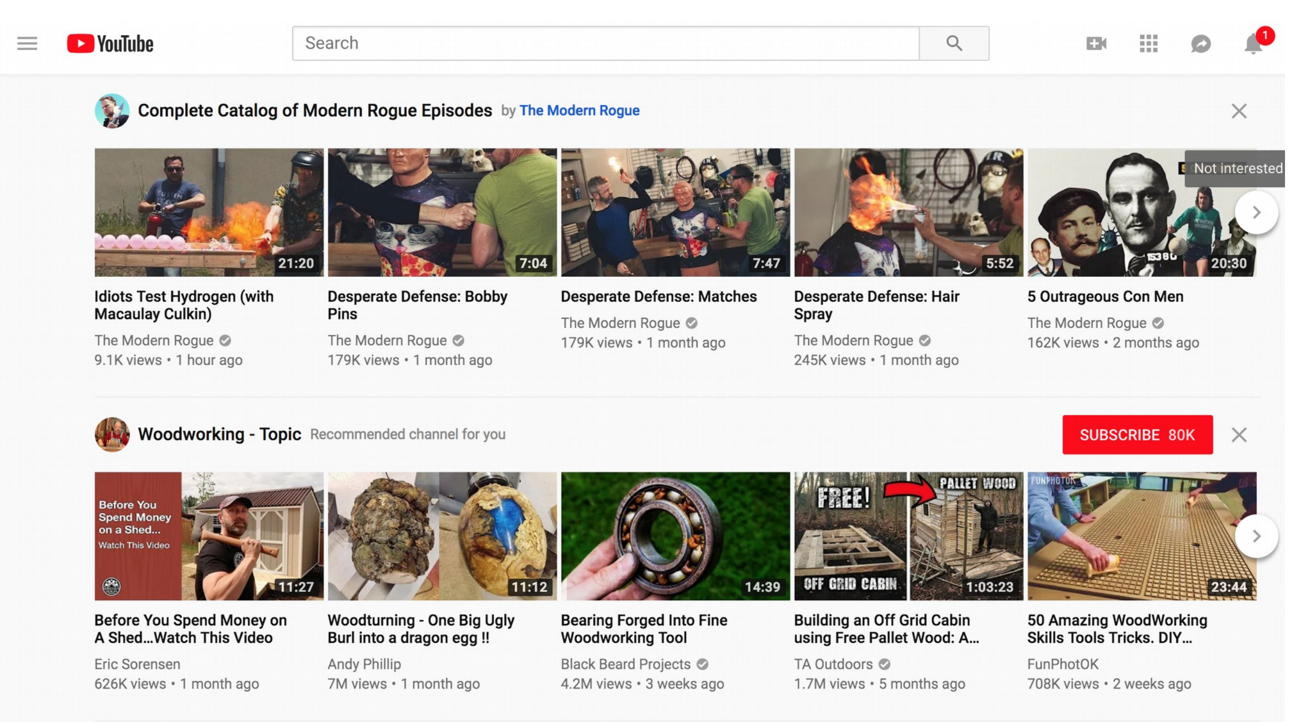
Cross-Stack Characterization and Solid State Drive-Based Near Data Processing for Recommendation Workloads

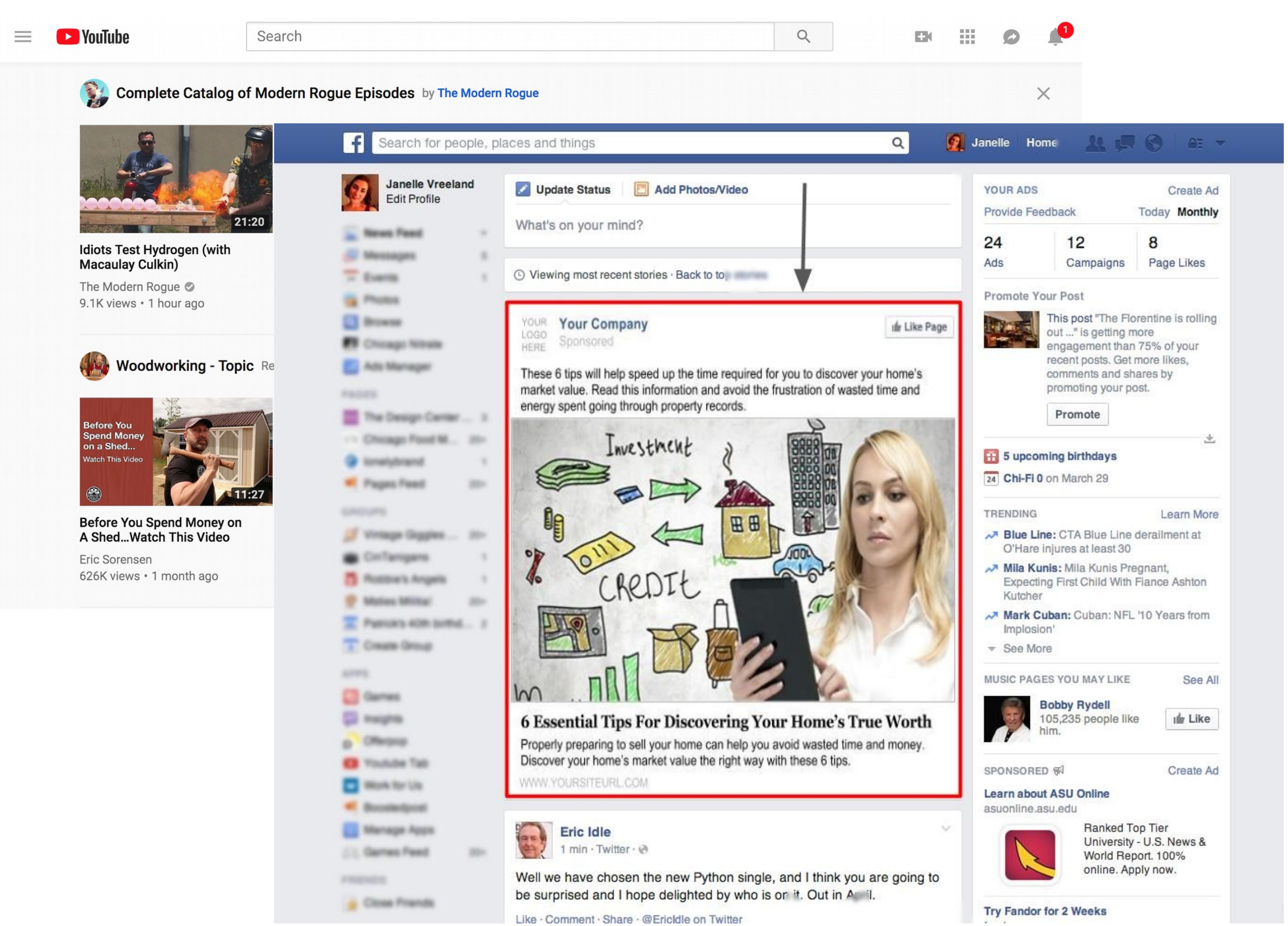
Samuel Hsia*, **Mark Wilkening***, Udit Gupta,

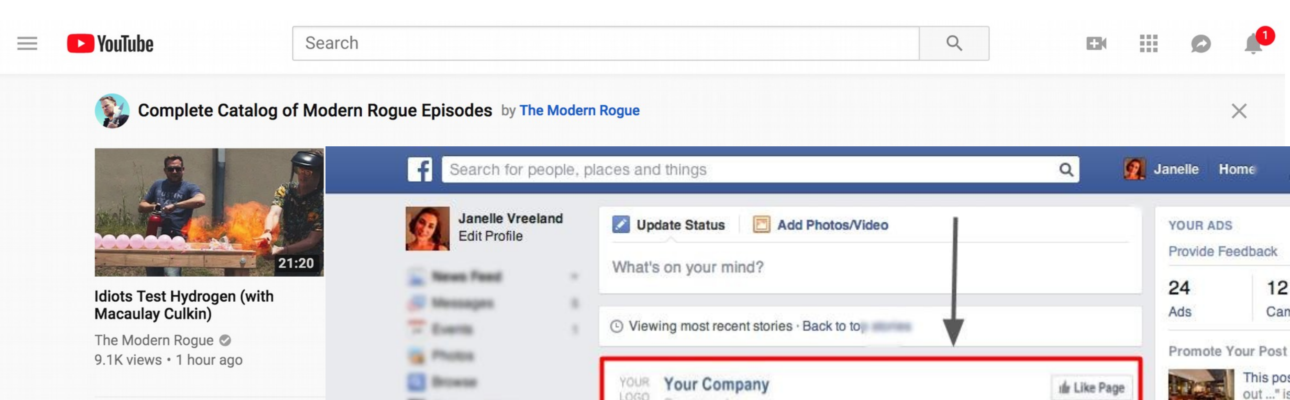
Caroline Trippel, Carole-Jean Wu, Gu-Yeon Wei, David Brooks

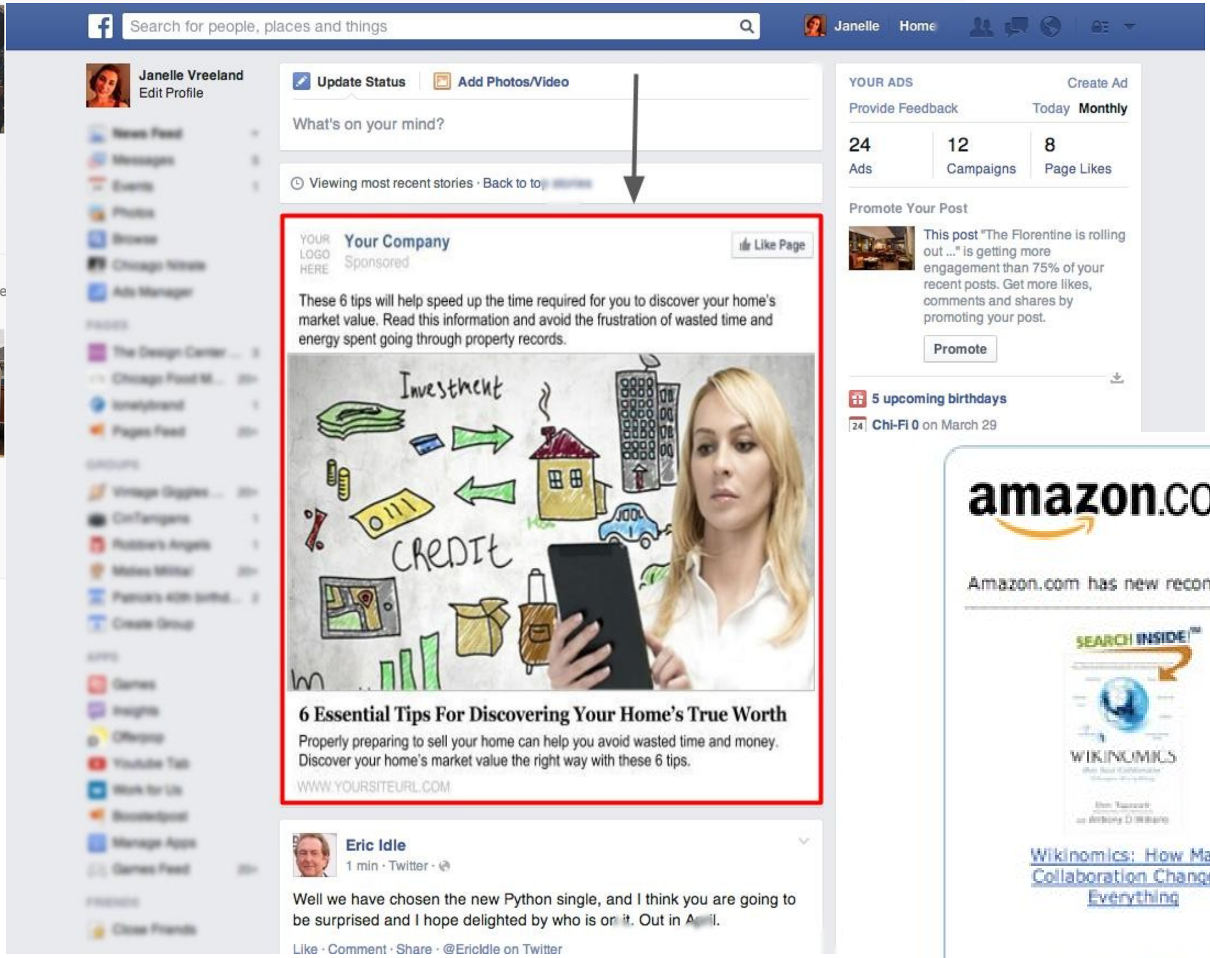
At Boston Area Architecture Workshop (*BARC 2021*)

* Indicates equal contribution

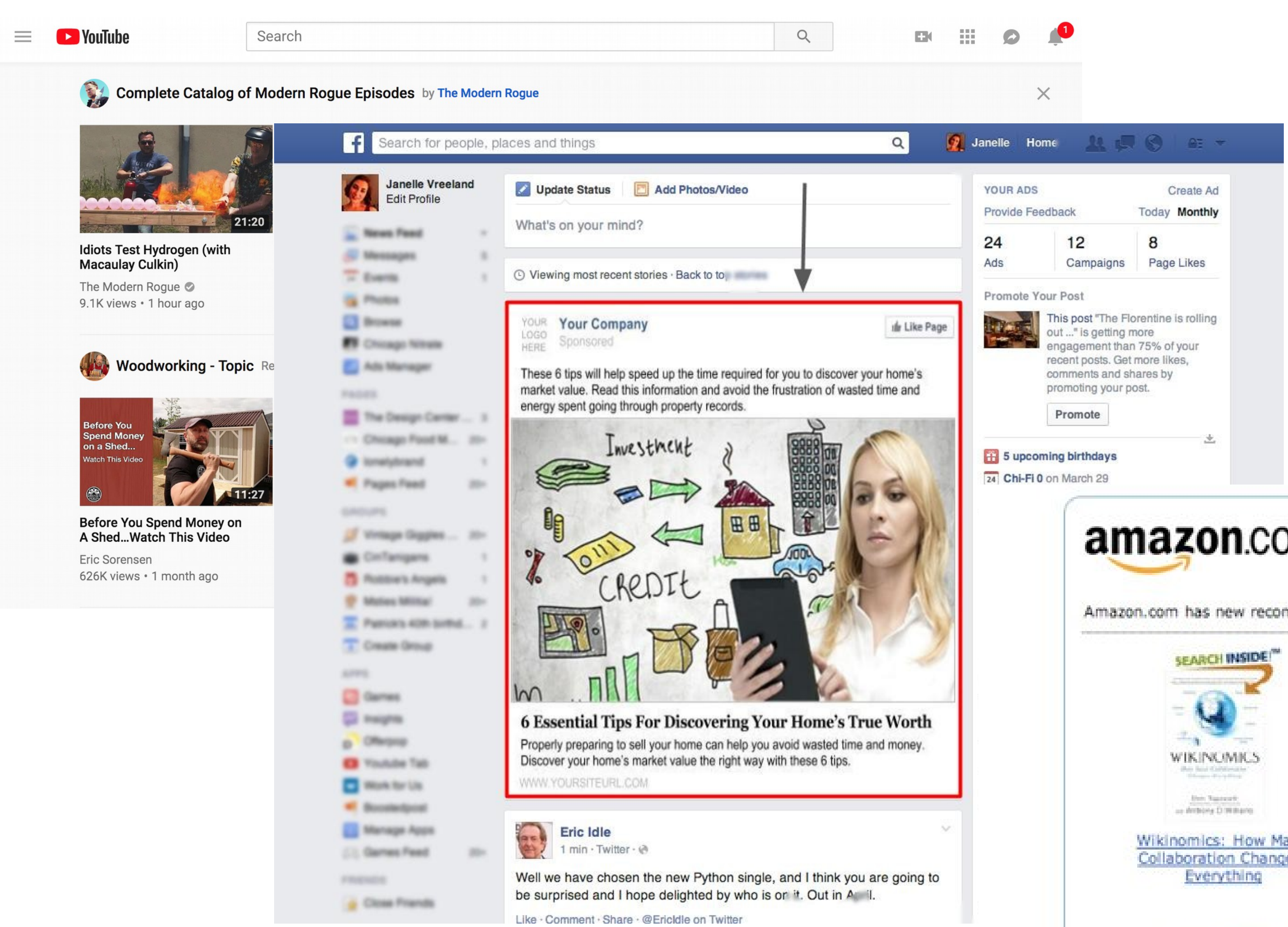

YouTube
Search
Complete Catalog of Modern Rogue Episodes by The Modern Rogue
Not interested
21:20
Idiots Test Hydrogen (with Macaulay Culkin)
The Modern Rogue
9.1K views • 1 hour ago
7:04
Desperate Defense: Bobby Pins
The Modern Rogue
179K views • 1 month ago
7:47
Desperate Defense: Matches
The Modern Rogue
179K views • 1 month ago
5:52
Desperate Defense: Hair Spray
The Modern Rogue
245K views • 1 month ago
20:30
5 Outrageous Con Men
The Modern Rogue
162K views • 2 months ago
Woodworking - Topic
Recommended channel for you
SUBSCRIBE 80K
Before You Spend Money on a Shed... Watch This Video
11:27
Before You Spend Money on A Shed...Watch This Video
Eric Sorensen
626K views • 1 month ago
11:12
Woodturning - One Big Ugly Burl into a dragon egg !!
Andy Phillip
7M views • 1 month ago
14:39
Bearing Forged Into Fine Woodworking Tool
Black Beard Projects
4.2M views • 3 weeks ago
FREE!
PALLET WOOD
OFF GRID CABIN
1:03:23
Building an Off Grid Cabin using Free Pallet Wood: A...
TA Outdoors
1.7M views • 5 months ago
23:44
50 Amazing WoodWorking Skills Tools Tricks. DIY...
FunPhotOK
708K views • 2 weeks ago


YouTube
Search
Complete Catalog of Modern Rogue Episodes by The Modern Rogue
21:20
Idiots Test Hydrogen (with Macaulay Culkin)
The Modern Rogue
9.1K views • 1 hour ago
Woodworking - Topic
Before You Spend Money on a Shed... Watch This Video
11:27
Before You Spend Money on A Shed...Watch This Video
Eric Sorensen
626K views • 1 month ago
Search for people, places and things
Janelle
Home
Janelle Vreeland
Edit Profile
Update Status
Add Photos/Video
What's on your mind?
Viewing most recent stories · Back to top
YOUR LOGO HERE
Your Company
Sponsored
Like Page
These 6 tips will help speed up the time required for you to discover your home's market value. Read this information and avoid the frustration of wasted time and energy spent going through property records.
Investment
CREDIT
6 Essential Tips For Discovering Your Home's True Worth
Properly preparing to sell your home can help you avoid wasted time and money. Discover your home's market value the right way with these 6 tips.
WWW.YOURSITEURL.COM
Eric Idle
1 min · Twitter
Well we have chosen the new Python single, and I think you are going to be surprised and I hope delighted by who is on it. Out in April.
Like · Comment · Share · @EricIdle on Twitter
YOUR ADS
Create Ad
Provide Feedback
Today
Monthly
24
Ads
12
Campaigns
8
Page Likes
Promote Your Post
This post "The Florentine is rolling out ..." is getting more engagement than 75% of your recent posts. Get more likes, comments and shares by promoting your post.
Promote
5 upcoming birthdays
Chi-Fi 0 on March 29
TRENDING
Learn More
Blue Line: CTA Blue Line derailment at O'Hare injures at least 30
Mila Kunis: Mila Kunis Pregnant, Expecting First Child With Fiance Ashton Kutcher
Mark Cuban: Cuban: NFL '10 Years from Implosion'
See More
MUSIC PAGES YOU MAY LIKE
See All
Bobby Rydell
105,235 people like him.
Like
SPONSORED
Create Ad
Learn about ASU Online
asuonline.asu.edu
Ranked Top Tier University - U.S. News & World Report. 100% online. Apply now.
Try Fandor for 2 Weeks


YouTube
Search
Complete Catalog of Modern Rogue Episodes by The Modern Rogue
21:20
Idiots Test Hydrogen (with Macaulay Culkin)
The Modern Rogue
9.1K views • 1 hour ago
Woodworking - Topic
Before You Spend Money on a Shed... Watch This Video
11:27
Before You Spend Money on A Shed...Watch This Video
Eric Sorensen
626K views • 1 month ago

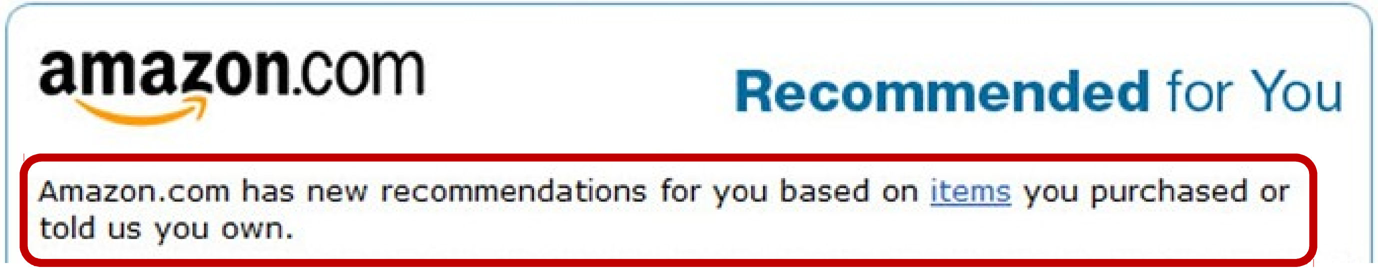

Search for people, places and things
Janelle
Home
Janelle Vreeland
Edit Profile
Update Status
Add Photos/Video
What's on your mind?
Viewing most recent stories · Back to top
YOUR LOGO HERE
Your Company
Sponsored
Like Page
These 6 tips will help speed up the time required for you to discover your home's market value. Read this information and avoid the frustration of wasted time and energy spent going through property records.
Investment
CREDIT
6 Essential Tips For Discovering Your Home's True Worth
Properly preparing to sell your home can help you avoid wasted time and money. Discover your home's market value the right way with these 6 tips.
WWW.YOURSITEURL.COM
Eric Idle
1 min · Twitter
Well we have chosen the new Python single, and I think you are going to be surprised and I hope delighted by who is on it. Out in April.
Like · Comment · Share · @EricIdle on Twitter
YOUR ADS
Create Ad
Provide Feedback
Today
Monthly
24
Ads
12
Campaigns
8
Page Likes
Promote Your Post
This post "The Florentine is rolling out ..." is getting more engagement than 75% of your recent posts. Get more likes, comments and shares by promoting your post.
Promote
5 upcoming birthdays
Chi-Fi 0 on March 29


amazon.com
Recommended for You
Amazon.com has new recommendations for you based on items you purchased or told us you own.
SEARCH INSIDE!
Wikinomics: How Mass Collaboration Changes Everything
Word of Mouth Marketing: How Smart Companies Get People Talking
Made to Stick: Why Some Ideas Survive and Others Die
Cut to the Chase: and 99 Other Rules to Liberate Yourself and Gain Back the Gift of Time
NEVER COLD CALL AGAIN!
DRUCKER
The Effective Executive
JAMES PATTERSON
7TH HEAVEN
AN ARMY OF ORDINARY PEOPLE

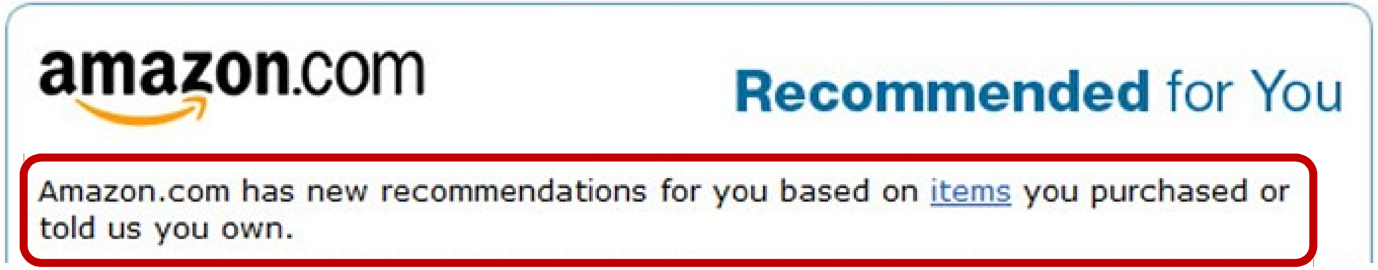

YouTube
Search
Complete Catalog of Modern Rogue Episodes by The Modern Rogue
21:20
Idiots Test Hydrogen (with Macaulay Culkin)
The Modern Rogue
9.1K views • 1 hour ago
Woodworking - Topic
Before You Spend Money on a Shed... Watch This Video
11:27
Before You Spend Money on A Shed...Watch This Video
Eric Sorensen
626K views • 1 month ago
Search for people, places and things
Janelle Home
Janelle Vreeland
Edit Profile
Update Status
Add Photos/Video
What's on your mind?
Viewing most recent stories · Back to top stories
YOUR LOGO HERE
Your Company
Sponsored
Like Page
These 6 tips will help speed up the time required for you to discover your home's market value. Read this information and avoid the frustration of wasted time and energy spent going through property records.
Investment
CREDIT
6 Essential Tips For Discovering Your Home's True Worth
Properly preparing to sell your home can help you avoid wasted time and money. Discover your home's market value the right way with these 6 tips.
WWW.YOURSITEURL.COM
Eric Idle
1 min · Twitter ·
Well we have chosen the new Python single, and I think you are going to be surprised and I hope delighted by who is on it. Out in April.
Like · Comment · Share · @EricIdle on Twitter
YOUR ADS
Create Ad
Provide Feedback
Today Monthly
24 Ads
12 Campaigns
8 Page Likes
Promote Your Post
This post "The Florentine is rolling out ..." is getting more engagement than 75% of your recent posts. Get more likes, comments and shares by promoting your post.
Promote
5 upcoming birthdays
Chi-Fi 0 on March 29


amazon.com
Recommended for You
Amazon.com has new recommendations for you based on items you purchased or told us you own.
Wikinomics: How Mass Collaboration Changes Everything
Word of Mouth Marketing: How Smart Companies Get People Talking
Made to Stick: Why Some Ideas Survive and Others Die
Cut to the Chase: and 99 Other Rules to Liberate Yourself and Gain Back the Gift of Time
SEARCH INSIDE!
NEVER COLD CALL AGAIN!
DRUCKER The Effective Executive
JAMES PATTERSON 7TH HEAVEN
AN ARMY OF ORDINARY PEOPLE

What is recommendation

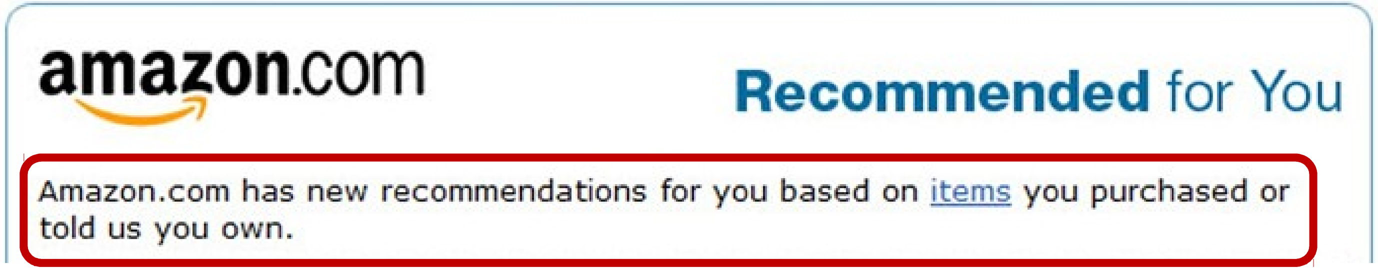


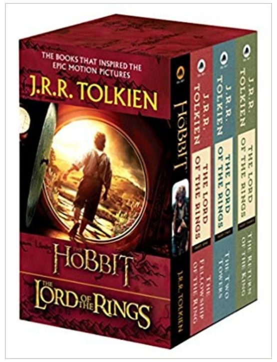
What is recommendation



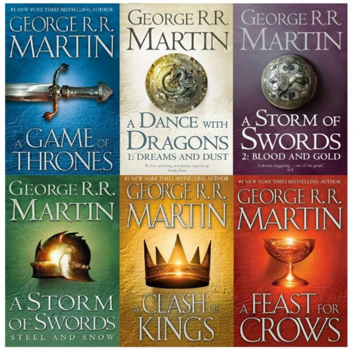
You **Item Preferences**

What is recommendation





Recommendation System

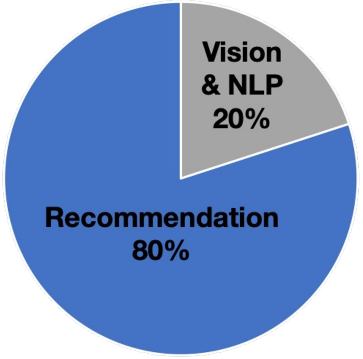


You **Item Preferences** **Item Recommendations**

Why should computer architects care

Why should computer architects care

Infrastructure Demands

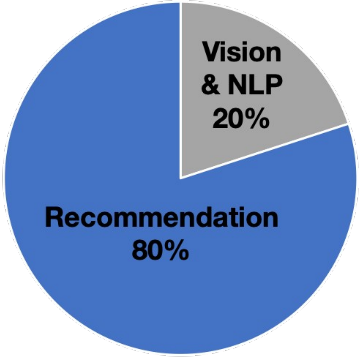


Facebook Datacenters' AI Inference Cycles [1]

The Architectural Implications of Facebook's DNN-based Personalized Recommendation Models" (HPCA 2020) *Gupta, et. al*.

Why should computer architects care

Infrastructure Demands



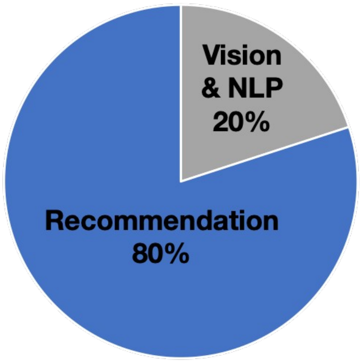
Facebook Datacenters' AI Inference Cycles [1]

Also accounts for 50% of training demand [2]

The Architectural Implications of Facebook's DNN-based Personalized Recommendation Models" (HPCA 2020) *Gupta, et. al*.
Deep Learning Training in Facebook Data Centers: Design of Scale-up and Scale-out Systems" (arXiv 2020) Naumov, et. al.

Why should computer architects care

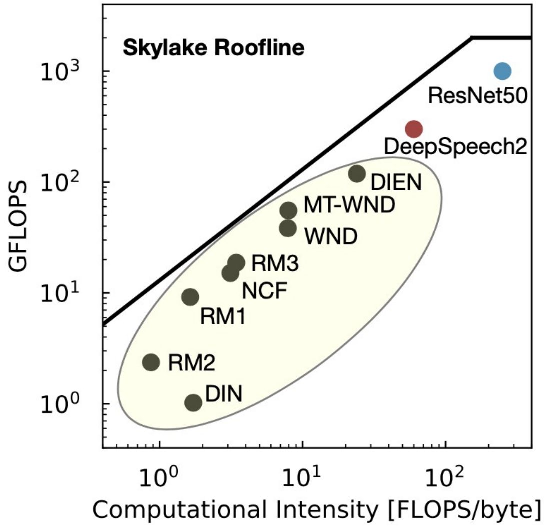
Infrastructure Demands



Facebook Datacenters' AI Inference Cycles [1]

Also accounts for 50% of training demand [2

Unique Compute Requirements

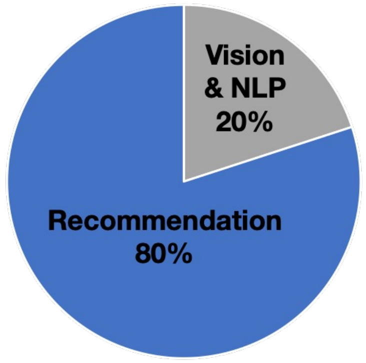


Different than CNNs and RNNs

The Architectural Implications of Facebook's DNN-based Personalized Recommendation Models" (HPCA 2020) *Gupta, et. al*.
Deep Learning Training in Facebook Data Centers: Design of Scale-up and Scale-out Systems" (arXiv 2020) Naumov, et. al.

Why should computer architects care

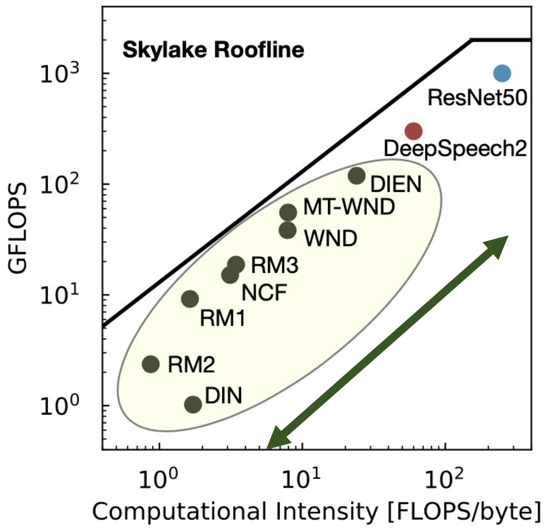
Infrastructure Demands



Facebook Datacenters' AI Inference Cycles [1]

Also accounts for 50% of training demand [2

Unique Compute Requirements



Different than CNNs and RNNs

Model Diversity

The Architectural Implications of Facebook's DNN-based Personalized Recommendation Models" (HPCA 2020) *Gupta, et. al*.
Deep Learning Training in Facebook Data Centers: Design of Scale-up and Scale-out Systems" (arXiv 2020) Naumov, et. al.

This talk

This talk

Characterization

[IISWC ‘20]

This talk

Characterization

[IISWC ‘20]

RecSSD

[ASPLOS ‘21]

This talk

Characterization

[IISWC ‘20]

RecSSD

[ASPLOS ‘21]

Improving recommendation requires **cross-stack characterization**

Improving recommendation requires **cross-stack characterization**

Algorithms

Application variety leads to **algorithm diversity**

Improving recommendation requires **cross-stack characterization**

Algorithms

Application variety leads to **algorithm diversity**

Software

Algorithms are implemented with **different software frameworks**

Improving recommendation requires **cross-stack characterization**

Algorithms

Application variety leads to **algorithm diversity**

Software

Algorithms are implemented with **different software frameworks**

Hardware

Recommendation is deployed on **heterogenous hardware**

Improving recommendation requires **cross-stack characterization**

Algorithms	Software	Hardware
 		
Application variety leads to **algorithm diversity**	Algorithms are implemented with **different software frameworks**	Recommendation is deployed on **heterogenous hardware**

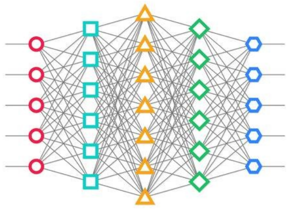
Different layers of the execution stack have different bottlenecks!

Characterization

Question: What are the bottlenecks of each layer and how do they affect one another?

Characterization

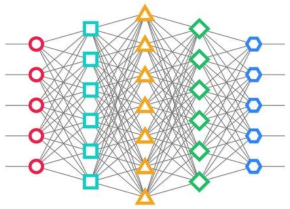
Question: What are the bottlenecks of each layer and how do they affect one another?



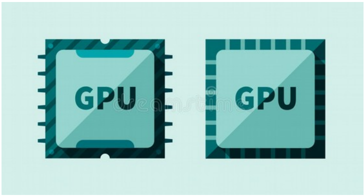
What do industry-representative algorithms (**model architectures**) look like?

Characterization

Question: What are the bottlenecks of each layer and how do they affect one another?



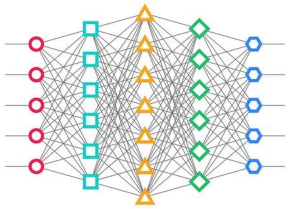
What do industry-representative algorithms (**model architectures**) look like?



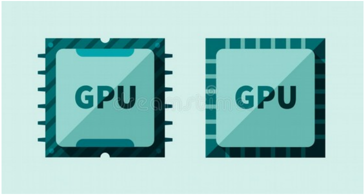
What are the **performance trends** of deploying recommendation on CPUs and GPUs?

Characterization

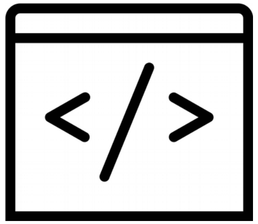
Question: What are the bottlenecks of each layer and how do they affect one another?



What do industry-representative algorithms (**model architectures**) look like?



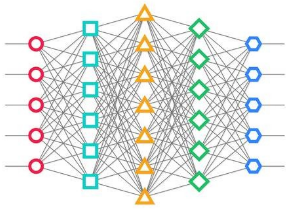
What are the **performance trends** of deploying recommendation on CPUs and GPUs?



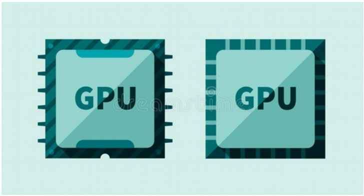
Can we explain the performance trends with **software level operators**?

Characterization

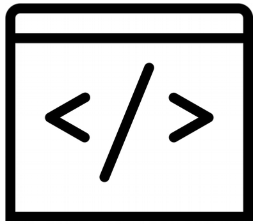
Question: What are the bottlenecks of each layer and how do they affect one another?



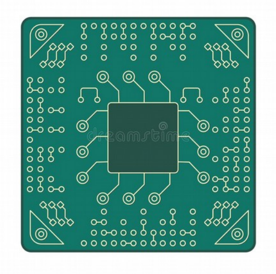
What do industry-representative algorithms (**model architectures**) look like?



What are the **performance trends** of deploying recommendation on CPUs and GPUs?



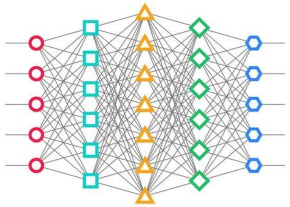
Can we explain the performance trends with **software level operators**?



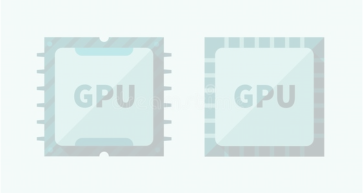
Can we explain the performance trends with **microarchitectural analysis**?

Characterization

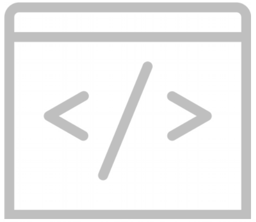
Question: What are the bottlenecks of each layer and how do they affect one another?



What do industry-representative algorithms (**<u>model architectures</u>**) look like?



What are the **<u>performance trends</u>** of deploying recommendation on CPUs and GPUs?

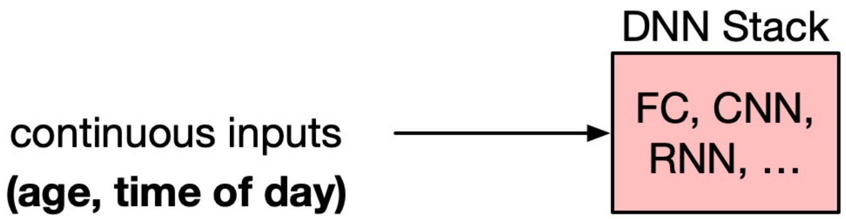


Can we explain the performance trends with **<u>software level operators</u>**?

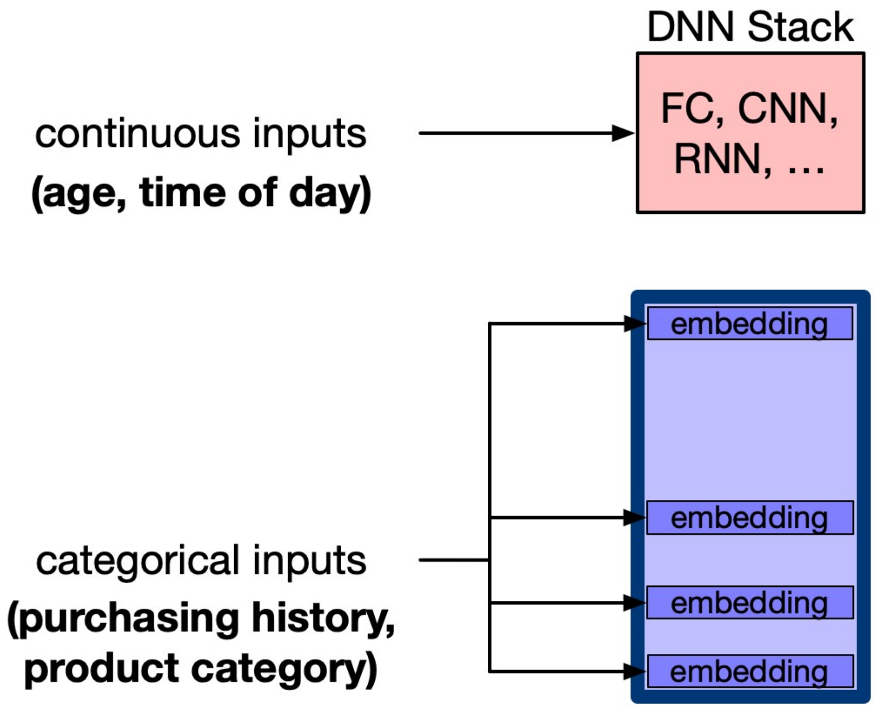
Can we explain the performance trends with **<u>microarchitectural analysis</u>**?

Deep Recommendation Model Architecture

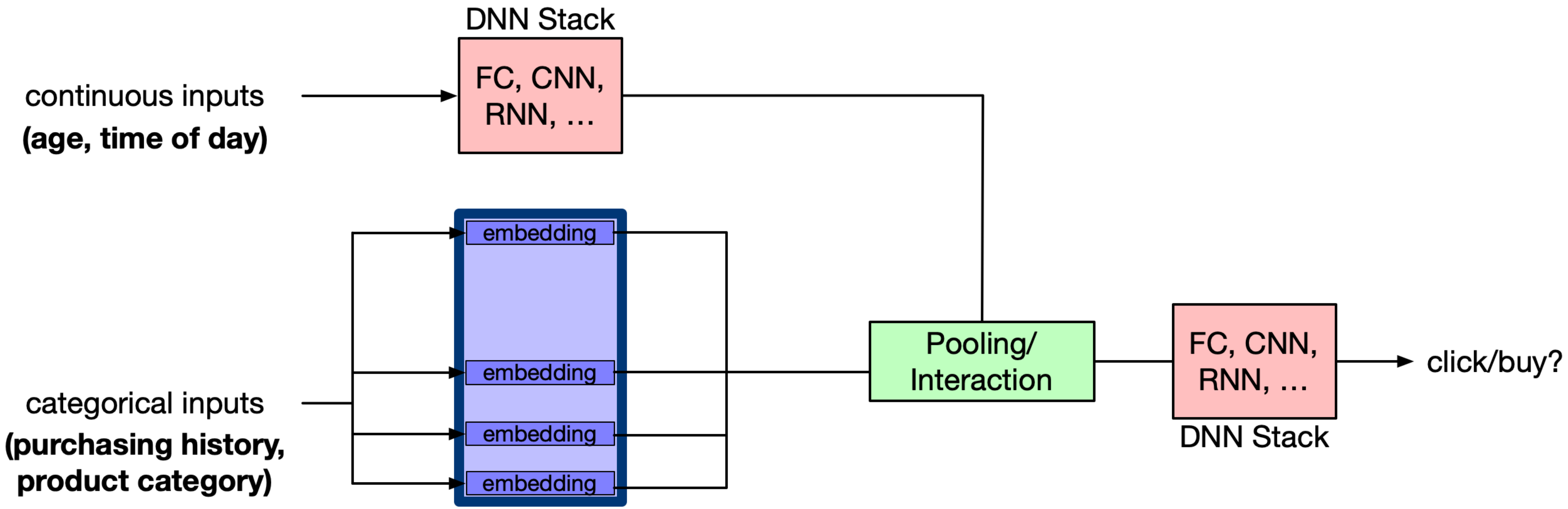
Deep Recommendation Model Architecture



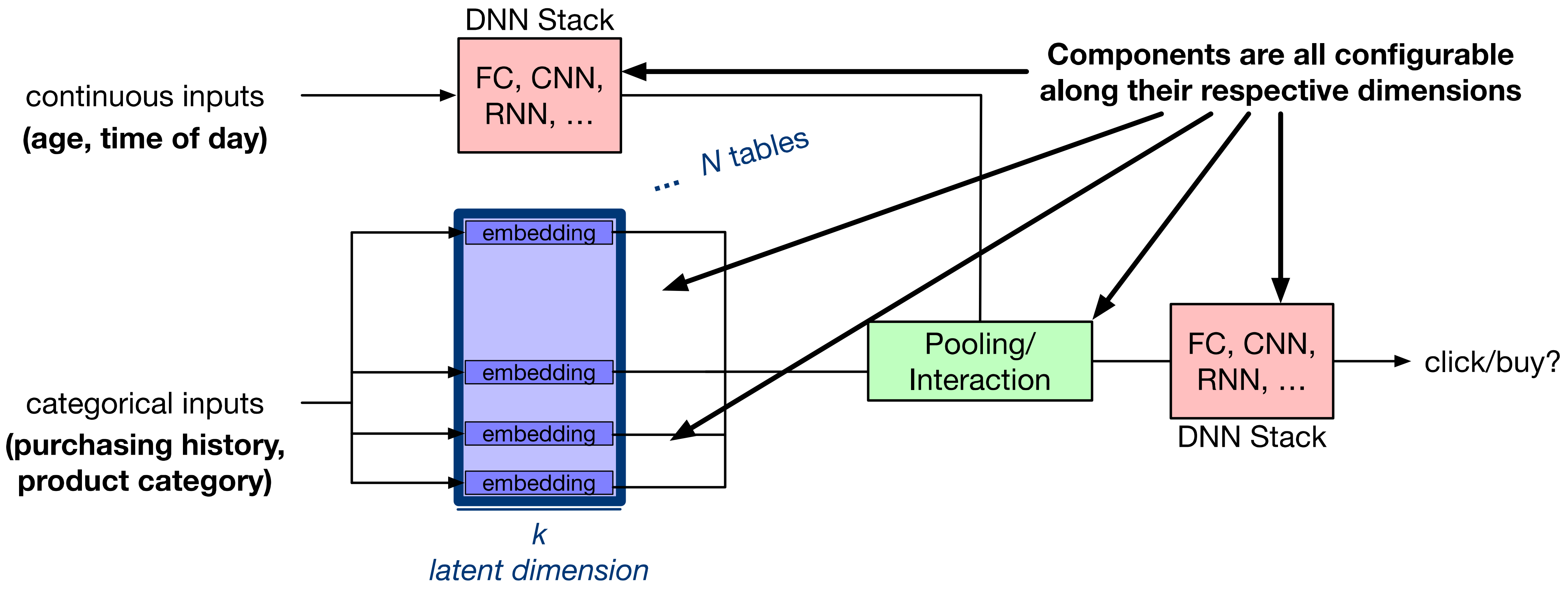
Deep Recommendation Model Architecture



Deep Recommendation Model Architecture



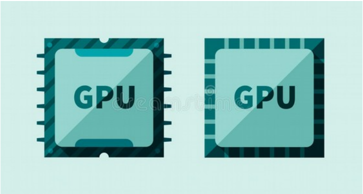
Deep Recommendation Model Architecture



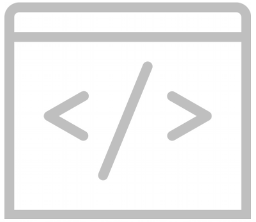
Characterization

Question: What are the bottlenecks of each layer and how do they affect one another?

What do industry-representative algorithms (**model architectures**) look like?



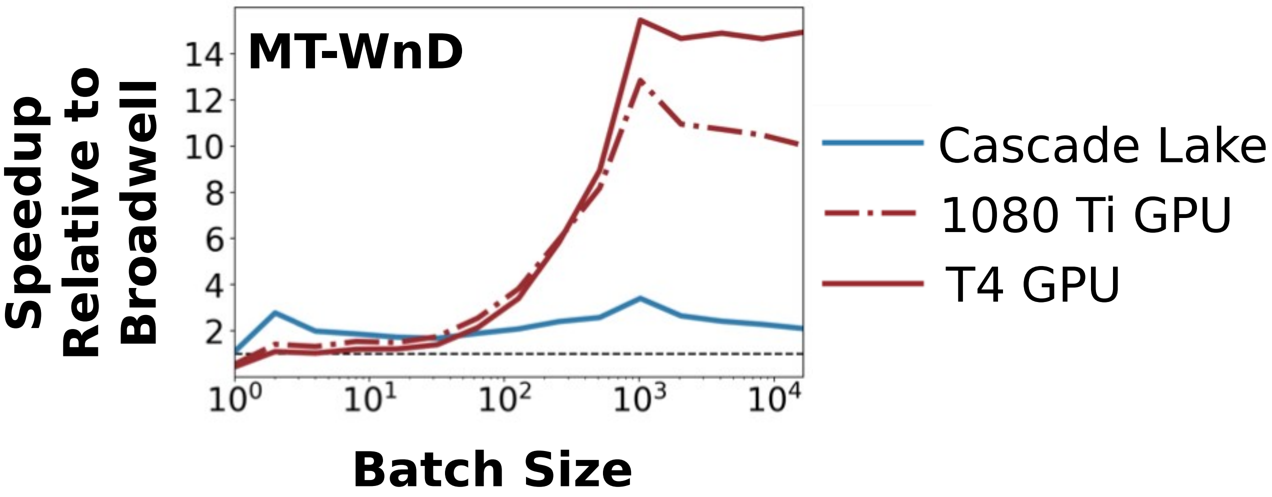
What are the **performance trends** of deploying recommendation on CPUs and GPUs?



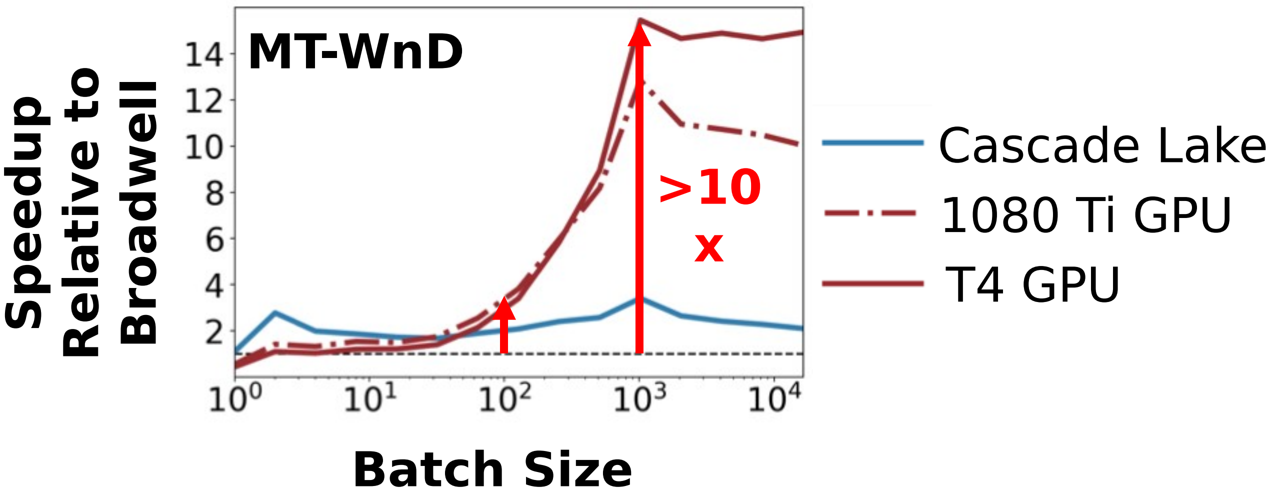
Can we explain the performance trends with **software level operators**?

Can we explain the performance trends with **microarchitectural analysis**?

Systems Platforms Evaluation

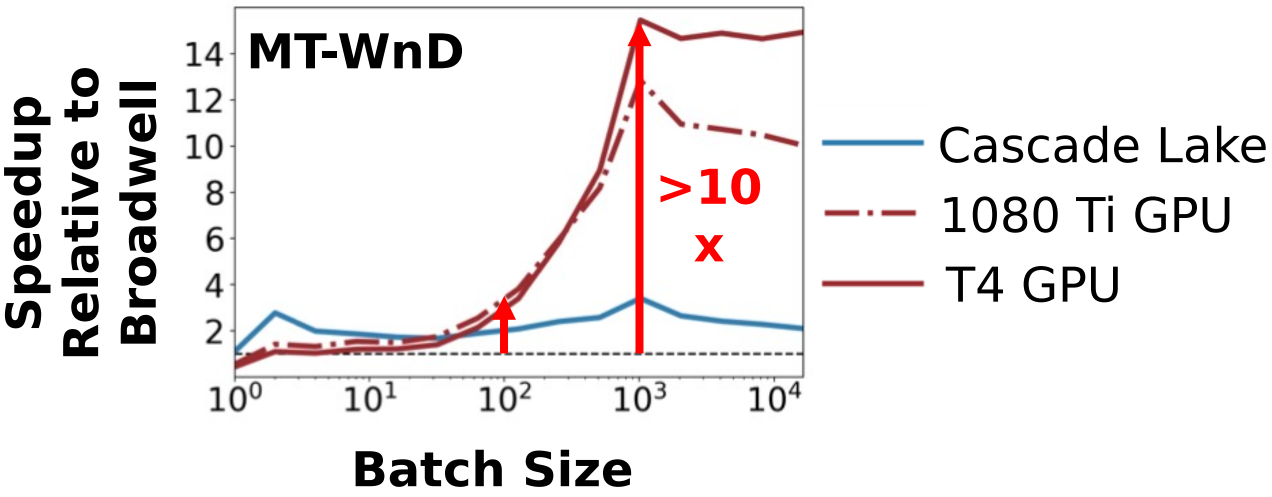


Systems Platforms Evaluation



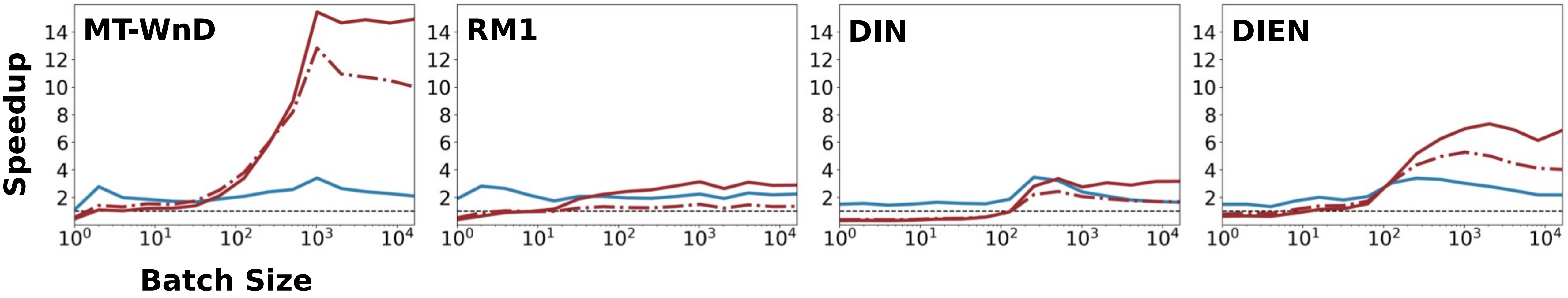
Great speedup on GPUs

Systems Platforms Evaluation



Great speedup on GPUs
Rely on Fully-Connected (FC) stacks

Systems Platforms Evaluation



Great speedup on GPUs

Rely on Fully-Connected (FC) stacks

Bad speedup on GPUs

Rely on Embedding lookups

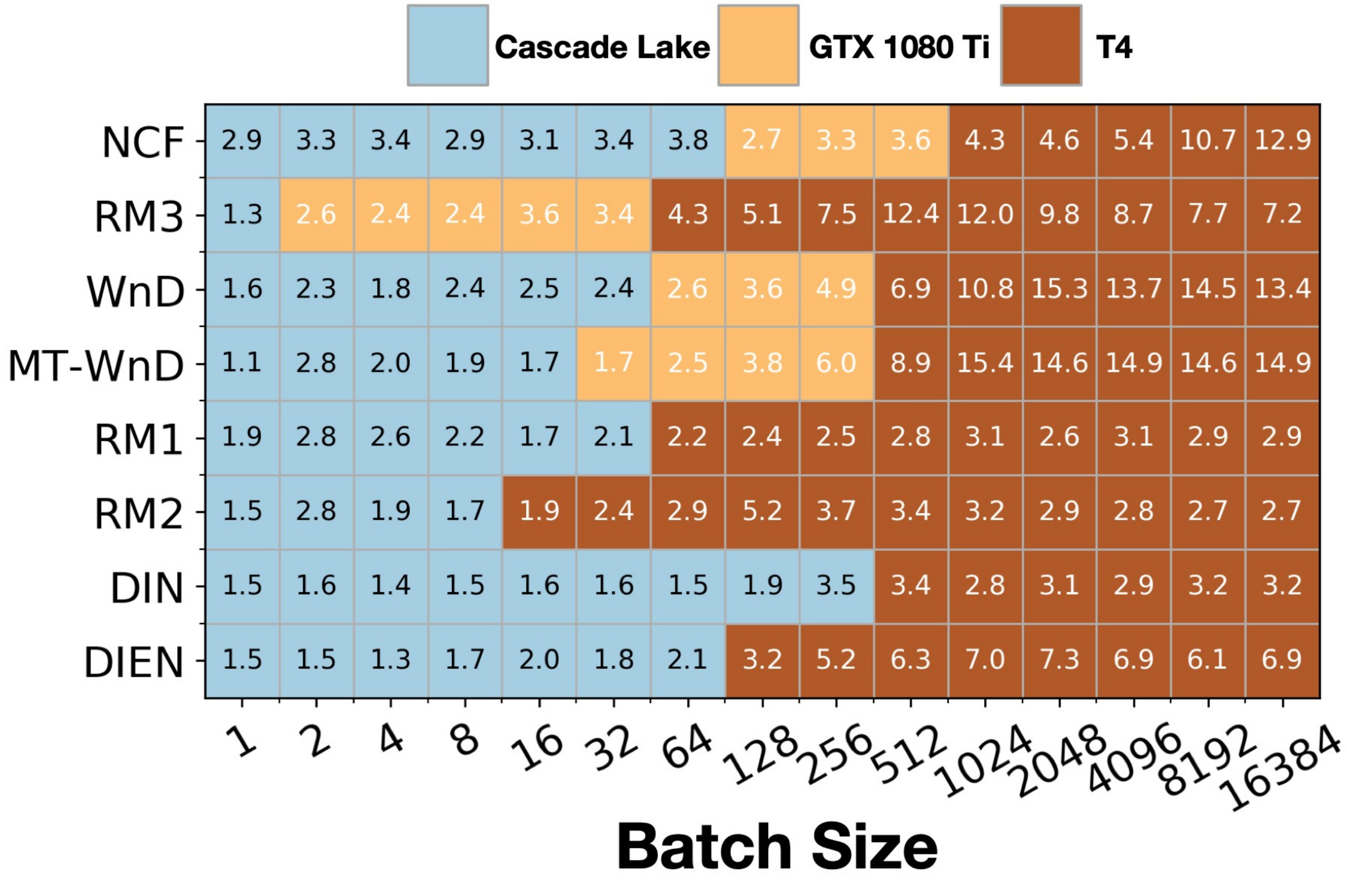
Implements Attention Mechanism

DIN bad on GPUs
DIEN good on GPUs

Cascade Lake
1080 Ti GPU
T4 GPU

Model architecture and use-case play important roles in determining acceleration

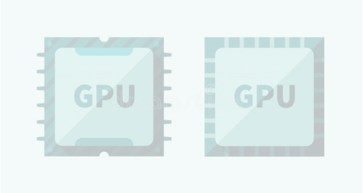
Optimal hardware varies based on **model architecture and input batch size**



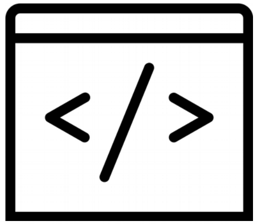
Characterization

Question: What are the bottlenecks of each layer and how do they affect one another?

What do industry-representative algorithms (**model architectures**) look like?



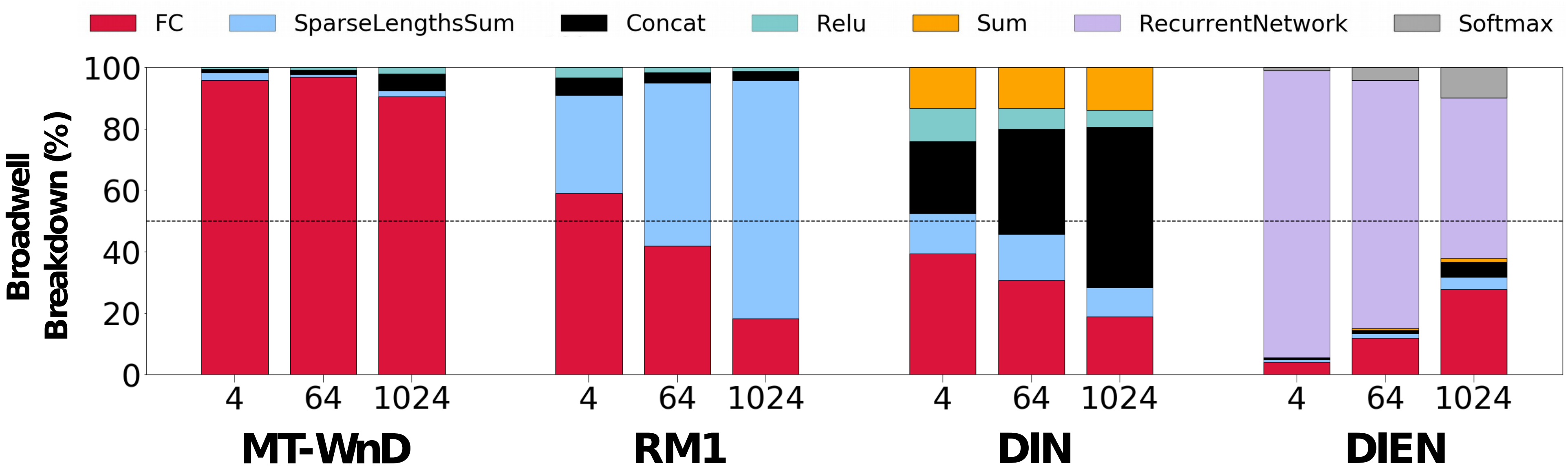
What are the **performance trends** of deploying recommendation on CPUs and GPUs?



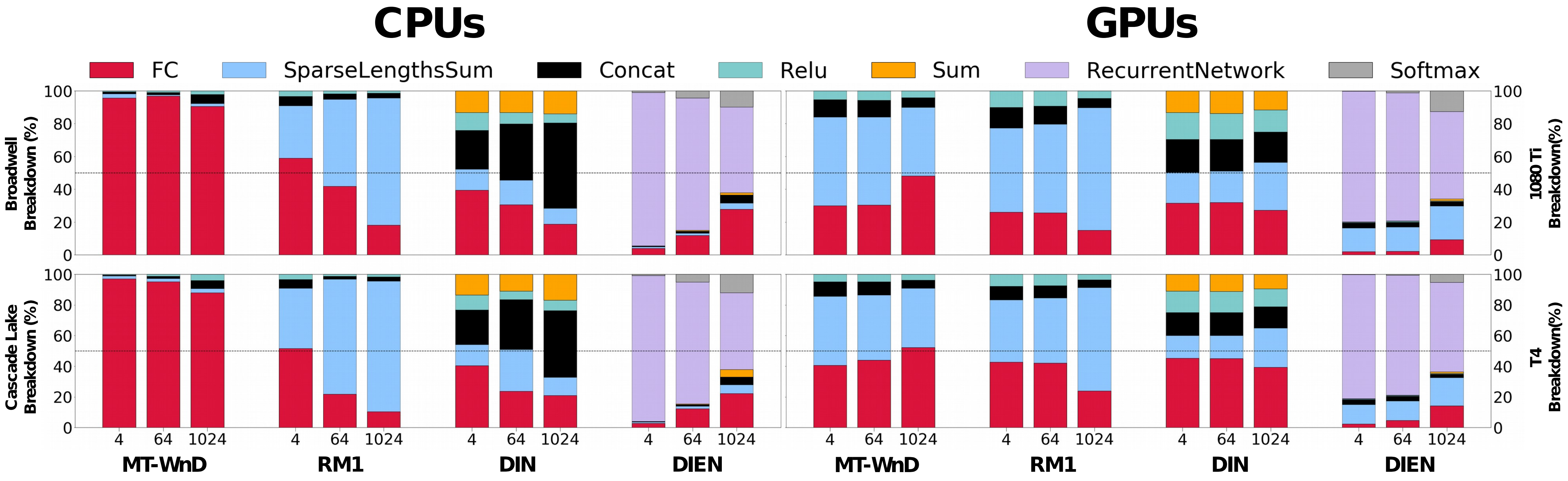
Can we explain the performance trends with **software level operators**?

Can we explain the performance trends with **microarchitectural analysis**?

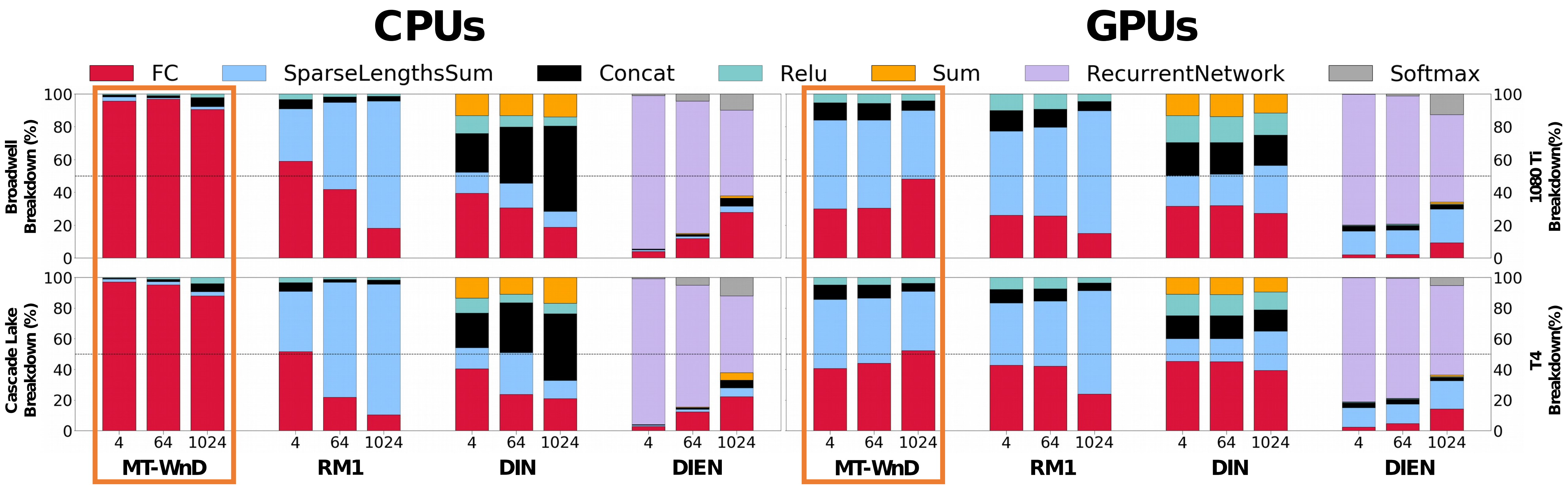
Operator Breakdowns



Operator Breakdowns

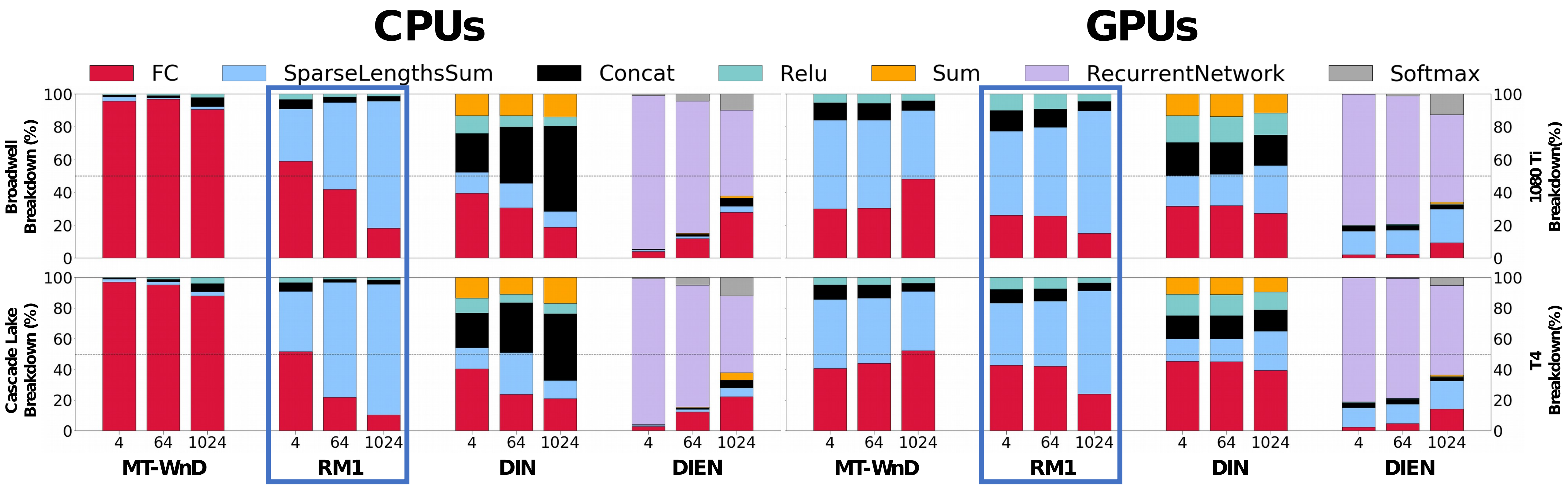


Operator Breakdowns



GPUs accelerate models dominated by **FC** operators on CPUs **by reducing the FC operator**

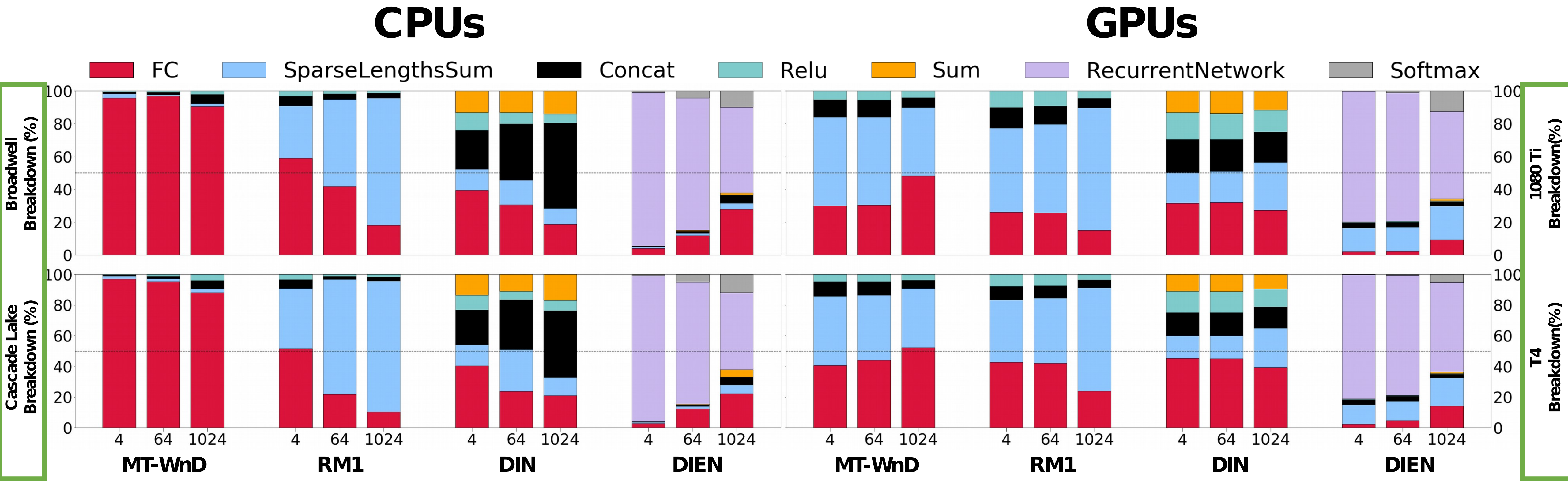
Operator Breakdowns



GPUs struggle with models dominated by **Embedding** operators on CPUs

due to data communication overheads

Operator Breakdowns

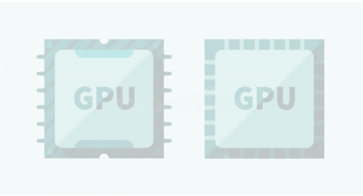


Different generations of the same platform type (i.e., CPU/GPU) **affect exact operator usages but retain general trends**.

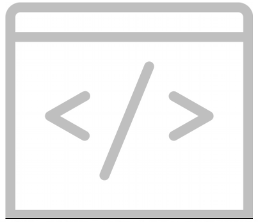
Characterization

Question: What are the bottlenecks of each layer and how do they affect one another?

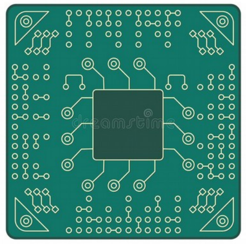
What do industry-representative algorithms (**model architectures**) look like?



What are the **performance trends** of deploying recommendation on CPUs and GPUs?

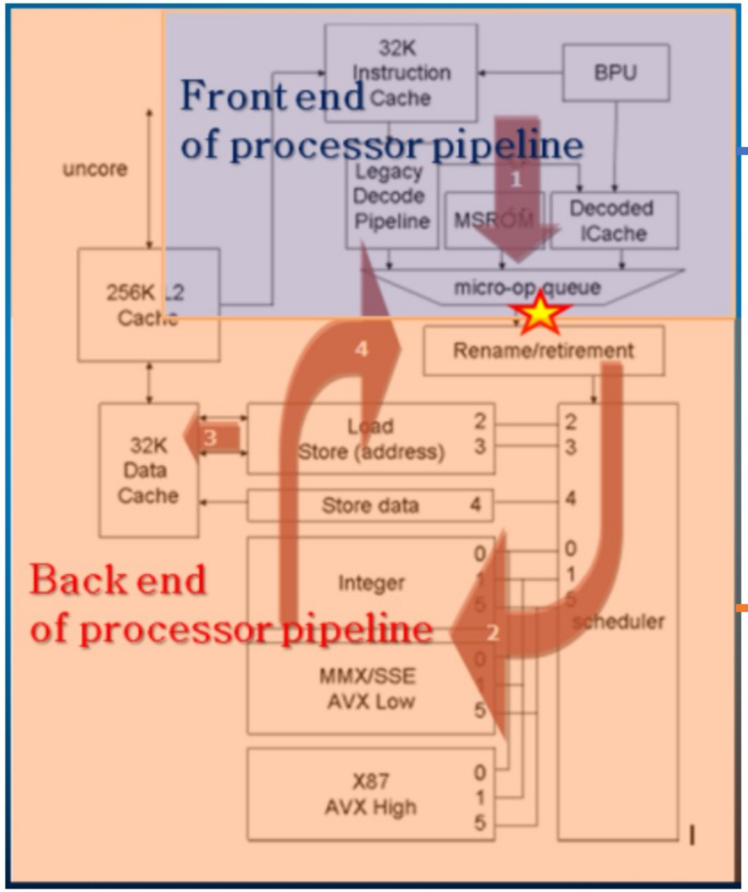


Can we explain the performance trends with **software level operators**?



Can we explain the performance trends with **microarchitectural analysis**?

TopDown Background



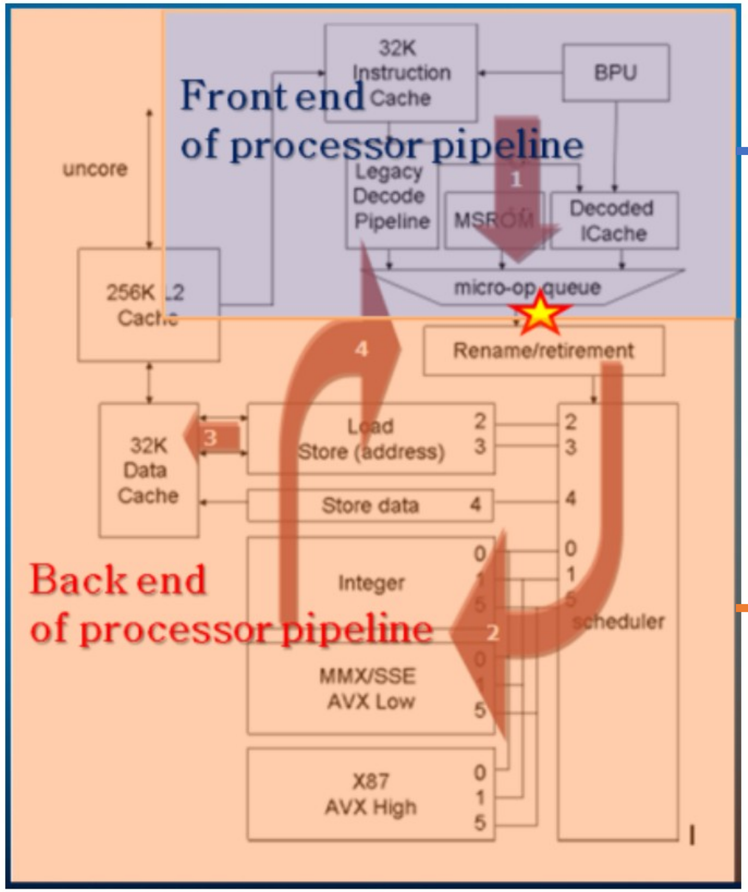
Latency Bound

Bandwidth Bound

Core Bound

Memory Bound

TopDown Background



Latency Bound

i-cache miss

Bandwidth Bound

instruction decoder inefficiency

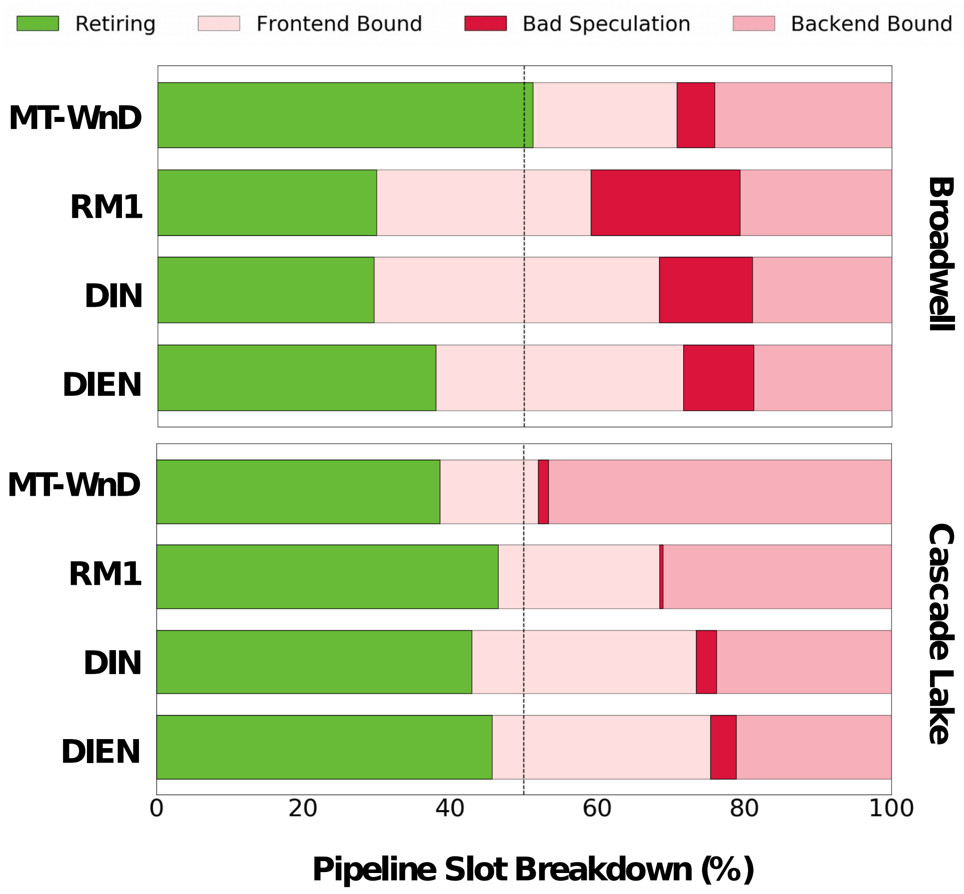
Core Bound

sub-optimal functional units

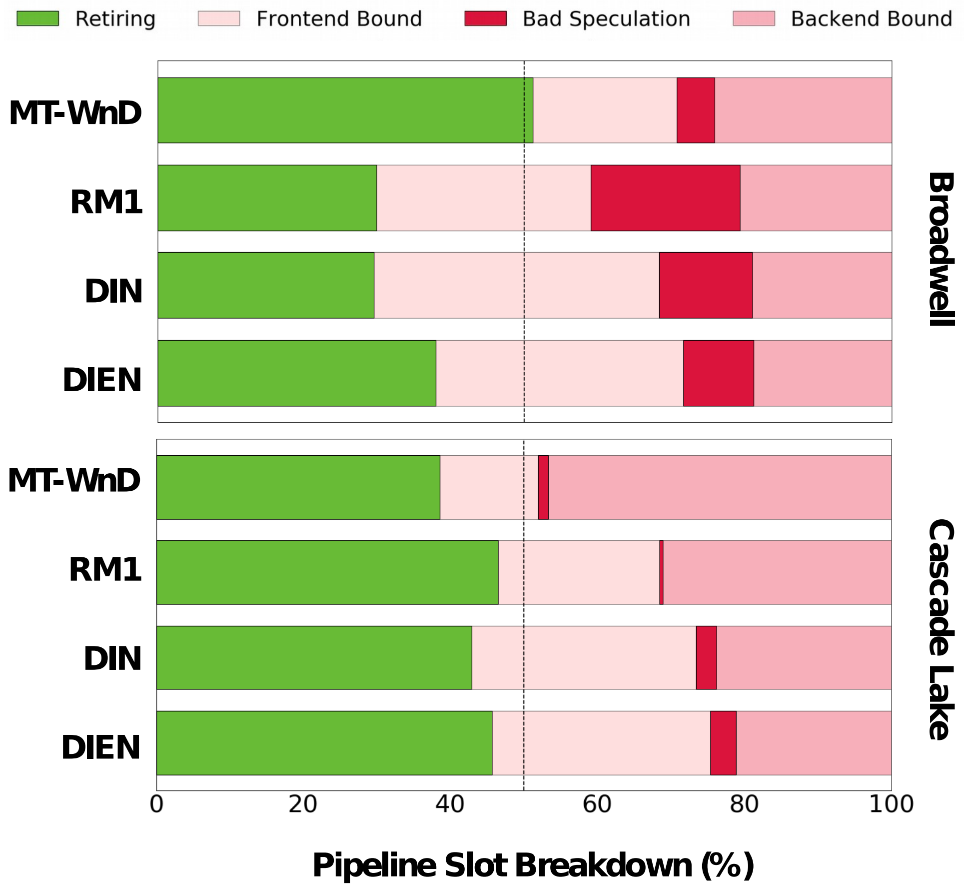
Memory Bound

d-cache miss/bandwidth

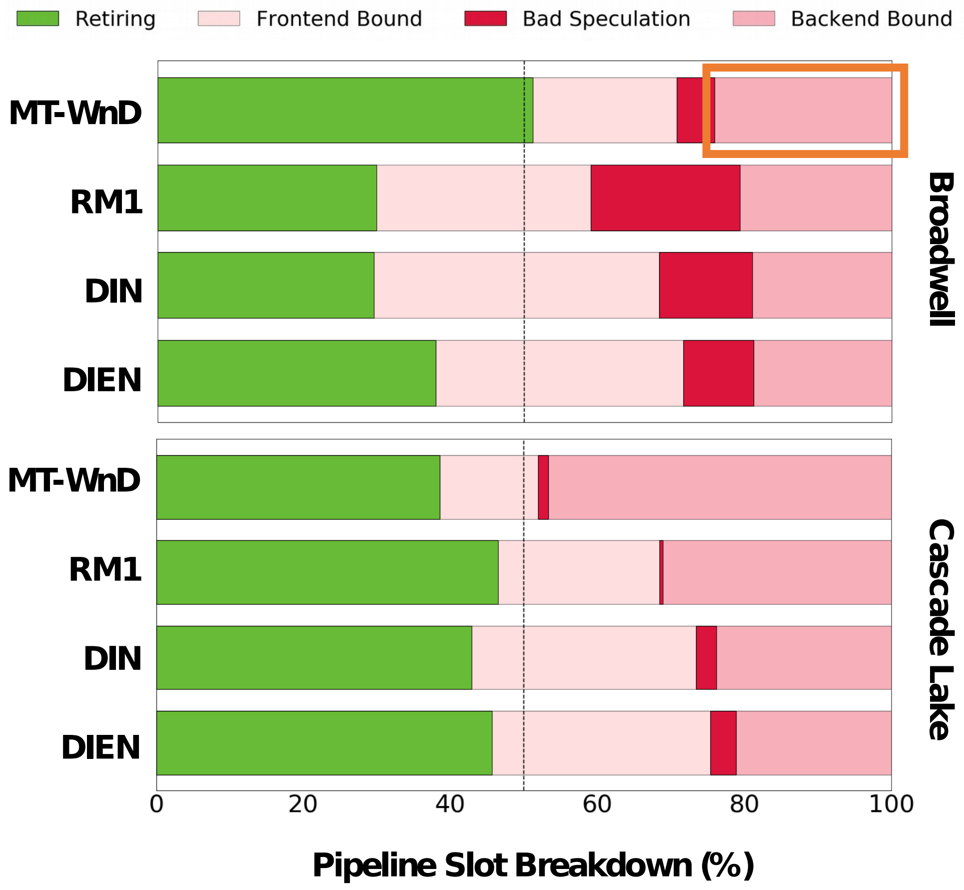
TopDown Pipeline Slot Breakdowns



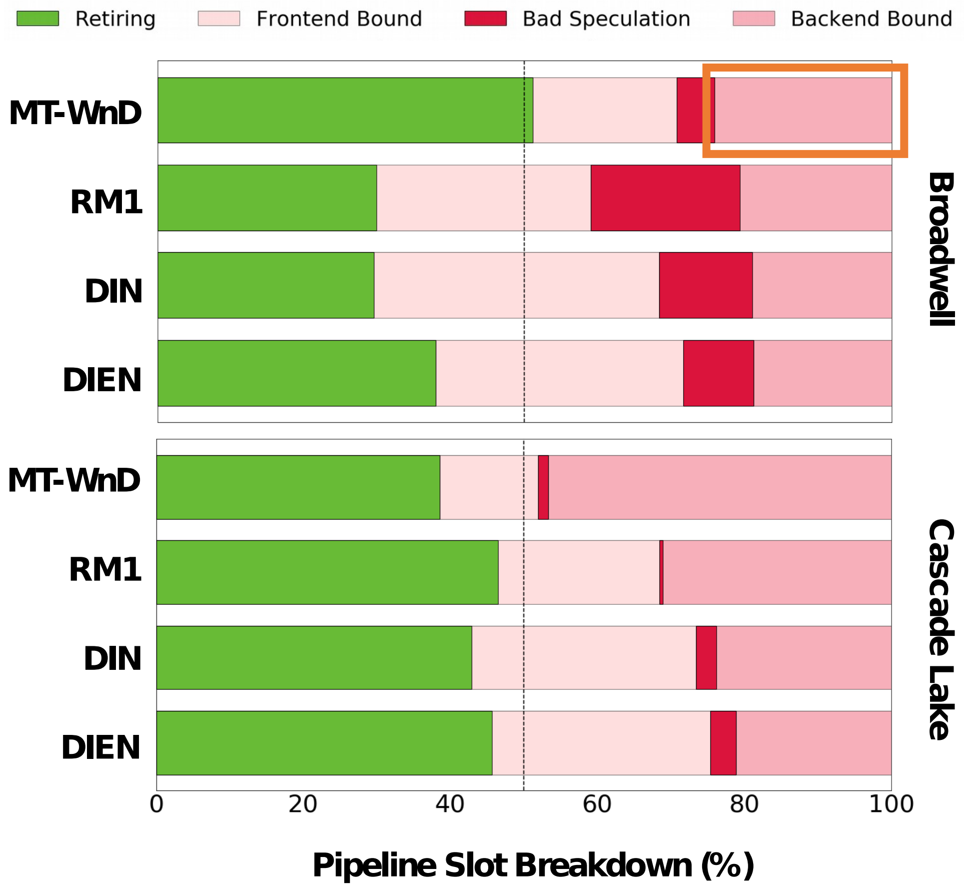
On Broadwell, FC-dominated models **are limited by insufficient functional units**

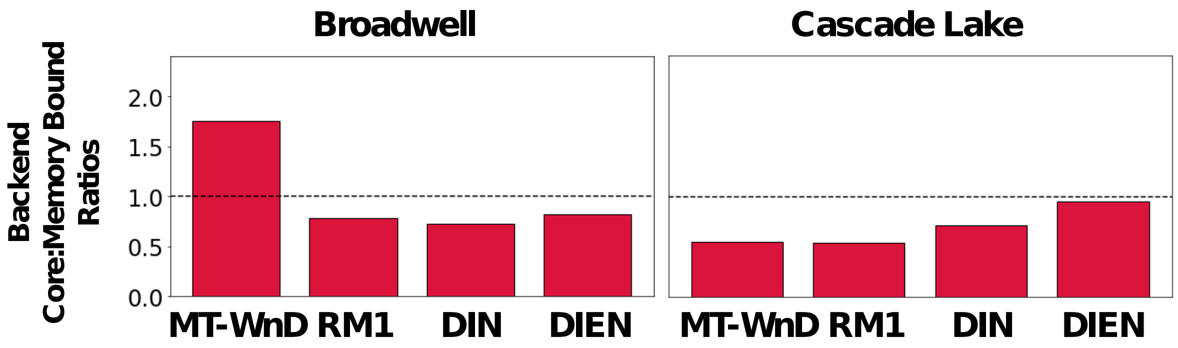


On Broadwell, FC-dominated models **are limited by insufficient functional units**

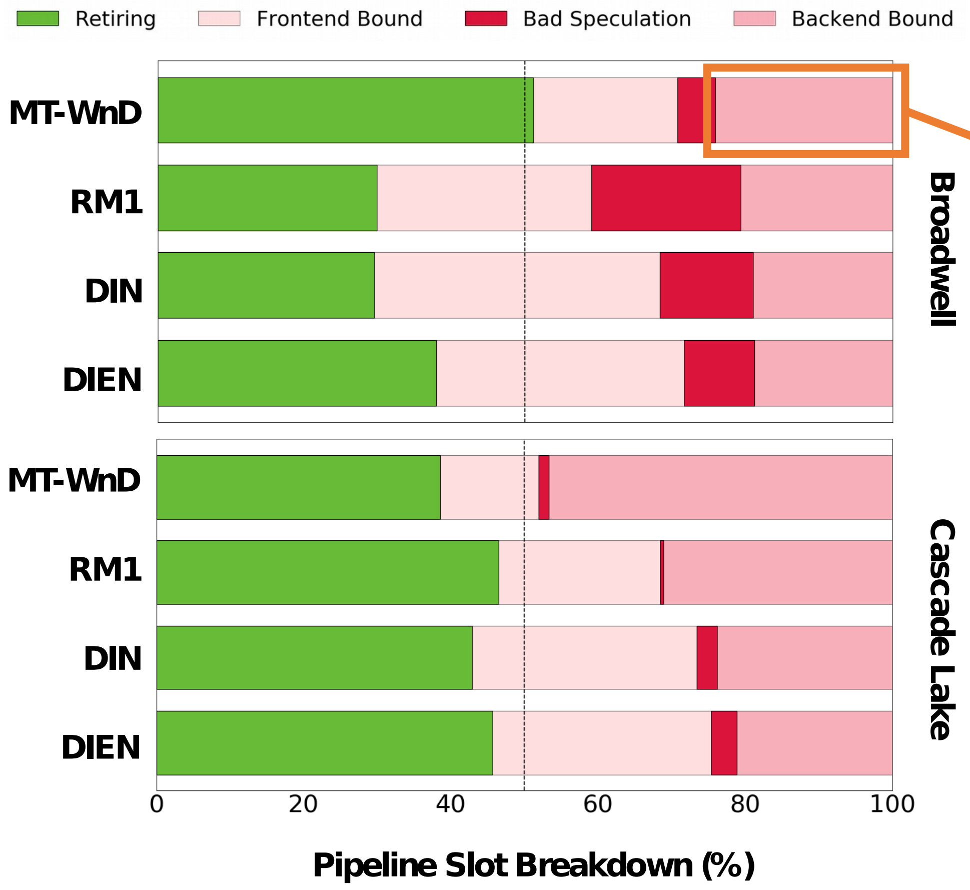


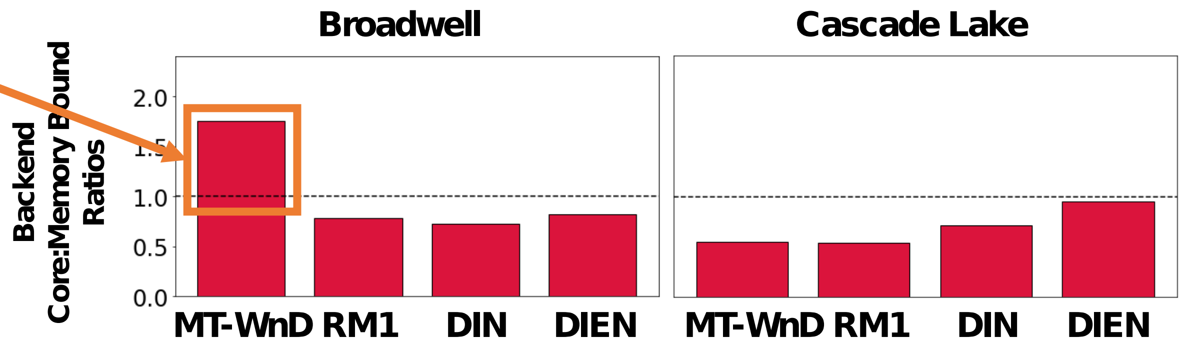
On Broadwell, FC-dominated models **are limited by insufficient functional units**





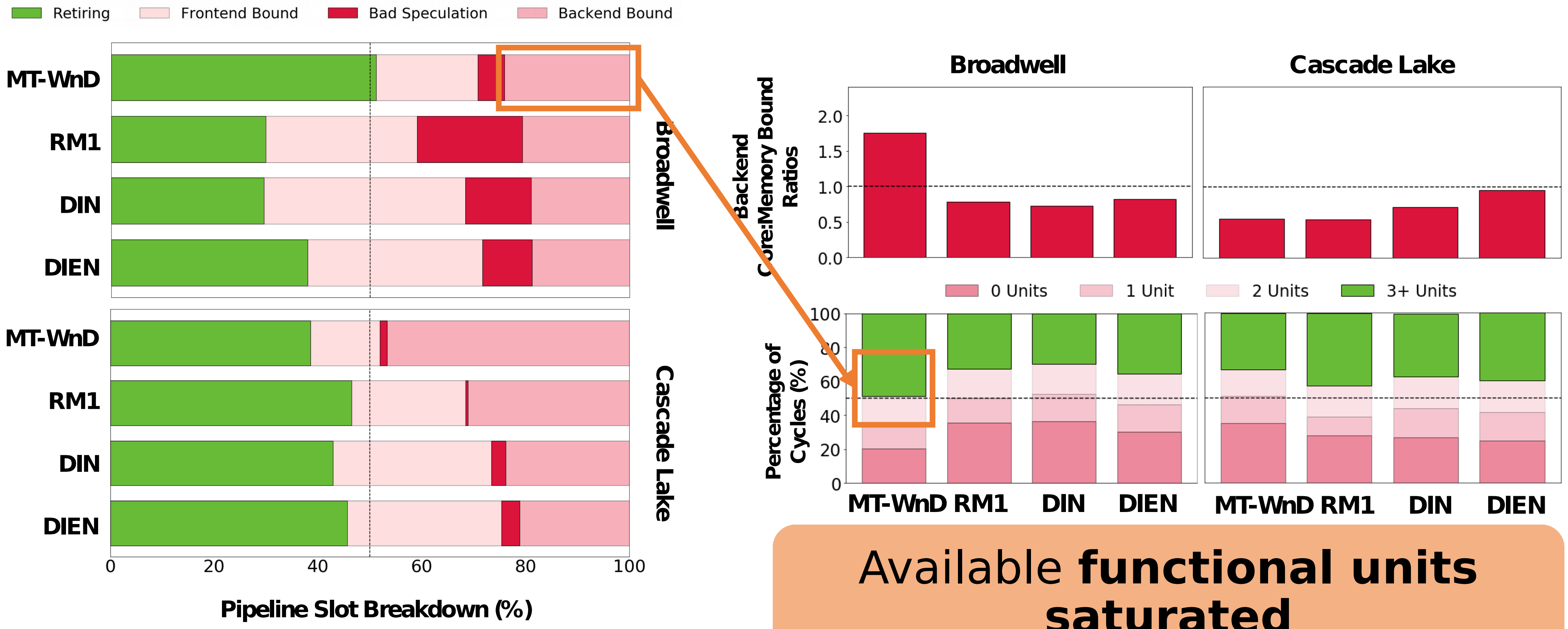
On Broadwell, FC-dominated models **are limited by insufficient functional units**





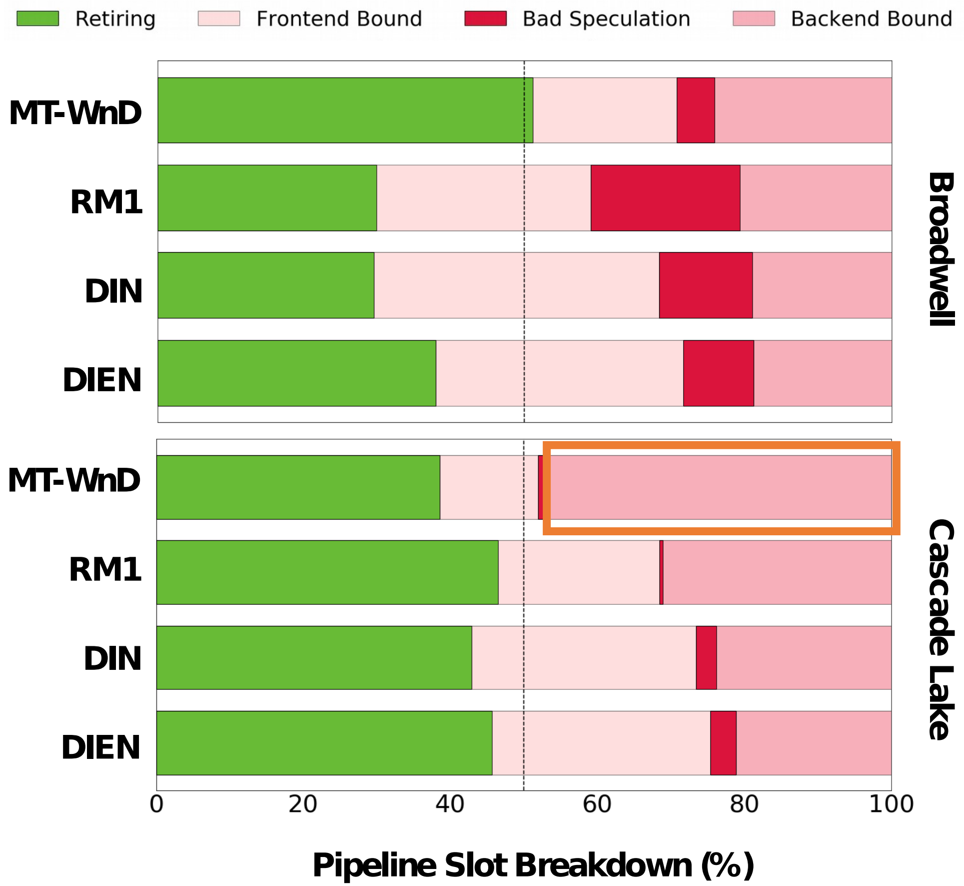
MT-WnD **core bound**!

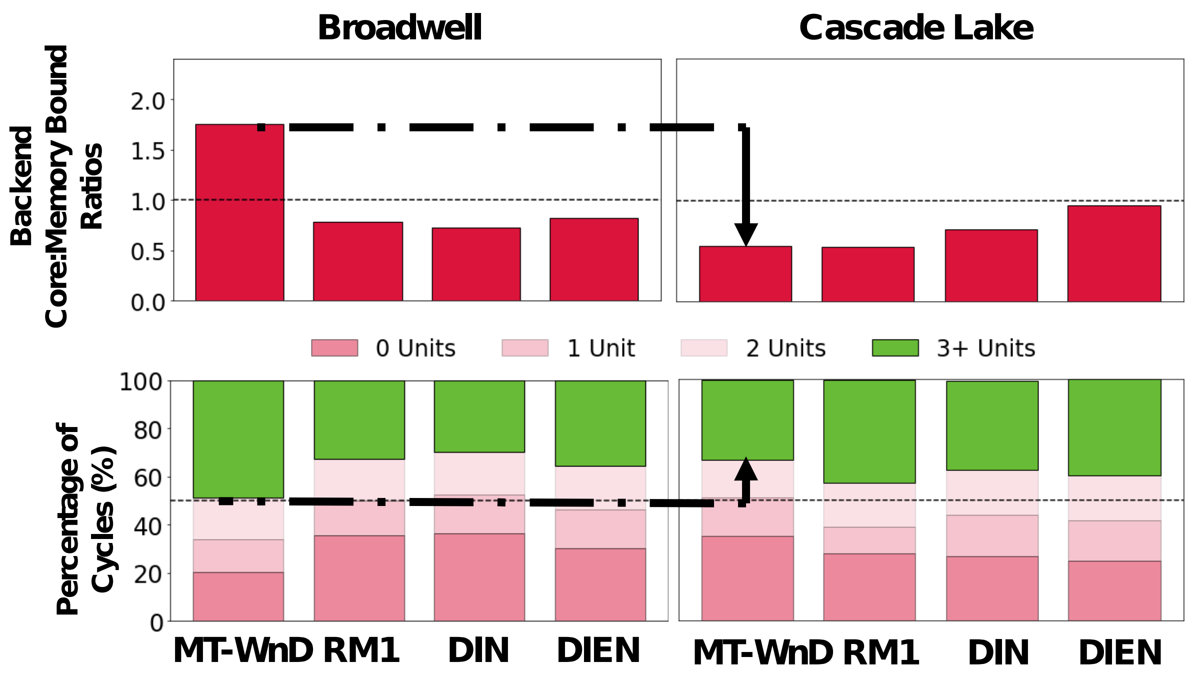
On Broadwell, FC-dominated models **are limited by insufficient functional units**



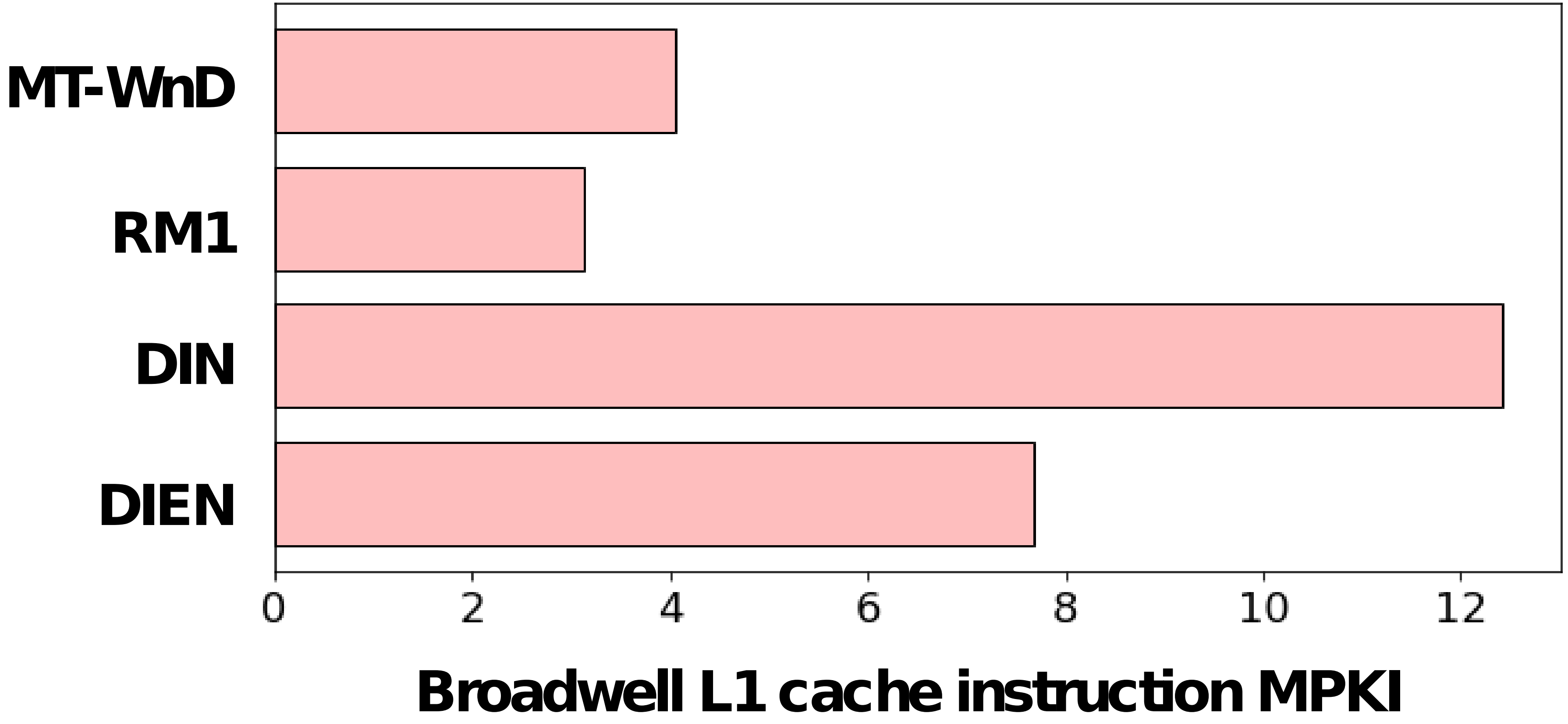
Available **functional units saturated**

On Cascade Lake, FC-dominated models **benefit from wider SIMD**, **shifting bottlenecks to memory subsystem**



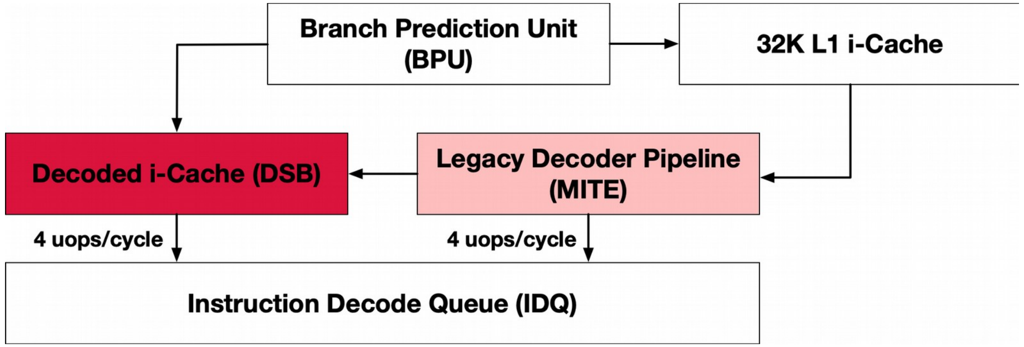


Attention-based models suffer from frontend latency (**L1 instruction-cache misses)**

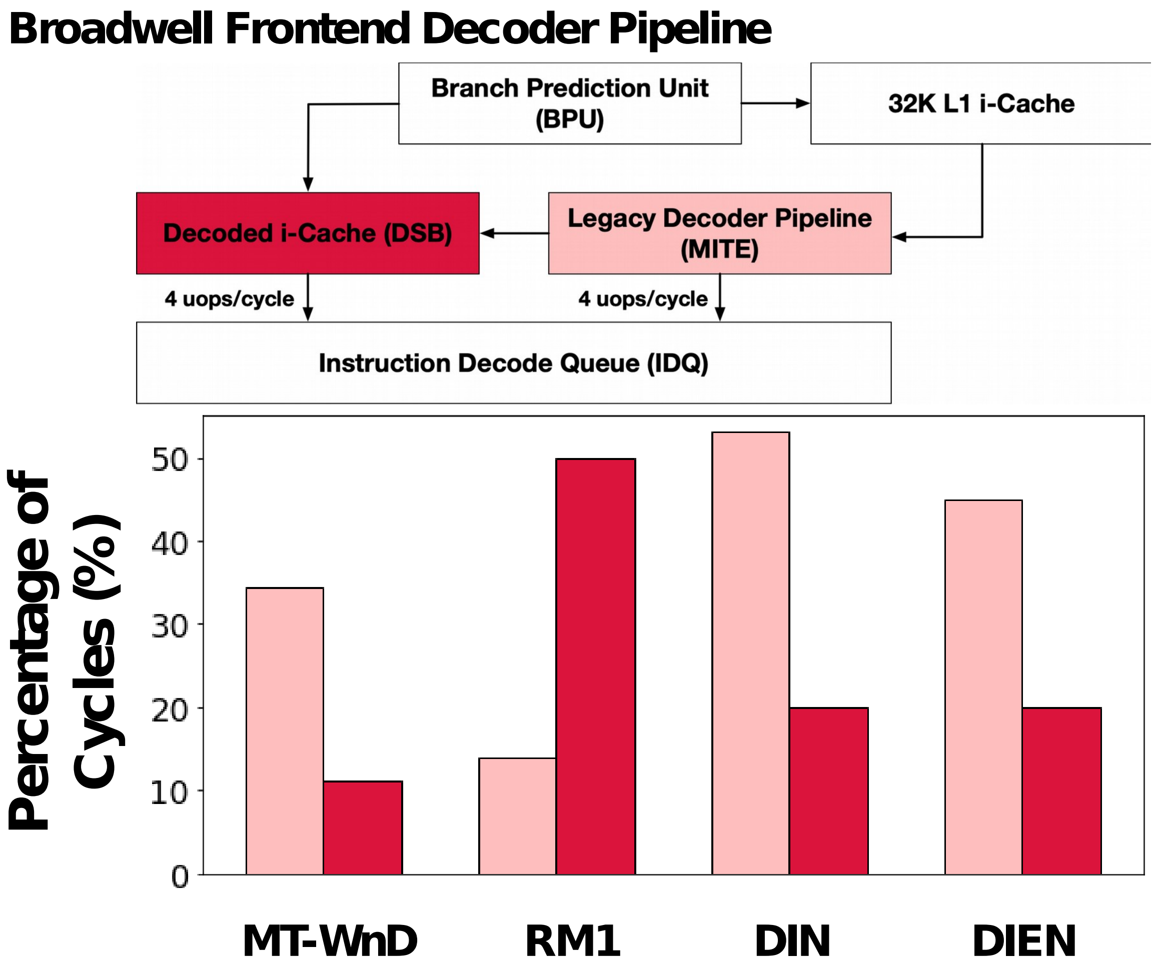


Models with more embedding table lookups suffer from **instruction decoder bottlenecks**

Broadwell Frontend Decoder Pipeline



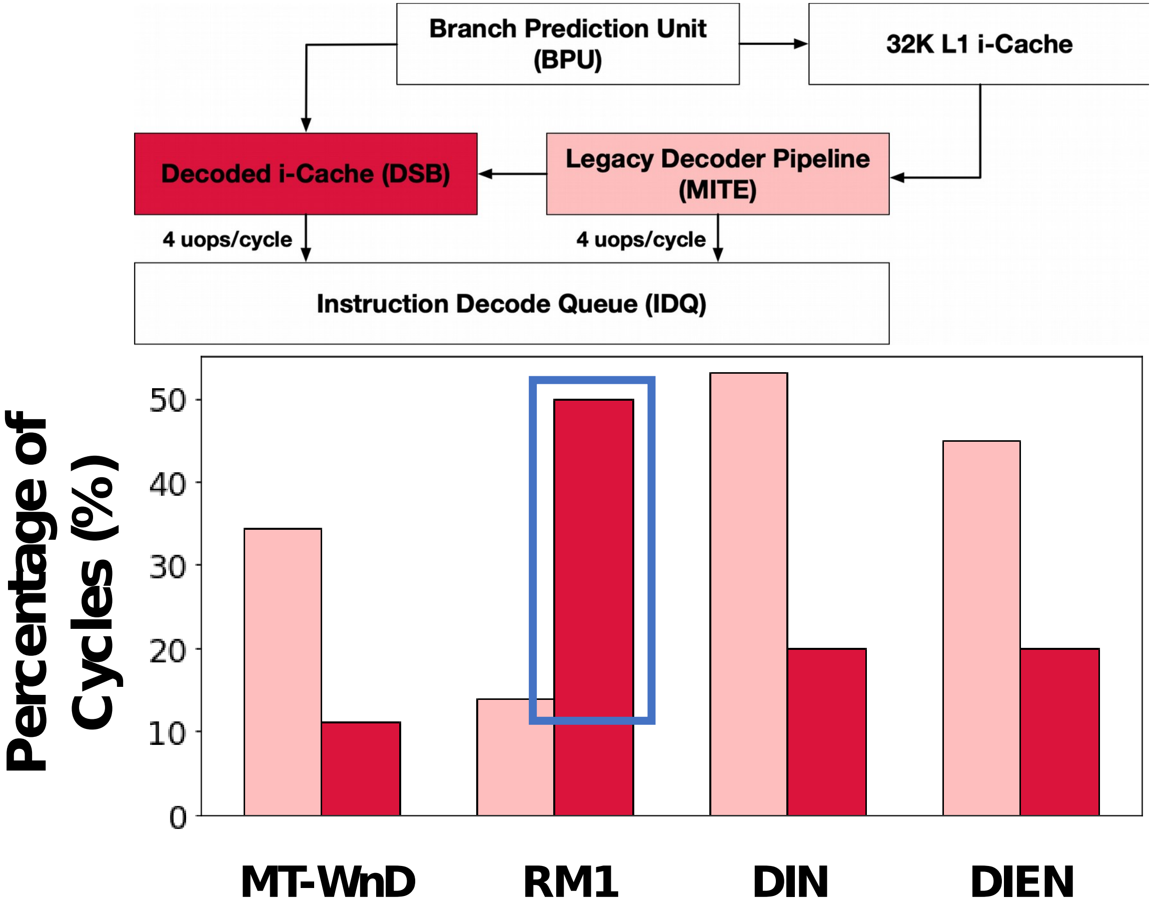
Models with more embedding table lookups suffer from **instruction decoder bottlenecks**



CPU Cycles limited due to Frontend Decoder Pipeline

Models with more embedding table lookups suffer from **instruction decoder bottlenecks**

Broadwell Frontend Decoder Pipeline



DSB ineffective for RM1!

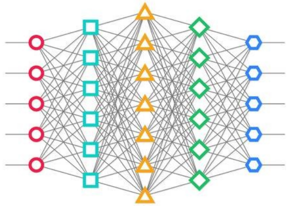
CPU Cycles limited due to Frontend Decoder Pipeline

Summary of Microarchitectural Effects

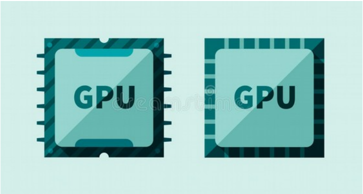
Type of Model	Microarchitectural Insight
FC Heavy	On Broadwell, **insufficient functional units** On Cascade Lake, **sub-optimal memory subsystem**
Attention Heavy	Frontend Latency L1 i-cache miss rate (**L1 i-MPKI**)
Embedding Heavy	Frontend Bandwidth Decoded i-cache (**DSB**)

Characterization

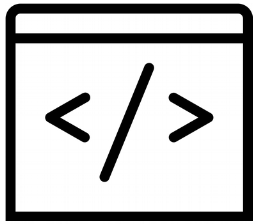
Question: What are the bottlenecks of each layer and how do they affect one another?



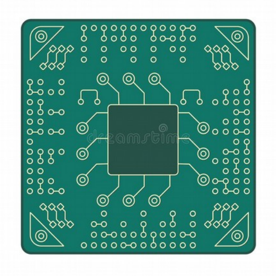
What do industry-representative algorithms (**<u>model architectures</u>**) look like?



What are the **<u>performance trends</u>** of deploying recommendation on CPUs and GPUs?



Can we explain the performance trends with **<u>software level operators</u>**?



Can we explain the performance trends with **<u>microarchitectural analysis</u>**?

Characterization

2020 IEEE International Symposium on Workload Characterization (IISWC)

Cross-Stack Workload Characterization of Deep Recommendation Systems

Samuel Hsia[1], Udit Gupta[1,2], Mark Wilkening[1],
Carole-Jean Wu[2], Gu-Yeon Wei[1], David Brooks[1]

More algorithms

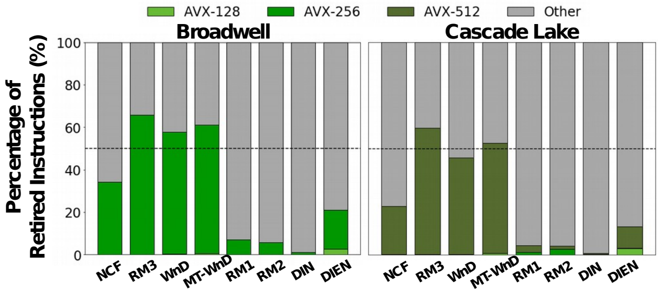
More characterization

Open-Source

Google Play

NETFLIX

More industry-representative deep recommendation models



More microarchitectural insights based on detailed PMU counter analysis

Model implementations and experiment scripts open-sourced: https://github.com/harvard-acc/DeepRecSys

This talk

Characterization

[IISWC '20]

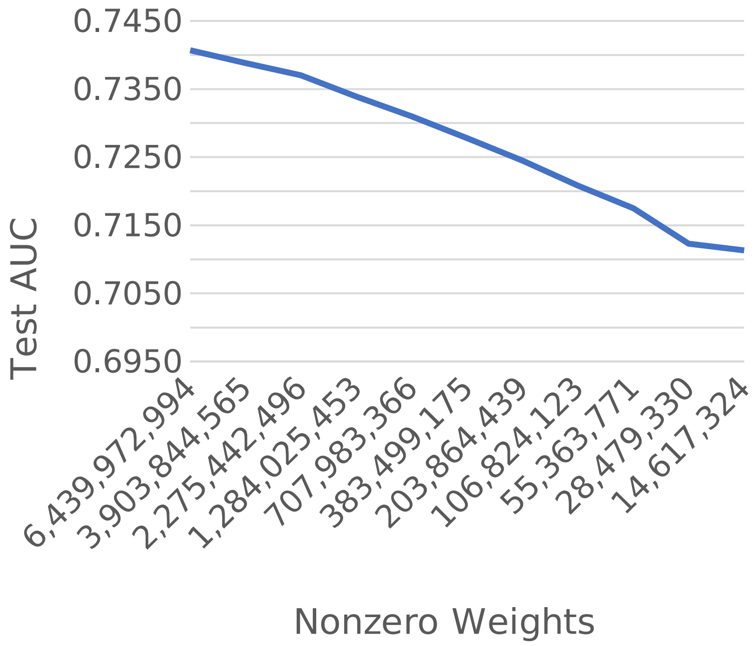
RecSSD

[ASPLOS '21]

Computational Trends in Recommendation

Computational Trends in Recommendation

More Features, More Accuracy

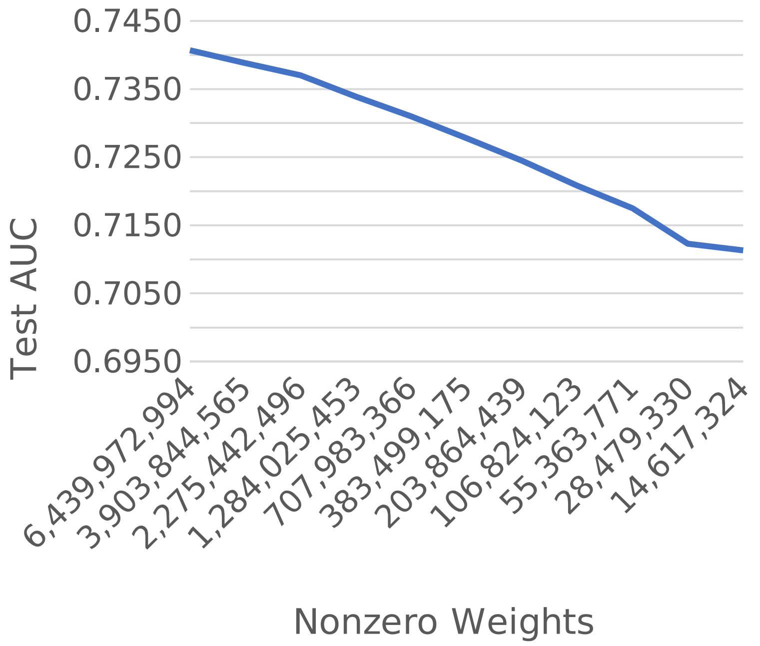


[1]

ing Massive Scale Deep Learning Ads Systems with GPUs and SSDs”, PeRSonAl at ISCA 2020, Weijie Zhao

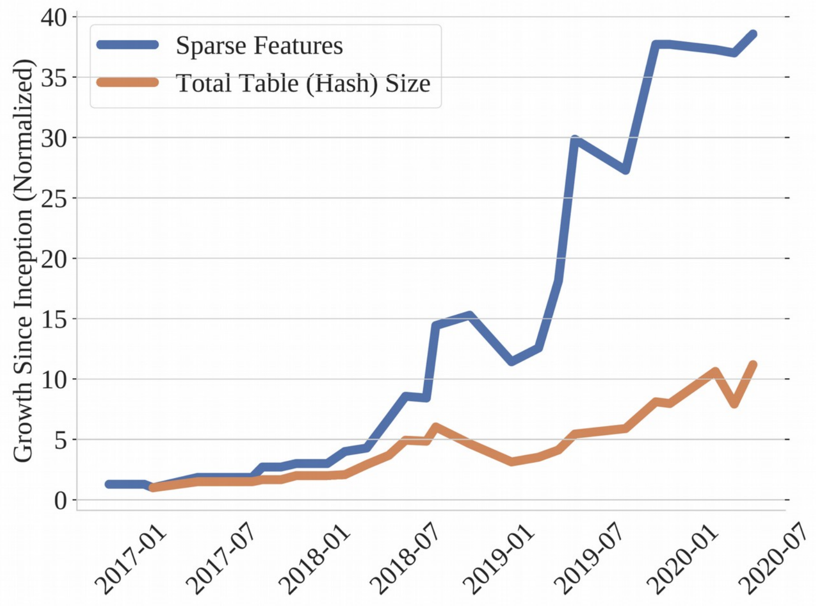
Computational Trends in Recommendation

More Features, More Accuracy



[1]

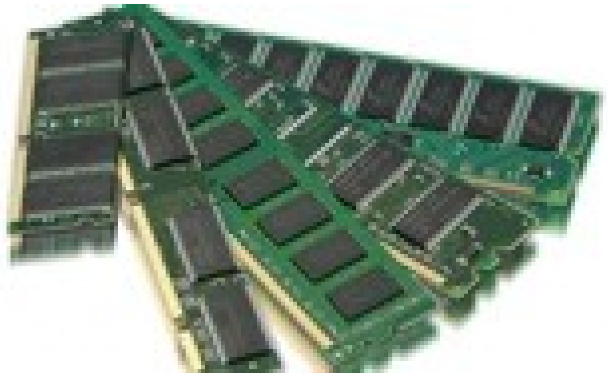
… And Memory Capacity



[2]

ing Massive Scale Deep Learning Ads Systems with GPUs and SSDs", PeRSonAl at ISCA 2020, Weijie Zhao
erstanding Capacity-Driven Scale-Out Neural Recommendation Inference", arXiv:2011.02084, Lui et. al.

High-Capacity Flash vs. DRAM



High-Capacity Flash vs. DRAM

Cost

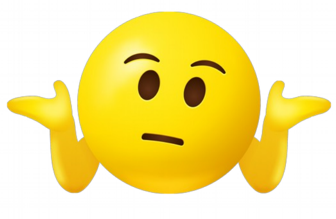
O(5-10X)

O(X)

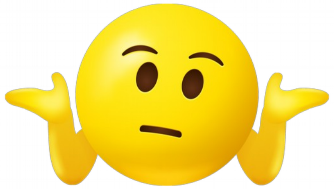
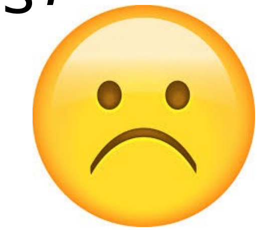
High-Capacity Flash vs. DRAM

	Cost	Read Latency
	O(5-10X)	O(10ns)
	O(X)	O(10us)
		

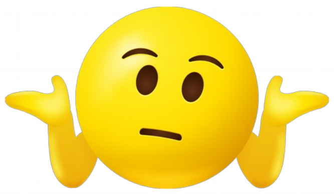
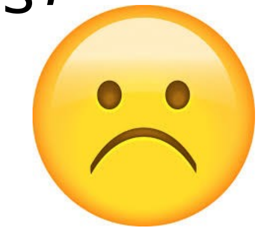
High-Capacity Flash vs. DRAM

	Cost	Read Latency	Write Latency
	O(5-10X)	O(10ns)	O(10ns)
	O(X)	O(10us)	O(1ms)
			

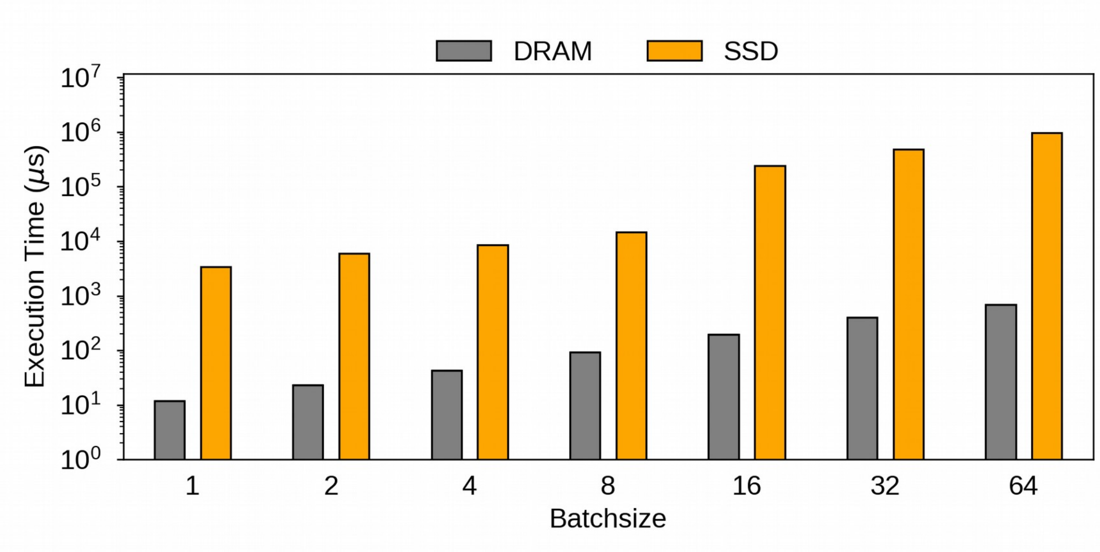
High-Capacity Flash vs. DRAM

	Cost	Read Latency	Write Latency	Random 4KB Read B/W
	O(5-10X)	O(10ns)	O(10ns)	O(75GB/s)
	O(X)	O(10us)	O(1ms)	O(2-3GB/s)
				

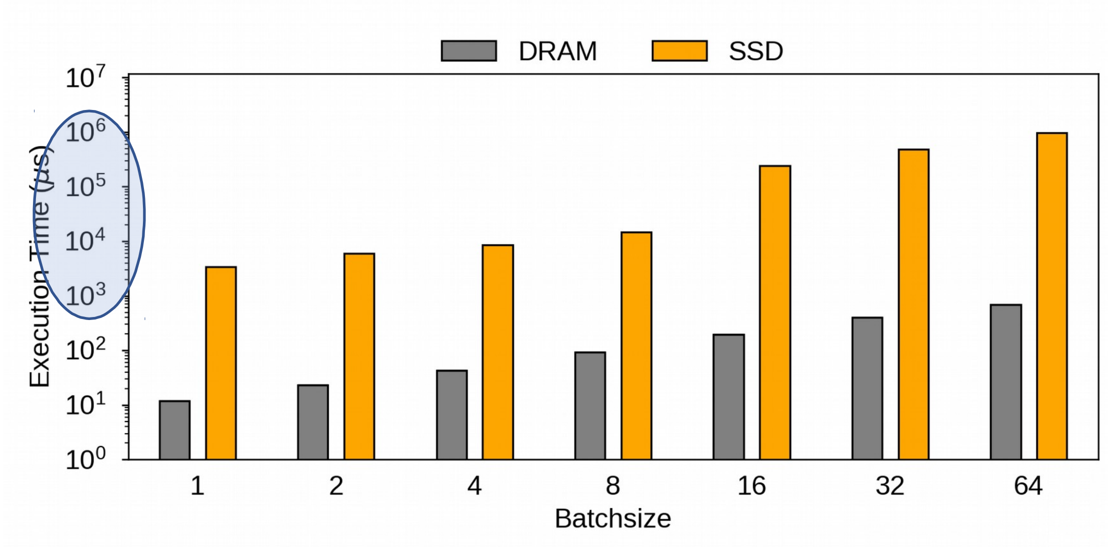
High-Capacity Flash vs. DRAM

	Cost	Read Latency	Write Latency	Random 4KB Read B/W	Random 128B
	O(5-10X)	O(10ns)	O(10ns)	O(75GB/s)	O(75GB/s)
	O(X)	O(10us)	O(1ms)	O(2-3GB/s)	O(10MB/s)
					

Flash SSDs for Recommendation



Flash SSDs for Recommendation



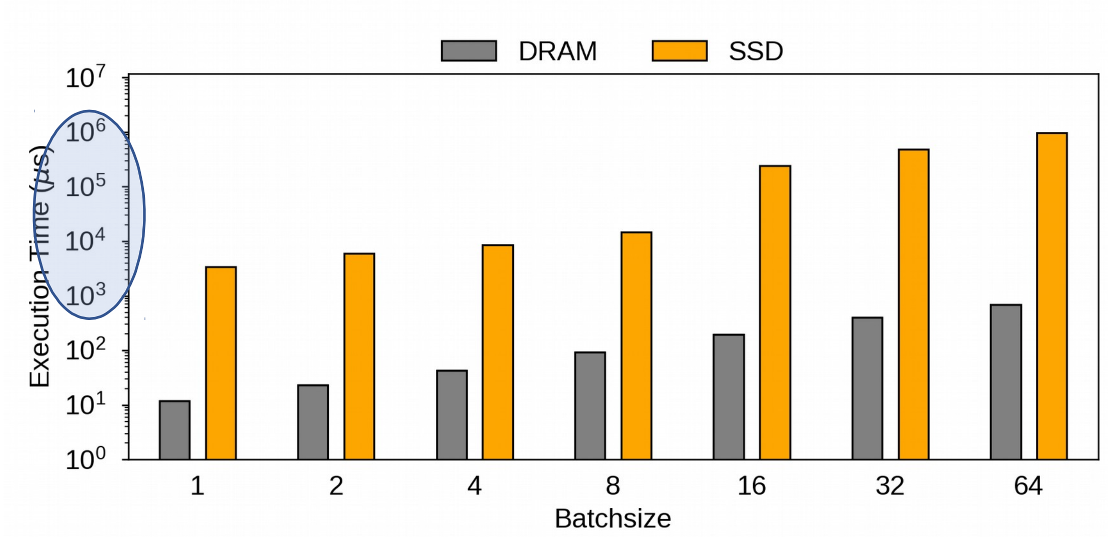
3 Orders of magnitude slower embedding operations

Low Bandwidth

Page Size vs. Access Size

Software Overheads in PCIe Access

Flash SSDs for Recommendation

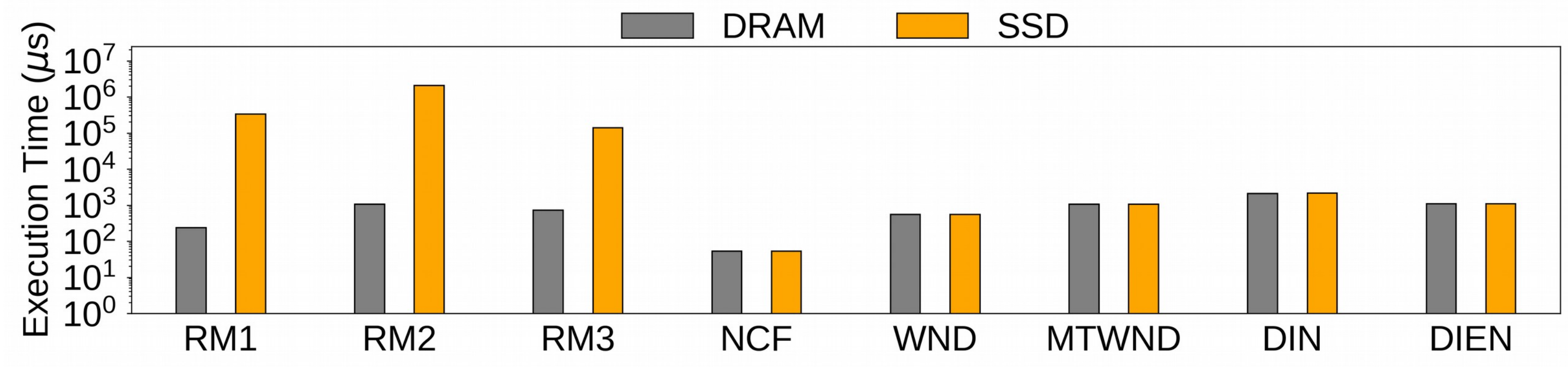


3 Orders of magnitude slower embedding operations

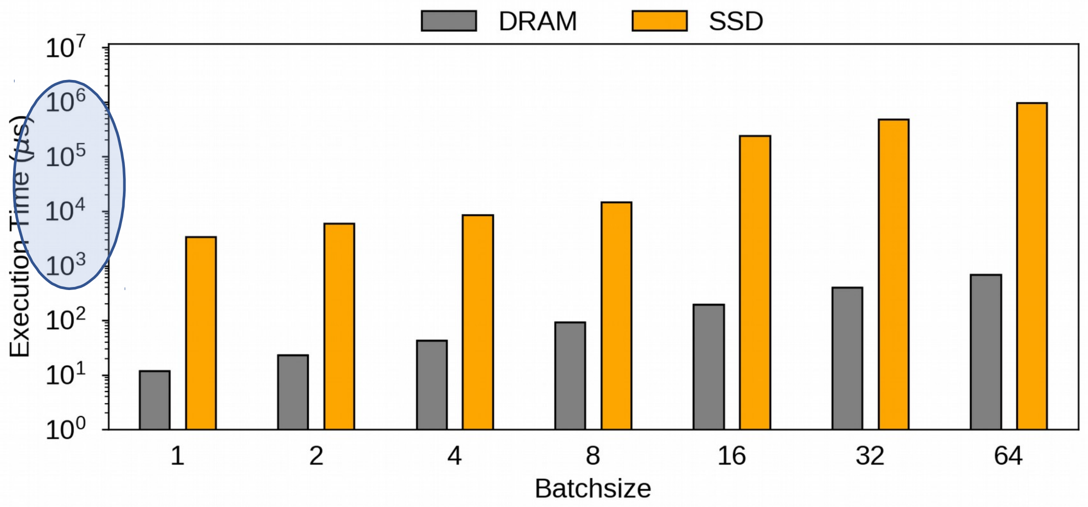
Low Bandwidth

Page Size vs. Access Size

Software Overheads in PCIe Access



Flash SSDs for Recommendation

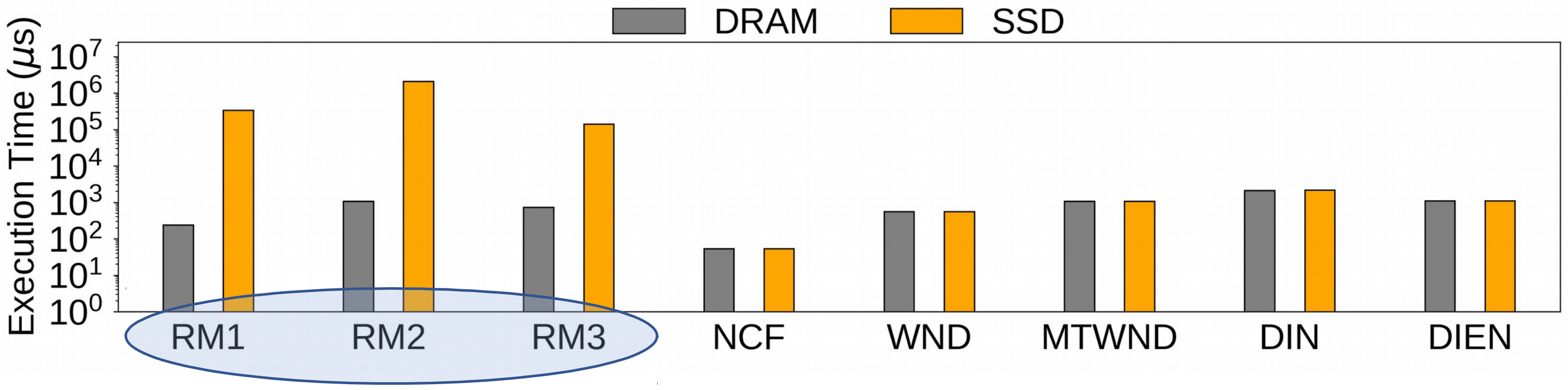


3 Orders of magnitude slower embedding operations

Low Bandwidth

Page Size vs. Access Size

Software Overheads in PCIe Access



Significant slowdown in embedding dominated models

Problems with Flash for Recommendation

Low Bandwidth

Page and Access Size Mismatch

Software Overheads in PCIe Access

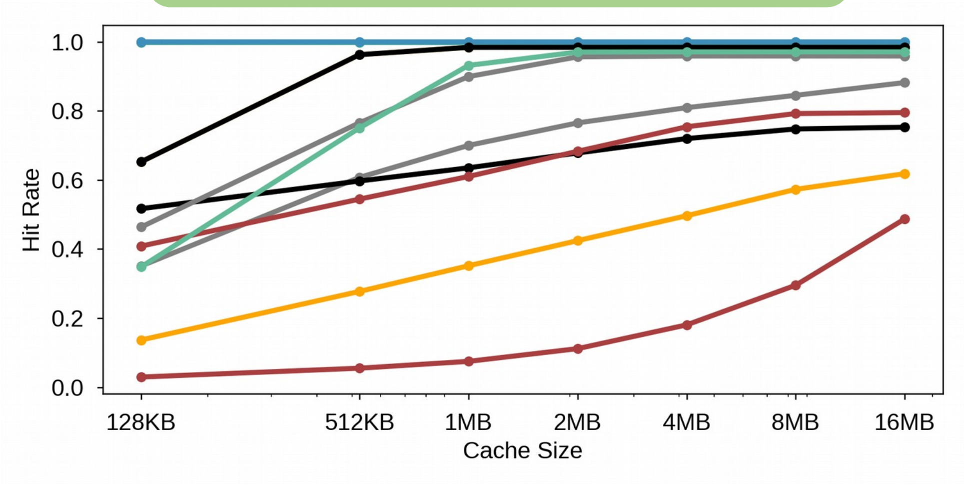
Problems with Flash for Recommendation

Low Bandwidth

Page and Access Size Mismatch

Software Overheads in PCIe Access

DRAM Caching

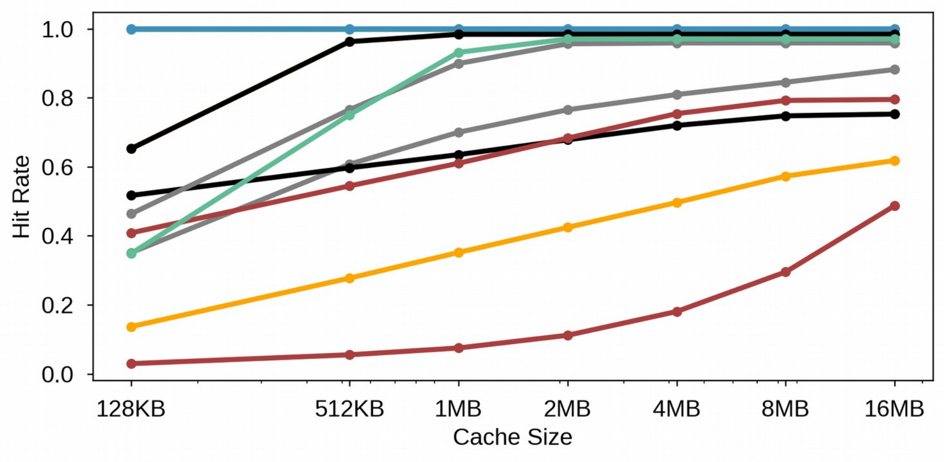


Hit-rates vary wildly across embedding tables from 10% to 90%

Problems with Flash for Recommendation

Low Bandwidth

DRAM Caching



Hit-rates vary wildly across embedding tables from 10% to 90%

Page and Access Size Mismatch

Table re-ordering, advanced caching

Bandana [1]

Smaller flash page sizes in SSD hardware, byte addressable NVM

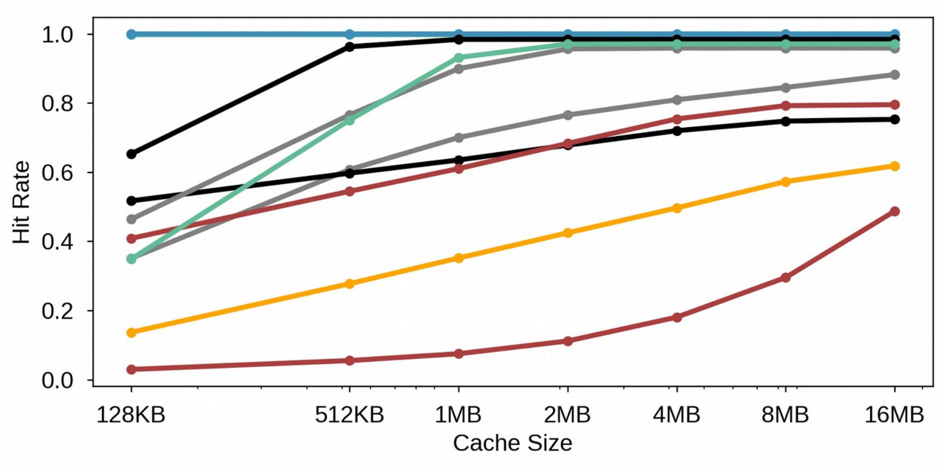
Software Overheads in PCIe Access

[1] "Bandana: Using Non-volatile Memory for Storing Deep Learning Models", SysML 19, Eisenman et. al.

Problems with Flash for Recommendation

Low Bandwidth

DRAM Caching



Hit-rates vary wildly across embedding tables from 10% to 90%

Page and Access Size Mismatch

Table re-ordering, advanced caching

Bandana [1]

Smaller flash page sizes in SSD hardware, byte addressable NVM

Software Overheads in PCIe Access

Near Data Processing

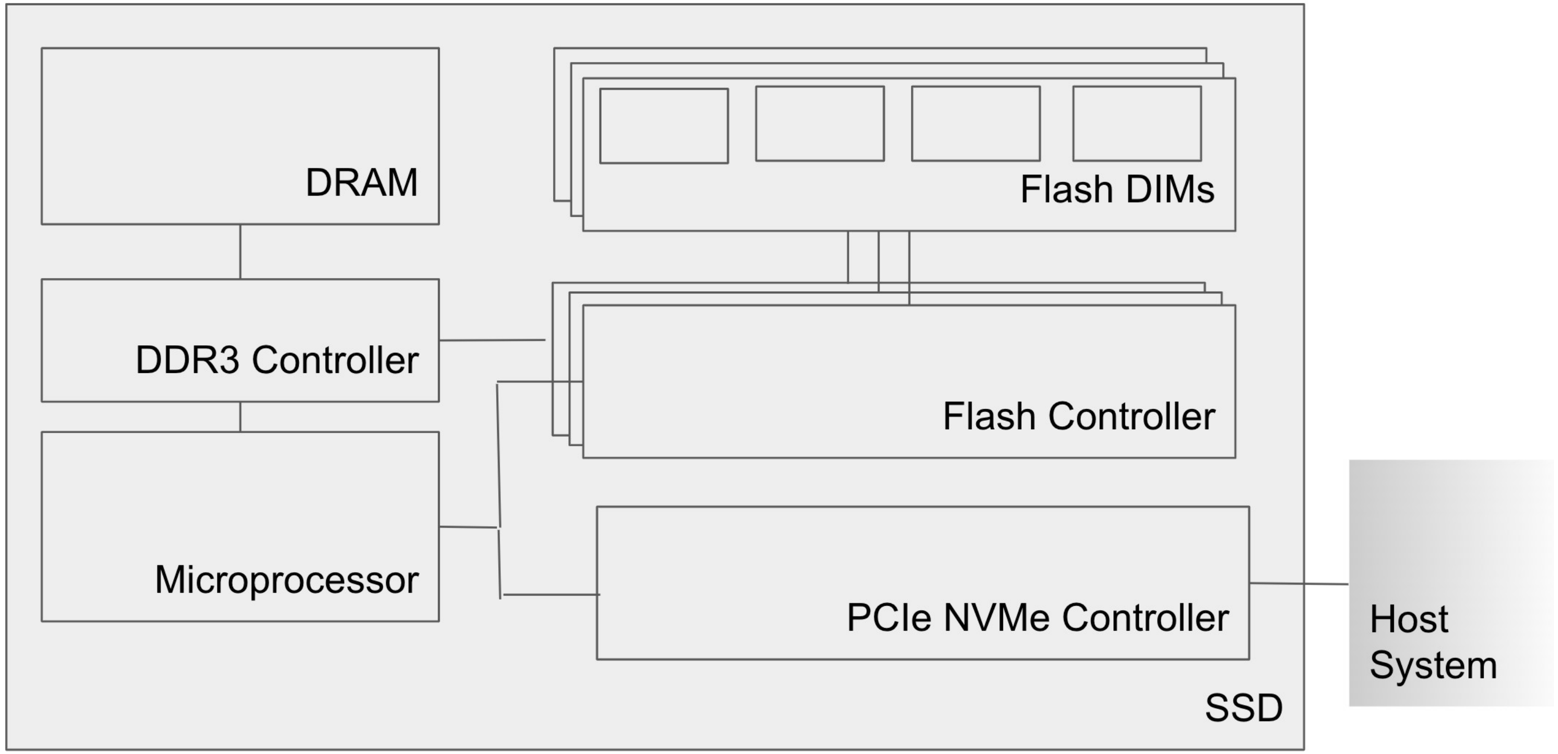
RecSSD [2]

[1] "Bandana: Using Non-volatile Memory for Storing Deep Learning Models", SysML 19, Eisenman et. al.

[2] "RecSSD: Near Data Processing forSolid State Drive Based Recommendation Inference", ASPLOS 2021, Wilkening et. al.

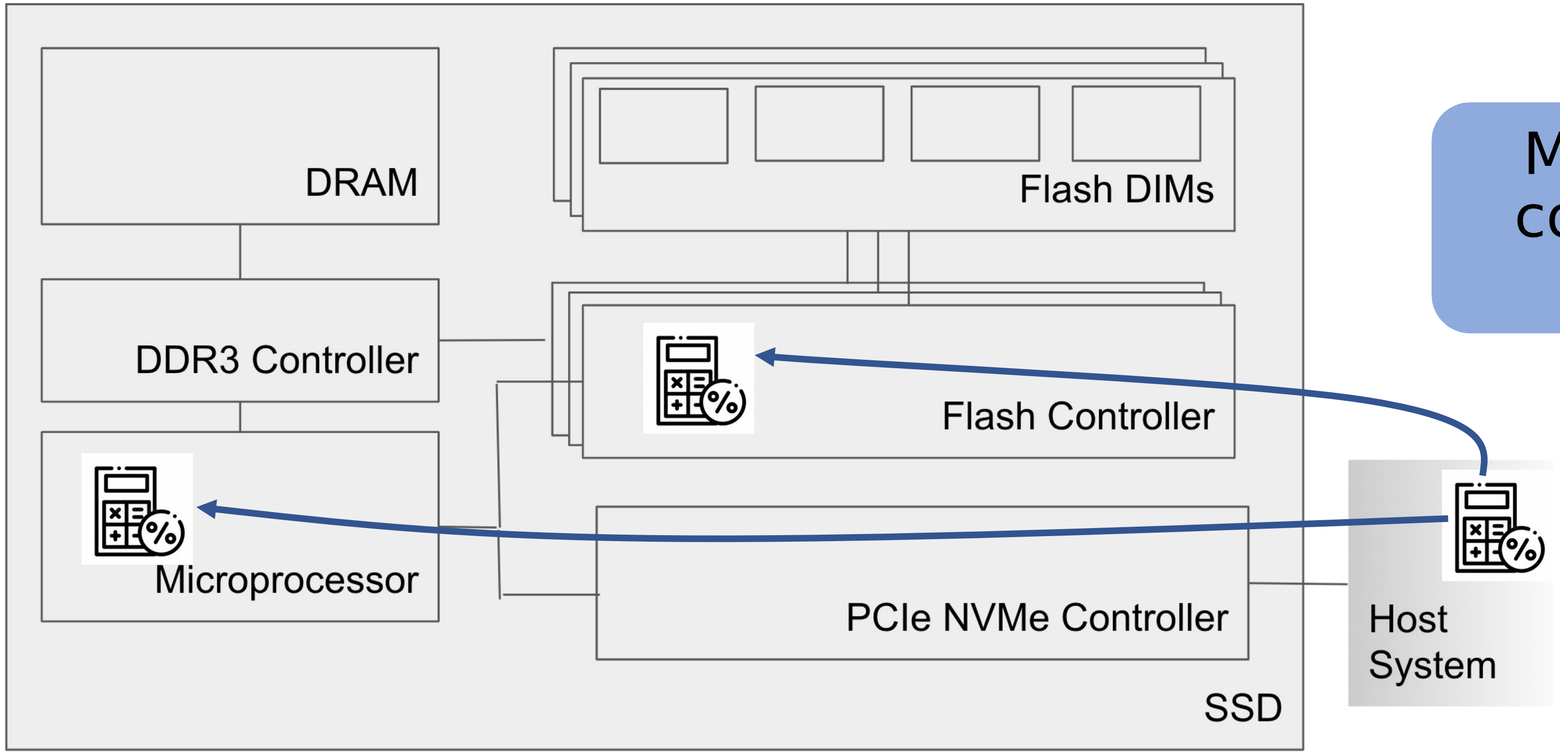
RecSSD: Efficient NDP for Recommendation

Question: What is NDP, why does it work, and when does it work?



RecSSD: Efficient NDP for Recommendation

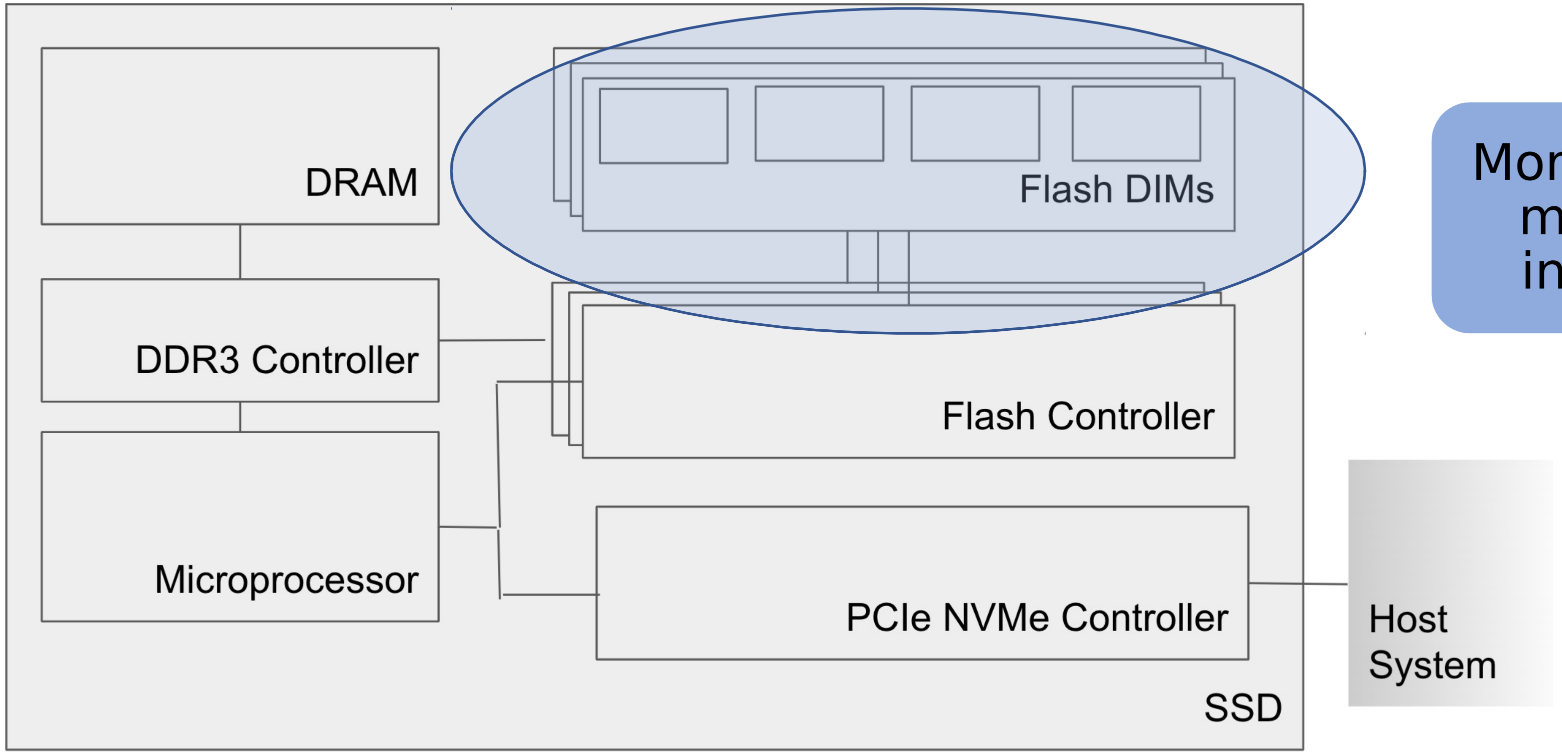
Question: What is Near Data Processing, why does it work, and when does it work?



Move application specific computation closer to the data

RecSSD: Efficient NDP for Recommendation

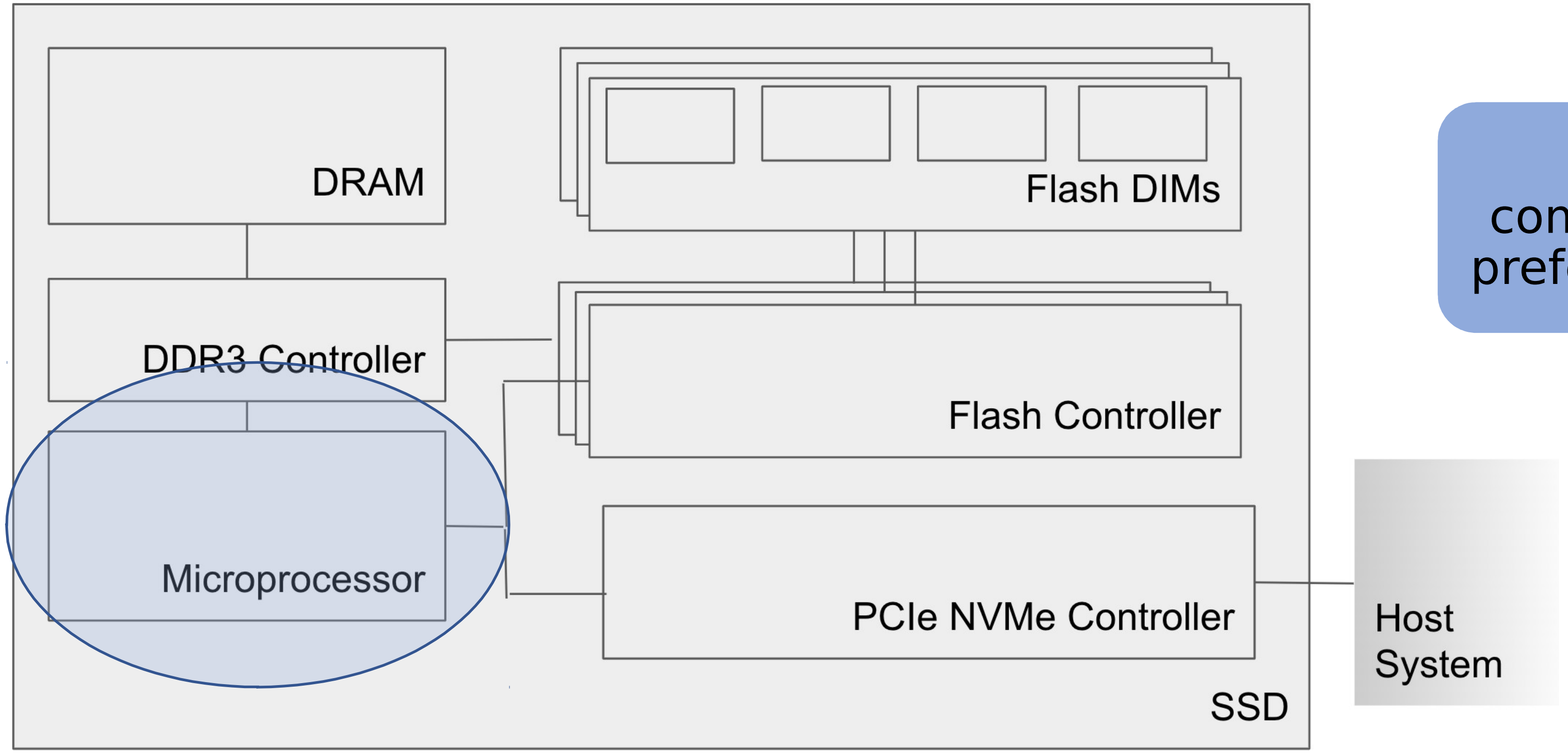
Question: What is NDP, why does it work, and when does it work?



More efficiently leverage internal memory level parallelism, for increased internal bandwidth

RecSSD: Efficient NDP for Recommendation

Question: What is NDP, why does it work, and when does it work?



Requires data intensive, computationally light tasks, which preferably reduce to simpler results

RecSSD: Efficient NDP for Recommendation

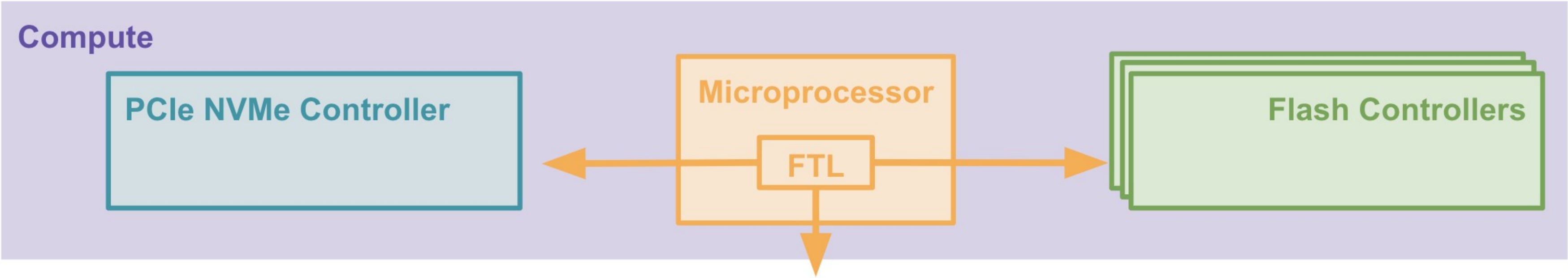
General Purpose NDP

- Built for a wide array of computational tasks
- Typically relies on highly customized hardware accelerators, SSD firmware, host drivers, and programming interfaces

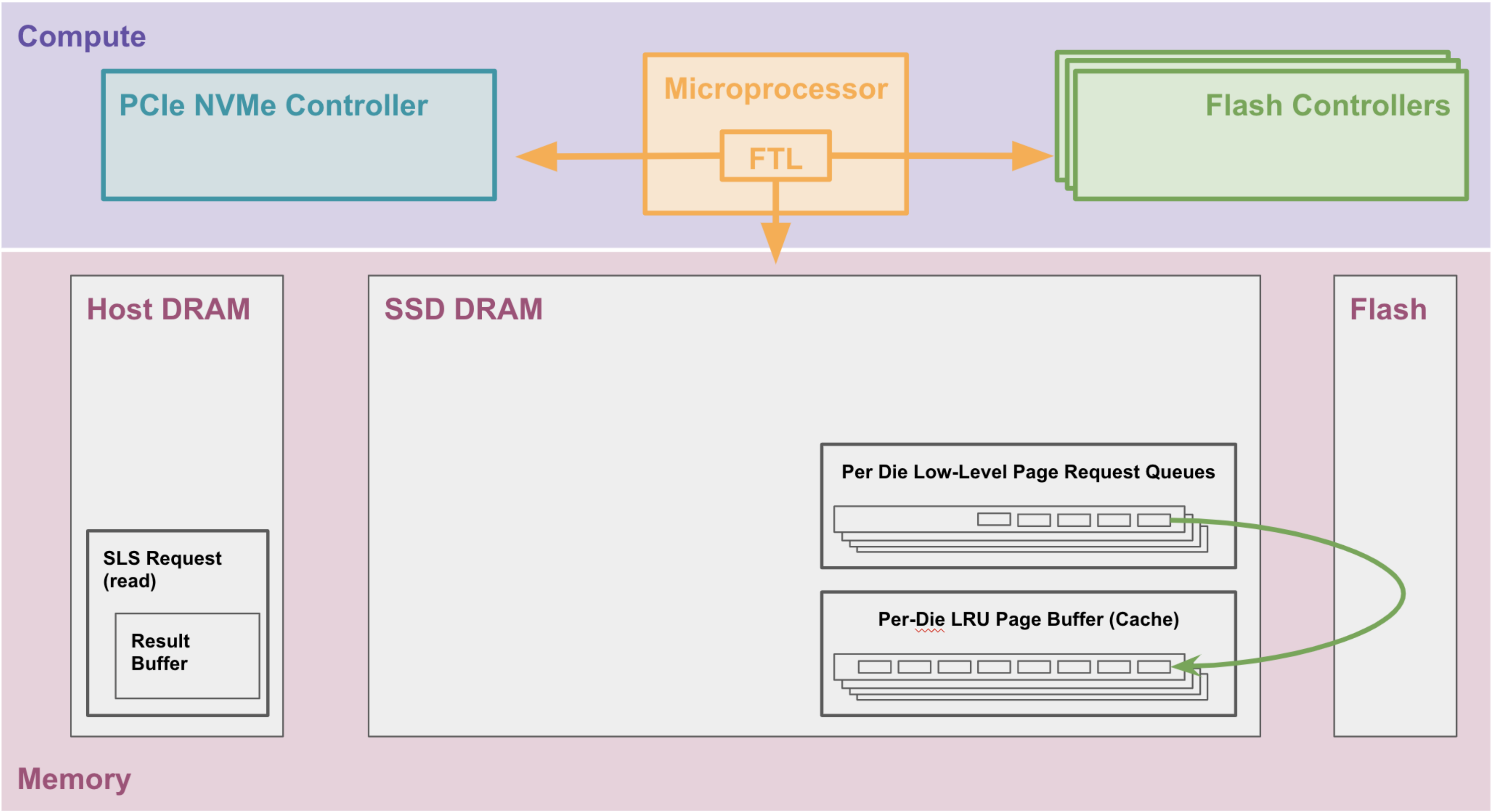
RecSSD

- Built for recommendation
- Uses commodity hardware
- Built entirely within the FTL
- Uses standard NVMe interfaces and minimal driver modifications
- Minimalist, cost efficient, low latency

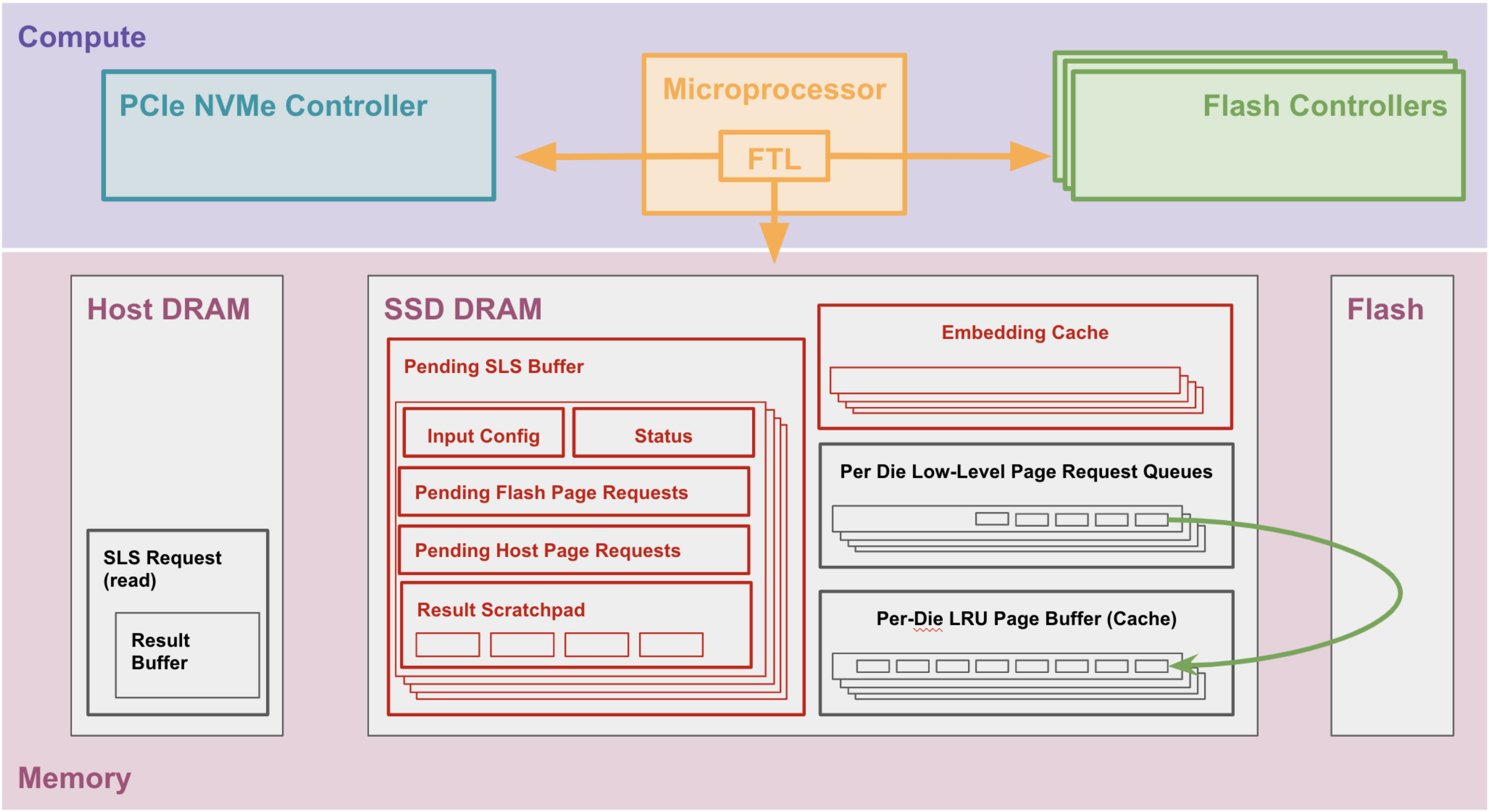
RecSSD Design Overview



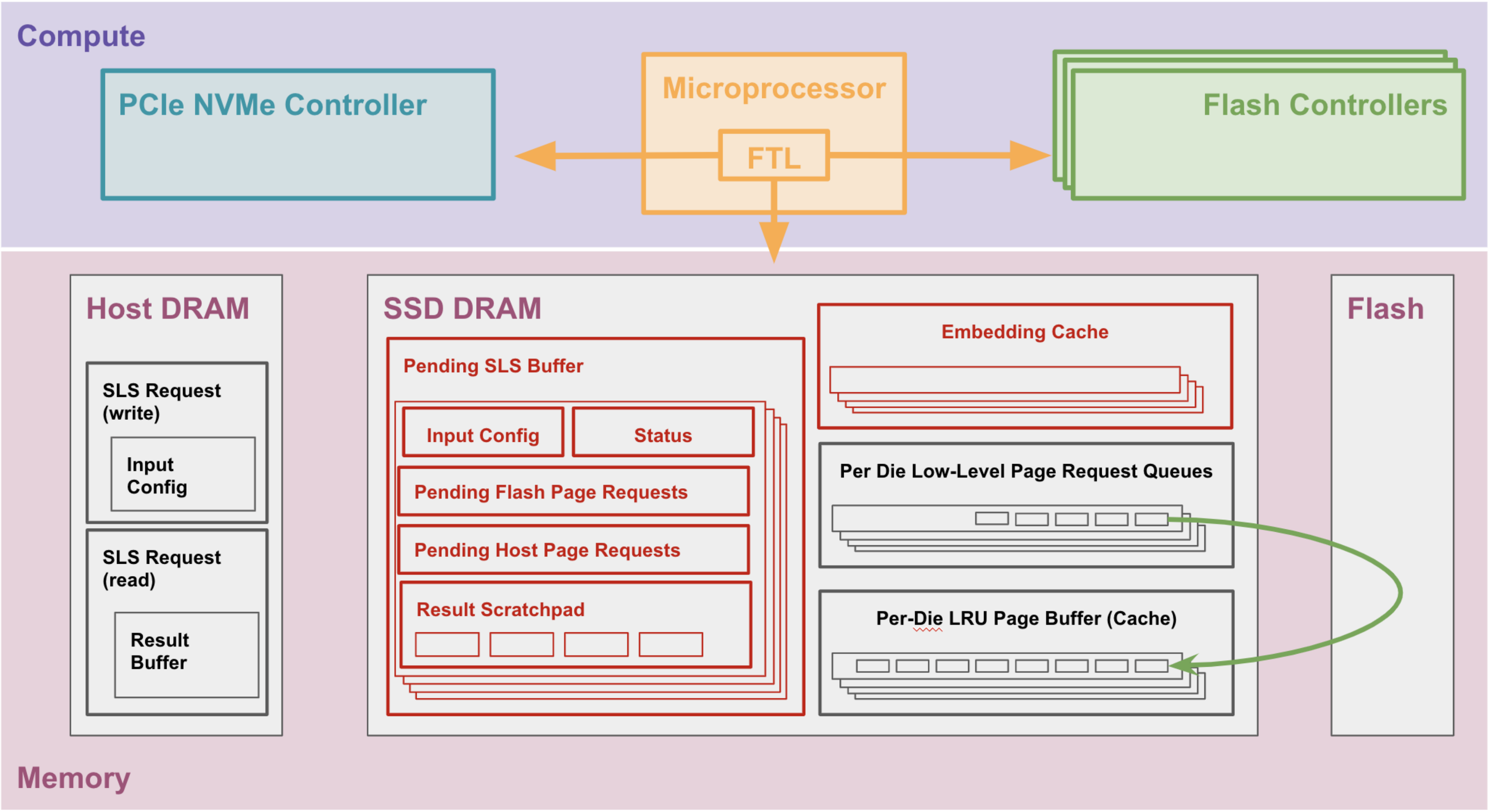
RecSSD Design Overview



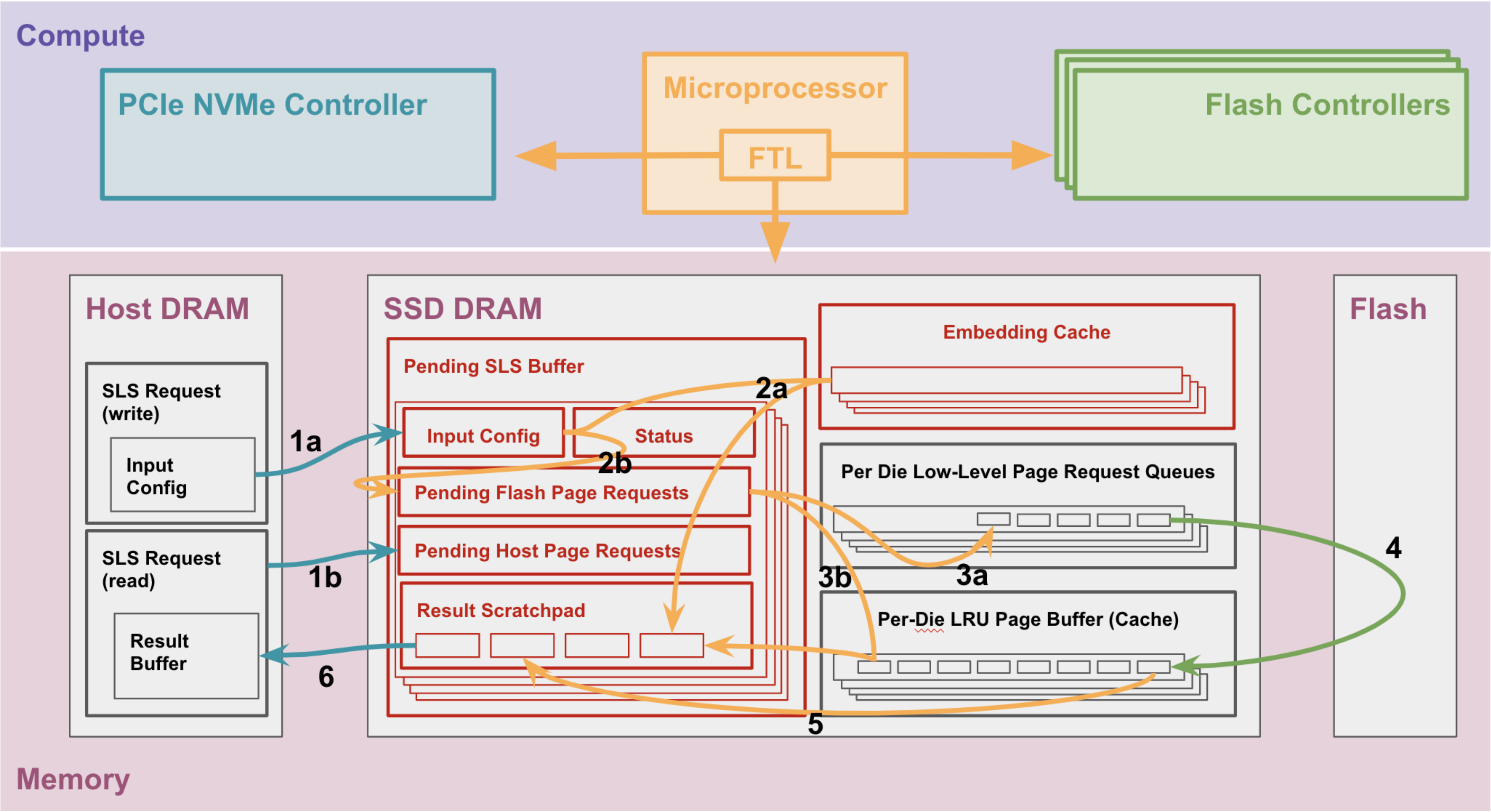
RecSSD Design Overview



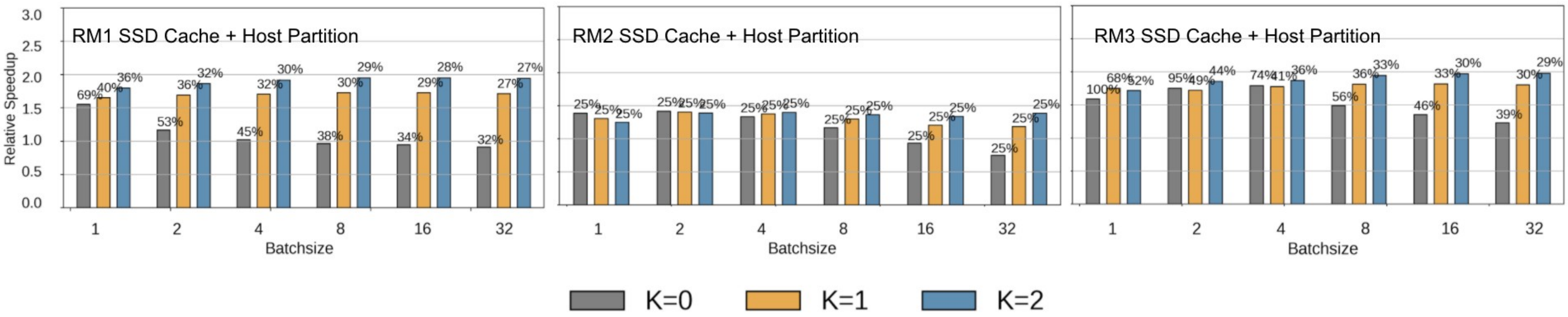
RecSSD Design Overview



RecSSD Design Overview



RecSSD Performance

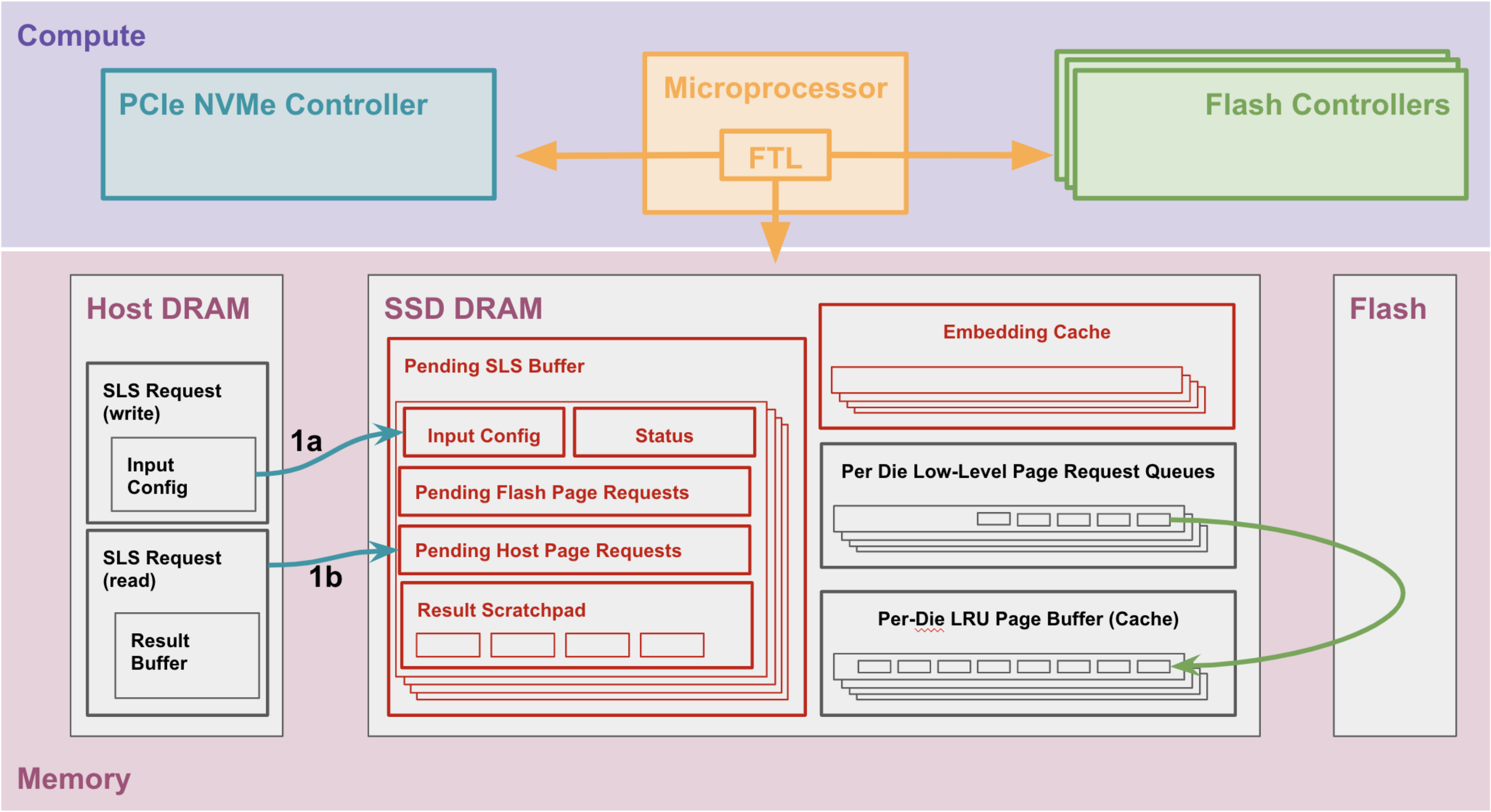


Up to 2x inference latency improvement alongside conventional caching techniques

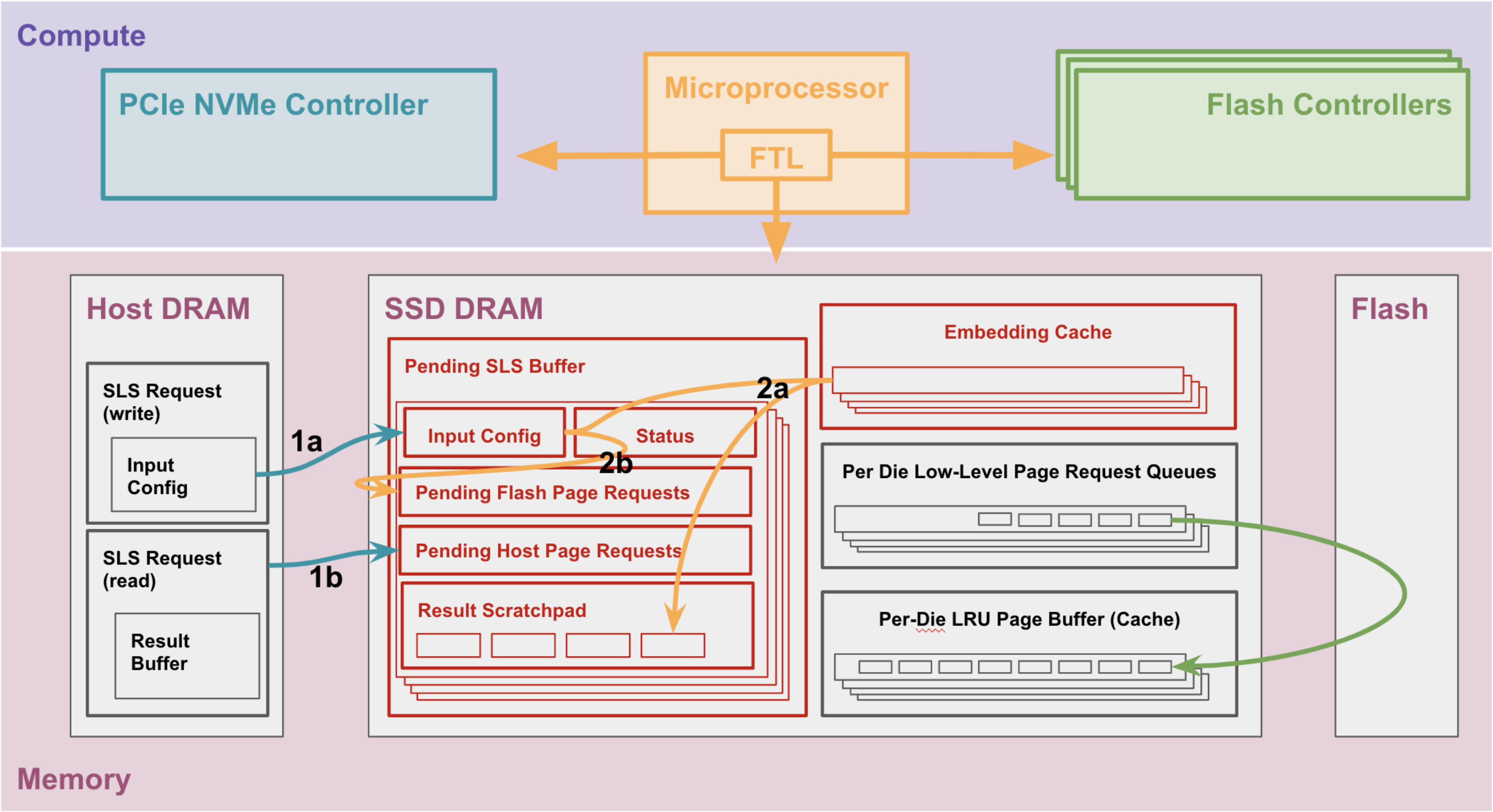
Thanks for listening!

Questions?

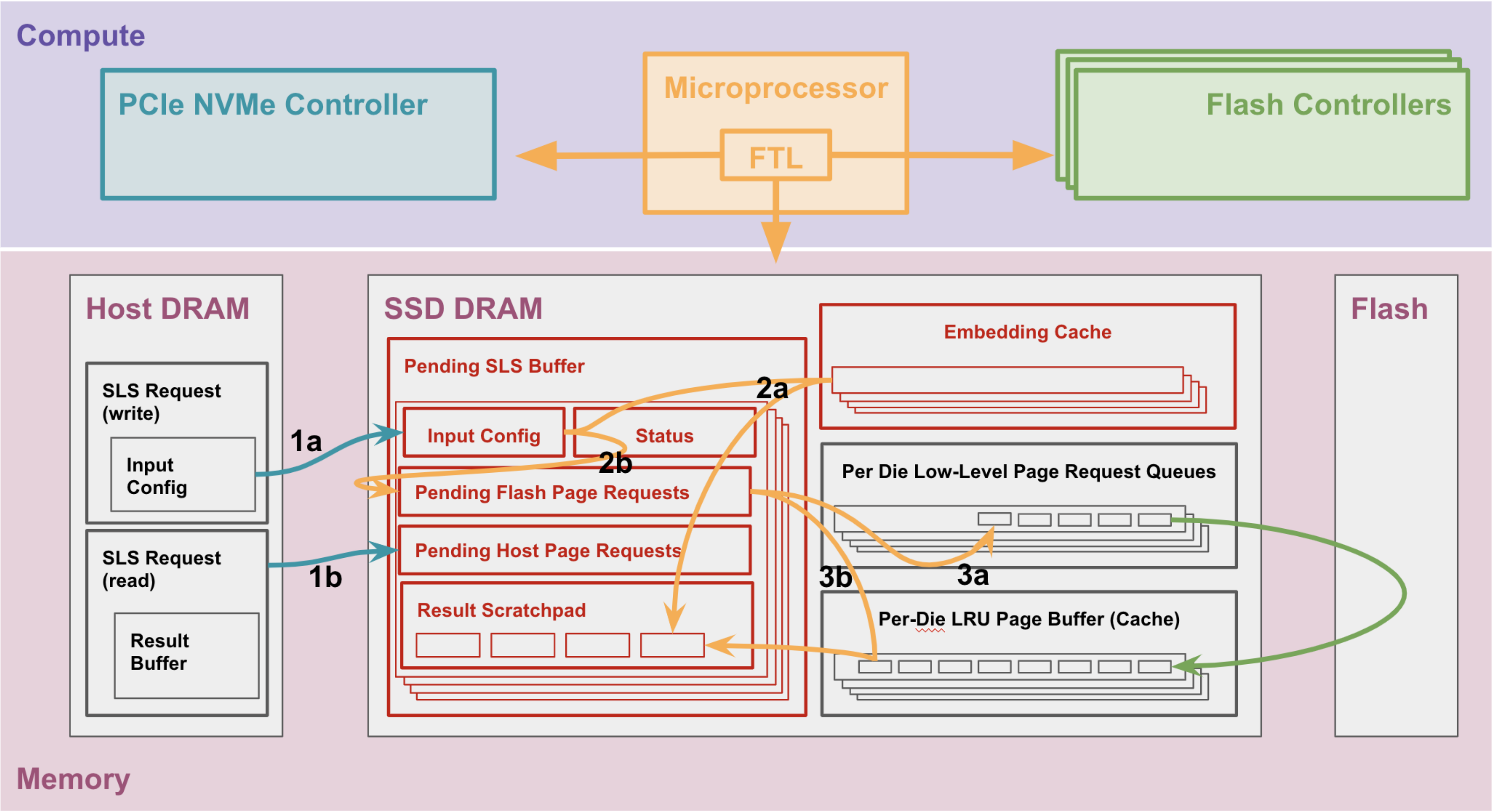
RecSSD Design Overview



RecSSD Design Overview



RecSSD Design Overview



RecSSD Design Overview