PMPM: Prediction by Combining Multiple Partial Matches

Hongliang Gao    Huiyang Zhou
Prediction by Partial Matching (PPM)

- Originally proposed for data compression by Cleary and Witten. Introduced to branch prediction by Chen et. al.
- For branch prediction:
  - Each static branch has a set of Markov predictors from order 0 to order $m$.
  - The “longest match” policy: use the $m$ immediately preceding history bits to search a pattern in the highest order Markov predictor.
- Assumptions of the PPM algorithm
  - Longer history provides a more accurate context (true).
  - A prediction counter associated with a more accurate context will provide higher prediction accuracy (false).
The “longest match” policy is not optimal

- The confidence-based PPM
  - Use the longest confident (ctr <> 0) match.
  - Misprediction rate (MPKI) reductions vs. PPM.
  - Max H = 40:

- Max H = 0 to 40:
Introduction

• Key observation on PPM
  – The “longest match” policy is not optimal for branch prediction.

• Our contributions
  – A novel algorithm: Prediction by combining Multiple Partial Matches (PMPM).
  – A PMPM-based idealistic branch predictor.
  – A PMPM-based realistic branch predictor.
Prediction by combining Multiple Partial Matches

- Different branches favor different history lengths.
  - Using a longer history than necessary:
    - Uncorrelated history information -> noise -> distribute useful information into more prediction counters.
  - Long history repeats less frequently -> only capture most recently behaviors
    - Especially harmful for “not-correlated / random-like” branches.

- Solution
  - Combining multiple matches
    - Why?
    - How: summation -> integrates both direction AND confidence.
    - Which: several longest confident matches with non-zero prediction counters.
Prediction accuracy of PMPM

- **Configuration**
  - Combine the $L$ longest **confident** matches.
  - Maximum global history length: 40.

- Prediction = $\left( \sum_{i=1}^{L} Ctr_i \geq 0 \right), \quad Ctr_i \neq 0$

- **Prediction accuracy**

![Graph showing prediction accuracy with L on the x-axis and Average MPKI on the y-axis. The graph compares PMPM-L and PPM, with the minimum MPKI highlighted.](image-url)
The idealistic PMPM predictor

Input

GHR
Path history

meta ctr & bias ctr
br tag
LHR

Short-GHR table
(Len 1-32)
tag | LRU | ctr | ubit | bf

Long-GHR table
(geometrical lengths)
tag | LRU | ctr | ubit | bf

LHR table
(Len 1-32)
tag | LRU | ctr | ubit | bf

select M (=7) longest matched useful counters

bias ctr
select N (=16) longest matched useful counters

meta ctr

Global prediction

Local prediction

final prediction
Prediction accuracy of the idealistic PMPM predictor

- **PPM**: Same predictor structure, but using the “longest match” prediction policy.
- **Average MPKI**
  - PPM: 3.330
  - PMPM: 2.824
The TAGE predictor

Bimodal table

pc, ghr, path

gtable6 (shortest)
tag  ctr  ubit

gtable5
tag  ctr  ubit

gtable0 (longest)
tag  ctr  ubit

select one matched counter (the longest match OR the 2nd longest match)

final prediction
The realistic PMPM predictor

Bimodal table

gtale ctrs:
4 groups: (gtable6, gtable5), (gtable4, gtable3), (gtable2, gtable1), (gtable0).

Total (max):
a 2-bit bimodal ctr, a 5-bit ltable ctr
four 5-bit gtable ctrs.

gtale0 (longest)
tag ctr ubit

select one matched counter
(the longest match or the 2nd longest match)

final prediction
**Ahead pipelining**

Initiate a 3-block ahead prediction

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycle1</td>
<td>Cycle2</td>
<td>Cycle3</td>
<td>Cycle4</td>
<td></td>
</tr>
</tbody>
</table>

Prediction of D is available

1. **Indexes.**
2. **Tags.**
3. **Calculate** 4 potential predictions.
4. Use information of B and C to **select** out one prediction.

1. **Tags.**
2. **Read** 4 adjacent entries.
Compared to the TAGE predictors

- **Configuration**
  - 32kB, same global history series (5 - 131), similar structures.
  - Compared to the TAGE predictor:
    - PMPM-G (GH only): 2-bit larger ctrs, 2-bit smaller tags.
  - Compared to the PMPM-G predictor:
    - PMPM-GL(GH and LH): one ltable, smaller bimodal table, smaller tags for 3 gtables.

- **Average MPKI:**
  - TAGE: 3.666
  - PMPM-G: 3.597 (higher aliasing, gcc +7.3%)
  - PMPM-GL: 3.441
The realistic PMPM predictor for CBP2

• Submitted configuration
  – Save some storage for miscellaneous registers, counters etc.
  – Empirically tuned inputs, tag widths etc.

• Several optimizations
  – Shared hysteresis bits in the bimodal table (proposed in the EV8 predictor).
  – Detect traces with high branch footprints and reset *ubits* periodically (borrowed from the TAGE predictor).
  – Limited *ubit* updates if all predictions from gtables are same.
The realistic PMPM predictor for CBP2 - accuracy

- Observations:
  - High accuracy: 3.416 MPKI
  - The local history is still important for some benchmarks (e.g., raytrace, mtrt and vortex) although we already use a very long (203) global history.

<table>
<thead>
<tr>
<th>Trace</th>
<th>CBP2-GL</th>
<th>CBP2-G</th>
<th>Trace</th>
<th>CBP2-GL</th>
<th>CBP2-G</th>
</tr>
</thead>
<tbody>
<tr>
<td>gcc</td>
<td>3.690</td>
<td>3.637</td>
<td>mcf</td>
<td>10.092</td>
<td>10.033</td>
</tr>
<tr>
<td>crafty</td>
<td>2.581</td>
<td>2.565</td>
<td>parser</td>
<td>5.215</td>
<td>5.244</td>
</tr>
<tr>
<td>compress</td>
<td>5.537</td>
<td>5.819</td>
<td>jess</td>
<td>0.393</td>
<td>0.433</td>
</tr>
<tr>
<td>raytrace</td>
<td>0.542</td>
<td>0.963</td>
<td>db</td>
<td>2.319</td>
<td>2.380</td>
</tr>
<tr>
<td>javac</td>
<td>1.107</td>
<td>1.159</td>
<td>mpegudio</td>
<td>1.102</td>
<td>1.159</td>
</tr>
<tr>
<td>mtrt</td>
<td>0.657</td>
<td>1.009</td>
<td>jack</td>
<td>0.688</td>
<td>0.763</td>
</tr>
<tr>
<td>eon</td>
<td>0.276</td>
<td>0.359</td>
<td>perlbench</td>
<td>0.314</td>
<td>0.484</td>
</tr>
<tr>
<td>gap</td>
<td>1.431</td>
<td>1.745</td>
<td>vortex</td>
<td>0.137</td>
<td>0.331</td>
</tr>
<tr>
<td>bzip2</td>
<td>0.037</td>
<td>0.042</td>
<td>twolf</td>
<td>13.551</td>
<td>13.616</td>
</tr>
</tbody>
</table>

Average PMPM-CBP2-GL: 3.416  PMPM-CBP2-G: 3.557
The realistic PMPM predictor for CBP2 – ahead pipelining

![Graph showing the comparison of PMPM-CBP2-GL and PMPM-CBP2-G with average MPKI values for 1-block, 2-block, 3-block, and 4-block. The graph indicates that PMPM-CBP2-GL has 0.11 MPKI, 3.0% and PMPM-CBP2-G has 0.15 MPKI, 4.4%.]
Summary

• Key observation on PPM
  – The “longest match” policy is not optimal for branch prediction.

• Solution
  – Prediction by combining Multiple Partial Matches (PMPM)

• PMPM-based predictor designs
  – Idealistic predictor: 2.824 MPKI.
  – Realistic predictor: 3.416 MPKI.
Thank you and Questions?