Introduction

- **Cache**
  - Fast but relatively small in capacity

- **Machine Learning**
  - Improves decision processes

**Cache management + ML**

Improved performance
Cache Replacement Algorithms

- **Non-adaptive**
  - Least Recently Used (LRU)
  - Least Frequently Used (LFU)
  - Low Inter-reference Recency Set (LIRS) [Jiang et al., ‘02]

- **Adaptive**
  - Adaptive Replacement Cache (ARC) [Modha, Megiddo, ‘02]
  - Dynamic LIRS (DLIRS) [Li, ‘18]

- **ML-based Adaptive**
  - Adaptive Caching Using Multiple Experts (ACME) [Ari et al., ‘02]
  - Reinforcement Learning On Cache Replacement (LeCaR) [Vietri et al., ‘18]
  - Reinforcement Learning (Cacheus) [This Work]
### WORKLOAD PRIMITIVES

- LRU-Friendly
- LFU-Friendly
- Churn
- Scan

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LRU</th>
<th>LFU</th>
<th>Churn</th>
<th>Scan</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARC</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>LIRS</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>LeCaR</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>DLIRS</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Diagram:**
- **LRU-friendly**
- **LFU-friendly**
- **Churn**
- **Scan**

**Time (hours):**
- 0
- 5
- 10
- 15
- 20

**Requested item:**
- 0
- 10K
- 20K
- 30K
- 40K
Prior Work: LeCaR

- ML-Based: Reinforcement Learning On Cache Replacement
  - Simple
    - LRU, LFU as experts
  - Adaptive
    - Update weights
  - Outperforms state-of-the-art
    - Small cache sizes

Limitations of LeCaR

- Fixed Learning rate
  - 0.45 (empirically chosen)
- Can’t handle Scan
CACHEUS: Improving LeCaR

- Adapt Learning Rate
- Improve experts
  - Introduce scan resistance
    - Replace LRU with
      - ARC (C1)
      - LIRS (C2)
      - DLIRS
    - Scan resistant LRU: SR-LRU (C3)
  - Improve churn resistance
    - Churn resistant LFU (CR-LFU)
CACHEUS: Learning Rate Adaptation

- Learning rate changed
  - Performance change
    - Positive, reinforce latest direction
    - Negative, reverse the latest direction
- Learning rate unchanged
  - Performance change
    - Positive, no update
    - Negative, random jump
- Performance zero for 10 intervals (Einziger et al., Middleware '18)
  - Restart Learning

\[ w_t = w_{t-1} e^{\eta_r} \]
Figure: Understanding SR-LRU. Actions taken to handle request x for: cache miss, cache miss with x in history, cache hit with x in SR, and cache hit with x in R.
CACHEUS: Churn Resistance LFU (CR-LFU)

Figure: Understanding CR-LFU. Actions taken to handle request x for: cache miss, cache miss with x in history, cache hit with x in SR, and cache hit with x in R.
Figure: SR-LRU with the scan workload primitive type. Two synthetic workloads: LFU-friendly pattern (left column) and LRU-Friendly pattern (right column). The working set is 175 items with a single inserted scan of size 60.

Figure: CR-LFU with the churn workload primitive type. Two synthetic workloads: a churn pattern (left column) and a combination of churn and LRU-friendly pattern (right column). The working set is 200 items.
Experiments

- **Datasets**: 5 different sources
- **Cache sizes**: 0.05, 0.1, 0.5, 1, 5, 10%
- **6+1 Algorithms compared**
  - LRU, LFU, ARC, LIRS, LeCaR, DLIRS, CACHEUS (3 variants)
- **Total experiments**: 17,766

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Traces</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIU</td>
<td>184</td>
</tr>
<tr>
<td>MSR</td>
<td>22</td>
</tr>
<tr>
<td>CloudPhysics</td>
<td>99</td>
</tr>
<tr>
<td>CloudVPS</td>
<td>18</td>
</tr>
<tr>
<td>CloudCache</td>
<td>6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>329</strong></td>
</tr>
</tbody>
</table>
Evaluation of CACHEUS

- Paired t-test used
- **Significance**: p-value
  - **Green**: Significantly better
  - **Red**: Significantly worse
  - **Gray**: Not significant
- **Magnitude**: Effect size (Cohen’s d-measure)
  - **Dark**: High effect
  - **Light**: Low effect

**Figure**: CACHEUS vs. others

<table>
<thead>
<tr>
<th></th>
<th>Effect size</th>
<th>Better</th>
<th>Insignificant</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>CACHEUS</td>
<td>[-0.31, 2.08]</td>
<td>47%</td>
<td>40%</td>
<td>13%</td>
</tr>
</tbody>
</table>
Contributions

- **Workload Primitive Types**
  - LRU-friendly, LFU-friendly, Churn, Scan

- **CACHEUS**: Improved Cache replacement algorithm
  - **Adaptive** learning rate
  - **Improved** experts: LRU and LFU algorithm
    - SR-LRU and
    - CR-LFU
  - **Comprehensive** evaluations (17,766 simulations)
  - **Outstanding** Performance
THANKS!!!

- SyLab Team, SCIS, FIU
- BioRG Team, SCIS, FIU
- Kenneth Salem
- FAST Organizing Committee

fyusu003@fiu.edu
https://sites.google.com/fiu.edu/farzana