1. Motivation / Background

- DRAM scaling is becoming difficult
- Memories like Phase Change Memory (PCM) offer scalability, but have drawbacks
- Use DRAM as a cache to PCM

<table>
<thead>
<tr>
<th></th>
<th>PCM</th>
<th>DRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data storage</td>
<td>Resistance</td>
<td>Charge</td>
</tr>
<tr>
<td>Scalability</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Latency (R/W)</td>
<td>~4x/~12x</td>
<td>1x</td>
</tr>
<tr>
<td>Energy (R/W)</td>
<td>~2x/~40x</td>
<td>1x</td>
</tr>
<tr>
<td>Endurance</td>
<td>10^8 writes</td>
<td>N/A</td>
</tr>
</tbody>
</table>

2. Key Insight

- DRAM and PCM both employ row buffers
- Similar row hit latency, different row miss latencies
- Store data which miss in the row buffer and are reused frequently in DRAM

3. Mechanism

- For recently accessed rows in PCM,
  - Track misses to predict future locality
  - Track accesses to predict future reuse
  - Cache data after a threshold number of misses and accesses in an interval
  - Dynamically adjust threshold to adapt to runtime characteristics

4. Evaluation

- 16-core system, 32/512 KB L1/L2 per core
- Separate DRAM and PCM controllers
- 1 GB DRAM, 16 GB PCM (both 8 banks)