Thermometer: Profile-Guided BTB Replacement for Data Center Applications

Shixin Song, Tanvir Ahmed Khan,
Sara Mahdizadeh Shahri, Akshitha Sriraman, Niranjan K Soundararajan,
Sreenivas Subramoney, Daniel A. Jiménez, Heiner Litz, Baris Kasikci
8 processor cores

~520 ms

Millions of dollars

Share of the population using the Internet

Global warming

[1] Memory Hierarchy for Web Search, HPCA 2018
CPU Performance of Google Web Search

Frontend stalls\(^1\)

- Retiring: 32.0%
- Backend core: 15.4%
- Backend memory: 20.5%
- Bad speculation: 8.5%
- Frontend latency: 13.8%
- Frontend bandwidth: 9.7%

\(^1\) AsmDB: understanding and mitigating front-end stalls in warehouse-scale computers, ISCA 2019
Contribution #1: Why Don’t Existing Techniques Avoid Frontend Stalls?

Decoupled Frontend

Branch Target Buffer (BTB)

Optimal replacement

10% speedup

Too many branches

Best Prior BTB replacement policy 2%
Contribution #2: BTB Replacement with Thermometer

1. Profile Collection
2. Replacement Analysis
3. Hint Injection
4. Hardware Replacement
Contribution #3: Demonstration of Thermometer’s Efficacy

Thermometer

+ BTB

Reduce 21% Misses

IPC Speedup 9%
Characterization
Why Do Existing Techniques Fall Short?

Thermometer
Profile-Guided BTB Replacement

Evaluation
Demonstration of Thermometer’s Efficacy
Decoupled Frontend + Fetch Directed Prefetching to Avoid Frontend Stalls

A Scalable Front-End Architecture for Fast Instruction Delivery

Fetch Directed Instruction Prefetching

Glenn Reinman†  Todd Austin‡  Brad Calder†

Arm  AMD  Samsung
What Stops Decoupled Frontend from Eliminating All Frontend Stalls?

![Bar graph showing speedup (%) for various applications]

**Optimal BTB replacement policy 10%**

**Best Prior BTB replacement policy 2%**

BTB replacement policy is crucial to improve data center applications’ performance.
Why Do Prior Replacement Policies Fall Short?

Hardware
- SRRIP
- GHRP
- Hawkeye

<table>
<thead>
<tr>
<th>Branch</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Unfriendly</td>
</tr>
<tr>
<td>B</td>
<td>Friendly</td>
</tr>
<tr>
<td>C</td>
<td>Friendly</td>
</tr>
<tr>
<td>D</td>
<td>Unfriendly</td>
</tr>
</tbody>
</table>

Evict A

Later...

A: friendly or not?

Reuse A

Keep

BTB friendly

A

C

D

Evict

BTB unfriendly

B

F

E

Too many branches

Existing replacement policies can not retain branch information after eviction
Which Branches Are Worth Retaining in BTB?

Only half of all unique branches are hot and consistently retained by the optimal replacement policy.
Which Branches Are Worth Inserting into BTB?

The optimal replacement policy does not insert half of all cold branches into BTB.
Characterization

Why Do Existing Techniques Fall Short?

Thermometer

Profile-Guided BTB Replacement

Evaluation

Demonstration of Thermometer’s Efficacy
Thermometer

1. Profile Collection
2. Replacement Analysis
3. Hint Injection
4. Hardware Replacement

Online Offline
Thermometer: Profile Collection

1. Profile Collection
2. Replacement Analysis
3. Hint Injection
4. Hardware Replacement

Online Profiling

<table>
<thead>
<tr>
<th>Branch IP</th>
<th>Taken or not</th>
<th>Branch Target IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x10</td>
<td>Taken</td>
<td>0x50</td>
</tr>
<tr>
<td>0x14</td>
<td>Taken</td>
<td>0x30</td>
</tr>
<tr>
<td>0x34</td>
<td>Not Taken</td>
<td>-</td>
</tr>
<tr>
<td>0x08</td>
<td>Taken</td>
<td>0x14</td>
</tr>
</tbody>
</table>

Binary + Intel PT
Thermometer: Offline Analysis

1. Profile Collection
2. Replacement Analysis
3. Hint Injection
4. Hardware Replacement

Branch Trace + Optimal BTB Replacement

Offline simulation: 4.18-167s

<table>
<thead>
<tr>
<th>Branch IP</th>
<th># Taken</th>
<th># Hits</th>
<th>Hit-to-taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x10</td>
<td>500</td>
<td>10</td>
<td>98%</td>
</tr>
<tr>
<td>0x14</td>
<td>60</td>
<td>65</td>
<td>48%</td>
</tr>
<tr>
<td>0x34</td>
<td>100</td>
<td>30</td>
<td>77%</td>
</tr>
<tr>
<td>0x08</td>
<td>300</td>
<td>30</td>
<td>91%</td>
</tr>
</tbody>
</table>
Thermometer: Hint Injection

1. Profile Collection
2. Replacement Analysis
3. Hint Injection
4. Hardware Replacement

Branch IP | Hit-to-taken | Temperature | Hint Bits
---|-------------|-------------|---------
0x10  | 98%         | hot         | 0b10    
0x14  | 48%         | warm        | 0b01    
0x34  | 77%         | cold        | 0b00    
0x08  | 91%         | hot         | 0b10    

3 categories
2-bit hints

Branch IP: 10100111 01010101 00001000 10001010

x86   aarch64

unused

hot 80% warm 50% cold 50%
Thermometer: Hardware Replacement

1. Profile Collection
2. Replacement Analysis
3. Hint Injection
4. Hardware Replacement

Overhead in BTB: 2.6%
0.2% performance impact

<table>
<thead>
<tr>
<th>Branch</th>
<th>Target</th>
<th>Temperature</th>
<th>LRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x10</td>
<td>0x50</td>
<td>0b10 (hot)</td>
<td>1</td>
</tr>
<tr>
<td>0x14</td>
<td>0x30</td>
<td>0b01 (warm)</td>
<td>0</td>
</tr>
<tr>
<td>0x34</td>
<td>0x00</td>
<td>0b00 (cold)</td>
<td>3</td>
</tr>
<tr>
<td>0x08</td>
<td>0x14</td>
<td>0b00 (cold)</td>
<td>2</td>
</tr>
</tbody>
</table>

New branch

| 0x78 | 0x28 | 0b10 (hot) |
Characterization
Importance of BTB Replacement

Thermometer
Profile-Guided BTB Replacement

Evaluation
Demonstration of Thermometer’s Efficacy
Evaluation Metrics for Thermometer

- **IPC speedup**
- **BTB miss reduction**

Optimal replacement policy
- SRRIP, GHRP, Hawkeye
IPC Speedup

Thermometer achieves 84% of the speedup provided by the optimal replacement policy
Thermometer achieves 63% of BTB miss reduction provided by the optimal replacement policy.
Thermometer: Profile-Guided BTB Replacement

9% (up to 65%) Speedup!

Shixin Song  shixins@umich.edu  Tanvir Ahmed Khan  takh@umich.edu