Lecture 3 of the

MLArchSys Seminar

Instructor: Thaleia Dimitra Doudali

Assistant Professor at IMDEA Software Institute

Universidad Politécnica de Madrid (UPM)

March 2023

Outline of Today's Lecture

Systems Software ← ML ***for*** Systems — Machine Learning

Today's Paper:

Kleio: A Hybrid Memory Page Scheduler with Machine Intelligence

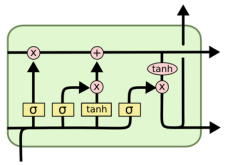
Thaleia Dimitra Doudali
Georgia Institute of Technology
thdoudali@gatech.edu

Sergey Blagodurov
Advanced Micro Devices, Inc.
Sergey.Blagodurov@amd.com

Abhinav Vishnu
Advanced Micro Devices, Inc.
Abhinav.Vishnu@amd.com

Sudhanva Gurumurthi
Advanced Micro Devices, Inc.
Sudhanva.Gurumurthi@amd.com

Ada Gavrilovska
Georgia Institute of Technology
ada@cc.gatech.edu



LSTMs

for Hybrid Memory Management (HMem Management)

Lecture Outline:

- Hybrid Memories
- Hybrid Memory Management
- How to Integrate ML?
- System Design
- Evaluation
- Lessons Learned

Outline of Today's Lecture

Systems Software ← ML *for* Systems — Machine Learning

Today's Paper:

Kleio: A Hybrid Memory Page Scheduler with Machine Intelligence

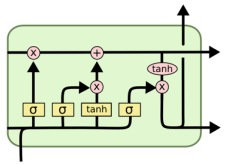
Thaleia Dimitra Doudali
Georgia Institute of Technology
thdoudali@gatech.edu

Sergey Blagodurov
Advanced Micro Devices, Inc.
Sergey.Blagodurov@amd.com

Abhinav Vishnu
Advanced Micro Devices, Inc.
Abhinav.Vishnu@amd.com

Sudhanva Gurumurthi
Advanced Micro Devices, Inc.
Sudhanva.Gurumurthi@amd.com

Ada Gavrilovska
Georgia Institute of Technology
ada@cc.gatech.edu



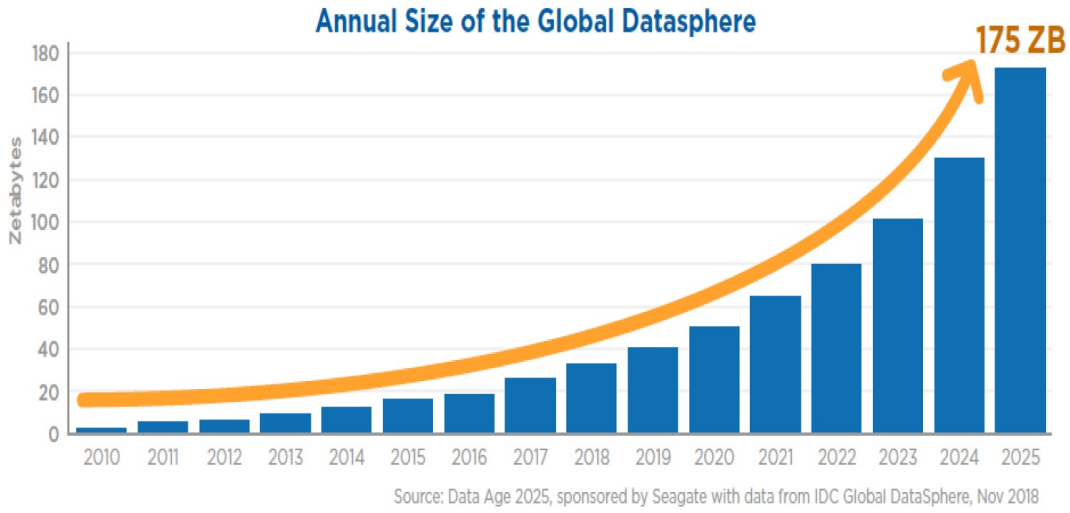
LSTMs

for Hybrid Memory Management (HMem Management)

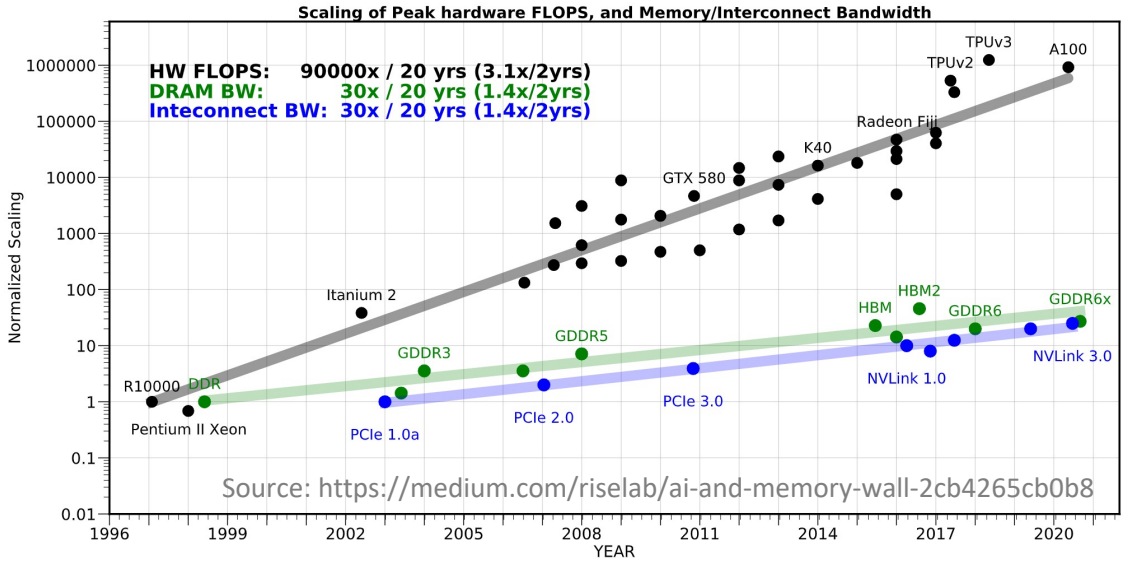
Lecture Outline:

- Hybrid Memories
- Hybrid Memory Management
- How to Integrate ML?
- System Design
- Evaluation
- Lessons Learned

Big Data hits the Memory Wall



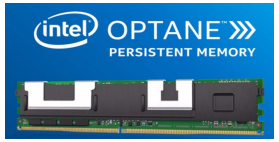
Applications generate huge amounts of data.



Processing speed grows faster than memory and data transfer speed.

Need for **larger** and **faster** memory configurations.

New Memory Technologies

Characteristic	Technology	Vendors
Low Latency	MRAM	
Uniform Latency	DRAM	
High Bandwidth	HBM	
Persistent / Non Volatile	PMEM / NVM	

Application Performance	200 PF
Number of Nodes	4,608
Node performance	42 TF
Memory per Node	512 GB DDR4 + 96 GB HBM2
NV memory per Node	1600 GB
Total System Memory	>10 PB DDR4 + HBM2 + Non-volatile
Processors	2 IBM POWER9™ 9,216 CPUs 6 NVIDIA Volta™ 27,648 GPUs
File System	250 PB, 2.5 TB/s, GPFS™
Power Consumption	13 MW
Interconnect	Mellanox EDR 100G InfiniBand
Operating System	Red Hat Enterprise Linux (RHEL) version 7.4

Example Configuration of a Supercomputer.

We are in the era of **Hybrid Memory** Systems. A mix of different technologies at different speeds / capacities / costs.

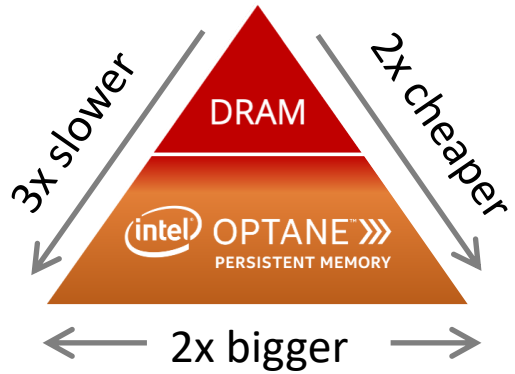
Hybrid Memory Configurations

In today's paper we assume a hybrid memory system with **DRAM** and **NVM** (Non Volatile Memory).

The NVM actual product was released by Intel in 2019, after the paper was published.

So, the paper had to assume various possible configurations, e.g., capacity ratios.

Intel Optane is packaged together with DRAM.



Source: memverge.com

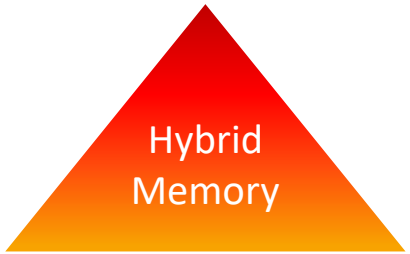
Guess what? The product got discontinued in 2022..

Intel kills off Optane Memory, writes off $559 million inventory

In a terrible quarter for the chip giant

July 29, 2022 By: Sebastian Moss 1 Comment

But this is not the end for hybrid memory configurations.
It is a mix of **any** number of **fast vs. slow**, **small vs. big** memories.



Outline of Today's Lecture

Systems Software ← ML ***for*** Systems — Machine Learning

Today's Paper:

Kleio: A Hybrid Memory Page Scheduler with Machine Intelligence

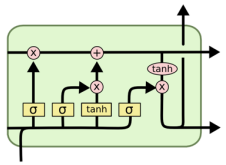
Thaleia Dimitra Doudali
Georgia Institute of Technology
thdoudali@gatech.edu

Sergey Blagodurov
Advanced Micro Devices, Inc.
Sergey.Blagodurov@amd.com

Abhinav Vishnu
Advanced Micro Devices, Inc.
Abhinav.Vishnu@amd.com

Sudhanva Gurumurthi
Advanced Micro Devices, Inc.
Sudhanva.Gurumurthi@amd.com

Ada Gavrilovska
Georgia Institute of Technology
ada@cc.gatech.edu



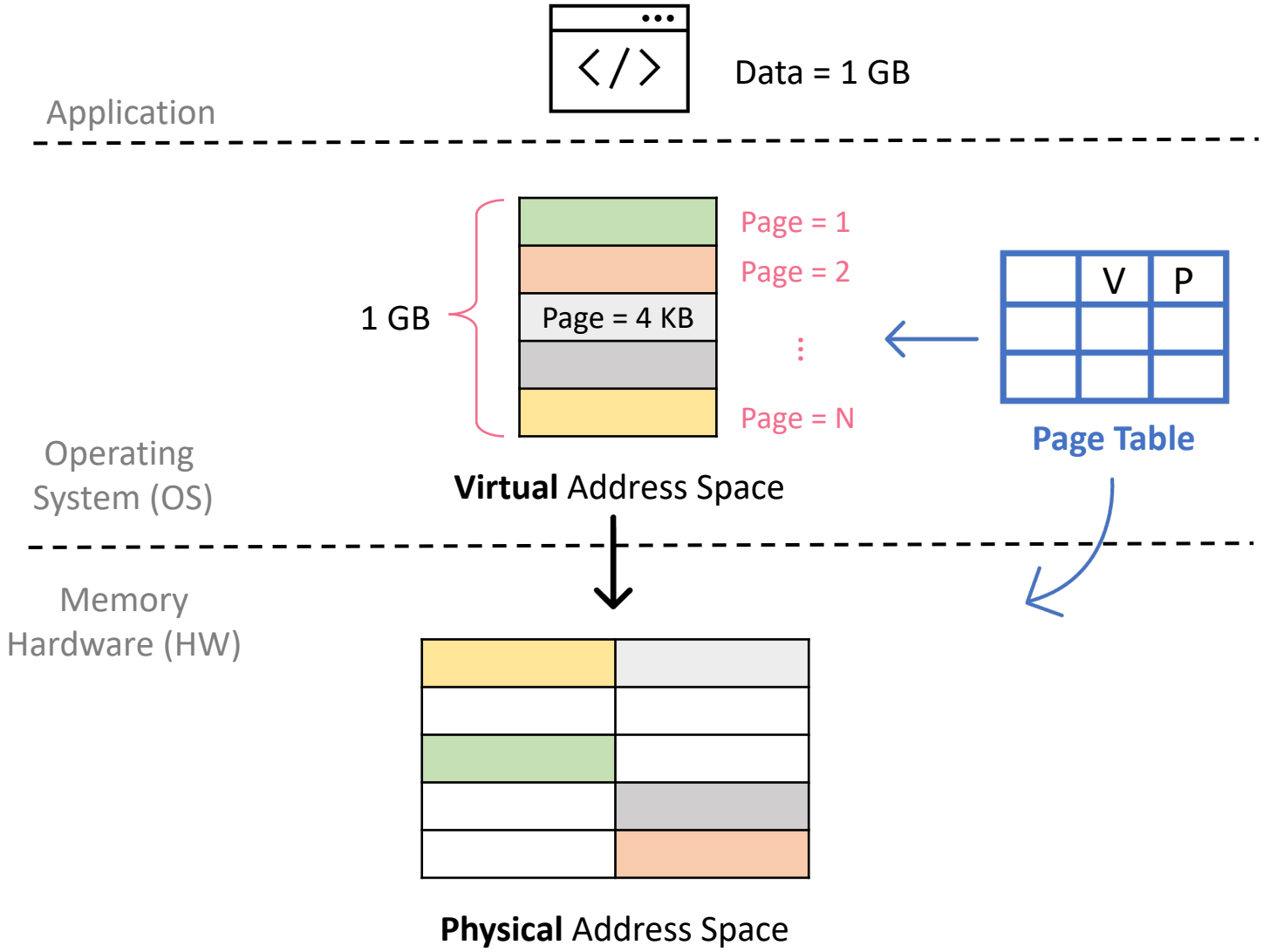
LSTMs

for Hybrid Memory Management (HMem Management)

Lecture Outline:

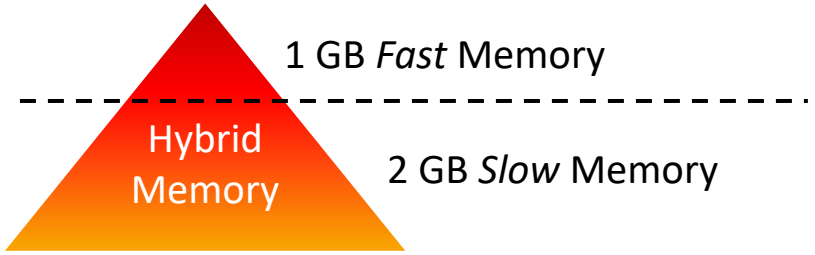
- Hybrid Memories
- **Hybrid Memory Management**
- How to Integrate ML?
- System Design
- Evaluation
- Lessons Learned

OS-level Memory Management

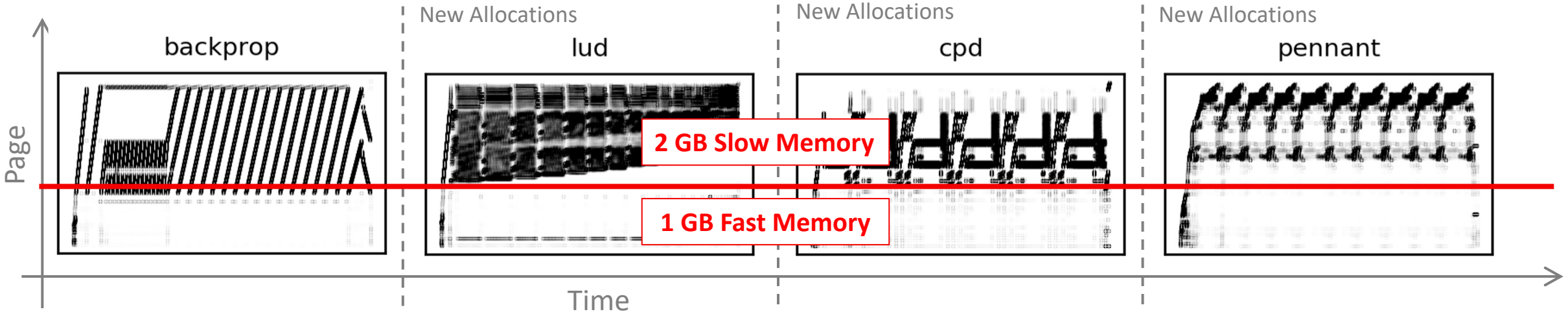


Hybrid Memory Page Allocation

Example: Let's assume each application uses 3 GB. And hybrid memory consists of:



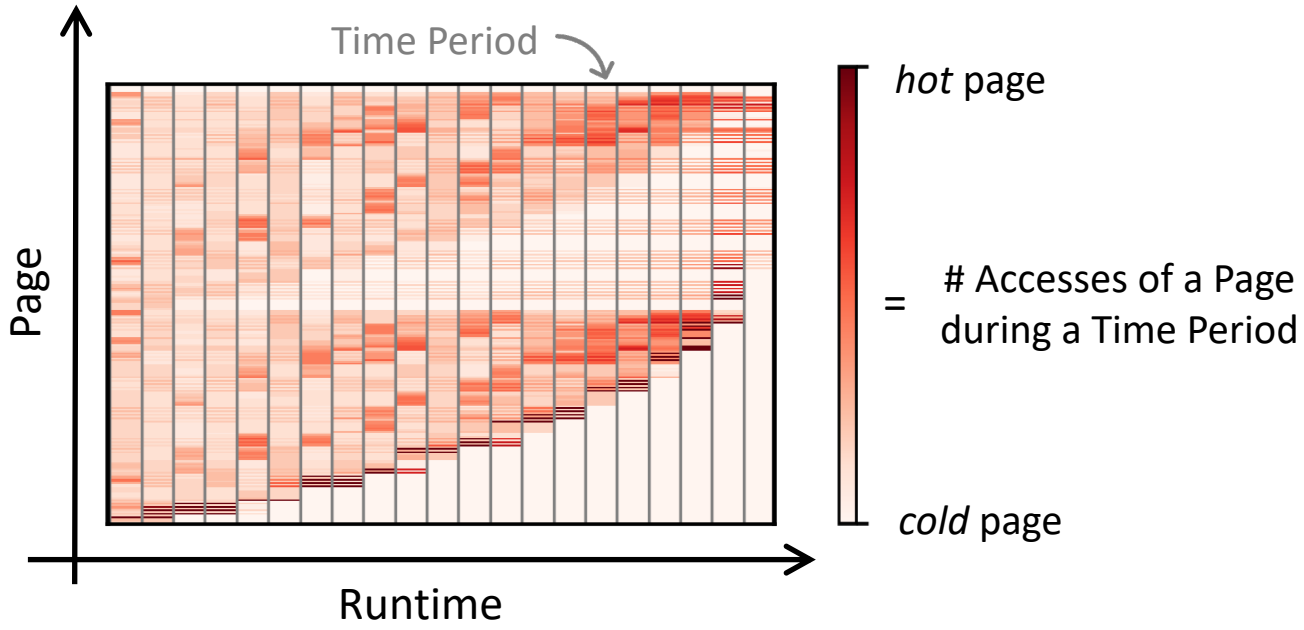
The OS allocates the pages in memory using the **"first touch"** policy.

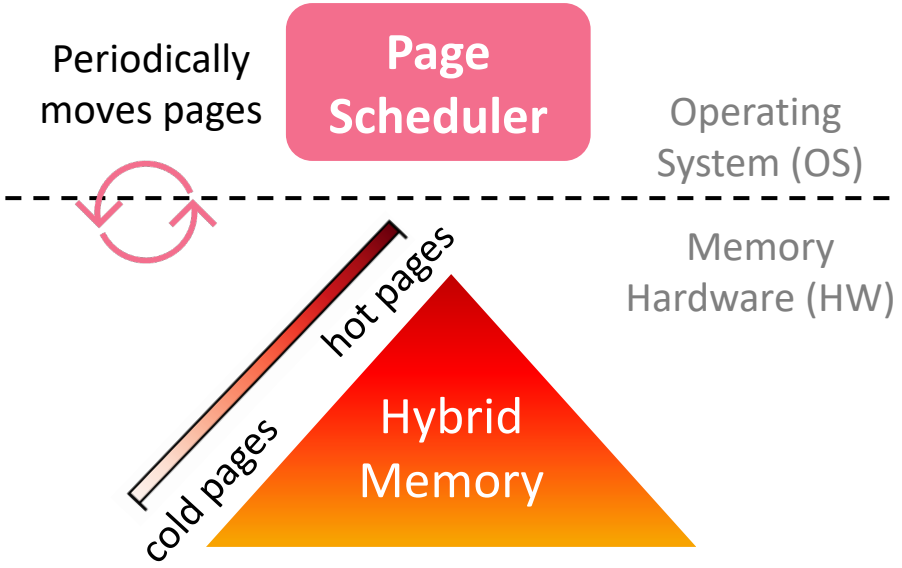


If the memory allocation doesn't change throughout time, then we get no use out of the *fast* memory.

Hybrid Memory Page Scheduling

The OS should move pages dynamically across hybrid memory to maximize the efficiency.



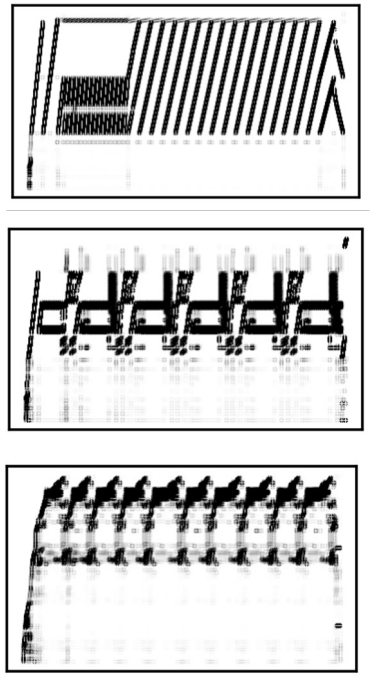


The page *hotness* changes through time.

Keep the hot pages in fast memory through time.

Page Scheduling as a Prediction Problem

Applications



OS-level Page Scheduler

1. Page Access Monitor

Keep track of page hotness.

Past
Page Hotness

In every
time period

2. Page Hotness Predictor

Predict **future** page hotness,
based on past access **history**.

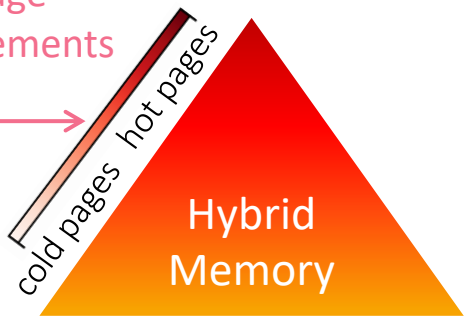
Future
Page Hotness

3. Page Movement Selector

Choose **which** pages to move
across hybrid memory.

Hardware

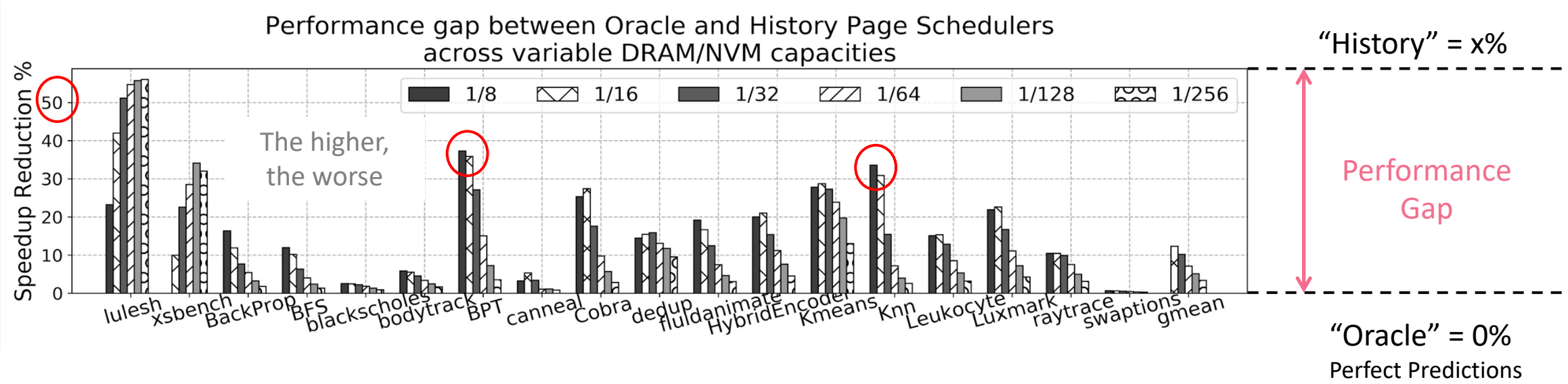
Page
Movements



Existing Predictors

OS uses a very simple **history**-based predictor, to minize operational overheads.

Predicts for all pages that page hotness at period p_N = page hotness at period p_{N+1}



Need something more clever, to close this big gap.. How effective would Machine Learning be?

Outline of Today's Lecture

Systems Software ← ML *for* Systems ← Machine Learning

Today's Paper:

Kleio: A Hybrid Memory Page Scheduler with Machine Intelligence

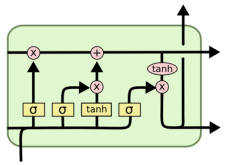
Thaleia Dimitra Doudali
Georgia Institute of Technology
thdoudali@gatech.edu

Sergey Blagodurov
Advanced Micro Devices, Inc.
Sergey.Blagodurov@amd.com

Abhinav Vishnu
Advanced Micro Devices, Inc.
Abhinav.Vishnu@amd.com

Sudhanva Gurumurthi
Advanced Micro Devices, Inc.
Sudhanva.Gurumurthi@amd.com

Ada Gavrilovska
Georgia Institute of Technology
ada@cc.gatech.edu



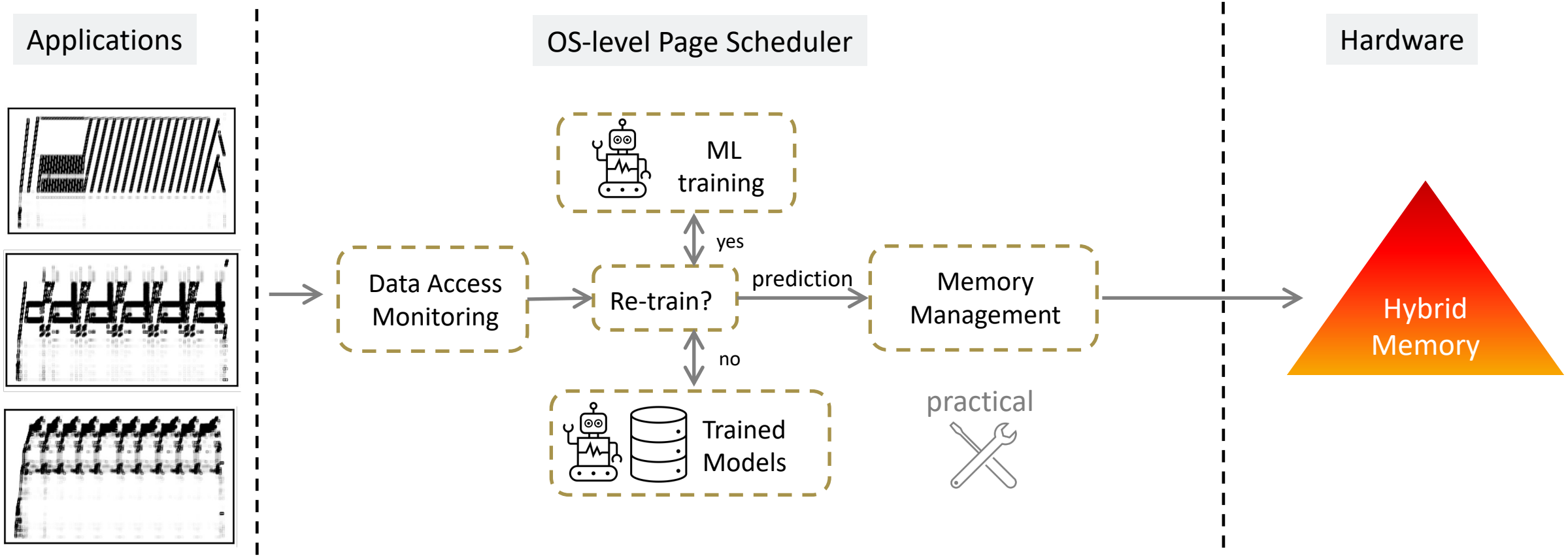
LSTMs

for Hybrid Memory Management (HMem Management)

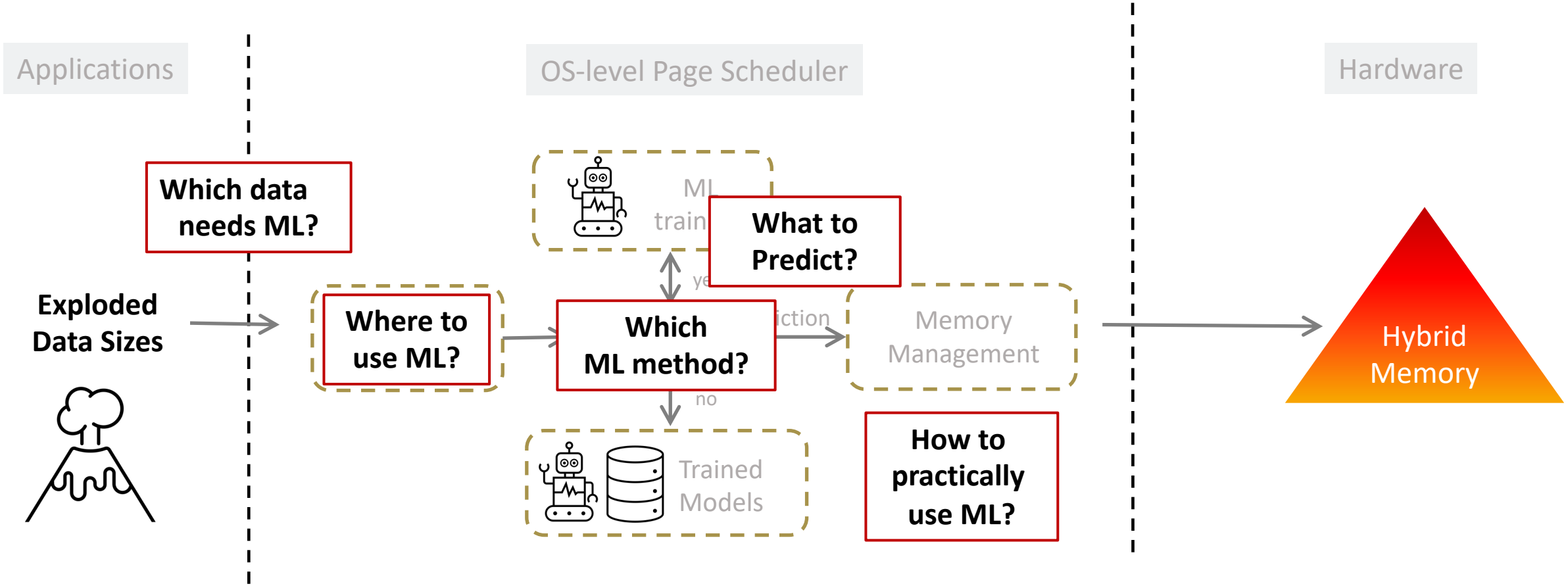
Lecture Outline:

- Hybrid Memories
- Hybrid Memory Management
- How to Integrate ML?
- System Design
- Evaluation
- Lessons Learned

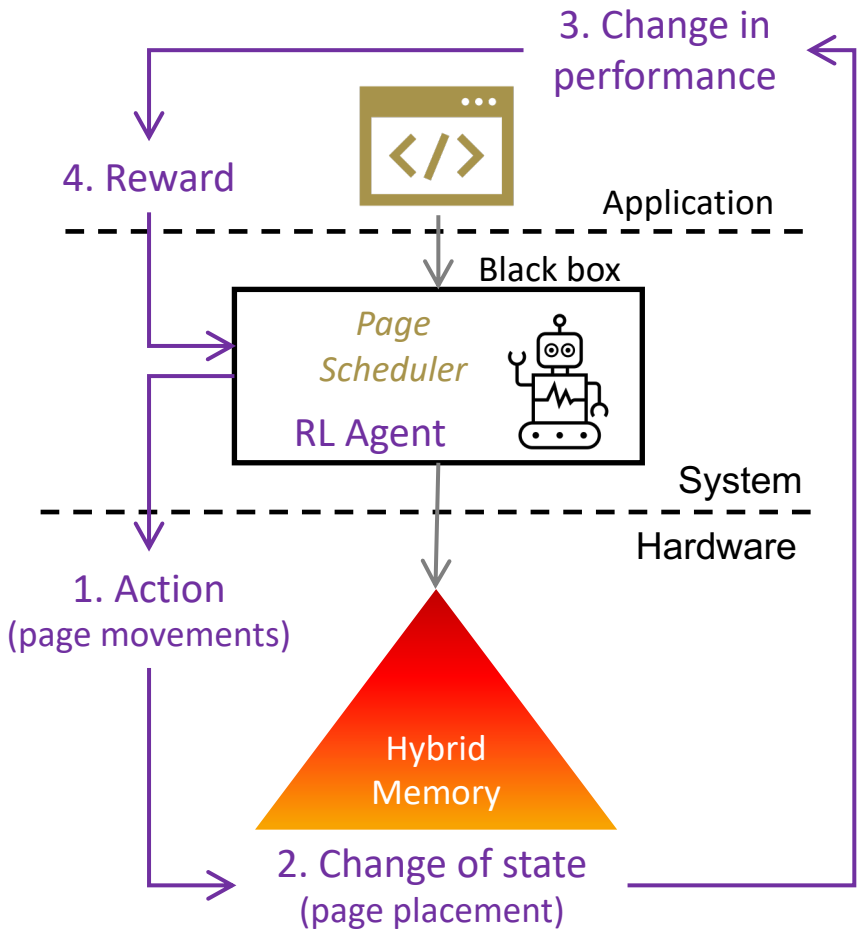
The Vision: Integrate ML in Page Sceduling



Toward Realizing the Vision: Questions to Ask



Where to Use ML? (1)



Replace the Page Scheduler with a Reinforcement Learning (RL) agent.

Learn the Action: Learn from moving pages across hybrid memory.

Learn from mistakes (e.g., cold pages in DRAM).

Why it is not a good fit:

- Exponential Action Space = 2^N, when moving N pages across 2 memories.
- Need to re-train if configuration of hybrid memory changes.
  - Number of memory units.
  - Difference in access speeds / capacities.

Not practical / scalable.

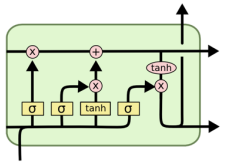
Don't learn the action!

Learning Memory Access Patterns

Don’t learn the action.. Learn the memory access pattern?

Learning Memory Access Patterns

Milad Hashemi [1] Kevin Swersky [1] Jamie A. Smith [1] Grant Ayers [2*] Heiner Litz [3*] Jichuan Chang [1]
Christos Kozyrakis [2] Parthasarathy Ranganathan [1]



LSTMs

for Cache Prefetching

Memory Access Trace

```
0x40001ee0: R 0xbfffe798
0x40001efd: W 0xbfffe7d4
0x40001f09: W 0xbfffe7d8
0x40001f20: W 0xbfffe864
0x40001f20: W 0xbfffe868
```

PC Physical address

input

PC t_N,
Address Delta(t_N, t_{N-1})

→ LSTM-based Deep Neural Network →

output

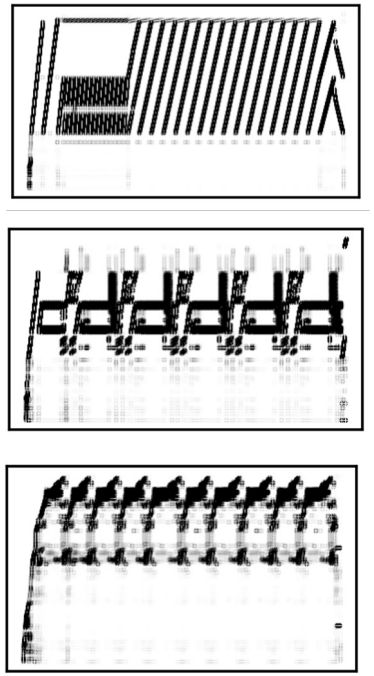
Delta(t_{N+1}, t_N)

LSTMs (Long Short Term Memory) Networks are used in timeseries forecasting.

Can we do something similar?

Where to Use ML? (2)

Applications



OS-level Page Scheduler

1. Page Access Monitor

Keep track of page hotness.

Past Page Hotness

In every time period

2. Page Hotness Predictor

Use **LSTMs** to predict **which** pages will be accessed in the next period. Then, calculate page hotness.

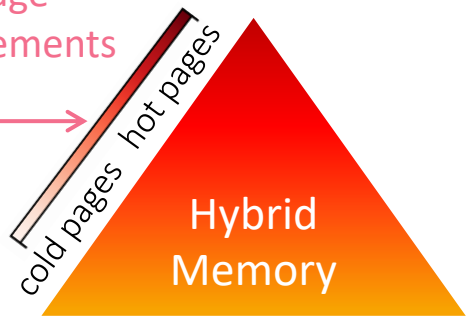
Future Page Hotness

3. Page Movement Selector

Choose **which** pages to move across hybrid memory.

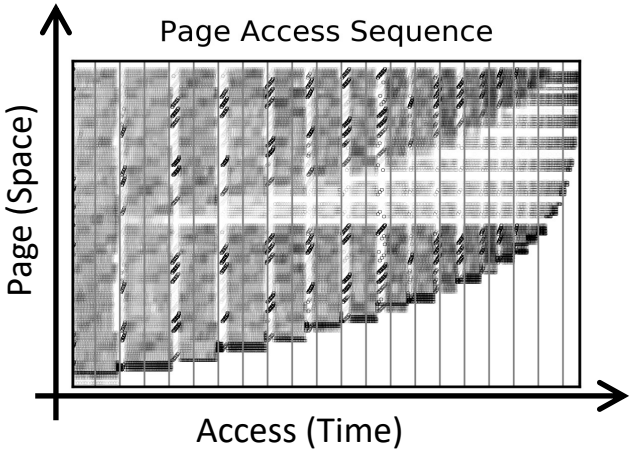
Hardware

Page Movements



Learning the Page Access Sequence

Learn **which** pages will be accessed in the next period of time, given a window of history.



input

Page ID t_{N-1},
Page ID t_{N-2},
...
Page ID t_{N-h}

→ LSTM-based Deep Neural Network →

output

Page ID t_N

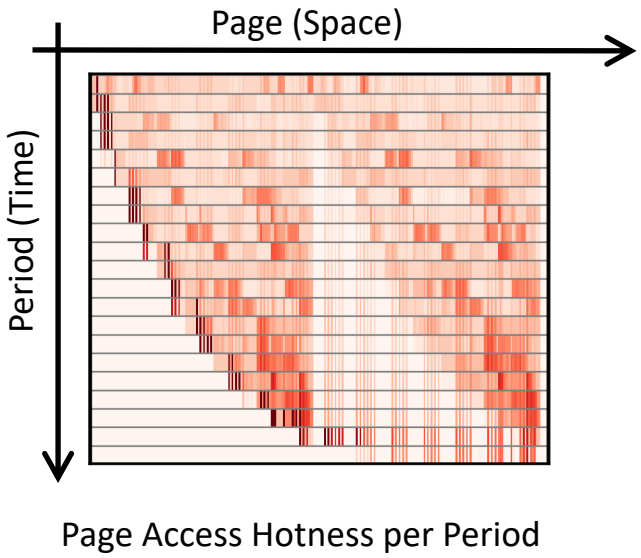
No. models	Overheads	Accuracy
1 per app	Huge Memory Trace Files. Days to train. Months to fine-tune.	Low, because millions of pages. Top-k predictions not useful.

Can we learn something else?

Learning the Page Hotness

Learn **how hot** a page will be in the next period of time, given a window of history.

Flip the way you look at a problem!



Page Access Hotness per Period

input

$$\begin{aligned} &\texttt{Page Hotness } t_{N-1}, \\ &\texttt{Page Hotness } t_{N-2}, \\ &\quad \ldots \\ &\texttt{Page Hotness } t_{N-h} \end{aligned}$$

→ LSTM-based Deep Neural Network →

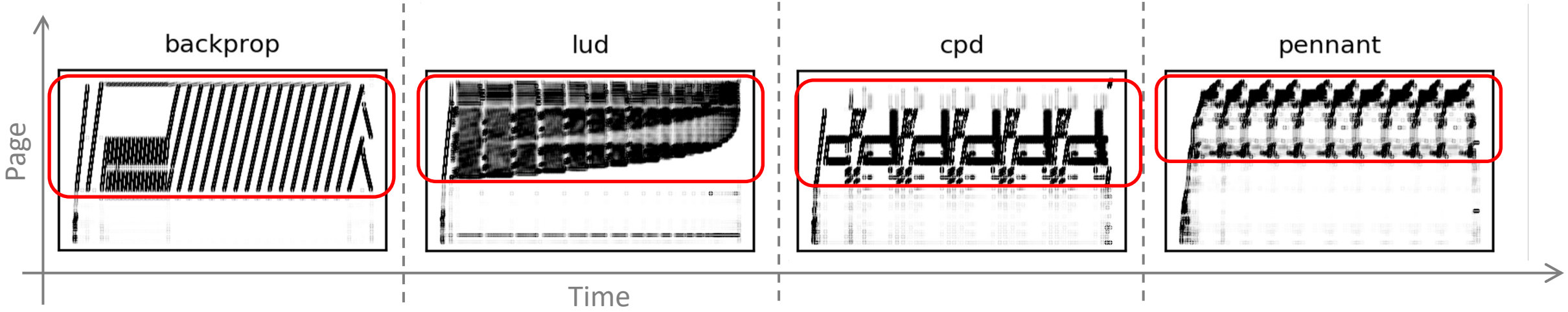
output

$$\texttt{Page Hotness } t_N$$

No. models	Overheads	Accuracy
Many per app. 1 per page.	Smaller models. Parallel Training. Faster to train and fine-tune.	High, because hotness value is smaller, depending on the period length.

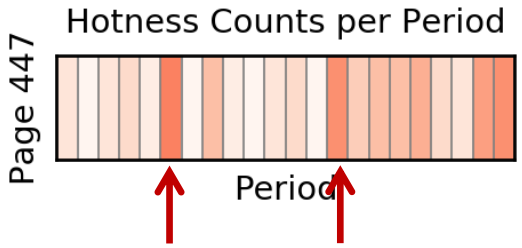
Challenge: 1 model per page, means millions of models..

Do all Pages Need ML?



Probably the pages that are part of a pattern, need ML-based management.
For the rest, the simple *history* page scheduler works well (page hotness at period p_N = page hotness at period p_{N+1}).

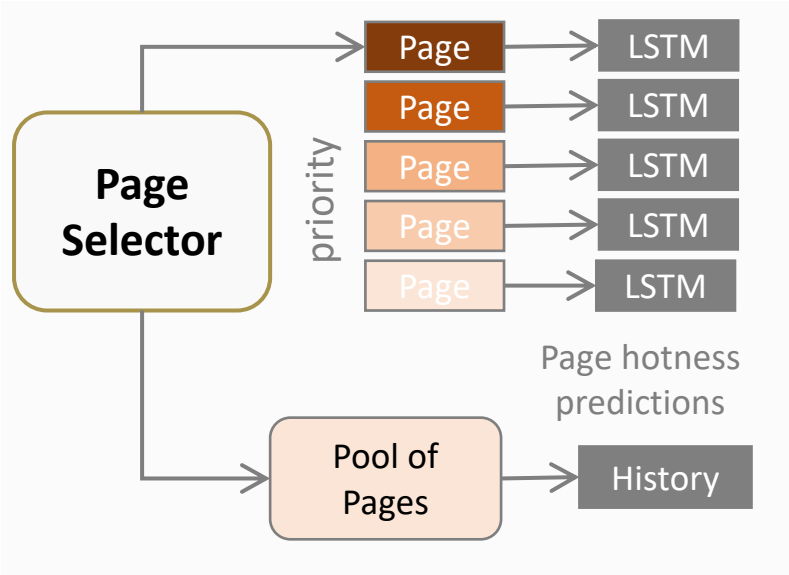
Which pages **really** need ML?



Page suddenly becomes very **hot**.
History scheduler will **misplace** it (not in DRAM when hot).

Use ML for subset of pages, and the existing *history* scheduler for the rest.

Proposed ML Integration



Apply ML on a small page subset.

↳ Foundations for practical use of ML.

Carefully select pages for ML.

↳ Application performance boost.

Outline of Today's Lecture

Systems Software ← ML *for* Systems — Machine Learning

Today's Paper:

Kleio: A Hybrid Memory Page Scheduler with Machine Intelligence

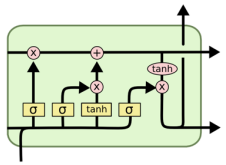
Thaleia Dimitra Doudali
Georgia Institute of Technology
thdoudali@gatech.edu

Sergey Blagodurov
Advanced Micro Devices, Inc.
Sergey.Blagodurov@amd.com

Abhinav Vishnu
Advanced Micro Devices, Inc.
Abhinav.Vishnu@amd.com

Sudhanva Gurumurthi
Advanced Micro Devices, Inc.
Sudhanva.Gurumurthi@amd.com

Ada Gavrilovska
Georgia Institute of Technology
ada@cc.gatech.edu



LSTMs

for Hybrid Memory Management (HMem Management)

Lecture Outline:

- Hybrid Memories
- Hybrid Memory Management
- How to Integrate ML?
- System Design
- Evaluation
- Lessons Learned

System Design of Kleio

Greek Trivia: According to the ancient Greek mythology, Kleio was the muse of history, daughter of Mnemosyne, goddess of memory.

Applications

OS-level Page Scheduler

Hardware

Page Selector

small subset

Pages for ML

bigger subset

Pages for History

1. Run Page Selector.
2. Train LSTMs.

Offline Phase

1. Page Access Monitor

Keep track of page hotness.

Past Page Hotness

2. Page Hotness Predictor

LSTM inference

History predictions

Future Page Hotness

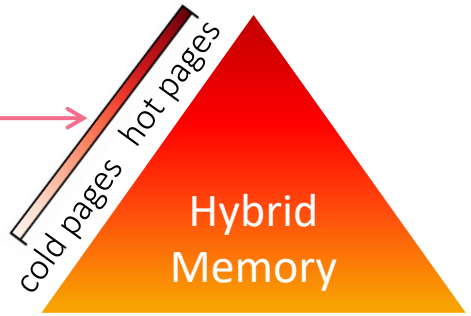
In every time period

3. Page Movement Selector

Choose **which** pages to move across hybrid memory.

Online Phase

Page Movements

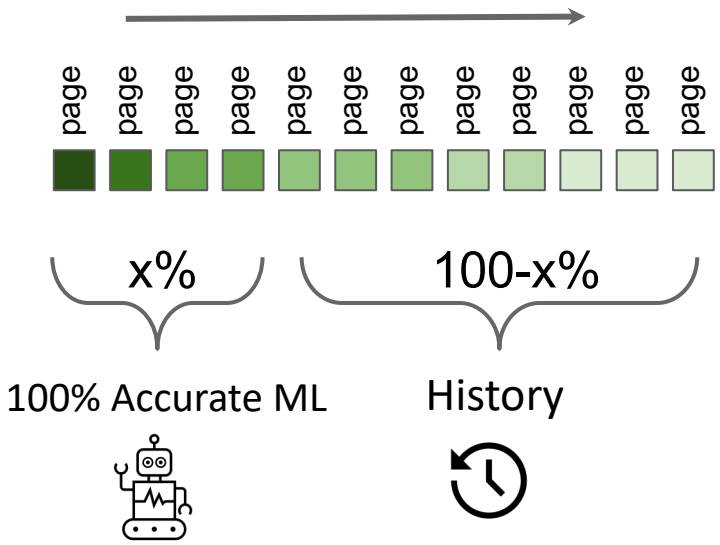


Which Pages to Select?

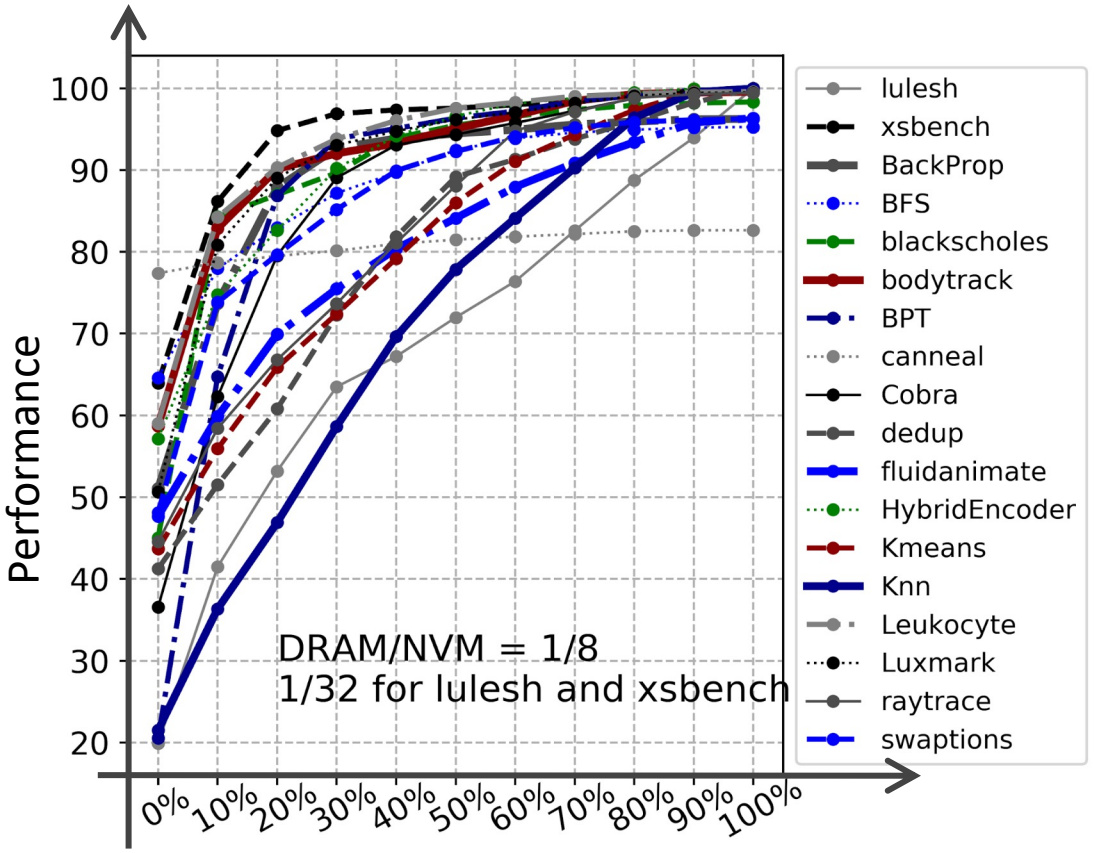
Need to create a metric to select pages.

Per Page Benefit Factor

benefit = # accesses x # misplacements



Page Hotness Prediction

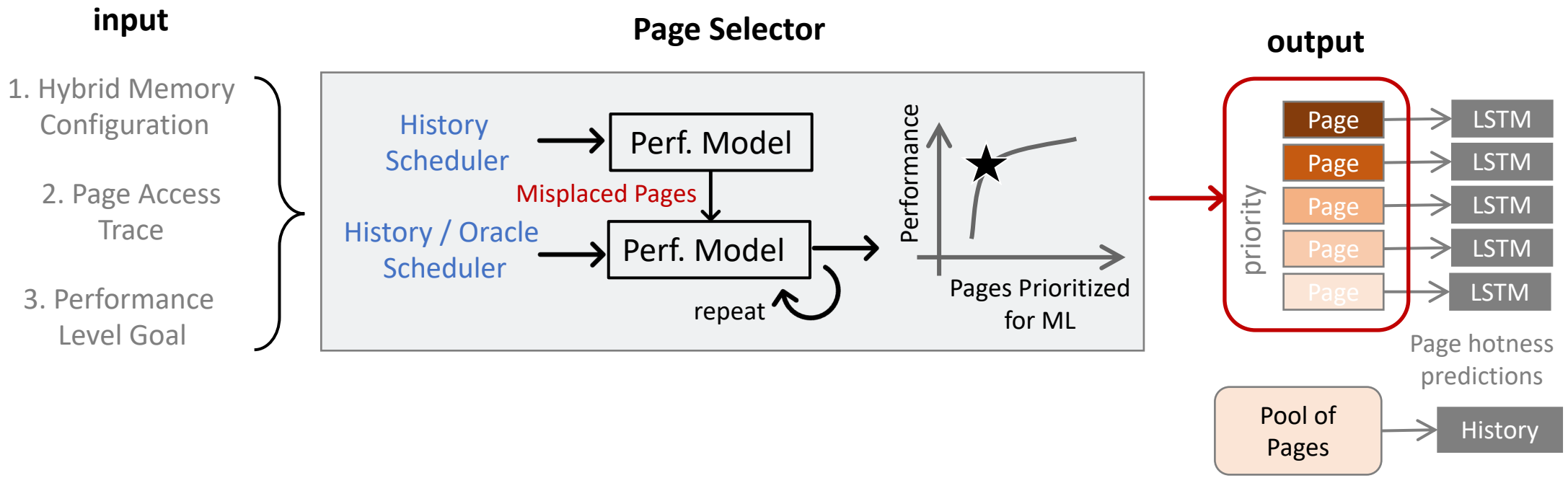


Number of pages managed by 100% Accurate ML

Select a **small subset** of pages in the **order** that brings the desired **performance** level.

Page Selector Design

Page Selector calculates internally the perfomance curve, using a performance estimate analytical model.



It is not a lightweight process, but necessary to deliver the desired application performance levels.

Outline of Today's Lecture

Systems Software ← ML *for* Systems — Machine Learning

Today's Paper:

Kleio: A Hybrid Memory Page Scheduler with Machine Intelligence

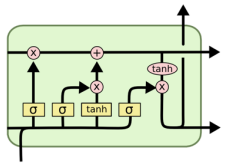
Thaleia Dimitra Doudali
Georgia Institute of Technology
thdoudali@gatech.edu

Sergey Blagodurov
Advanced Micro Devices, Inc.
Sergey.Blagodurov@amd.com

Abhinav Vishnu
Advanced Micro Devices, Inc.
Abhinav.Vishnu@amd.com

Sudhanva Gurumurthi
Advanced Micro Devices, Inc.
Sudhanva.Gurumurthi@amd.com

Ada Gavrilovska
Georgia Institute of Technology
ada@cc.gatech.edu



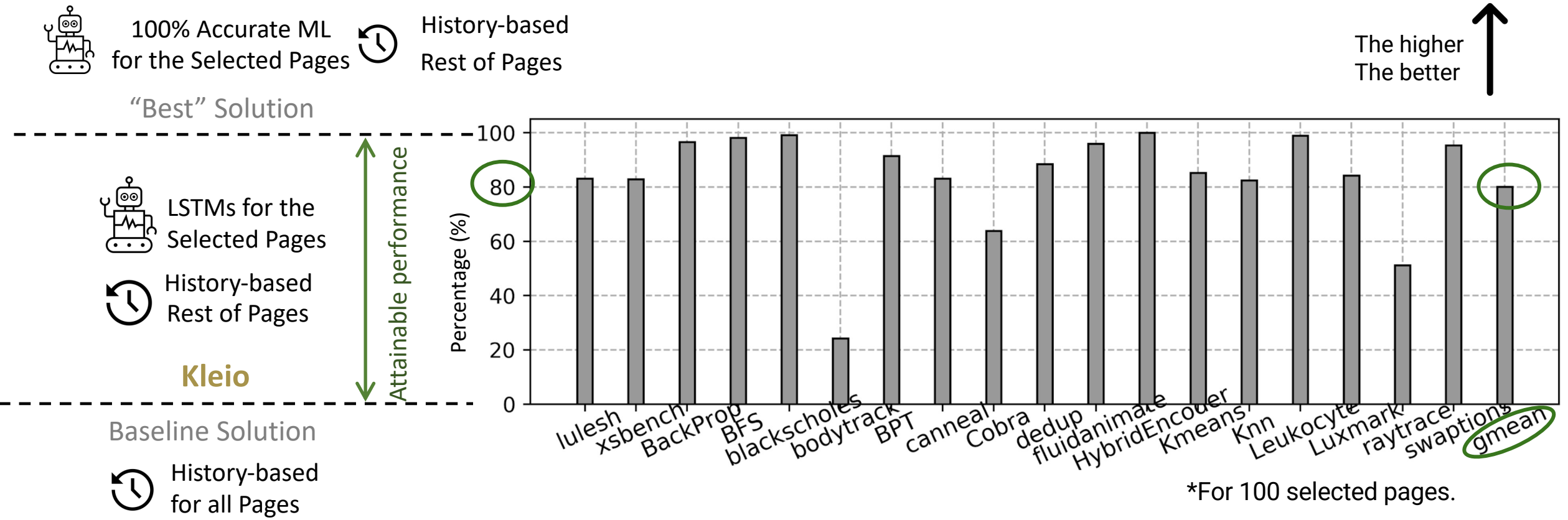
LSTMs

for Hybrid Memory Management (HMem Management)

Lecture Outline:

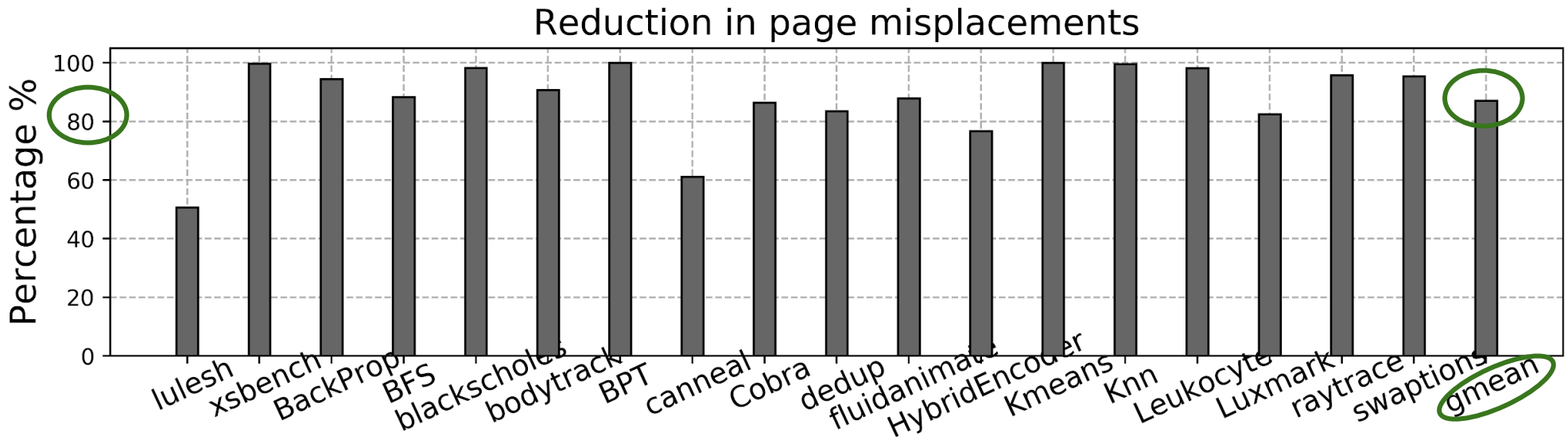
- Hybrid Memories
- Hybrid Memory Management
- How to Integrate ML?
- System Design
- Evaluation
- Lessons Learned

Effect on Application Performance



For half of the applications, Kleio reaches **95%** the possible performance levels!

Effect on Quality of Page Placement

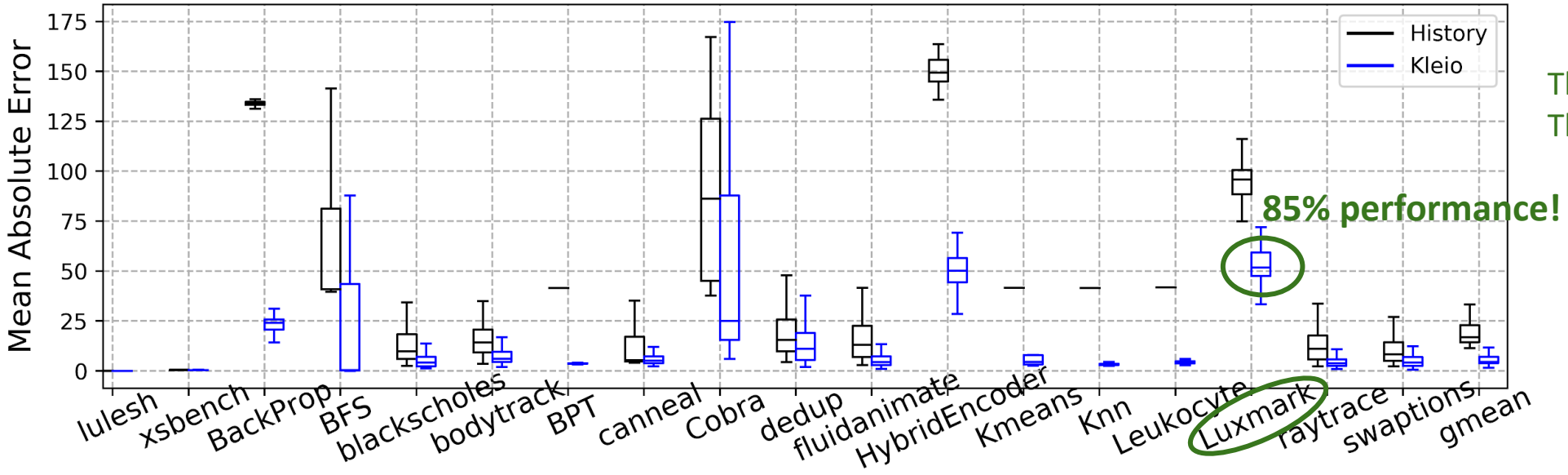


The higher
The better

For the pages managed with ML.

Kleio reduces more than 80% of the page misplacements, due to the improved **page movement decisions**, via the more **accurate** page hotness **predictions**.

LSTM Prediction Accuracy



e.g. MAE = 50 means that the RNN predicted on average 50 more accesses per scheduling epoch per page.

High prediction error **does not impact** application performance, when it does not affect the quality of page movement decision.

Outline of Today's Lecture

Systems Software ← ML *for* Systems — Machine Learning

Today's Paper:

Kleio: A Hybrid Memory Page Scheduler with Machine Intelligence

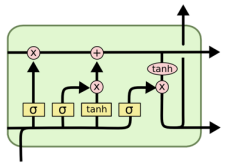
Thaleia Dimitra Doudali
Georgia Institute of Technology
thdoudali@gatech.edu

Sergey Blagodurov
Advanced Micro Devices, Inc.
Sergey.Blagodurov@amd.com

Abhinav Vishnu
Advanced Micro Devices, Inc.
Abhinav.Vishnu@amd.com

Sudhanva Gurumurthi
Advanced Micro Devices, Inc.
Sudhanva.Gurumurthi@amd.com

Ada Gavrilovska
Georgia Institute of Technology
ada@cc.gatech.edu



LSTMs

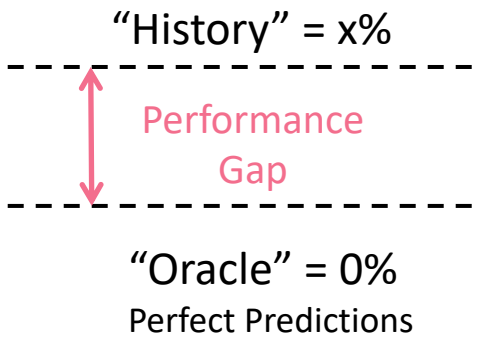
for Hybrid Memory Management (HMem Management)

Lecture Outline:

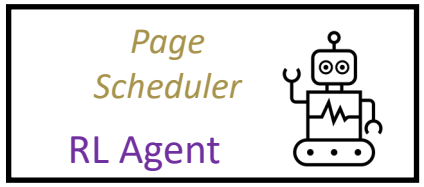
- Hybrid Memories
- Hybrid Memory Management
- How to Integrate ML?
- System Design
- Evaluation
- Lessons Learned

Lessons Learned

1. Understand what would be the benefit from using ML.



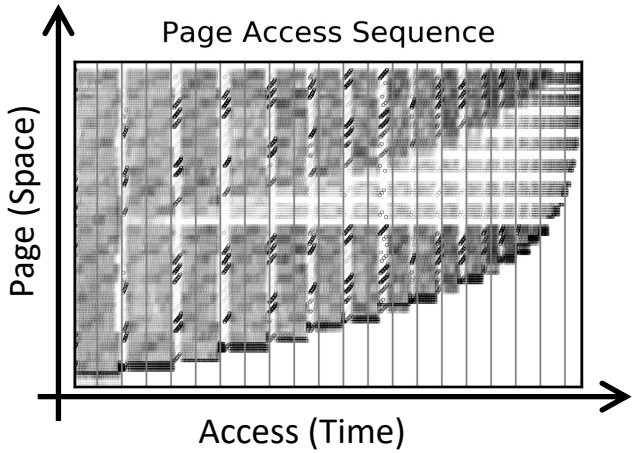
2. Learn the Behavior, not the Action, for the use case of hybrid memory management.

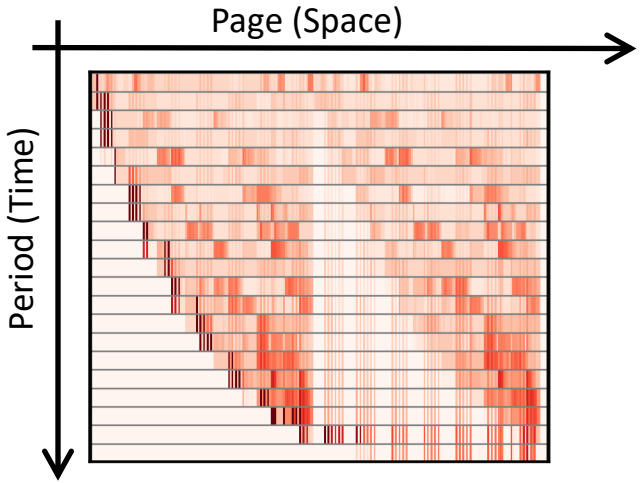


Replacing the Page Scheduler with an RL agent would not be practical, nor scalable.

Lessons Learned

3. Look at the same problem from a different angle.





Page Access Hotness per Period

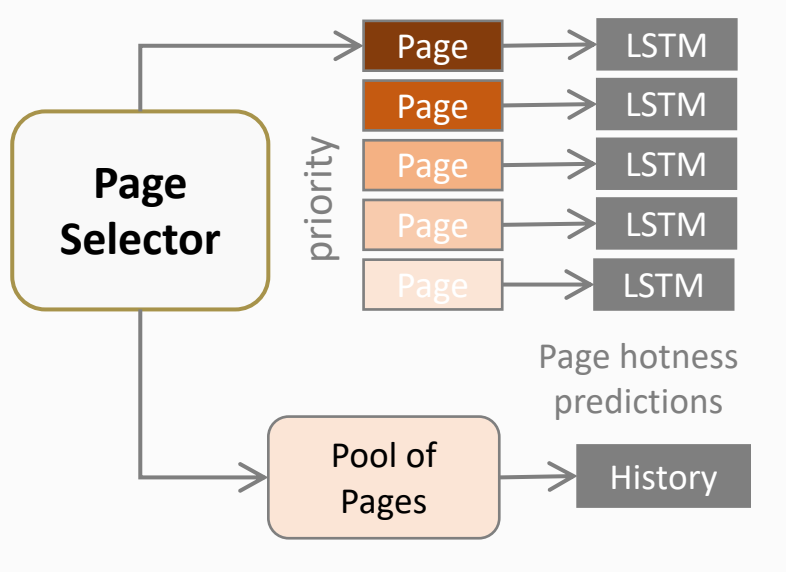
Learn **which** pages will be accessed next.

Learn **how hot** a page will be next.

4. High prediction error of page hotness **does not impact** application performance, when it does not affect the quality of page movement decision.

Lessons Learned

5. Use existing solutions to the best of their ability, and deploy ML where necessary.



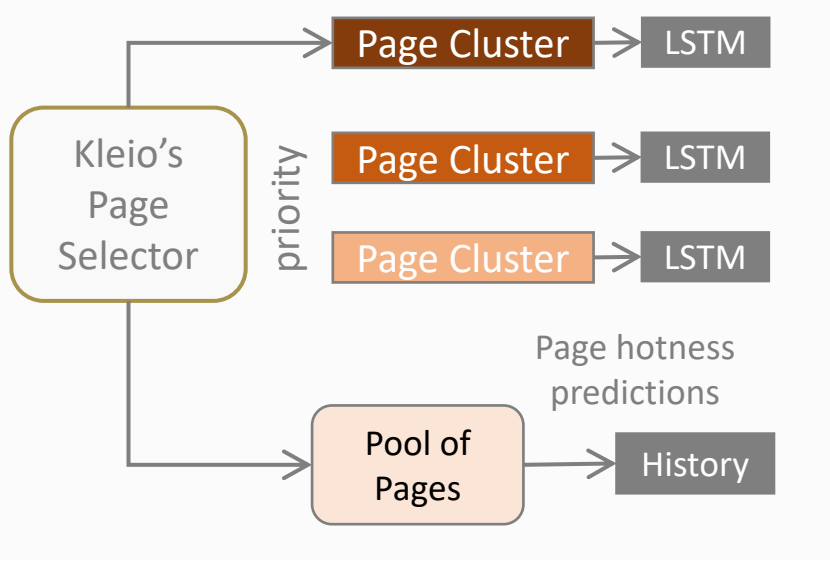
Apply ML on a small page subset.

↳ Foundations for practical use of ML.

Carefully select pages for ML.

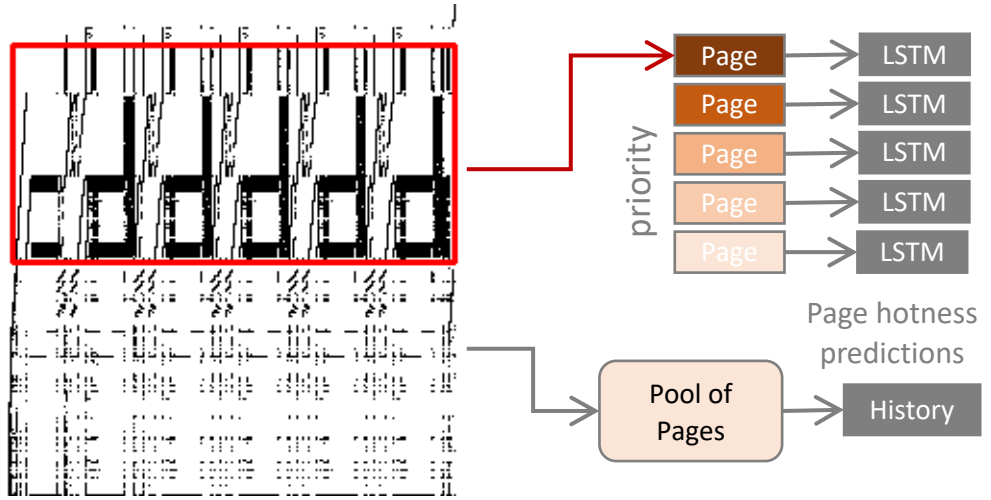
↳ Application performance boost.

Improvements on Kleio



Can we reduce the number of LSTMs via clustering?

Coeus: Clustering (A)like Patterns for Practical Machine Intelligent Hybrid Memory Management. [CCGrid 2022]



Can we accelerate the page selection process?

Cronus: Computer Vision-based Machine Intelligent Hybrid Memory Management. [MEMSYS 2022]

Report Due April 4 at 18.00

Answer / expand upon these 4 questions:

1. What problem is the paper addressing and why is it important?
2. How do they approach to solve the problem?
3. What are the main evaluation results?
4. What are 2 things you will remember from this paper?

Contact

- Via email: `thaleia.doudali@imdea.org`

https://thaleia-dimitradoudali.github.io/

Website