Berti: A Per-Page Best-Request-Time Delta Prefetcher

Alberto Ros

Universidad de Murcia, Spain
aros@ditec.um.es
OUTLINE

1 Motivation

2 The Berti Prefetcher

3 Evaluation

4 Conclusions and Future Work
1 Motivation

2 The Berti Prefetcher

3 Evaluation

4 Conclusions and Future Work
Motivation

- Stride prefetcher, a simple technique but...
  - The stride can be difficult to find
    - The order of accesses can be altered by out-of-order cores, lower cache levels, or even prefetchers of lower cache levels
  - May lead to late prefetches
    - No timing considerations, although throttling can help
Motivation

- **Best-offset prefetcher (BOP)** addresses the limitations of stride prefetchers
- Example: pattern found in `mcf`

<table>
<thead>
<tr>
<th>Blocks accessed [0..63]</th>
<th>Stride</th>
</tr>
</thead>
<tbody>
<tr>
<td>...110110110110</td>
<td></td>
</tr>
</tbody>
</table>
**Motivation**

- **Best-offset prefetcher (BOP)** addresses the limitations of stride prefetchers.
- Example: pattern found in mcf.

<table>
<thead>
<tr>
<th>Blocks accessed [0..63]</th>
<th>Stride</th>
</tr>
</thead>
<tbody>
<tr>
<td>...1101101101100</td>
<td>-1?, -2?</td>
</tr>
</tbody>
</table>
**Motivation**

- Best-offset prefetcher (BOP) addresses the limitations of stride prefetchers
- Example: pattern found in `mcf`

<table>
<thead>
<tr>
<th>Blocks accessed [0..63]</th>
<th>Stride</th>
</tr>
</thead>
<tbody>
<tr>
<td>...110110110110</td>
<td>-1?, -2?, -3?</td>
</tr>
</tbody>
</table>
Motivation

- **Best-offset prefector (BOP)** addresses the limitations of stride prefetchers
- Example: pattern found in \texttt{mcf}

<table>
<thead>
<tr>
<th>Blocks accessed [0..63]</th>
<th>Stride</th>
<th>BO</th>
</tr>
</thead>
<tbody>
<tr>
<td>...1101101101100</td>
<td>-1?, -2?, -3?</td>
<td>-6</td>
</tr>
</tbody>
</table>
Motivation

- **Best-offset prefetcher** (BOP) addresses the limitations of stride prefetchers
- Example: pattern found in `mcf`

<table>
<thead>
<tr>
<th>Blocks accessed [0..63]</th>
<th>Stride</th>
<th>BO</th>
</tr>
</thead>
<tbody>
<tr>
<td>...110110110110</td>
<td>-1?, -2?, -3?</td>
<td>-6</td>
</tr>
</tbody>
</table>
Motivation

- **Best-offset prefetcher (BOP)** addresses the limitations of stride prefetchers
- Example: pattern found in mcf

<table>
<thead>
<tr>
<th>Blocks accessed [0..63]</th>
<th>Stride</th>
<th>BO</th>
</tr>
</thead>
<tbody>
<tr>
<td>...110110110110</td>
<td>-1?, -2?, -3?</td>
<td>-6</td>
</tr>
</tbody>
</table>
Motivation

- **Best-offset prefetcher (BOP)** addresses the limitations of stride prefetchers.
- Example: pattern found in `mcf`:

<table>
<thead>
<tr>
<th>Blocks accessed [0..63]</th>
<th>Stride</th>
<th>BO</th>
</tr>
</thead>
<tbody>
<tr>
<td>...110110110110</td>
<td>-1?, -2?, -3?</td>
<td>-6</td>
</tr>
</tbody>
</table>

- But BOP detects a best offset per application phase:
  - And different pages have different best deltas!
OUTLINE

1 Motivation

2 The Berti Prefetcher

3 Evaluation

4 Conclusions and Future Work
**Terminology**

- **Offset vs Delta (or Stride)**

  Blocks accessed [0..63]
  
  ...110110110110

  Offset    Delta
**Key Concept**

- Per-page Berti *(Best request time)* Delta
  - Berti: the delta that would achieve more timely prefetches
  - It is calculated for each page

Alberto Ros
Third Data Prefetching Championship
Jun 23rd, 2019
8 / 19
**Key Concept**

- Per-page Berti *(Best request time)* Delta
  - Berti: the delta that would achieve more timely prefetches
  - It is calculated for each page
- We track hot pages and their offsets accessed

<table>
<thead>
<tr>
<th>Blocks accessed [0..63]</th>
</tr>
</thead>
<tbody>
<tr>
<td>...110110110110</td>
</tr>
</tbody>
</table>
**Key Concept**

- Per-page Berti (**Best request time**) Delta
  - Berti: the delta that would achieve more timely prefetches
  - It is calculated for each page
- We track hot pages and their offsets accessed
- When pages become cold, Berti is calculated

<table>
<thead>
<tr>
<th>Blocks accessed [0..63]</th>
<th>Berti</th>
</tr>
</thead>
<tbody>
<tr>
<td>...110110110110</td>
<td>-6</td>
</tr>
</tbody>
</table>

Delta
Two prefetching modes

<table>
<thead>
<tr>
<th>Blocks accessed [0..63]</th>
<th>Berti</th>
</tr>
</thead>
<tbody>
<tr>
<td>...110110110110</td>
<td>-6</td>
</tr>
</tbody>
</table>
Two prefetching modes

- Berti: One prefetch (timely)

<table>
<thead>
<tr>
<th>Blocks accessed [0..63]</th>
<th>Berti</th>
</tr>
</thead>
<tbody>
<tr>
<td>...110110110110</td>
<td>-6</td>
</tr>
</tbody>
</table>
Two prefetching modes

- Berti: One prefetch (timely)
- Burst: Several prefetches (first access to the page)
  - Expected to be late prefetches

<table>
<thead>
<tr>
<th>Blocks accessed [0..63]</th>
<th>Berti</th>
</tr>
</thead>
<tbody>
<tr>
<td>...110110110110</td>
<td>-6</td>
</tr>
</tbody>
</table>

Berti Access

Burst

Alberto Ros

Third Data Prefetching Championship

Jun 23rd, 2019
IP CLUSTERING

- Pages can be accessed by different load instructions
- We want all loads to a page to agree in the same Berti
  - Solution: Cluster the loads that access the same page
BERTI PREFETCHER OVERVIEW & EXAMPLE

Record
IP
Recorded pages

Current
Previous prefetches
Previous demand requests

Current pages

Miss for block 62 resolved
Latency 35 cycles
Miss for blocks 61, 59, and 58 resolved
Latency 35 cycles
Miss for block 56 resolved
Latency 35 cycles

Hits due to prefetches

Recording cold pages
**BERTI PREFETCHER OVERVIEW & EXAMPLE**

- **New page accessed and new block (offset 62)**

  - **Record**
    - IP
      - Recorded pages
        - ... 
        - ... 

  - **Current**
    - Current pages
      - ... 

- **Previous prefetches**
  - Offset: 62, Cycle: 10

- **Previous demand requests**

- **Current pages**

- **Vector**
  - First Offset: 62
  - Berti: 0
  - Counter: 0

- **Miss for block 62 resolved**
  - Latency: 35 cycles

- **Miss for blocks 61, 59, and 58 resolved**
  - Latency: 35 cycles

- **Miss for block 56 resolved**
  - Latency: 35 cycles

- **Hit due to prefetches**
  - Offset: 56
  - Cycle/Lat: 1/35

- **Recording cold pages**

- **Clustering IPs**

Alberto Ros

Third Data Prefetching Championship

Jun 23rd, 2019
**BERTI Prefetcher Overview & Example**

Record

Access to block with offset 61

Current

Previous prefetches

Previous demand requests

Current pages

<table>
<thead>
<tr>
<th>Offset</th>
<th>Cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>10</td>
</tr>
<tr>
<td>61</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vector</th>
<th>First Offset</th>
<th>Berti</th>
<th>Counter</th>
</tr>
</thead>
<tbody>
<tr>
<td>...000000000110</td>
<td>62</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Berti Prefetcher Overview & Example

Accesses to blocks 59, 58 and 56

Record

Recorded pages

IP

Current

Previous prefetches

Current pages

Previous demand requests

Offset | Cycle
---|---
62 | 10
61 | 20
59 | 30
58 | 40
56 | 50

Miss for block 62 resolved
Latency 35 cycles
Miss for blocks 61, 59, and 58 resolved
Latency 35 cycles
Miss for block 56 resolved
Latency 35 cycles

Hits due to prefetches
Offset/Lat: 56/1, 55/0, 60/6

Recording cold pages

Vector...000010110110

...000000000010

62

0

0
Berti Prefetcher Overview & Example

Miss for block 62 resolved
Latency 35 cycles

Record

Recorded pages

Current

Previous prefetches

Previous demand requests

Current pages

Vector

First Offset

Berti

Counter

...000010110110

62

0

0

Offset  Cycle

62  
61  
59  
58  
56  

10  
20  
30  
40  
50  

Miss for blocks 61, 59, and 58 resolved
Latency 35 cycles

Miss for block 56 resolved
Latency 35 cycles

Hits due to prefetches
Offset  Cycle/Lat

56  
55  
50  

1  
35  
60

Recording cold pages

Clustering IPs

Alberto Ros

Third Data Prefetching Championship

Jun 23rd, 2019
Berti Prefetcher Overview & Example

Record

Current

Miss for blocks 61, 59, and 58 resolved
Latency 35 cycles

IP

Recorded pages

Vector

First Offset | Berti | Counter
---|---|---
...000010110110 | 62 | 0 | 0

Current pages

Previous prefetches

Previous demand requests

Offset | Cycle
---|---
62 | 10
61 | 20
59 | 30
58 | 40
56 | 50

Clustering IPs

Alberto Ros  
Third Data Prefetching Championship  
Jun 23rd, 2019
**BERTI Prefetcher Overview & Example**

- Miss for block 56 resolved
  - Latency 35 cycles

- Record
  - IP
  - Recorded pages

- Current
  - Previous prefetches
  - Previous demand requests

- Current pages

- Vector
  - First Offset
  - Berti
  - Counter

<table>
<thead>
<tr>
<th>Offset</th>
<th>Cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>10</td>
</tr>
<tr>
<td>61</td>
<td>20</td>
</tr>
<tr>
<td>59</td>
<td>30</td>
</tr>
<tr>
<td>58</td>
<td>40</td>
</tr>
<tr>
<td>56</td>
<td>50</td>
</tr>
</tbody>
</table>

- Recording cold pages

- Clustering IPs

Alberto Ros

Third Data Prefetching Championship

Jun 23rd, 2019
Berti Prefetcher Overview & Example

Record
Recorded pages

IP

...  ...

Offset  C  Cycle/Lat
56     1    35
55     0    60

Current

Previous prefetches

...  ...

Previous demand requests

Current pages

Miss for block 62 resolved
Latency 35 cycles

Miss for blocks 61, 59, and 58 resolved
Latency 35 cycles

Miss for block 56 resolved
Latency 35 cycles

Hits due to prefetches

Offset  C  Cycle/Lat
56  1  35
55  0  60

Recording cold pages

Clustering IPs

Alberto Ros
Third Data Prefetching Championship
Jun 23rd, 2019
Berti Prefetcher Overview & Example

Recording cold pages

Record

Recorded pages

Current

Previous prefetches

Previous demand requests

Current pages

Vector | First Offset | Berti
--- | --- | ---
...000010110110 | 62 | -6

Hit for block 62 resolved
Latency 35 cycles

Miss for blocks 61, 59, and 58 resolved
Latency 35 cycles

Miss for block 56 resolved
Latency 35 cycles

Hits due to prefetches
Offset | Cycle/Lat
--- | ---
56 | 1 | 35
55 | 0 | 60

Recording cold pages

Vector | First Offset | Berti | Counter
--- | --- | --- | ---
...000010110110 | 62 | -6 | 1
**Berti Prefetcher Overview & Example**

Clustering IPs

Record

- IP
- Recorded pages

Current

- Previous prefetches
- Previous demand requests

- Current pages

Miss for block 62 resolved
Latency 35 cycles

Miss for blocks 61, 59, and 58 resolved
Latency 35 cycles

Miss for block 56 resolved
Latency 35 cycles

Hits due to prefetches

Offset Cycle/Lat
- 56 1 35
- 55 0 60

Recording cold pages
On a cache access the prefetcher checks if:

1. Matches page and its first access → high confidence
2. Matches IP and the first page access → high confidence
3. Page is hot and Berti counter > 2 → medium confidence
4. Matches page → low confidence
5. Matches IP → low confidence

If confidence is:

- High and first access → Berti & burst
- Medium or low → Berti
OUTLINE

1. MOTIVATION

2. THE BERTI PREFETCHER

3. EVALUATION

4. CONCLUSIONS AND FUTURE WORK
CONFIGURATIONS AND BENCHMARKS

CONFIGURATIONS

- Cores: 1 and 4
- Prefetchers: None, NextLine, Stride, SPP, KPCP, and Berti
  - BOP not considered (porting it from DCP-2 did not give the expected results)

BENCHMARKS

- SPEC CPU 2017
  - 1 core: > 1 MPKI at the LLC without prefetching
  - 4 core: random mixes
  - 50M instructions warm-up + 200M instructions stats
## Results

Normalized with respect to no prefetch

<table>
<thead>
<tr>
<th>Configuration (L1D, L2C, LLC)</th>
<th>1 core</th>
<th>4 cores</th>
</tr>
</thead>
<tbody>
<tr>
<td>NextLine, NextLine, NextLine</td>
<td>1.2688</td>
<td>1.1045</td>
</tr>
<tr>
<td>NextLine, Stride, NextLine</td>
<td>1.2946</td>
<td>1.1147</td>
</tr>
<tr>
<td>NextLine, SPP, NextLine</td>
<td>1.3083</td>
<td>1.1043</td>
</tr>
<tr>
<td>NextLine, KPCP, NextLine</td>
<td>1.3142</td>
<td>1.1032</td>
</tr>
<tr>
<td>NextLine, Berti, NextLine</td>
<td>1.3211</td>
<td>1.1042</td>
</tr>
<tr>
<td>Berti, Berti, Berti</td>
<td>1.3347</td>
<td>1.1087</td>
</tr>
<tr>
<td>Berti+, Berti, NextLine</td>
<td>1.3471</td>
<td>1.1186</td>
</tr>
<tr>
<td>Berti+, Berti, None</td>
<td>1.3303</td>
<td>1.1268</td>
</tr>
</tbody>
</table>
Comments/questions

- Merged prefetches do not call prefetcher_operate
  - Implication: L2 and LLC prefetchers may not see the request of a block by the previous level!
  - How can this affect performance?

- Cache contention model
  - Cache port for prefetches?
  - Impact in filtering prefetches that hit in cache negligible
OUTLINE

1 MOTIVATION

2 THE BERTI PREFETCHER

3 EVALUATION

4 CONCLUSIONS AND FUTURE WORK
CONCLUSIONS AND FUTURE WORK

CONCLUSIONS

- Per-page Berti Delta Prefetcher finds more timely prefetches

FUTURE WORK

- Adding confidence counters
- Explore Berti for LLC: late prefetches?
BERTI: A PER-PAGE BEST-REQUEST-TIME DELTA PREFETCHER

Alberto Ros

Universidad de Murcia, Spain
aros@ditec.um.es