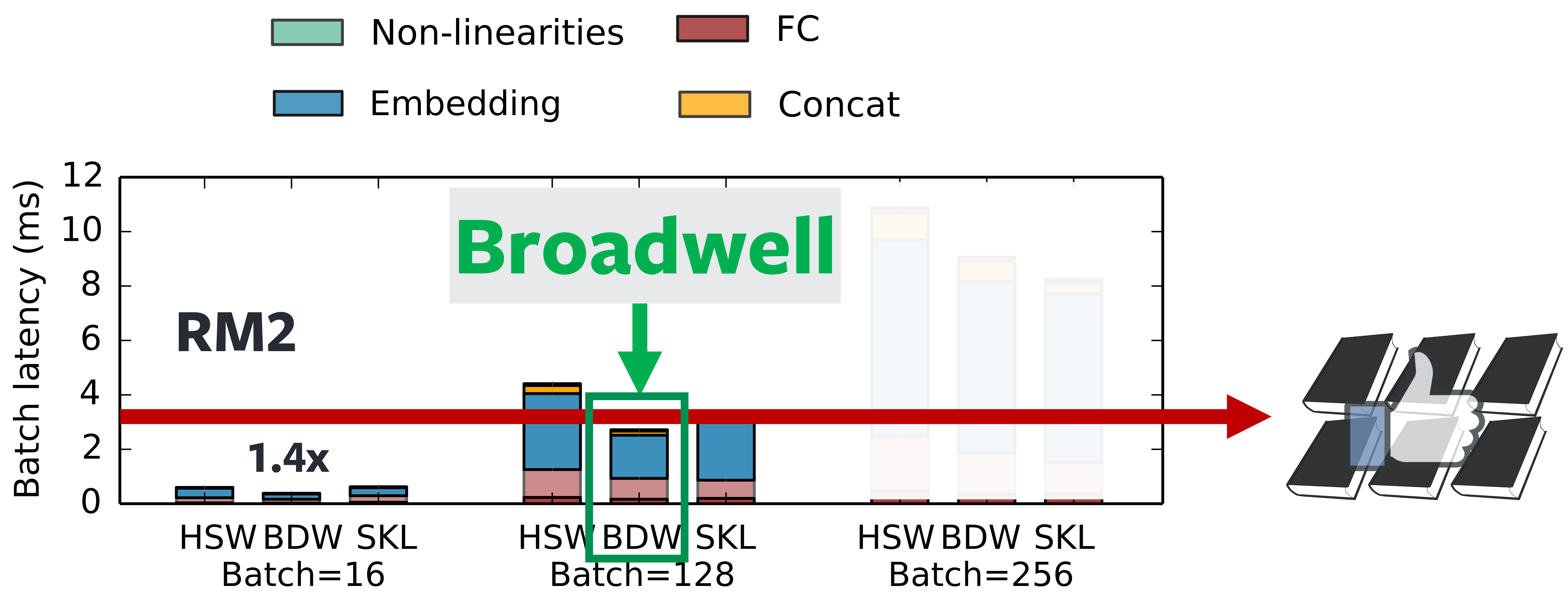


Increasing data-level parallelism
(batch-size)

Data parallelism: Characterizing latency bounded throughput design space

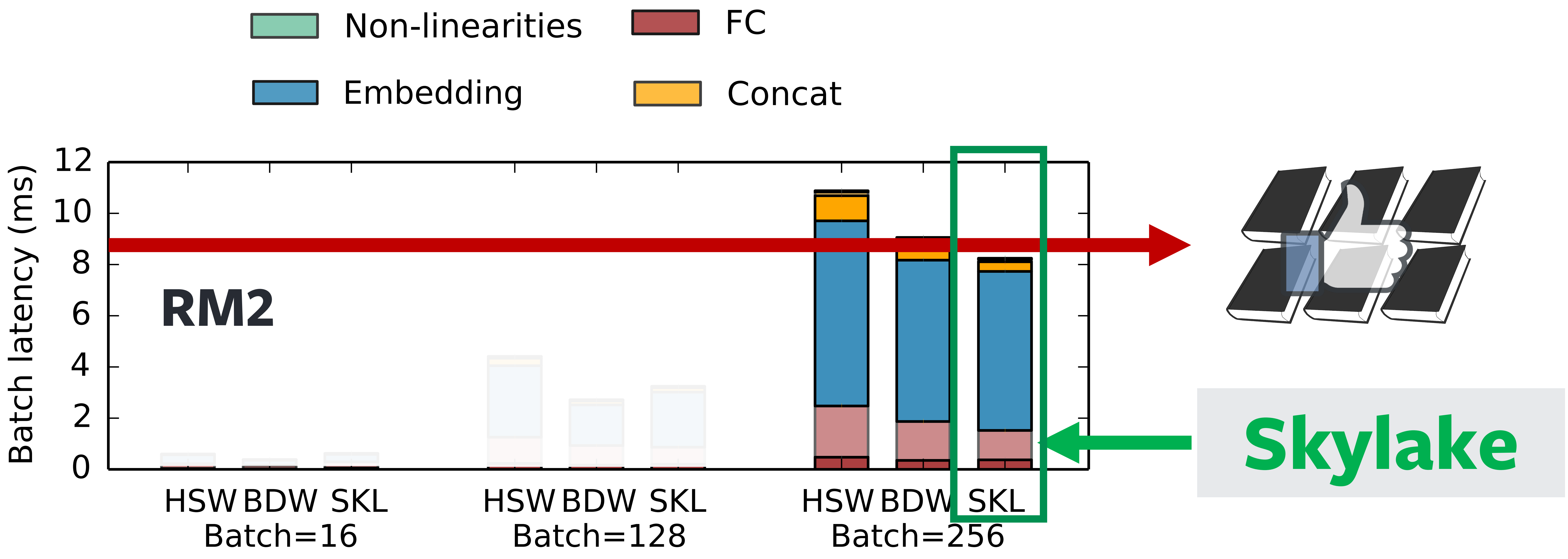


Data parallelism: Characterizing latency bounded throughput design space



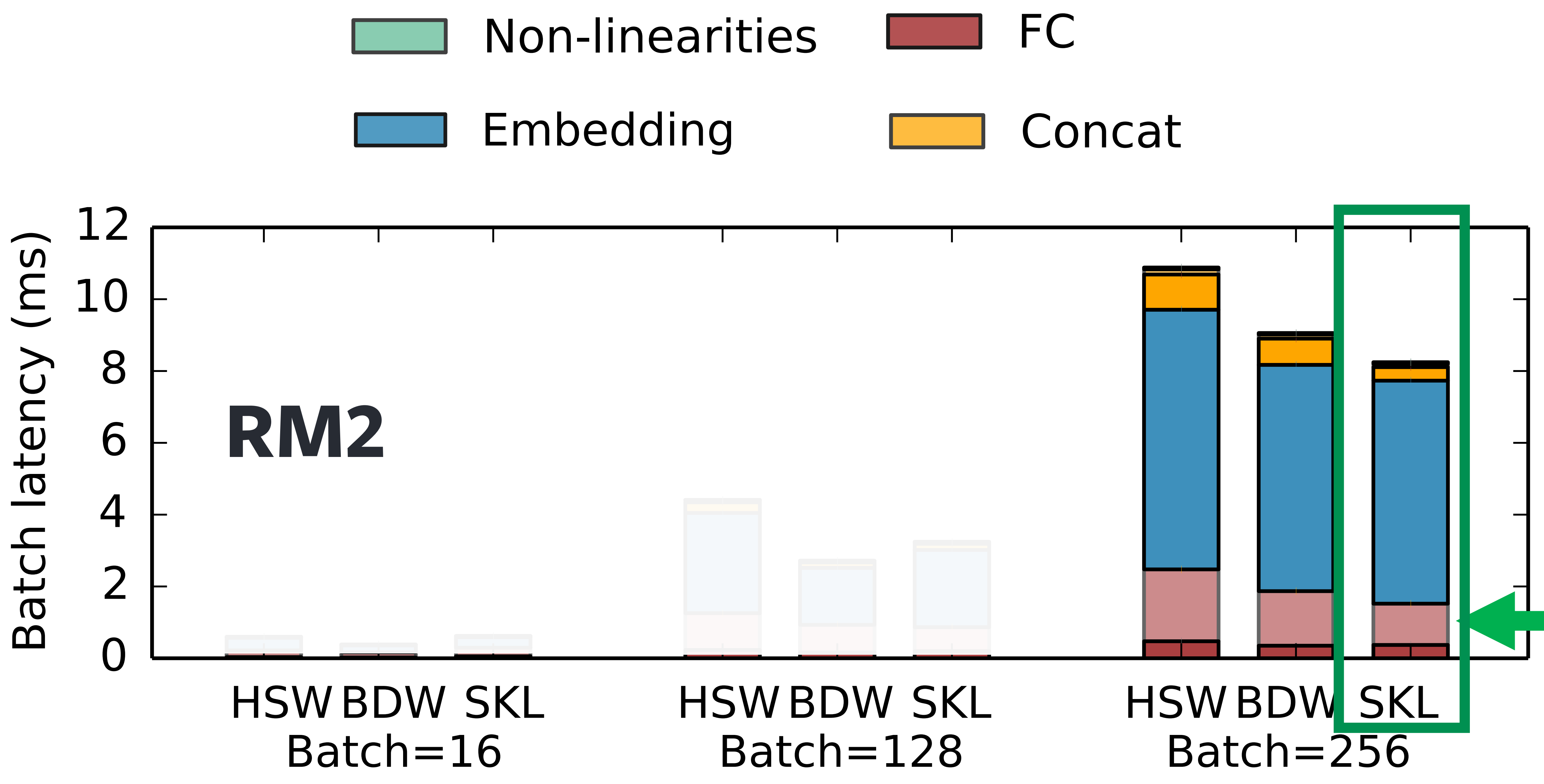
- At smaller batch-sizes Broadwell has 1.4x lower batch latency
  - Haswell: 50% lower DRAM frequency
  - Skylake: 20% lower CPU frequency and lower AVX-512 utilization (70%)

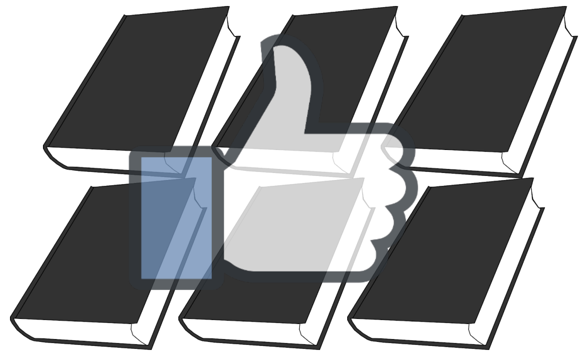
Data parallelism: Characterizing latency bounded throughput design space



- At higher batch-sizes Skylake has lower batch latency
  - Wider vector width and higher AVX-512 utilization (90%)

Data parallelism: Characterizing latency bounded throughput design space





Skylake

Solutions must co-design data-level parallelism with application target, recommendation models, and hardware platforms

Characterizing latency bounded throughput design space

2.5 GHz
AVX-2
Inclusive L2/L3 caches
DDR3

Haswell

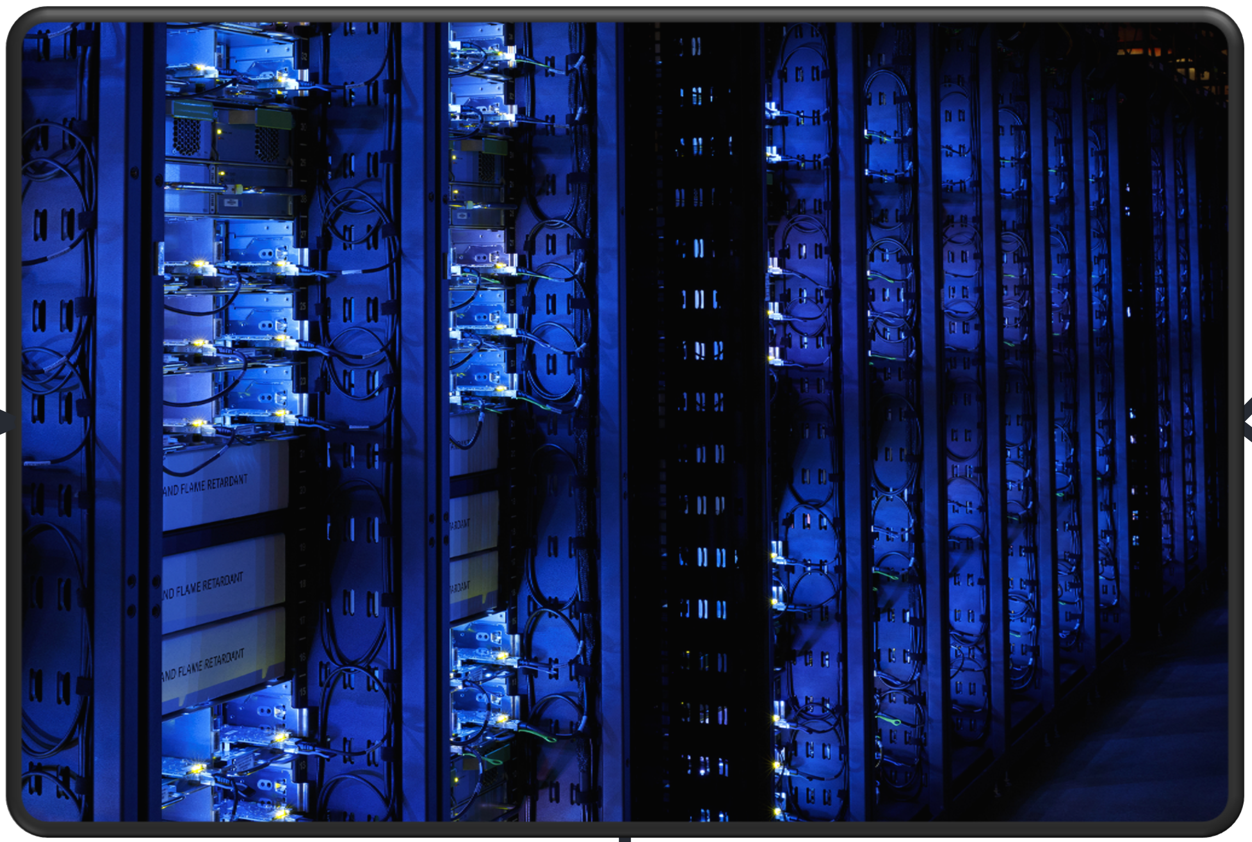
2.4 GHz
AVX-2
Inclusive L2/L3 caches
DDR4

Broadwell

2.0 GHz
AVX-512
Exclusive L2/L3 caches
DDR4

Skylake

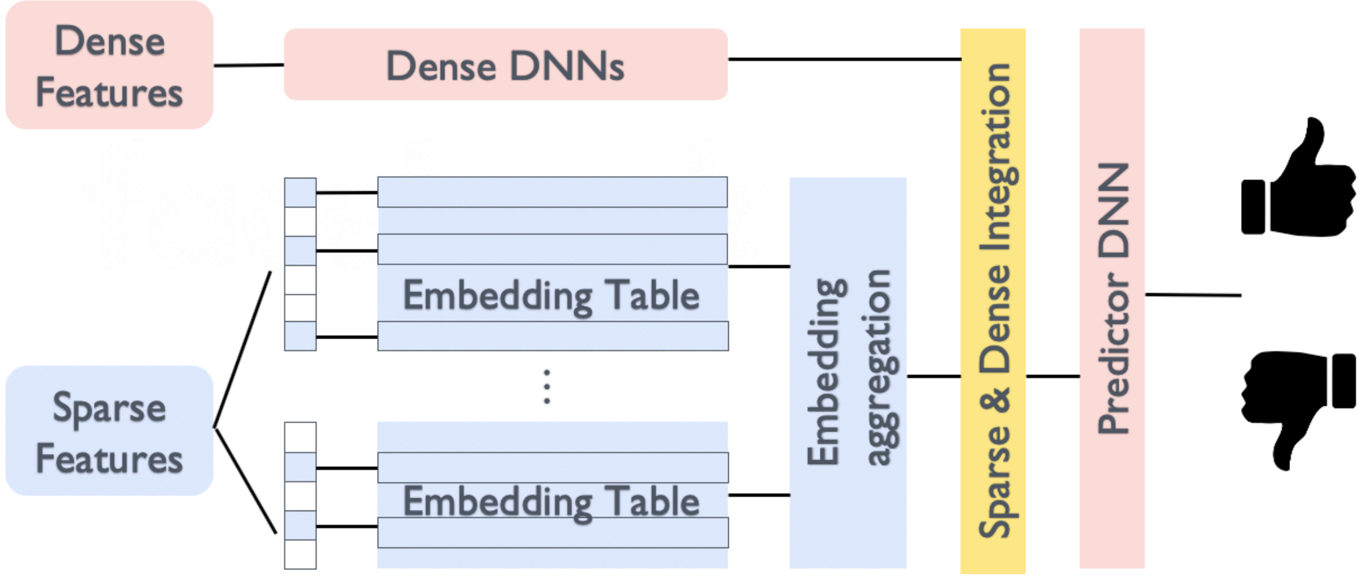
Hardware



Data level parallelism
(i.e., batch-size)

Task level parallelism
(i.e., co-locating models)

Parallelization

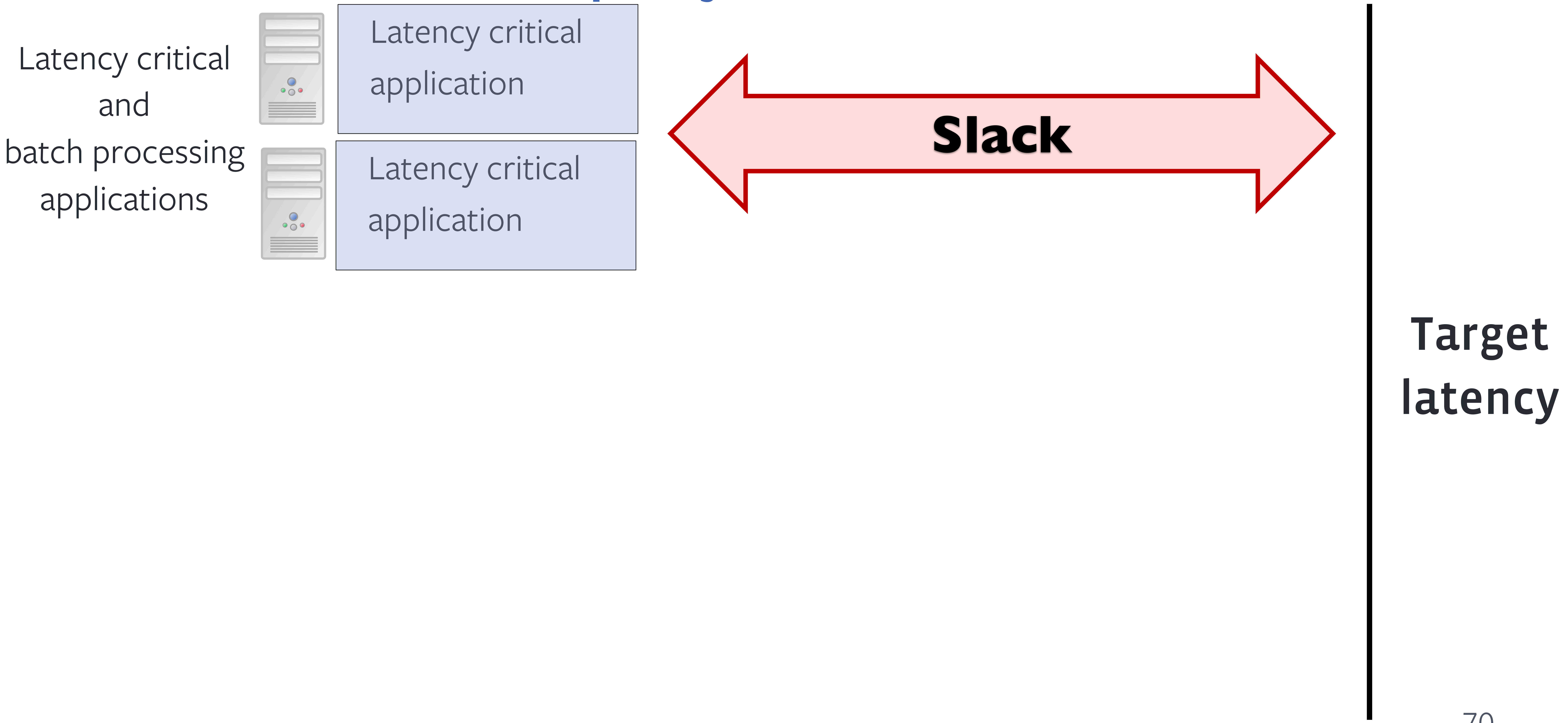


Models

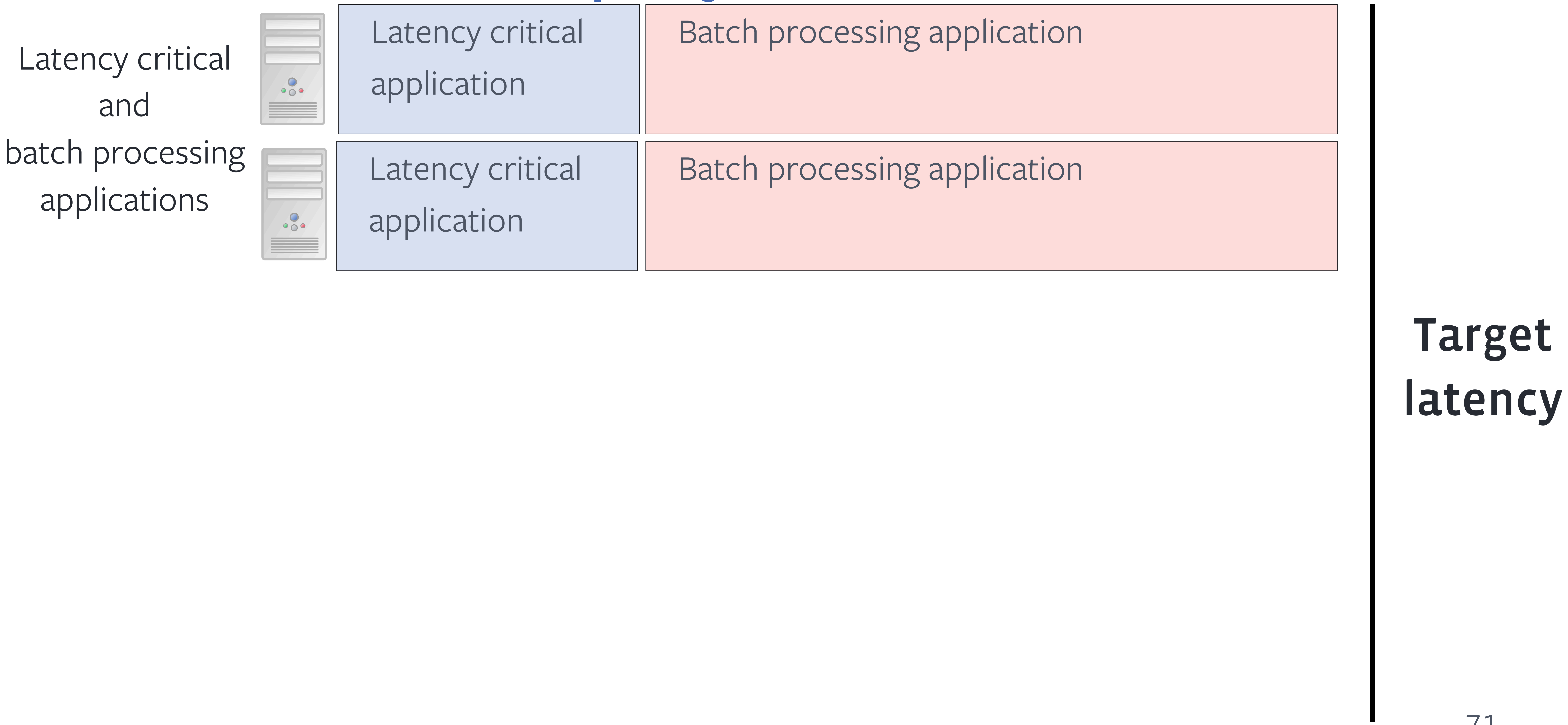
Co-locating models improves recommendation quality and reduces infrastructure capacity

Latency critical
and
batch processing
applications

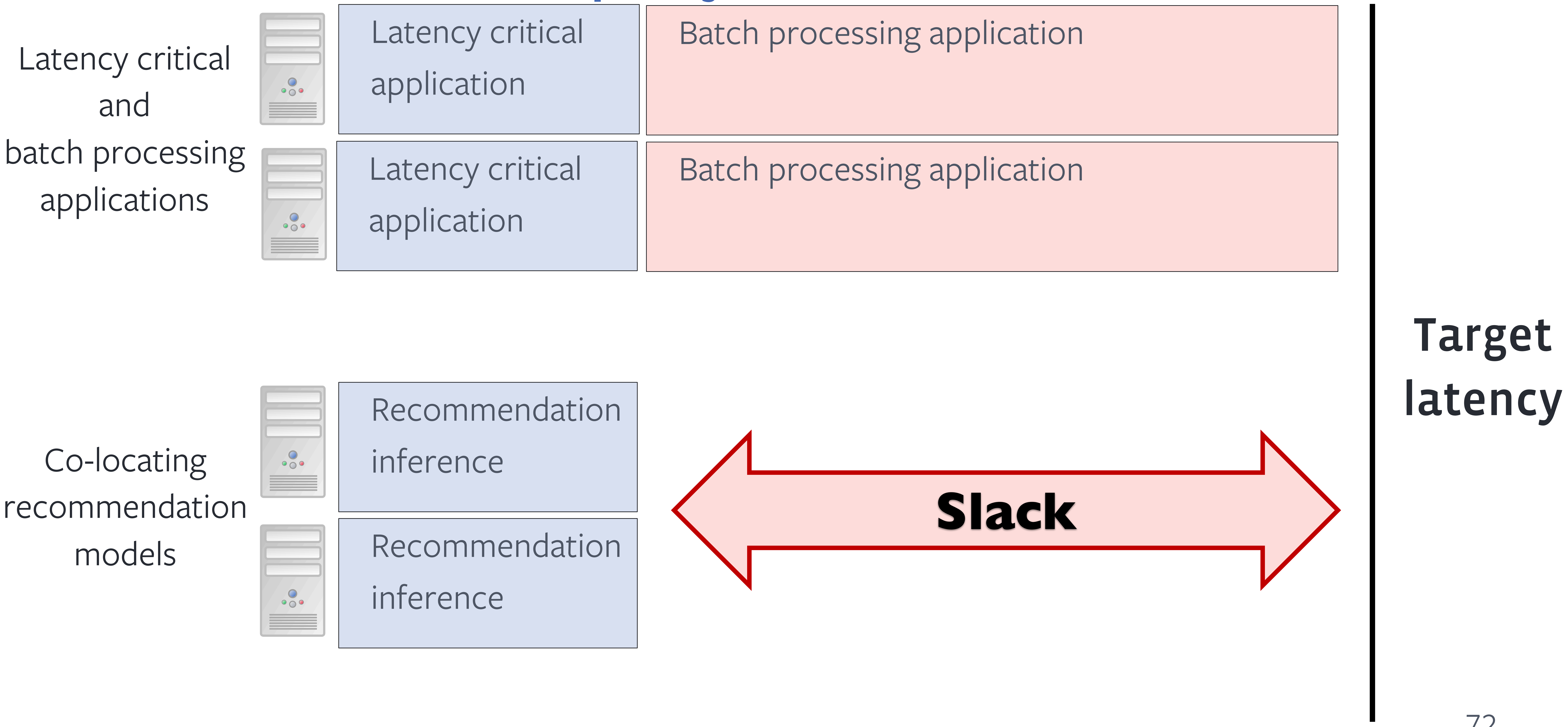
Co-locating models improves recommendation quality and reduces infrastructure capacity



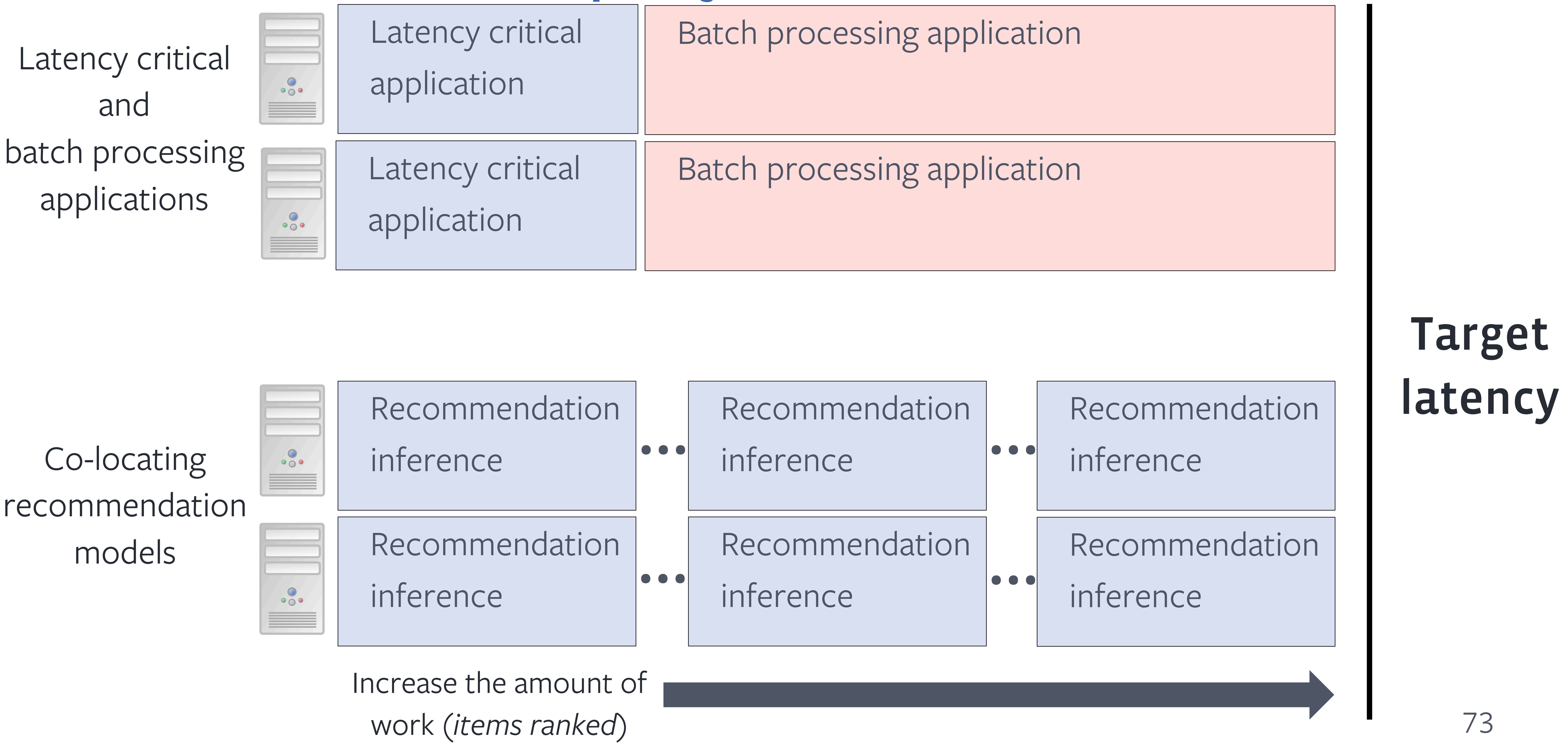
Co-locating models improves recommendation quality and reduces infrastructure capacity



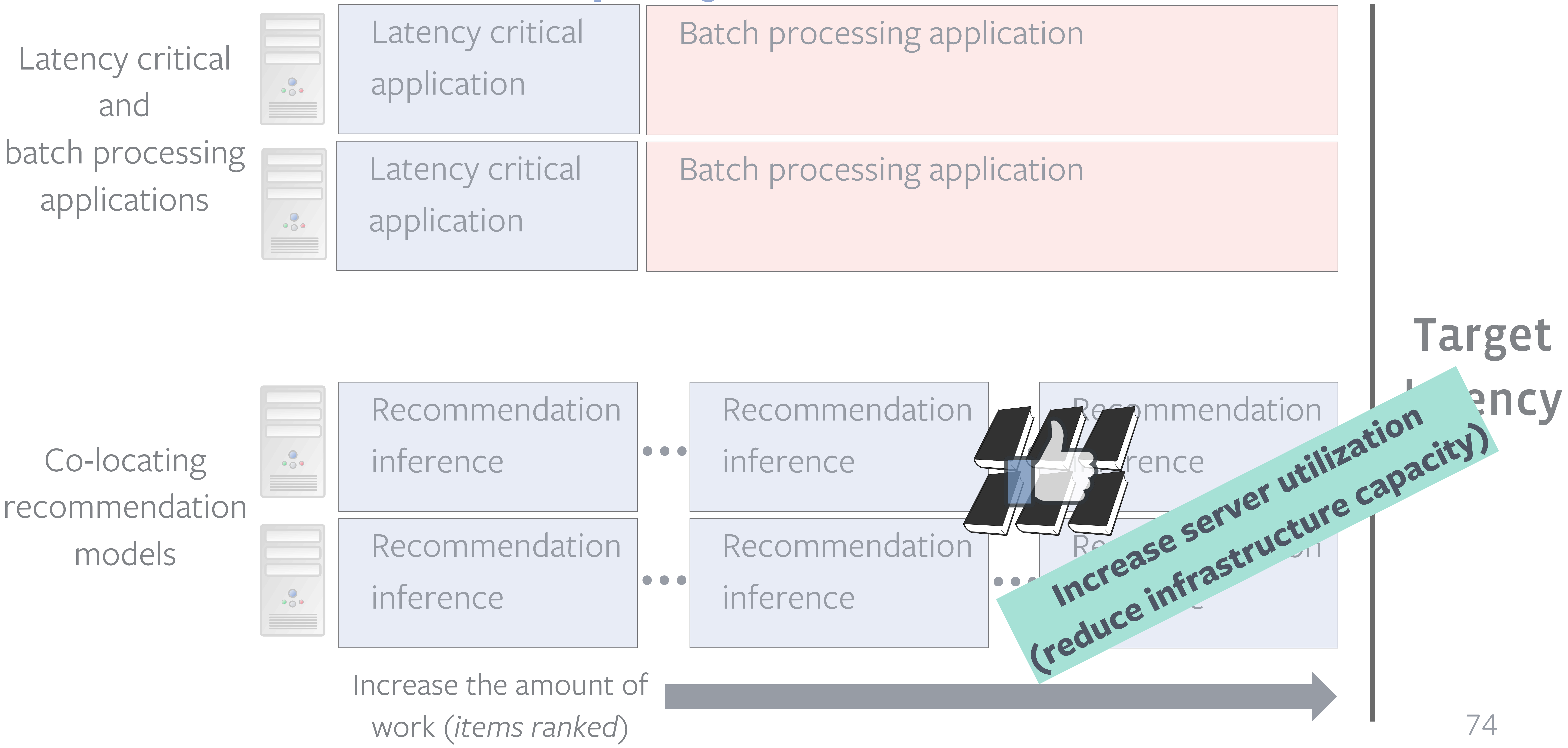
Co-locating models improves recommendation quality and reduces infrastructure capacity



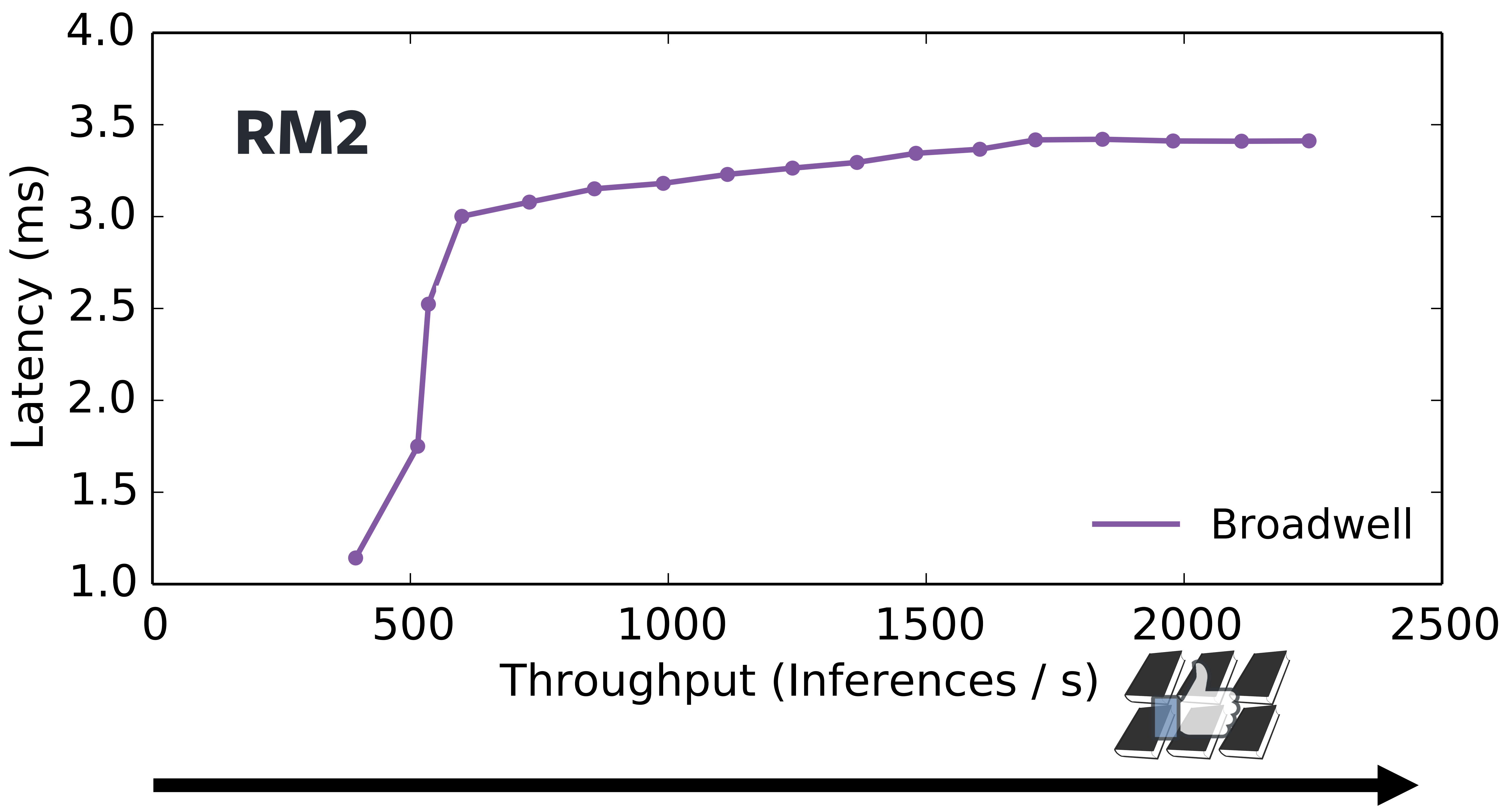
Co-locating models improves recommendation quality and reduces infrastructure capacity



Co-locating models improves recommendation quality and reduces infrastructure capacity

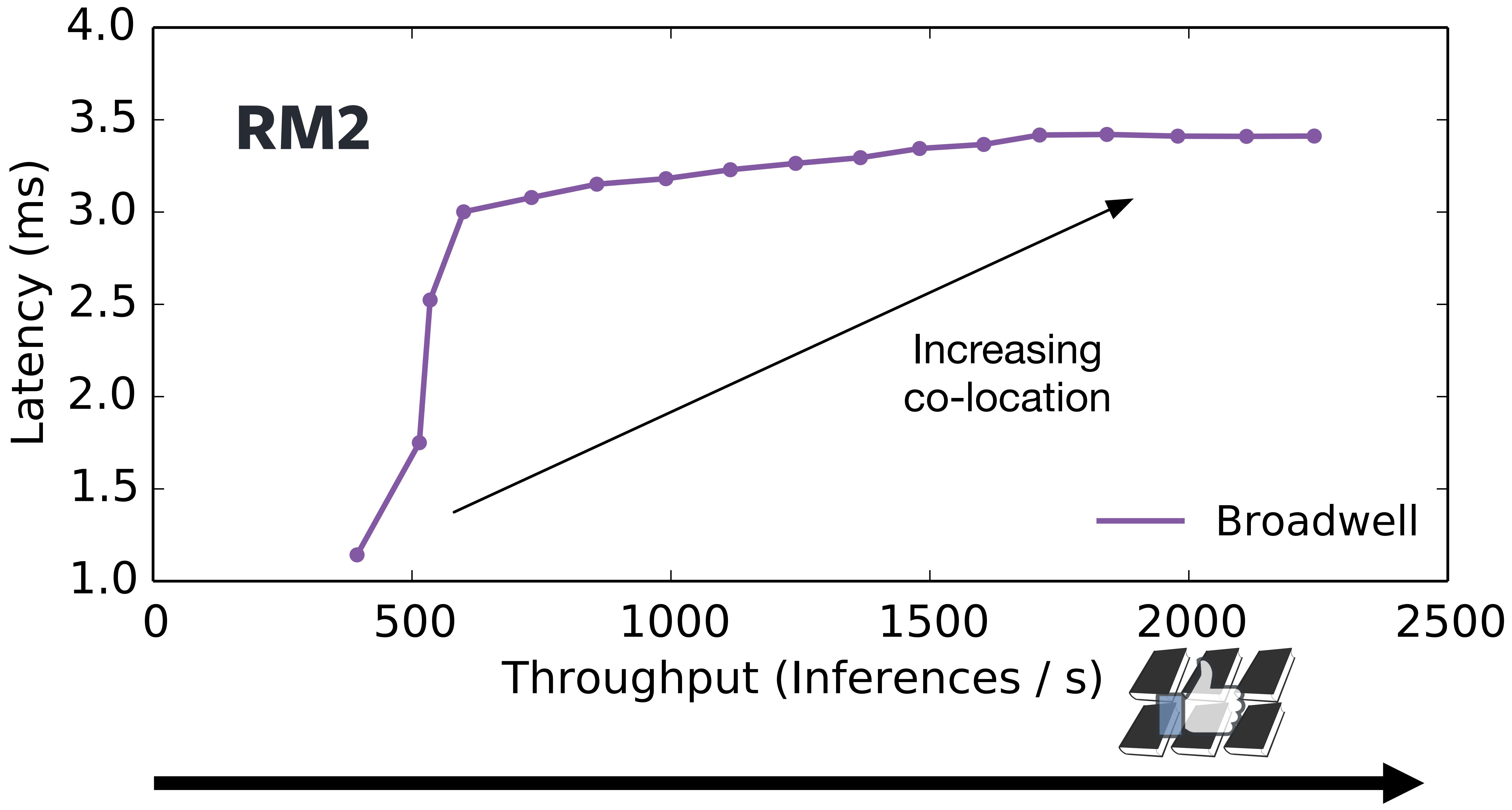


Task parallelism: Characterizing latency bounded throughput



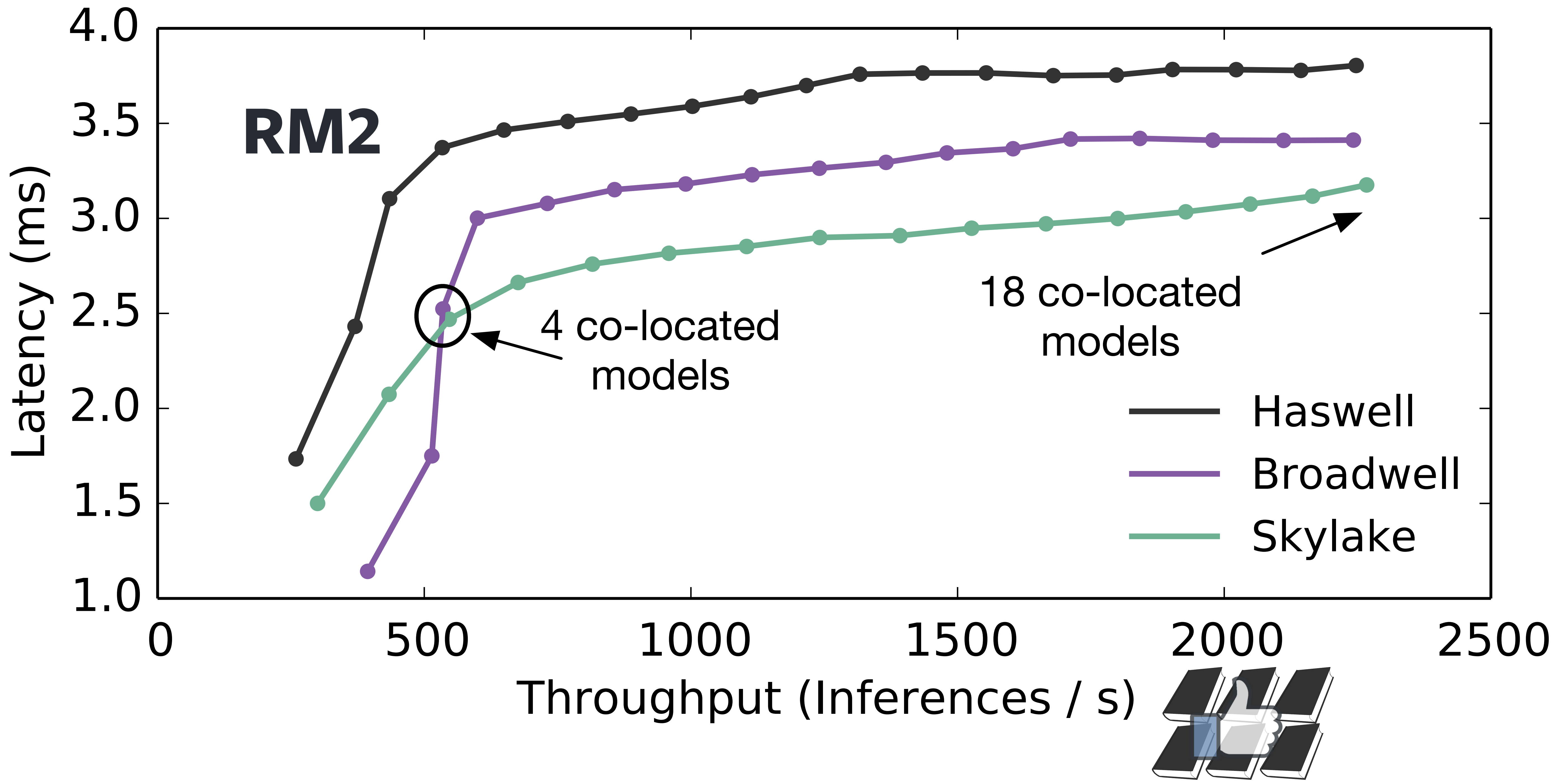
Items ranked (Higher is better)

Task parallelism: Characterizing latency bounded throughput

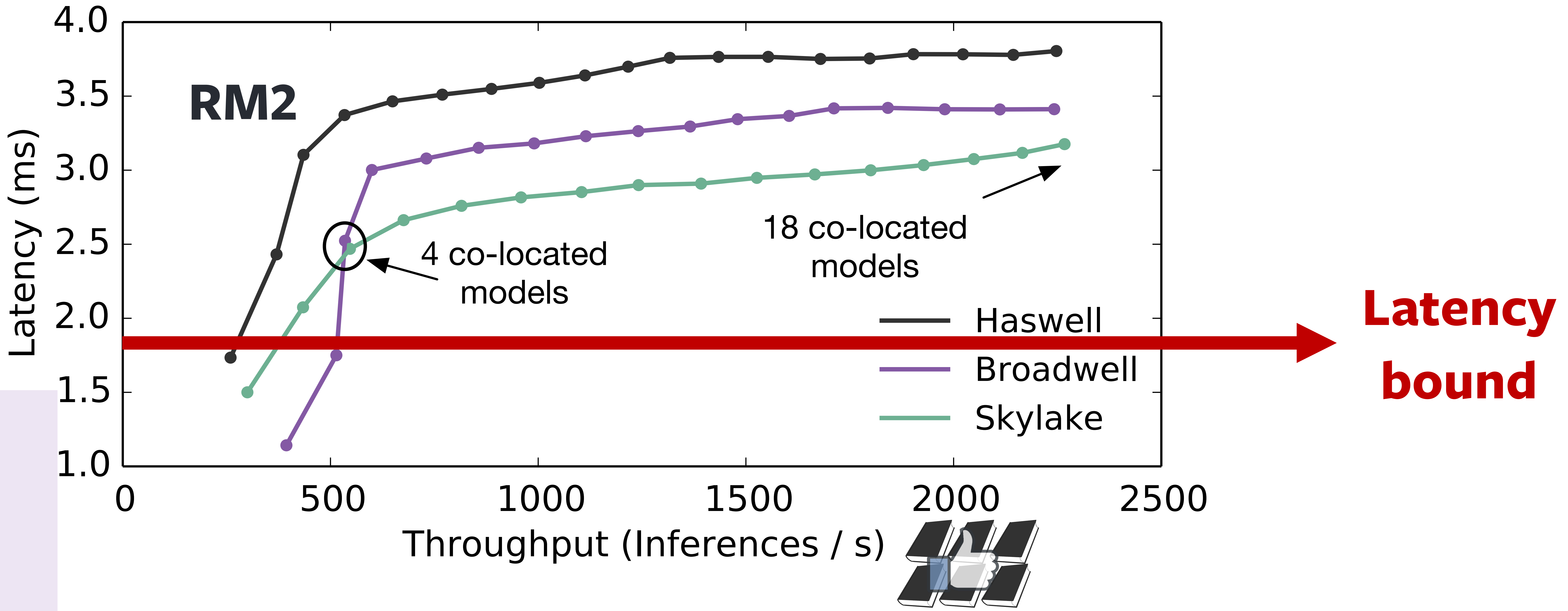


Items ranked (Higher is better)

Task parallelism: Characterizing latency bounded throughput



Task parallelism: Characterizing latency bounded throughput



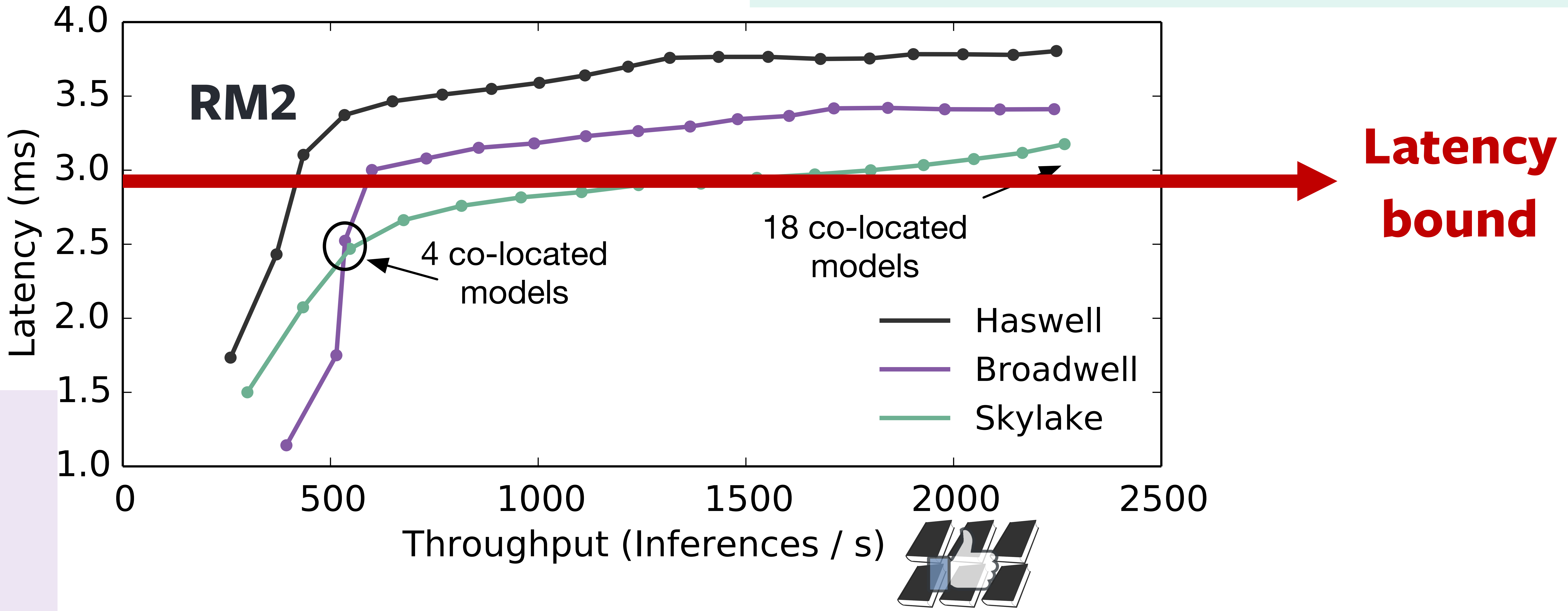
Broadwell is latency optimal

- Higher CPU frequency
- Inclusive L2/L3 caches

Task parallelism: Characterizing latency bounded throughput

Skylake is throughput optimal

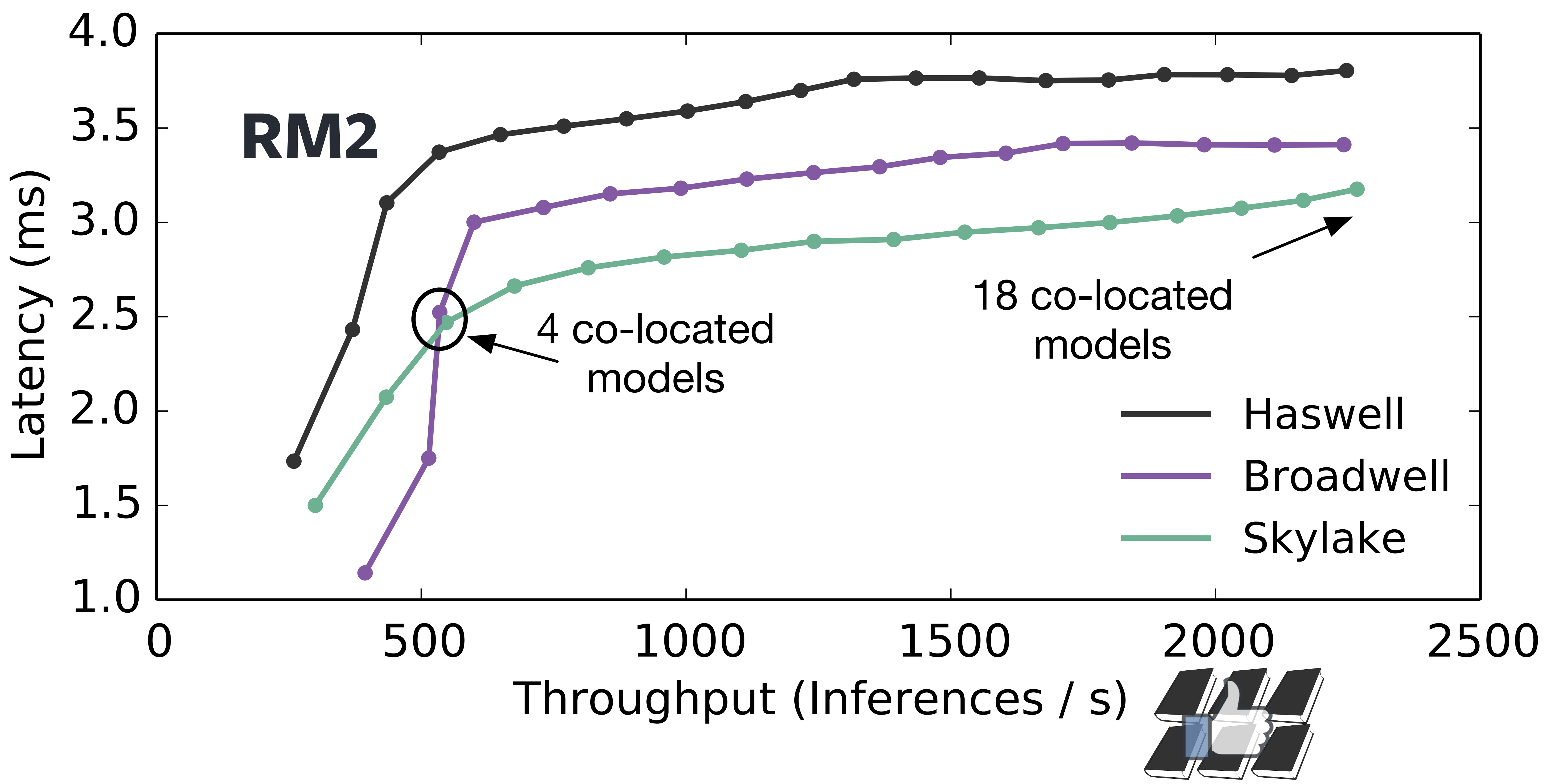
- Wider AVX width
- Exclusive L2/L3 caches



Broadwell is latency optimal

- Higher CPU frequency
- Inclusive L2/L3 caches

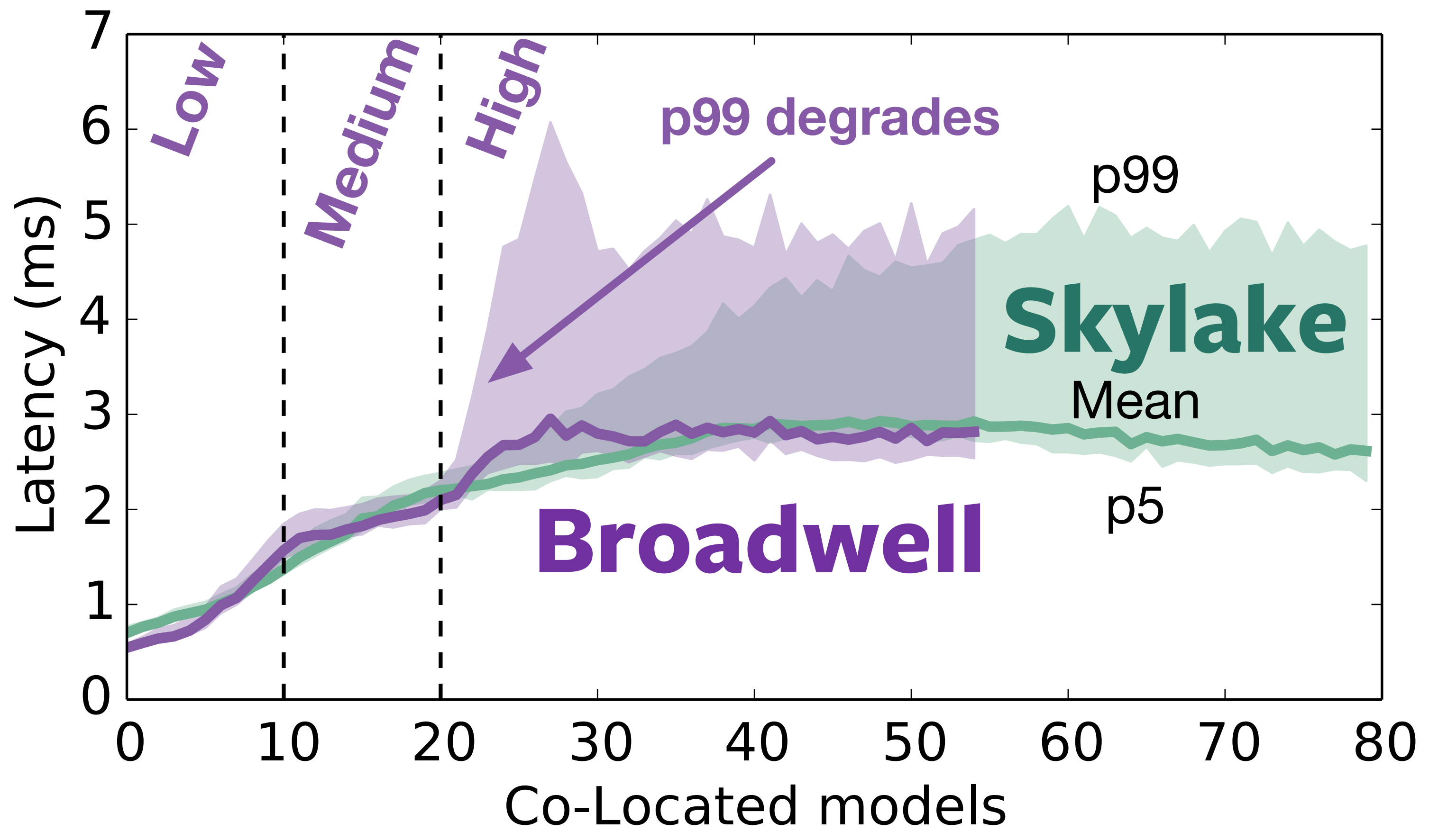
Task parallelism: Characterizing latency bounded throughput



Solutions must co-design task-level parallelism with application target, recommendation models, and hardware platforms

See the paper for more details!

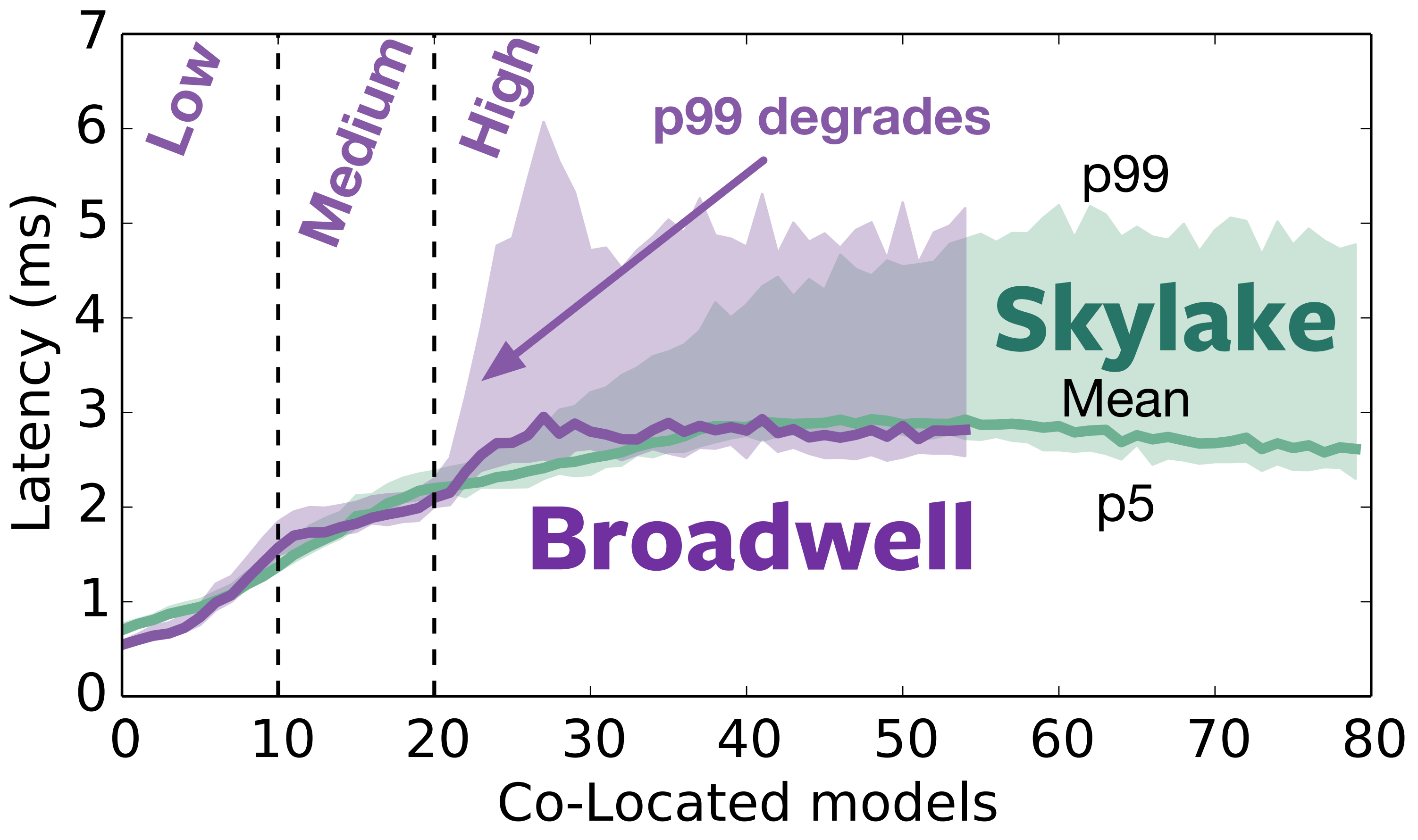
Performance variability



Impact of co-locating models on performance variability

See the paper for more details!

Performance variability



Impact of co-locating models on performance variability

Open-source

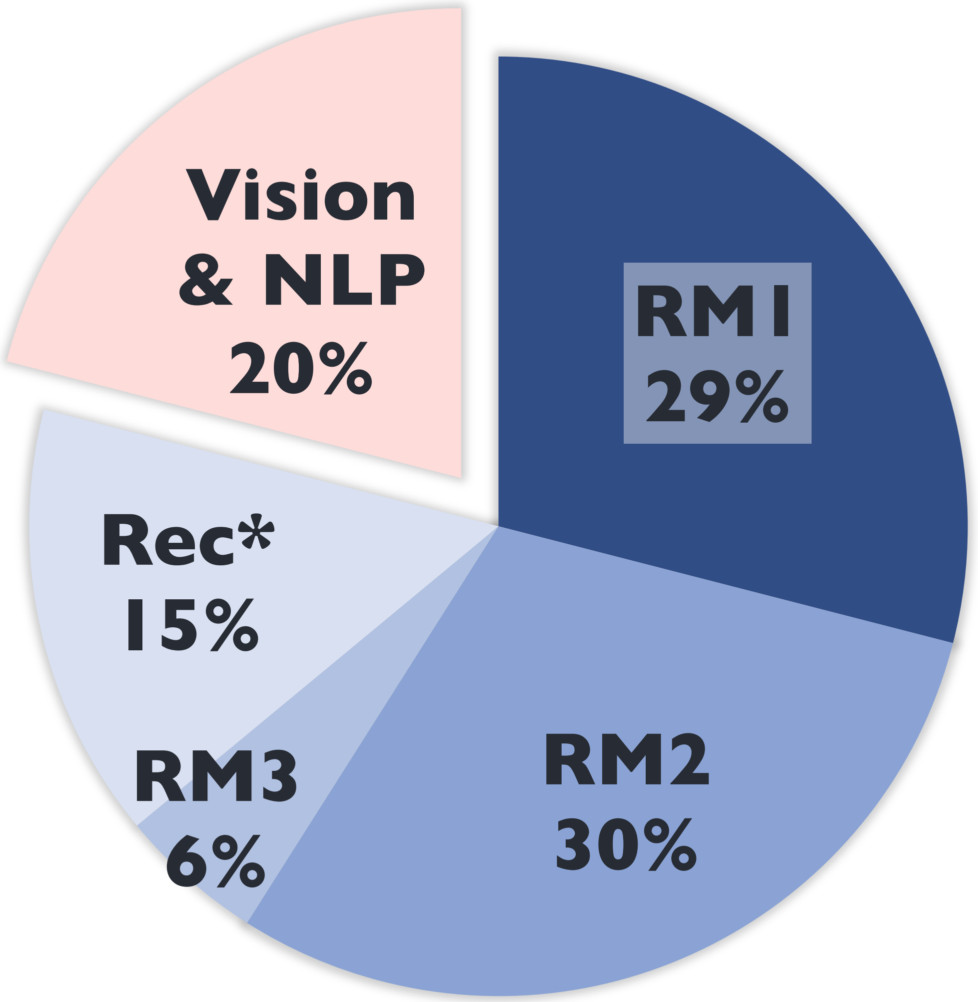
Model configurations using Facebook's open-source DLRM

Open-source data sets

"Deep Learning Recommendation Model for Personalization and Recommendation Systems" Naumov, et. al.

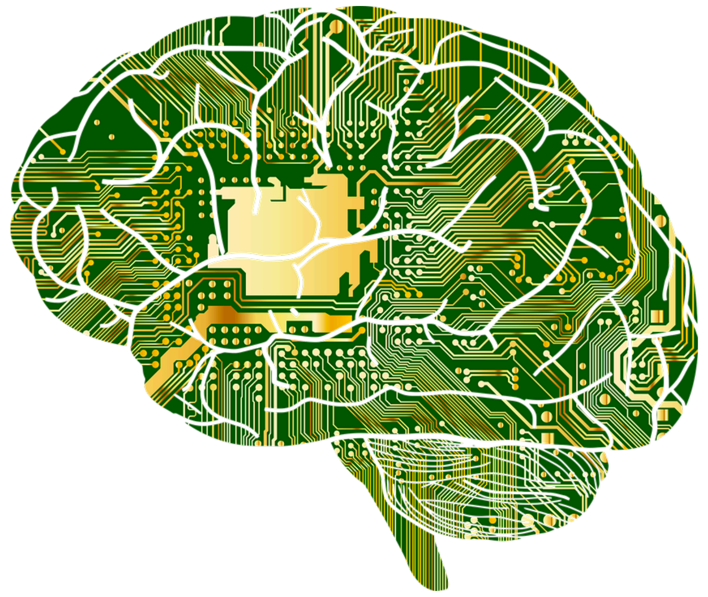
In this talk: Architectural Implications of Facebook's DNN-based Personalized Recommendation

Recommendation models



Importance of recommendation models

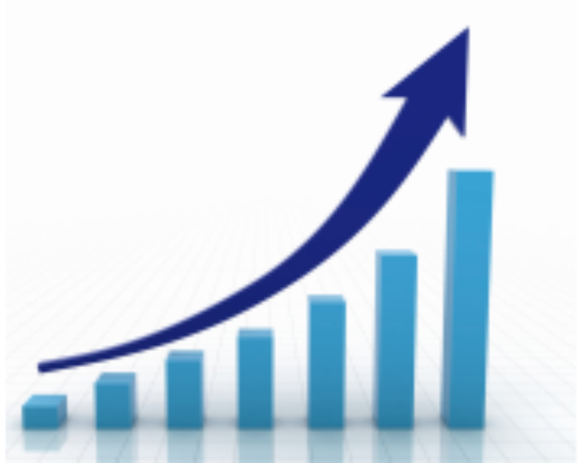
Diversity of recommendation models



Models have varying and unique performance characteristics

Design flexible hardware solutions

At-scale inference



Optimize for latency-bounded throughput

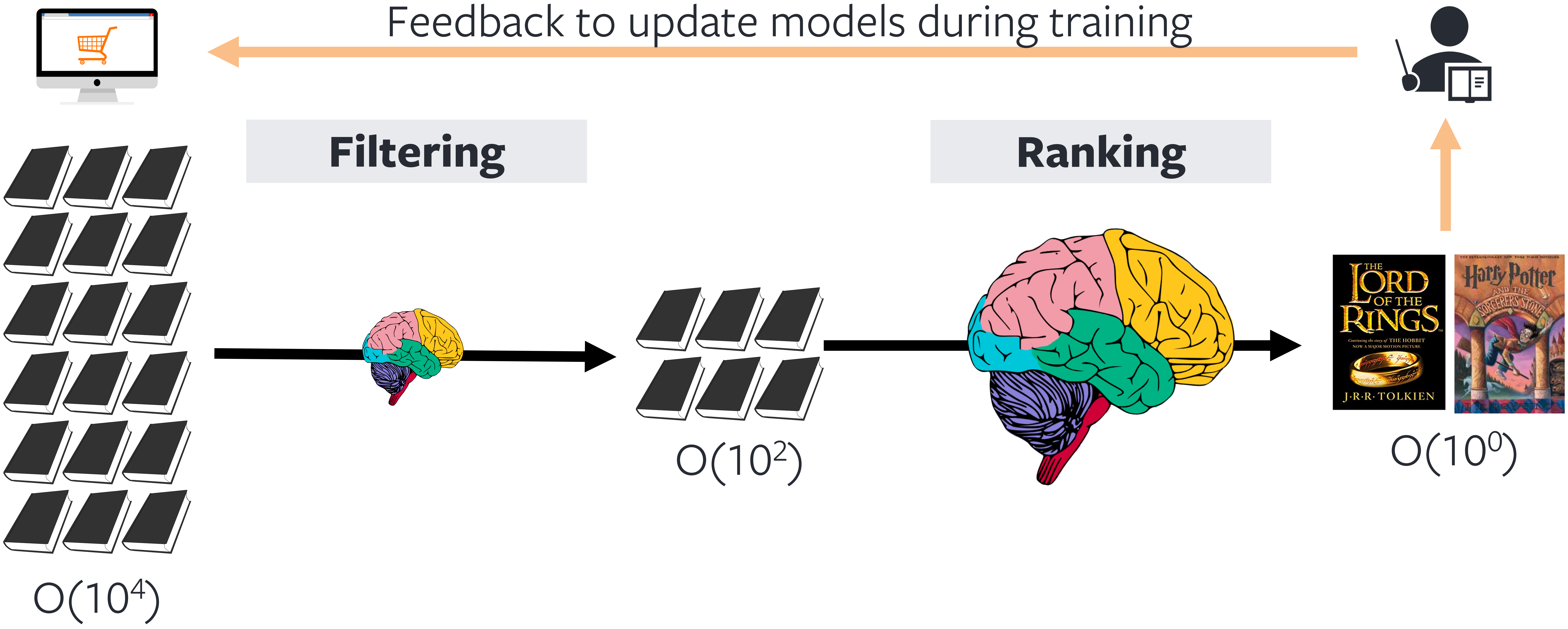
Co-design parallelism, application target, models, and hardware

MLPerf

ASPLOS 2020

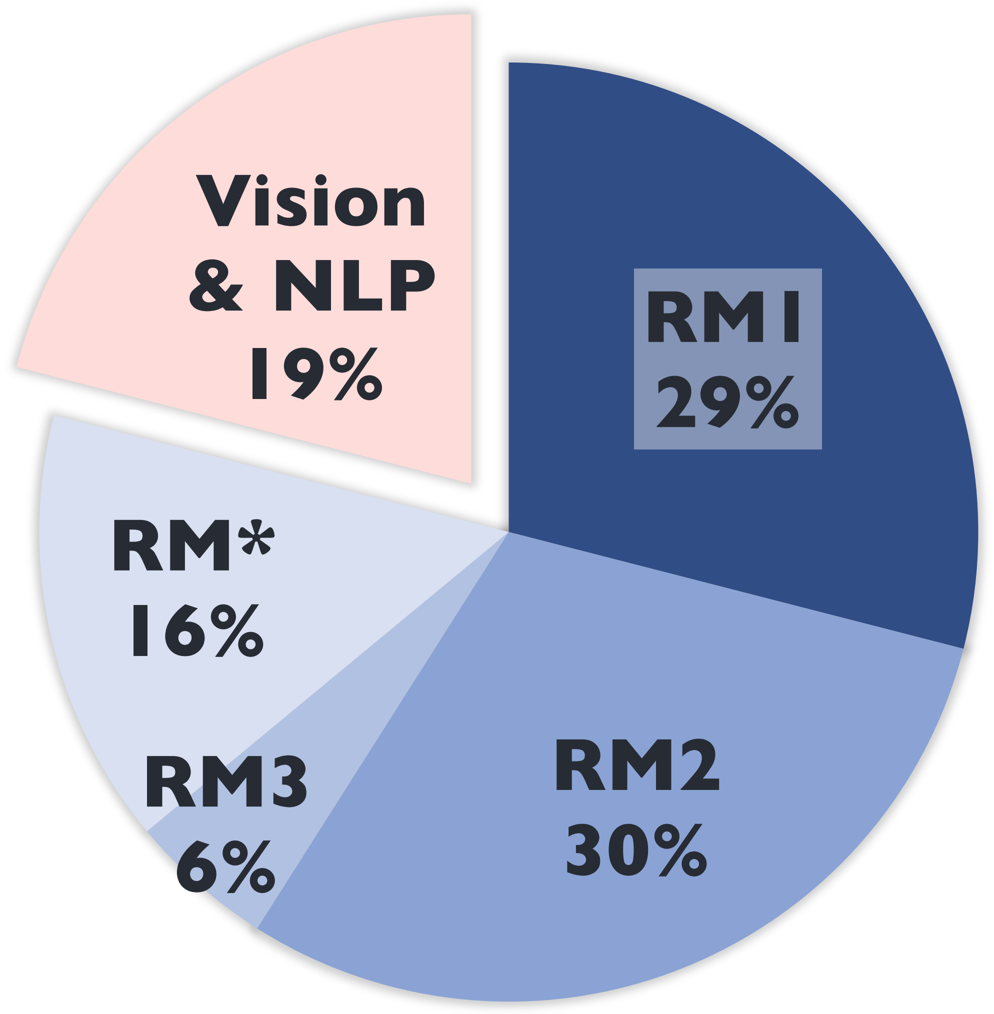
https://personal-tutorial.com/

Ranking thousands of items at-scale



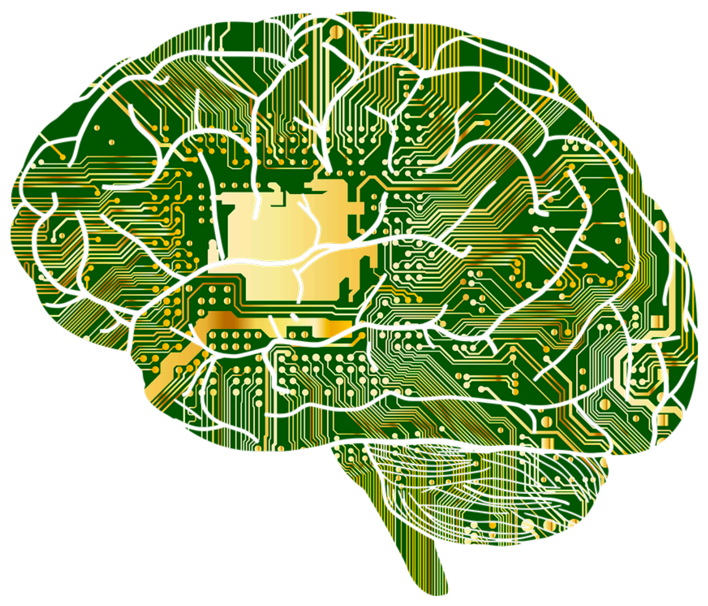
In this talk: Architectural Implications of Facebook's DNN-based Personalized Recommendation

Recommendation models



Importance of recommendation models

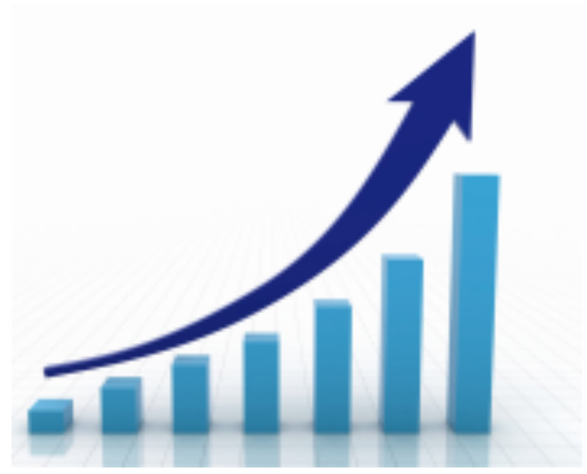
Diversity of recommendation models



Models have varying and unique performance characteristics

Design flexible hardware solutions

At-scale inference



Optimize for latency-bounded throughput

Co-design parallelism, application target, models, and hardware

MLPerf

ASPLOS 2020 https://personal-tutorial.com/