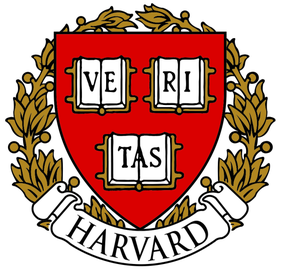
# The Architectural Implications of Facebook's DNN-based Personalized Recommendation

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



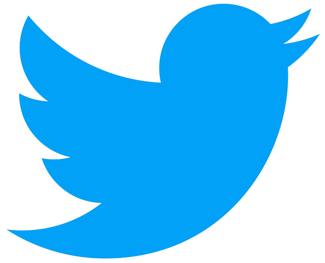

HPCA 2020

# Personalized Recommendation is everywhere

# Personalized Recommendation is everywhere

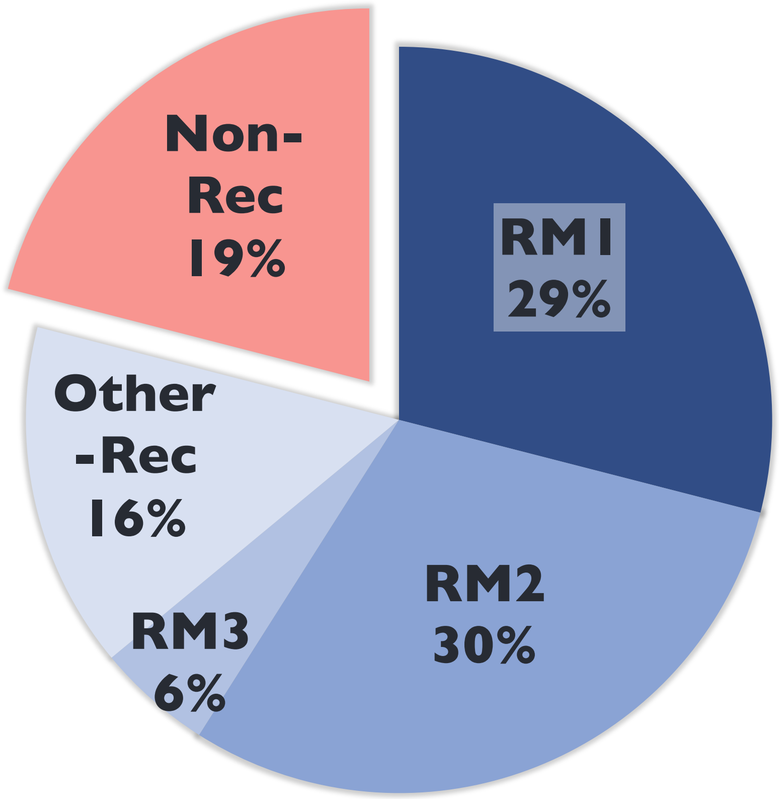


"35% of purchases on Amazon and 75% of videos on Netflix are powered by recommendation algorithms"
McKinsey & Co

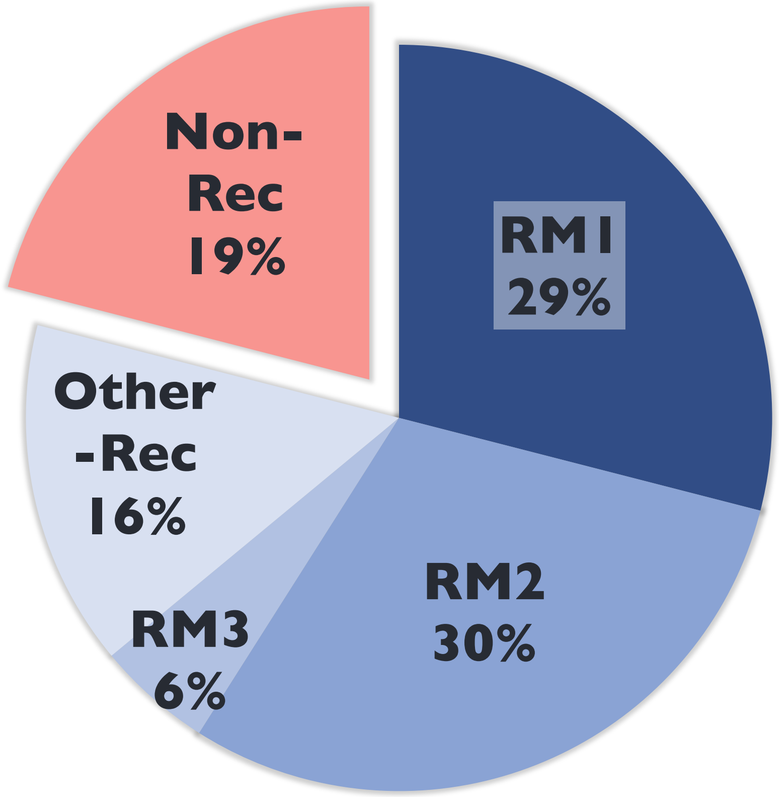
# Optimizing DNN-based recommendation is key for improving datacenter efficiency

AI inference cycles in Facebook's datacenter



# Optimizing DNN-based recommendation is key for improving datacenter efficiency

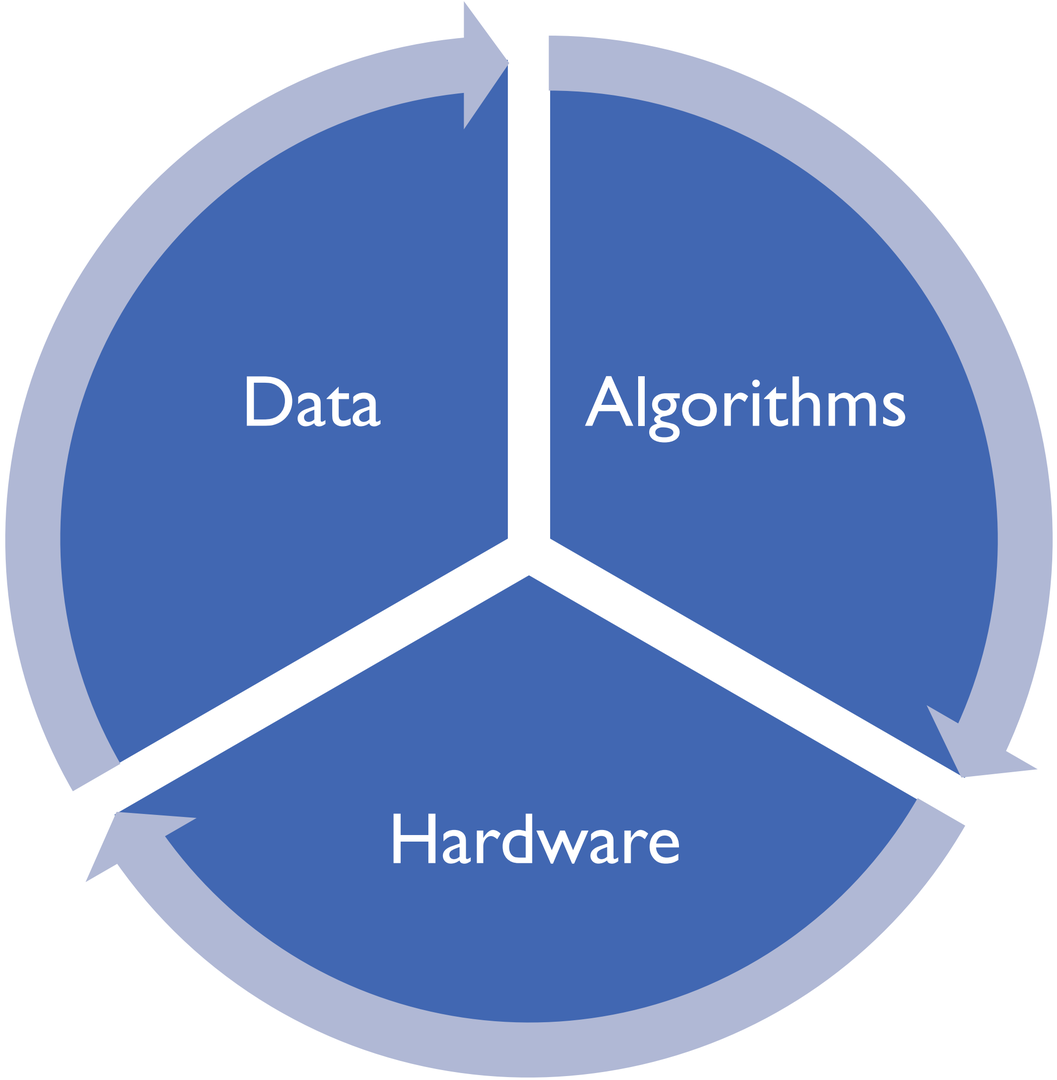
AI inference cycles in Facebook's datacenter

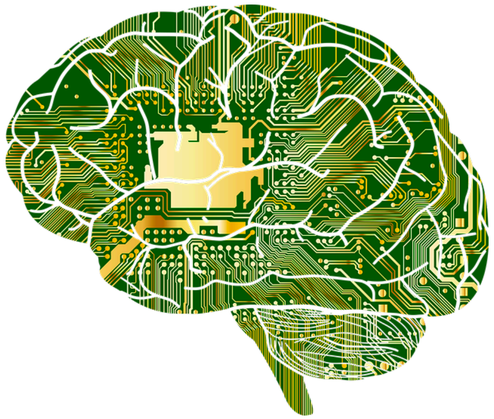



**Recommendation uses cases account for over 80% of all AI inference cycles in Facebook's datacenter**

# Lots of opportunities for HW research in recommendation

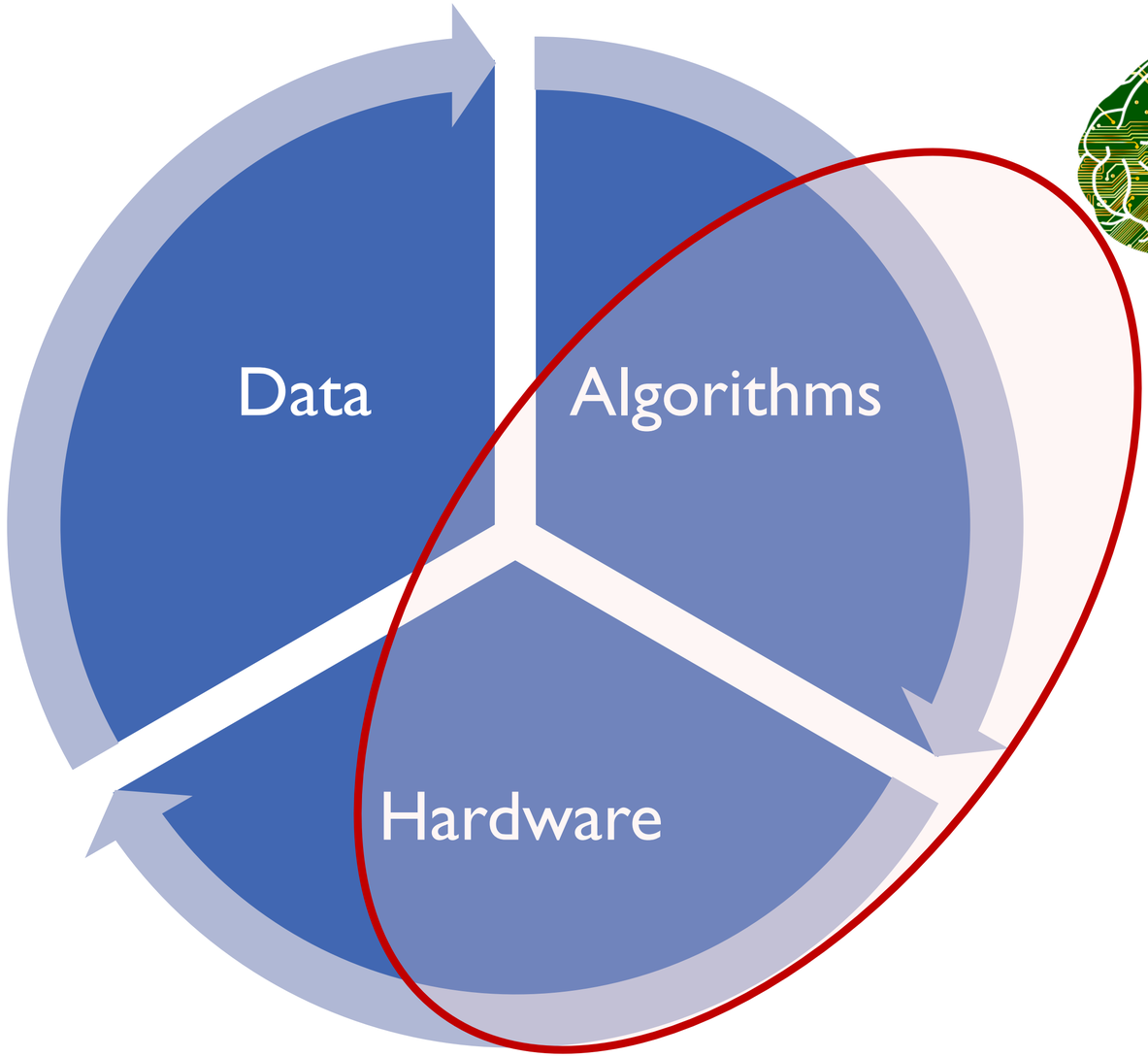
HCI

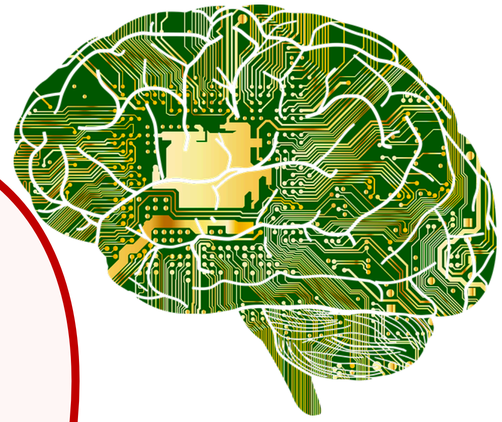


# Lots of opportunities for HW research in recommendation

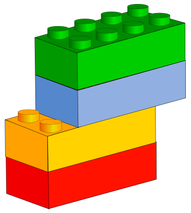
HCI



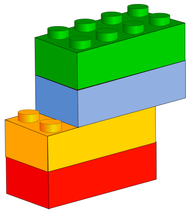
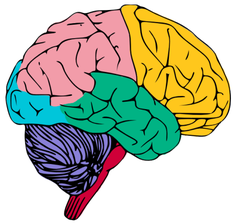



**This work!**

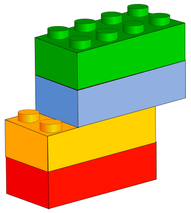
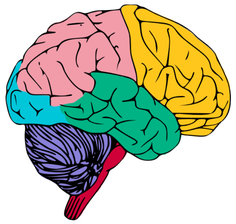
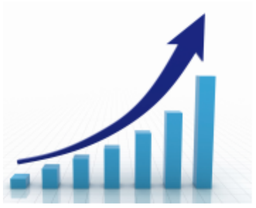
# Hardware insights of recommendation

| Algorithmic | Hardware |
| --- | --- |
|  General model structure | Requires optimizing operators with new storage, compute, and memory access requirements |

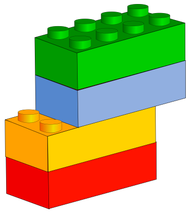
# Hardware insights of recommendation

| Algorithmic | Hardware |
| --- | --- |
|  General model structure | Requires optimizing operators with new storage, compute, and memory access requirements |
|  Diverse model architectures | Accelerating recommendation needs flexible and diverse system solutions |

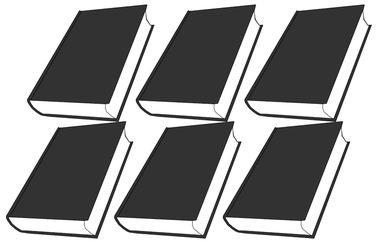
# Hardware insights of recommendation

| Algorithmic | Hardware |
| --- | --- |
|  General model structure | Requires optimizing operators with new storage, compute, and memory access requirements |
|  Diverse model architectures | Accelerating recommendation needs flexible and diverse system solutions |
|  Processing queries at-scale | Exploiting hardware heterogeneity and parallelism can optimize latency-bounded throughput |

# Hardware insights of recommendation

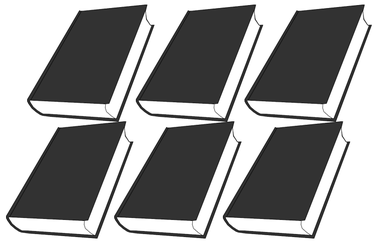
| | Algorithmic | Hardware |
|---|---|---|
|  | General model structure | Requires optimizing operators with new storage, compute, and memory access requirements |
|  | Diverse model architectures | Accelerating recommendation needs flexible and diverse system solutions |
|  | Processing queries at-scale | Exploiting hardware heterogeneity and parallelism can optimize latency-bounded throughput |

# DNNs for Recommendation



?

# DNNs for Recommendation



?

***Continuous (dense) features***

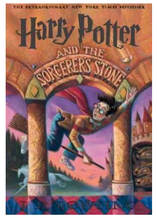
Age

Time of day

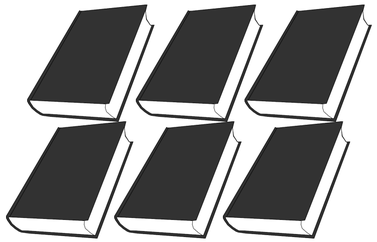
***Categorical (sparse) features***

User search history

Book's genre



# DNNs for Recommendation



?

***Continuous (dense) features***

Age
Time of day

Dense DNNs

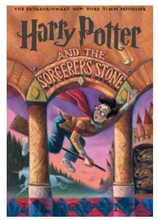
***Categorical (sparse) features***

User search history

User

Visited
- Inkheart
- Moby Dick
- Hunger Games

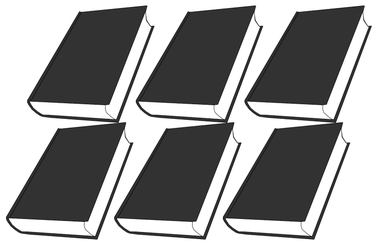
Book's genre




Item (Book)

Genre
- Magic
- Series

# DNNs for Recommendation



?

***Continuous (dense) features***

Age
Time of day

Dense DNNs

***Categorical (sparse) features***

User search history

User

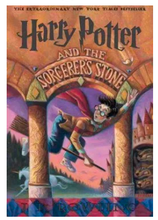
Visited

Inkheart
Moby Dick
Hunger Games

Embedding Table

0.9 0.7
-0.5 0.4
0.7 0.8

Book's genre



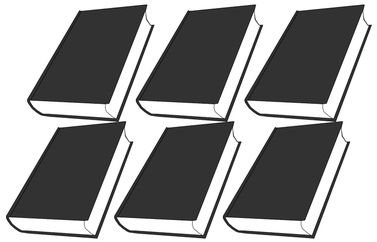

Item (Book)

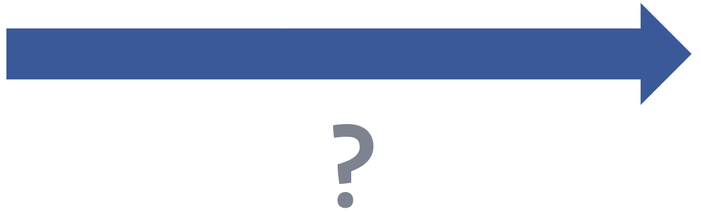
Magic
Series

Embedding Table

Genre

# DNNs for Recommendation





?

***Continuous (dense) features***

Age
Time of day

Dense DNNs

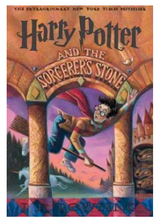
Visited
Inkheart
Moby Dick
Hunger Games
Embedding Table

User

User search history

***Categorical (sparse) features***

Book's genre



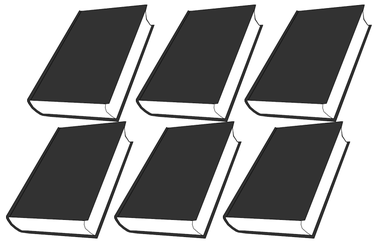

Item (Book)

Magic
Series
Genre
Embedding Table

Embedding aggregation

Sparse & Dense Integration

# DNNs for Recommendation



?

***Continuous (dense) features***

Age
Time of day

Dense DNNs

***Categorical (sparse) features***

User search history

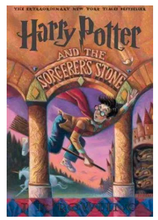
User

Visited

Inkheart
Moby Dick
Hunger Games

Embedding Table

Book's genre




Item (Book)

Genre

Magic
Series

Embedding Table

Embedding aggregation

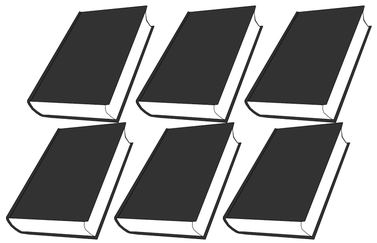
Sparse & Dense Integration

Predictor DNN

90%

10%

# DNNs for Recommendation



?

***Continuous (dense) features***

Age
Time of day

Dense DNNs

***Categorical (sparse) features***

User search history

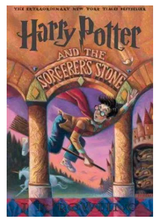
User

Visited

Inkheart
Moby Dick
Hunger Games

Embedding Table

Book's genre




Item (Book)

Genre

Magic
Series

Embedding Table

Embedding aggregation

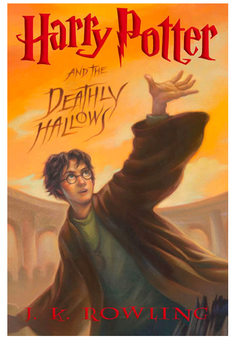
Sparse & Dense Integration

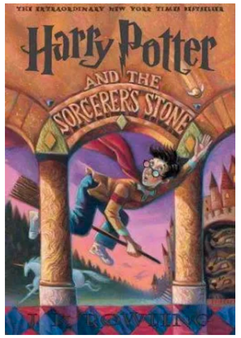
Predictor DNN

90% 84% 28%

12% 3% 57%

# DNNs for Recommendation




**Recommend to user**

***Continuous (dense) features***

Age
Time of day

Dense DNNs

***Categorical (sparse) features***

User search history

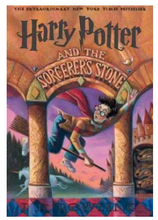
User

Visited

- Inkheart
- Moby Dick
- Hunger Games

Embedding Table

Book's genre




Item (Book)

- Magic
- Series

Genre

Embedding Table

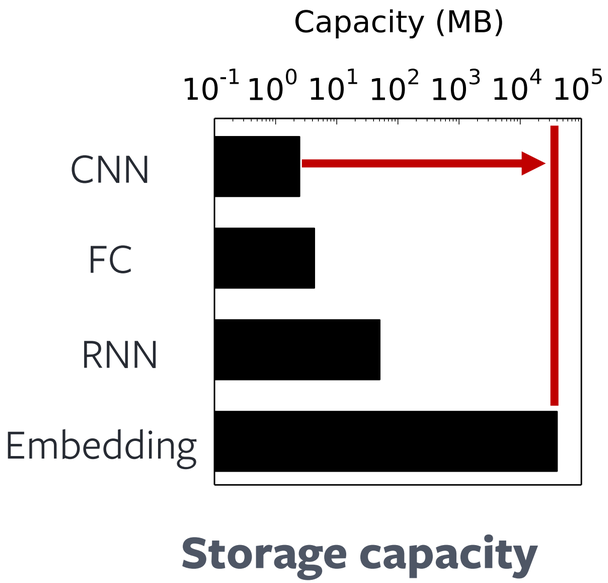
Embedding aggregation

Sparse & Dense Integration

Predictor DNN

# Embedding tables pose new challenges

Log Scale!



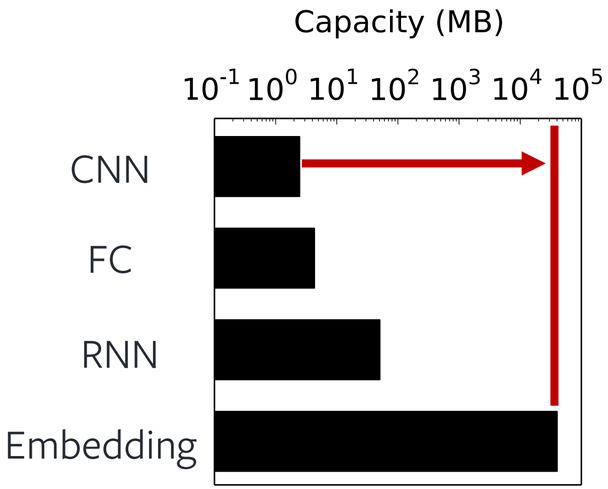

**Storage capacity**

Up to tens of GBs

Off-chip memory
(DRAM, NVM)

# Embedding tables pose new challenges

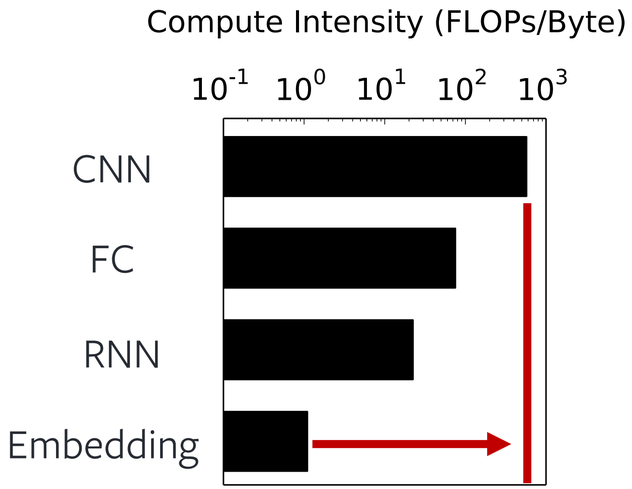
Log Scale!




**Storage capacity**

Up to tens of GBs

Off-chip memory
(DRAM, NVM)



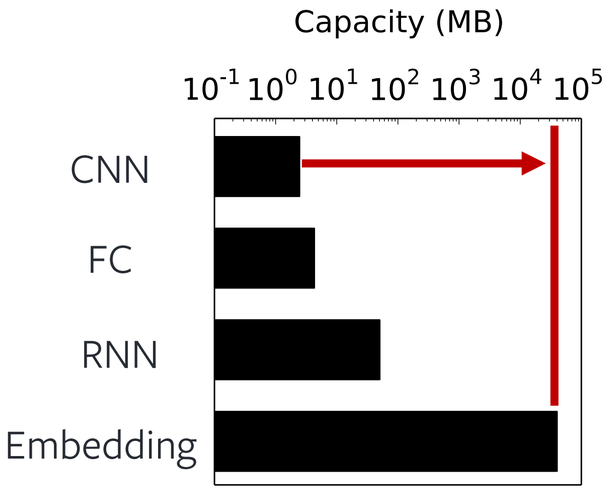

**Compute intensity**

Orders of magnitude lower
FLOPs/Byte

Unique acceleration
opportunities
(Near memory computing)

# Embedding tables pose new challenges

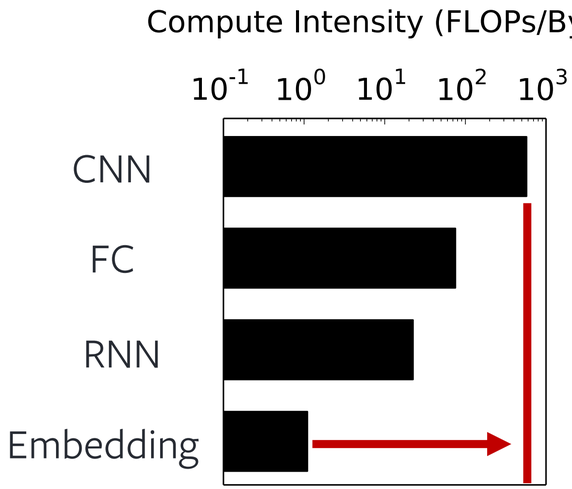
Log Scale!




**Storage capacity**

Up to tens of GBs

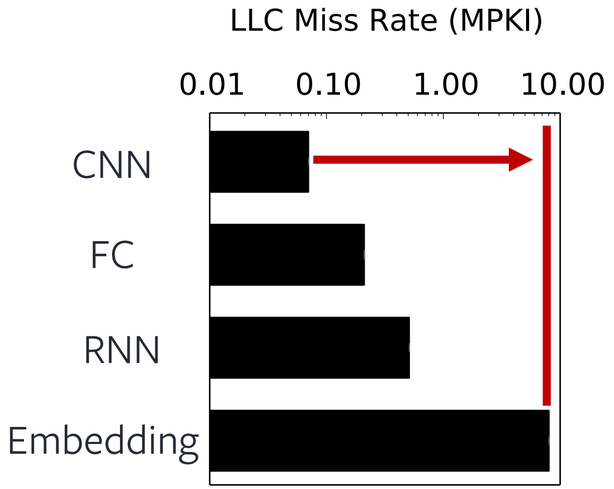
Off-chip memory
(DRAM, NVM)




**Compute intensity**

Orders of magnitude lower
FLOPs/Byte

Unique acceleration
opportunities
(Near memory computing)

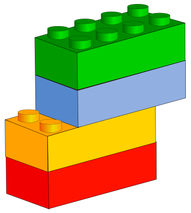
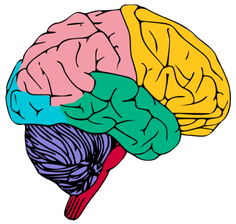



**Memory access pattern**

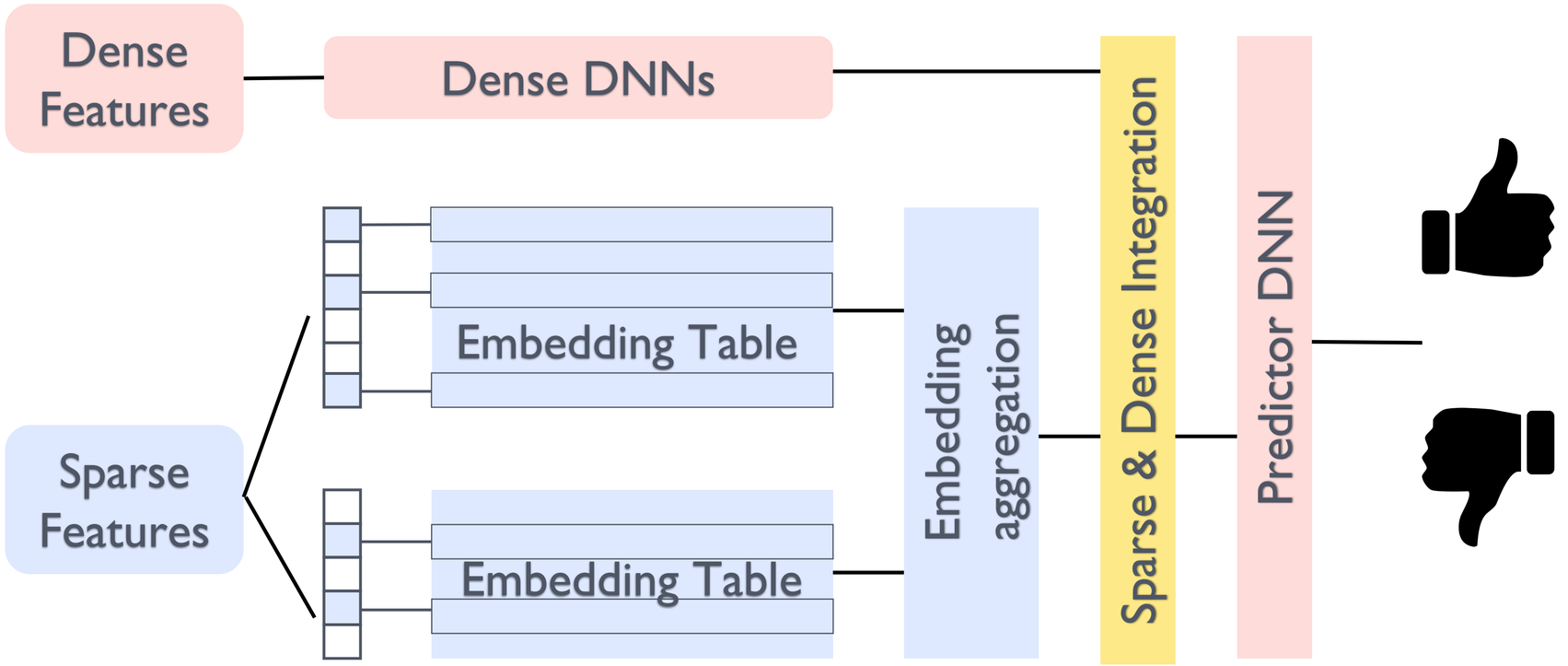
Sparse, irregular memory
accesses

Specialized caching and
pre-fetching capabilities

# Hardware insights of recommendation

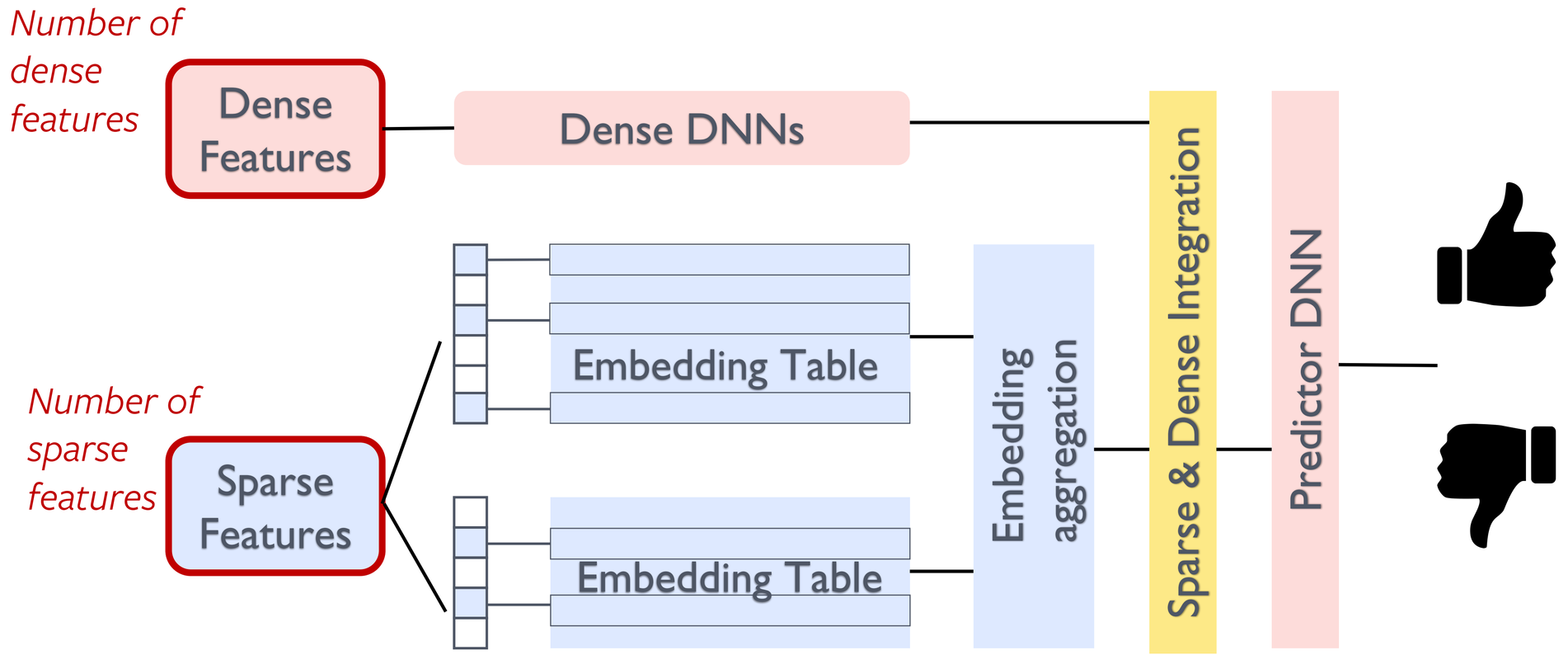
| Algorithmic | Hardware |
| --- | --- |
|  General model structure | Requires optimizing operators with new storage, compute, and memory access requirements |
|  Diverse model architectures | Accelerating recommendation needs flexible and diverse system solutions |
|  Processing queries at-scale | Exploiting hardware heterogeneity and parallelism can optimize latency-bounded throughput |

# DLRM: Configurable benchmark for end to end models



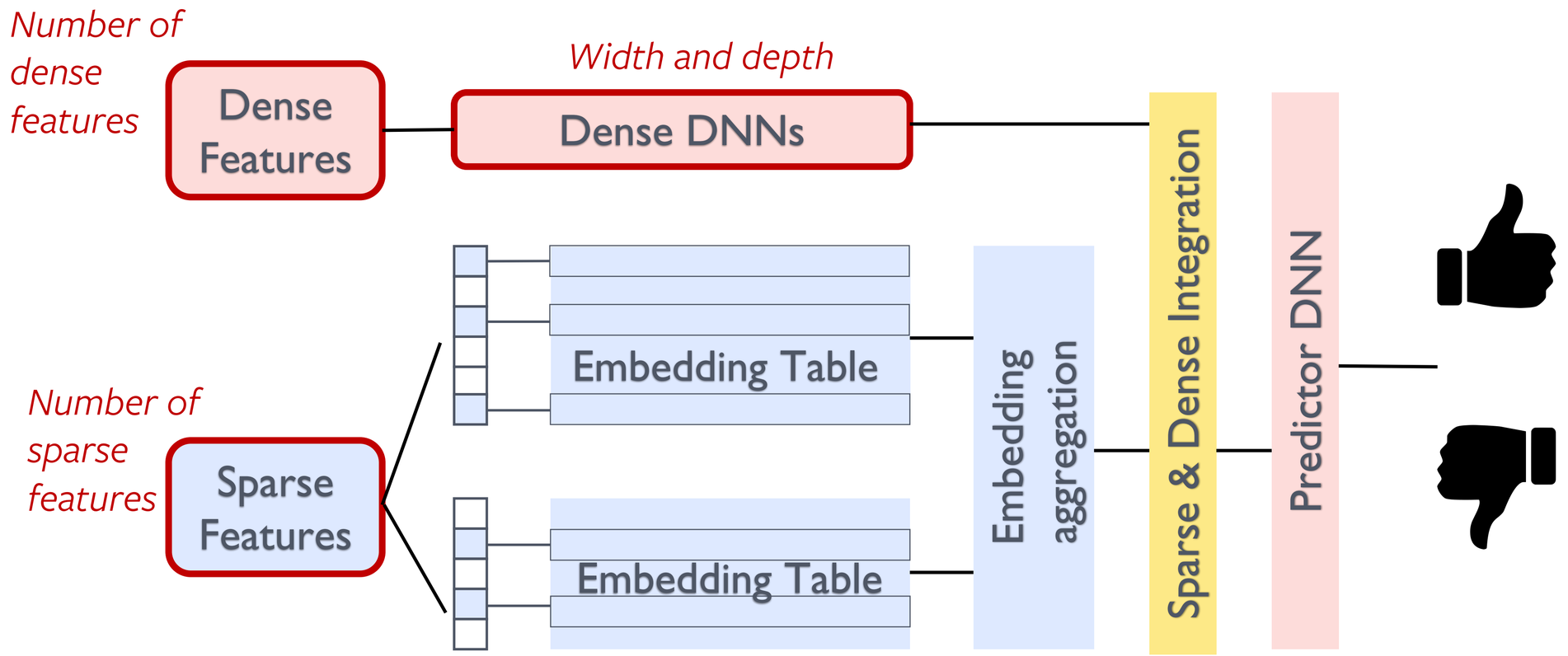

"Deep Learning Recommendation Model for Personalization and Recommendation Systems" Naumov, et. al. (https://arxiv.org/abs/1906.00091, https://github.com/facebookresearch/dlrm)

# DLRM: Configurable benchmark for end to end models



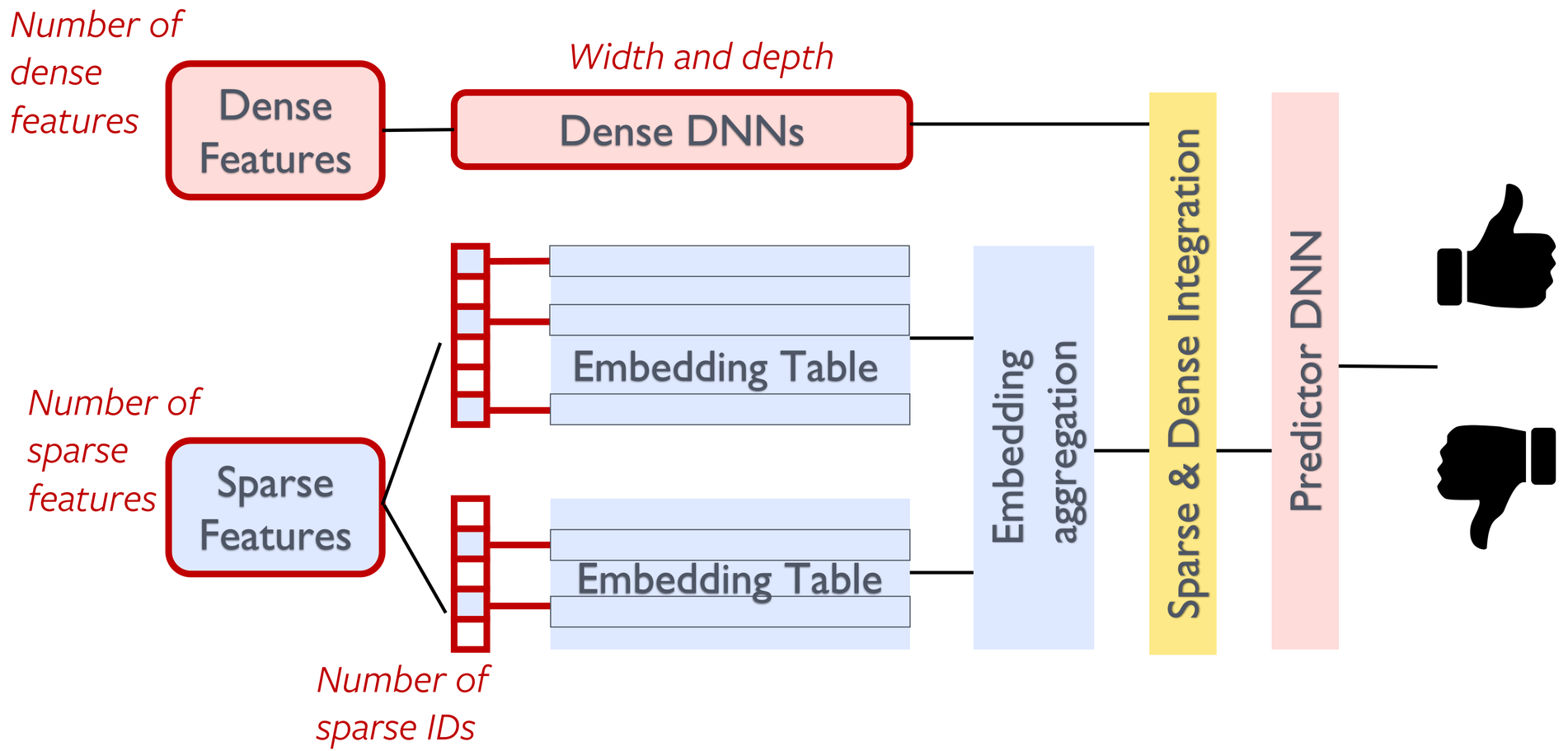

"Deep Learning Recommendation Model for Personalization and Recommendation Systems" Naumov, et. al. (https://arxiv.org/abs/1906.00091, https://github.com/facebookresearch/dlrm)

# DLRM: Configurable benchmark for end to end models



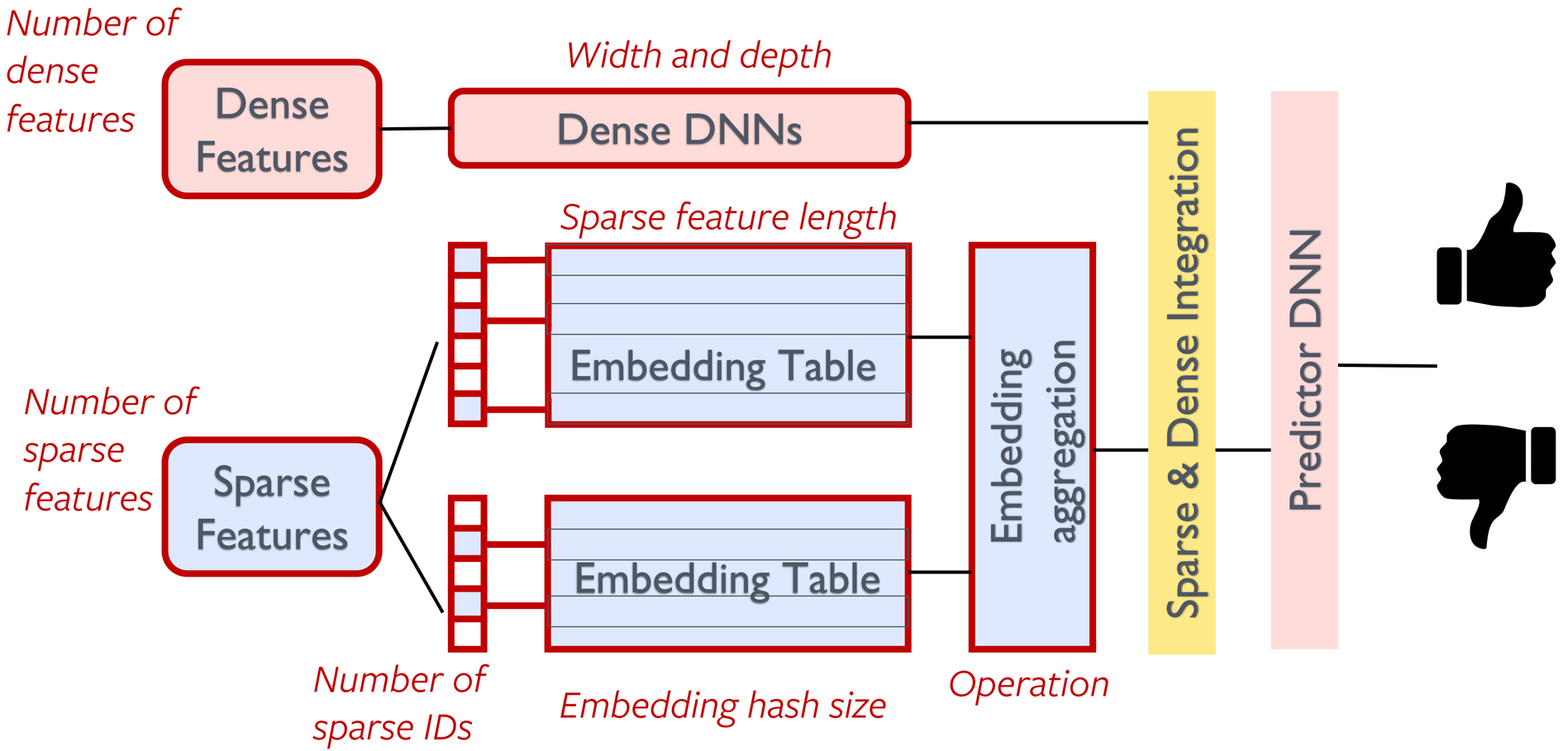

"Deep Learning Recommendation Model for Personalization and Recommendation Systems" Naumov, et. al. (https://arxiv.org/abs/1906.00091, https://github.com/facebookresearch/dlrm)

# DLRM: Configurable benchmark for end to end models



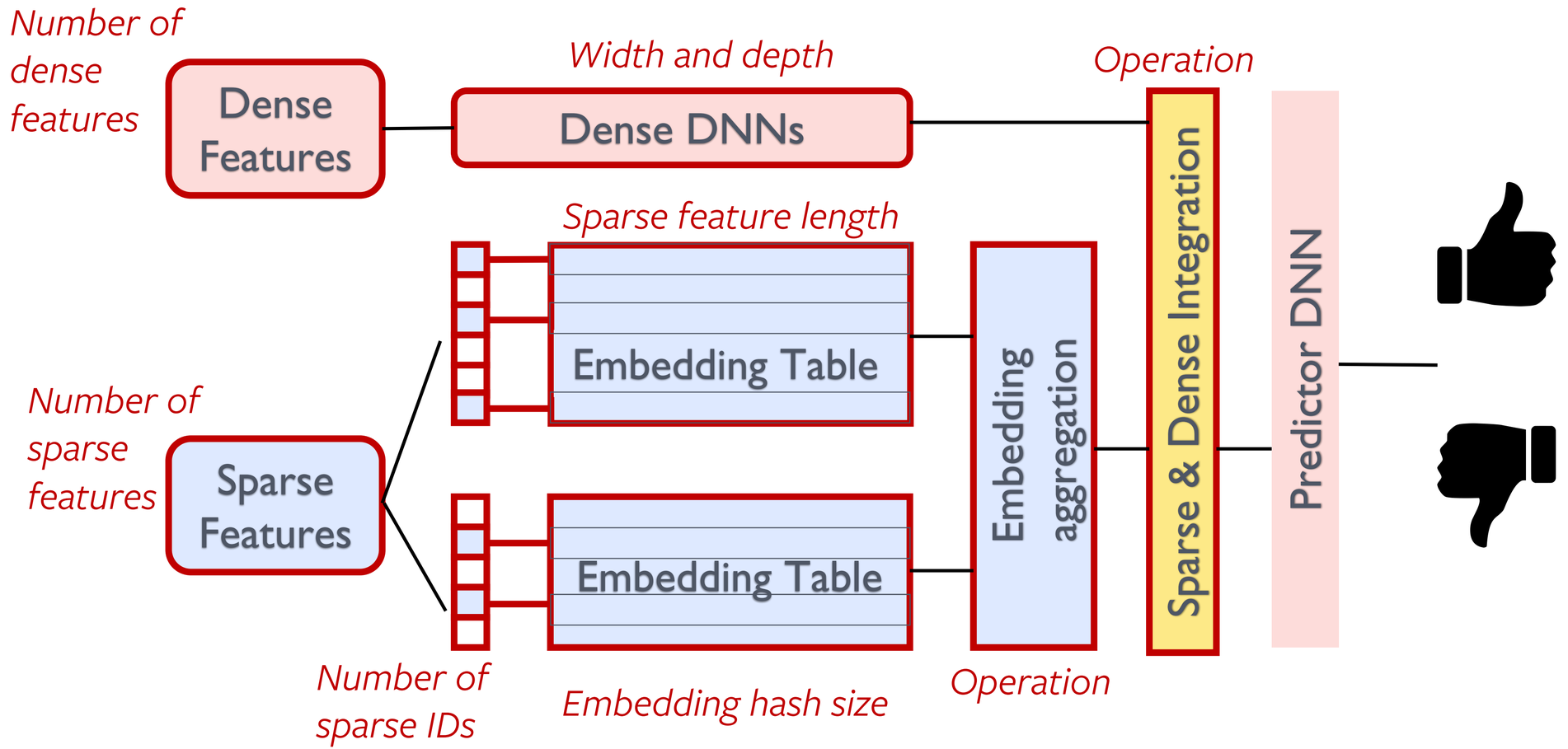

"Deep Learning Recommendation Model for Personalization and Recommendation Systems" Naumov, et. al. (https://arxiv.org/abs/1906.00091, https://github.com/facebookresearch/dlrm)

# DLRM: Configurable benchmark for end to end models



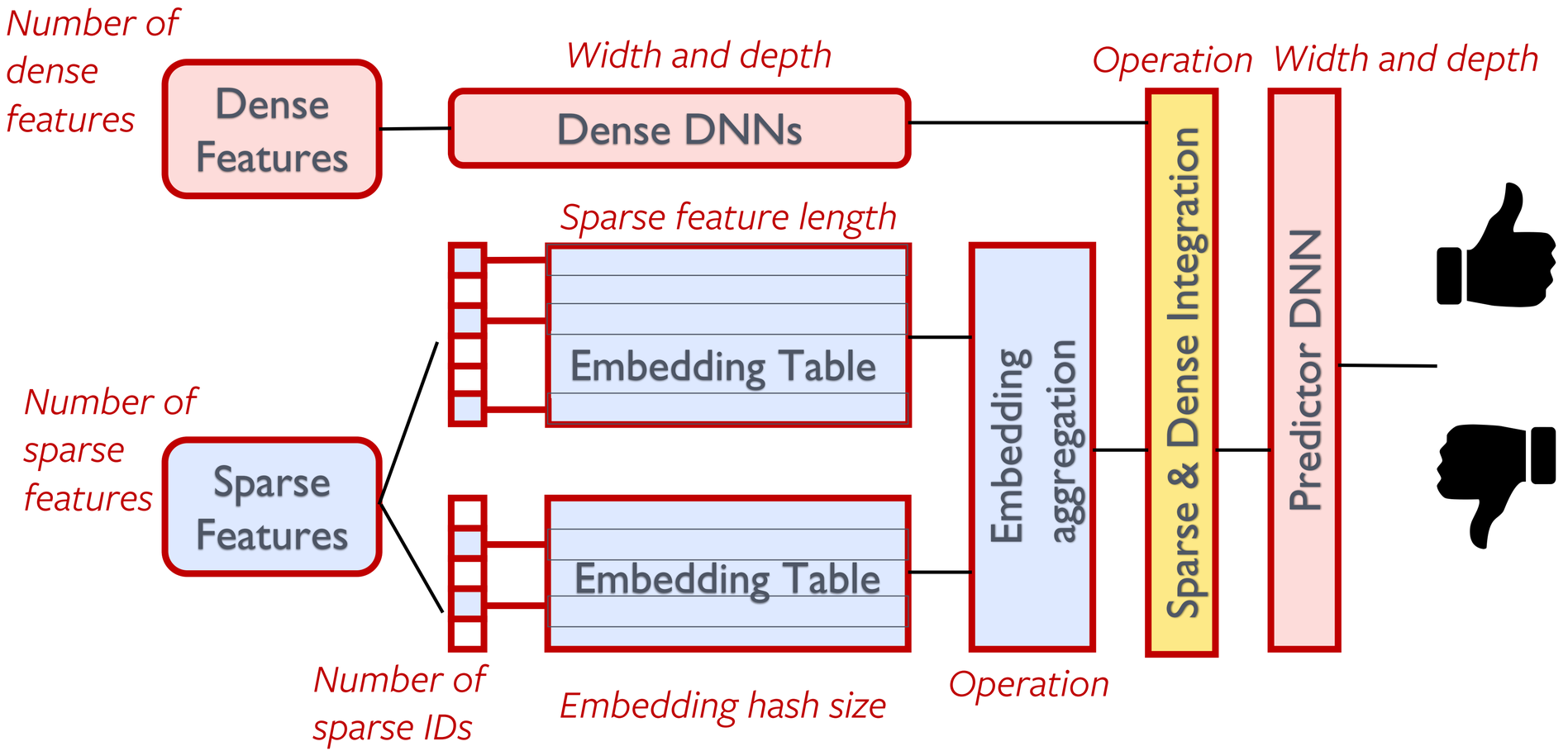

"Deep Learning Recommendation Model for Personalization and Recommendation Systems" Naumov, et. al. (https://arxiv.org/abs/1906.00091, https://github.com/facebookresearch/dlrm)

# DLRM: Configurable benchmark for end to end models




"Deep Learning Recommendation Model for Personalization and Recommendation Systems" Naumov, et. al. (https://arxiv.org/abs/1906.00091, https://github.com/facebookresearch/dlrm)

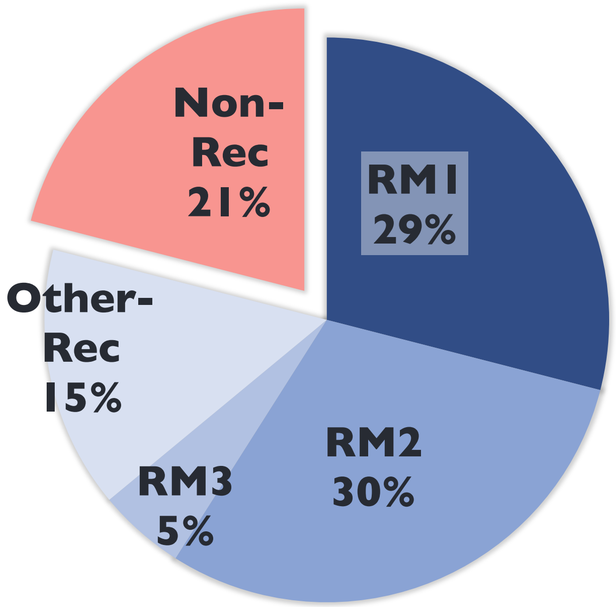
# DLRM: Configurable benchmark for end to end models

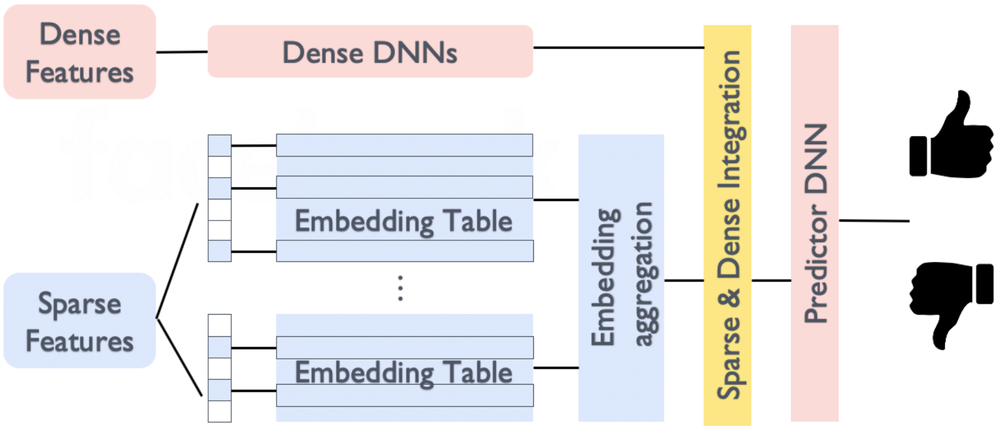



"Deep Learning Recommendation Model for Personalization and Recommendation Systems" Naumov, et. al. (https://arxiv.org/abs/1906.00091, https://github.com/facebookresearch/dlrm)

# Benchmarks represent key models in Facebook's datacenter

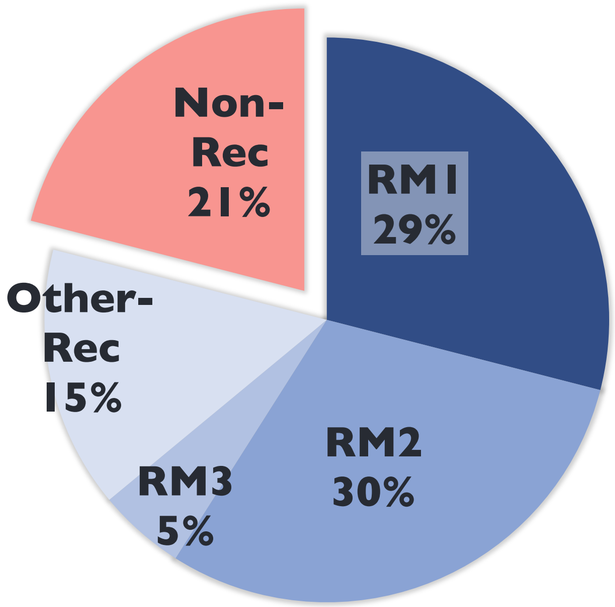
AI inference cycles in Facebook's datacenter

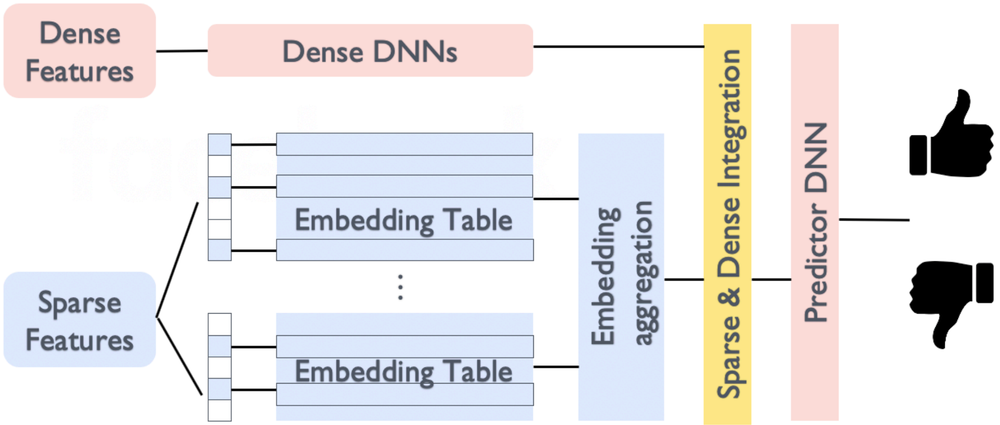


# Benchmarks represent key models in Facebook's datacenter

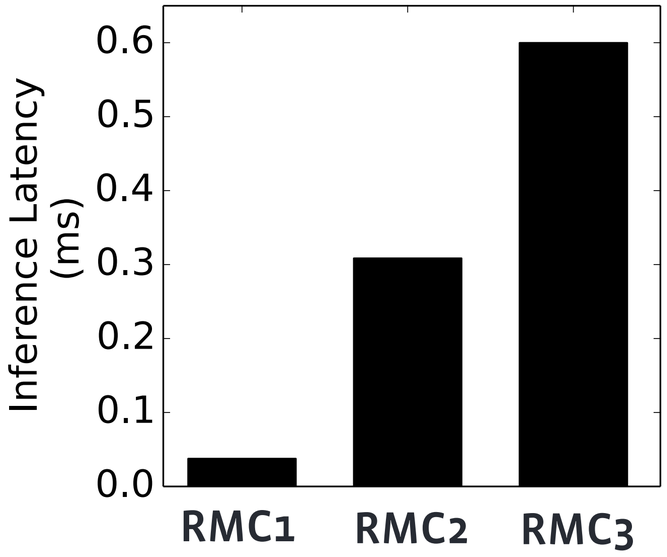
AI inference cycles in Facebook's datacenter



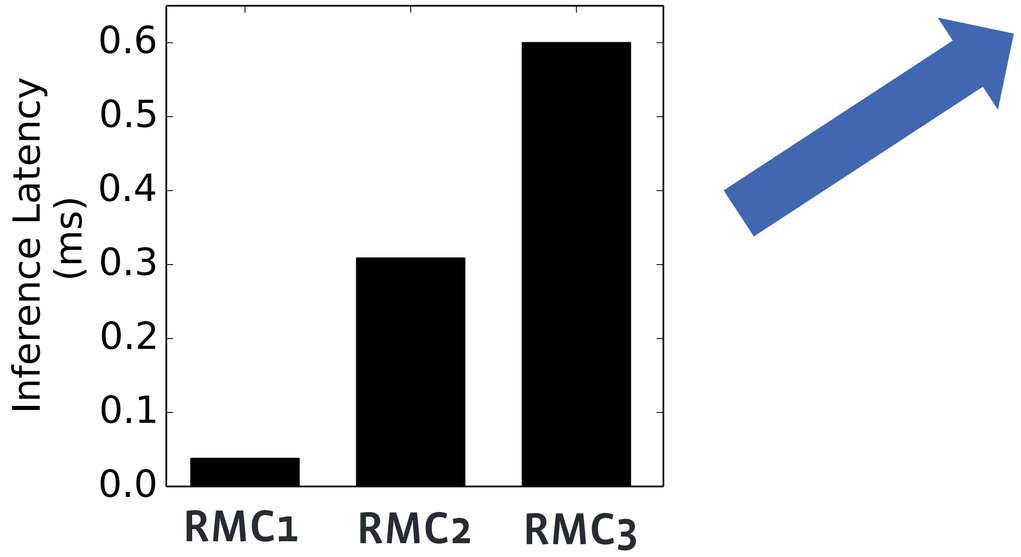



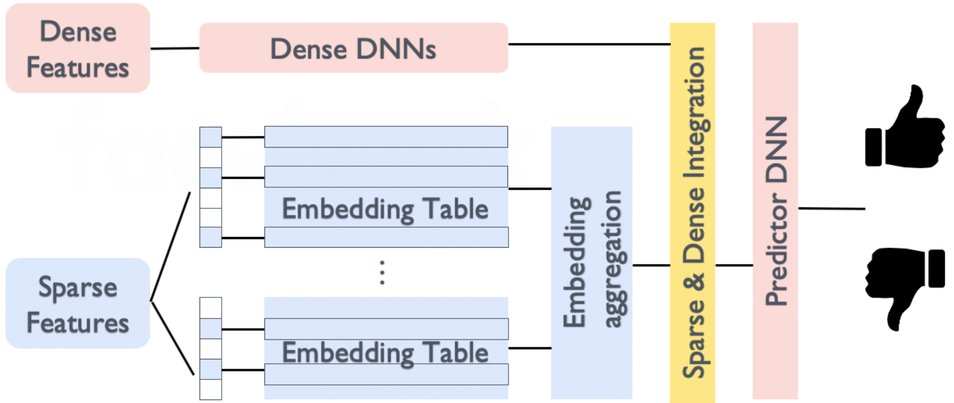

| | RM1 | RM2 | RM3 |
|---|---|---|---|
| FC sizes | Small | Medium | Large |
| Number of embedding tables | O(10) | O(50) | O(10) |
| Size of embeddings | Small | Medium | Large |
| Number of lookups per table | O(100) | O(100) | O(10) |

# Diverse solutions are needed to optimize recommendation

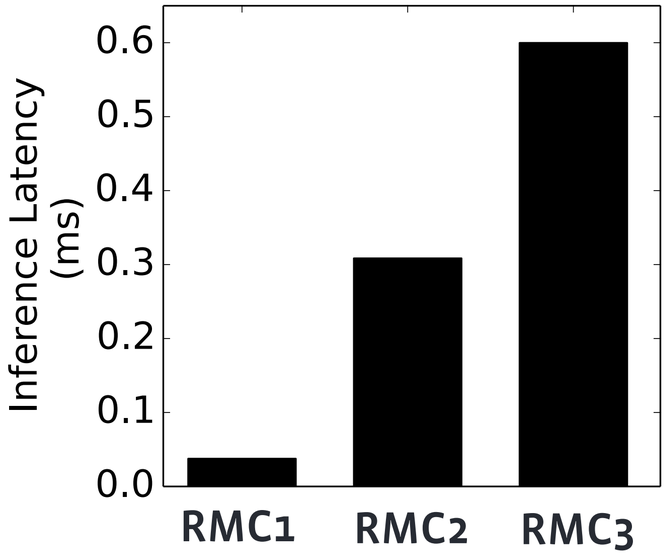


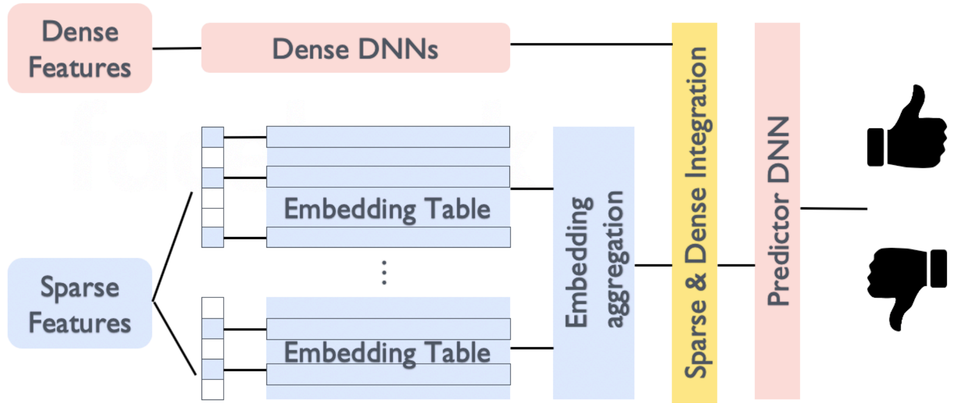
# Diverse solutions are needed to optimize recommendation

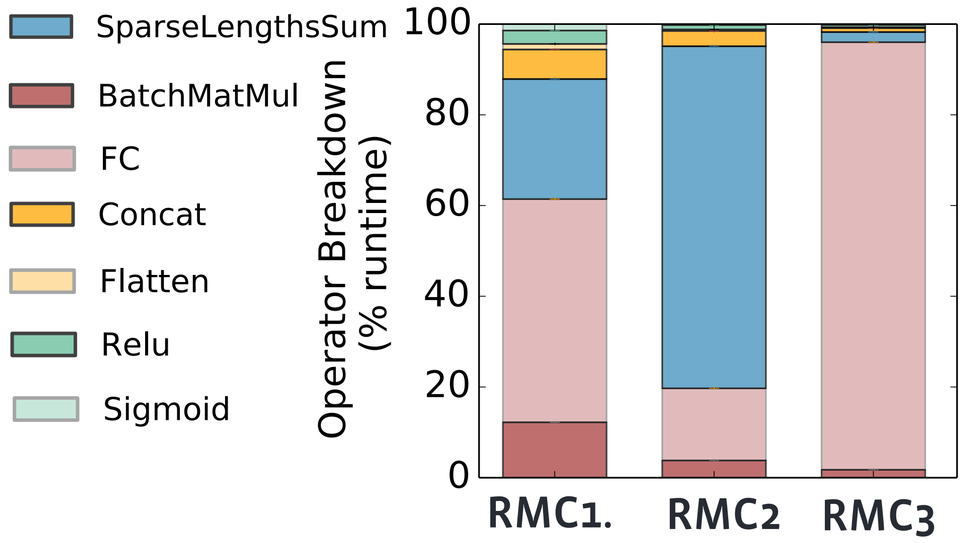


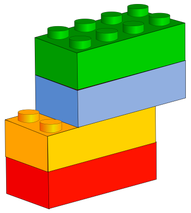
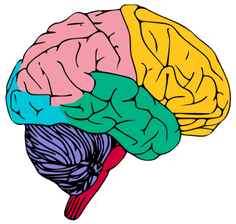
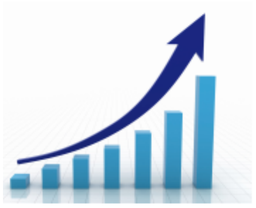
# Diverse solutions are needed to optimize recommendation



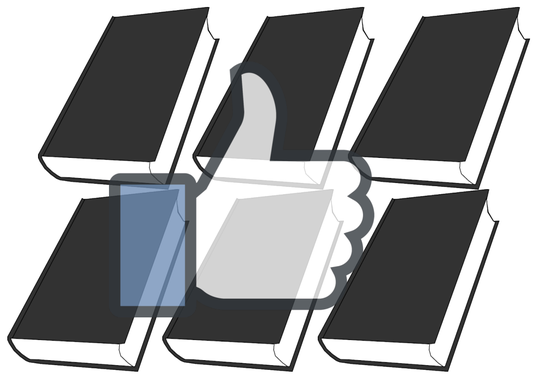



# Hardware insights of recommendation

| Algorithmic | Hardware |
| --- | --- |
|  General model structure | Requires optimizing operators with new storage, compute, and memory access requirements |
|  Diverse model architectures | Accelerating recommendation needs flexible and diverse system solutions |
|  Processing queries at-scale | Exploiting hardware heterogeneity and parallelism can optimize latency-bounded throughput |

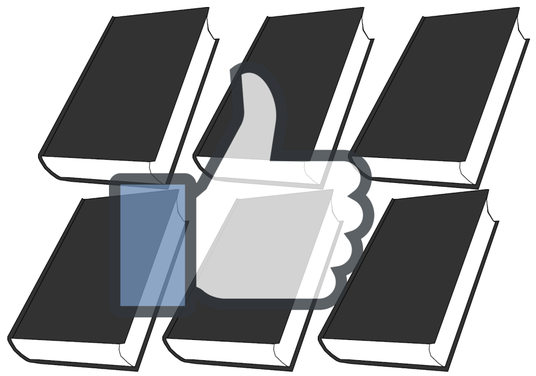
# Ranking more items leads to better recommendation quality

High throughput!



# Ranking more items leads to better recommendation quality

High throughput!



Low latency!

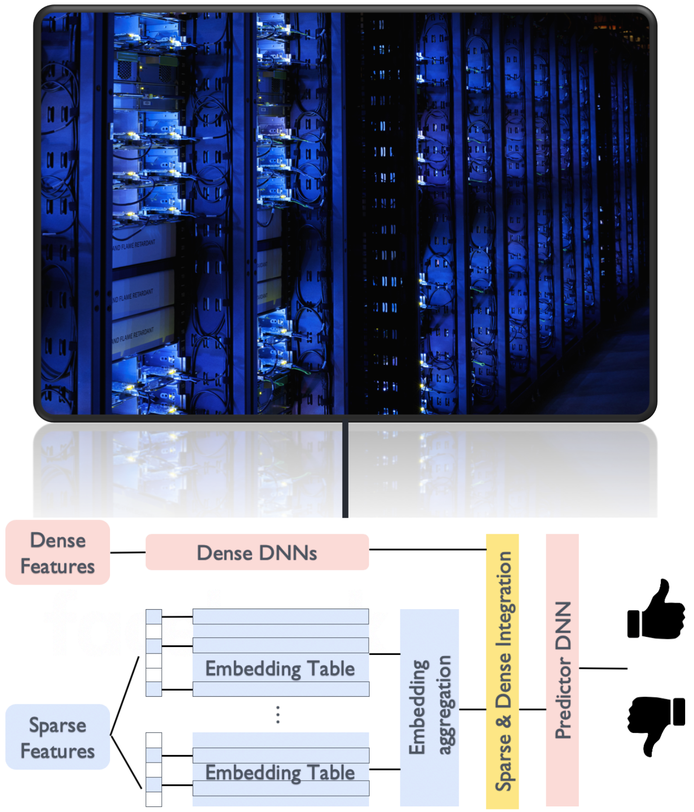
# Ranking more items leads to better recommendation quality

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

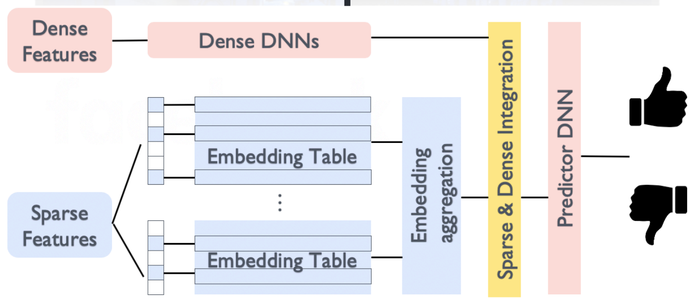
2.4 GHz
AVX-2
Inclusive L2/L3 caches
DDR4

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

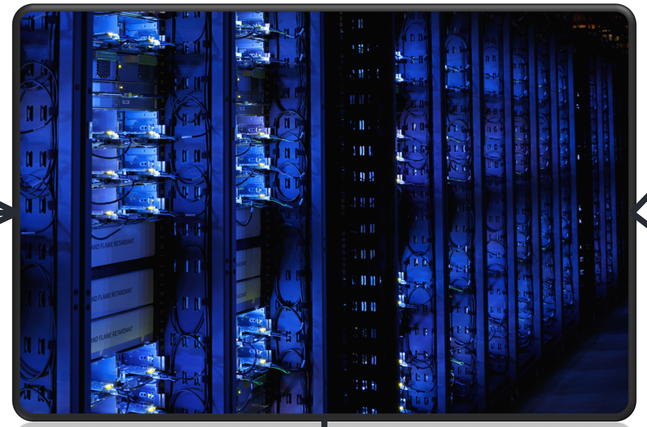
2.4 GHz
AVX-2
Inclusive L2/L3 caches
DDR4

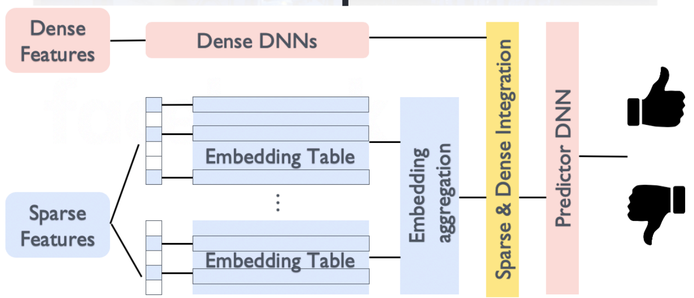
Broadwell

2.0 GHz
AVX-512
Exclusive L2/L3 caches
DDR4

Skylake

Hardware






Models

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

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

Parallelization

# Characterizing latency bounded throughput design space

2.5 GHz
AVX-2
Inclusive L2/L3 caches
DDR3

Haswell

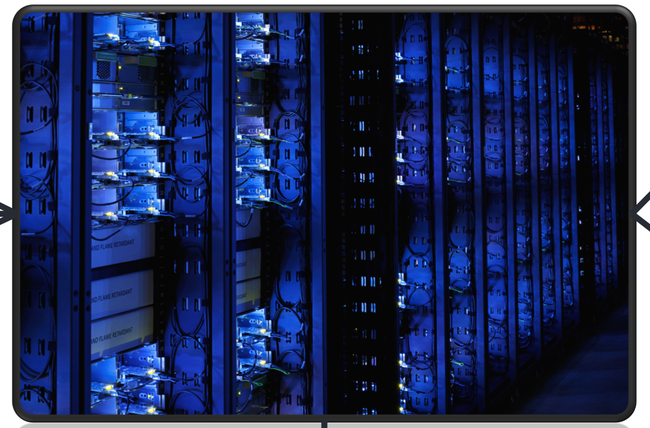
2.4 GHz
AVX-2
Inclusive L2/L3 caches
DDR4

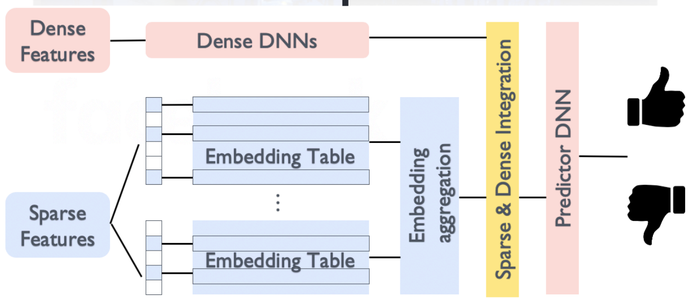
Broadwell

2.0 GHz
AVX-512
Exclusive L2/L3 caches
DDR4

Skylake

Hardware





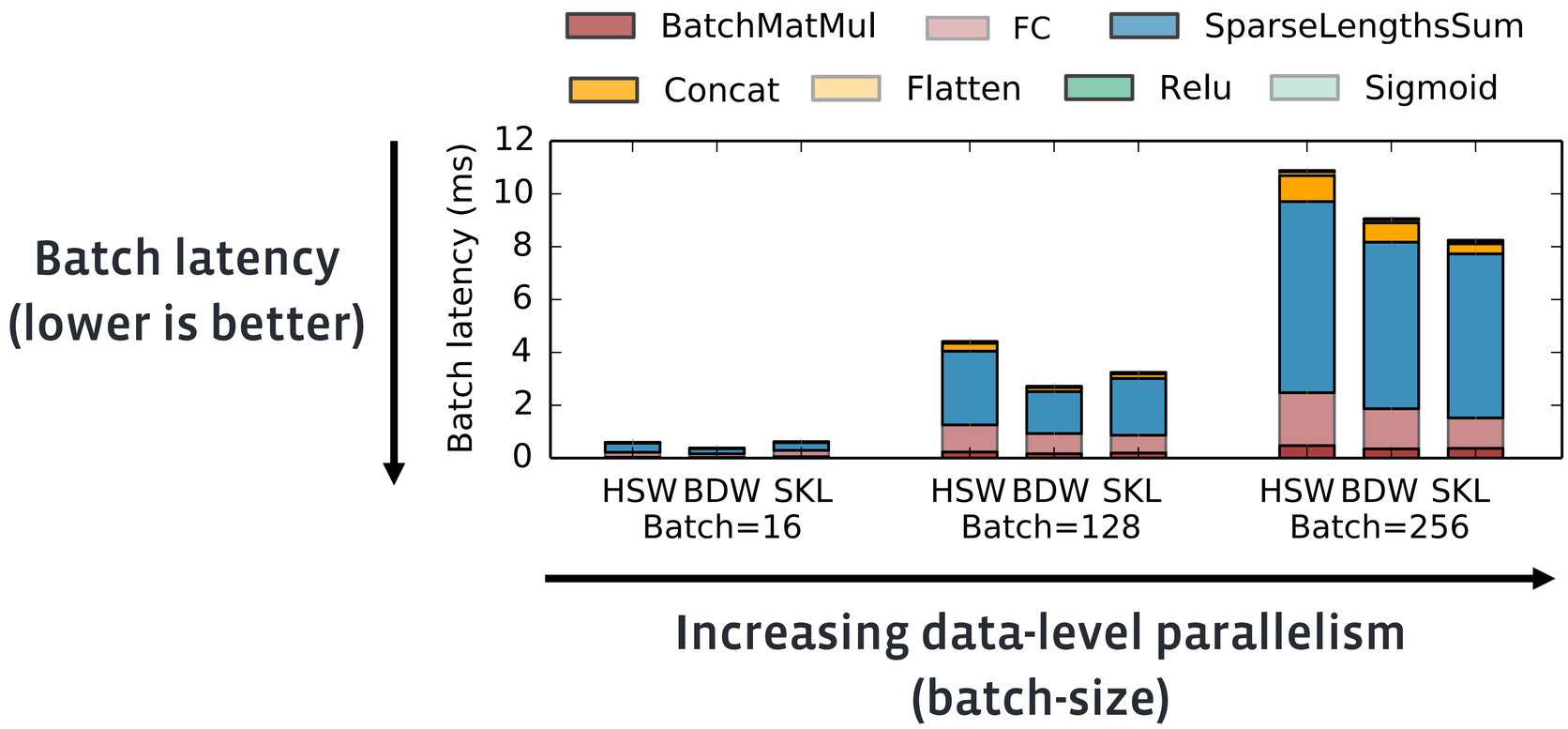

Models

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

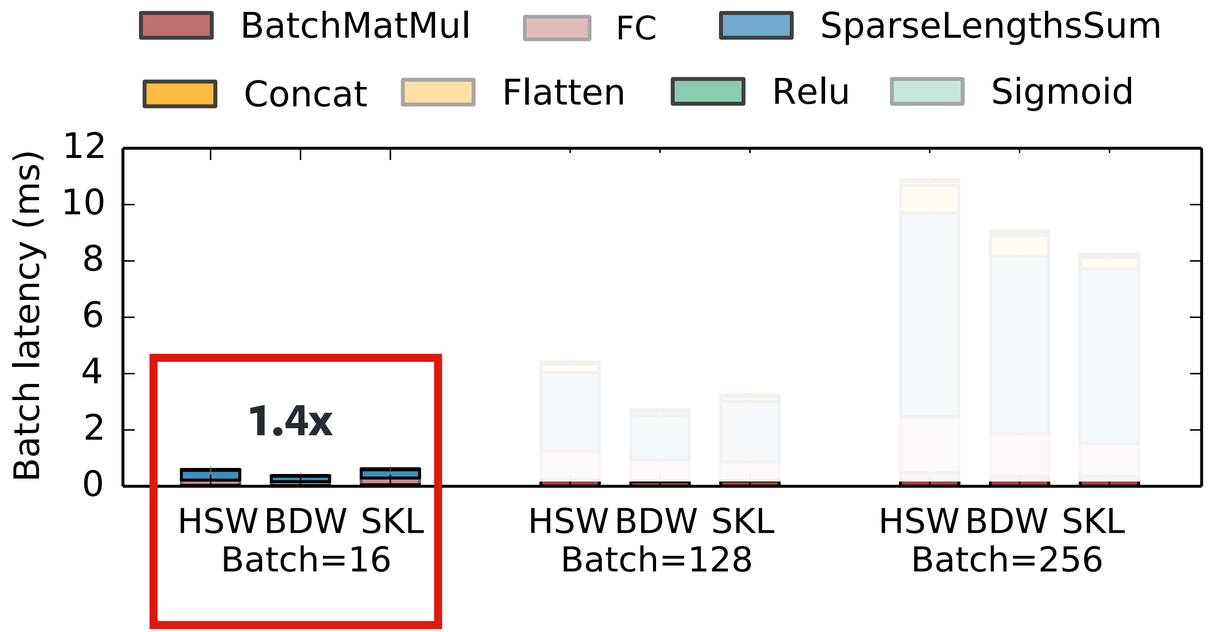
Task level parallelism
(i.e., co-locating models)

Parallelization

# Data parallelism: Characterizing latency bounded throughput design space

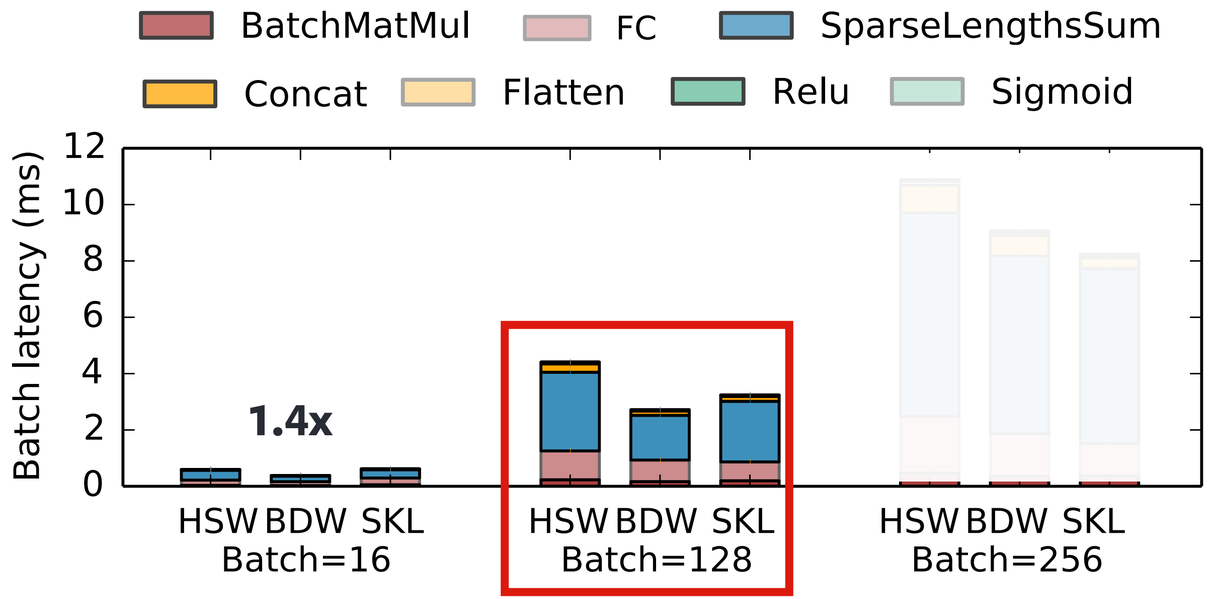


# Data parallelism: Characterizing latency bounded throughput design space



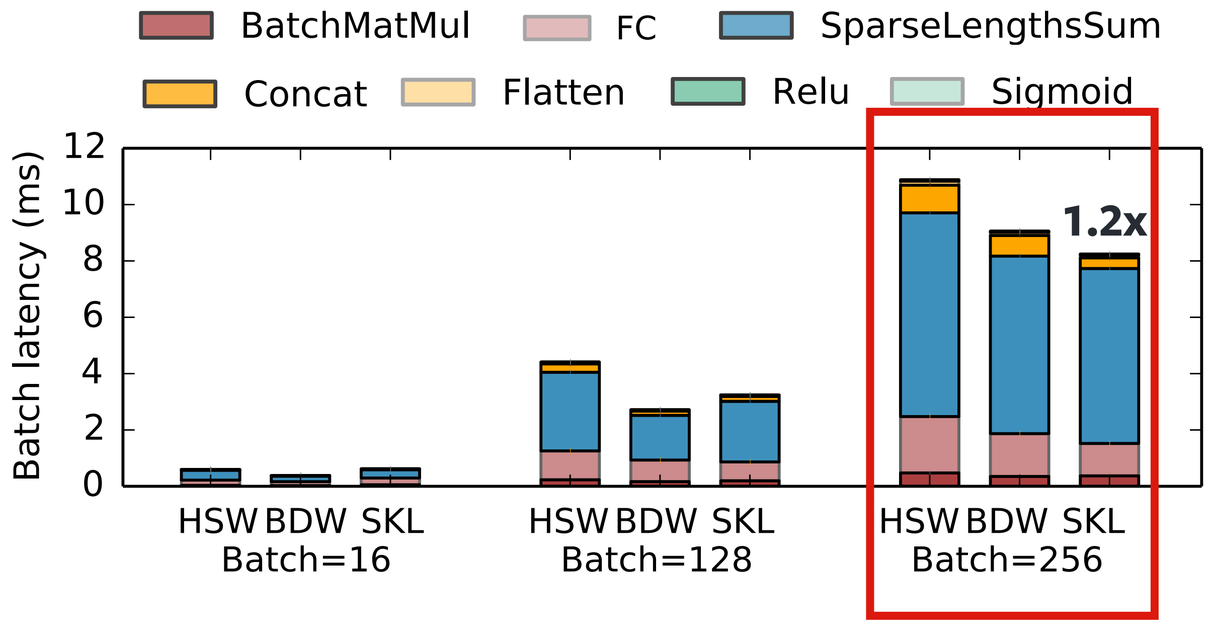

- At smaller batch-sizes Broadwell has 1.4x lower batch latency
  - Skylake: 20% lower CPU frequency and lower AVX-512 utilization (70%)

# Data parallelism: Characterizing latency bounded throughput design space



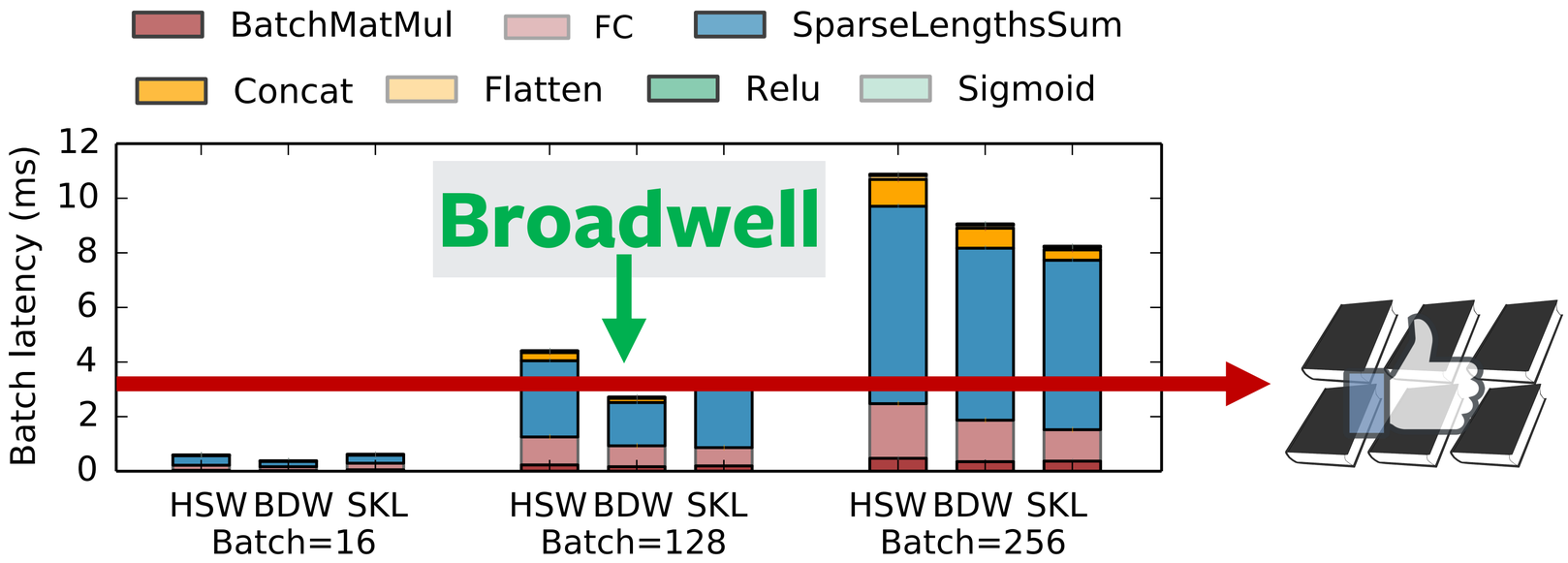

- At smaller batch-sizes Broadwell has 1.4x lower batch latency
  - Haswell: 50% lower DRAM frequency

# Data parallelism: Characterizing latency bounded throughput design space



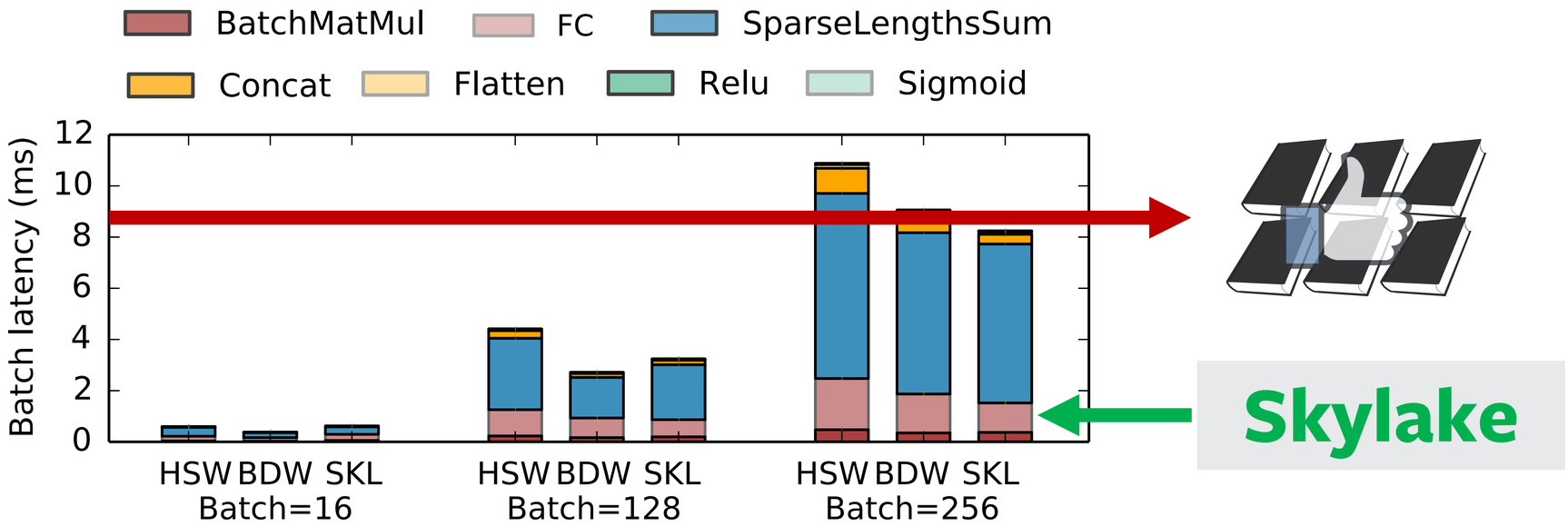

- At higher batch-sizes Skylake has lower batch latency
  - Wider vector width and higher AVX-512 utilization (90%)

# Data parallelism: Characterizing latency bounded throughput design space




- 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




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

# Characterizing latency bounded throughput design space

2.5 GHz
AVX-2
Inclusive L2/L3 caches
DDR3

Haswell

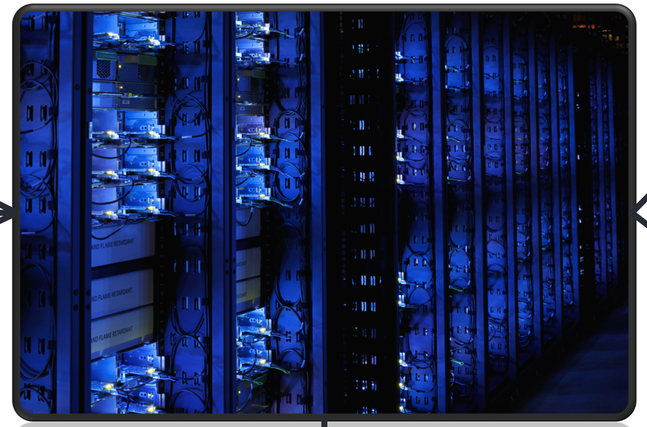
2.4 GHz
AVX-2
Inclusive L2/L3 caches
DDR4

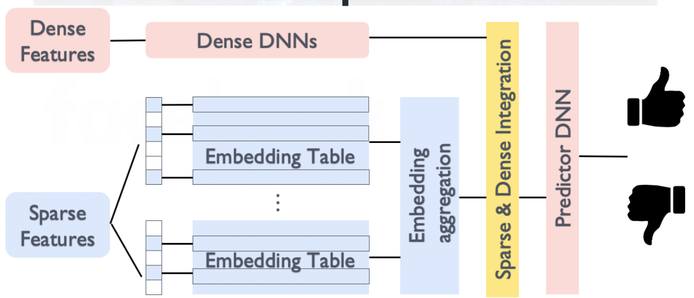
Broadwell

2.0 GHz
AVX-512
Exclusive L2/L3 caches
DDR4

Skylake

Hardware






Models

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

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

Parallelization

# Co-locating models improves recommendation quality and reduces infrastructure capacity

Latency and batch critical application

Latency critical application

Latency critical application

**Target latency**

# Co-locating models improves recommendation quality and reduces infrastructure capacity

Latency and batch critical application

Latency critical application

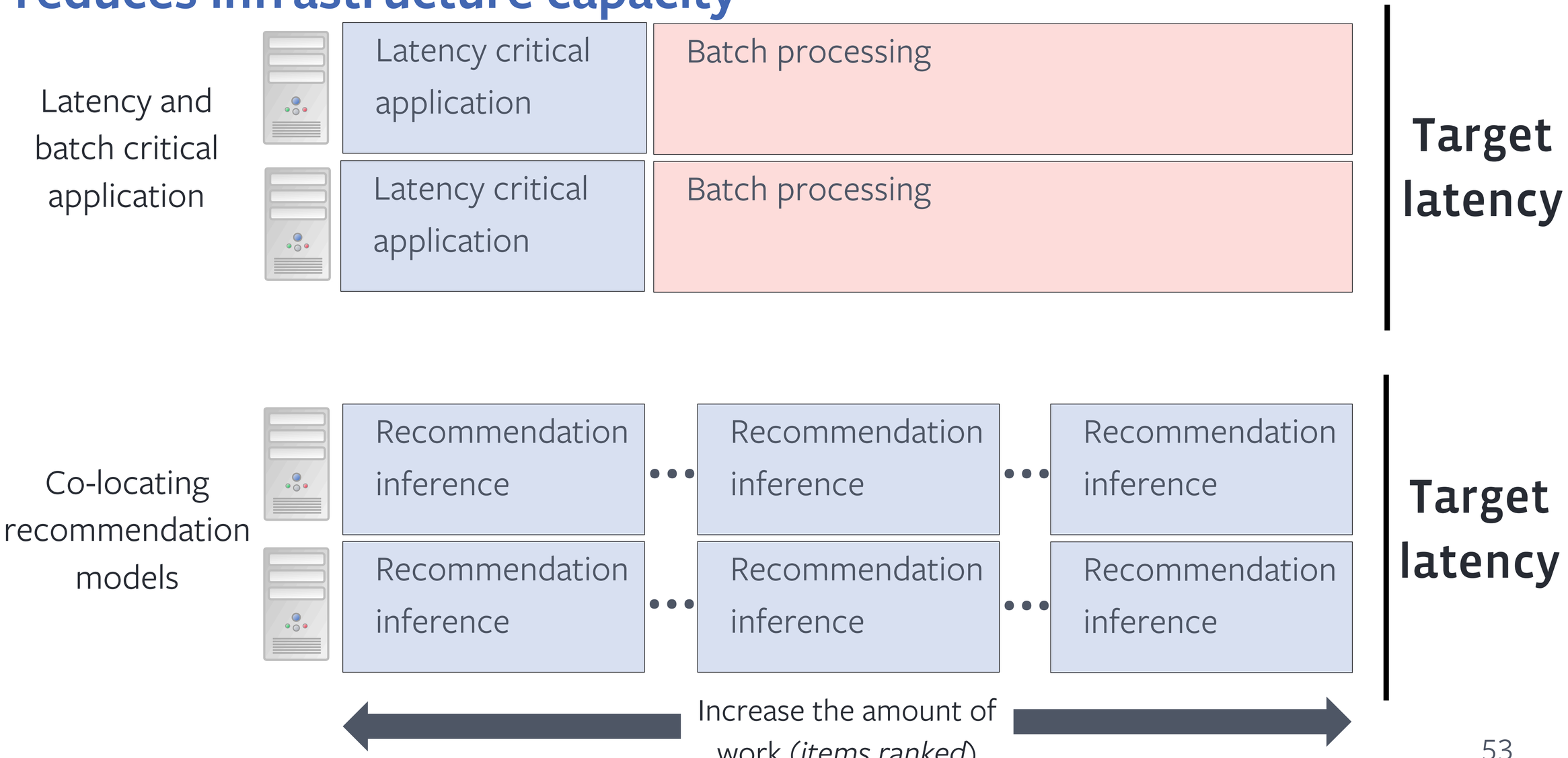
Batch processing

Latency critical application

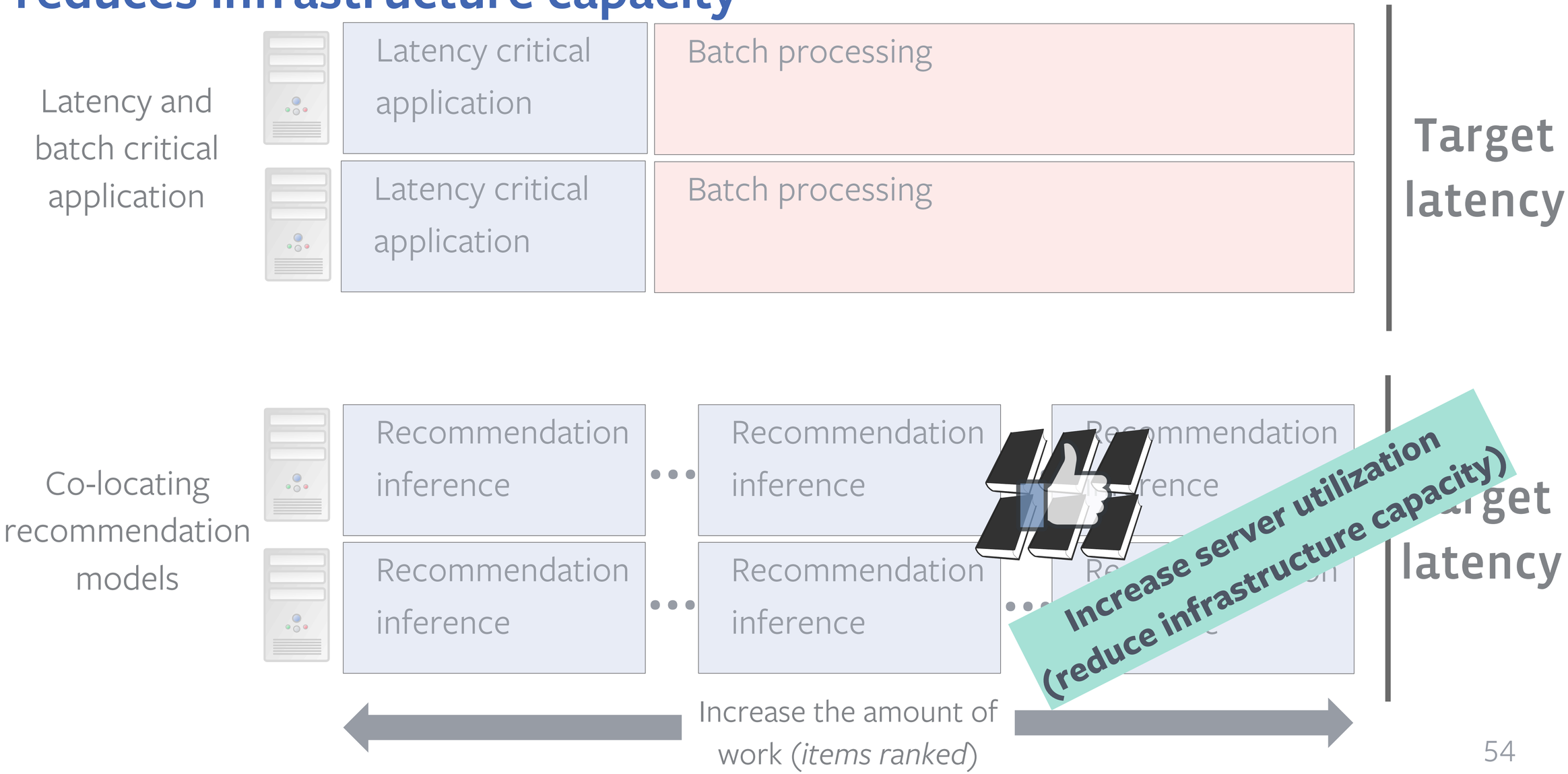
Batch processing

**Target latency**

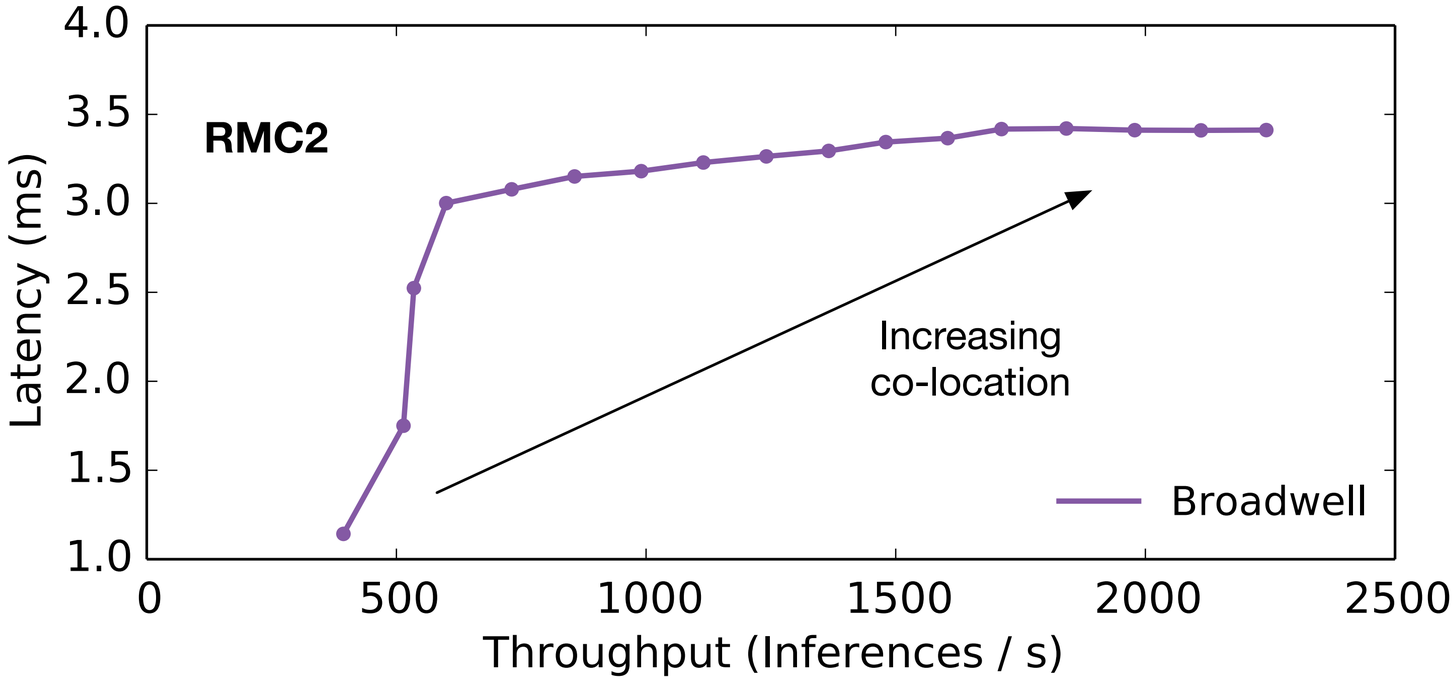
# Co-locating models improves recommendation quality and reduces infrastructure capacity



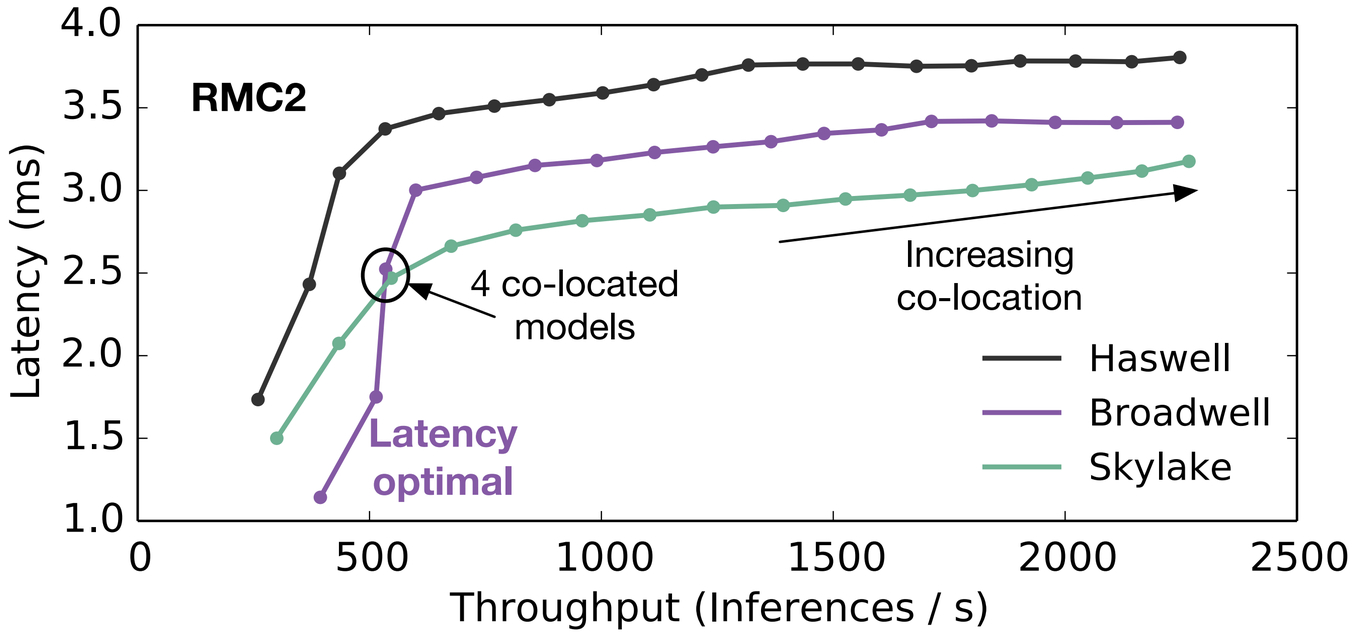
# Co-locating models improves recommendation quality and reduces infrastructure capacity



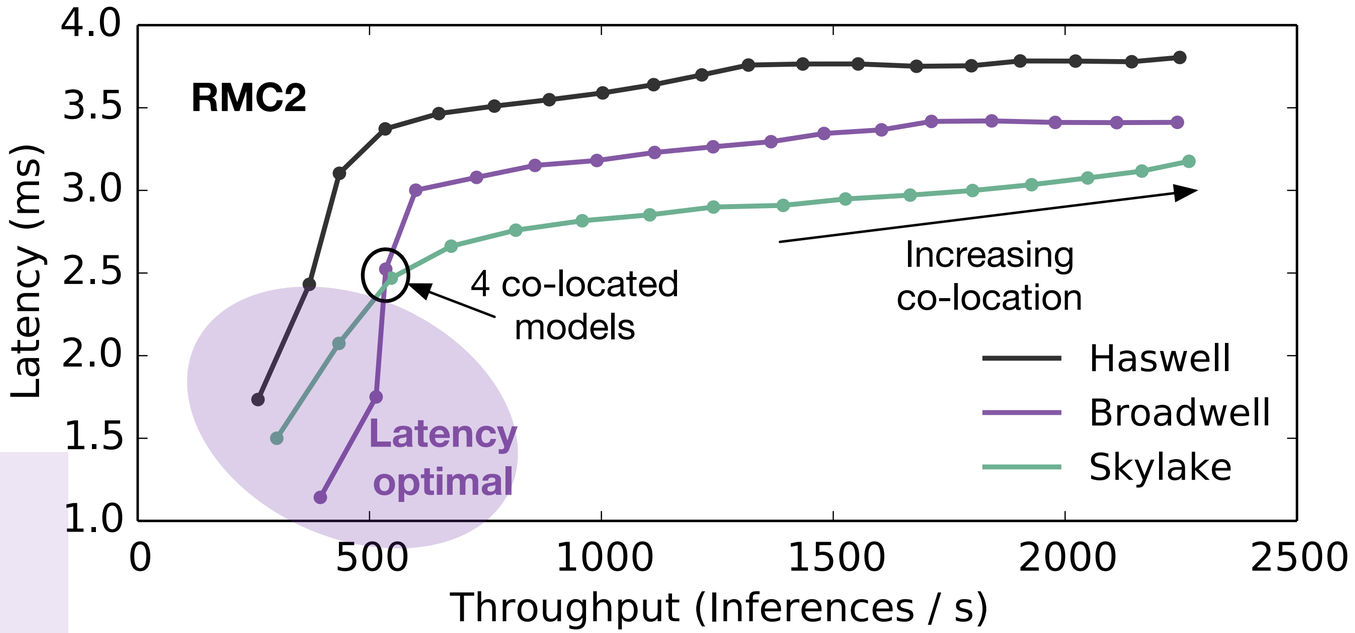
# Task parallelism: Characterizing latency bounded throughput



# Task parallelism: Characterizing latency bounded throughput



# Task parallelism: Characterizing latency bounded throughput



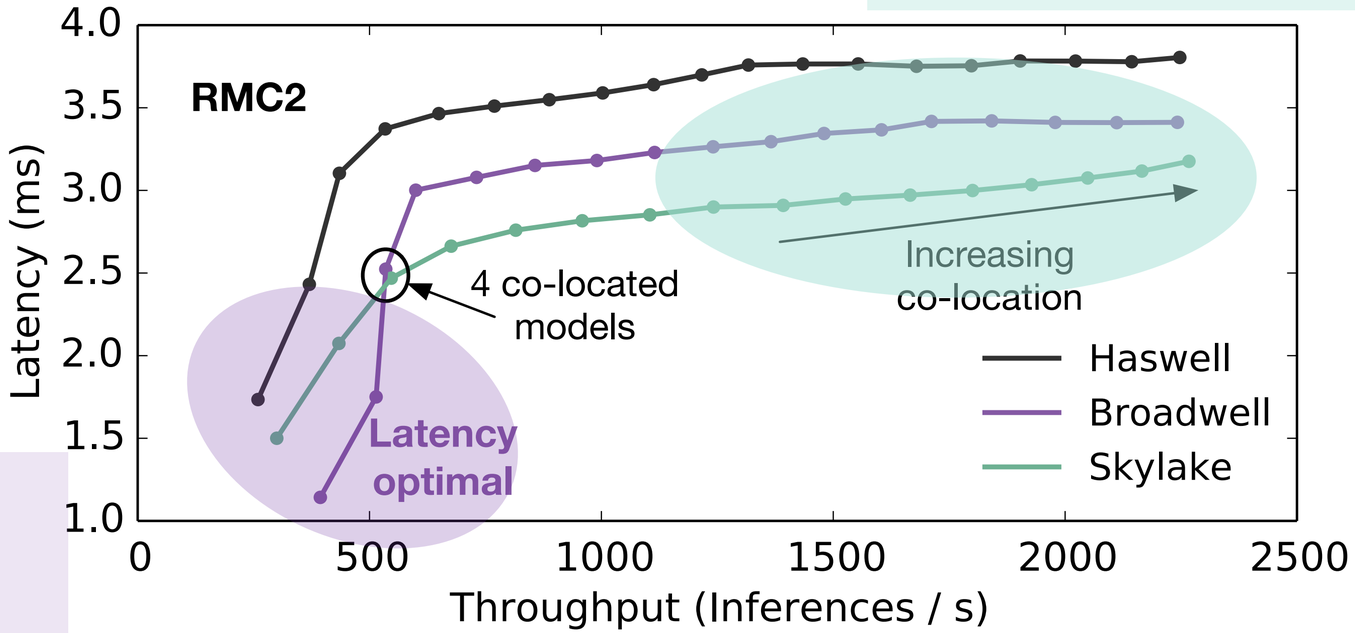

Broadwell is latency optimal

- Higher CPU frequency
- Inclusive L2/L3 caches

# Task parallelism: Characterizing latency bounded throughput

Skylake is throughput optimal

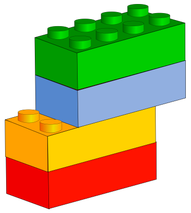
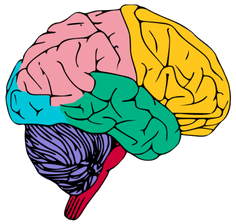
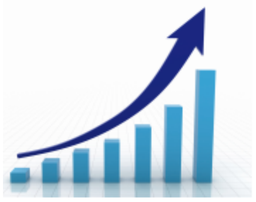
- Wider AVX width
- Exclusive L2/L3 caches




Broadwell is latency optimal

- Higher CPU frequency
- Inclusive L2/L3 caches

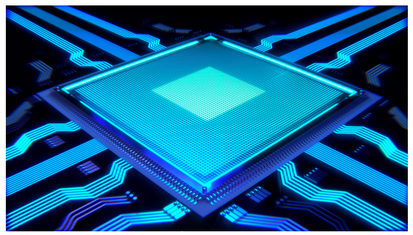
# Hardware insights of recommendation

| Algorithmic | Hardware |
| --- | --- |
|  General model structure | Requires optimizing operators with new storage, compute, and memory access requirements |
|  Diverse model architectures | Accelerating recommendation needs flexible and diverse system solutions |
|  Processing queries at-scale | Exploiting hardware heterogeneity and parallelism can optimize latency-bounded throughput |

**Hardware opportunities ahead**

# Hardware opportunities ahead

Hardware acceleration

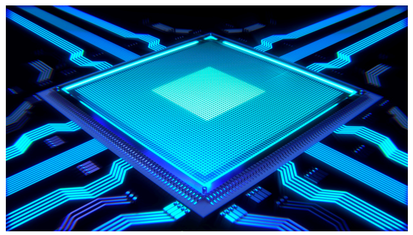


Evaluating current accelerator proposals

Designing new hardware solutions

# Hardware opportunities ahead

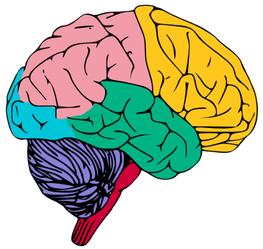
Hardware acceleration



Evaluating current accelerator proposals

Designing new hardware solutions

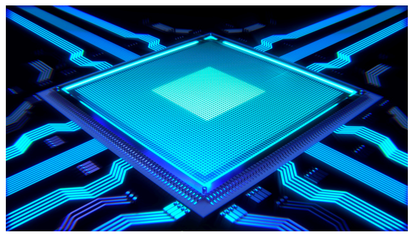
Model optimizations



Designing new compression methods (i.e., quantization)

# Hardware opportunities ahead

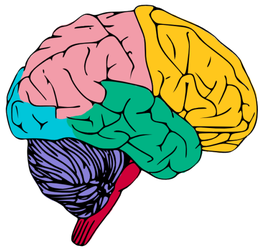
## Hardware acceleration



Evaluating current accelerator proposals

Designing new hardware solutions

## Model optimizations



Designing new compression methods (i.e., quantization)

## Large scale systems

Optimizing system level latency-bounded throughput

Performance variability

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

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

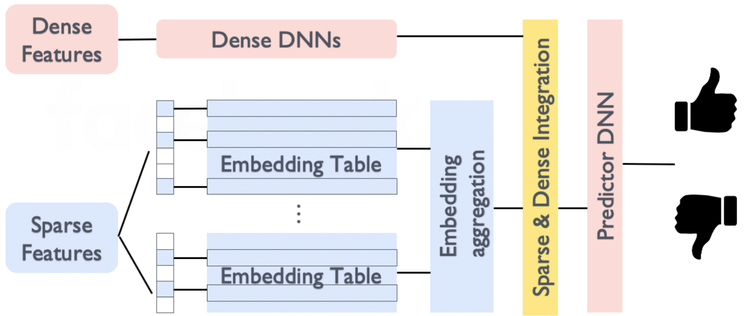
David Brooks, Bradford Cottel, Kim Hazelwood, Mark Hempstead, Bill Jia, Hsien-Hsin S. Lee, Andrey Malevich, Dheevatsa Mudigere, Mikhail Smelyanskiy, Liang Xiong, Xuan Zhang

DLRM (Deep learning recommendation model) is open source!

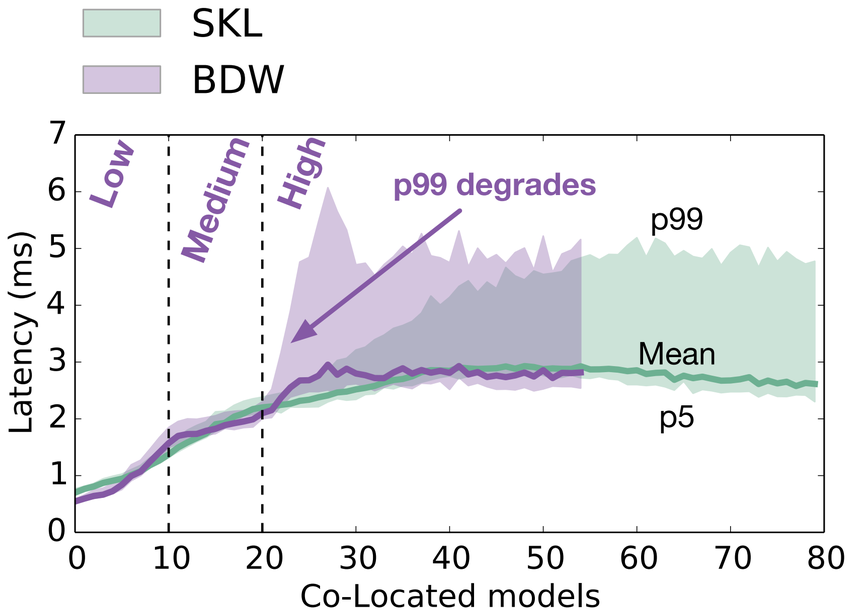
arXiv.org "Deep Learning Recommendation Model for Personalization and Recommendation Systems" (Naumov, et. al.)

https://github.com/facebookresearch/dlrm

MLPerf https://github.com/mlperf/training/tree/master/recommendation

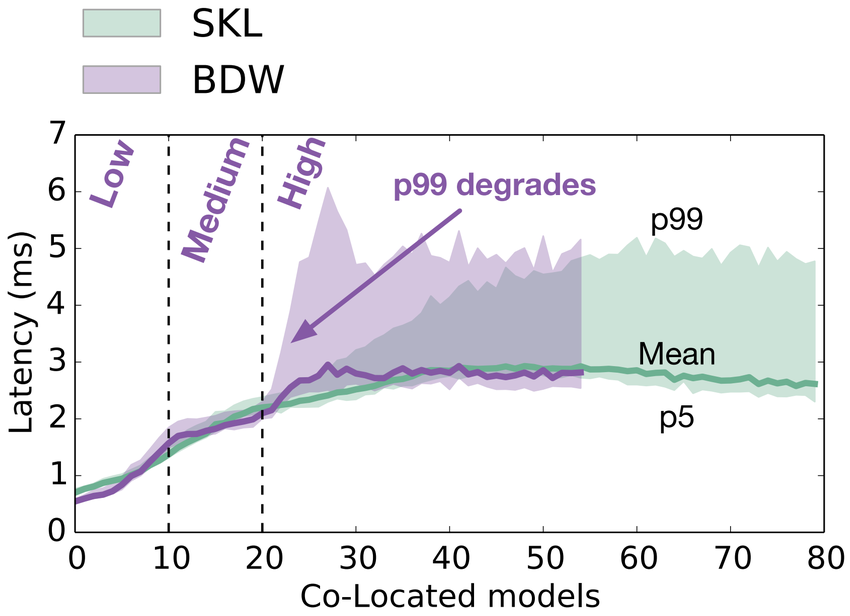


# Cost of co-locating models: Variability



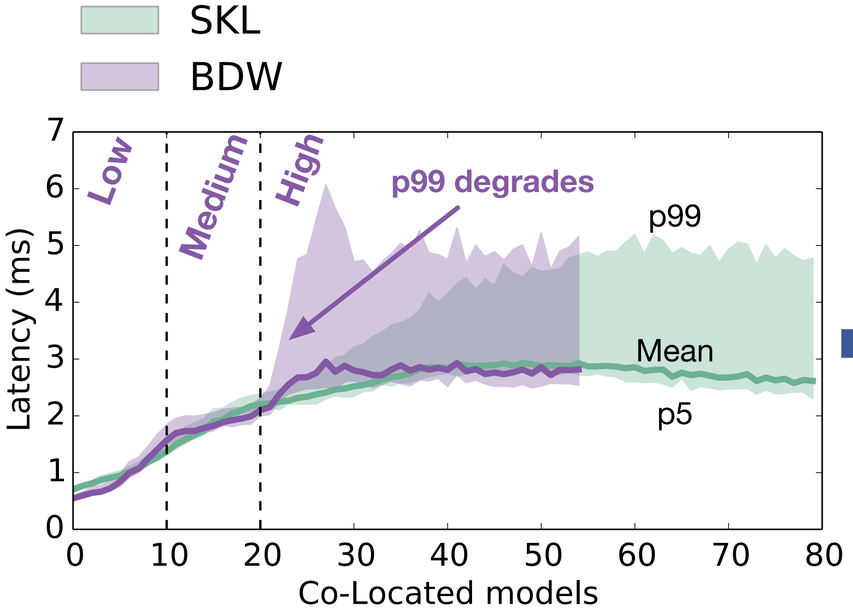

Broadwell and Skylake follow unique distribution as we increase degree of co-location

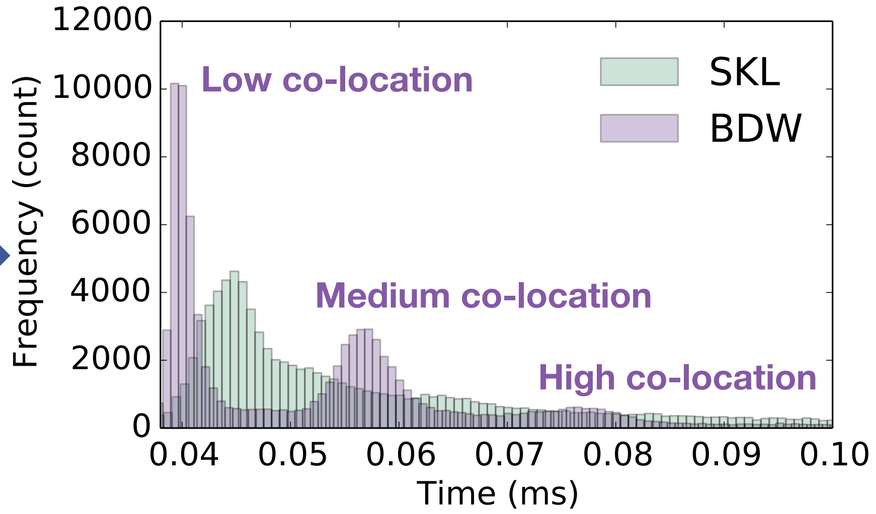
# Cost of co-locating models: Variability




Broadwell and Skylake follow unique distribution as we increase degree of co-location

# Cost of co-locating models: Variability




Broadwell and Skylake follow unique distribution as we increase degree of co-location

Distinct distributions found in production datacenters as well