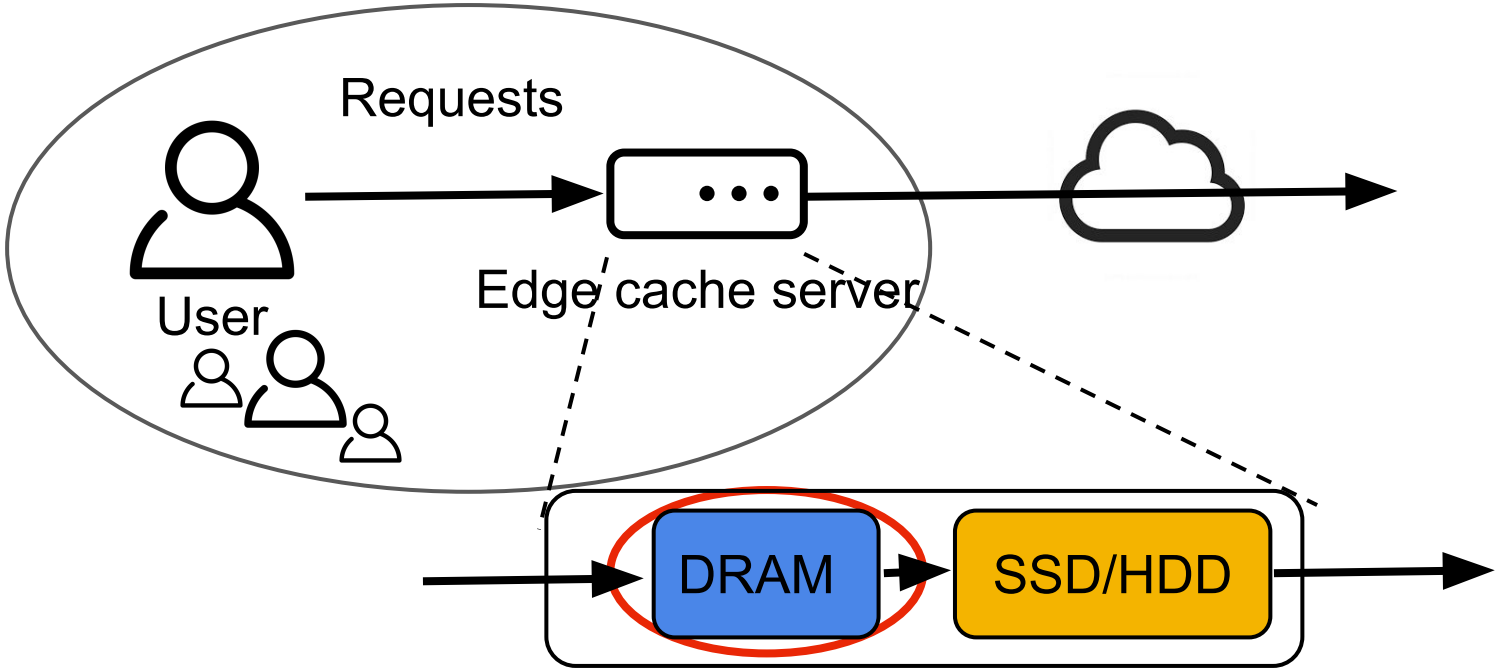
HALP: Heuristic Aided Learned Preference Eviction Policy for YouTube Content Delivery Network

Zhenyu Song, Kevin Chen, Nikhil Sarda, Deniz Altınbüken, Eugene Brevdo, Jimmy Coleman, Xiao Ju, Pawel Jurczyk, Richard Schooler, Ramki Gummadi

Google Research YouTube

CDN Cache Levels: **DRAM**, SSD, HDD, Origin



Metric: P95 DRAM byte miss ratio

Three Challenges for Learned Eviction Algorithm Deployment

Many recently learned eviction algorithms beat heuristic.

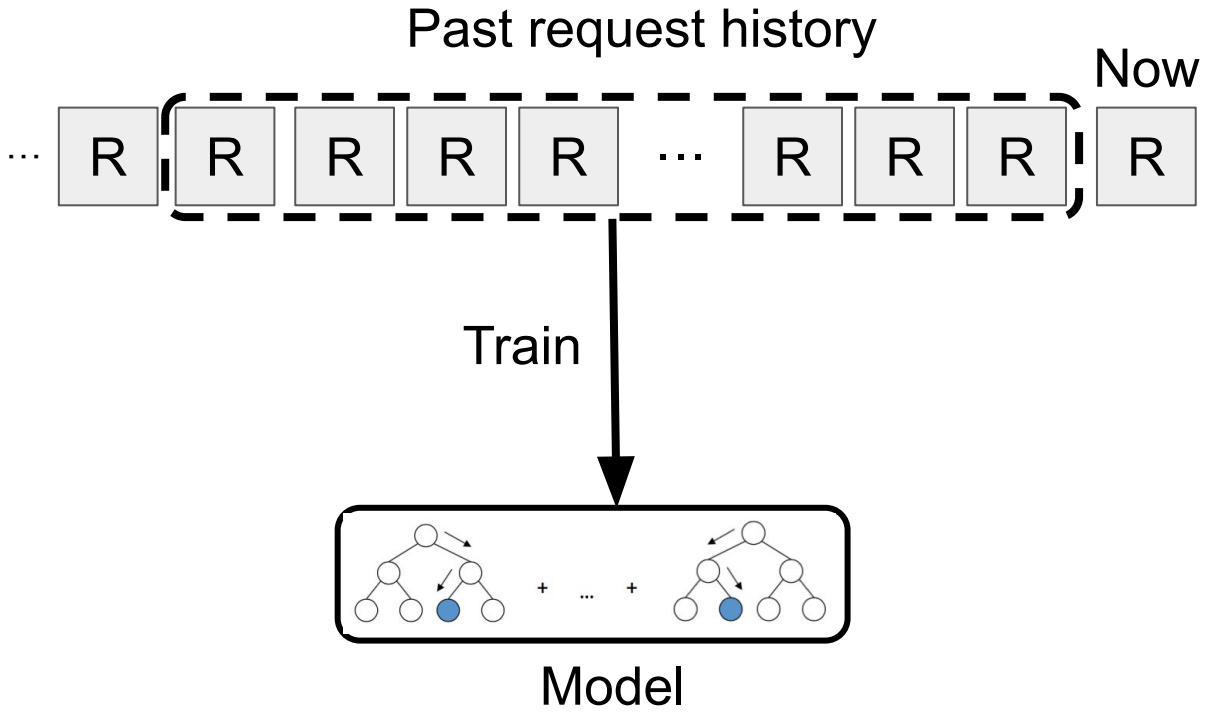
- LearningDistributedTraces (Zhou & Maas, 2021), CACHEUS (Rodriguez et al., 2021), LRB (Song et al., 2020).

Challenge 1: ML computation overhead

Challenge 2: reducing avg BMR w.o making any location worse

Challenge 3: measuring new alg impact under production noise

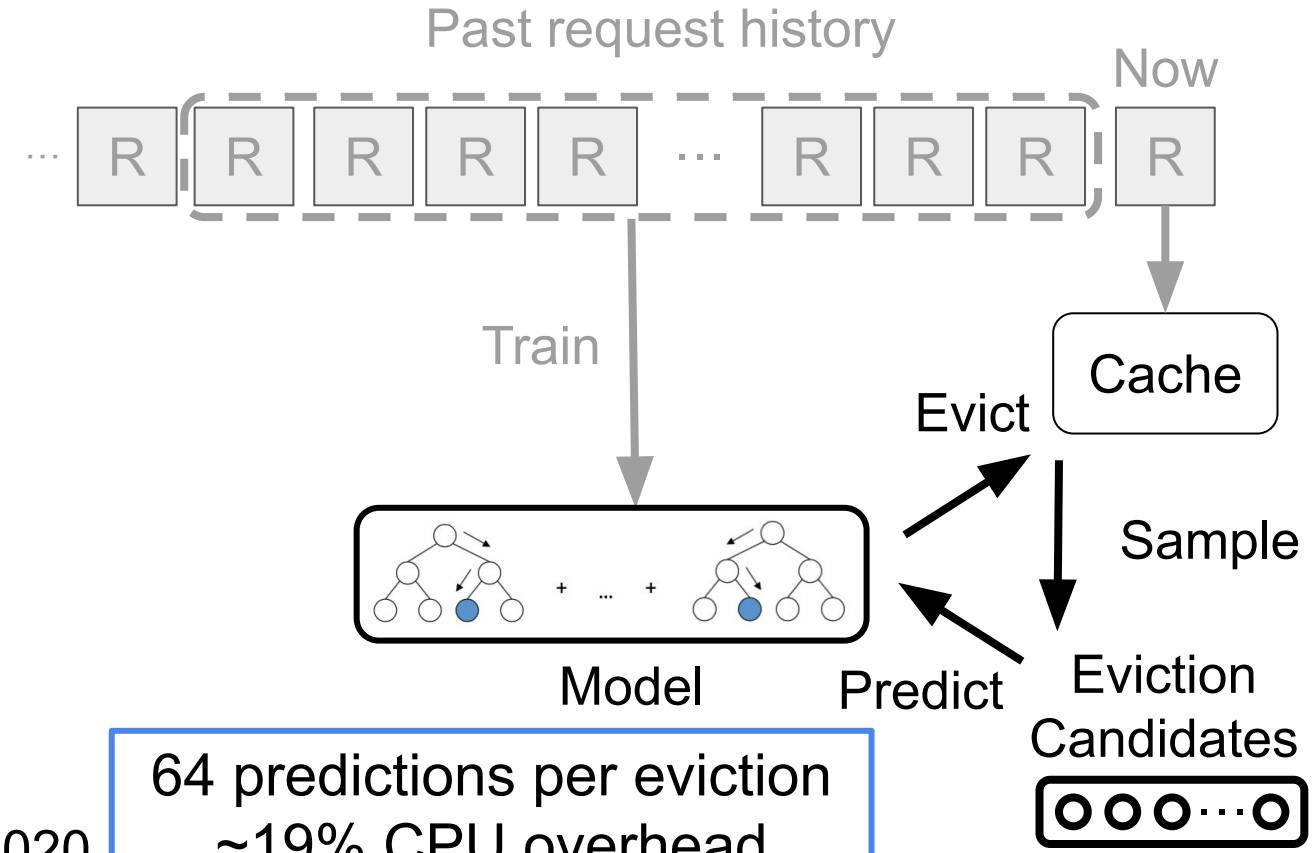
Challenge 1: ML Computation Overhead



Song et al., 2020

Need to learn all objects in the past window

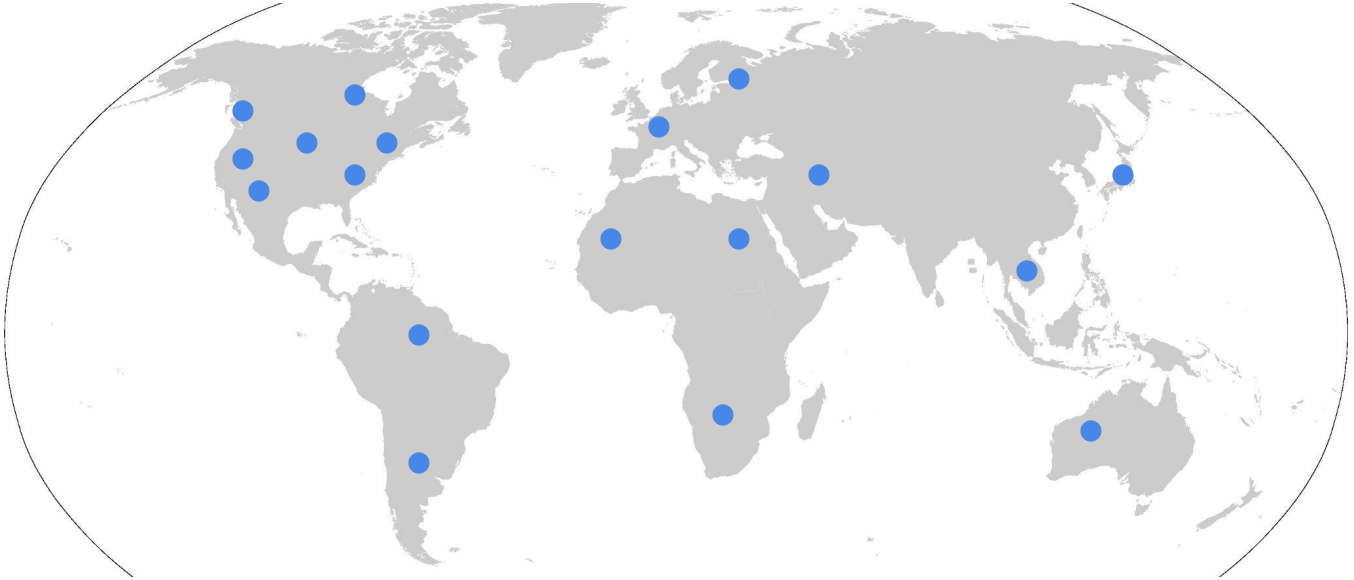
Challenge 1: ML Computation Overhead



Song et al., 2020

64 predictions per eviction
~19% CPU overhead

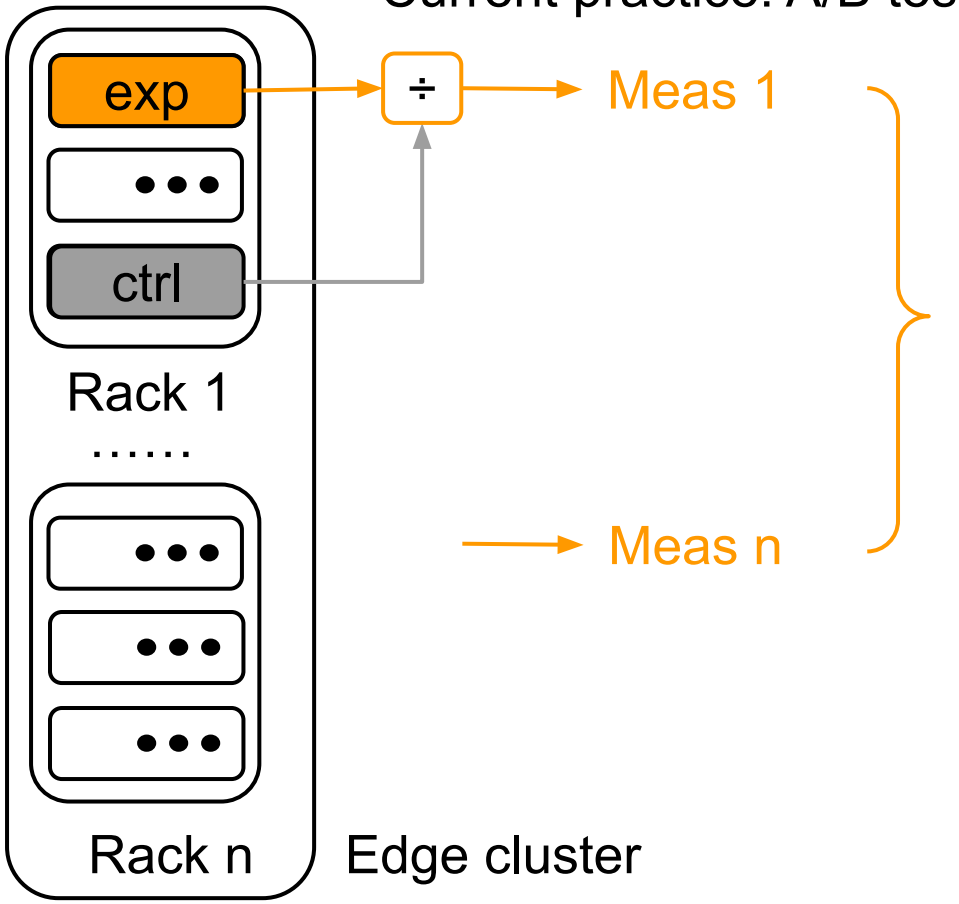
Challenge 2: Reducing Avg BMR W.O Making Any Location Worse

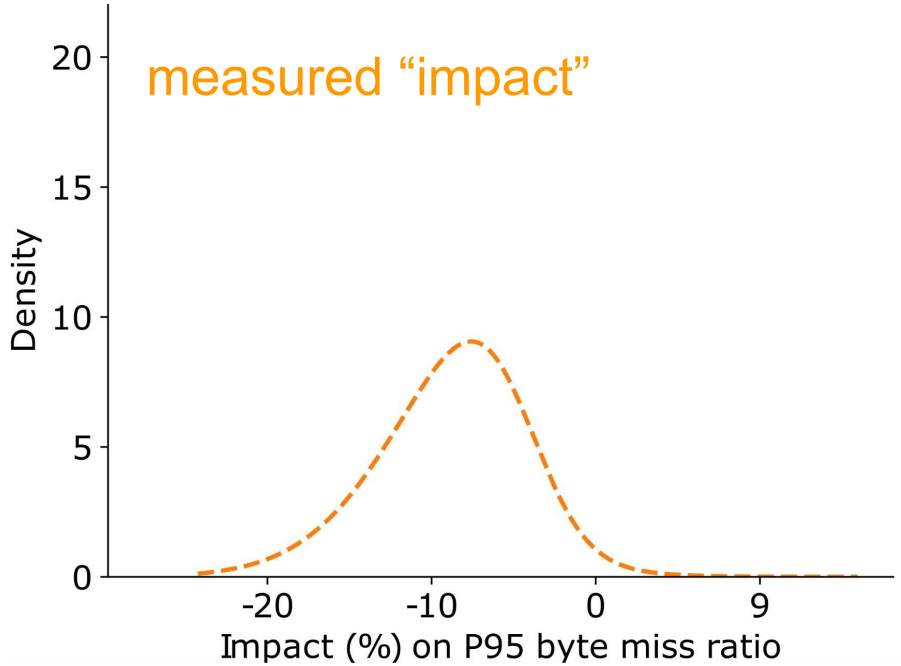


Regressions in a few locations could degrade user experience

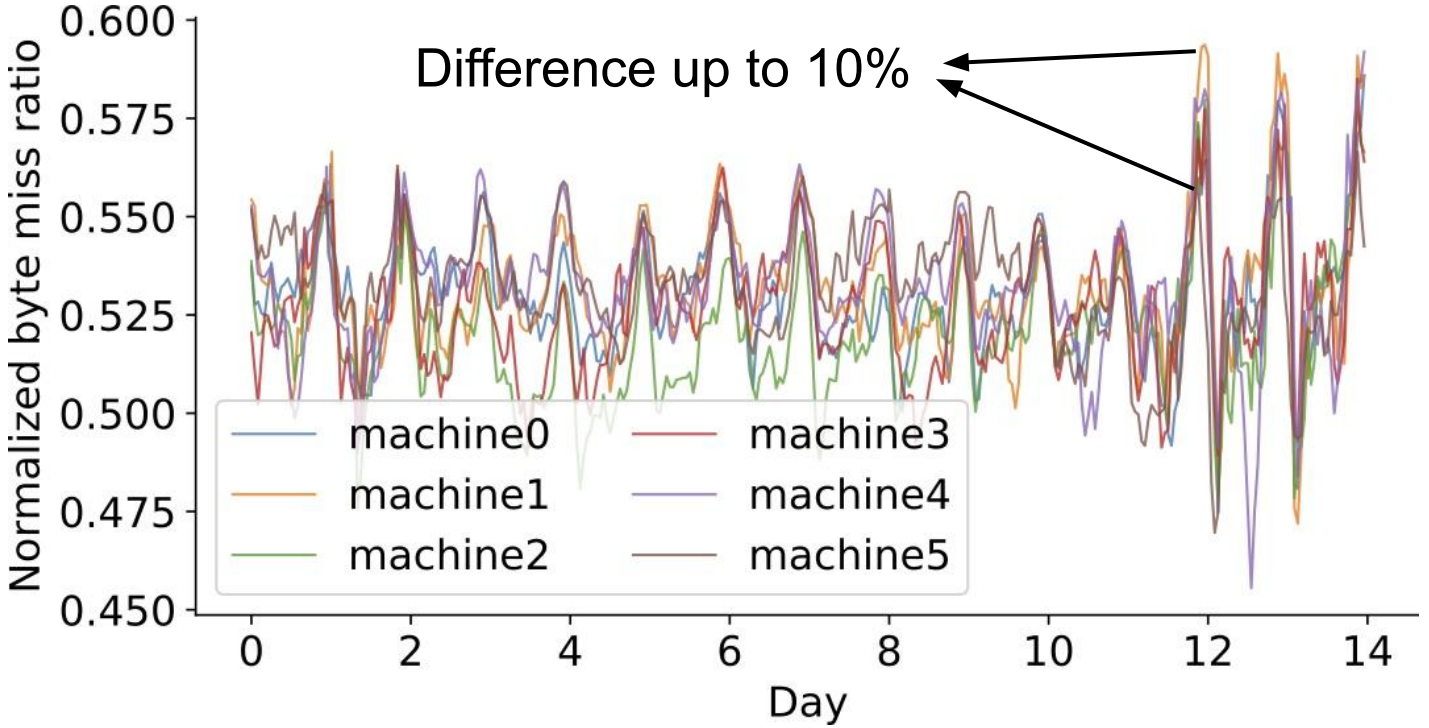
Challenge 3: Measuring New Alg Impact Under Production Noise

Current practice: A/B test

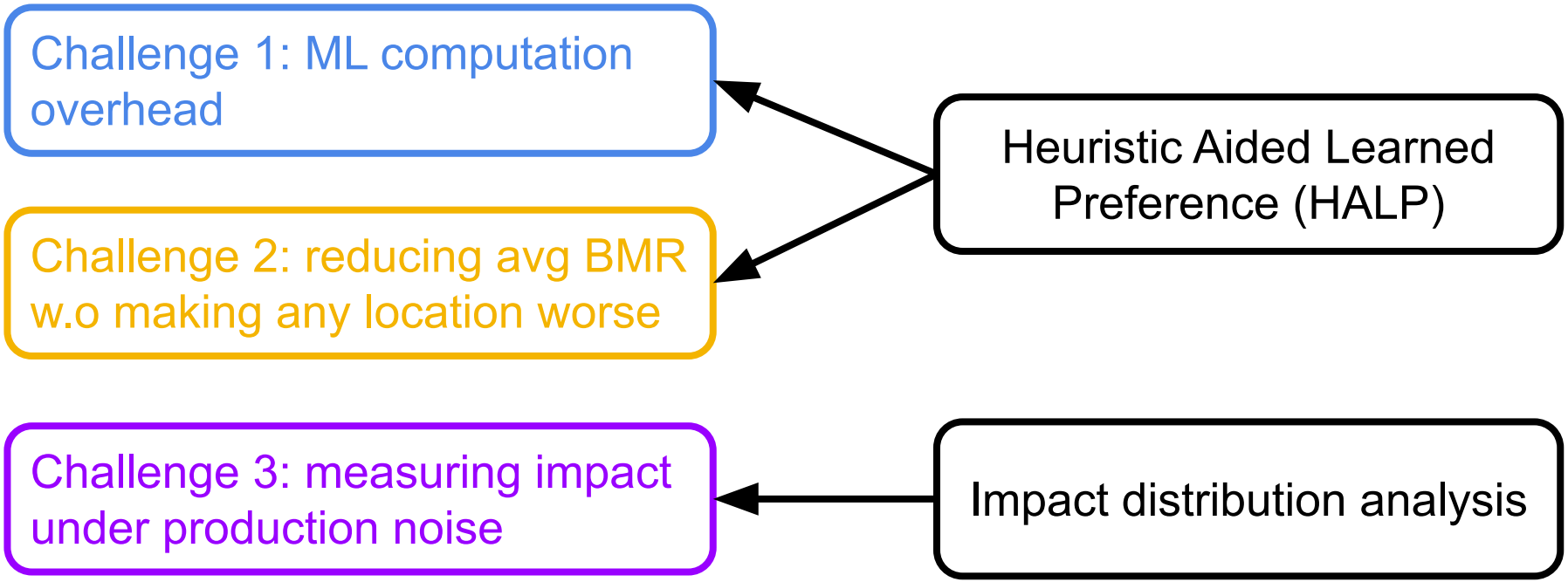




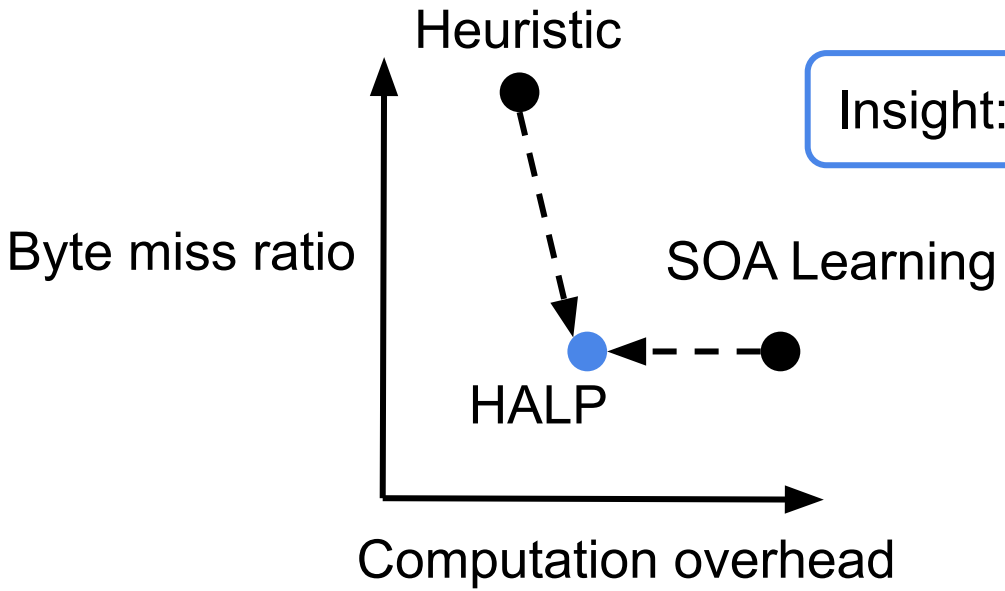
Challenge 3: Measuring New Alg Impact Under Production Noise



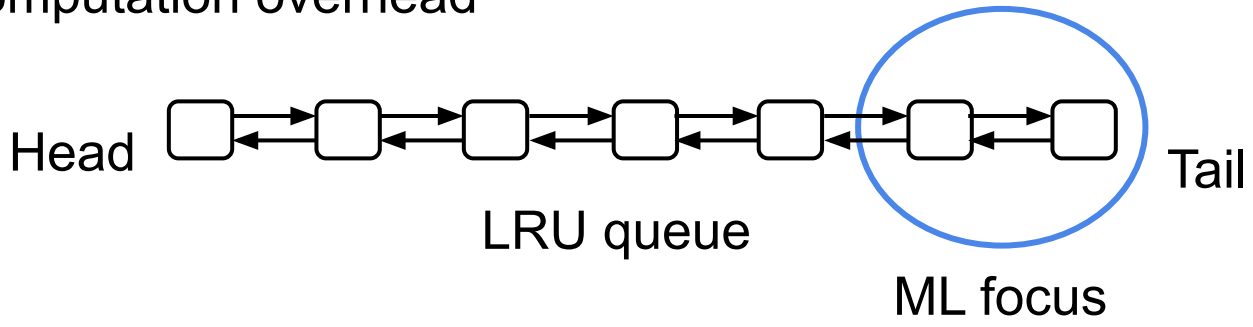
Solutions


Challenge 1: ML computation overhead
Challenge 2: reducing avg BMR w.o making any location worse
Challenge 3: measuring impact under production noise
Heuristic Aided Learned Preference (HALP)
Impact distribution analysis

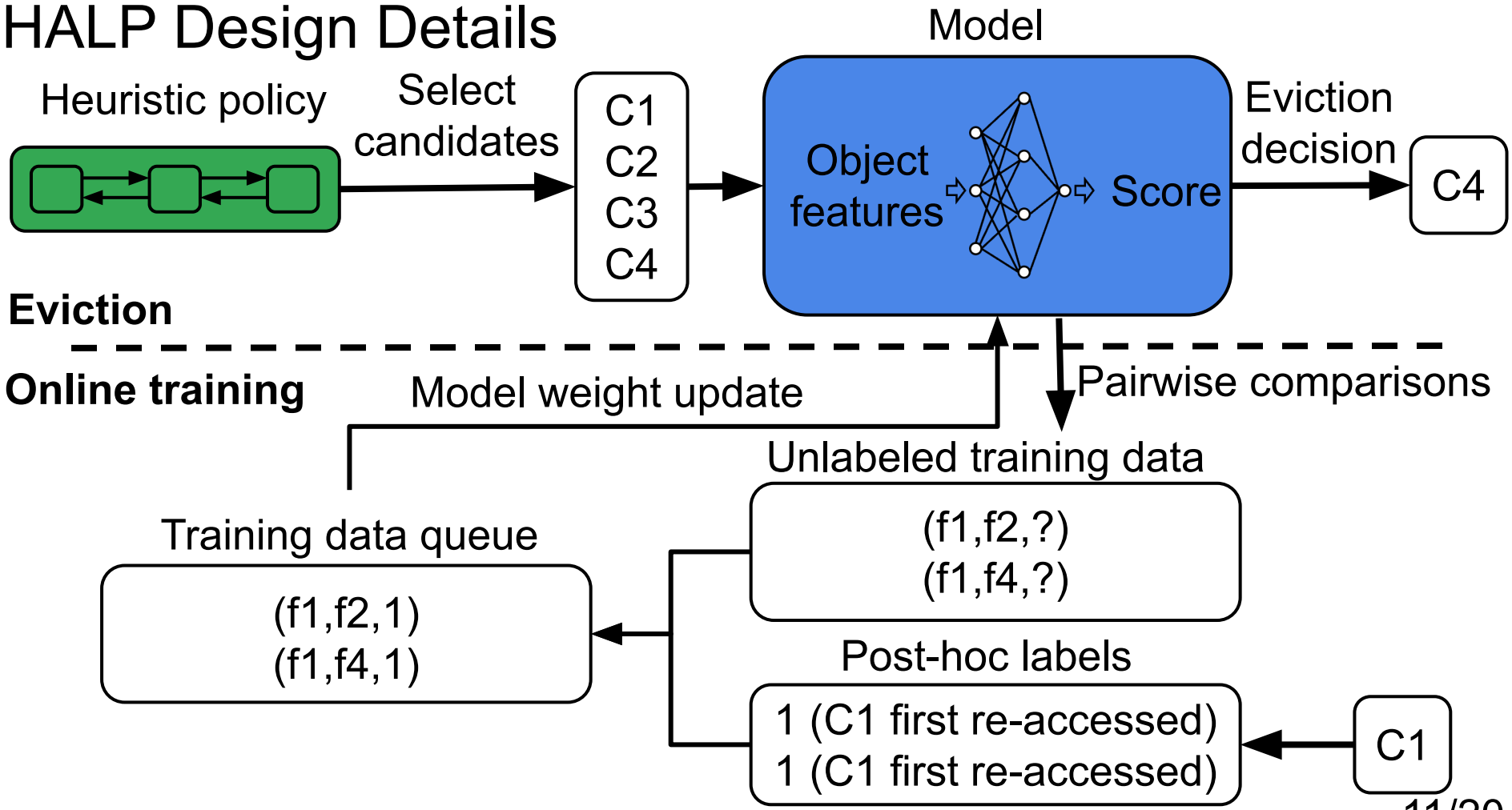
Heuristic Aided Learned Preference (HALP)



Insight: heuristic + preference learning



HALP Design Details



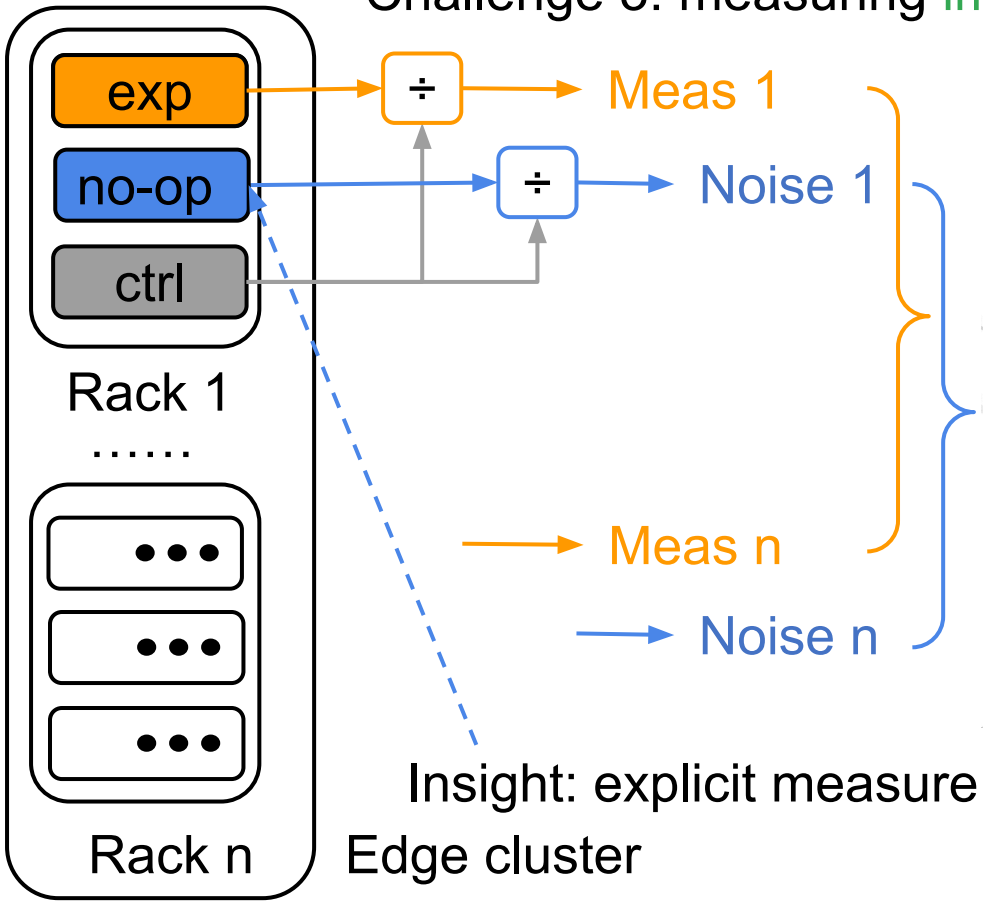
Model & Features

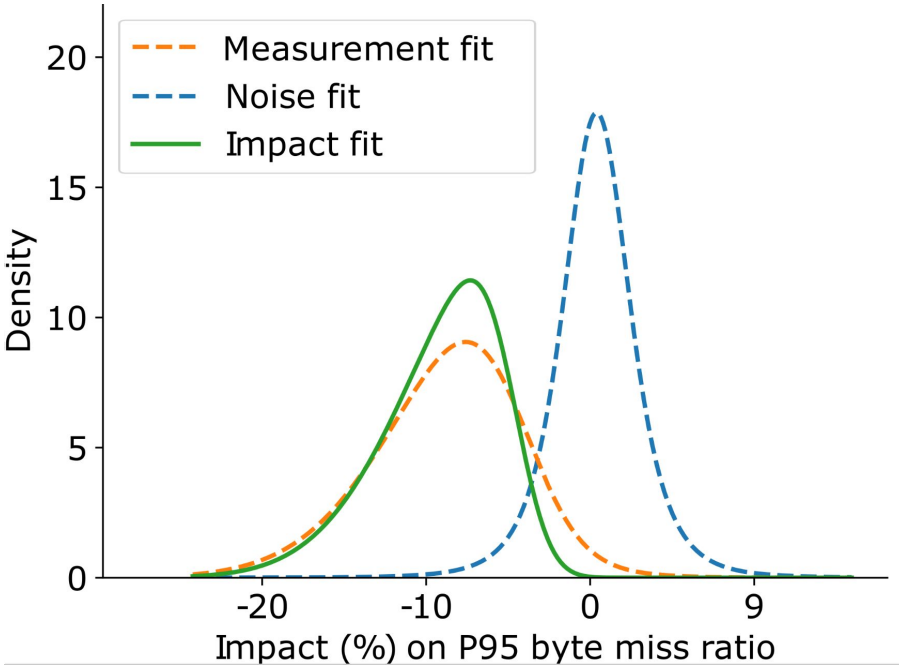
- Model: two-layer MLP.
- A pairwise prediction is 720 ns, and each training is several ms.
- Loss: cross entropy.

Feature name	Dimension
Access-based	
Time between accesses	32
Exponential decay counters	10
Number of accesses	1
Average time between accesses	1
Time since last access	1
Video-specific	
End of chunk	1

Recover Impact Distribution from Measurement and Noise

Challenge 3: measuring impact under production noise



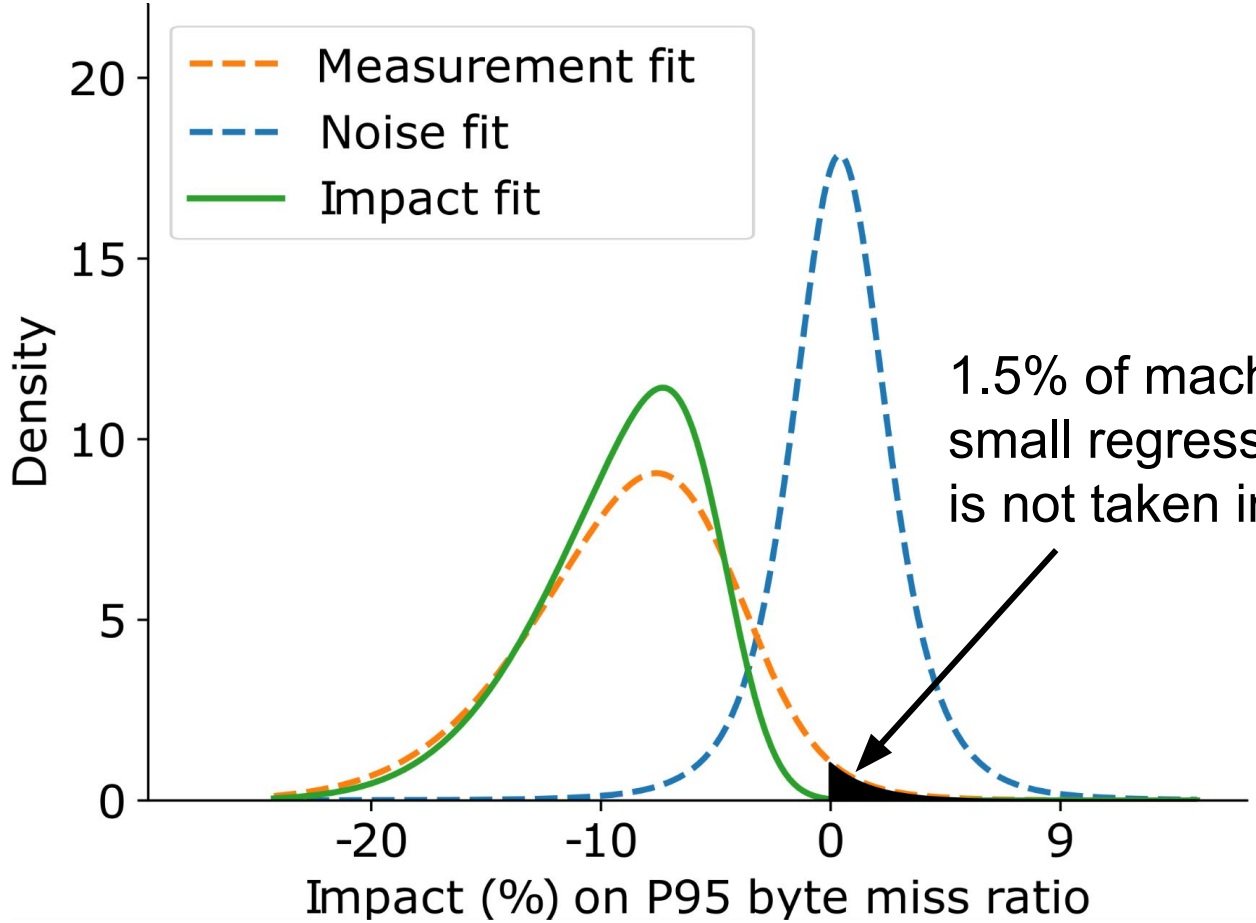


Insight: explicit measure noise distribution by a no-op group

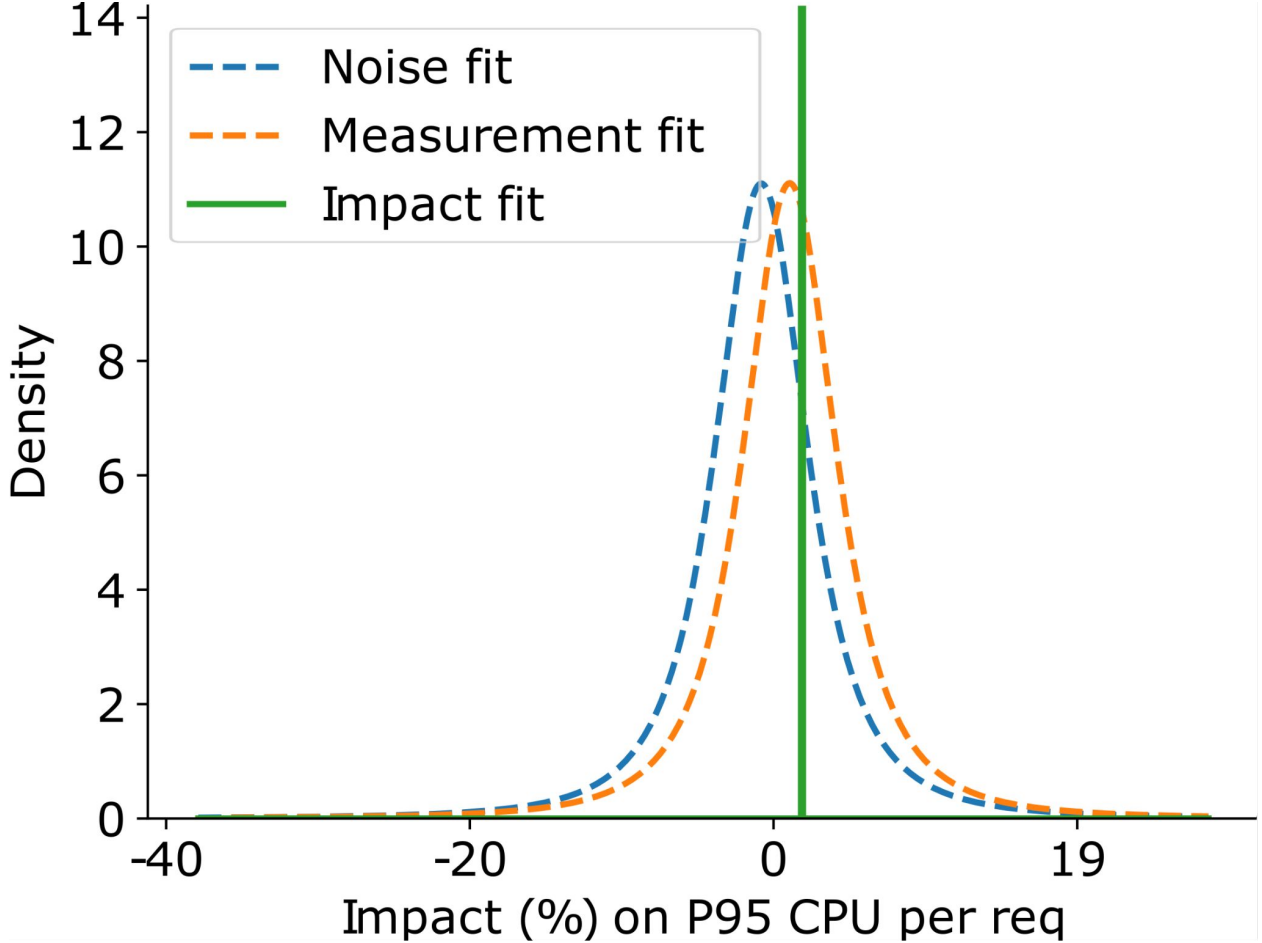
Evaluation Setup

- Implementation based on Google's SmartChoices ML service.
- Q1: Can HALP reduce the byte miss ratio without causing regression?
- Q2: What is the computation overhead of HALP?
- Q3: How does HALP compare with SOA cache algorithms?

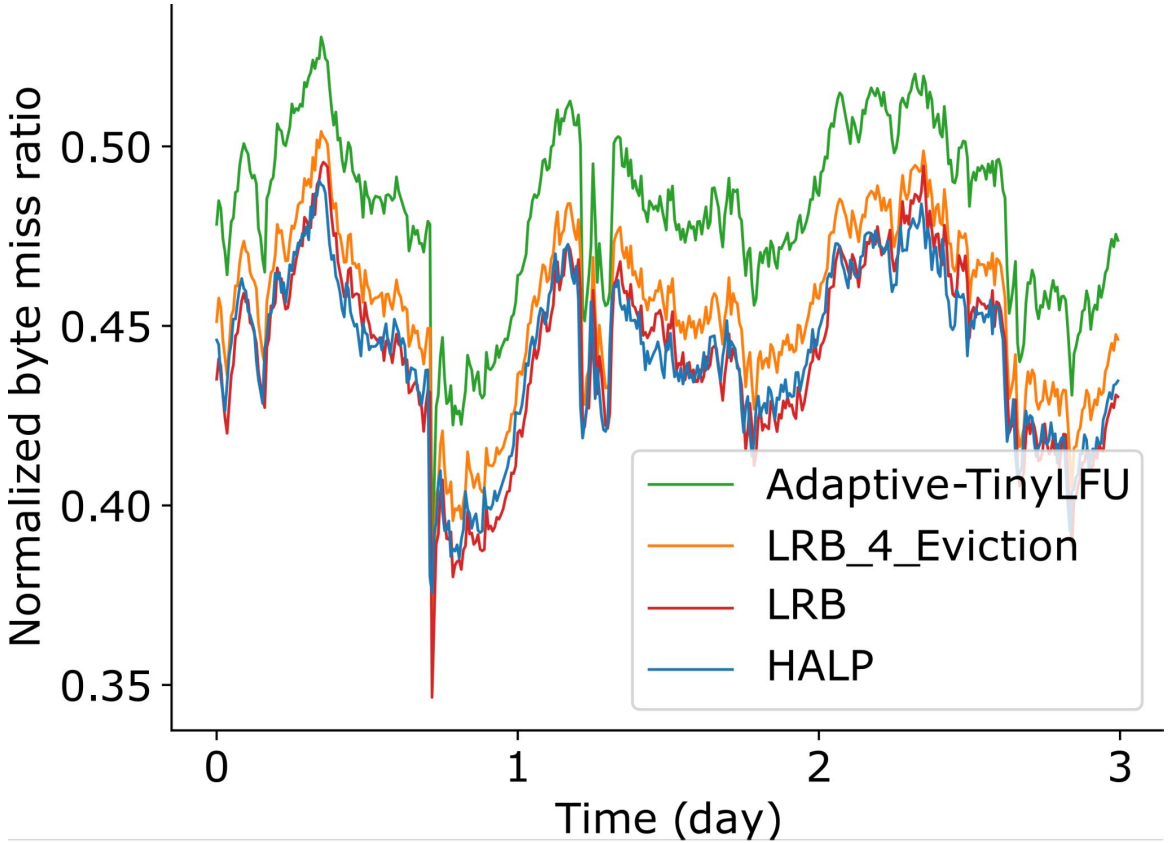
HALP Robustly Improves P95 BMR by 9.1% With Negligible Regression



HALP Has a Modest CPU Overhead of 1.8%



HALP Has the Best BMR/CPU Overhead Combination over SOA



Developed market trace simulation

Conclusion

- 9.1% P95 byte miss ratio reduction without making any location becomes noticeably worse.
- Insight: heuristic + preference learning.
- Deployed in production since early 2022.

Google Research YouTube