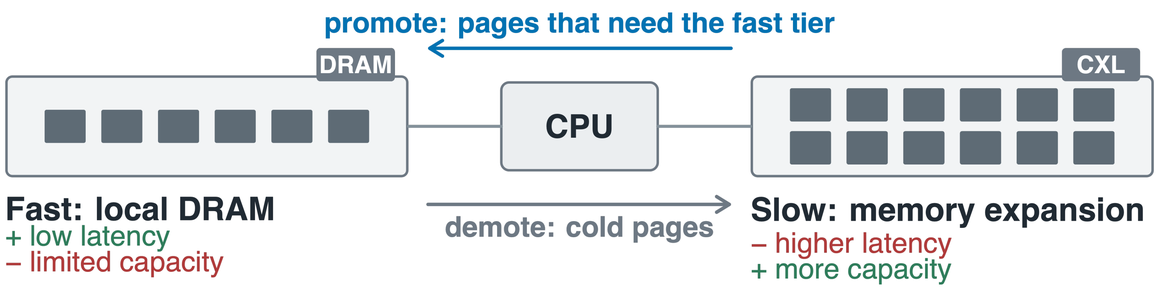
Tierce: Observability-Driven Tiered Memory Management for Colocated Workloads

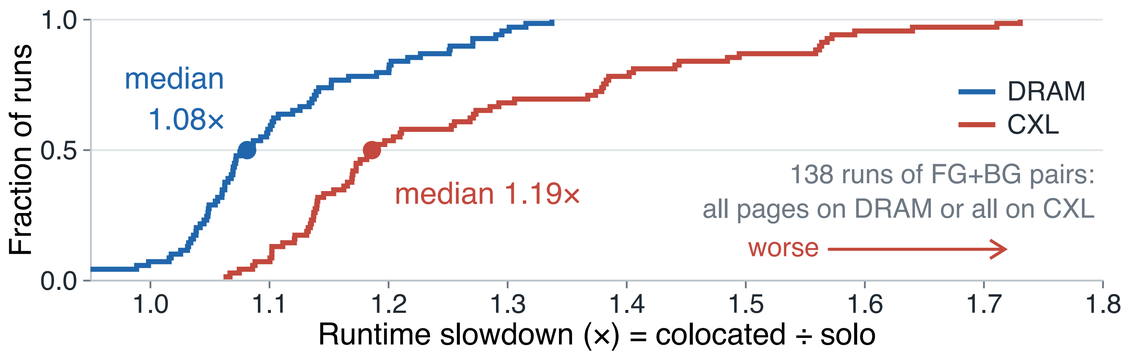
Hanchen Xu*, Berkay Inceisci*, Hao Li, Zhenyu Zhang, Huaicheng Li

CXL Amplifies Colocation Slowdown

- Memory is scarce and expensive; CXL expands capacity beyond local DRAM, at much higher latency.
- The question: which pages stay in fast memory?



- Colocation: an operator-designated foreground (FG) shares the LLC and memory with best-effort background (BG) co-runners.
- Colocation on CXL slows workloads over 2× as much as on local DRAM: median 19% vs. 8%, worst case 73% vs. 34%.



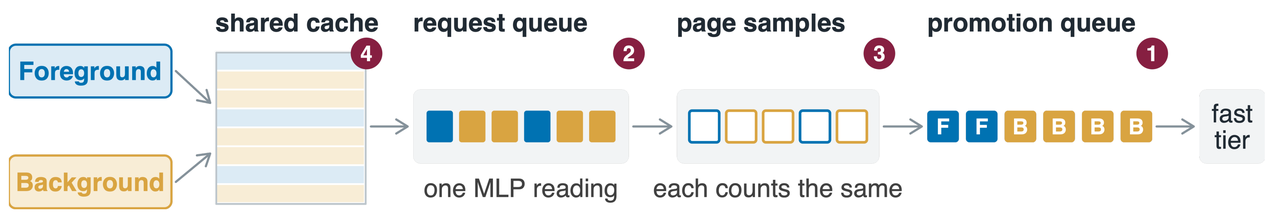
Colocation Aliases Tiering Signals

- Existing systems (NBT, Memtis, Colloid, PACT) rank pages from aggregate observations.
- Under colocation, shared cache and memory interference distort these observations.

Page Access Criticality (PAC), from PACT

$$\text{P Score} \propto \frac{\text{Accesses} \times \text{Fixed cost}}{\text{MLP}}$$

Estimates the stall fast memory could save: each miss's cost discounted by MLP, not access frequency alone.

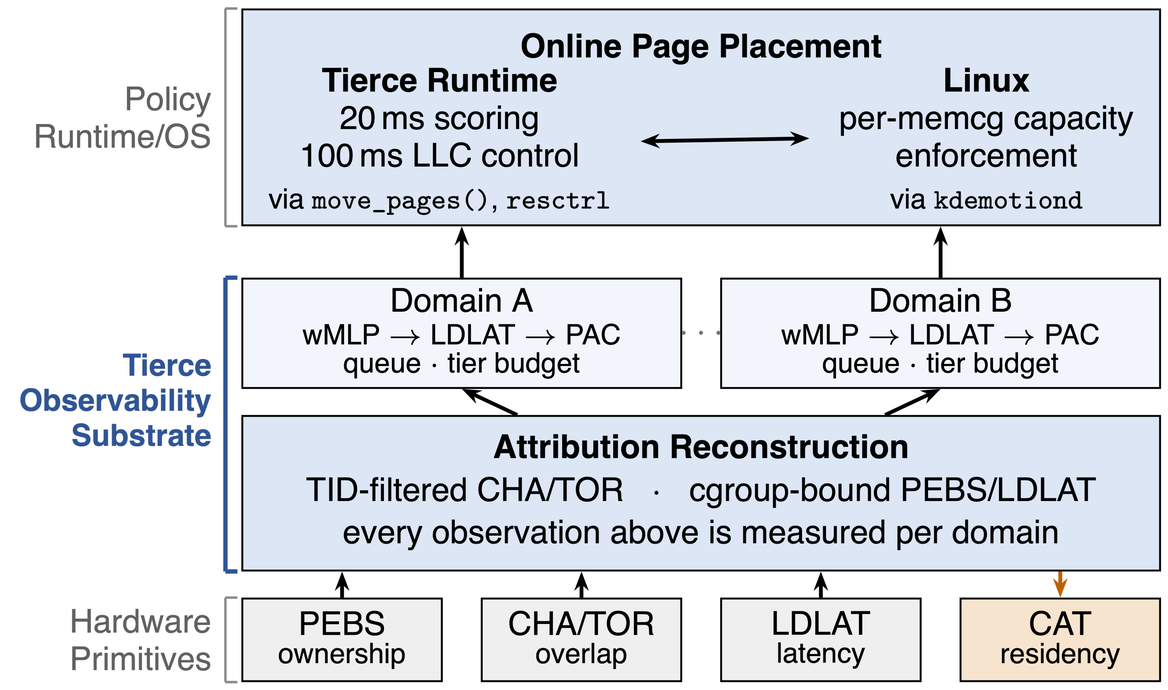


Gap	Failures of existing systems under colocation
① Shared promotions	one global queue: BG takes up to 95% of promotions
② Mixed MLP	aggregate MLP: BG compresses FG's criticality scale
③ Contention	fixed cost per miss: contention delays go unseen
④ Cache interference	no LLC control: BG evictions add FG misses

Colocation turns page placement into an ownership problem: the signals used to estimate criticality blend colocated workloads' behavior.

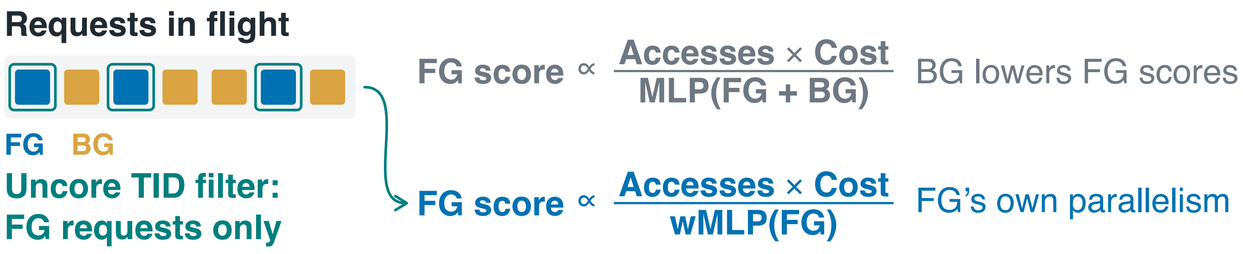
Tierce: An Observability Substrate for Colocated Tiering

Tierce keeps Page Access Criticality (PAC) as the placement abstraction; it makes measurement, attribution, admission, and LLC control workload-local.

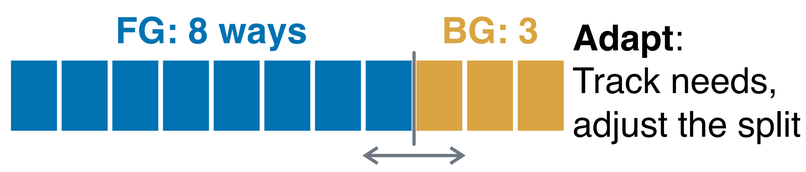


Domain: a workload's cores and memory cgroup.

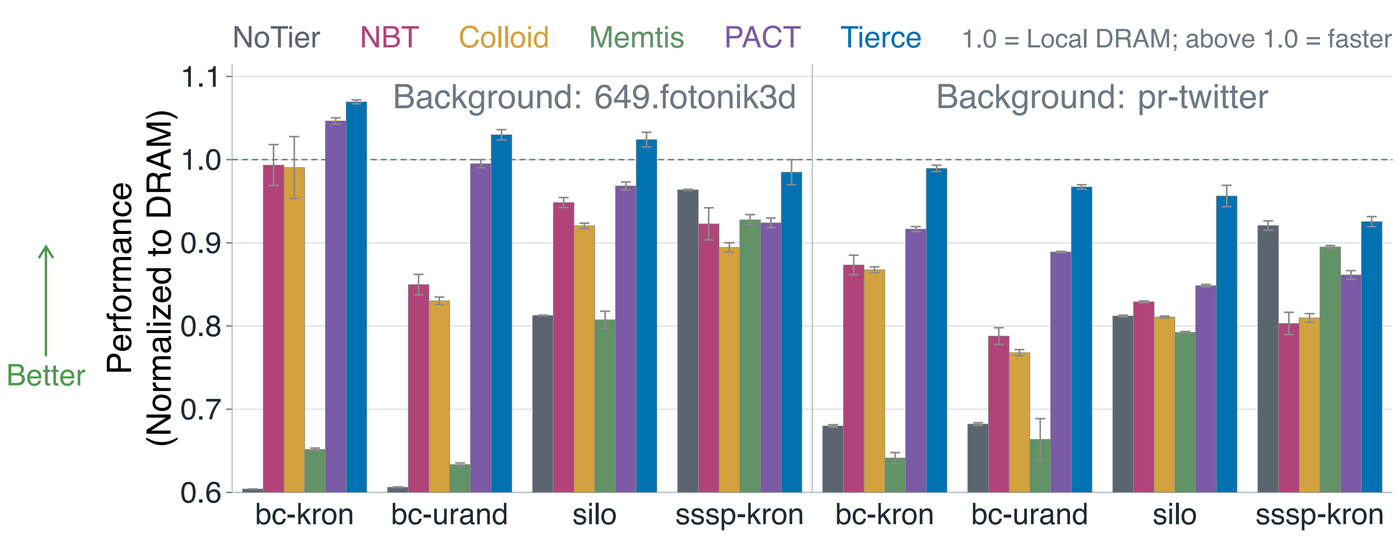
1. **Isolated Placement**: per-domain threshold, queue, and tier budget
2. **Workload-Specific MLP (wMLP)**: measured on the domain's own cores

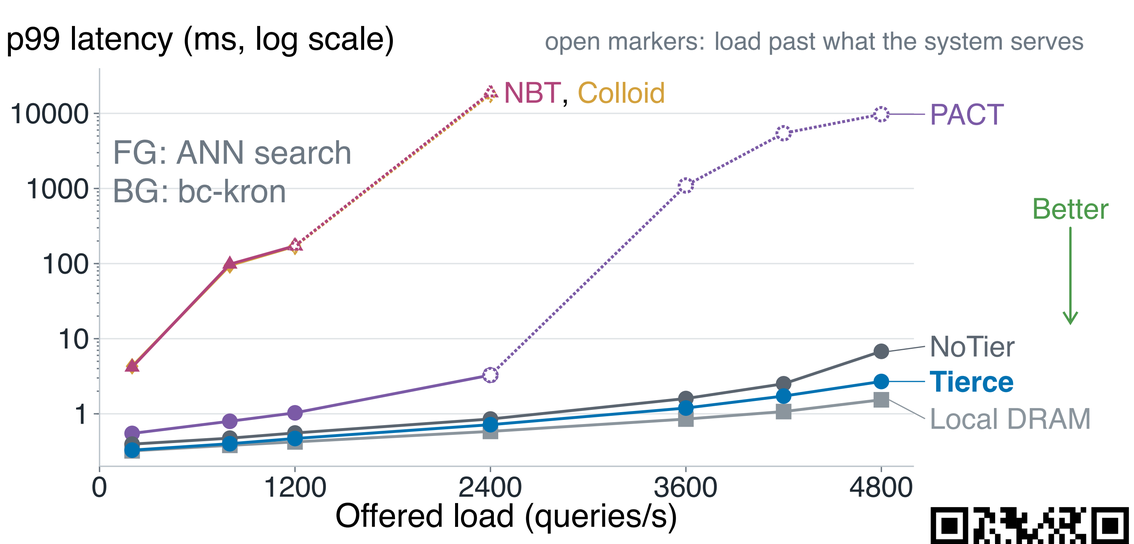


3. **Latency Selection**: admit only samples above a latency threshold
   - uses online per-access latency to attribute stall cost to each access
   - more accurate than a fixed cost per miss, so promotions target costly pages
4. **Predictive LLC Allocation**: bound LLC interference with cache ways



Up to 12.6% Higher Throughput and 60% Lower Tail Latency than Best Prior Art





Preserving workload ownership from observation through placement ties each decision to each workload's own costs, improving both throughput and tail latency under colocation.

https://doi.org/10.1145/3830418.3843905

Appears in the ACM SIGOPS 32nd Symposium on Operating Systems Principles (SOSP '26)