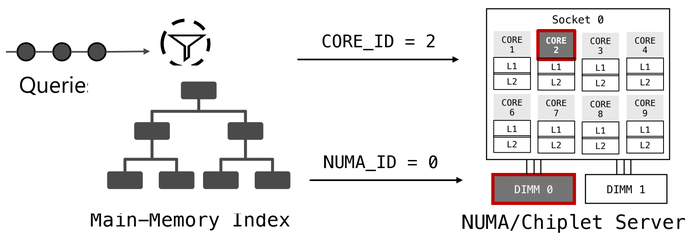
P-MOSS: Scheduling Main-Memory Indexes Over NUMA Servers Using Next Token Prediction

Yeasir Rayhan, Walid G. Aref

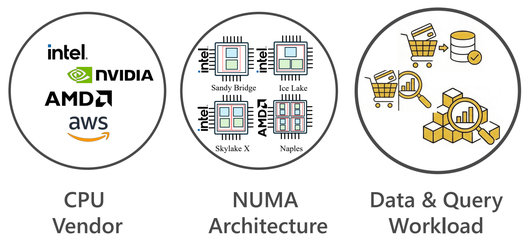
[PDF]

PURDUE UNIVERSITY

P-MOSS ≡ A Learned Performance MOnitoring Unit (PMU) driven Spatial Query Scheduling Framework



Spatial Scheduling for a Main-Memory Index



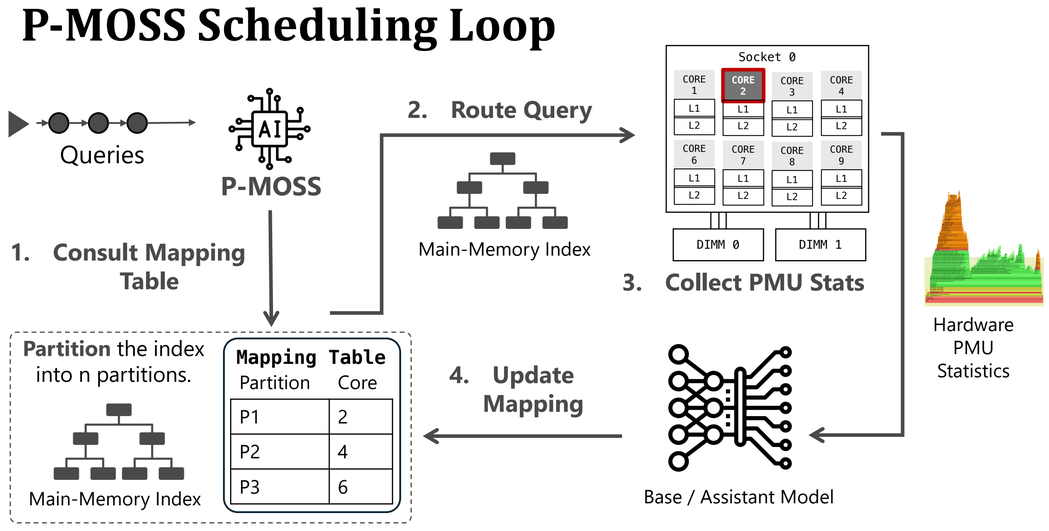
Seen and Unseen

1. Core Scheduling Strategy
   Selects the CORE_ID that executes a query for a given index partition.
2. Data Partitioning Strategy
   Assigns each index partition to a NUMA_NODE (its memory location).

Core Learning Principles of P-MOSS

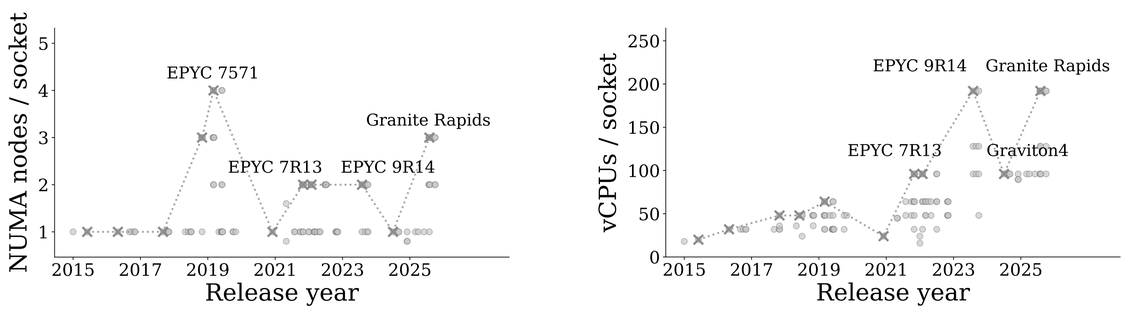
The LLM Recipe	Profile-Guided Optimization	Offline RL
Next Token prediction, GPT backbone, 2-stage training.	HW performance statistics from HW PMU guide scheduling.	Decouples training from the DMBS critical path.

P-MOSS Scheduling Loop



Motivation: Hardware Trends Are Shifting!

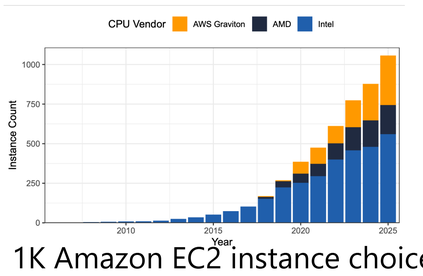
1. Chiplet Era
   4× more NUMA domains per socket.
2. Many-Core Era
   8× more vCPUs per socket.



Amazon EC2 Hardware Landscape

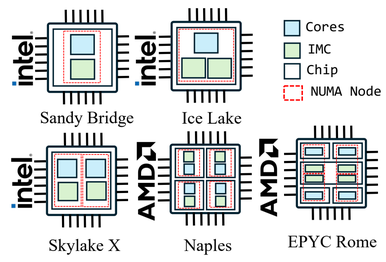
Challenges

1. Diverse CPU vendors with an exponential number of instance choices.

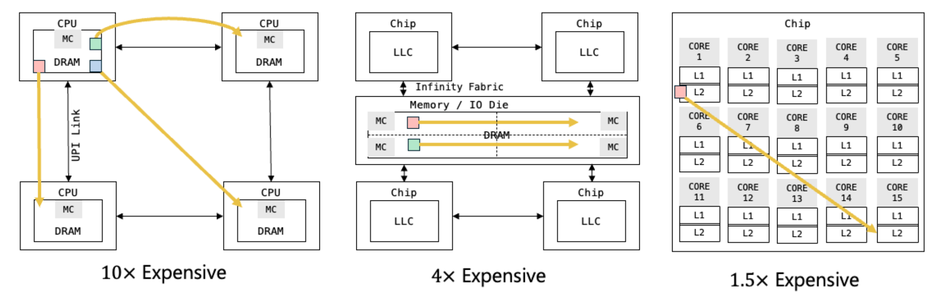


1K Amazon EC2 instance choices in 2025.

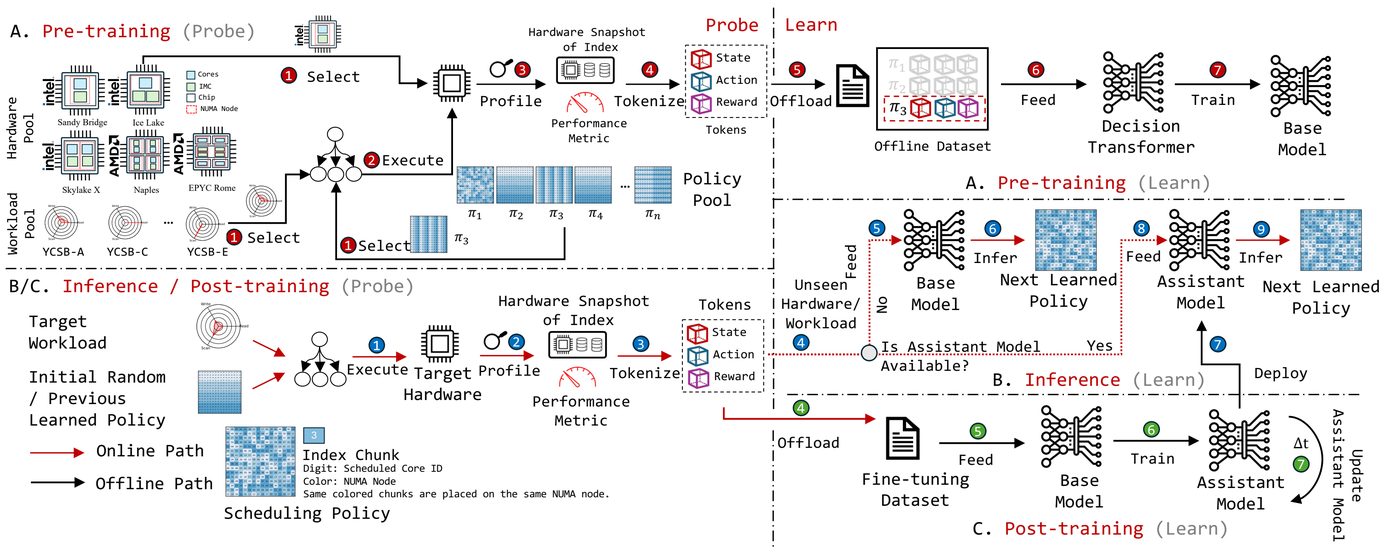
2. Diverse CPU Architecture with different CPU topology



3. Hardware heterogeneity



The LLM Recipe: Next Token Prediction (NTP) & Multi-Stage Training Pipeline



Next Token Prediction. Formulate spatial scheduling as an NTP problem.

- Predict the CORE_ID of the next slice
- Data Partitioning is implicit.



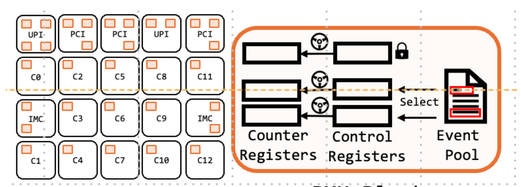
CORE_ID *of the next index partition* = ?

Multi-Stage Training Pipeline.

- **Pre-Training.** Train on 8K distinct scheduling policies across diverse HW and query workload.
- **Post-Training.** Train on 200-300 scheduling policies from the target HW and query workload.

Profile-Guided Optimization (PGO)

- Use execution profiles collected from prior runs to optimize future execution.
- Provides the abstraction across seen and unseen hardware and workload.



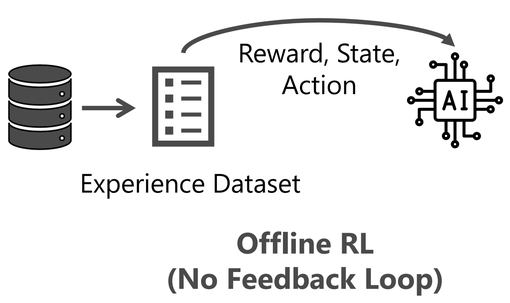
Each hardware component contains PMU blocks with 1 - 8 counters.

P-MOSS profiles 39 HW PMU counters from the HW PMU.

- The number of instructions executed
- L1-I (L1-Instruction) cache misses
- L1D (L1-Data) cache misses
- LLC (Last Level Cache) cache misses
- …

Offline Reinforcement Learning

- The agent learns a policy from an experience dataset, without any active interaction with the database.
- Decouples training from the DBMS critical path.
- Enables scalable training over large datasets.

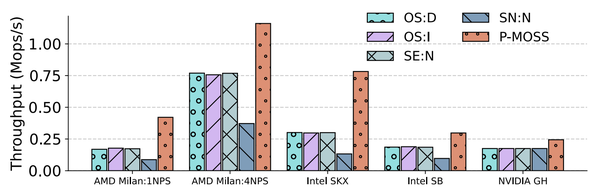


Offline RL (No Feedback Loop)

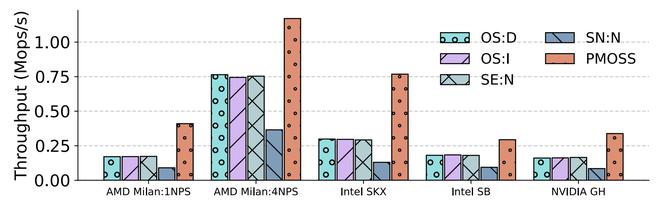
Questions?

Experiments

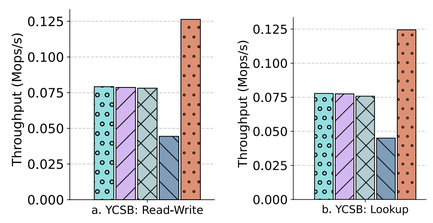
- YCSB A: 50% Read 50% Write
  1.90× speedup (Seen HW)



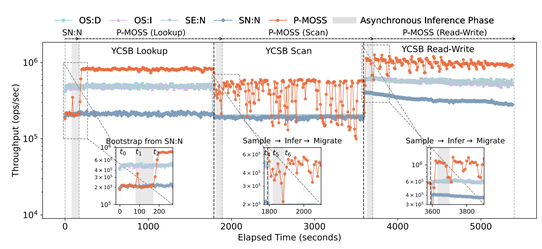
- YCSB C: 100% Lookup
  2.57× speedup (Seen HW)



- YCSB A, YCSB-C
  1.46× speedup (Unseen HW)
  IBM Power System S822LC



- YCSB-C → YCSB-E → YCSB-A
  1.72× speedup
  Intel Skylake X



PMOSS's Learned Scheduling Policies

- YCSB A: 50% Read 50% Write

