Prefetching branch metadata through hardware/software co-design and profiling significantly reduces MPKI

- The Context Cache stores encoded global history patterns that were deemed hard-to-predict from application profiles
- Contexts are encoded by concatenating the two least significant bits of the last 32 branches in the Last Branch Record

### PROPOSED MECHANISM

- Branch Predictor
- LBR
- Encode
- Context Cache
- Hit?
- Yes
- No
- Final Prediction

### DATACENTER WORKLOADS

<table>
<thead>
<tr>
<th>Workload</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cassandra</td>
<td>NoSQL DBMS used by Netflix/Uber</td>
</tr>
<tr>
<td>Drupal</td>
<td>PHP-based CMS (Facebook's OSS-performance)</td>
</tr>
<tr>
<td>Finagle-</td>
<td>Chirper Microblogging service (Java Renaissance)</td>
</tr>
<tr>
<td>Finagle-HTTP</td>
<td>HTTP server (Java Renaissance)</td>
</tr>
<tr>
<td>Kafka</td>
<td>Apache's stream-processing platform used by Uber/LinkedIn/Arbmb</td>
</tr>
<tr>
<td>Mediawiki</td>
<td>Open-source Wiki engine (Facebook's OSS-performance)</td>
</tr>
<tr>
<td>Tomcat</td>
<td>Apache's open-source Java web server</td>
</tr>
<tr>
<td>Verilator</td>
<td>Tool for simulating custom hardware designs used by Intel/ARM</td>
</tr>
<tr>
<td>Wordpress</td>
<td>PHP-based CMS (Facebook's OSS-performance)</td>
</tr>
</tbody>
</table>

---

**CONSIDERING OVERHEAD**

- ~50% of the ideal MPKI reduction can be gained by targeting the most impactful contexts, reducing storage overhead
- Re-evaluating the size of each context (64 bits) can also significantly reduce storage

### ADDITIONAL EVALUATION

- Evaluating our datacenter traces on the software-only Big-BranchNet [2] provided 10.41% avg. MPKI reduction compared to our design's 12.36% avg. on 64KB TAGE-SC

### WORKLOAD DESCRIPTION

- Cassandra: NoSQL DBMS used by Netflix/Uber
- Drupal: PHP-based CMS (Facebook's OSS-performance)
- Finagle-Chirper: Microblogging service (Java Renaissance)
- Finagle-HTTP: HTTP server (Java Renaissance)
- Kafka: Apache's stream-processing platform used by Uber/LinkedIn/Arbmb
- Mediawiki: Open-source Wiki engine (Facebook's OSS-performance)
- Tomcat: Apache's open-source Java web server
- Verilator: Tool for simulating custom hardware designs used by Intel/ARM
- Wordpress: PHP-based CMS (Facebook's OSS-performance)

---
