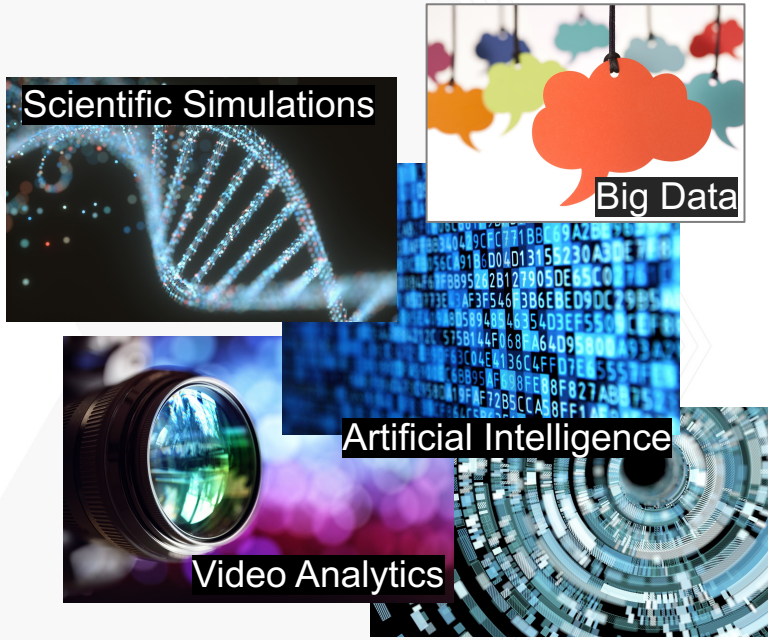
Cori: Dancing to the Right Beat of Periodic Data Movements over Hybrid Memory Systems

Thaleia Dimitra Doudali, Daniel Zahka, Ada Gavrilovska

@ IPDPS '21

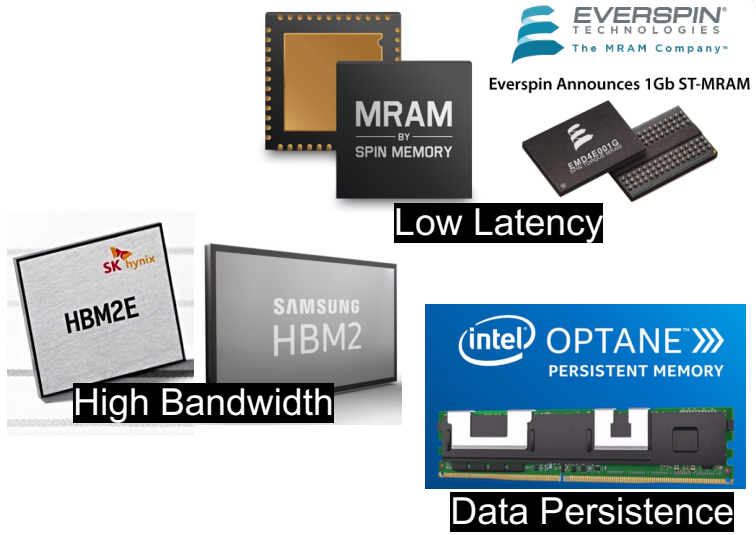
Heterogeneous (Hybrid) Memory Hardware



Application Classes

Exploded Data Sizes

Need for more and faster memory.

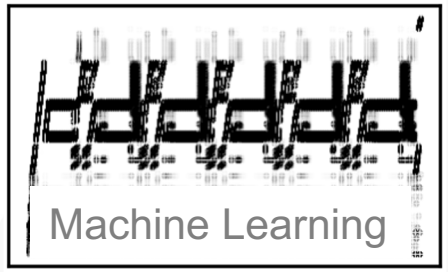


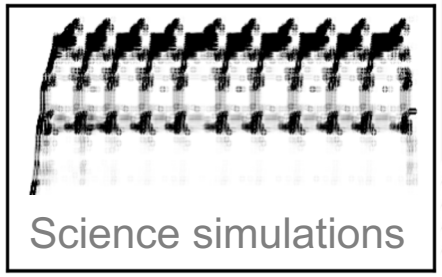
Emerging Memory Hardware

Georgia Tech
CREATING THE NEXT

Hybrid Memory Management Systems

Applications





Data access patterns

*System-level Memory Manager (**Page Scheduler**)*

Data Access Monitoring

Page access counts

Data Access Characterization

Repeat periodically

Data tiering decision

Data Management

Move pages across memories

Policy

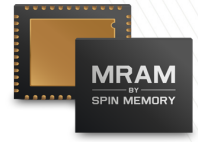
Reactive

Based only on past access history.
"Reacting" to behaviors.

Predictive

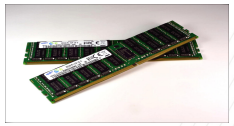
Using robust models (e.g., Machine Learning).
"Predicting" behaviors.

Hybrid Memory

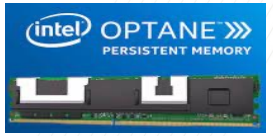


MRAM

HBM



DRAM



PMEM

Georgia Tech
CREATING THE NEXT

Plethora of Existing Solutions

Heterogeneous Memory Architectures: A HW/SW Approach for Mixing Die-stacked and Off-package Memories

Mitesh R. Meswani Sergey Blagodurov David Roberts
John Slice Mike Ignatowski Gabriel H. Loh
AMD Research
Advanced Micro Devices, Inc.
{mitesh.meswani, sergey.blagodurov, david.roberts, john.slice, mike.ignatowski, gabriel.loh}@amd.com

HPCA'15

Thermostat: Application-transparent Page Management for Two-tiered Main Memory

Neha Agarwal Thomas F. Wenisch
University of Michigan
nehaag@umich.edu, twenisch@umich.edu

ASPLOS '17

Nimble Page Management for Tiered Memory Systems

Zi Yan
Rutgers University & NVIDIA
ziy@nvidia.com

Daniel Lustig
NVIDIA
dlustig@nvidia.com

David Nellans
NVIDIA
dnellans@nvidia.com

Abhishek Bhattacharjee
Yale University
abhishek@cs.yale.edu

ASPLOS '19

Coordinated and Efficient Huge Page Management with Ingens

Youngjin Kwon, Hangchen Yu, Simon Peter, Christopher J. Rossbach[1], Emmett Witchel
The University of Texas at Austin
[1]The University of Texas at Austin and VMware Research Group

OSDI '16

HeteroOS - OS Design for Heterogeneous Memory Management in Datacenter

Sudarsun Kannan[1] Ada Gavrilovska[2] Vishal Gupta[3] Karsten Schwan[2]
[1]Department of Computer Sciences, University of Wisconsin-Madison
[2]School of Computer Science, Georgia Tech,
[3]VMWare
{sudarsun@cs.wisc.edu},{ada@cc.gatech.edu},{vishalg@vmware.com}

ISCA '17

High Performance Distributed Systems (Best Paper Nominees) HPDC '19, June 22–29, 2019, Phoenix, AZ, USA

HPDC '19

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

All these systems make ***periodic*** memory management decisions, based on reactive or predictive policies.

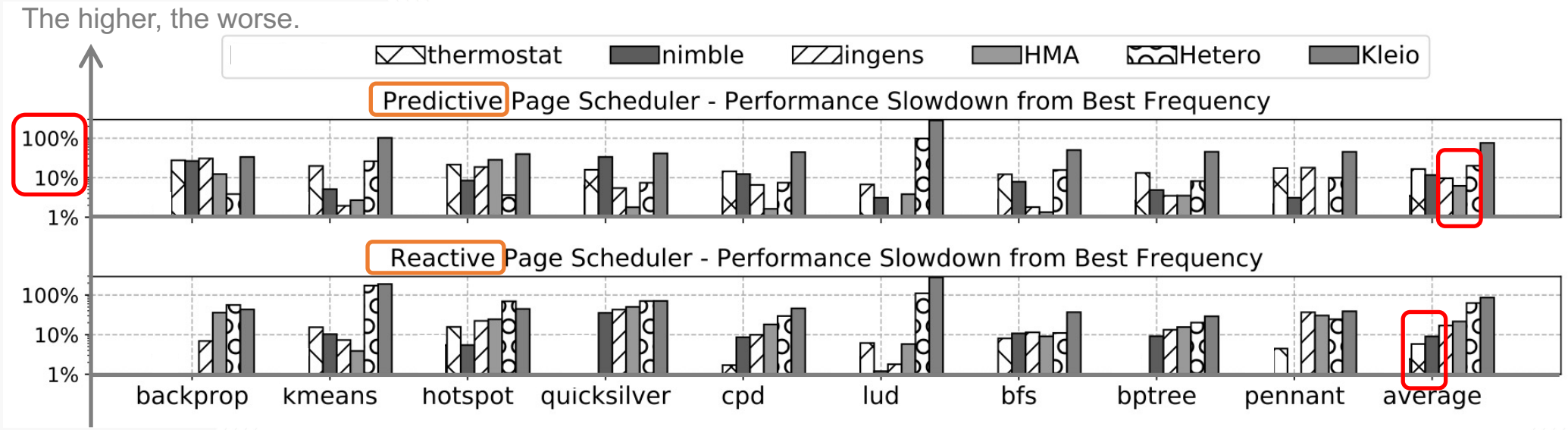
Lost Opportunity for Performance
Due to empirical configuration.

Systems are empirically tuned.

Periodicity differs by orders of magnitude!

Which period duration to use? Which one maximizes performance?

System	Periodicity
Thermostat	10 sec
Nimble	5 sec
Ingens	2 sec
HMA	1 sec
Hetero-OS	0.1 sec
Kleio	0.01 sec

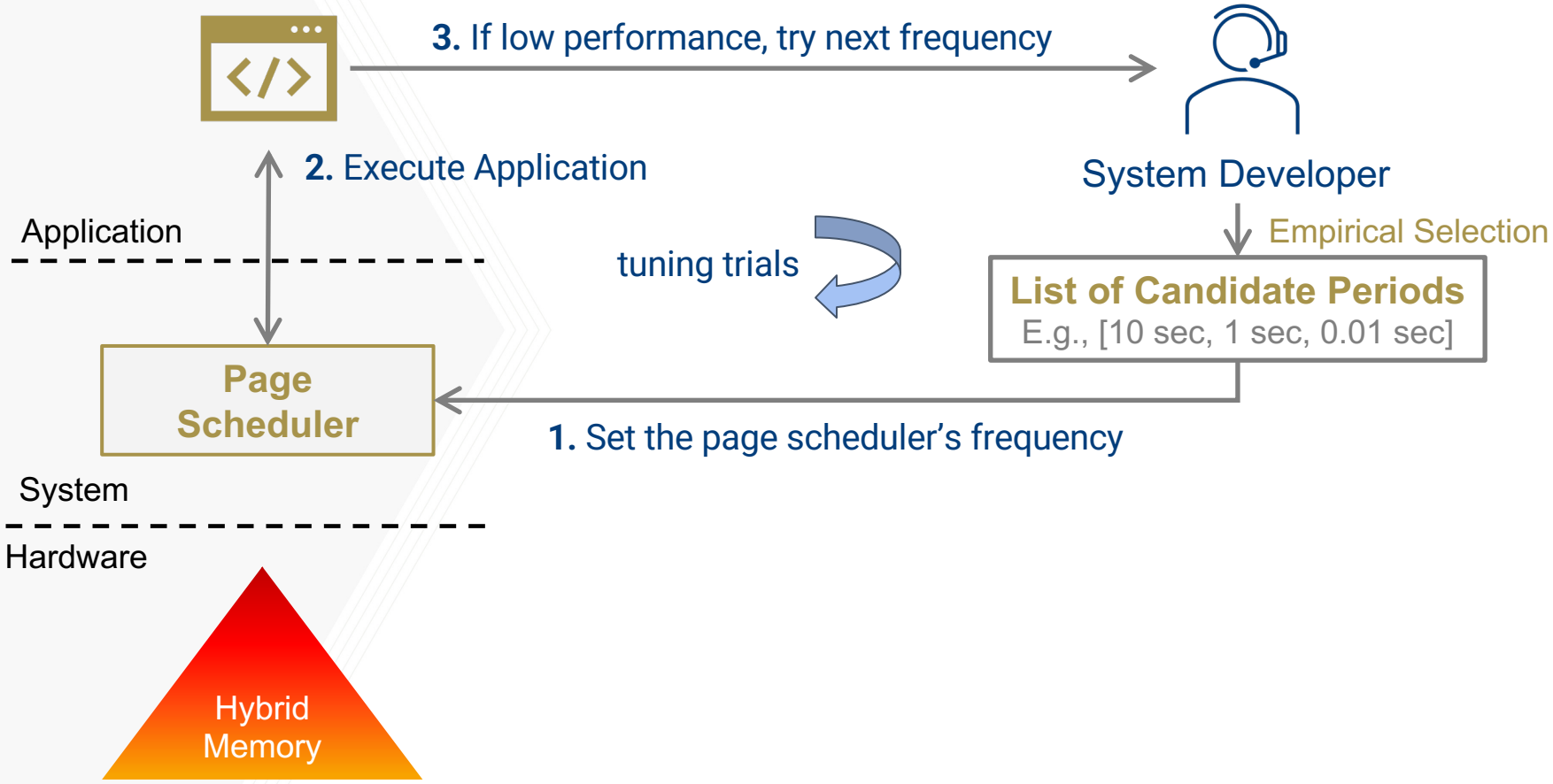


No single proposed period value maximizes performance **across applications and schedulers**.
10% - 100% performance slowdown.

Georgia Tech
CREATING THE NEXT

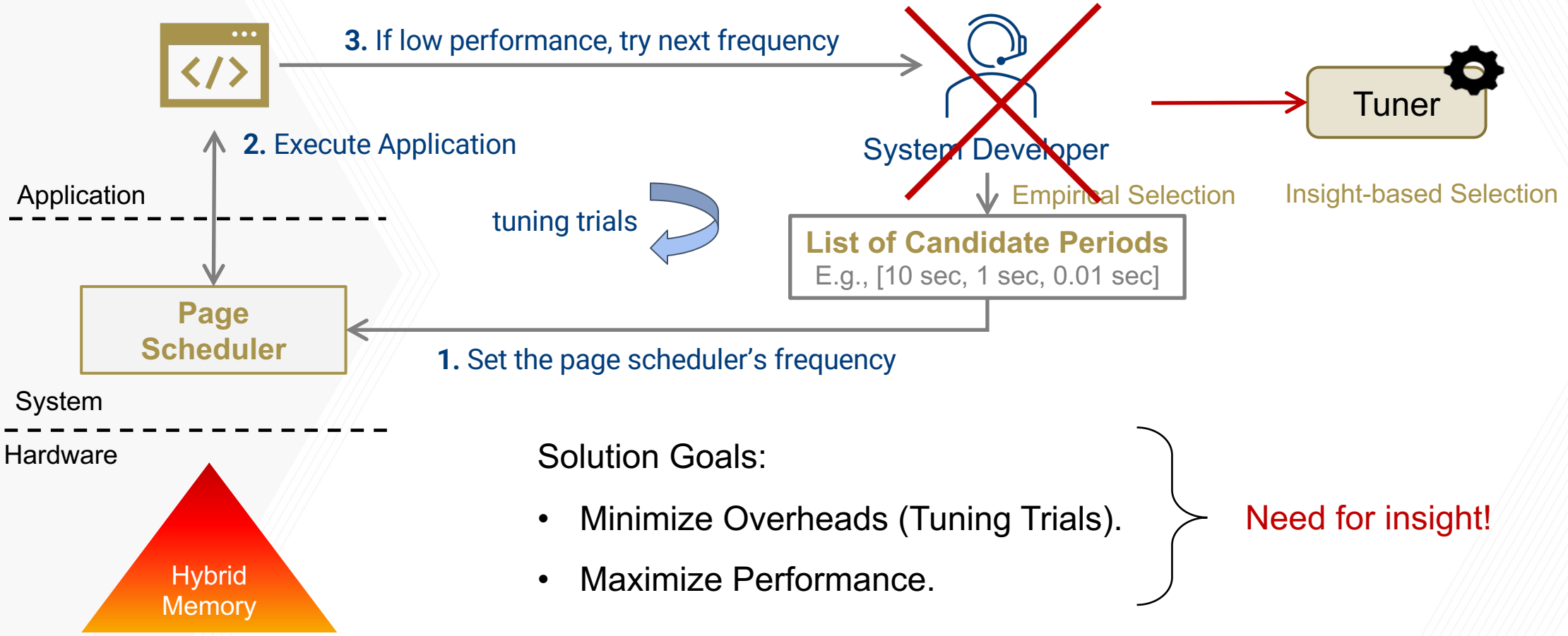
Empirical Configuration

Execution-based tuning of the periodicity.



Replacing Empirical with Insight-based Configuration

Execution-based tuning of the periodicity.

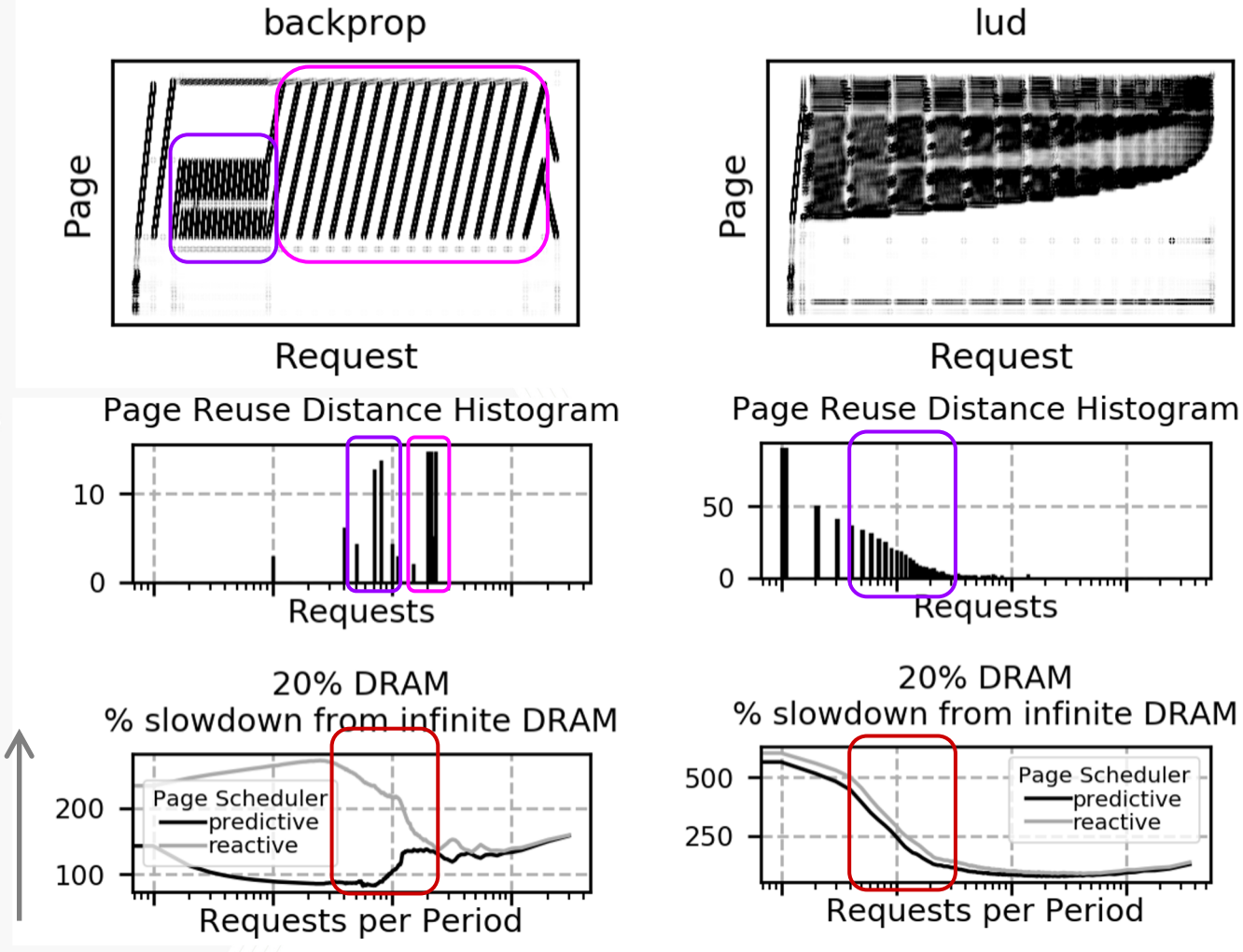


Solution Goals:

- Minimize Overheads (Tuning Trials).
- Maximize Performance.

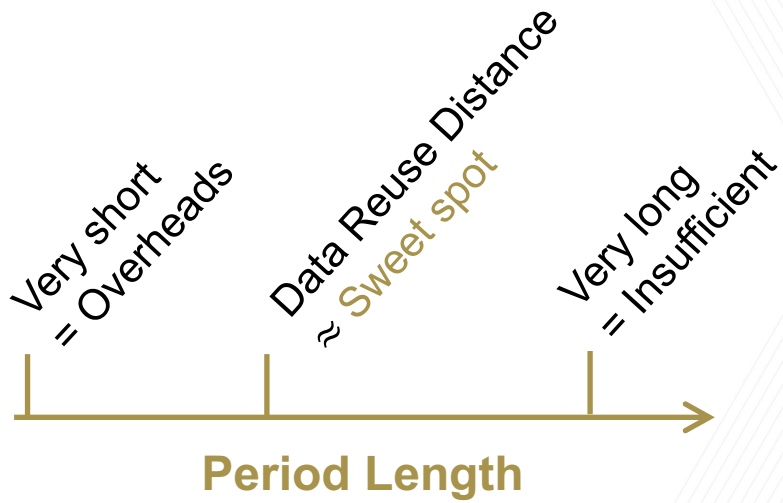
Need for insight!

“Don’t Break the Data Reuse” Insight



Results from exhaustive performance experimentation.

Page Reuse Distance = The time gap between two accesses to the same page.

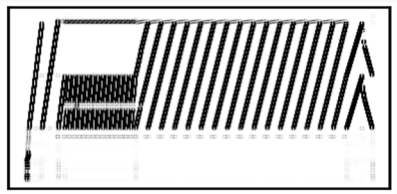


Insight: Periods that align with the data reuse distance, maximize performance.

Georgia Tech
CREATING THE NEXT

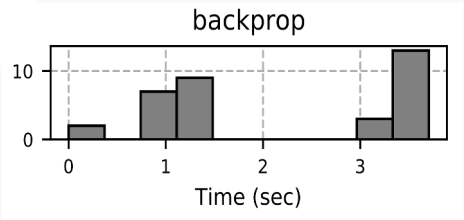
System Design of “Cori”

Cori is an insight-based system-level solution for tuning the frequency of periodic page schedulers.



Memory Access Trace

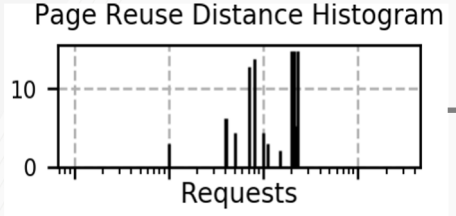
or



Loop time durations

Input

1. Reuse Collector



2. Frequency Generator

Calculate “Dominant Reuse (DR) Time”

$$DR = \frac{\sum_{i=1}^{N}(N-i) \times repeat_i \times reuse_i}{\sum_{i=1}^{N}(N-i) \times repeat_i}$$

$$CandidatePeriods = [DR,\ 2 \times DR,\ ...,\ \frac{Runtime}{2}]$$

Order of tuning trials

3. Tuner

Cori

tuning trials

Page Scheduler

Application Execution

Georgia Tech
CREATING THE NEXT

Evaluation Methodology

Metrics

- Application performance.
  Slowdown from optimally selected frequency (identified via extensive experimentation).
- Tuning Overheads.
  Number of trials to find the frequency that delivers best performance.

Comparison

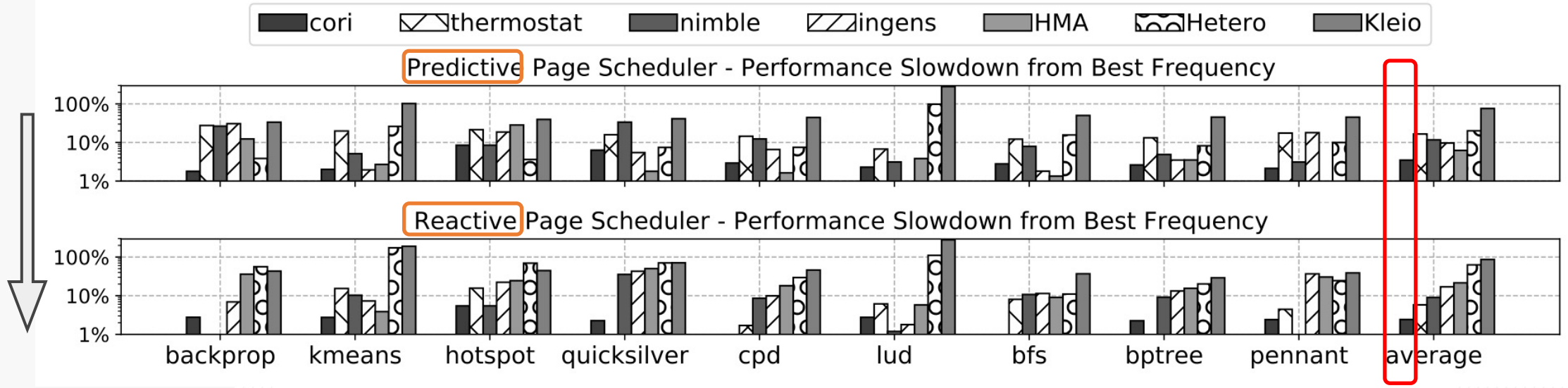
- Proposed values from existing solutions.
  HMA [HPCA '15], **Ingens** [OSDI '16], **Hetero-OS** [ISCA '17], **Thermostat** [ASPLOS '17], **Nimble** [ASPLOS '19], **Kleio** [HPDC '19].
- Cori's selection of period values that differ by the dominant reuse time.
  Tuning trials in increasing order of values.
- "Baseline" selection of period values that differ by a constant time step.
  Tuning trials in increasing, decreasing and random order of values.

Methodology

- Python-based simulation of hybrid memory system and page scheduler.
  https://github.com/GTkernel/cori-sim
- Validation using a hardware testbed with DRAM and Intel's Optane persistent memory.
  https://github.com/GTkernel/x86-Linux-Page-Scheduler

Evaluation (1)

Application performance.

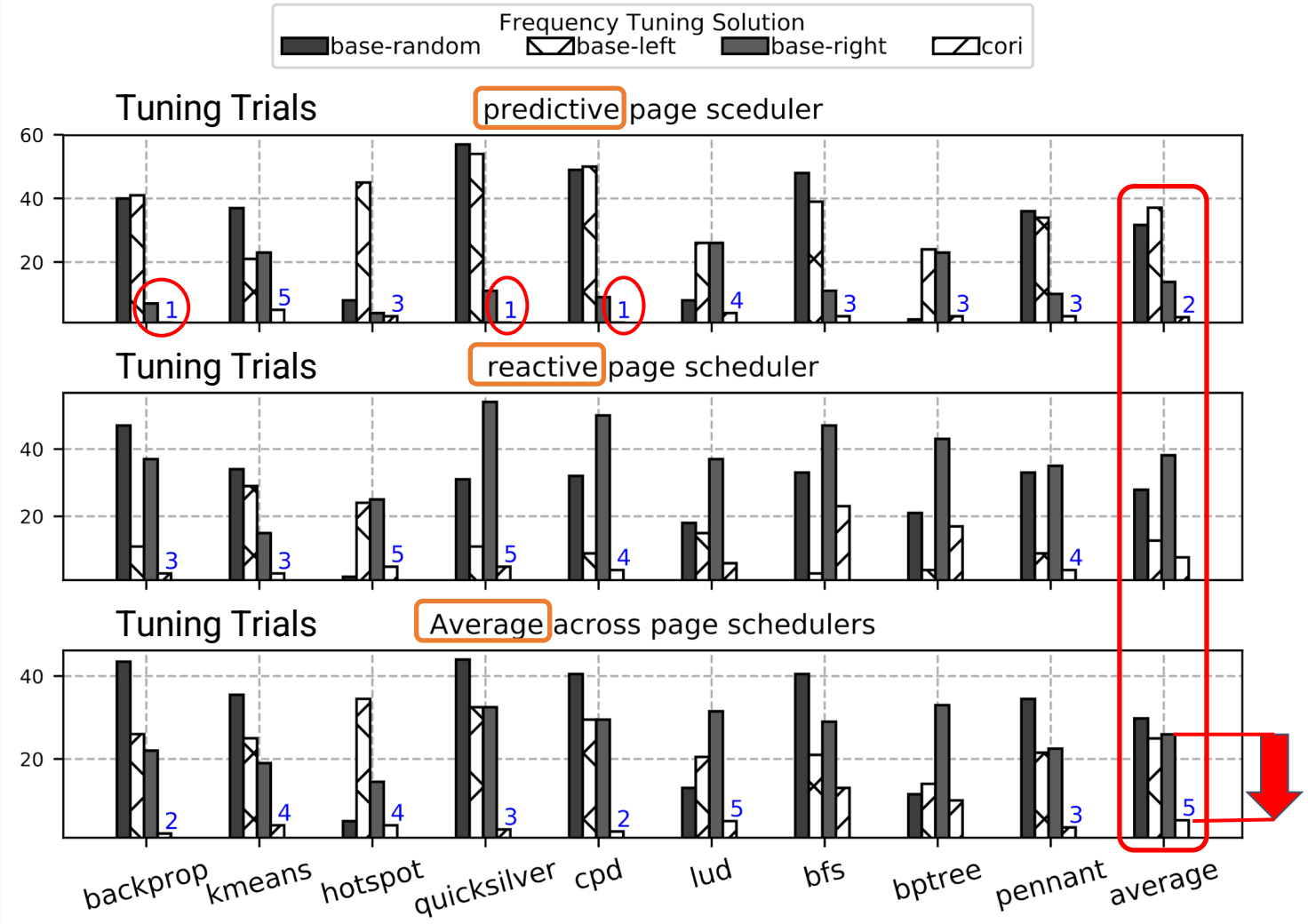


The lower the better

Cori reduces the performance slowdown down to only **3%** across applications and page schedulers, closing the 10% -100% gap.

Evaluation (2)

Number of tuning trials needed to find best performance.



The lower the better

Cori only requires **2 trials** on average for *predictive* ML-based schedulers.

Insight-less baselines are sensitive to the trial ordering (left, right, random).

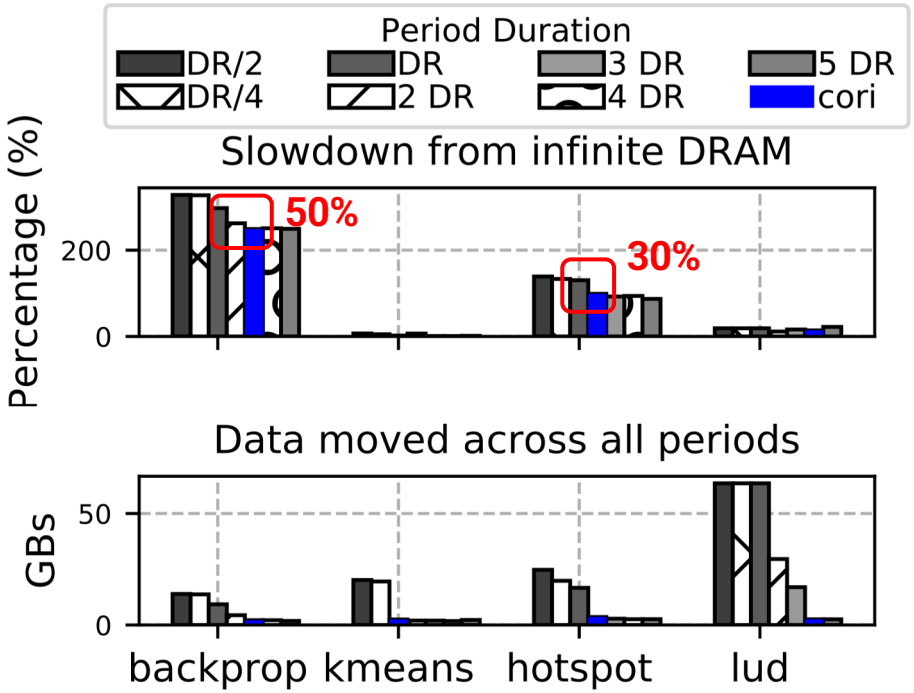
5x reduction in tuning trials

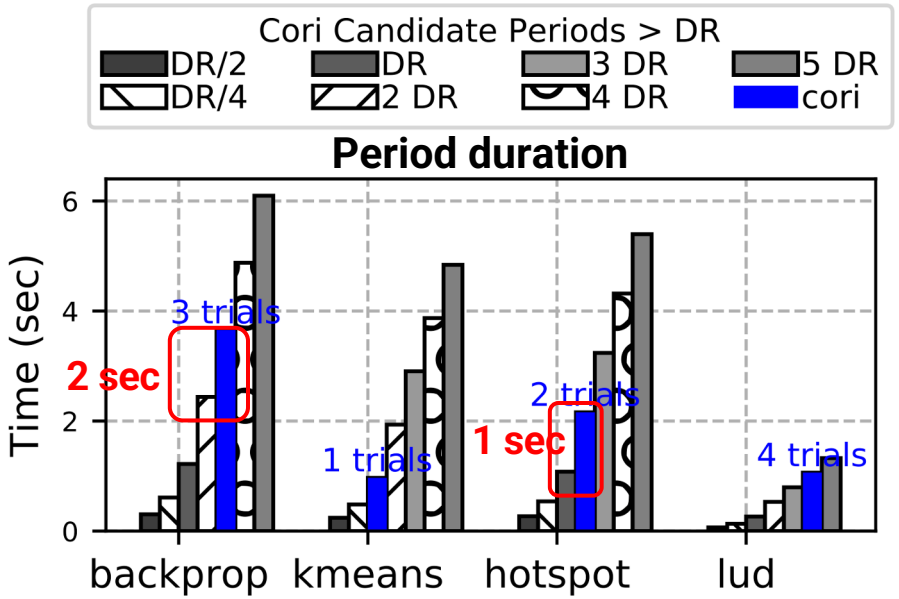
Georgia Tech
CREATING THE NEXT

Evaluation (3)

Validation on Optane persistent memory.

The lower the better

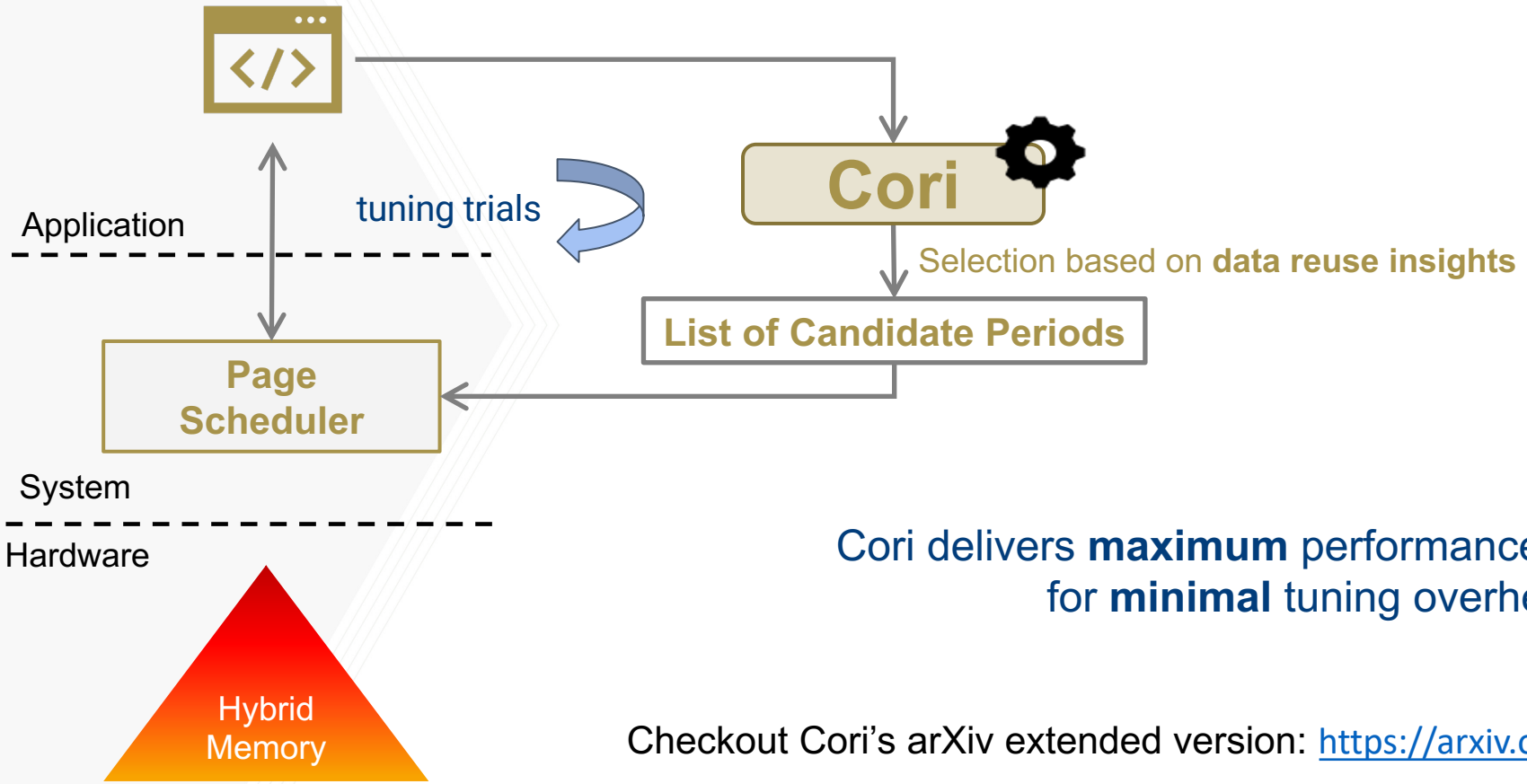




Even a difference of 1-2 seconds in period duration can reduce performance by 30%-50%.

Summary of Cori

Greek Trivia: According to the ancient Greek mythology, Cori (short for Terpsichore) was the muse of dance, sister of Kleio, daughter of Mnemosyne, goddess of memory.



Cori is open source.

Cori delivers **maximum** performance improvements for **minimal** tuning overheads.

Checkout Cori's arXiv extended version: https://arxiv.org/abs/2101.07200