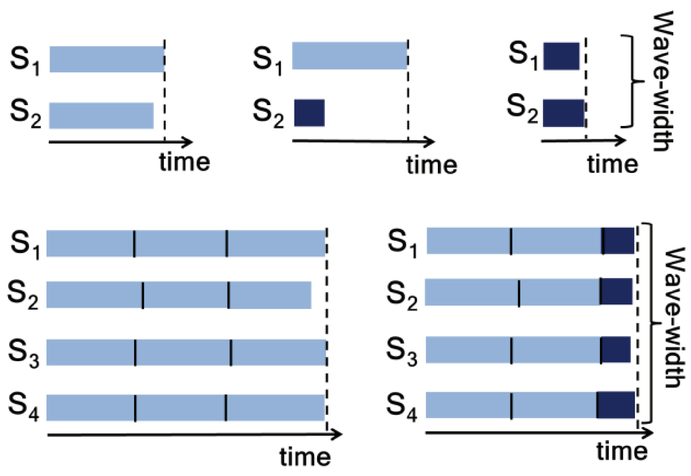
PACMAN: COORDINATED MEMORY CACHING FOR PARALLEL JOBS

Ganesh Ananthanarayanan, Ali Ghodsi, Andrew Wang, Dhruba Borthakur, Srikanth Kandula,
Scott Shenker , Ion Stoica (UC Berkeley)

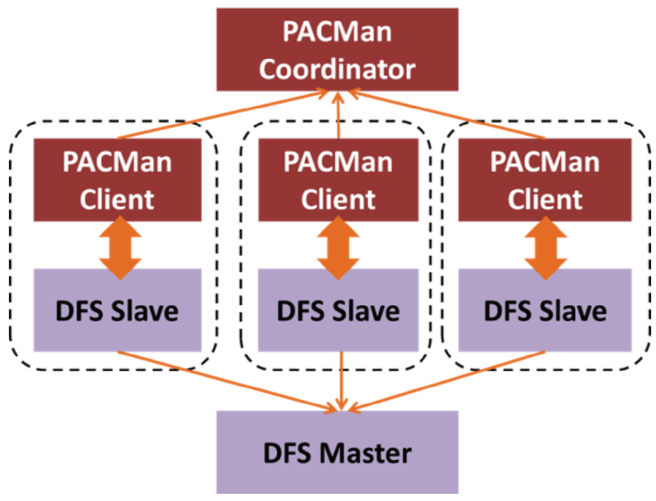
MOTIVATION

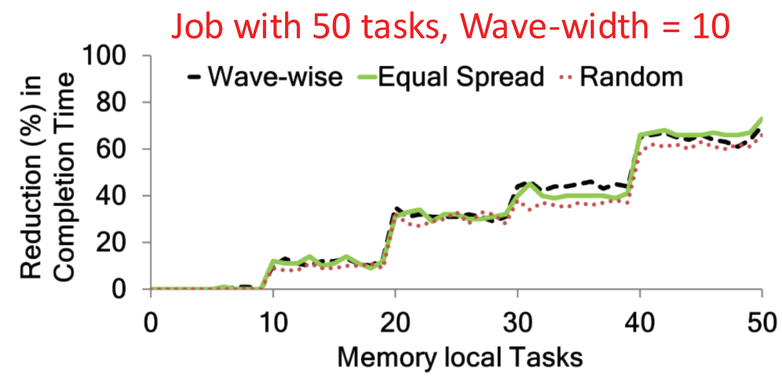
- MAnalytics jobs are parallel and process large amounts of data
- Machines have tens of gigabytes of memory
  - Falling memory prices
  - Median utilization of 19%
- Heavy-tailed Input Sizes
  - Elephant and mice jobs
  - 92% of smallest job inputs can fit in memory

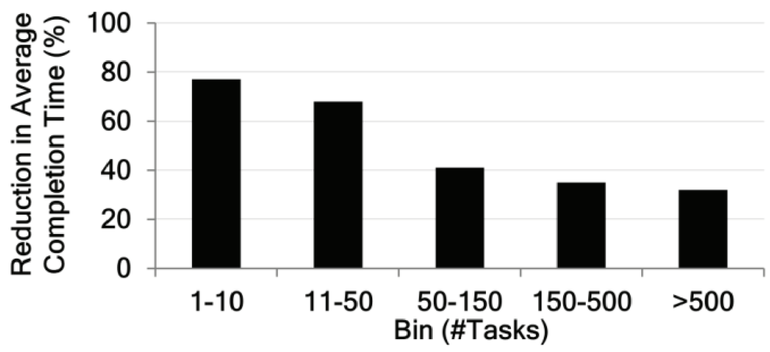


ALL-OR-NOTHING

- Jobs speed up when multiples of its wave-width are cached
  - Wave-width: #parallel executing tasks
  - Small single-waved jobs require 100% memory locality
- Cache hit-ratio insufficient; even MIN speeds up jobs by only 13%
- *Sticky* policy: Focus replacements on incompletely cached waves







COORDINATED CACHING

- Coordinator has global view of cache
  - Eviction and task placement
- Average Completion Time
  - LIFE: Evict from file with highest wave-width
  - Learn wave-width across multiple runs; file size correlates with wave-width
- Cluster Utilization
  - LFU-F: Evict from file with highest frequency
  - Overlap across map and reduce phases → sticky policy is important

EVALUATION

- Replayed Facebook and Bing workloads
- LIFE reduces average completion time by 53% and 51% in Facebook and Bing workloads
  - Small jobs see 77% improvement
- LFU-F improves cluster utilization by 47% and 53% in the Facebook and Bing workloads
- LIFE and LFU-F beat Belady's MIN despite lower cache hit-ratio
- Pre-fetch & Pre-replace → Ideal (87%) speedup
- Pre-replacement ~ Oracle cache eviction

PRE-FETCH AND PRE-REPLACE

- Oracle cache eviction and singly-accessed inputs
- *Preparation* Jobs: Large and multi-waved
- Pre-fetch for later waves of preparation jobs
- Evict inputs after multi-waved job ends
  - If singly-accessed, good!
  - If not, pre-fetch all but first wave
- Pre-replace with files of lowest wave-width

FUTURE WORK

- Proof of optimal cache eviction
- Hierarchical caching to include SSDs
- Details: HotOS 2011, NSDI 2012

Intel Science & Technology Center for Cloud Computing

amplab

Carnegie Mellon University · Georgia Tech · intel · Princeton University · UC Berkeley