Why is Application Aware Prefetch Prioritization needed?

**Motivation**
Aggressive prefetching causes severe network congestion as prefetcher injects significant number of inaccurate prefetches into the network.

**Problem**:
1) Prefetches are prioritized over demands in NoC routers
2) Useless or less beneficial prefetches are prioritized over more useful & timely prefetches.

**Goal**
Prioritize prefetches from those applications where prefetching
1) is more likely to improve its performance
2) is likely to not cause significant interference to other applications.

---

**Prioritization Table (rank)**

<table>
<thead>
<tr>
<th>High Prefetch Count</th>
<th>Low Prefetch Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Prefetch Accuracy</td>
<td>Mid High(2)</td>
</tr>
<tr>
<td>Low Prefetch Accuracy</td>
<td>Min(4)</td>
</tr>
</tbody>
</table>

**Static Vs Dynamic Prioritization**

**Static**: Profile run determines priority of application. Rank remains static throughout the execution.

**Dynamic**: Captures applications dynamic behavior. Hardware collects prefetch accuracy & count periodically based on which applications are ranked.

---

**Need for Dynamism**

Applications have dynamic behavior.

**Results**

- System Config = 64-core CMP, mesh n/w, XY-routing, 5 VCs, demands not differentiated from prefetches.
- AVG: across 12 heterogeneous multi-programmed SPEC workloads.
- Ideal = prefetches do not content with each other or demand requests in the network.

**Future work**: Includes analyzing our techniques and combining them with other application-aware prioritization techniques in NoC routers.

Contact: nachi@cse.psu.edu