BERTI: A PER-PAGE BEST-REQUEST-TIME DELTA PREFETCHER

Alberto Ros

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

UNIVERSIDAD DE MURCIA

Outline

UNIVERSIDAD DE MURCIA

Outline

UNIVERSIDAD DE MURCIA

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

UNIVERSIDAD DE MURCIA

MOTIVATION

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

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

UNIVERSIDAD DE MURCIA

MOTIVATION

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

Blocks accessed [0..63]	Stride
...1101101**10**1**1**0	-1?, -2?

UNIVERSIDAD DE MURCIA

MOTIVATION

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

Blocks accessed [0..63]	Stride
...1101101**1**01**1**0	-1?, -2?, -3?

UNIVERSIDAD DE MURCIA

MOTIVATION

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

Blocks accessed [0..63]	Stride	BO
...1101**1**0110110**1**0	-1?, -2?, -3?	-6

UNIVERSIDAD DE MURCIA

MOTIVATION

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

Blocks accessed [0..63]	Stride	BO
...1101101101**1**10	-1?, -2?, -3?	-6

UNIVERSIDAD DE MURCIA

MOTIVATION

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

Blocks accessed [0..63]	Stride	BO
...1**1**0110110**1**0110	-1?, -2?, -3?	-6

UNIVERSIDAD DE MURCIA

MOTIVATION

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

Blocks accessed [0..63]	Stride	BO
...110110110110	-1?, -2?, -3?	-6

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

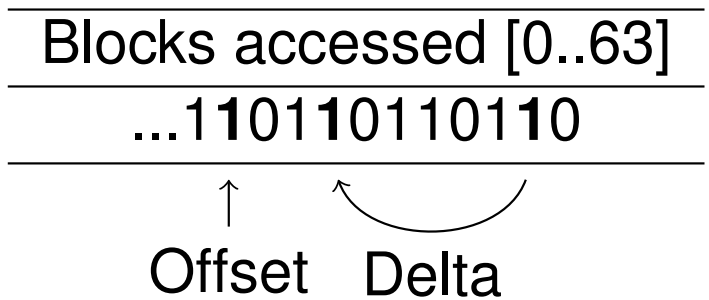
UNIVERSIDAD DE MURCIA

OUTLINE

UNIVERSIDAD DE MURCIA

TERMINOLOGY

- Offset vs Delta (or Stride)



KEY CONCEPT

- Per-page Berti (**Be**st **r**equest **ti**me) Delta
  - Berti: the delta that would achieve more timely prefetches
  - It is calculated for each page

UNIVERSIDAD DE MURCIA

KEY CONCEPT

- Per-page Berti (**Be**st **r**equest **ti**me) Delta
  - Berti: the delta that would achieve more timely prefetches
  - It is calculated for each page
- We track hot pages and their offsets accessed

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

KEY CONCEPT

- Per-page Berti (**Be**st **r**equest **ti**me) 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

Blocks accessed [0..63]	Berti
...1101**1**0110110**1**0	-6

Delta

UNIVERSIDAD DE MURCIA

PREFETCHING MODES

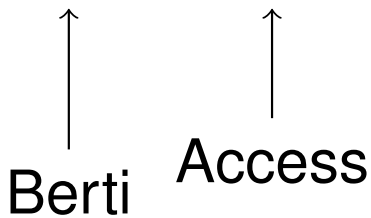
- Two prefetching modes

Blocks accessed [0..63]	Berti
...110110110110	-6

PREFETCHING MODES

- Two prefetching modes
  - Berti: One prefetch (timely)

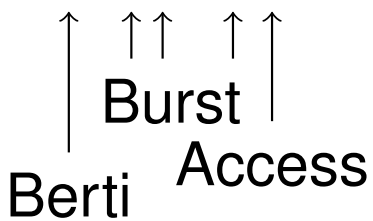
Blocks accessed [0..63]	Berti
...1101**1**0110110**1**0	-6



PREFETCHING MODES

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

Blocks accessed [0..63]	Berti
...1101**1**0110110**1**0	-6

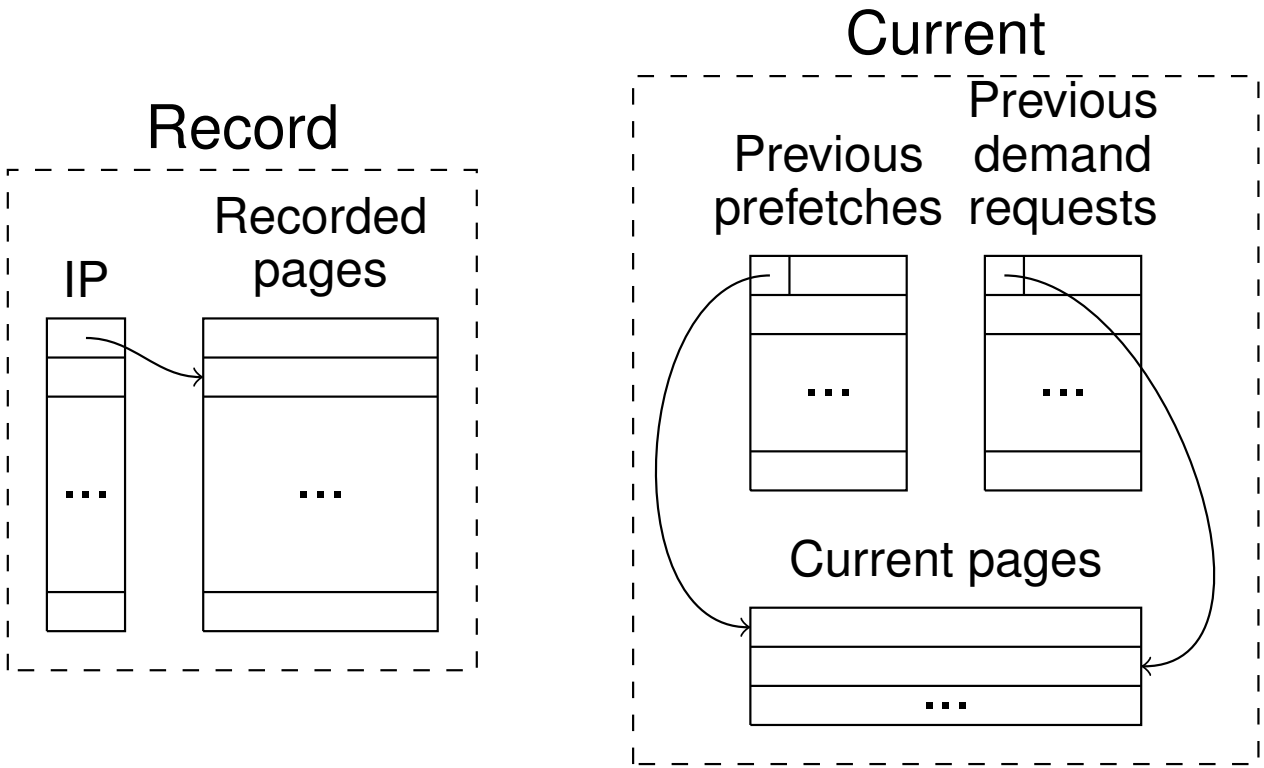


IP CLUSTERING

- 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

UNIVERSIDAD DE MURCIA

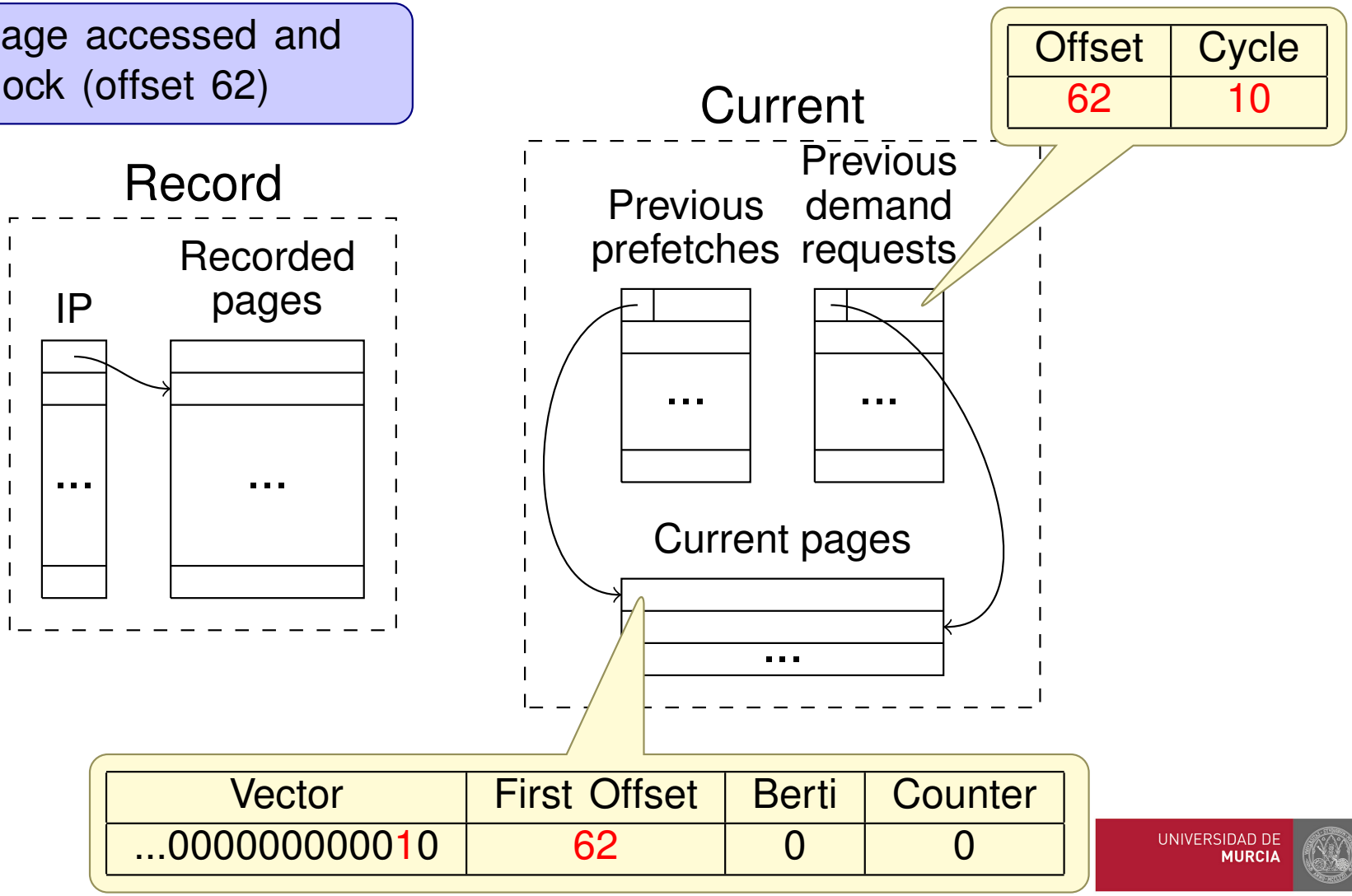
BERTI PREFETCHER OVERVIEW & EXAMPLE



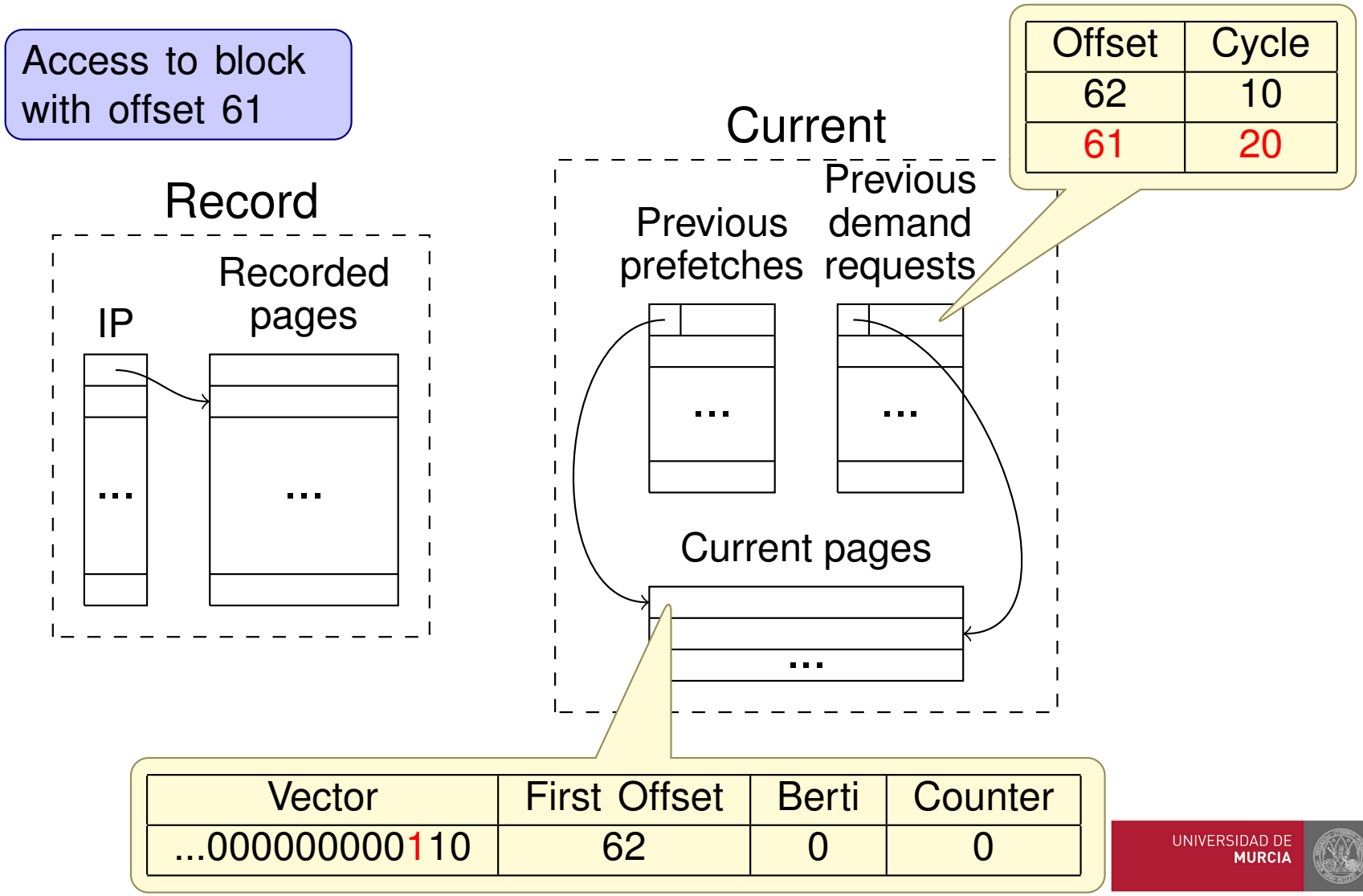
UNIVERSIDAD DE MURCIA

BERTI PREFETCHER OVERVIEW & EXAMPLE

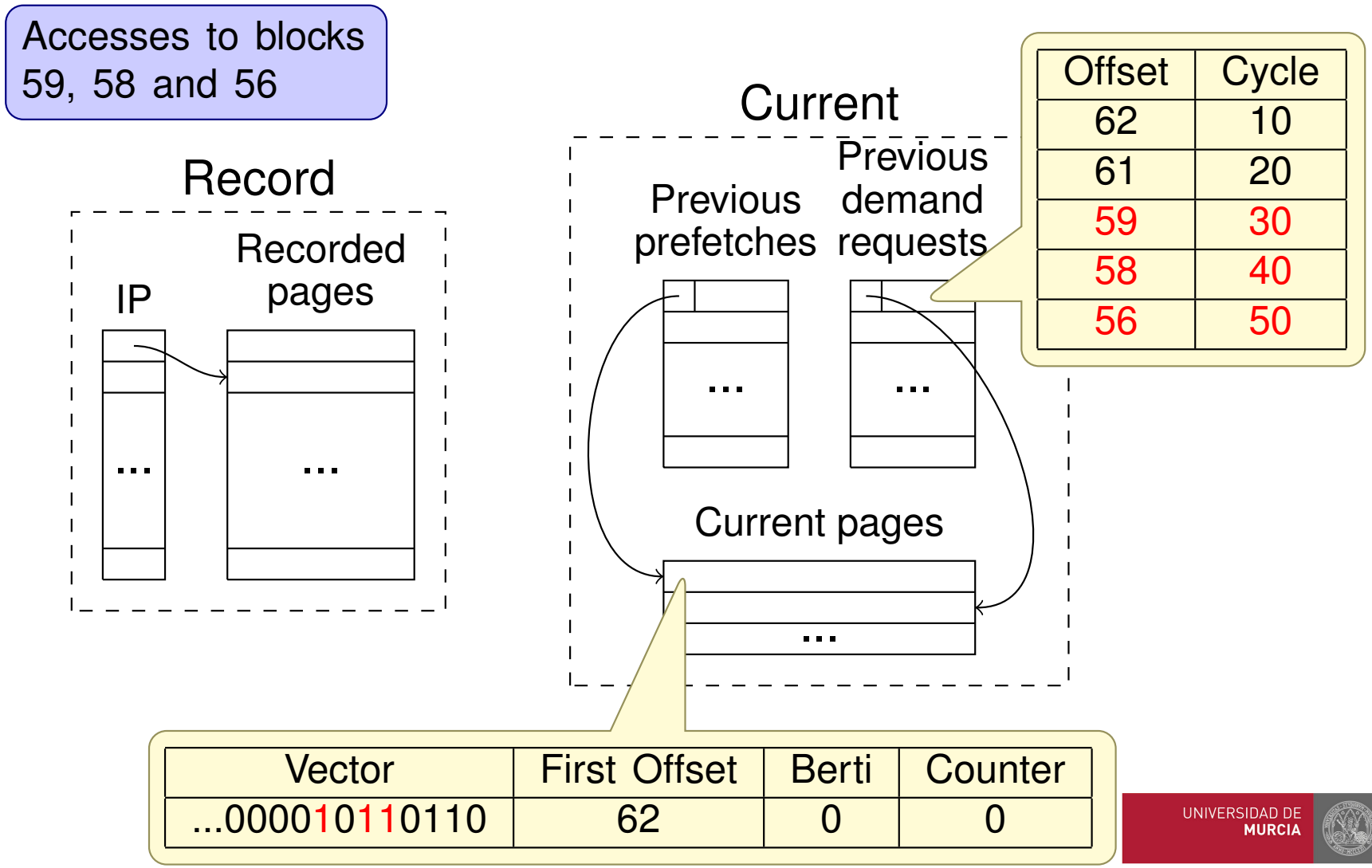
New page accessed and new block (offset 62)



BERTI PREFETCHER OVERVIEW & EXAMPLE

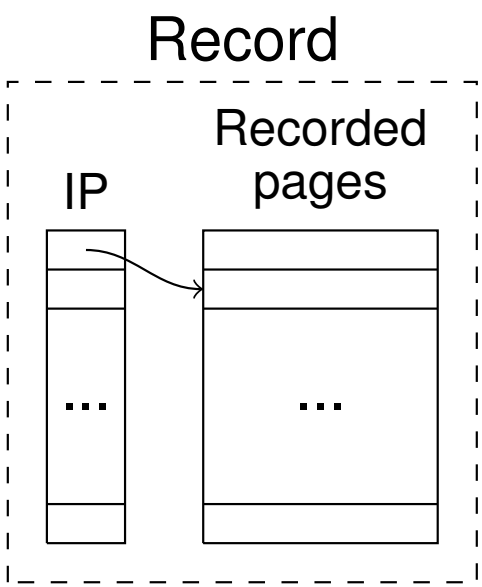


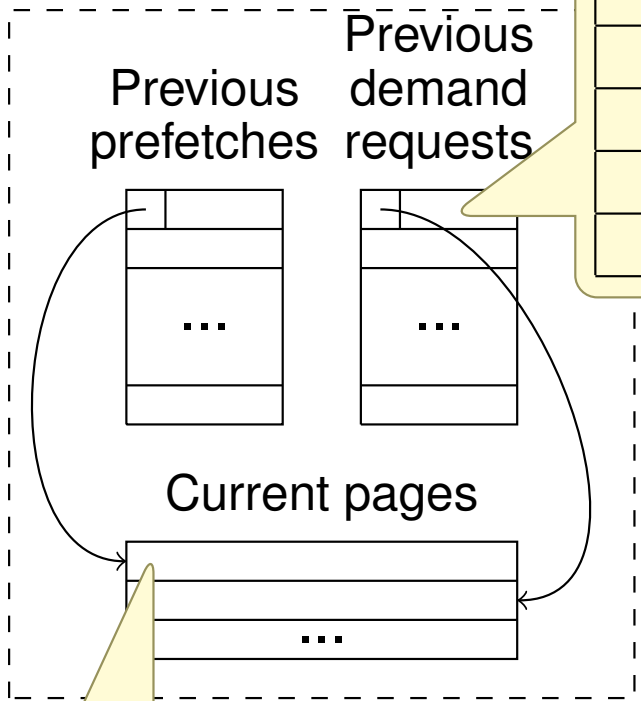
BERTI PREFETCHER OVERVIEW & EXAMPLE



BERTI PREFETCHER OVERVIEW & EXAMPLE

Miss for block 62 resolved
Latency 35 cycles





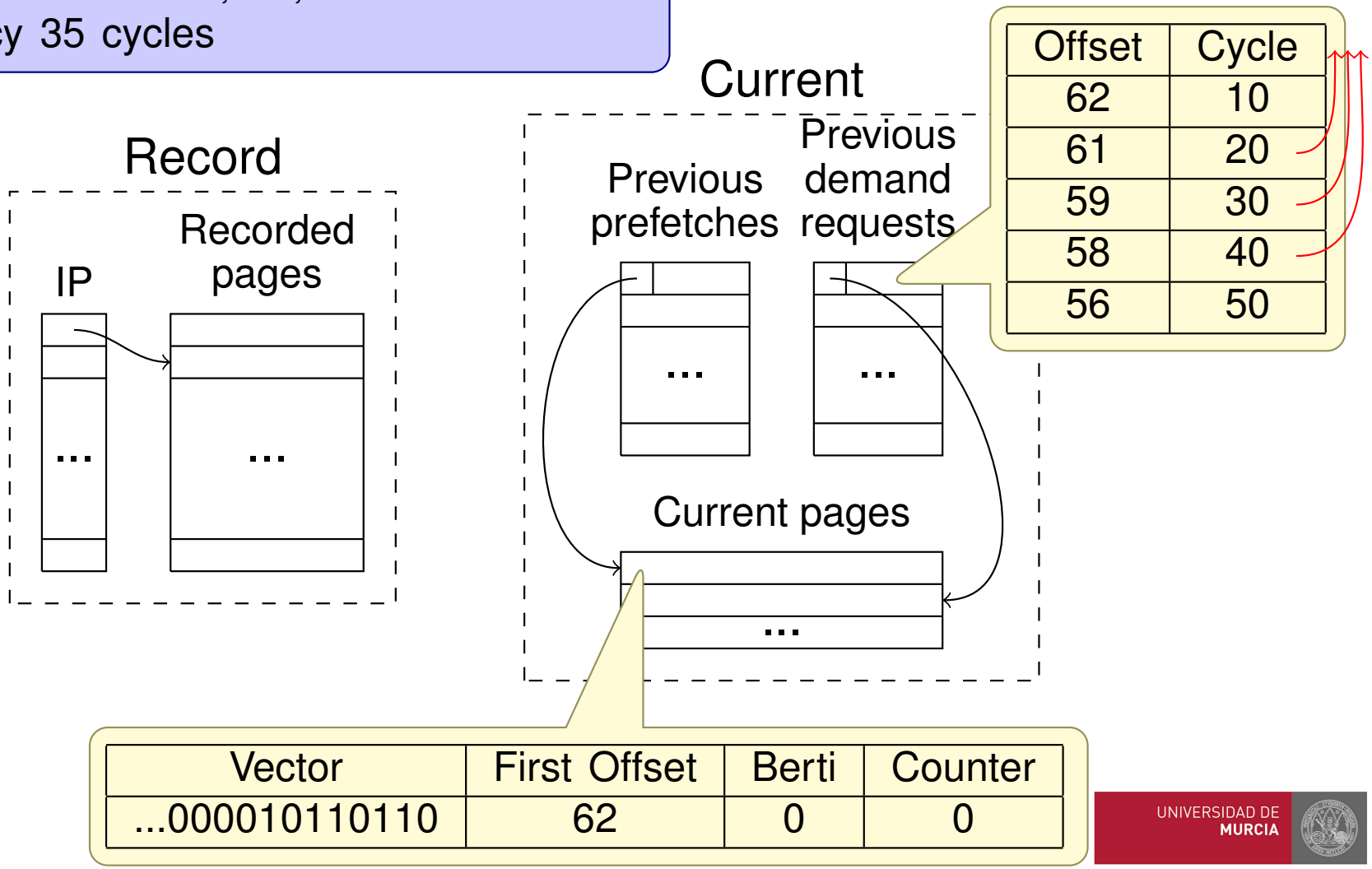
Offset	Cycle
62	10
61	20
59	30
58	40
56	50

Vector	First Offset	Berti	Counter
...000010110110	62	0	0

UNIVERSIDAD DE MURCIA

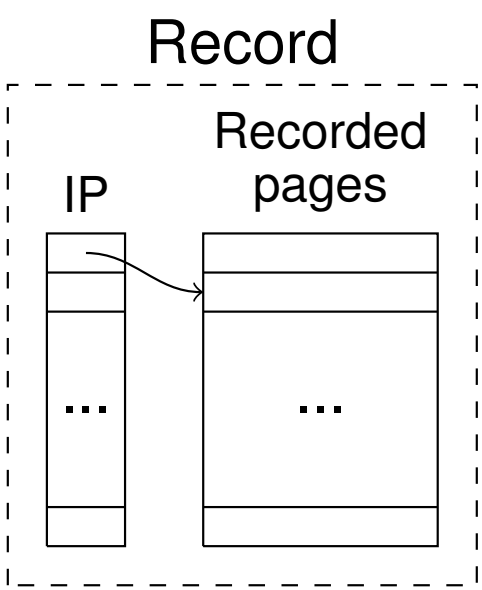
BERTI PREFETCHER OVERVIEW & EXAMPLE

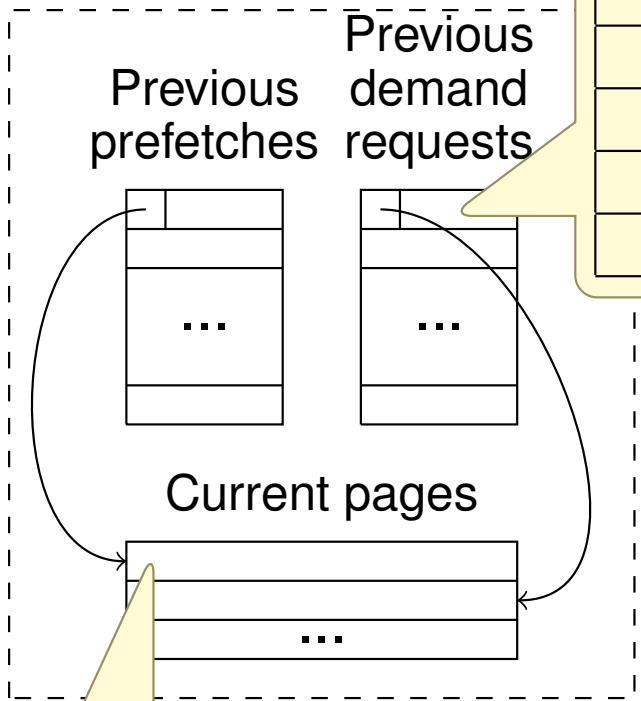
Miss for blocks 61, 59, and 58 resolved
Latency 35 cycles



BERTI PREFETCHER OVERVIEW & EXAMPLE

Miss for block 56 resolved
Latency 35 cycles





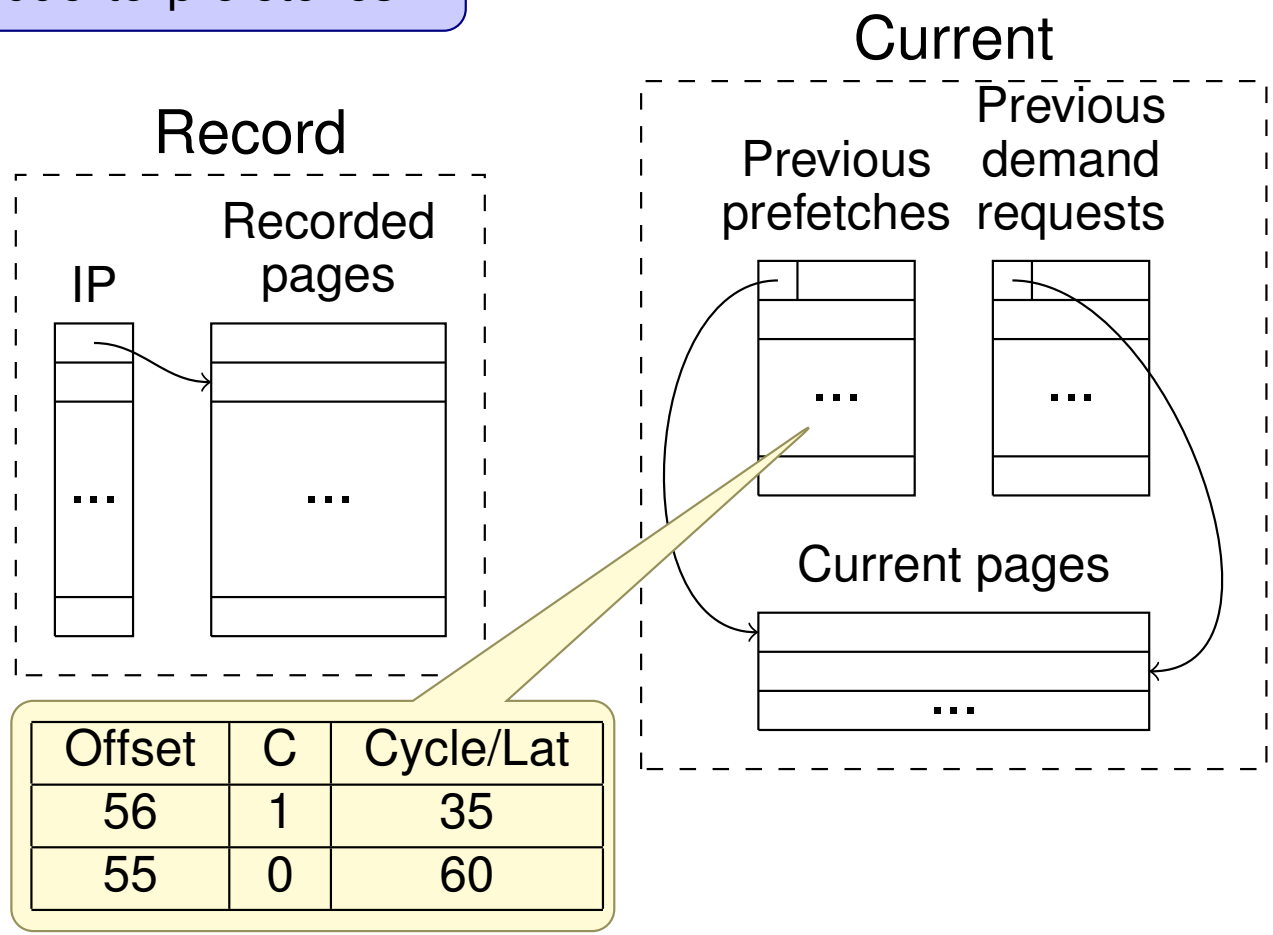
Offset	Cycle
62	10
61	20
59	30
58	40
56	50

Vector	First Offset	Berti	Counter
...000010110110	62	-6	1

UNIVERSIDAD DE MURCIA

BERTI PREFETCHER OVERVIEW & EXAMPLE

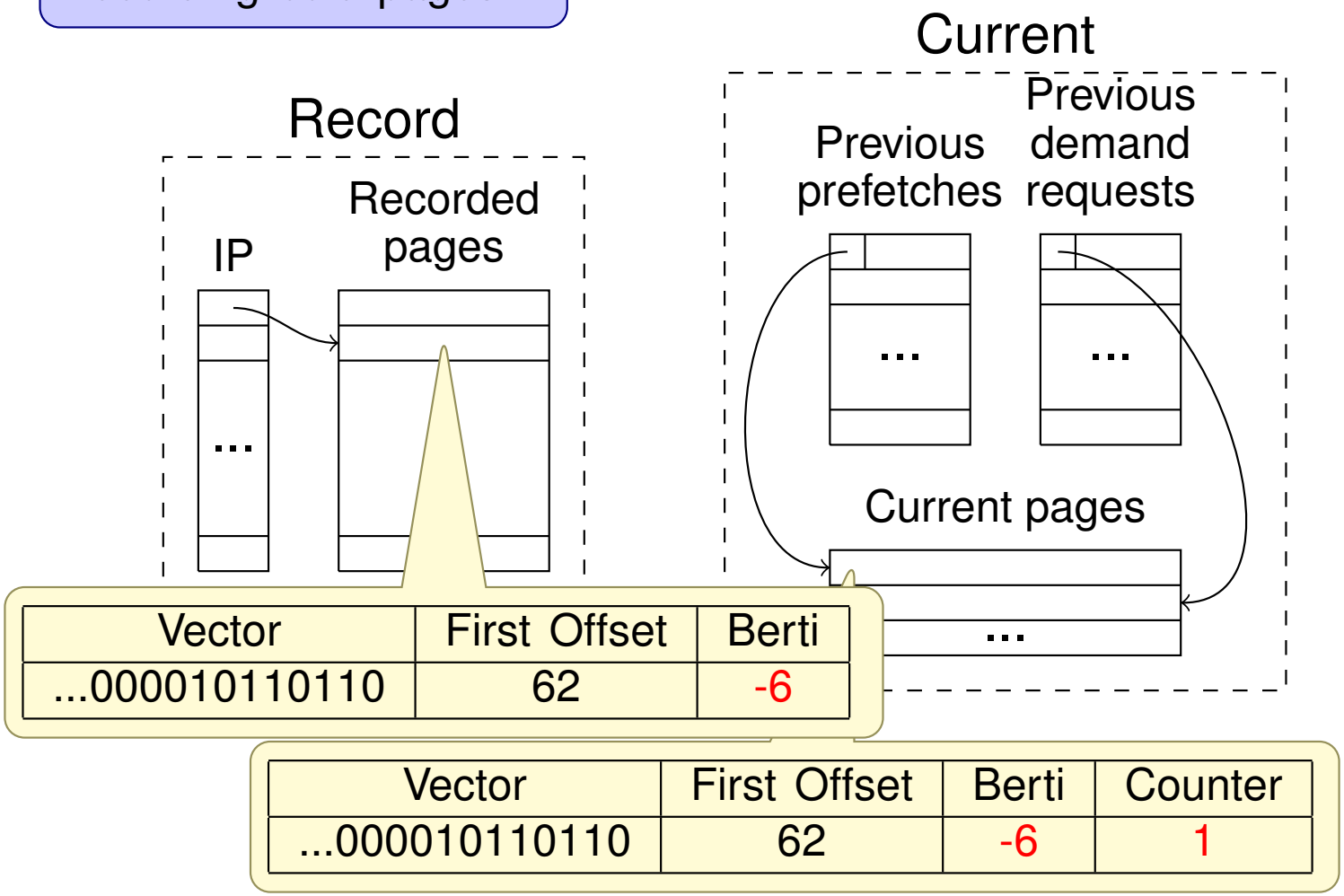
Hits due to prefetches



Offset	C	Cycle/Lat
56	1	35
55	0	60

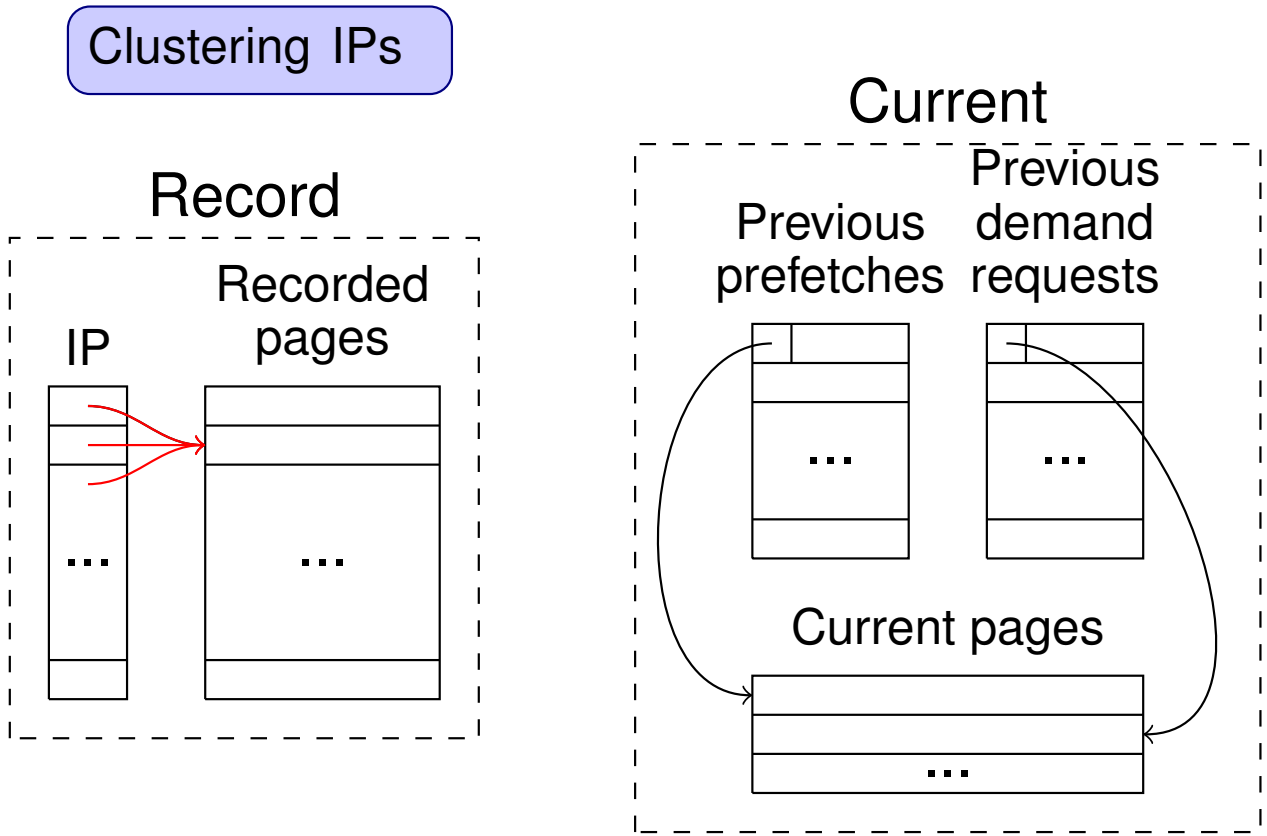
BERTI PREFETCHER OVERVIEW & EXAMPLE

Recording cold pages



UNIVERSIDAD DE MURCIA

BERTI PREFETCHER OVERVIEW & EXAMPLE



BUILDING CONFIDENCE

- On a cache access the prefetcher checks if:
  1. Matches page and its first access $\rightarrow$ high confidence
  2. Matches IP and the first page access $\rightarrow$ high confidence
  3. Page is hot and Berti counter > 2 $\rightarrow$ medium confidence
  4. Matches page $\rightarrow$ low confidence
  5. Matches IP $\rightarrow$ low confidence

- If confidence is:
  - High and first access $\rightarrow$ Berti & burst
  - Medium or low $\rightarrow$ Berti

UNIVERSIDAD DE MURCIA

Outline

UNIVERSIDAD DE MURCIA

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

UNIVERSIDAD DE MURCIA

RESULTS

- Normalized with respect to no prefetch

Configuration (L1D, L2C, LLC)	1 core	4 cores
NextLine, NextLine, NextLine	1.2688	1.1045
NextLine, Stride, NextLine	1.2946	1.1147
NextLine, SPP, NextLine	1.3083	1.1043
NextLine, KPCP, NextLine	1.3142	1.1032
NextLine, Berti, NextLine	1.3211	1.1042
Berti, Berti, Berti	1.3347	1.1087
Berti+, Berti, NextLine	1.3471	1.1186
Berti+, Berti, None	1.3303	1.1268

UNIVERSIDAD DE MURCIA

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

UNIVERSIDAD DE MURCIA

Outline

UNIVERSIDAD DE MURCIA

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?

UNIVERSIDAD DE MURCIA

BERTI: A PER-PAGE BEST-REQUEST-TIME DELTA PREFETCHER

Alberto Ros

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

UNIVERSIDAD DE MURCIA