# Berti: An Accurate Local-Delta Data Prefetcher

**Agustín Navarro-Torres**[1], Biswabandan Panda[2], Jesús Alastruey-Benedé[1], Pablo Ibáñez[1], Víctor Viñals-Yúfera[1], Alberto Ros[3]

[1]Universidad de Zaragoza
{**agusnt**, jalastru, imarin, victor}@unizar.es

[2]Indian Institute of Technology Bombay
biswa@cse.iitb.ac.in

[3]University of Murcia
aros@ditec.um.es

MICRO'55, Session Microarchitecture II, October 4th, 2022

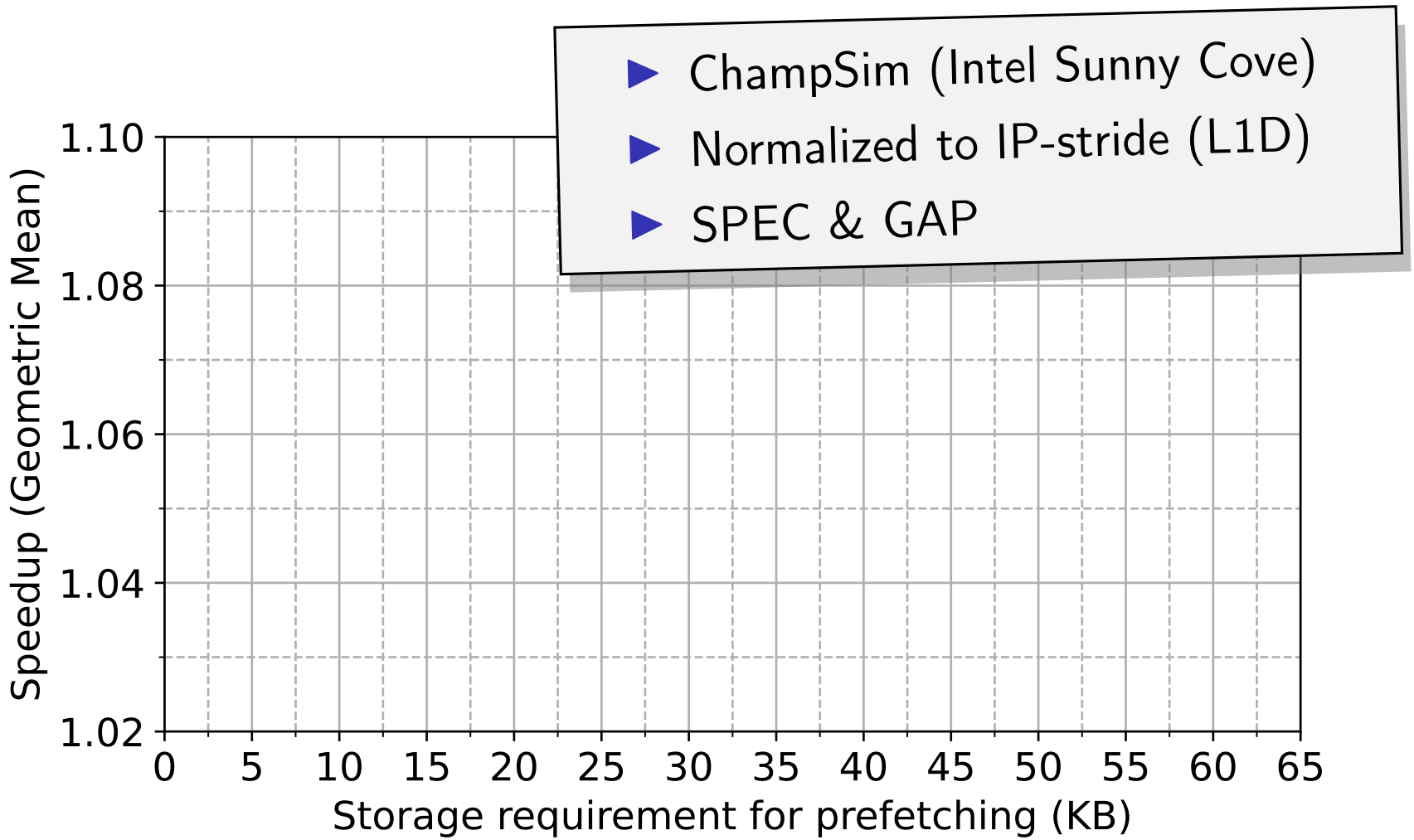
# Berti: Introduction

- Accurate and timely local delta L1D prefetcher
- Outperforms state-of-the-art prefetchers
- Only 2.55 KB of storage

# Berti: Introduction

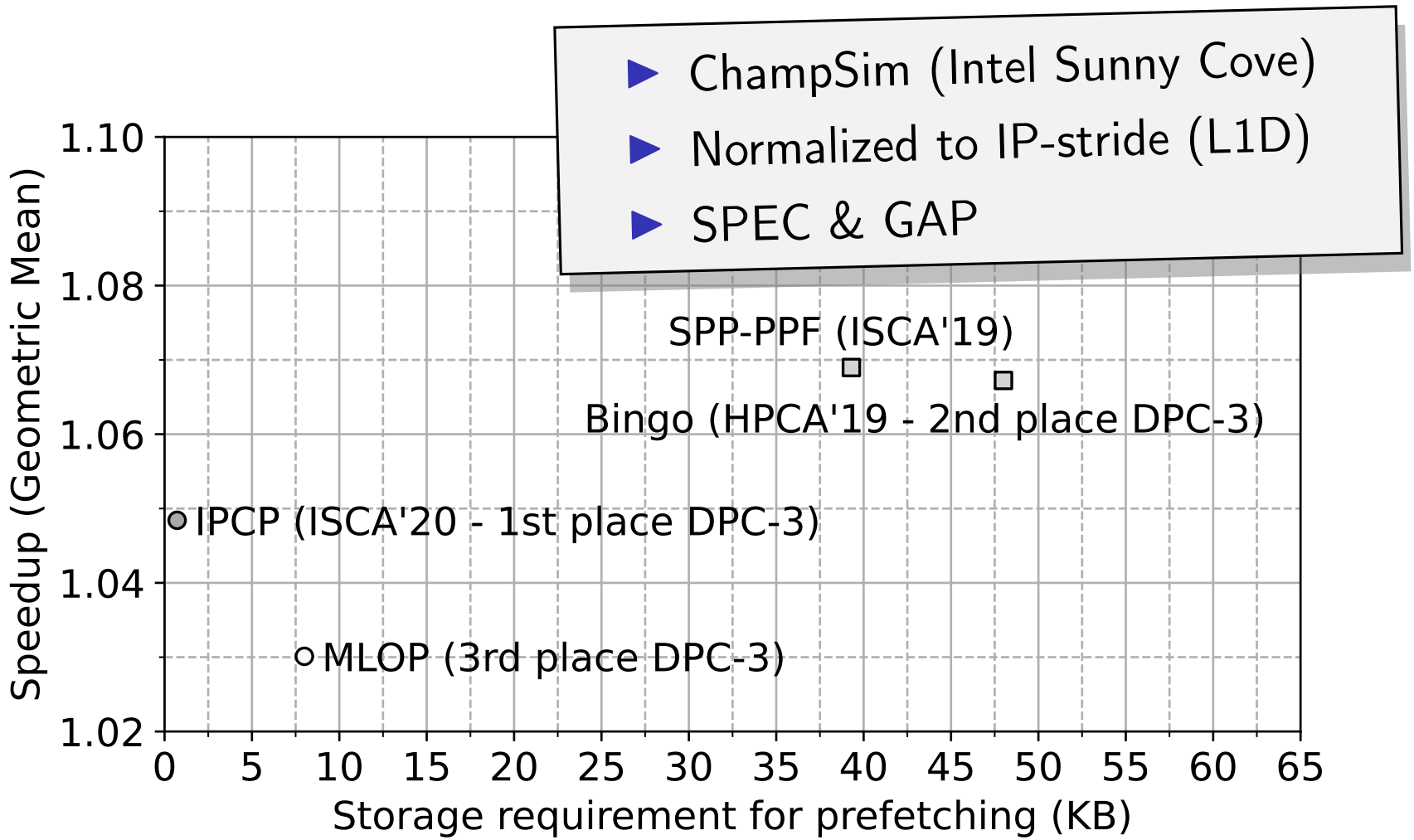
- Accurate and timely local delta L1D prefetcher
- **Outperforms state-of-the-art prefetchers**
- **Only 2.55 KB of storage**

# Berti: Evaluation




DPC-3 = 3rd Data Prefetching Championship

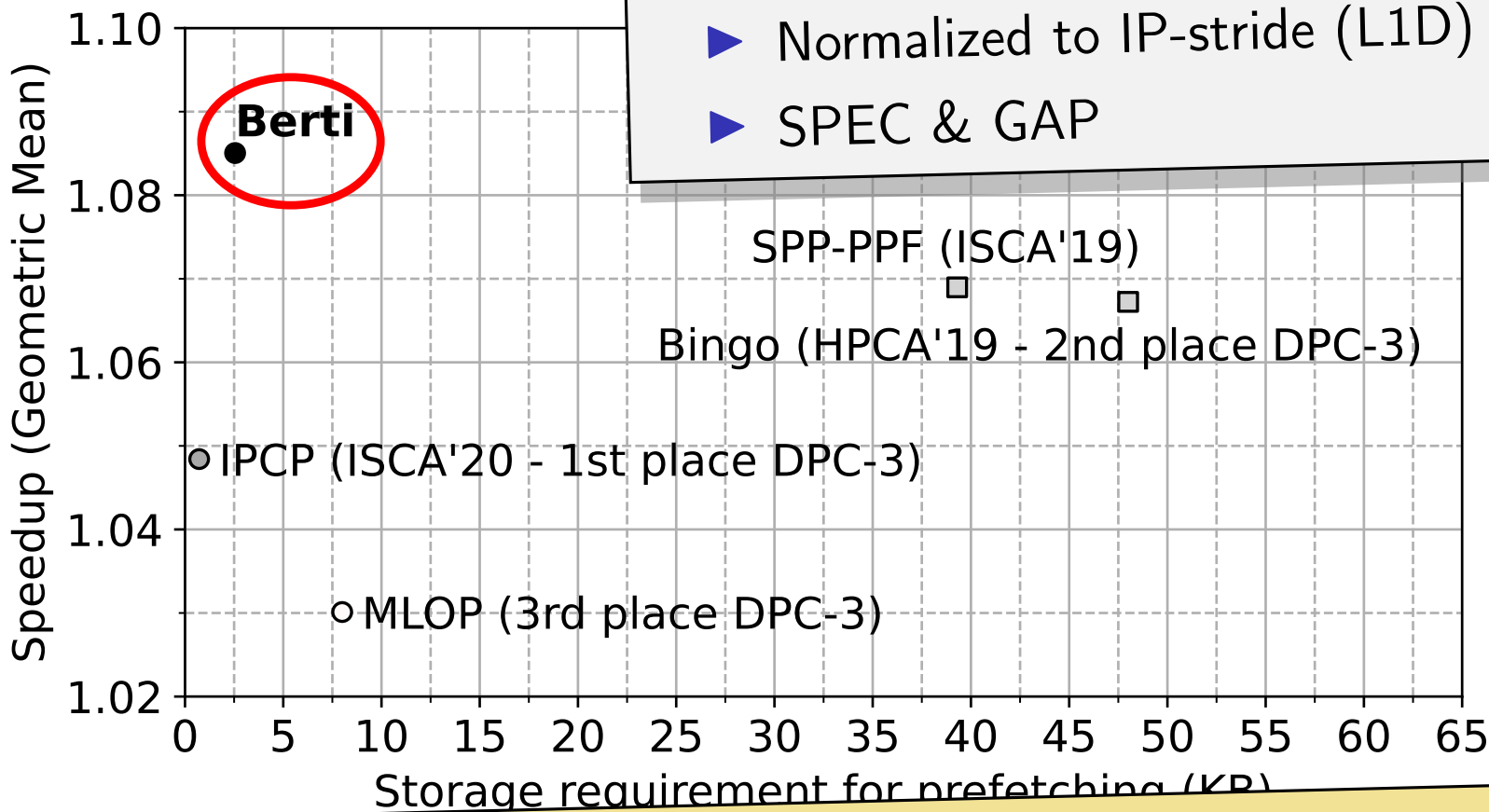
# Berti: Evaluation




DPC-3 = 3rd Data Prefetching Championship

# Berti: Evaluation

- ChampSim (Intel Sunny Cove)
- Normalized to IP-stride (L1D)
- SPEC & GAP

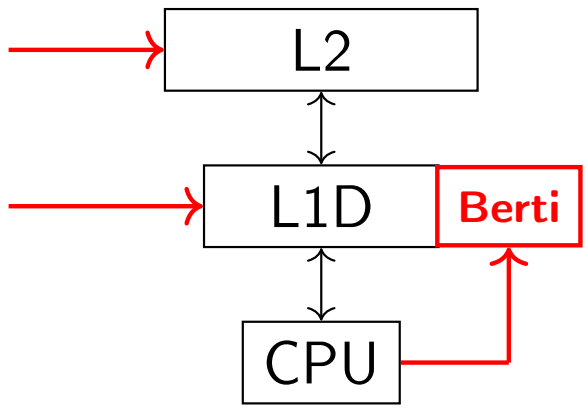



**Berti outperforms state-of-the-art prefetchers with less storage**

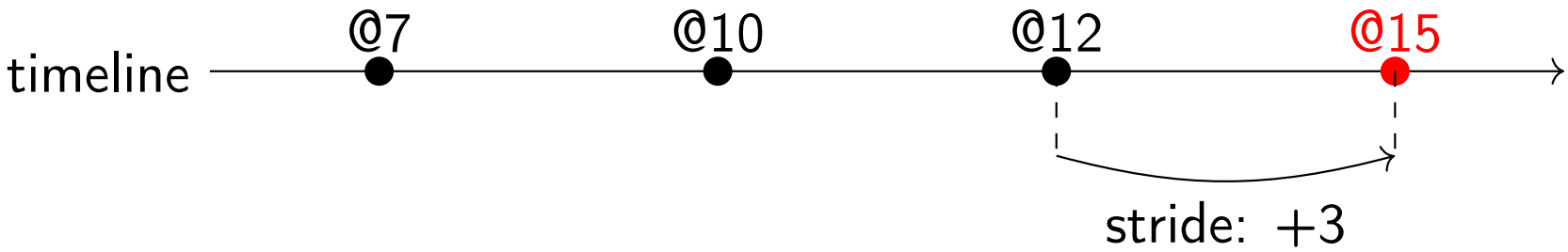
DPC-3 = 3rd Data Prefetching Championship

# Berti: Accurate and timely local delta L1D prefetcher

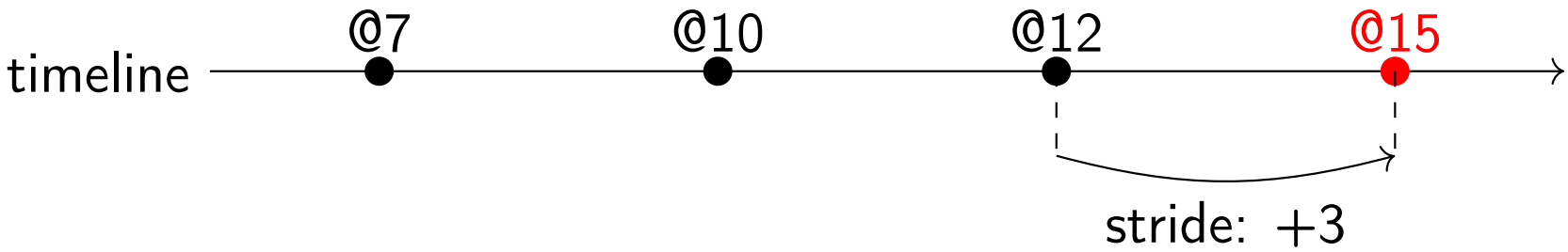
- Orchestrates prefetch requests across L1D/L2
- Instruction Pointer (IP) and unfiltered memory references
- Virtual addresses: cross-page prefetching

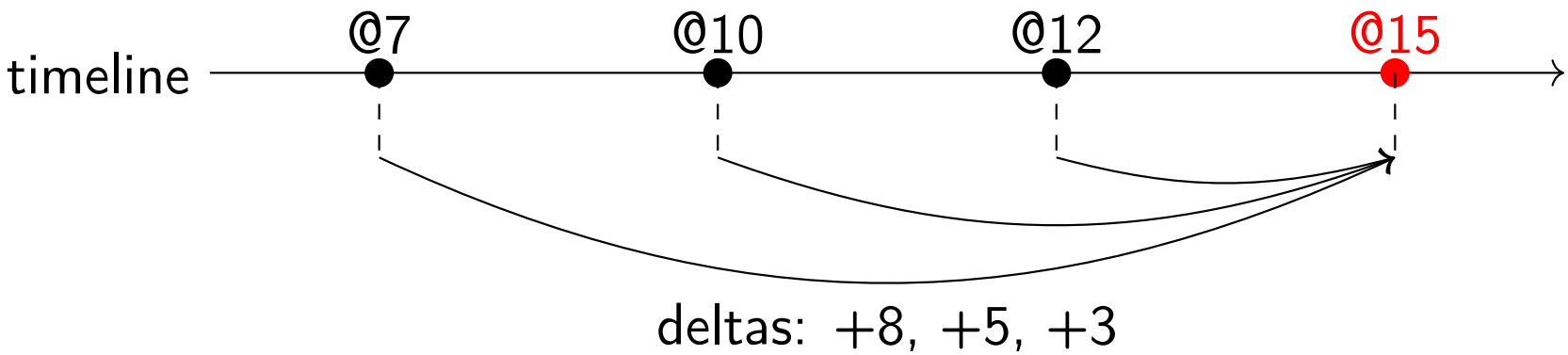


# Definition of delta

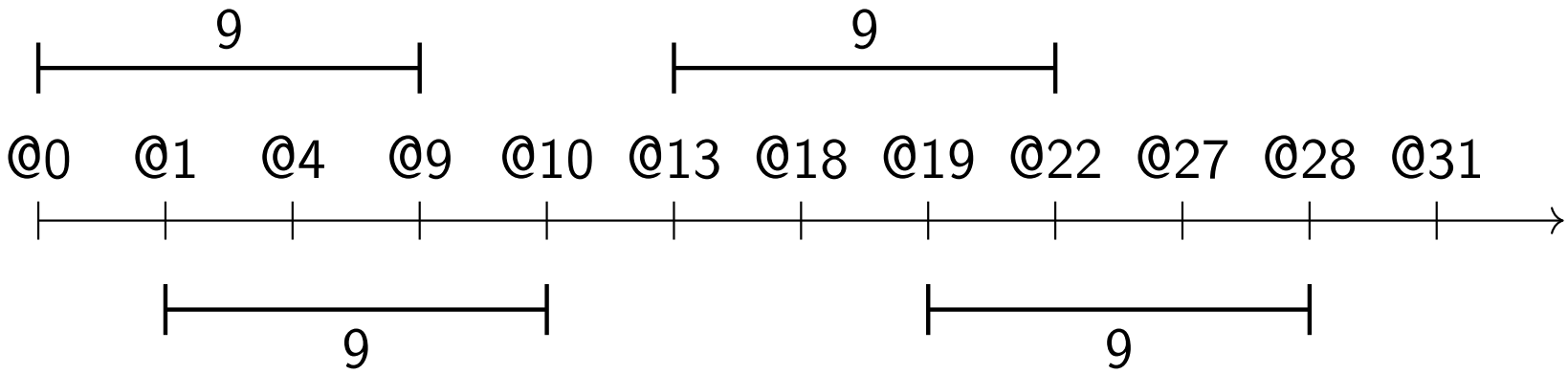


# Definition of delta



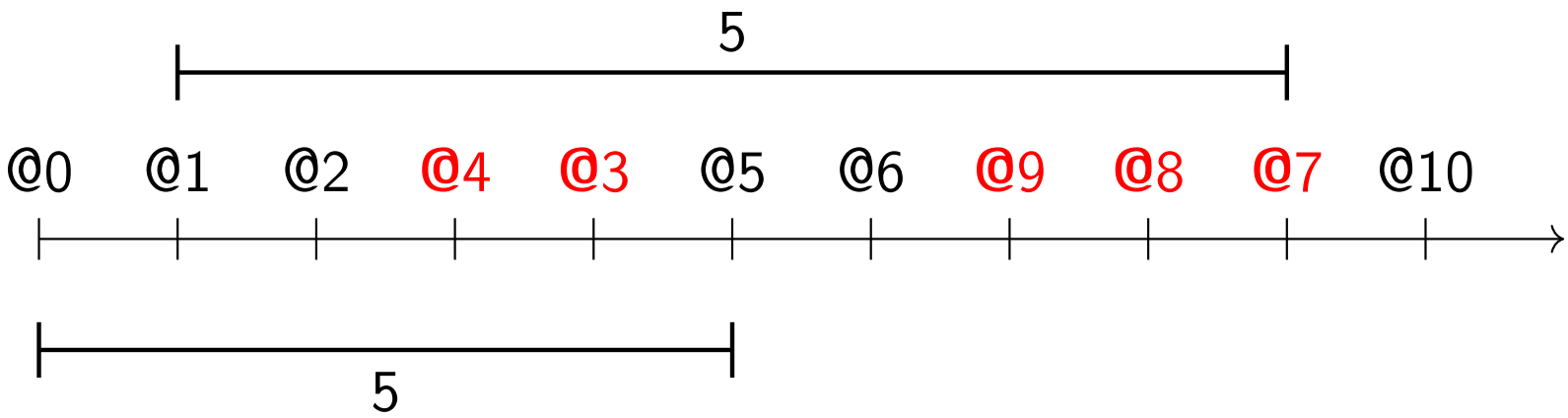
Stride: **+1, +3, +5**



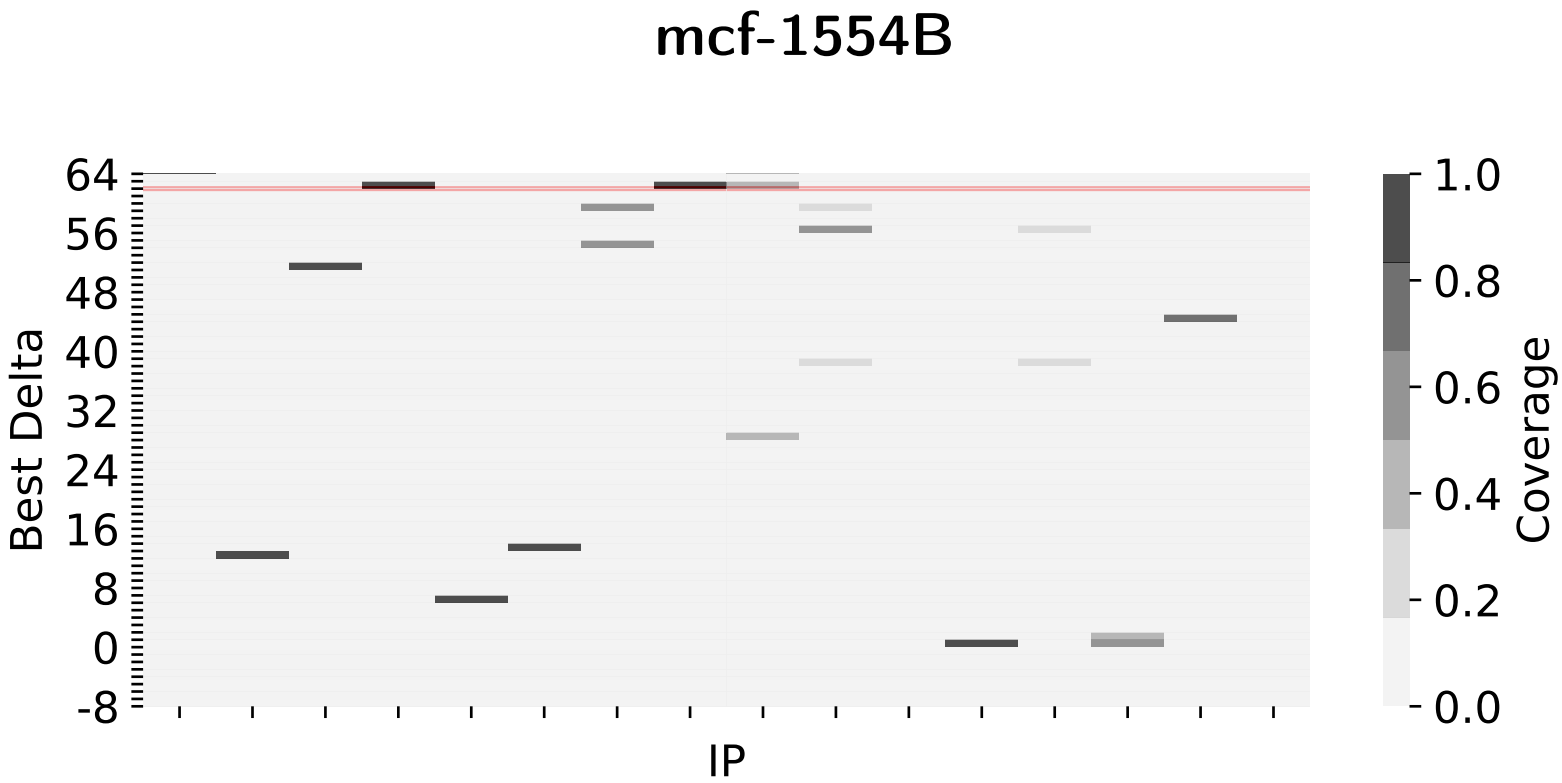

With which delta should I prefetch?

**delta = 1 + 3 + 5 = 9 → always hit**

Addresses reordered by **out-of-order** processor

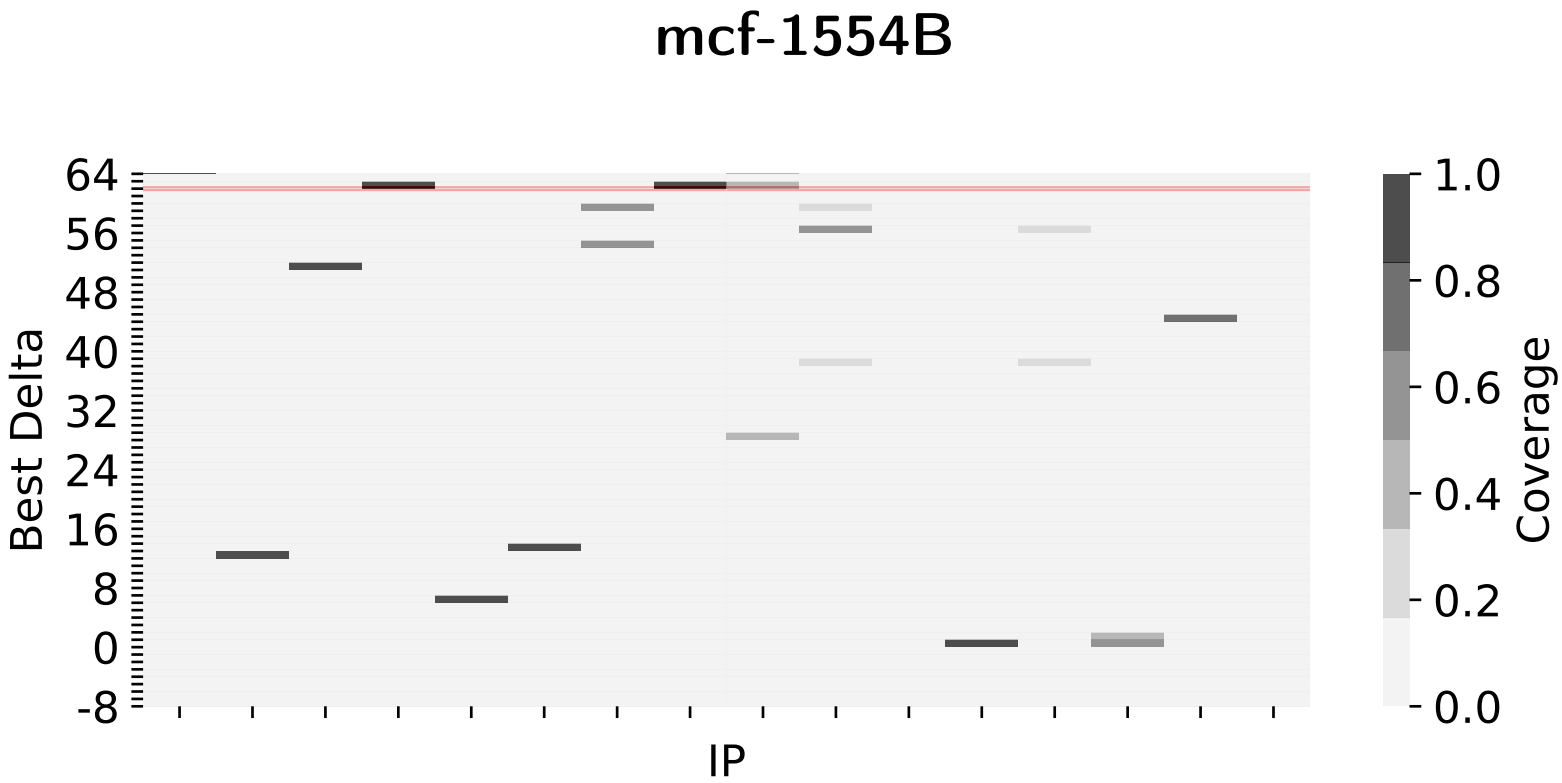



Stride prefetch requires specific order
**We can prefetch with delta = 5**, for example




- **Red line**: best delta by BOP[1], coverage: 2%
- **Black lines**: per-IP local deltas, coverage: 10%

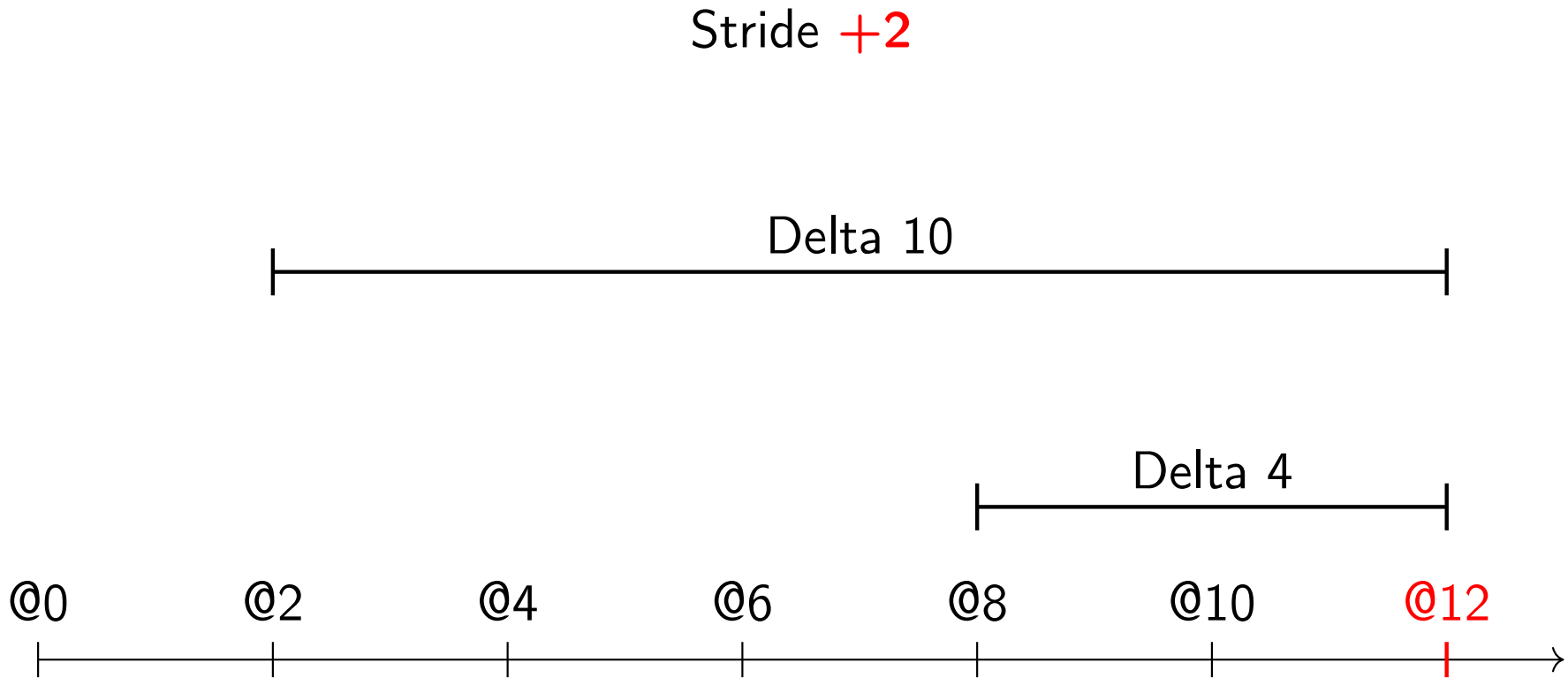
[1]Winner of 2nd Data Prefetching Championship (DPC2)



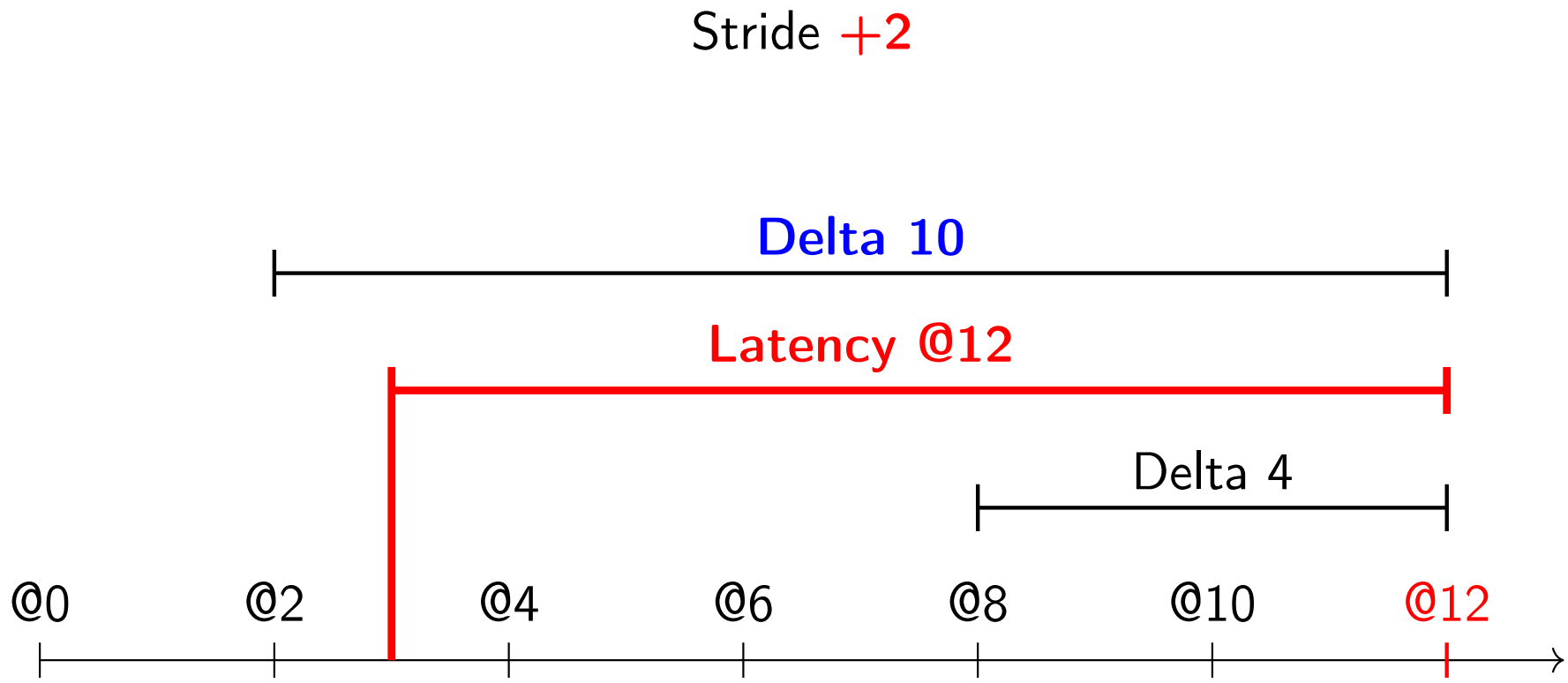

- **Red line**: best delta by I
- **Black lines**: per-IP local

**Every IP can have its own deltas!**

[1]Winner of 2nd Data Prefetching Championship (DPC2)




How far in advance should I prefetch address 12?




How far in advance should I prefetch address 12?
**Depends on its latency**

## Training

1. Measure fetch latency
2. Learn timely and accurate deltas
3. Compute coverage of deltas

**History table**

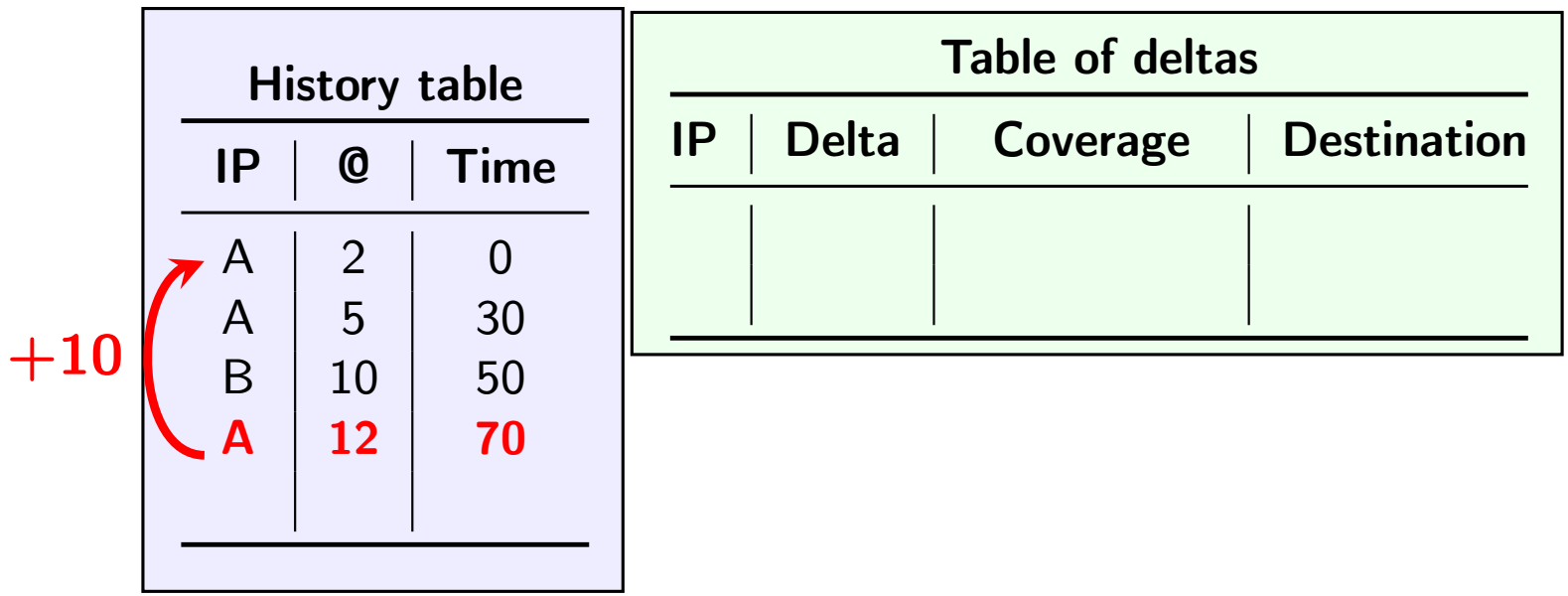
| IP | @ | Time |
|---|---|---|
| A | 2 | 0 |
| A | 5 | 30 |
| B | 10 | 50 |
| | | |

**Table of deltas**

| IP | Delta | Coverage | Destination |
|---|---|---|---|
| | | | |

## Training

1. Measure fetch latency
2. Learn timely and accurate deltas
3. Compute coverage of deltas



## Training

1. Measure fetch latency
2. Learn timely and accurate deltas
3. Compute coverage of deltas

**History table**

| IP | @ | Time |
|---|---|---|
| A | 2 | 0 |
| A | 5 | 30 |
| B | 10 | 50 |
| **A** | **12** | **70** |

**Table of deltas**

| IP | Delta | Coverage | Destination |
|---|---|---|---|
| **A** | **+10** | **1/1 (100%)** | |

## TRAINING

1. Measure fetch latency
2. Learn timely and accurate deltas
3. Compute coverage of deltas

+10, +13

**History table**

| IP | @ | Time |
|---|---|---|
| A | 2 | 0 |
| A | 5 | 30 |
| B | 10 | 50 |
| A | 12 | 70 |
| **A** | **15** | **140** |

**Table of deltas**

| IP | Delta | Coverage | Destination |
|---|---|---|---|
| **A** | **+10** | **2/2 (100%)** | |
| **A** | **+13** | **1/2 (50%)** | |

## Issuing prefetch requests

1. Select deltas
2. Orchestration

**History table**

| IP | @ | Time |
|---|---|---|
| A | 2 | 0 |
| A | 5 | 30 |
| B | 10 | 50 |
| A | 12 | 70 |
| A | 15 | 140 |

**Table of deltas**

| IP | Delta | Coverage | Destination |
|---|---|---|---|
| A | +10 | 2/2 (100%) | |
| A | +13 | 1/2 (50%) | |

## Issuing prefetch requests

1. Select deltas
2. Orchestration

**History table**

| IP | @ | Time |
|---|---|---|
| A | 2 | 0 |
| A | 5 | 30 |
| B | 10 | 50 |
| A | 12 | 70 |
| A | 15 | 140 |

**Table of deltas**

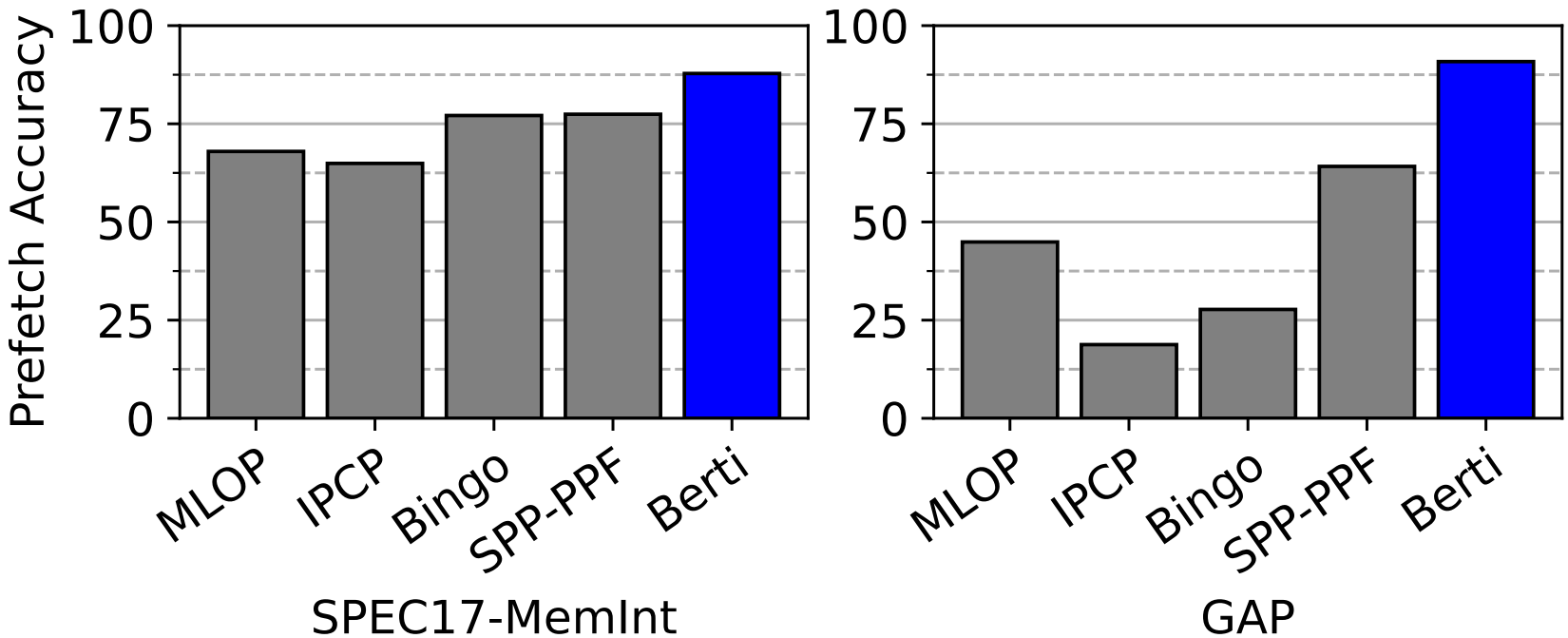
| IP | Delta | Coverage | Destination |
|---|---|---|---|
| A | +10 | 2/2 (100%) | L1D |
| A | +13 | 1/2 (50%) | L2 |

*Coverage*

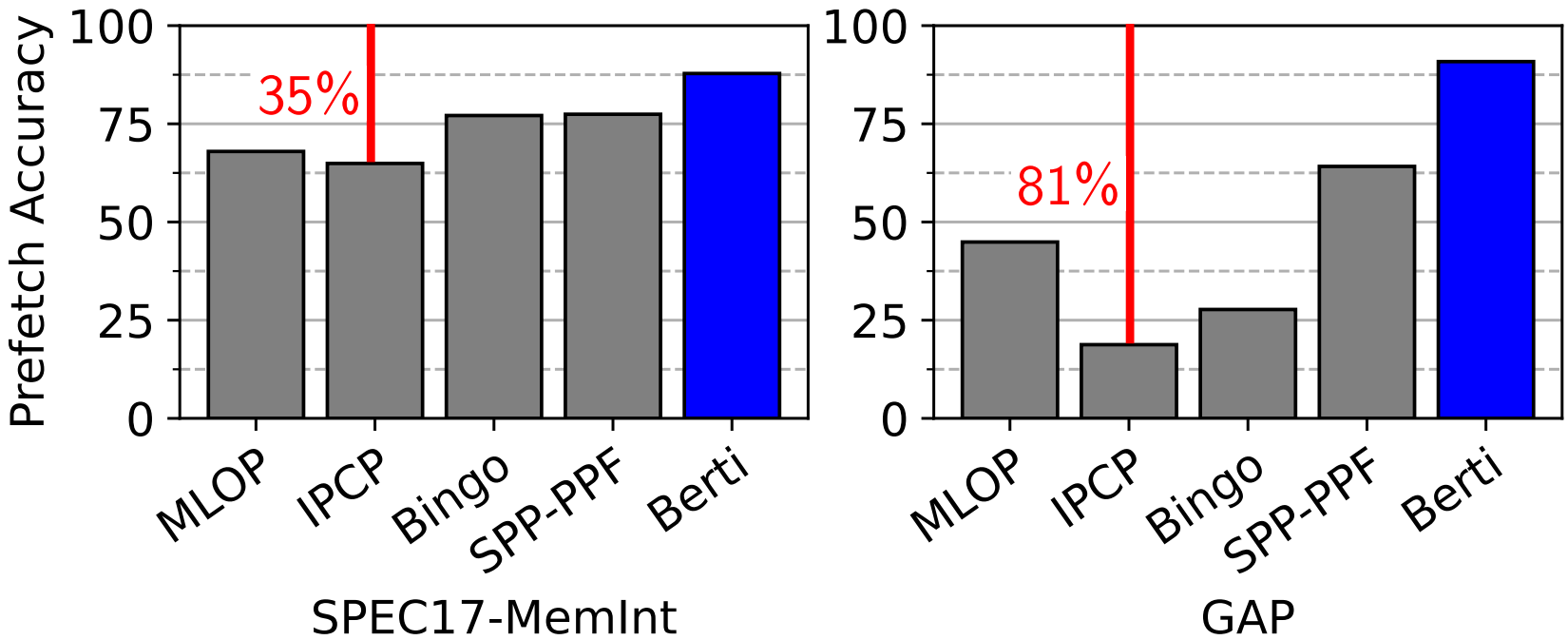
$> 65\% \rightarrow L1D$

$> 35\% \rightarrow L2$

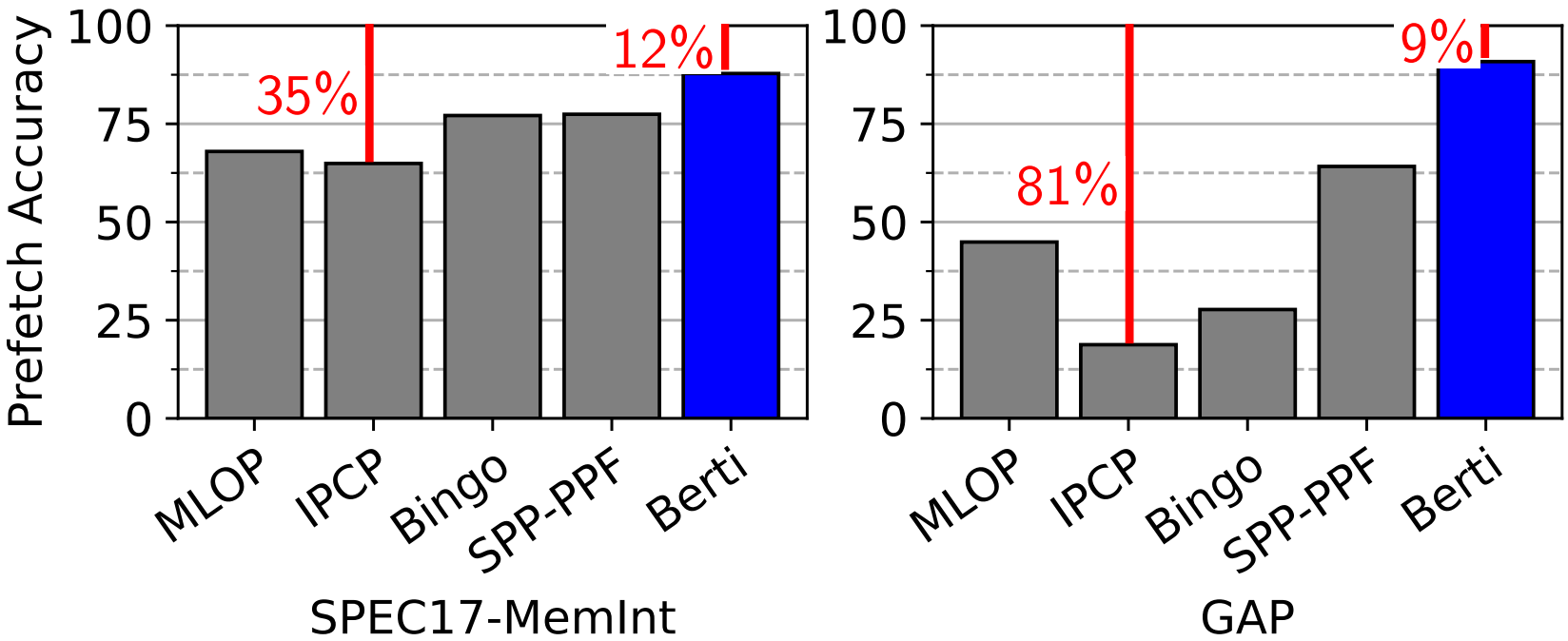
# Berti: Accurate and timely local delta L1D prefetcher



# Berti: Accurate and timely local delta L1D prefetcher



# Berti: Accurate and timely local delta L1D prefetcher




**Improved accuracy reduces BW overhead and L1D pollution**

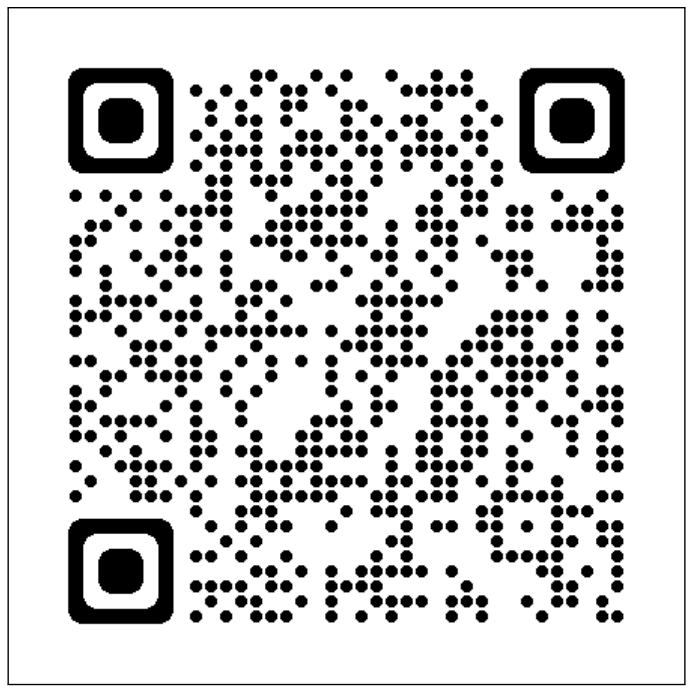
# Berti: Conclusions

- **Accurate** and **timely local delta L1D** prefetcher
- Outperforms state-of-the-art prefetchers
- Learns the best **deltas** to prefetch
- Artifact available (Webpage QR)

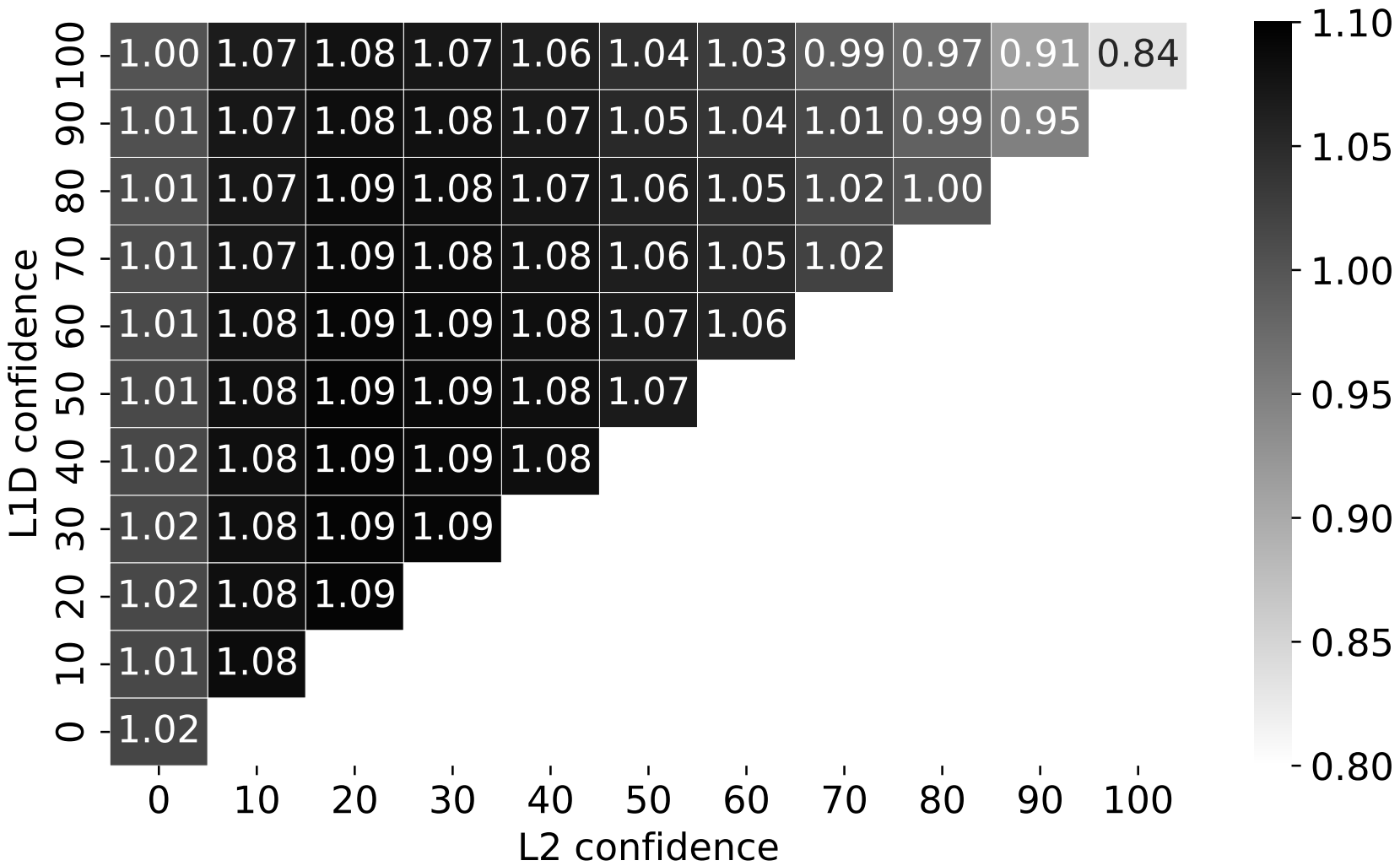
# Thanks!

# Questions?

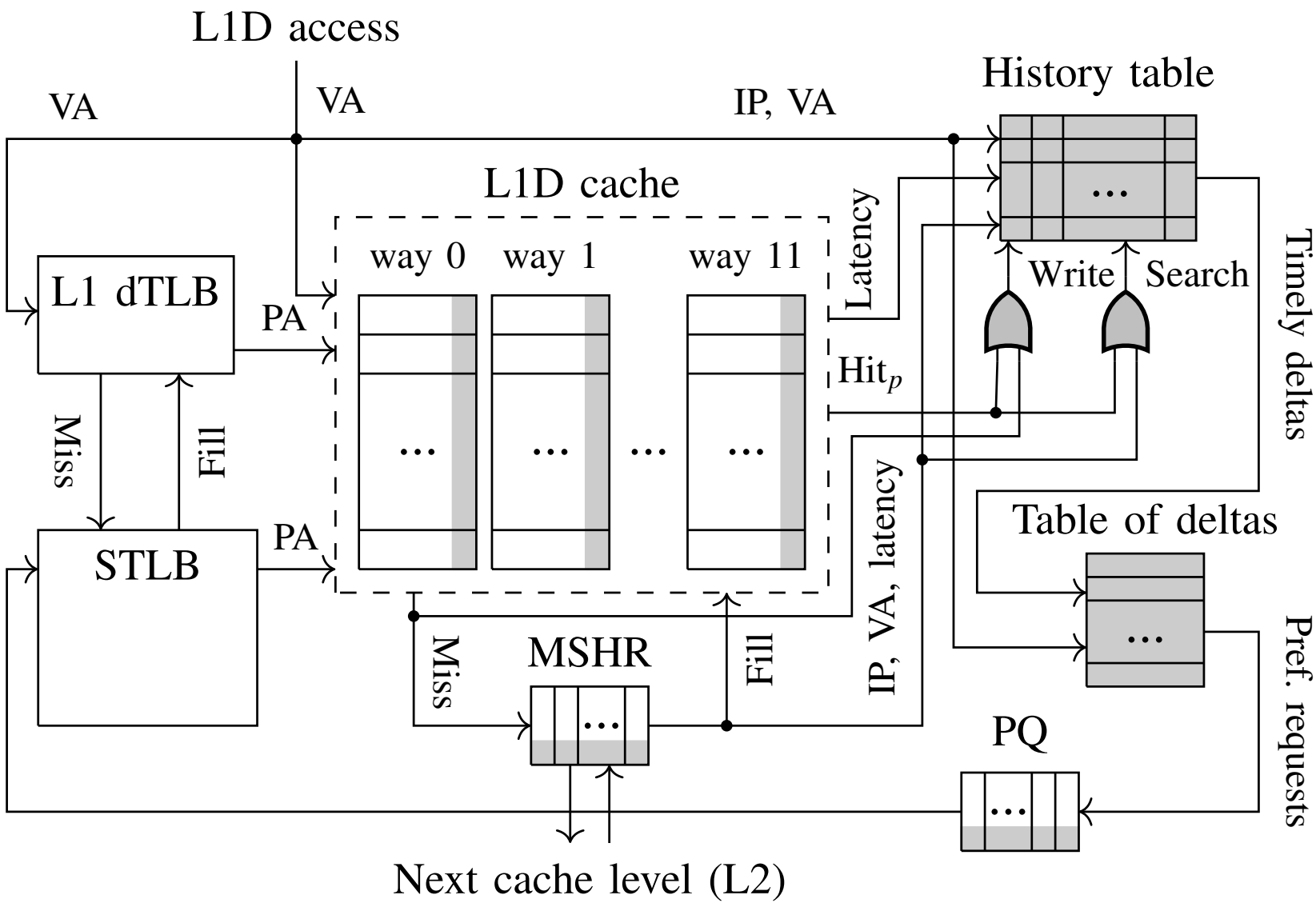
agusnt@unizar.es



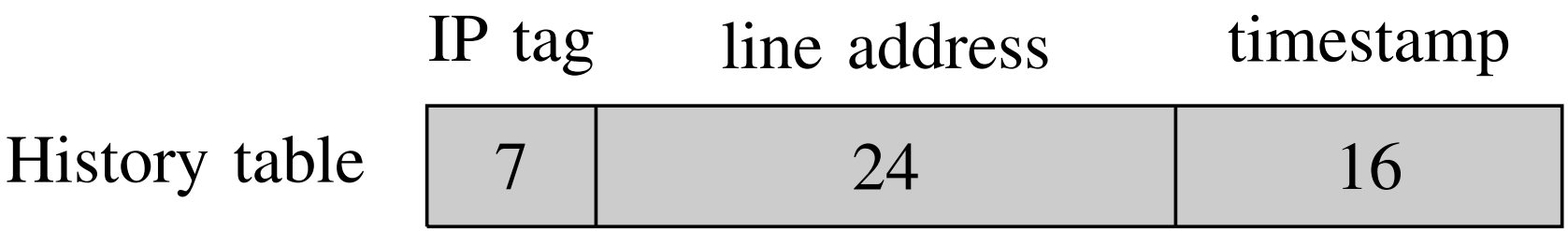
# Watermark Sensitivity



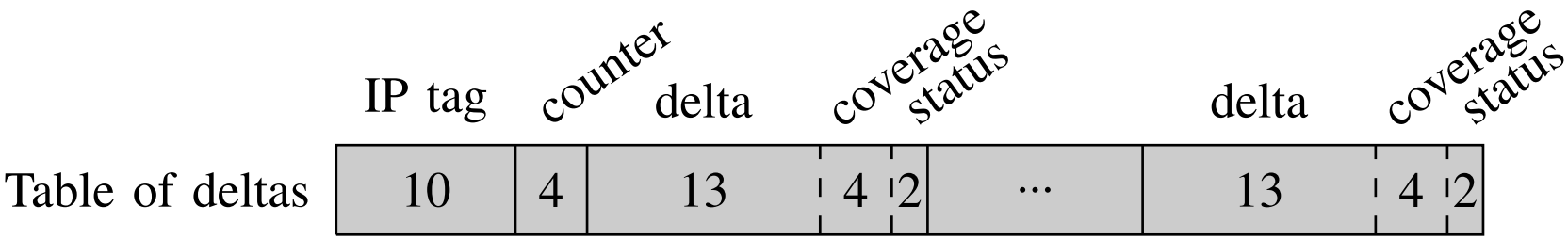
# Hardware Implementation



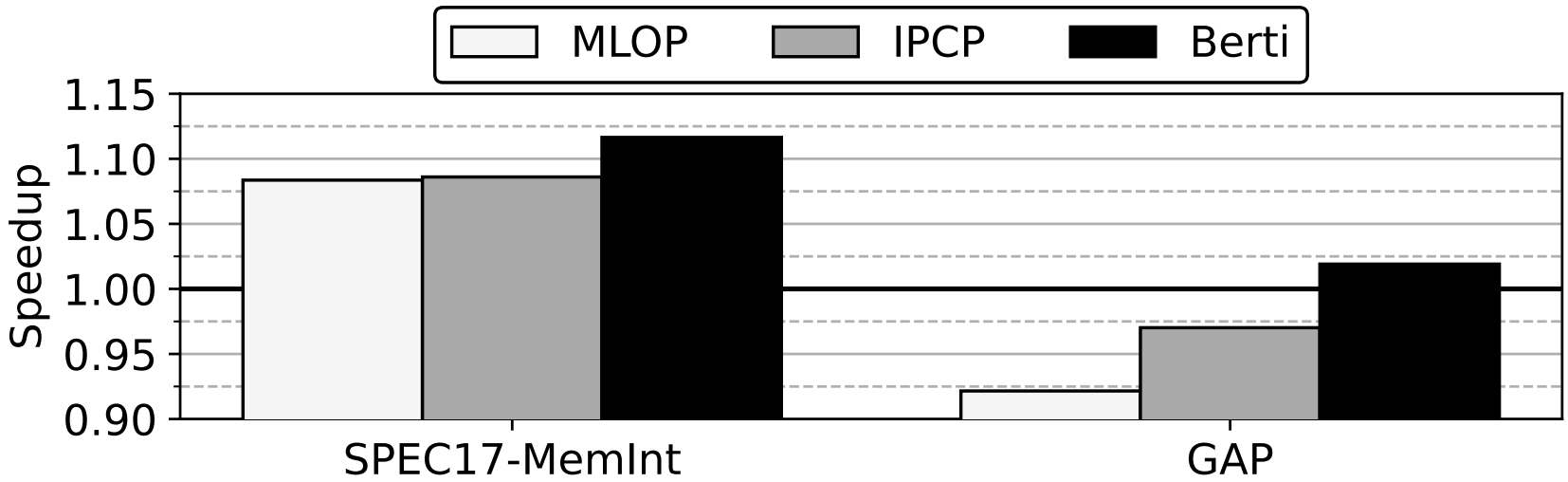
# History Table



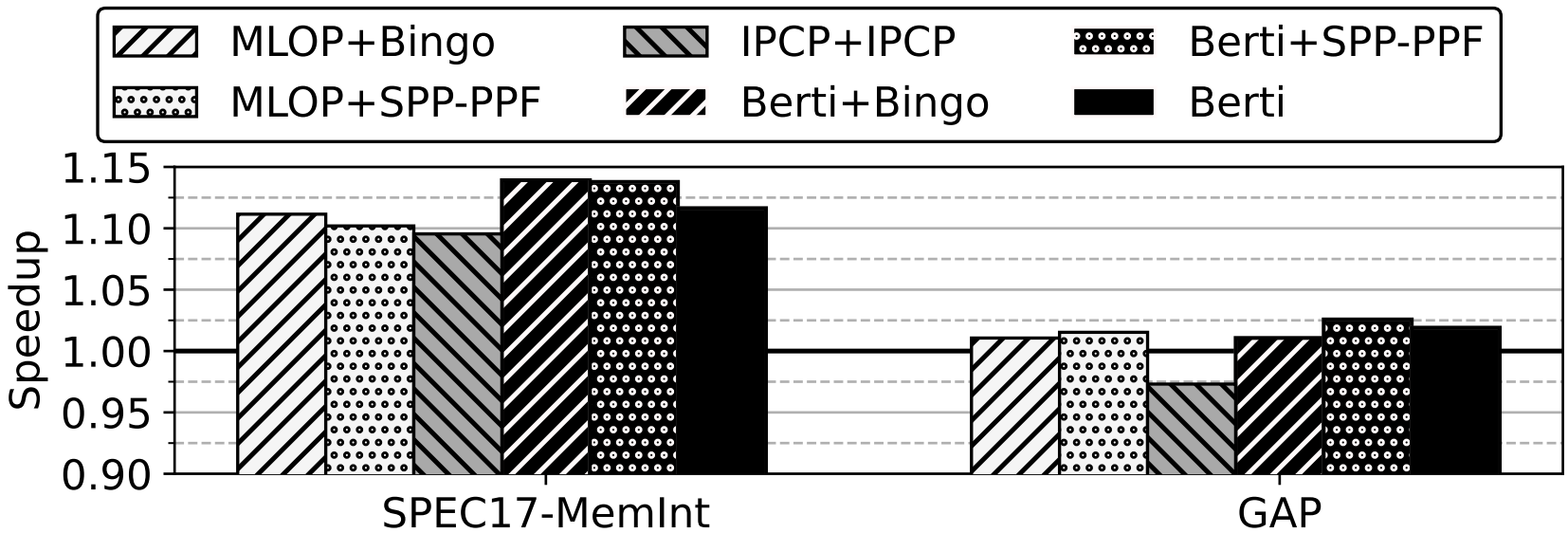
# Delta Table


IP tag
counter
delta
coverage
status
delta
coverage
status
Table of deltas
10
4
13
4
2
···
13
4
2

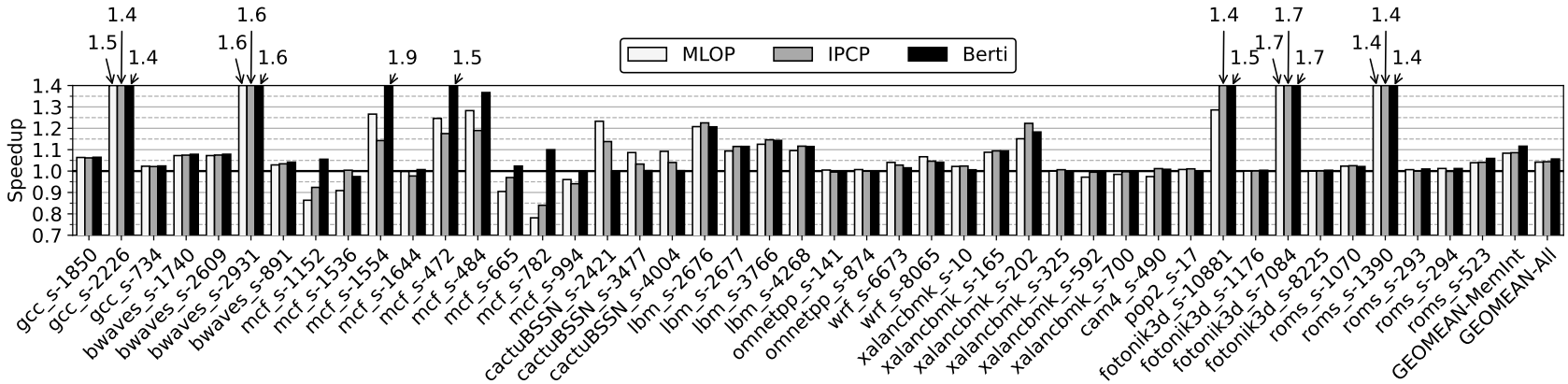
# L1D Prefetchers Performance



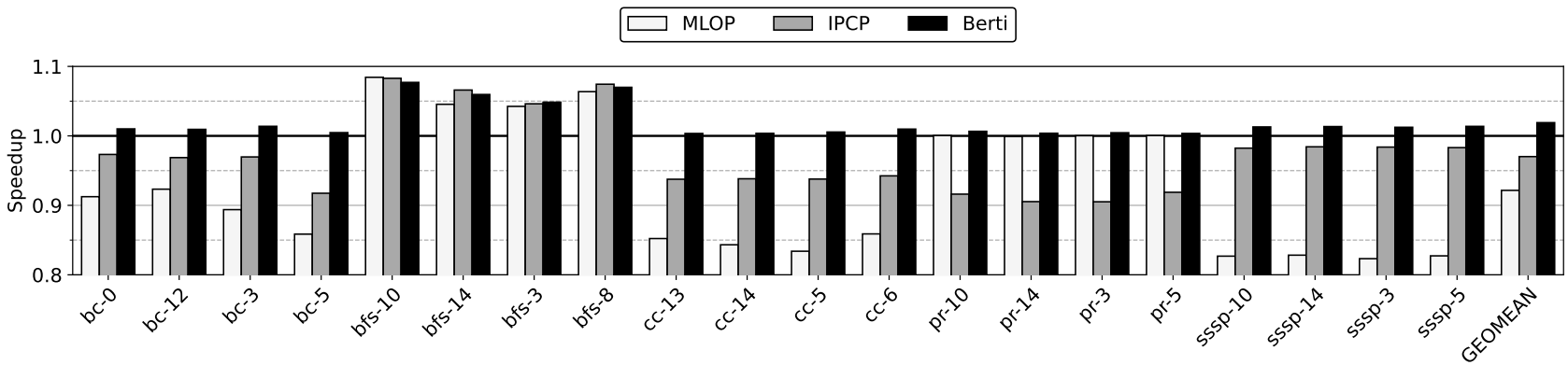
# Multi-level Prefetchers Performance



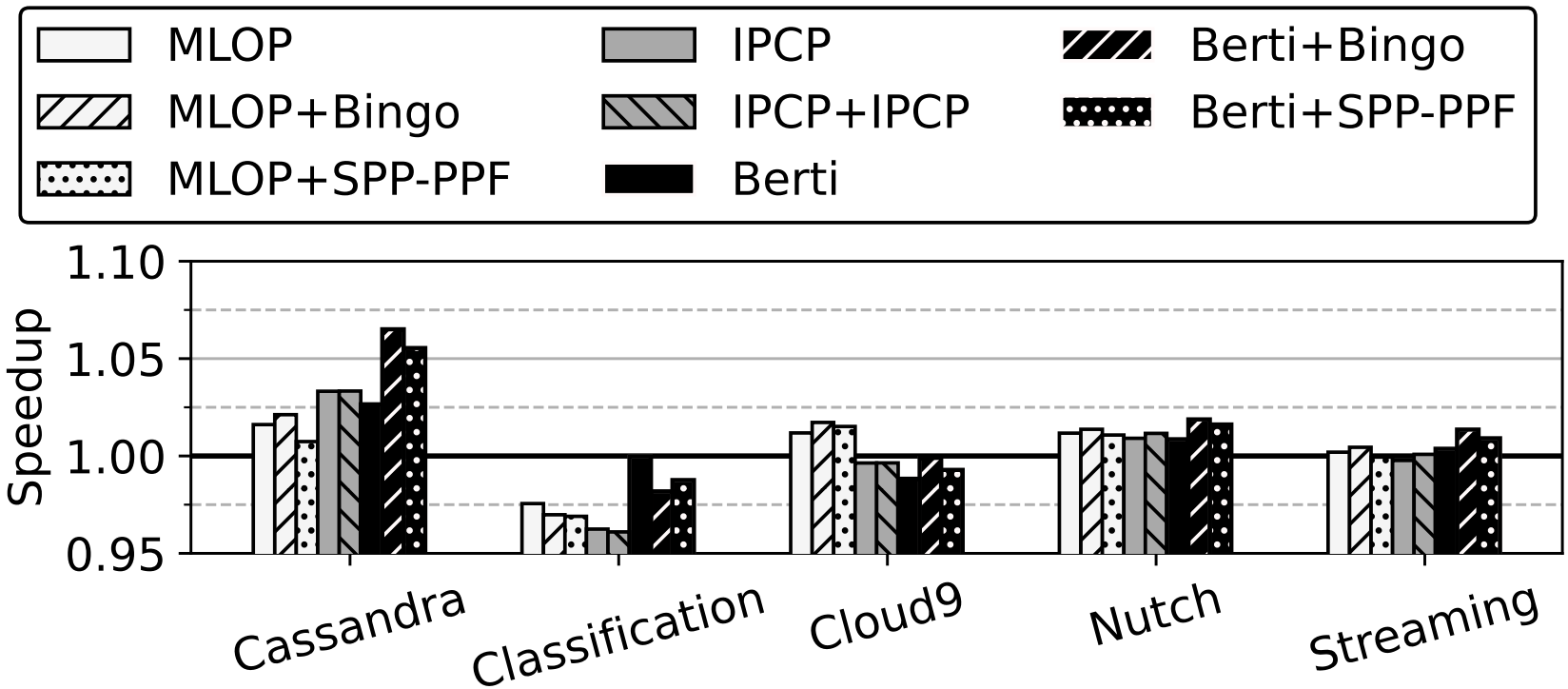
# L1D Prefetchers SPEC Performance



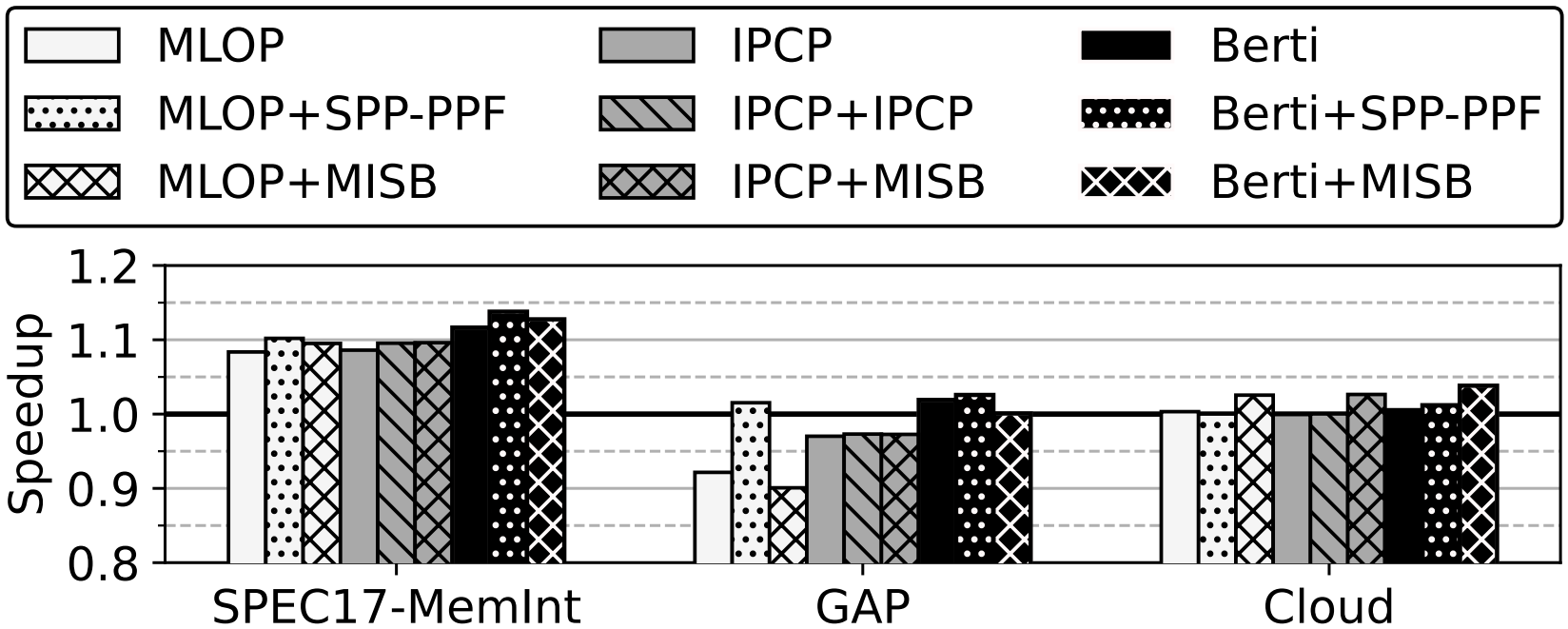
# L1D Prefetchers GAP Performance



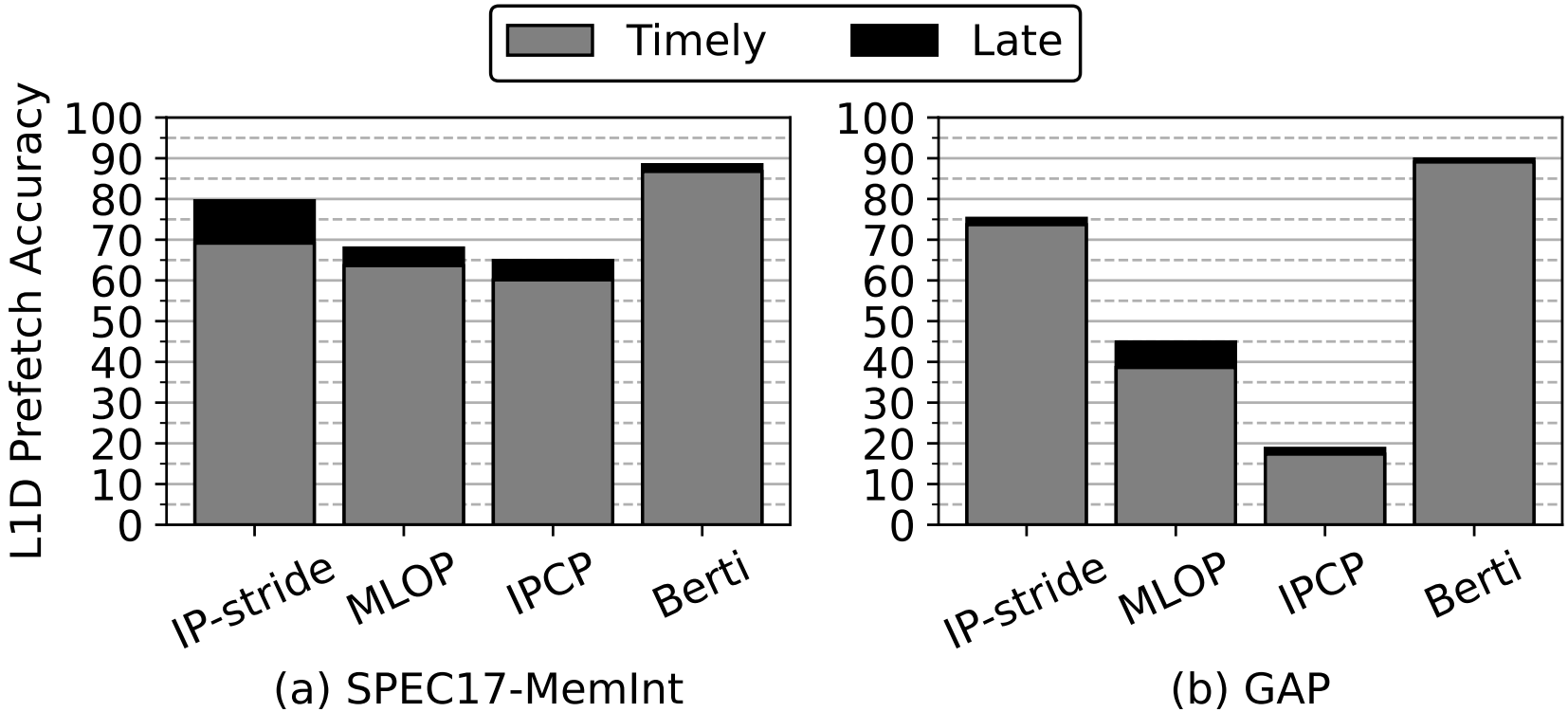
# CloudSuite Performance



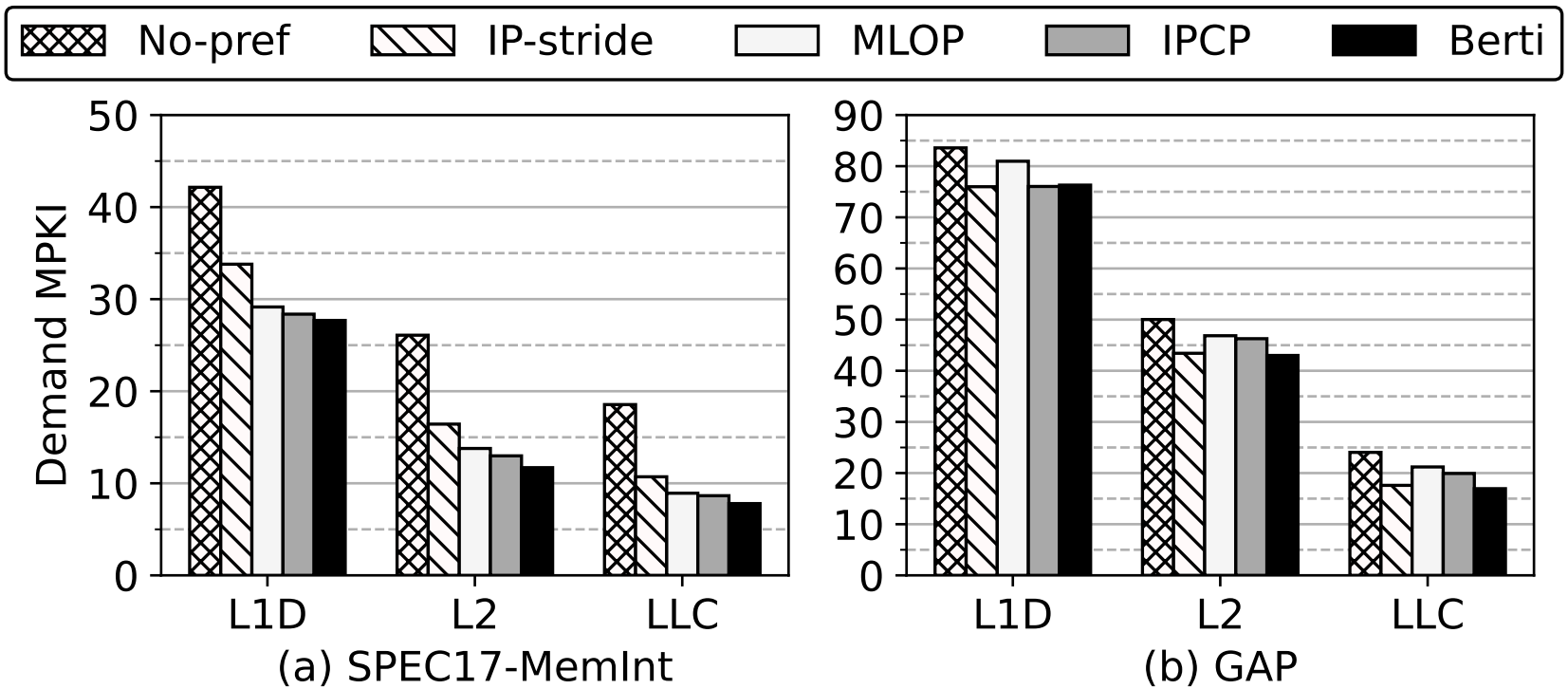
# ISB Performance



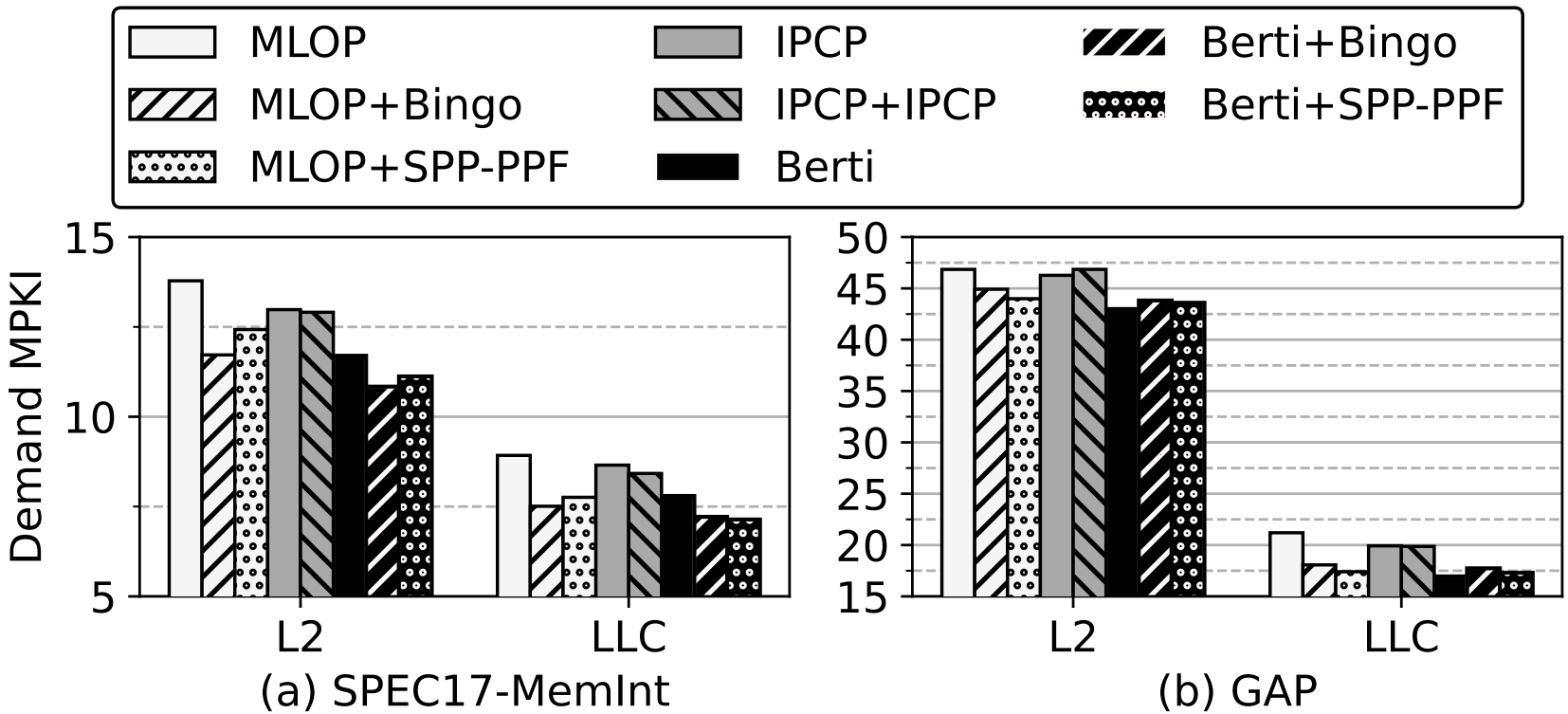
# L1D Prefetchers Accuracy



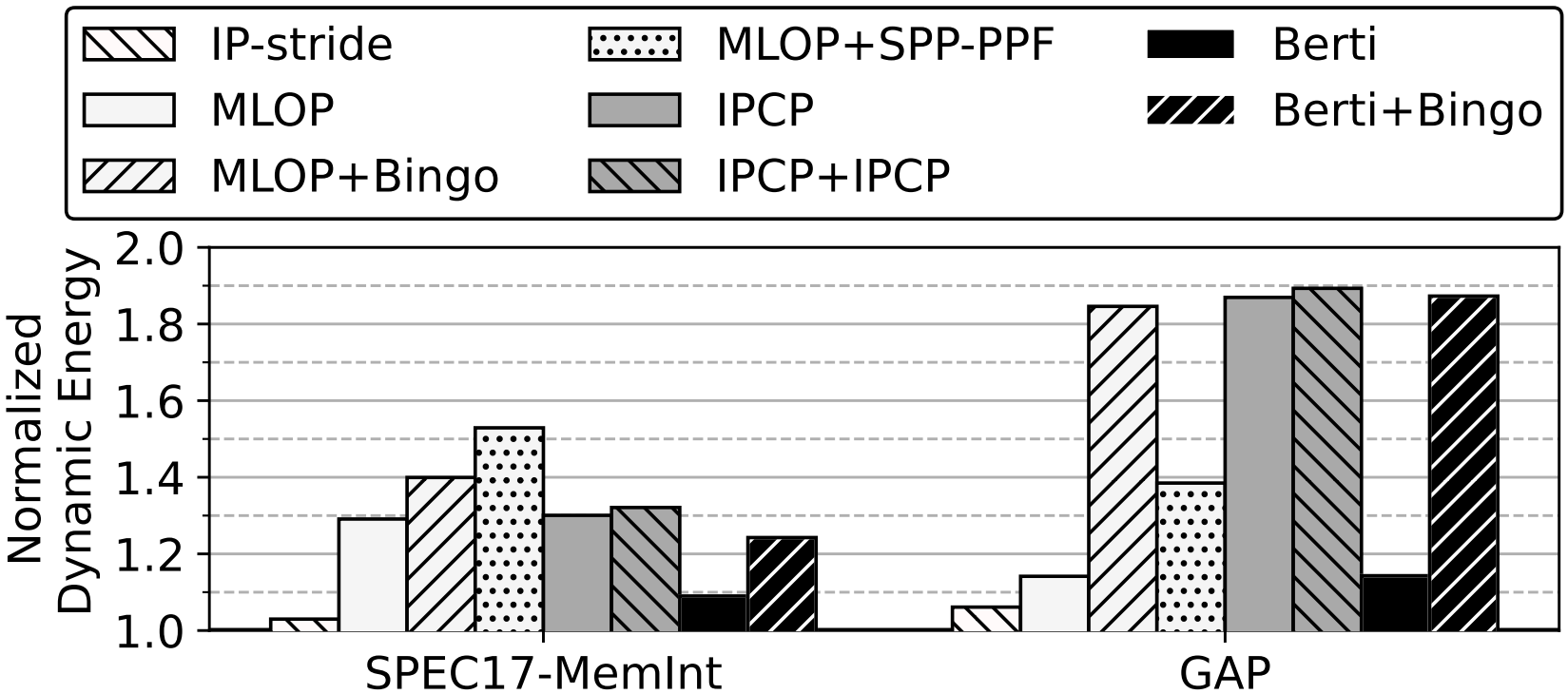
# L1D Prefetchers Demand MPKI



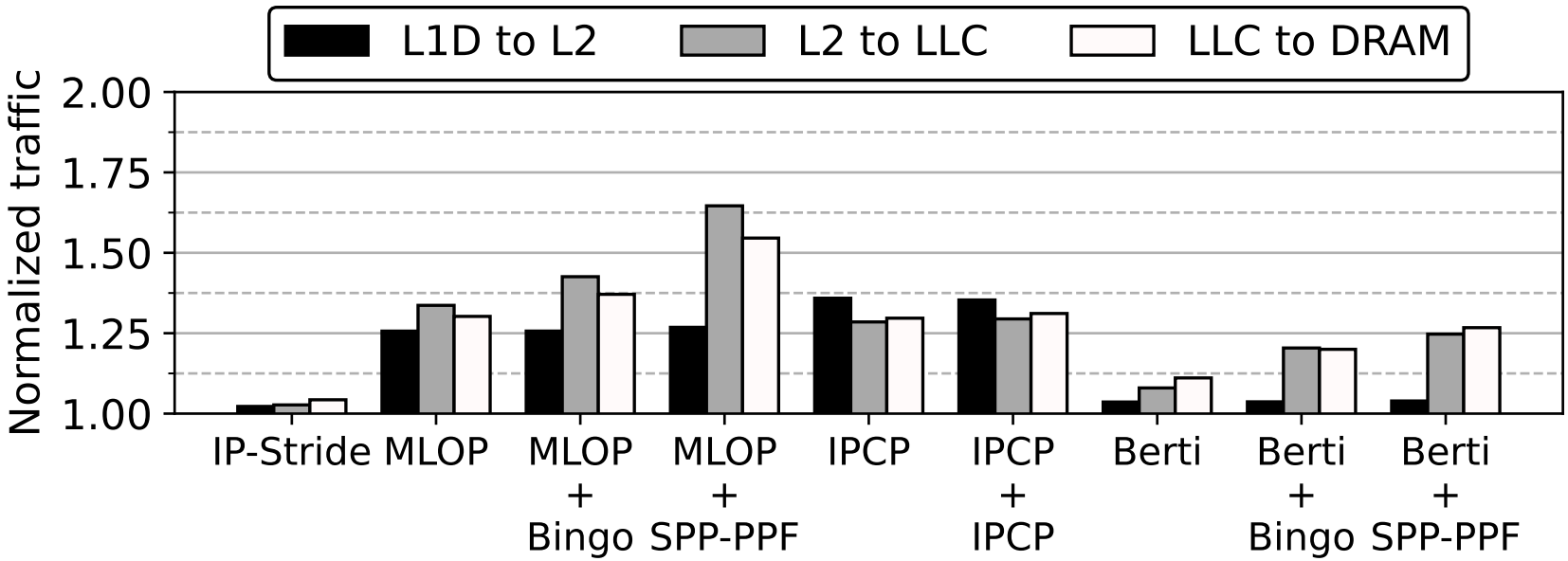
# Multi-Level Prefetcher Demand MPKI



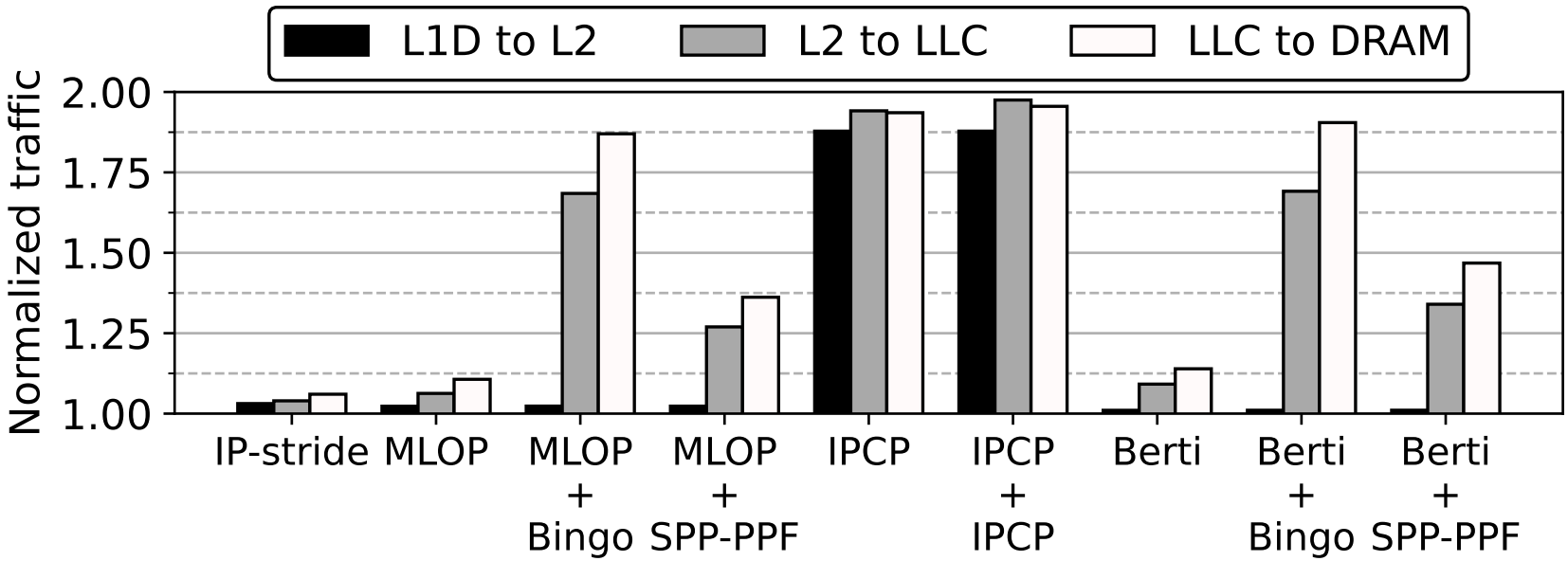
# Energy


IP-stride
MLOP
MLOP+Bingo
MLOP+SPP-PPF
IPCP
IPCP+IPCP
Berti
Berti+Bingo
Normalized Dynamic Energy
2.0
1.8
1.6
1.4
1.2
1.0
SPEC17-MemInt
GAP

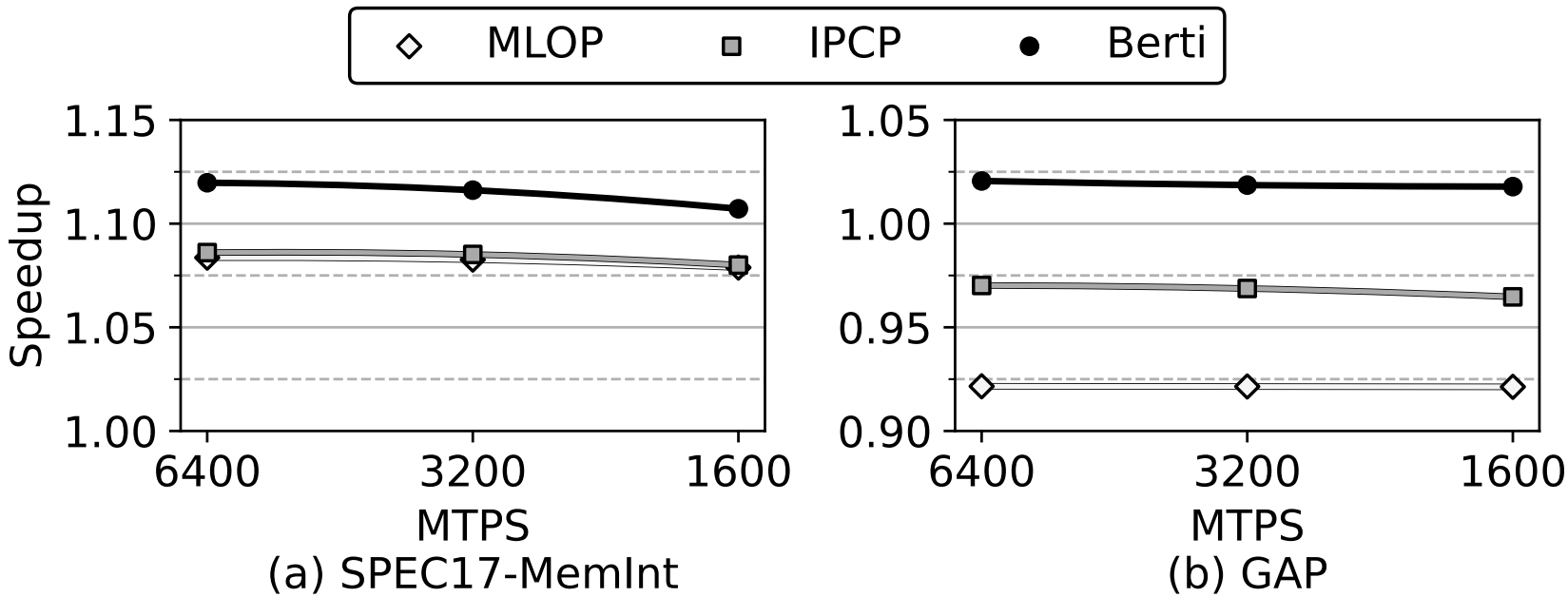
# SPEC Normalized Traffic



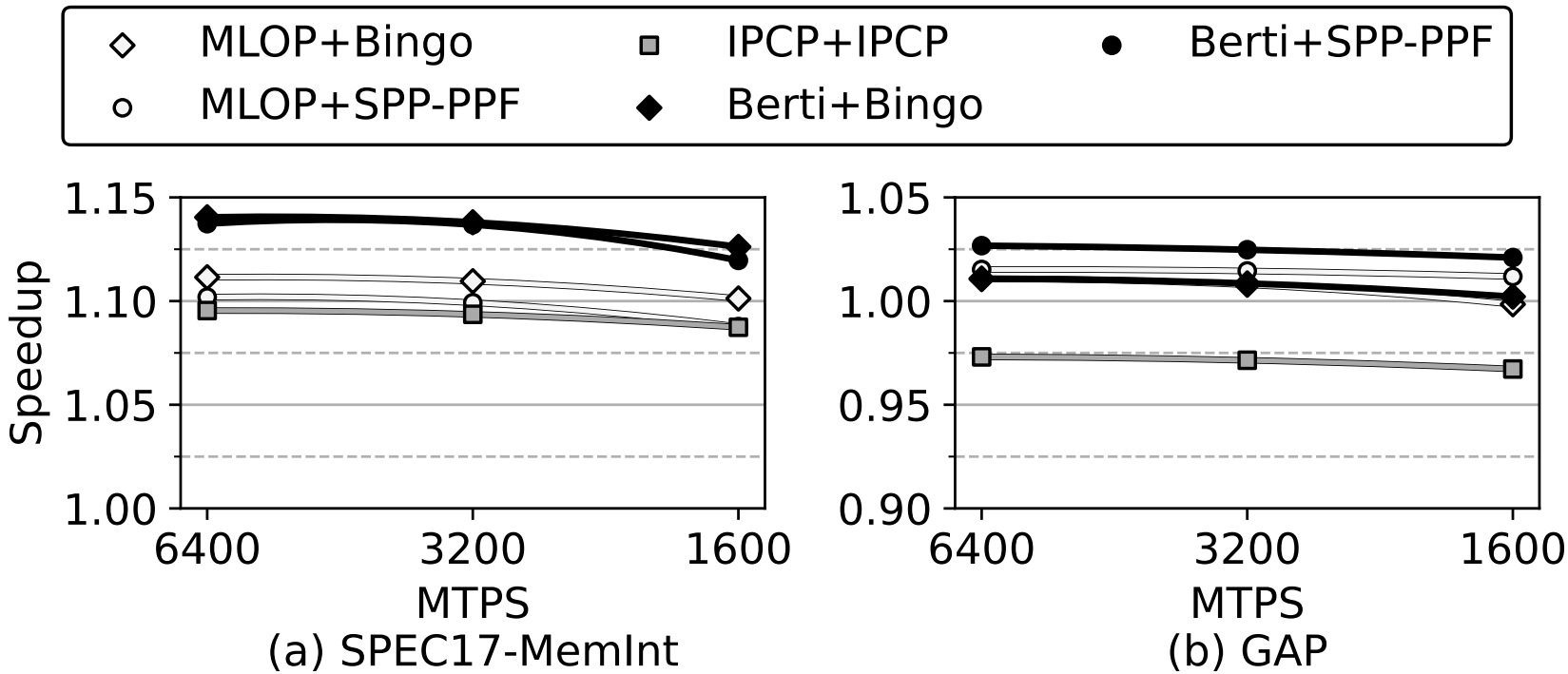
# GAP Normalized Traffic



# L1D Prefetcher Low Bandwidth



# Multi-level Prefetcher Low Bandwidth



# Performance vs. Size