ATC'24

ARTIFACT EVALUATED — AVAILABLE

ARTIFACT EVALUATED — FUNCTIONAL

Telescope

Telemetry for Applications with Gargantuan Memory Footprints

Alan Nair (presenting)

Sandeep Kumar, Aravinda Prasad, Ying Huang
Andy Rudoff, Srinivas Subramoney

Intel Labs

Background

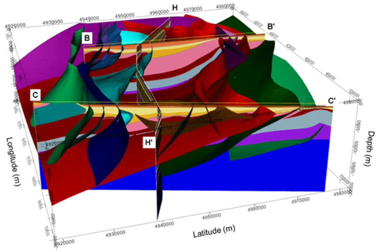
Terabyte-scale Applications

GPT-4

Amazon
Elasticache

Weather
Prediction

Seismic
Modeling



Machine Learning
LLMs

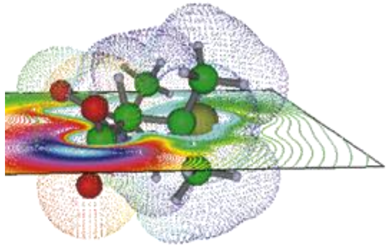
In-Memory Databases

High-Performance Computing

PaLM2

Google's Next-Generation
Large Language Model

SAP HANA

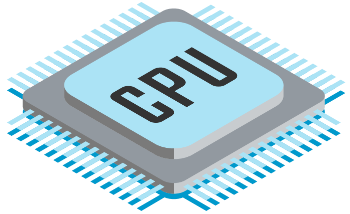


Manufacturing
Materials
Simulations

Drug Development
Cancer Analysis

Qubit
PHARMACEUTICALS

Terabyte-scale Memory Systems



SRAM Caches

Latency low
Bandwidth high
Capacity low
Cost/byte high

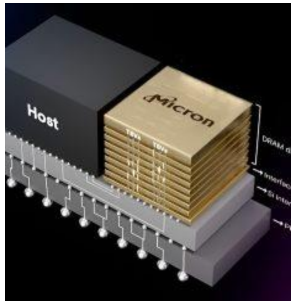
Memory

Latency high
Bandwidth low
Capacity high
Cost/byte low

Storage

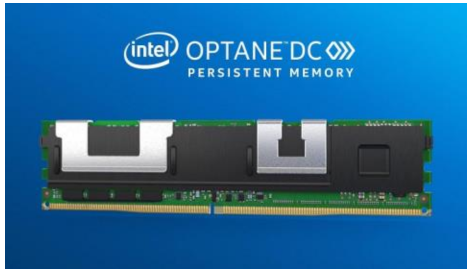
HOT MEMORY TIERS

COLD MEMORY TIERS



HBM

DDR-DRAM

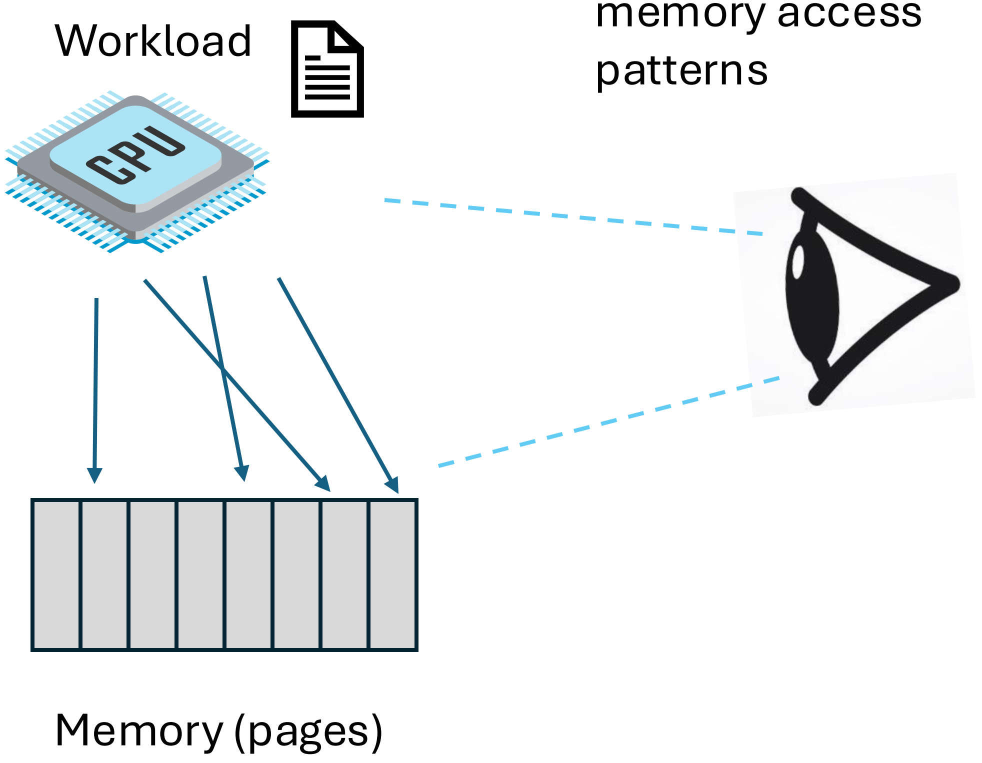


Persistent Memory

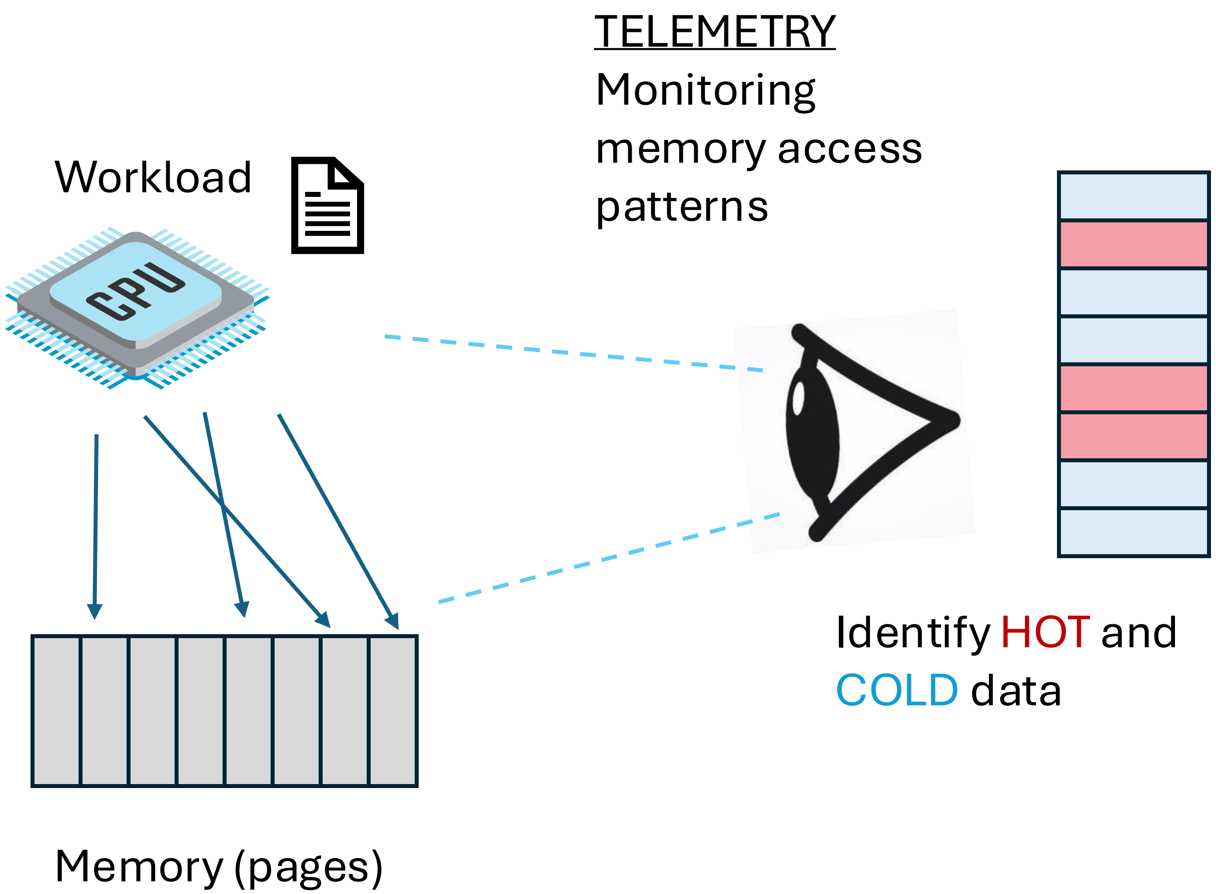
CXL Memory Pooling/Expansion

Tiered Memory needs Telemetry

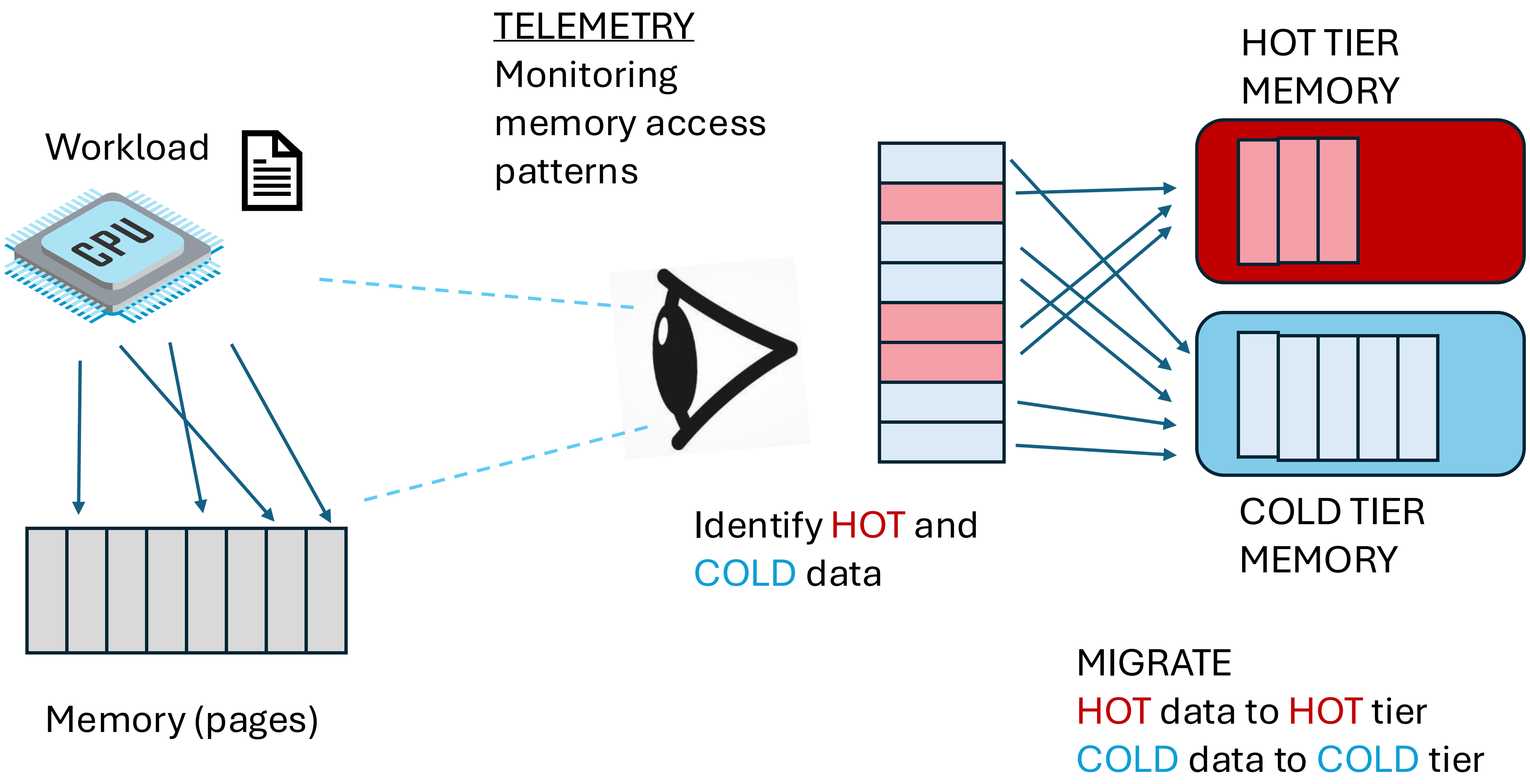
<u>TELEMETRY</u>
Monitoring memory access patterns



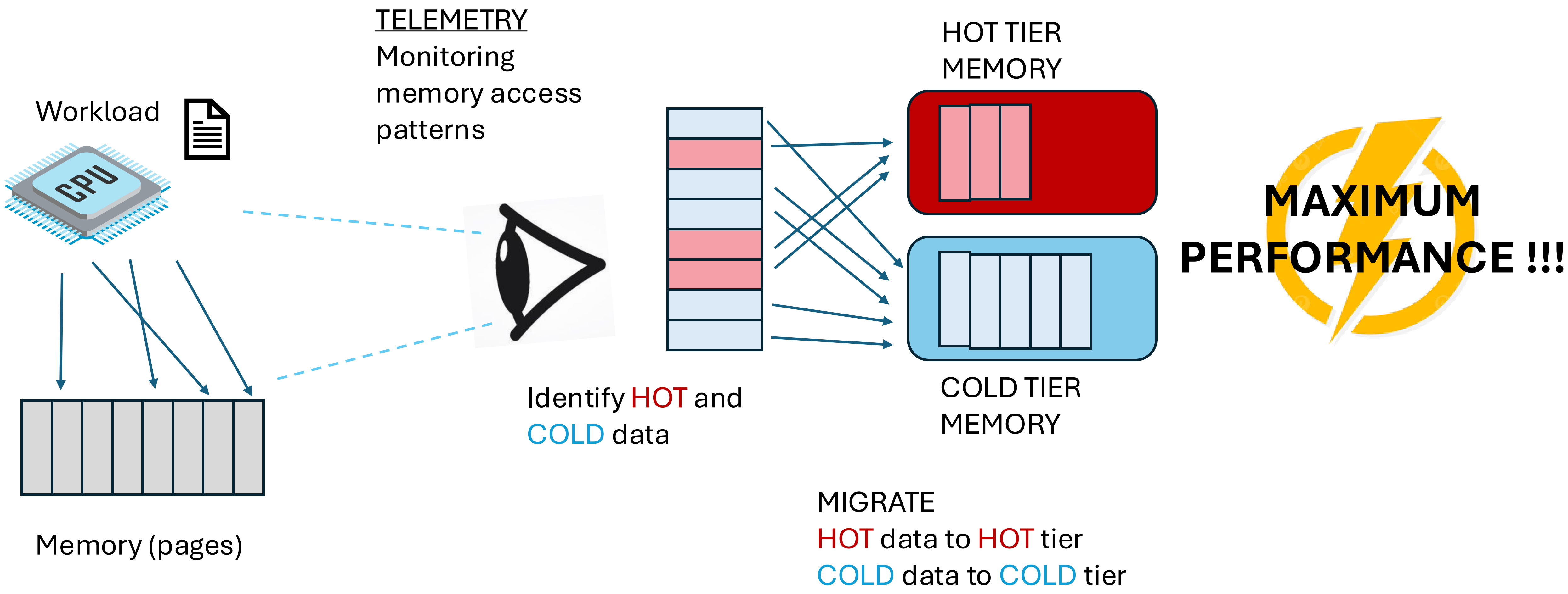
Tiered Memory needs Telemetry



Tiered Memory needs Telemetry

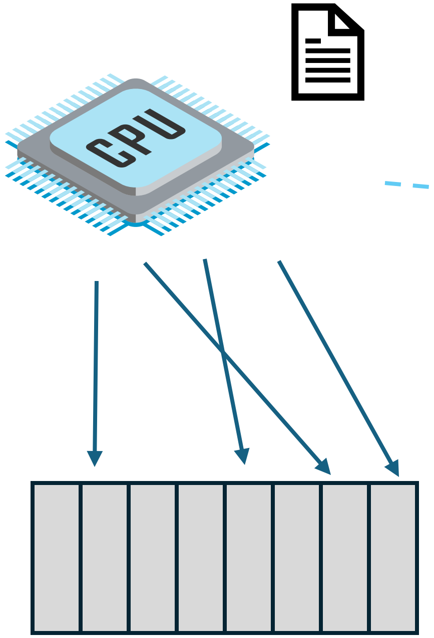


Tiered Memory needs Telemetry



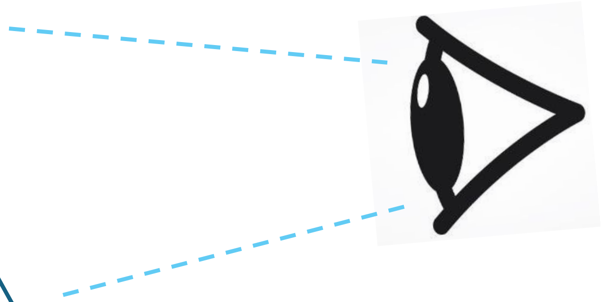
Data Placement needs Telemetry

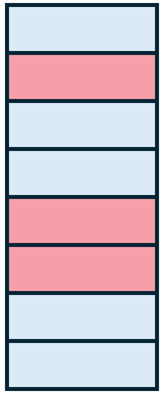
Workload



Memory (pages)

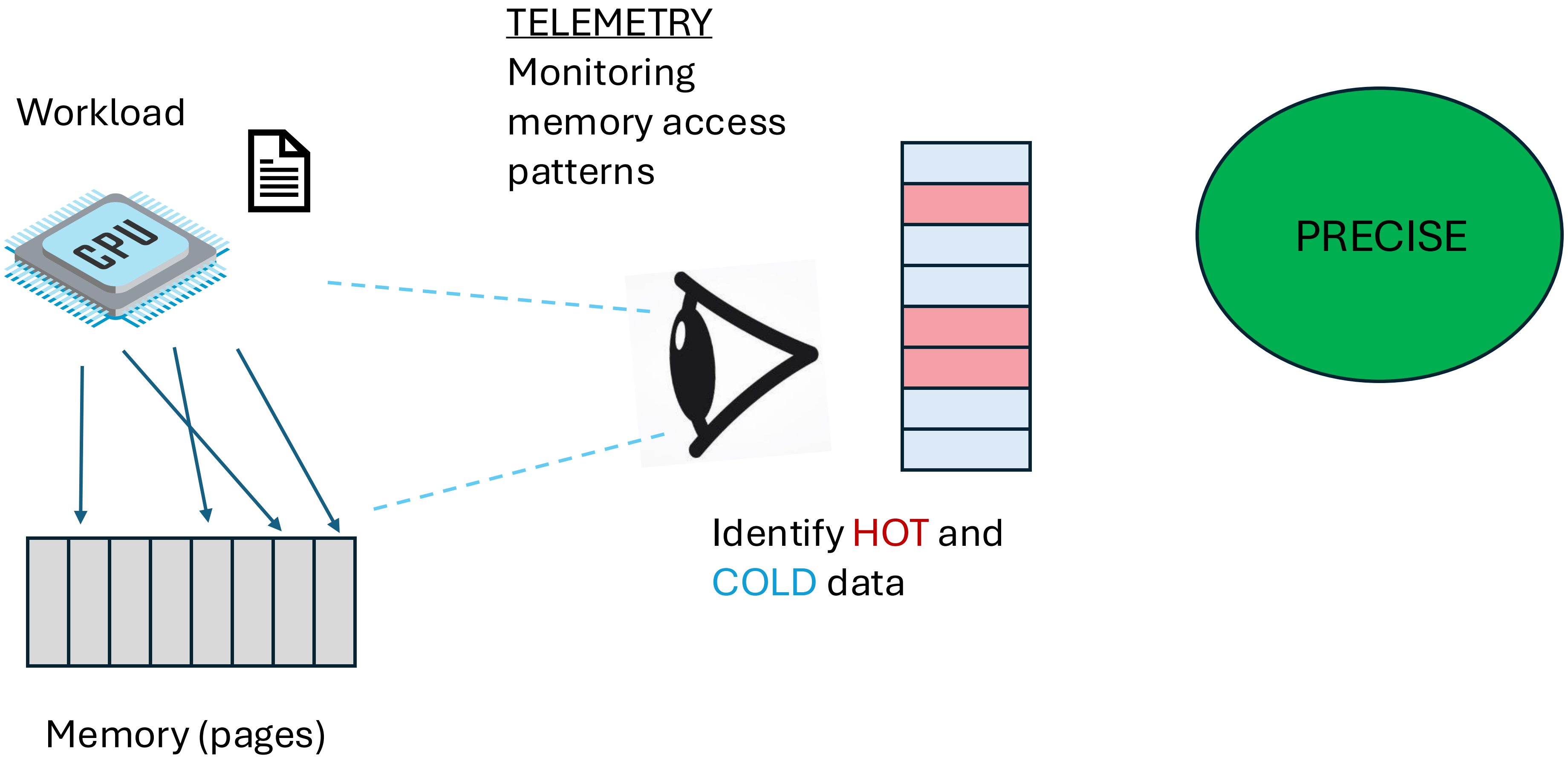
<u>TELEMETRY</u>
Monitoring memory access patterns



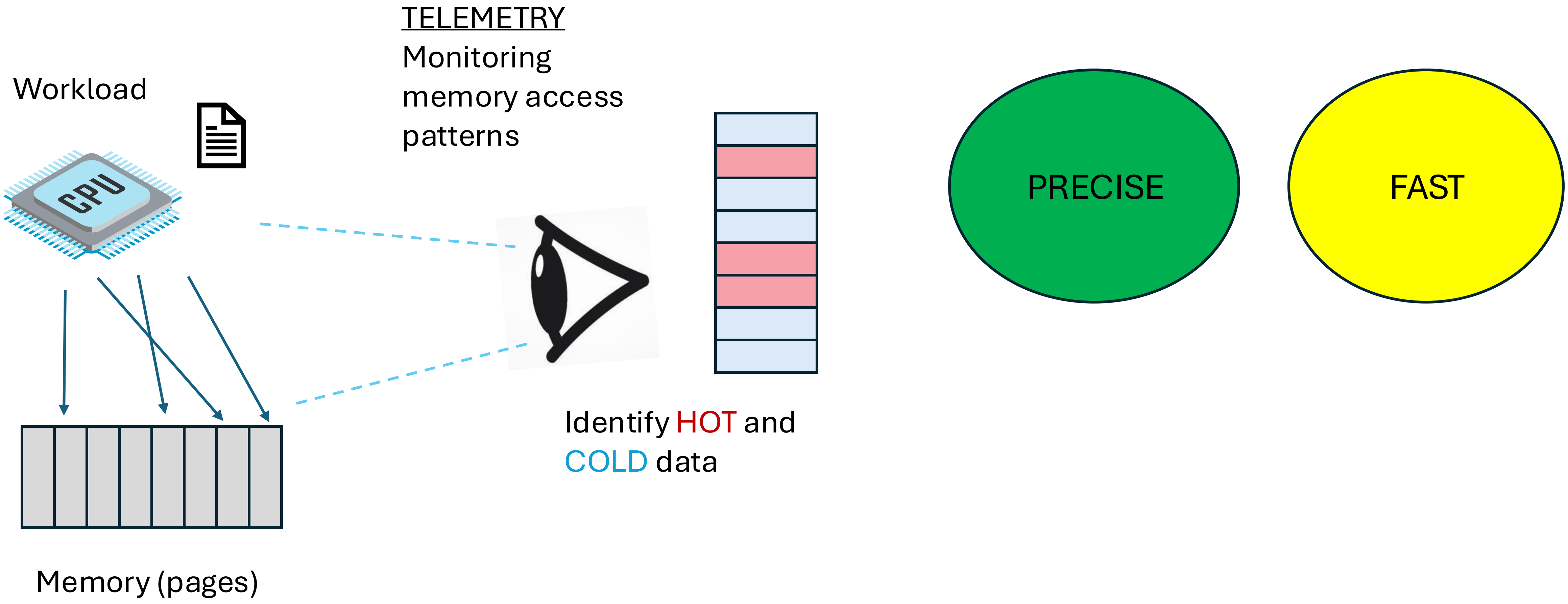


Identify HOT and COLD data

Data Placement needs Telemetry

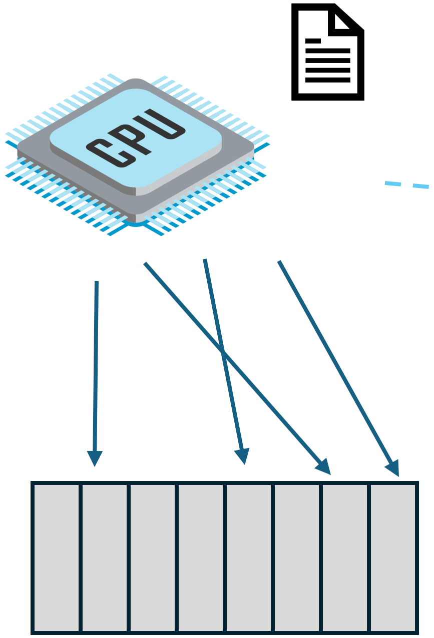


Data Placement needs Telemetry



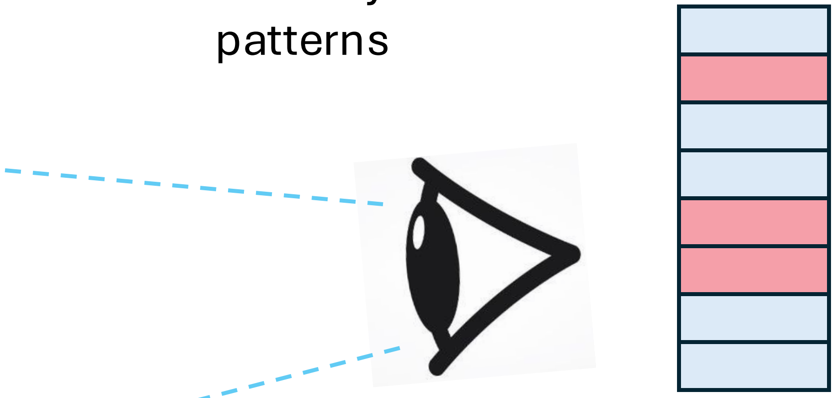
Data Placement needs Telemetry

Workload



Memory (pages)

TELEMETRY
Monitoring memory access patterns



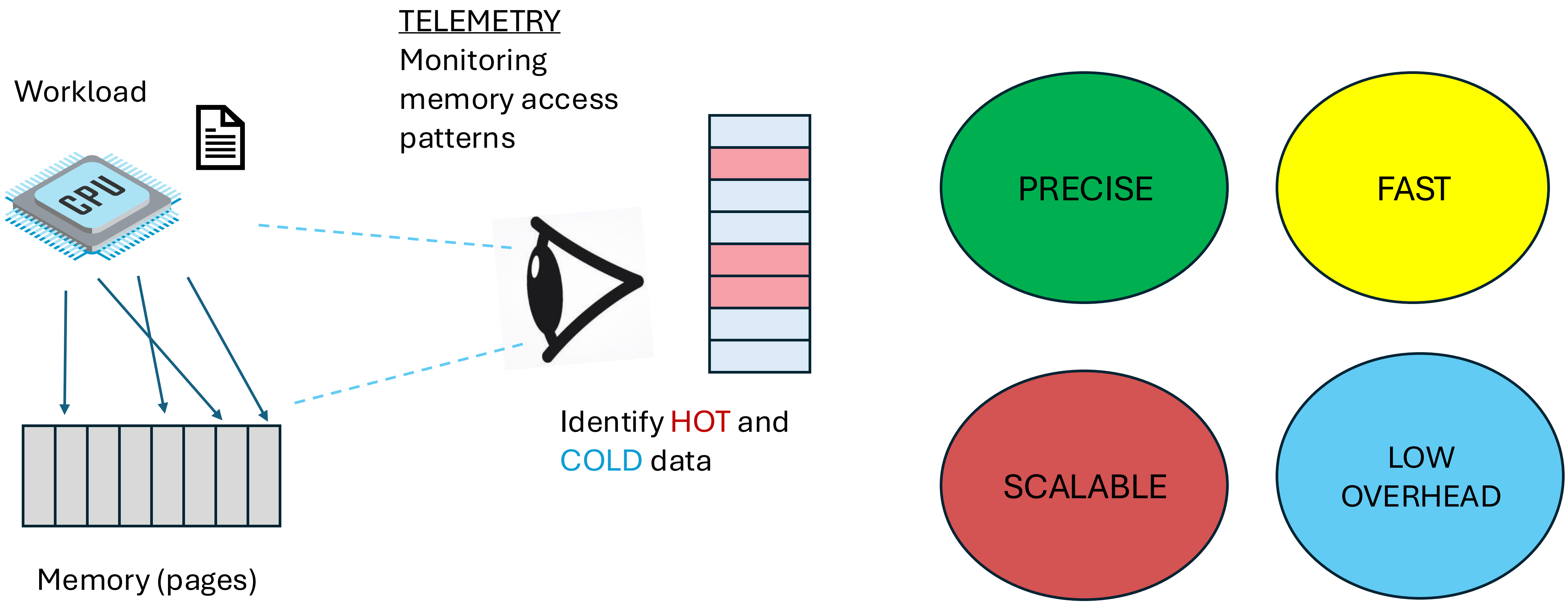
Identify HOT and COLD data

PRECISE

FAST

SCALABLE

Data Placement needs Telemetry



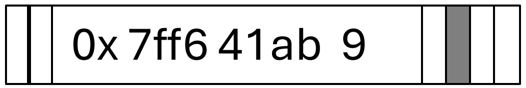
Prior Approaches

1. LINEAR SCANNING OF PAGE TABLE ENTRIES (PTEs)
2. REGION-BASED SAMPLING
3. PERFORMANCE COUNTERS

Linear Scanning of PTEs

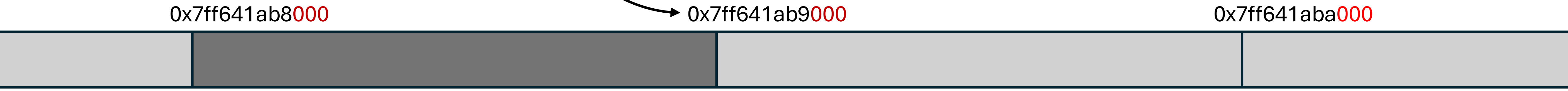
Status Bits

1. PRESENT
2. PROTNONE
3. READ/WRITE
4. USER-ACCESSIBLE
5. ACCESSED
6. DIRTY

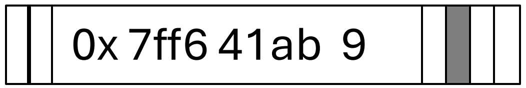


Physical
Page Frame
Number (PFN)

Physical Memory



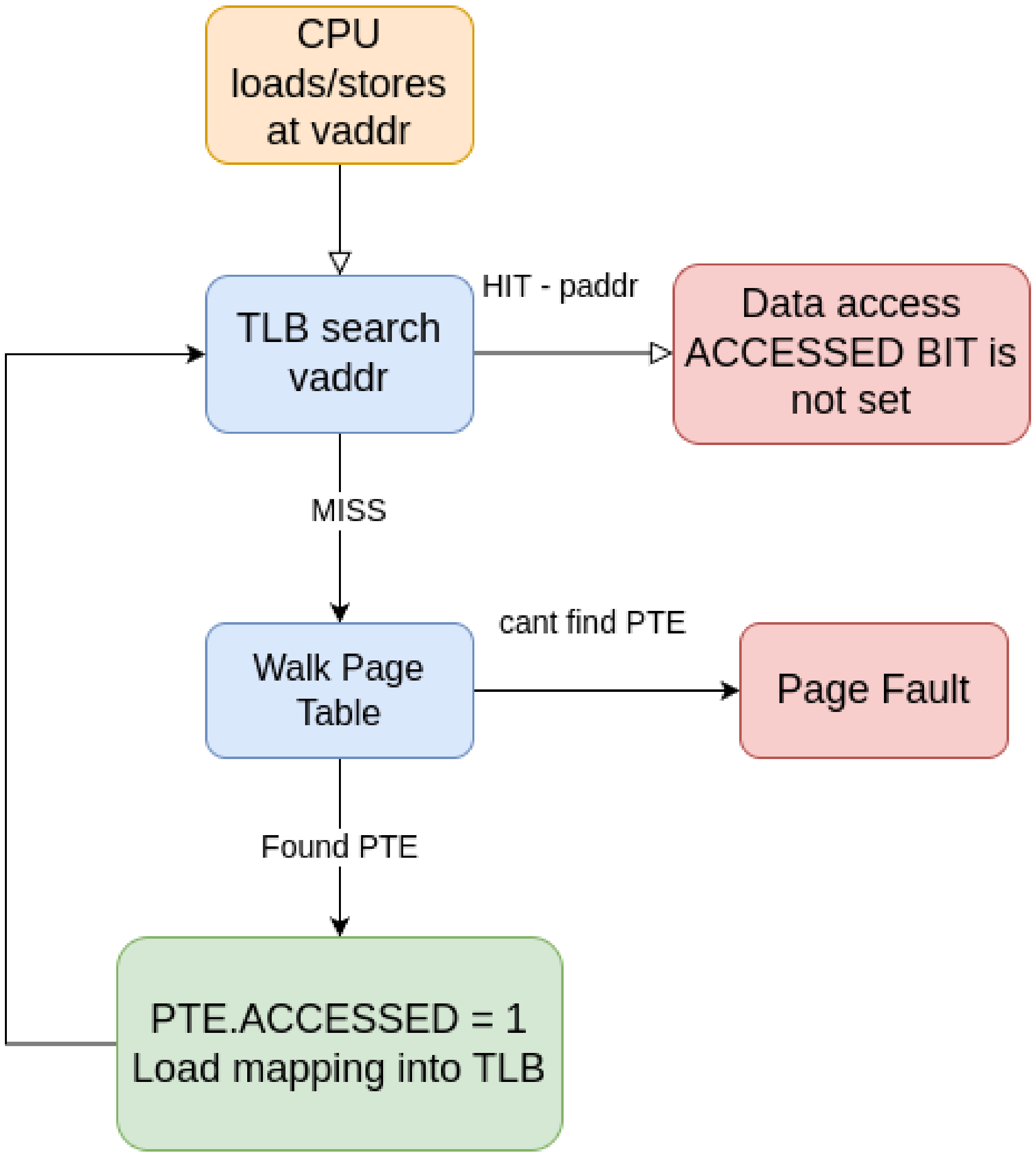
#1 Linear Scanning of PTEs



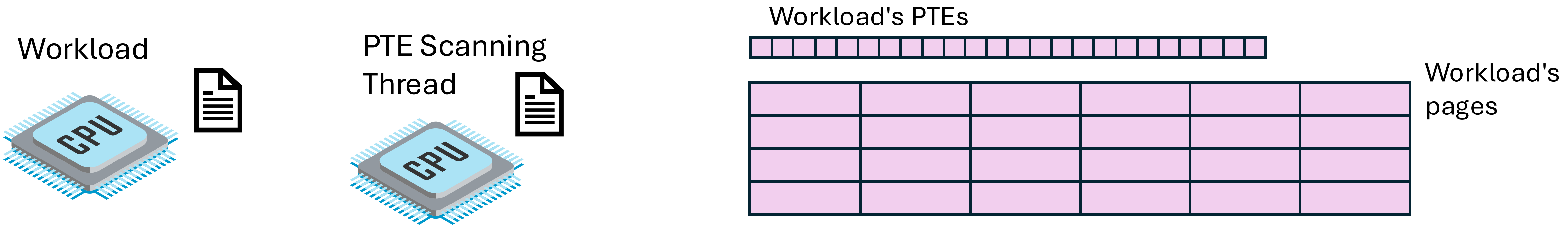
Physical
Page Frame
Number (PFN)

<u>Status Bits</u>

1. PRESENT
2. PROTNONE
3. READ/WRITE
4. USER-ACCESSIBLE
5. ACCESSED
6. DIRTY

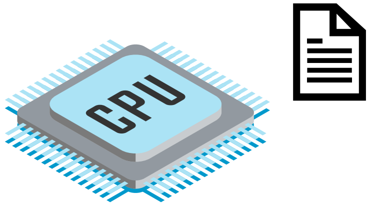


Linear Scanning of PTEs

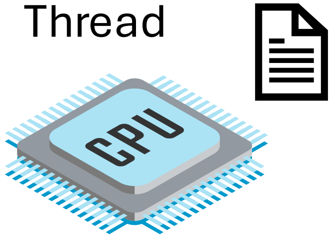


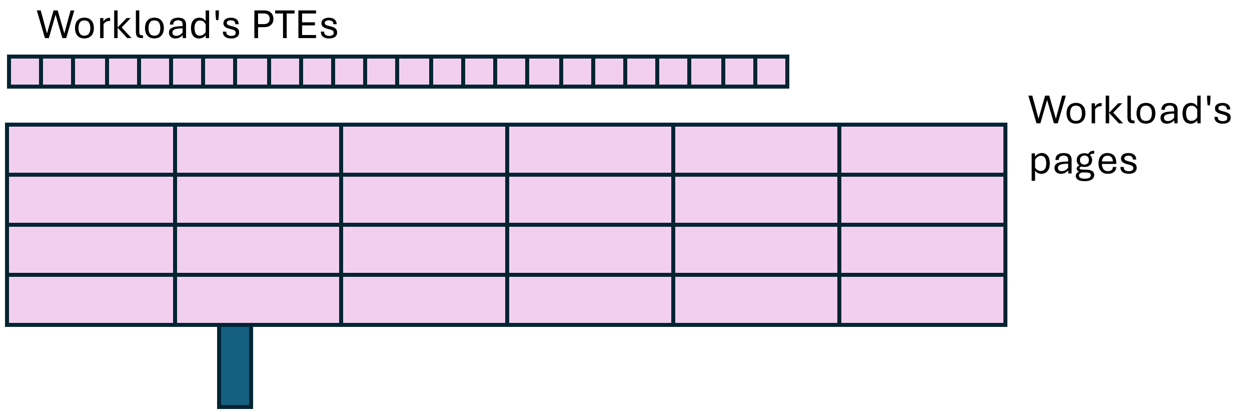
Linear Scanning of PTEs

Workload



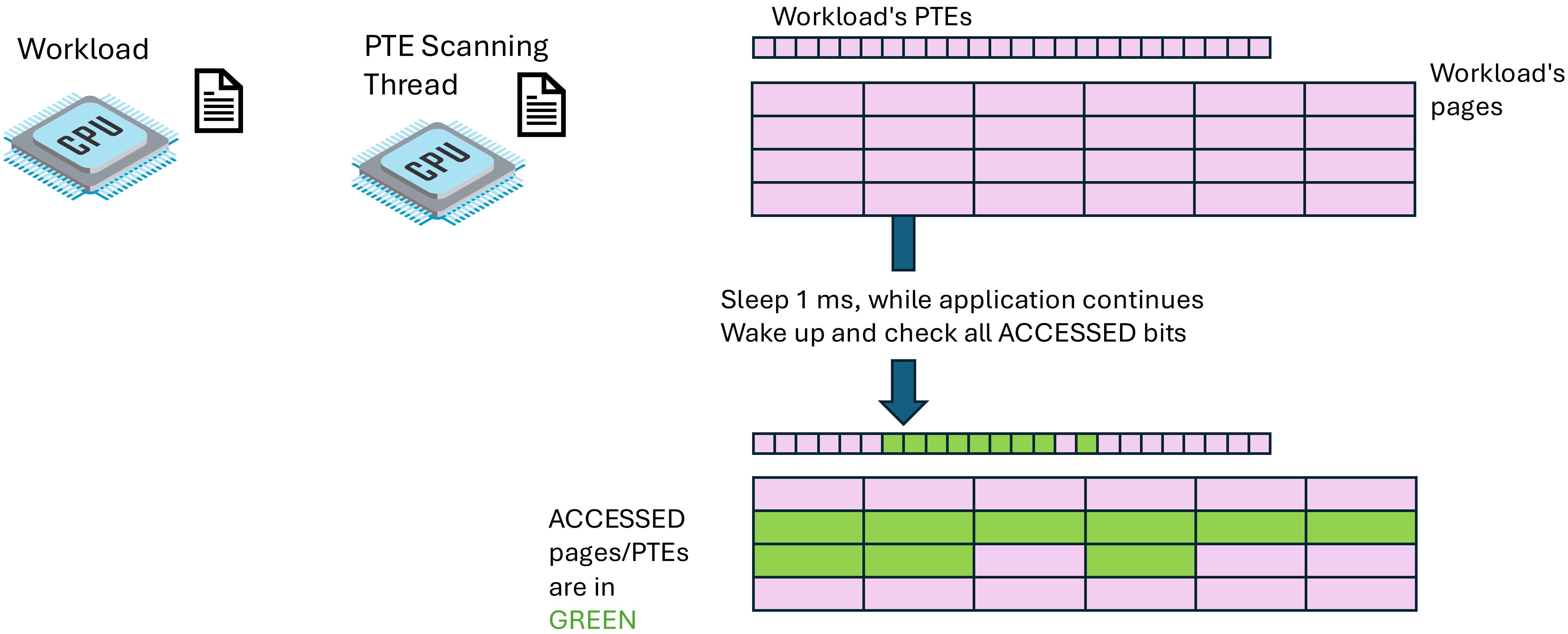
PTE Scanning
Thread



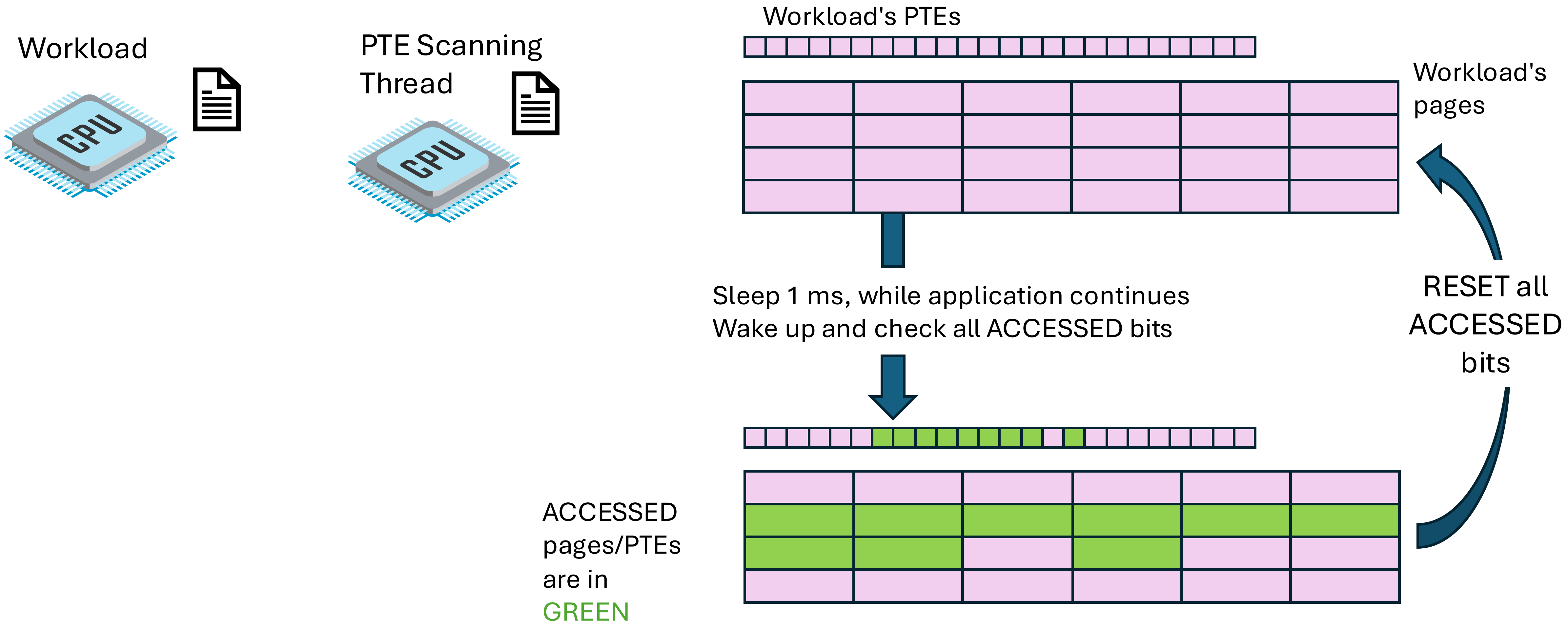


Sleep 1 ms, while application continues
Wake up and check all ACCESSED bits

Linear Scanning of PTEs

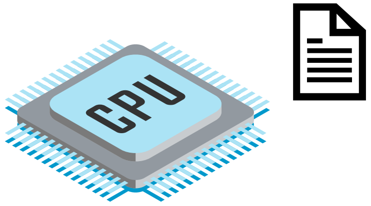


Linear Scanning of PTEs

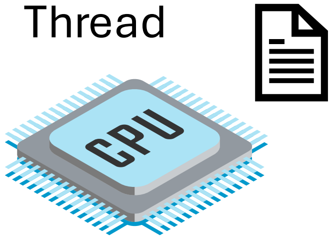


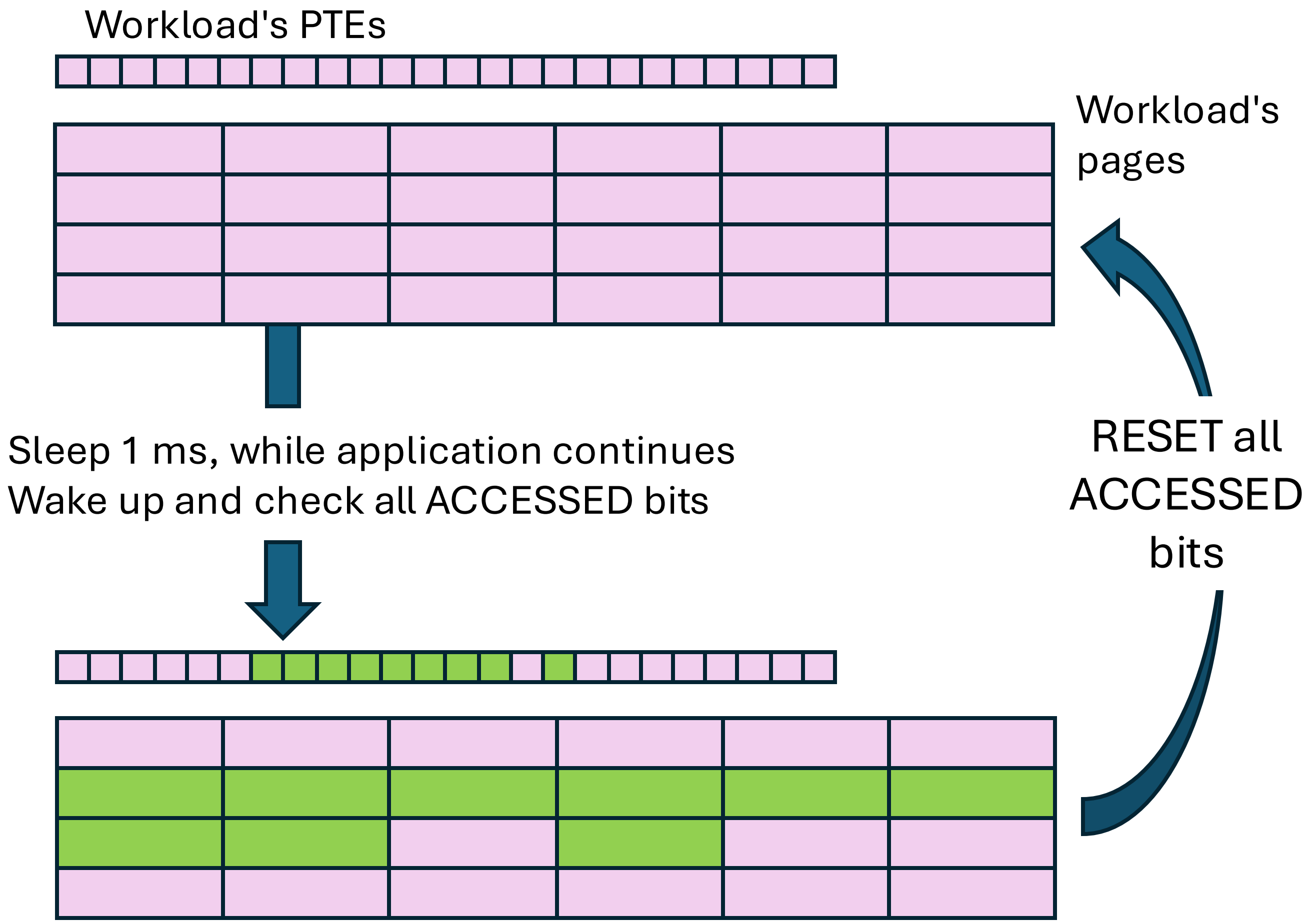
Linear Scanning of PTEs

Workload



PTE Scanning Thread





Example: MGLRU, kstaled

ACCESSED pages/PTEs are in GREEN

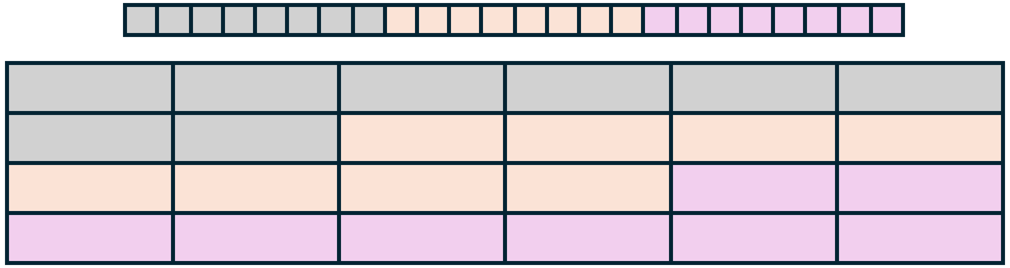
Limitation

1 TB memory = 256M pages (4K each)
One full scan takes > 1 minute

Does not scale well with increasing number of pages !

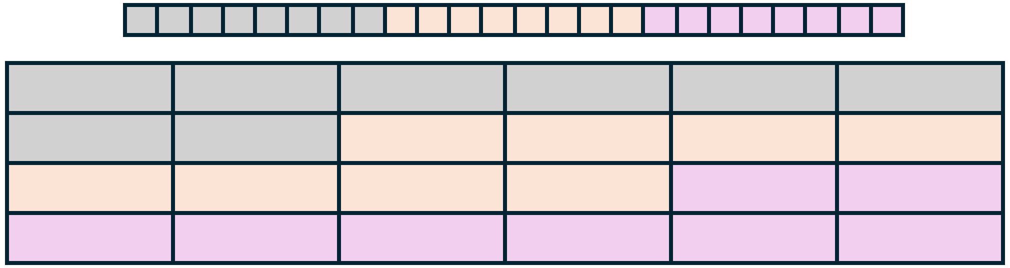
Need Telemetry that converges to access pattern in SECONDS

#2 Region-Based Sampling



Workload's Mapped Address
Space is divided into Regions.

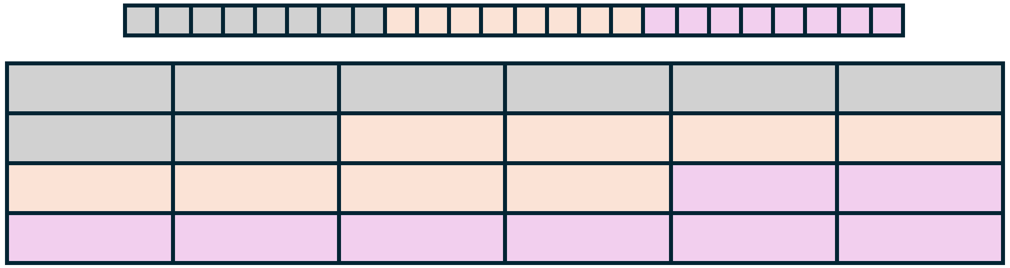
#2 Region-Based Sampling



Workload's Mapped Address Space is divided into Regions.

Randomly pick one PTE per region
RESET these PTEs' ACCESSED bits

#2 Region-Based Sampling

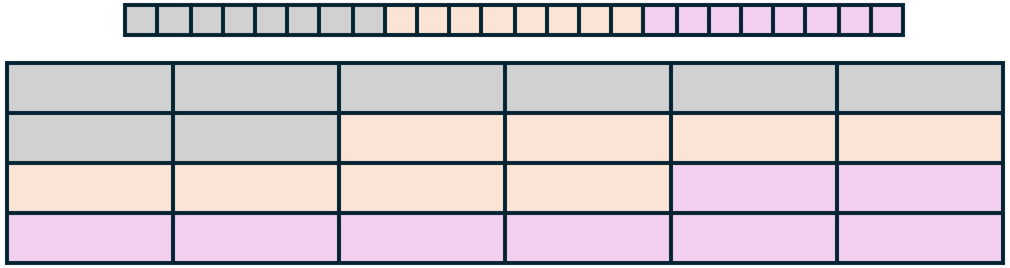


Workload's Mapped Address
Space is divided into Regions.

Randomly pick one PTE per region
RESET these PTEs' ACCESSED bits

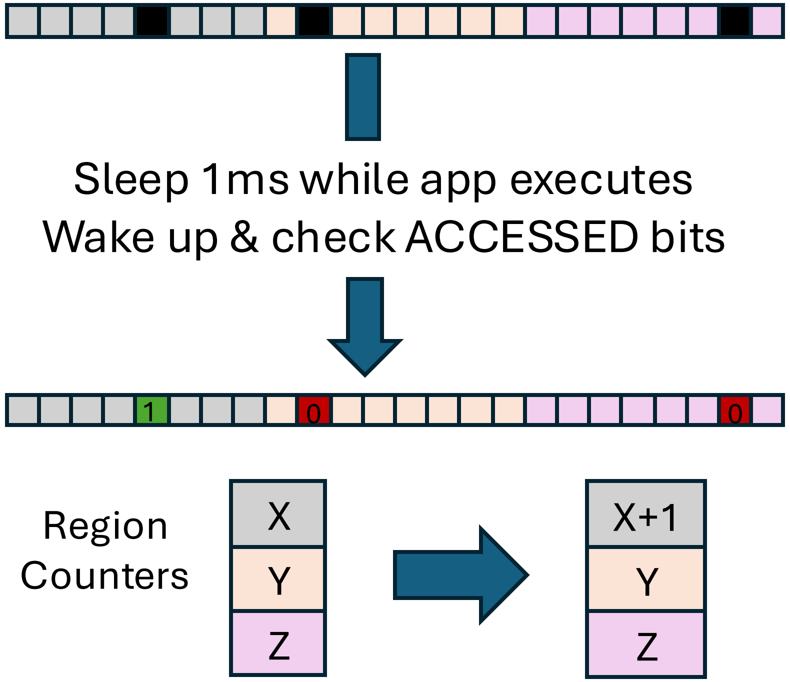
Sleep 1ms while app executes
Wake up & check ACCESSED bits

#2 Region-Based Sampling

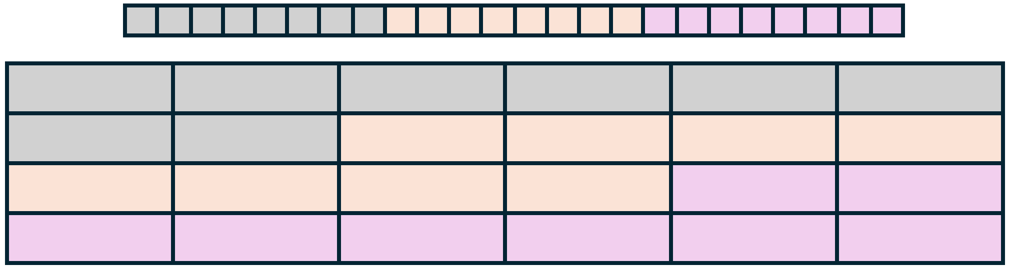


Workload's Mapped Address Space is divided into Regions.

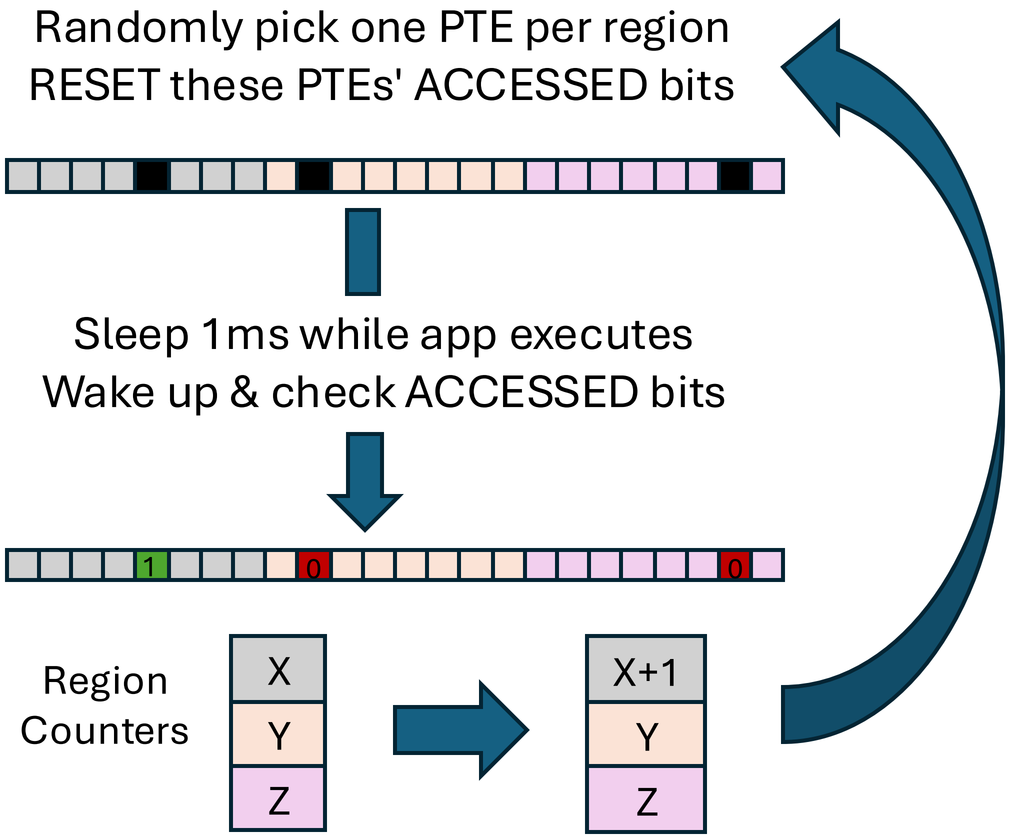
Randomly pick one PTE per region
RESET these PTEs' ACCESSED bits



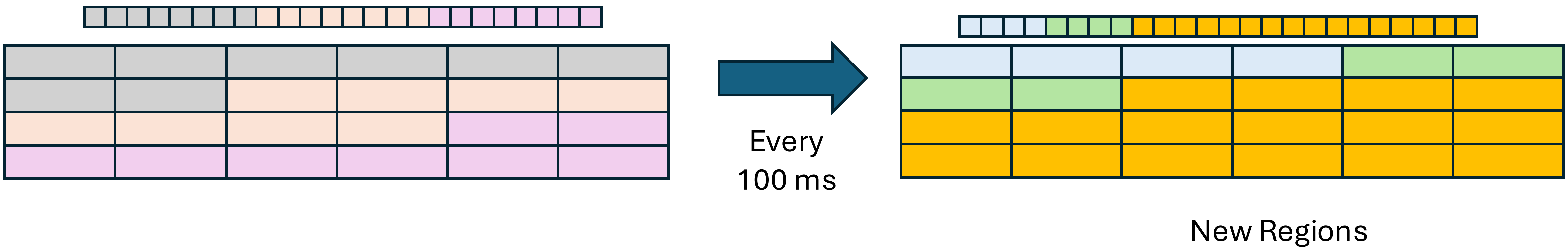
#2 Region-Based Sampling



Workload's Mapped Address Space is divided into Regions.

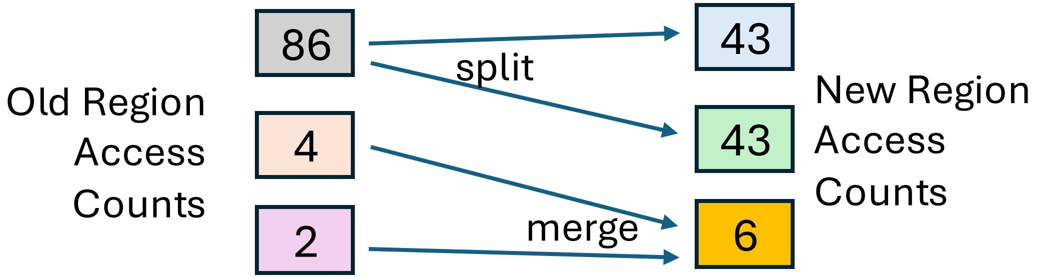


Region-Based Sampling



Adjacent Regions with identical access patterns (access counts) are MERGED.

Regions are split for precision.



Example: DAMON

Limitation

Does not scale well with increasing number of pages per region !

PREMISE

"Sampled page represents its whole region"

Premise holds only if
(#hot pages) / (#pages in region)
is significant

10 GB hot data in a 1 TB address space
DAMON fails to detect ANYTHING in 1 minute

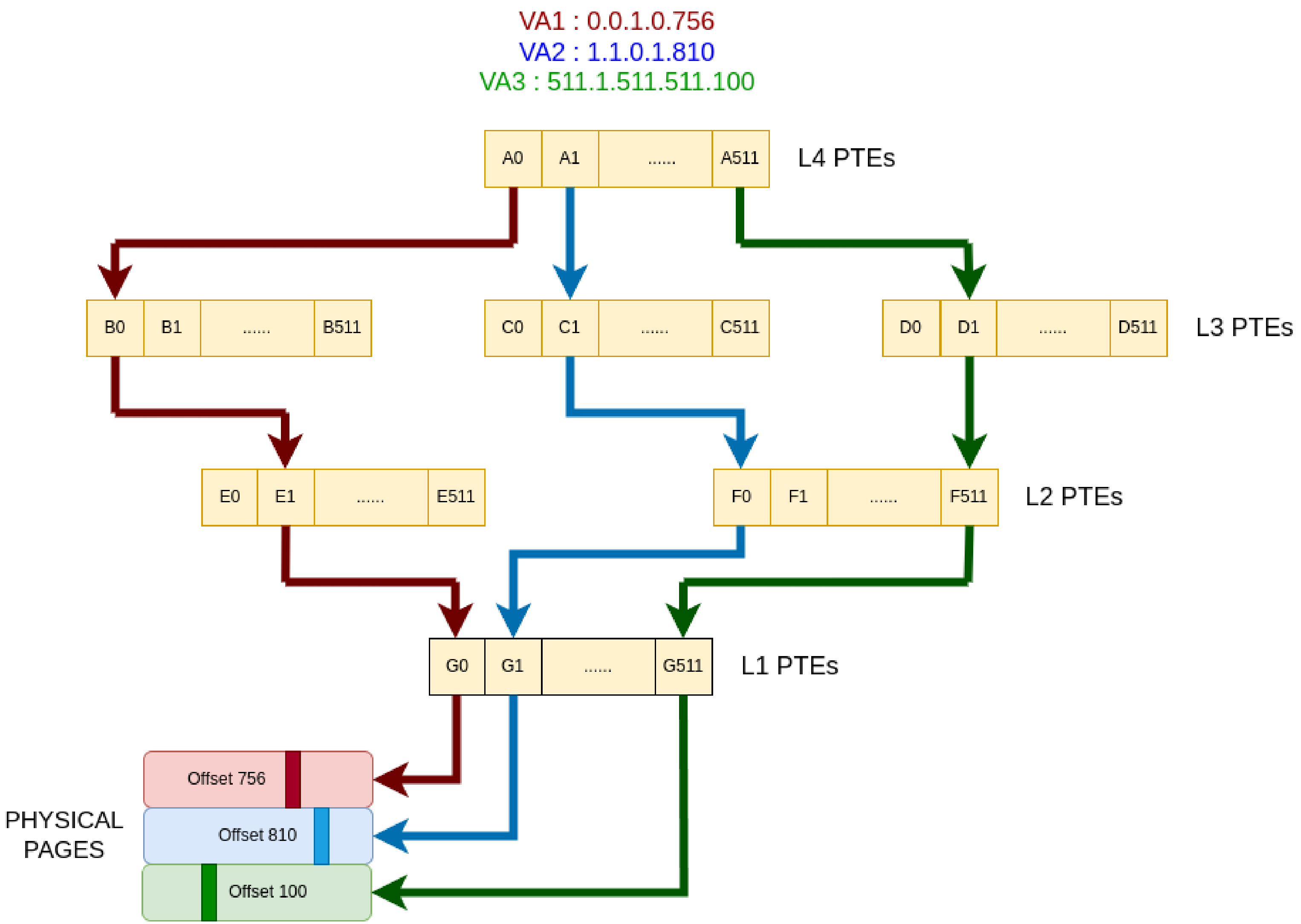
#3 Performance Counters

- Sample hardware events - LLC Miss / TLB Miss / …
- Sample contains data address
- Example: Intel PEBS (Processor Event-Based Sampling) on x86_64

- Limitation: HIGH OVERHEADS
- Huge memory workload => more samples needed => high sampling frequency => Workload Slowdown => UNACCEPTABLE !

Telescope

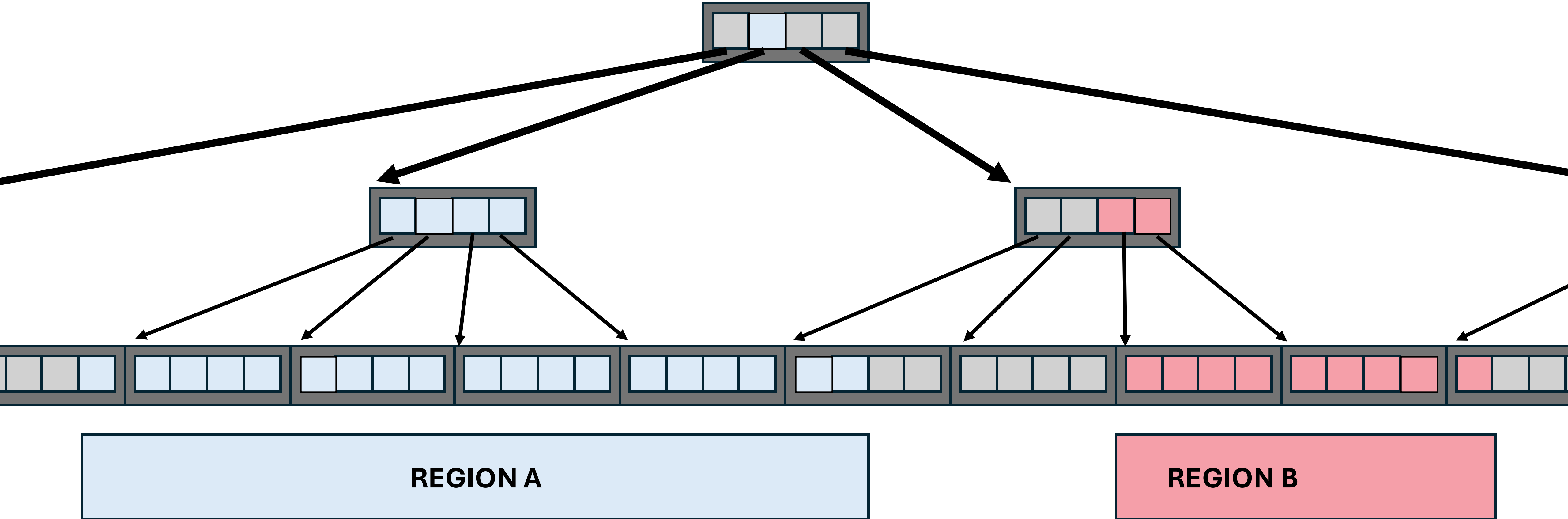
Radix Tree Structure of Page Table



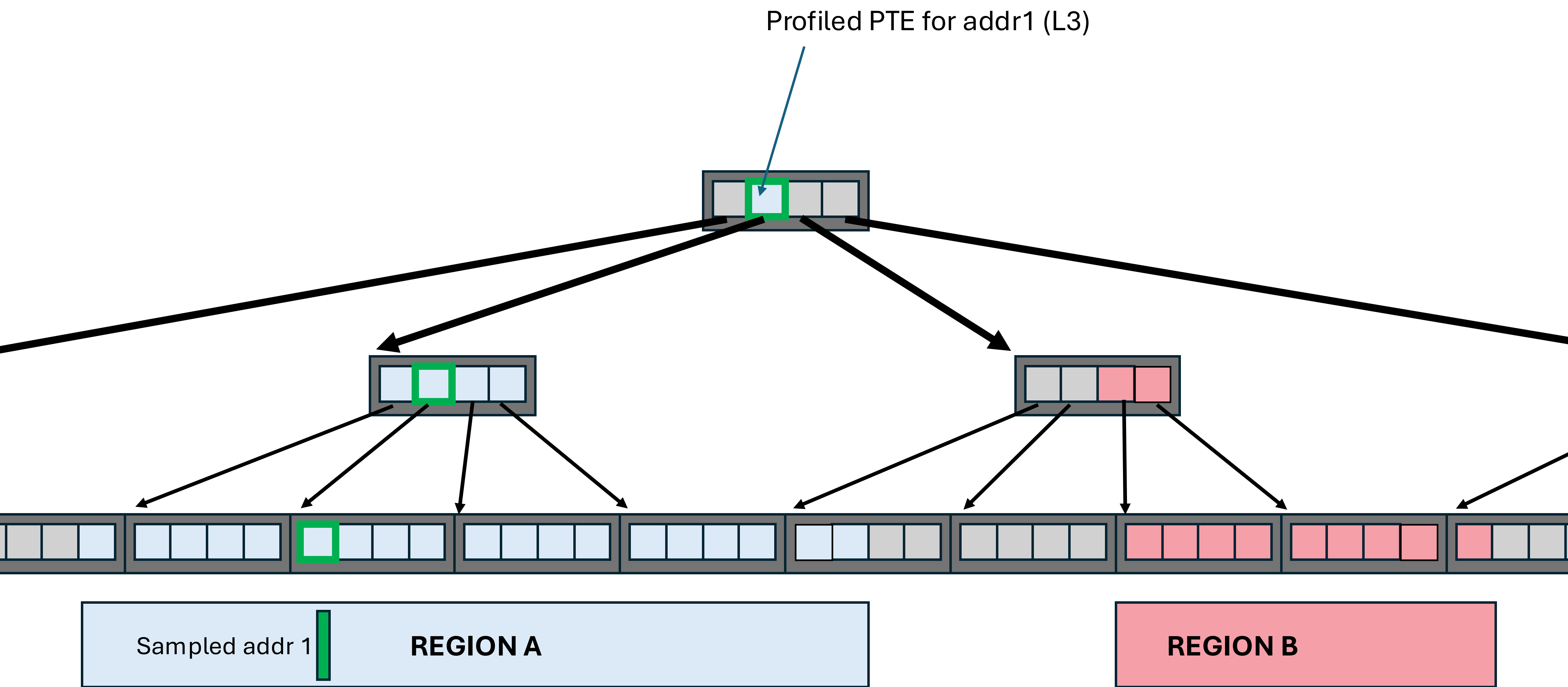
Key Insight

- On Page Walk, Page Table Walker sets the ACCESSED bit at every level of the Page Table tree.
- If at a higher level PTE, ACCESSED=0, then all its lower PTEs will have ACCESSED=0.
- Check the ACCESSED bit at higher level for fast but coarse-grained profiling of access patterns.

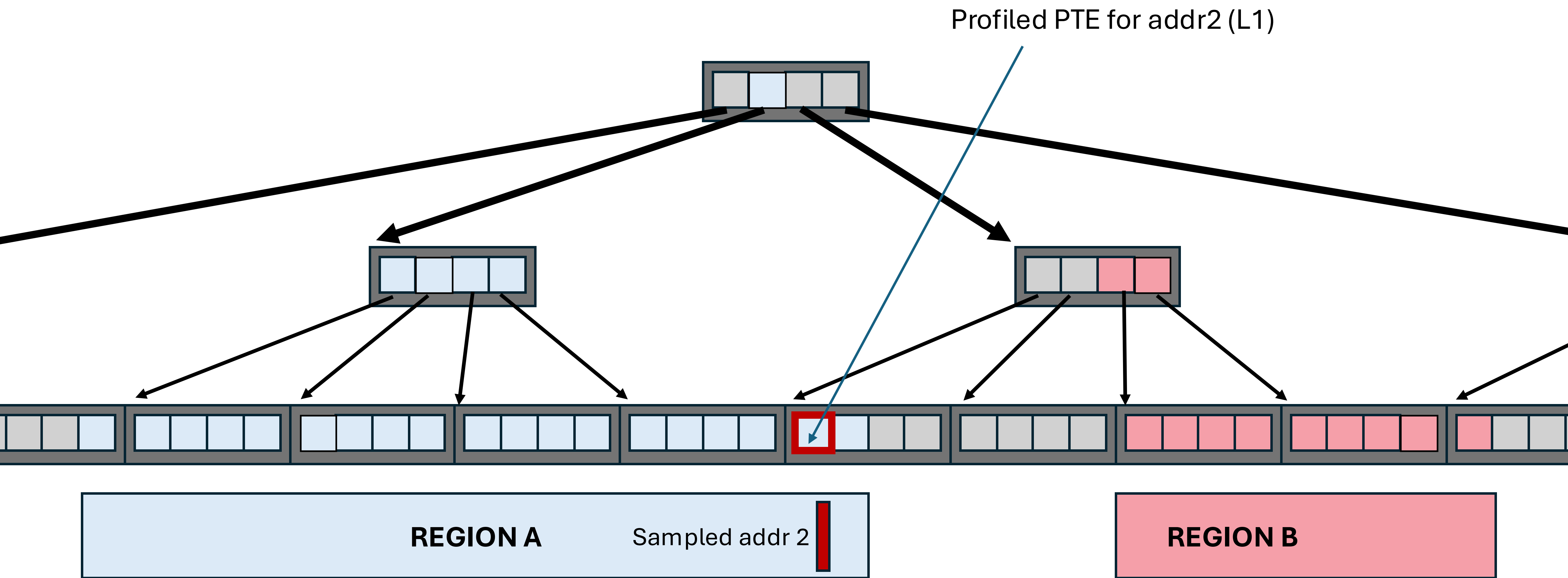
Telescope - Profiling



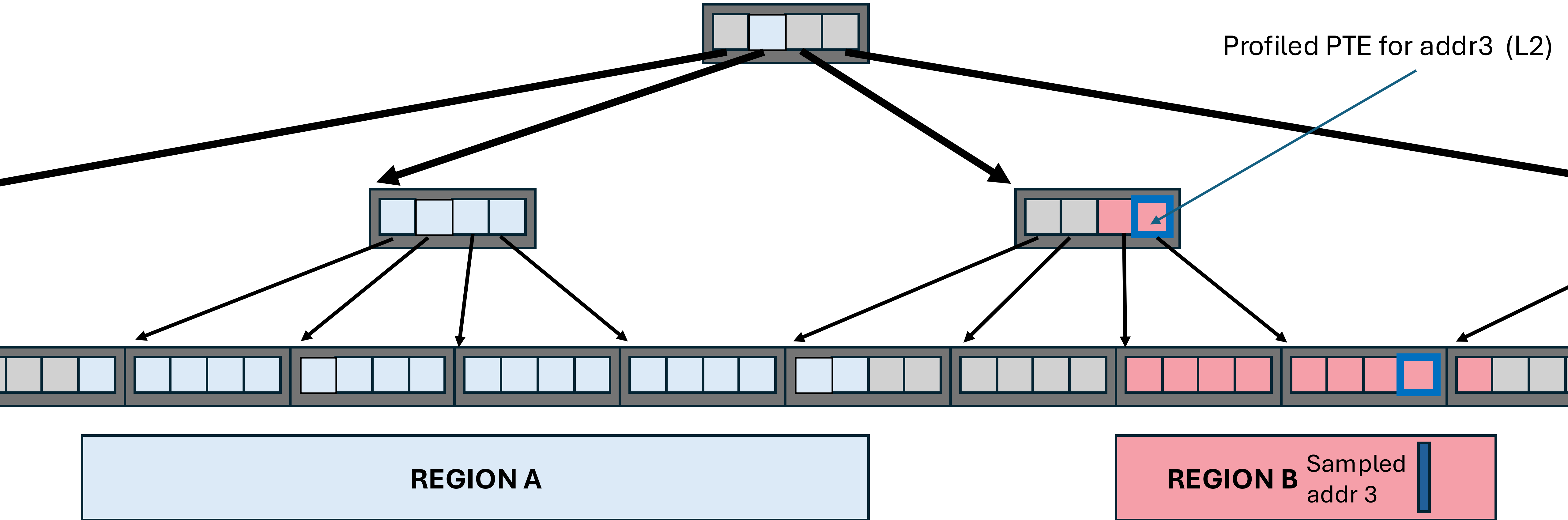
Telescope - Profiling



Telescope - Profiling



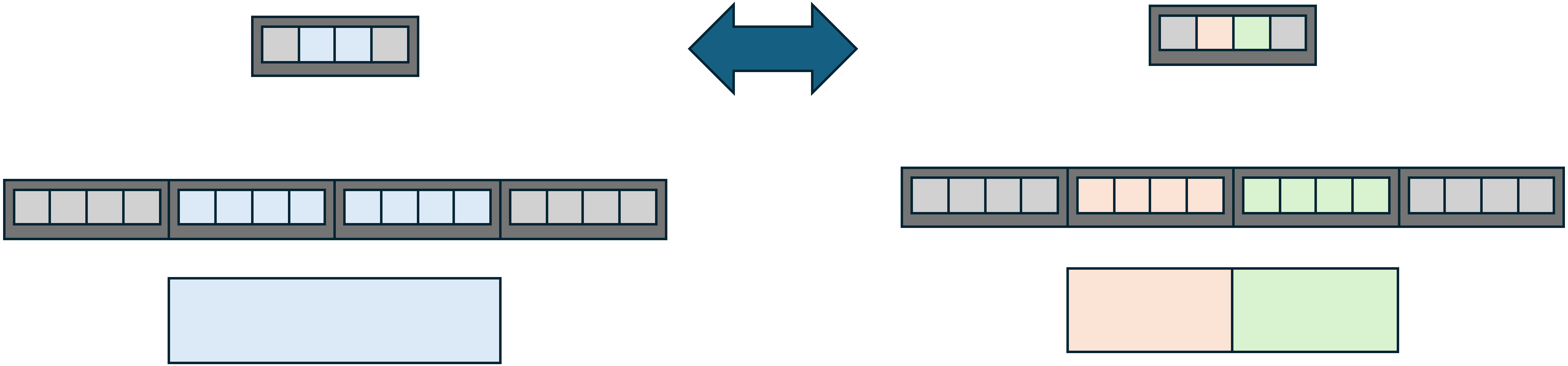
Telescope - Profiling



Telescope - FLEX variant

- Can pick a higher PTE that overshoots the region boundary.
- The overshooted region must be within an error threshold.
- Called FLEXIBLE Telescope, as opposed to BOUNDED Telescope.

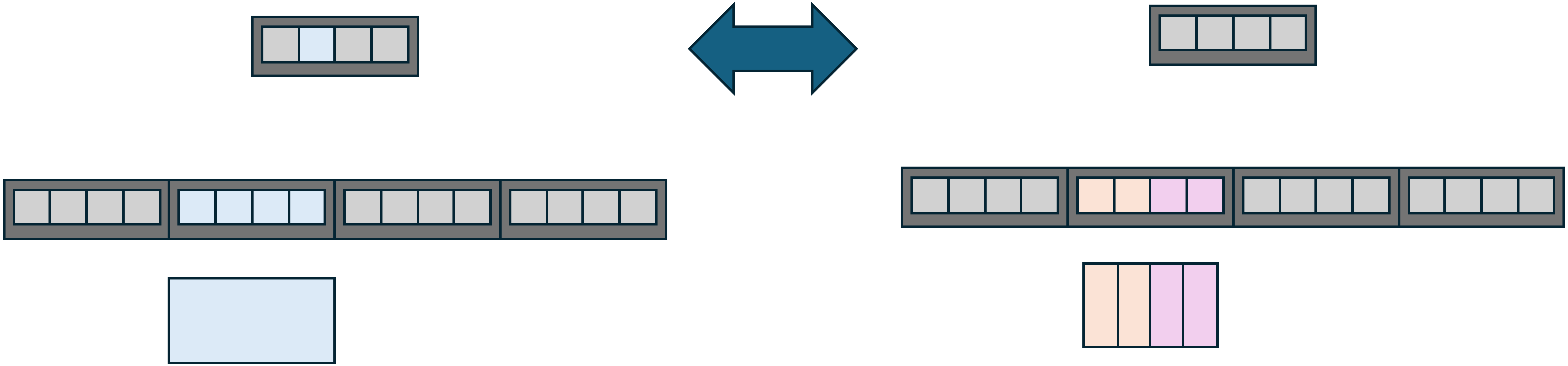
Telescope - Region Split/Merge



Before Split / After Merge

After Split / Before Merge

Telescope - Region Split/Merge



Before Split / After Merge

After Split / Before Merge

Evaluation

Telemetry Techniques Evaluated

DAMON

Region-based Sampling

MODERATE Config
5 ms sampling window

AGGRESSIVE Config
1 ms sampling window

200 ms profiling window
1 sec vma-scan interval

PEBS

Performance Counters

MODERATE Config
5 kHz sampling freq

AGGRESSIVE Config
10 kHz sampling freq

Events sampled:

- MEM_INST_RETIRED.ALL_LOADS
- MEM_INST_RETIRED.ALL_STORES

TELESCOPE

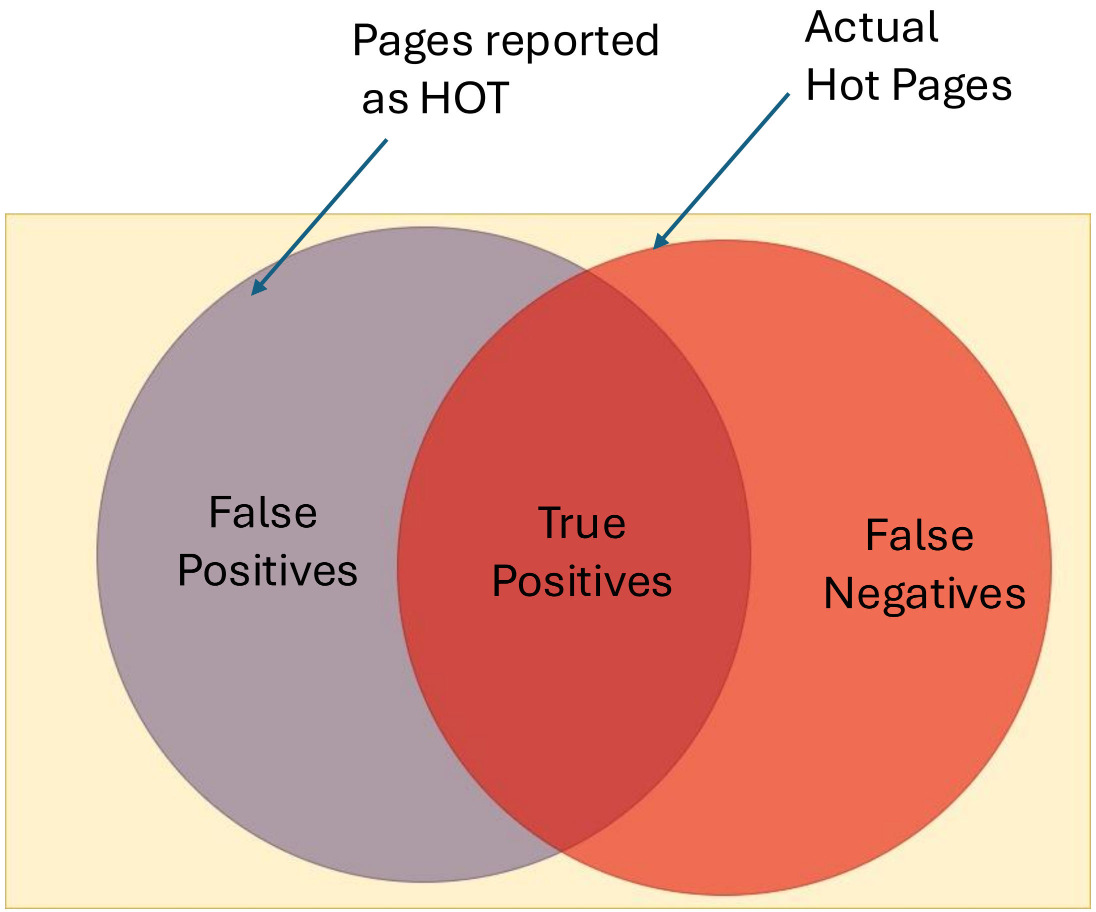
Our Work

BOUNDED Config

FLEX Config
L2: n = 25%
L3: n = 15%

5 ms sampling window
200 ms profiling window
1 sec vma-scan interval

Key Metrics



PRECISION = $\dfrac{\text{(Actual HOT pages reported as HOT)}}{\text{(Total pages reported as HOT)}}$

OR (True Positives) / (True Positives + False Positives)

A measure of Accuracy

RECALL = $\dfrac{\text{(Actual HOT pages reported as HOT)}}{\text{(Actual HOT pages)}}$

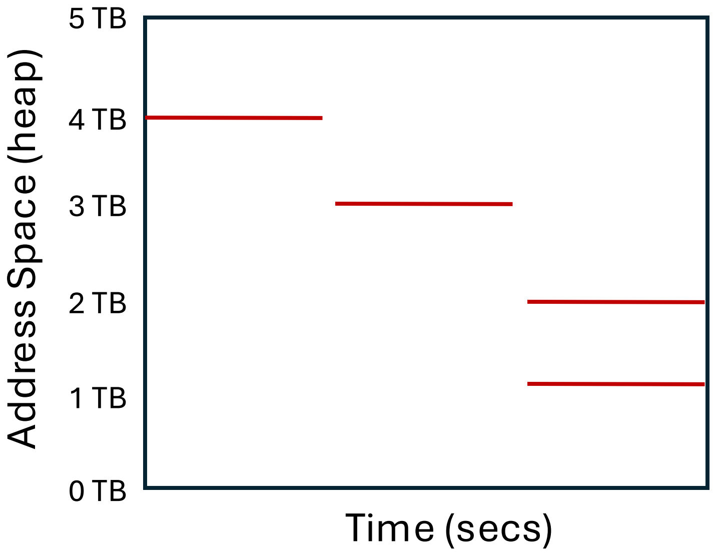
OR (True Positives) / (True Positives + False Negatives)

A measure of Coverage

Microbenchmark

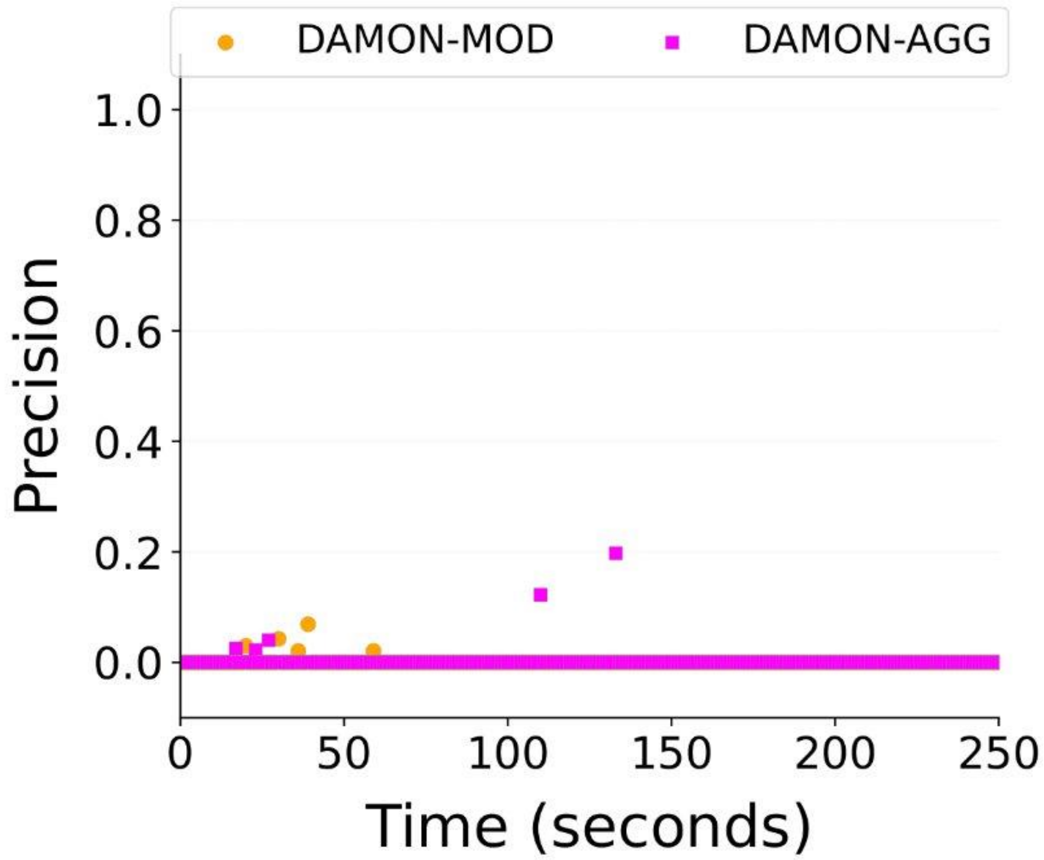
- MASIM (Memory Access Simulator)

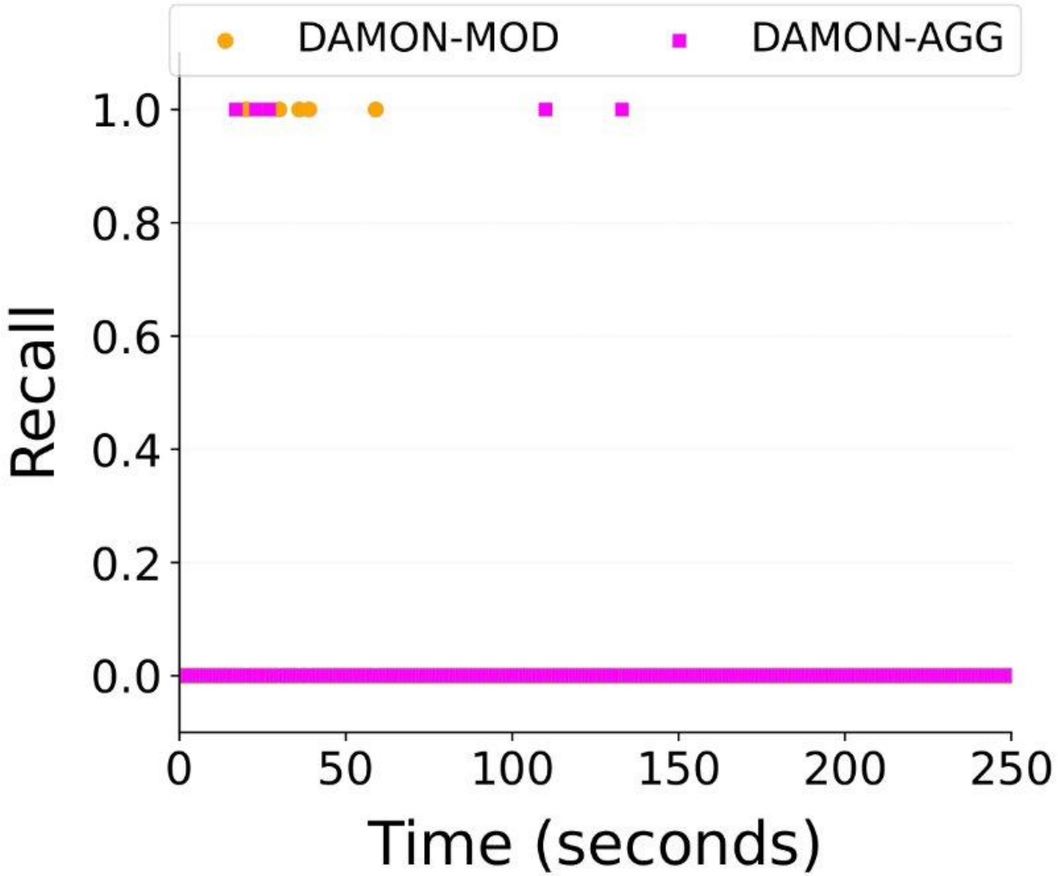
sjp38/masim



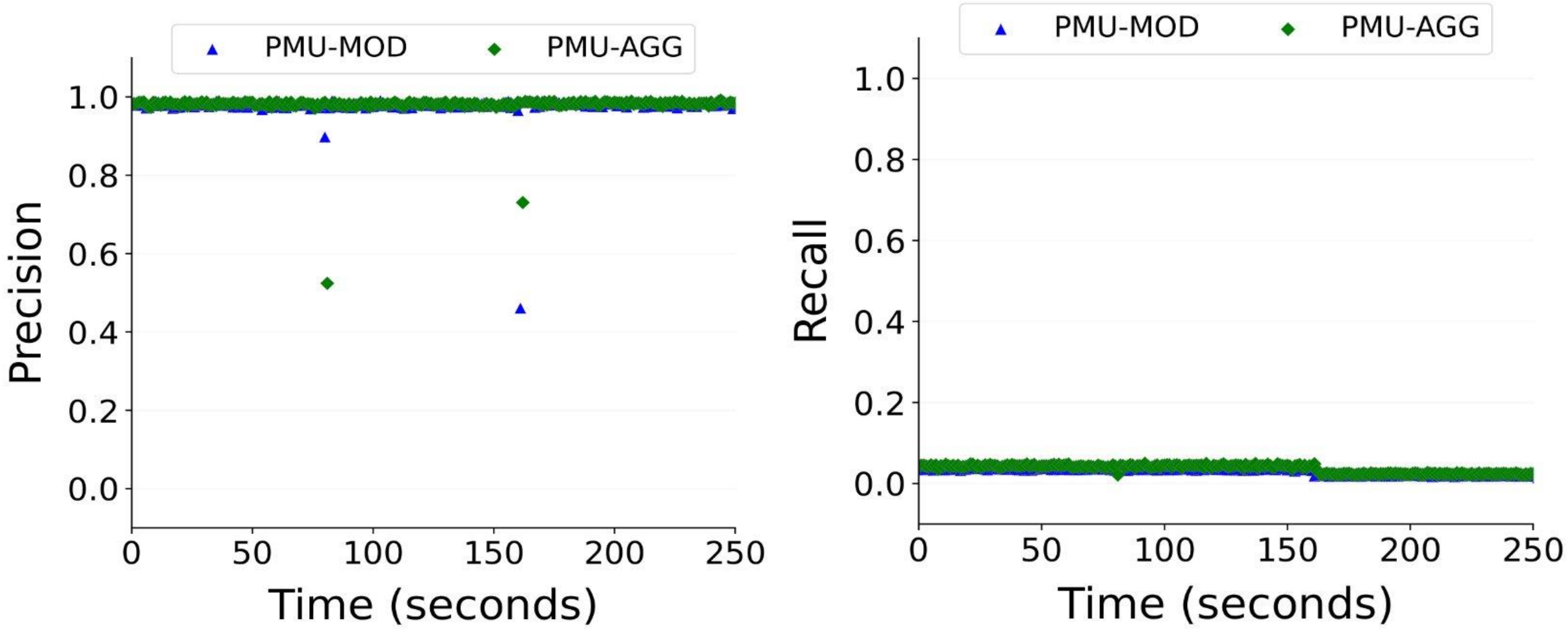
<u>Multi-Phase</u>
10 GB hot *needles* in a 5 TB heap
4 KB pages

Results - DAMON

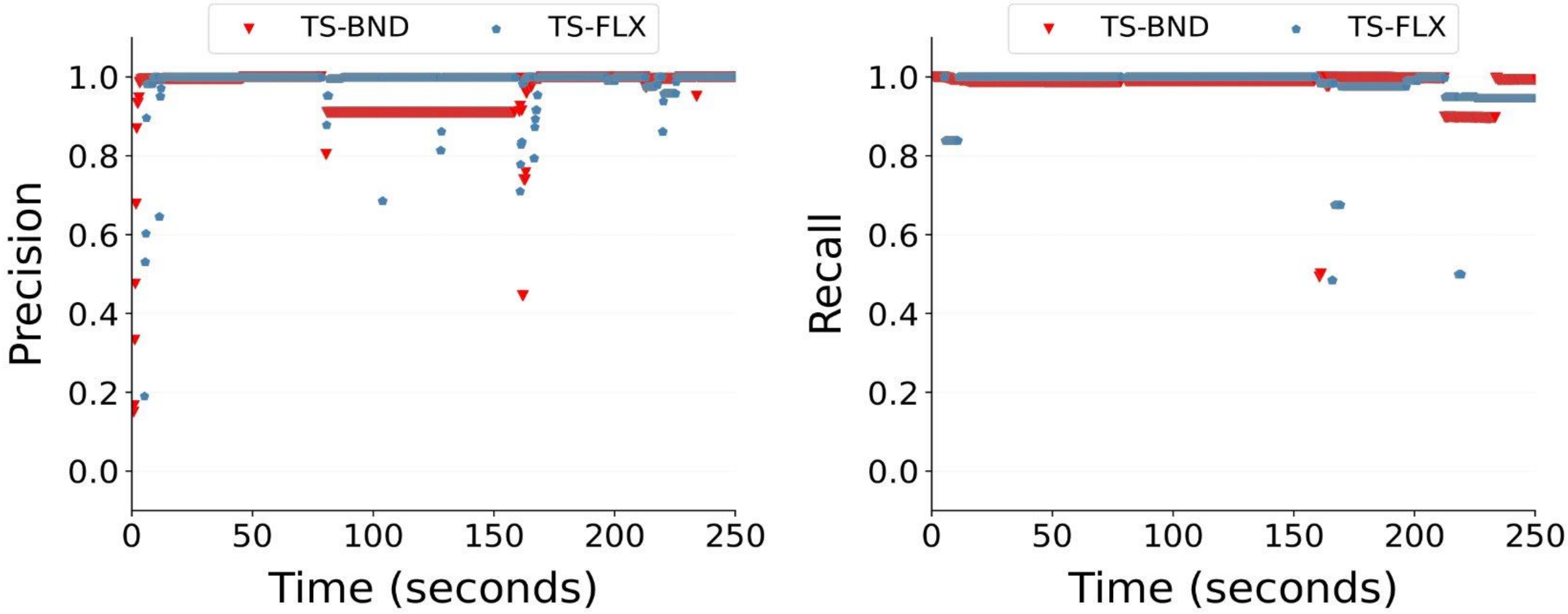




Results - PEBS



Results - Telescope

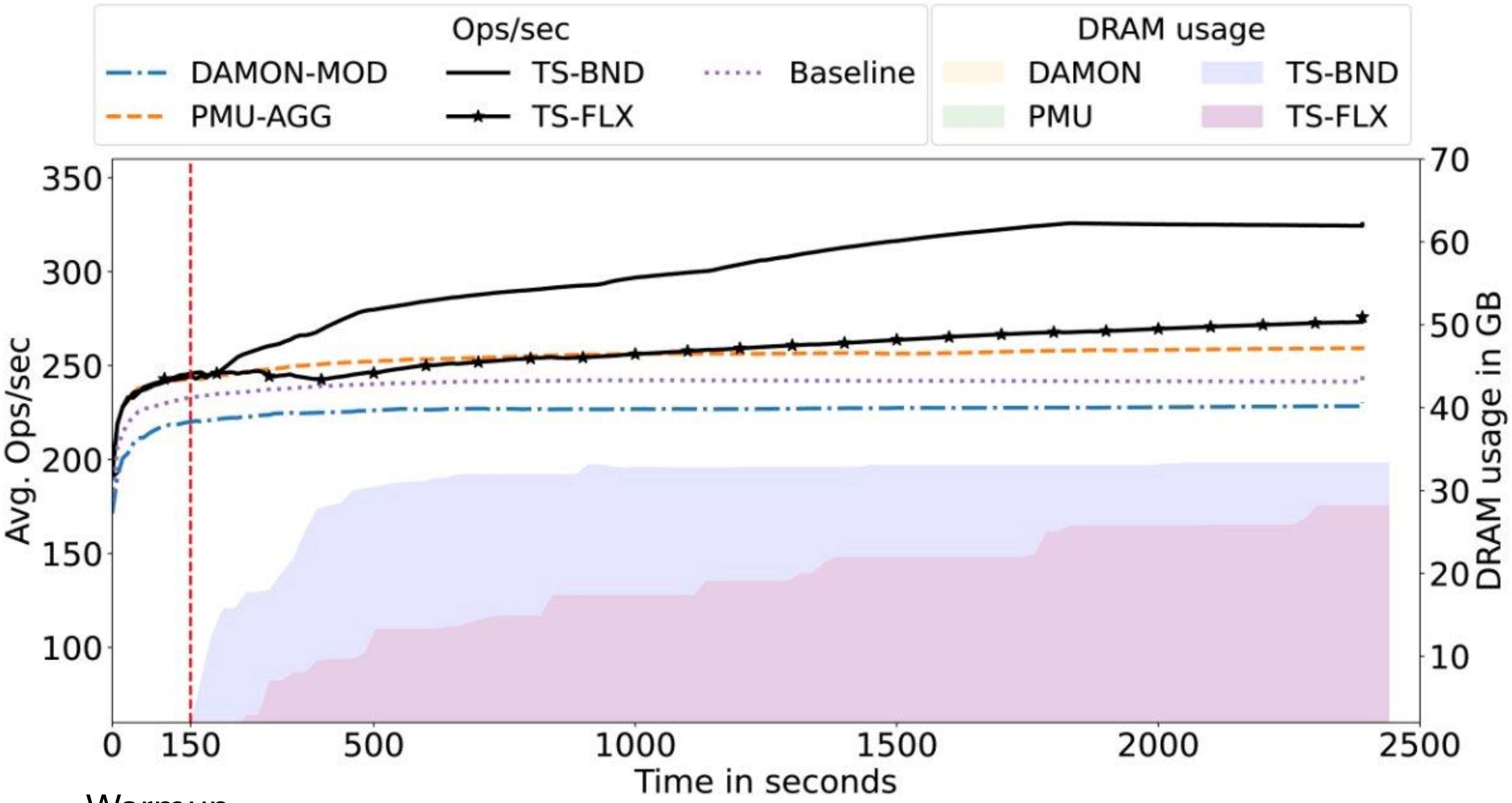


Real World Benchmark - Data Tiering

- Redis with YCSB for Load Generation
- 2TB data initialized in Optane NVM (cold tier)
- Telemetry outputs list of hot regions
- We migrate hottest regions to DRAM (hot tier)

- Metrics
  - Redis Throughput (ops/sec)
  - Tail Request Latency (95p)
  - DRAM Usage (GB)

Real World Benchmark - Results



Warmup
150s

Real World Benchmark - Application Impact

	Config.	95th percentile latency (ms)	
Redis	DAMON-MOD	850	59.13
	PMU-AGG	757	57.50
	Telescope-BND	696	54.01
	Telescope-FLX	741	55.55

Conclusion

- Effectiveness of terabyte-scale tiered memory systems depends on precise and timely identification of hot/cold data.
- Telescope introduces a novel page table profiling technique to quickly converge upon memory access patterns for workloads with huge memory footprints.
- We evaluate Telescope and compare it with other State-Of-The-Art telemetry techniques, and demonstrate its benefits for various applications with large memory footprints.
- Telescope future-proofs memory access telemetry for tiered memory systems with memory capacity up to and beyond the terabyte scale.

We are in the process of
upstreaming Telescope into the Linux Kernel
!!!
Scan QR code to see activity on
lore.kernel.org

THANK YOU!

Backup Slides

Region-Based Sampling

Successive Merge-and-Split ensures convergence to real access pattern over time.

Example: DAMON