GENERACIÓN
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Financiado por la Unión Europea
NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
Plan de Recuperación, Transformación y Resiliencia

AI4S
BSC AI 4 Science Fellowships

Enhancing Instruction Prefetching via Cache and TLB Management

Alexandre Valentin Jamet[1], *Georgios Vavouliotis*[2], *Martí Torrents*[1],
Dimitrios Chasapis[1], *Marc Casas*[1]

[1]*Barcelona Supercomputing Center*, [2]*Computing Systems Lab, Huawei Technologies AG, Switzerland*

alexandre.jamet@bsc.es

Executive Summary

- Contemporary server workloads have massive instruction footprint
  - Significant pressure on front-end structures
- Mitigation techniques:
  - **L1i prefetching** (*e.g.*, FNL+MMA, Barça, EPI, FDiP)
  - **TLB prefetching** (*e.g.*, Morrigan)
  - **Prefetch-aware cache replacement policy** (*e.g.*, PACIPV, PACMAN)

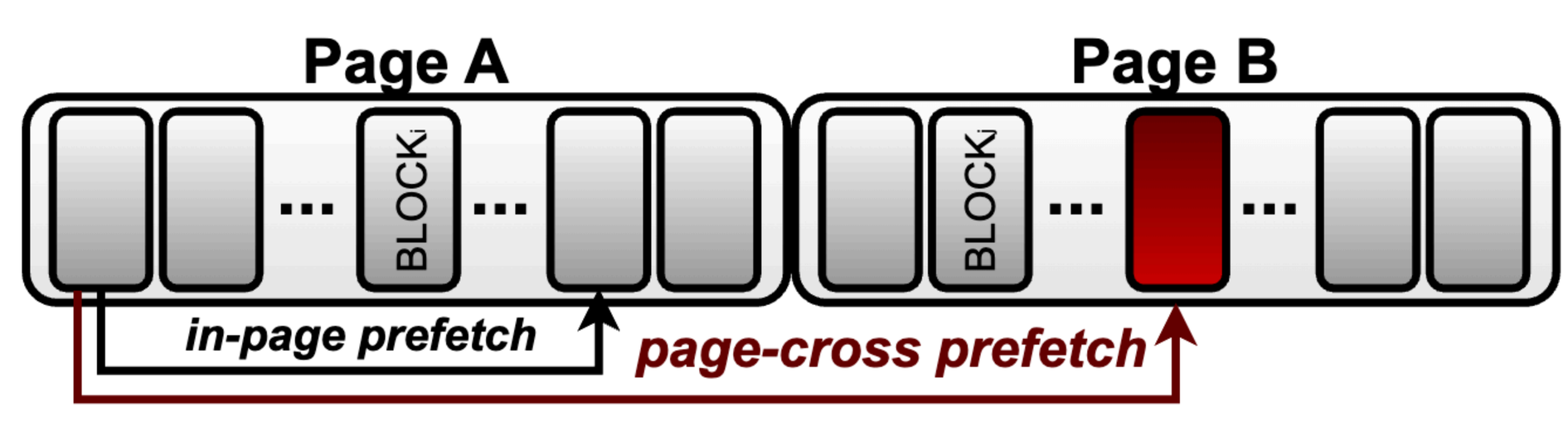
- **Limitations** in **L1i page-cross prefetching** mechanisms:
  - Address translation latency becomes a bottleneck
  - Insufficient utilization of lines brought by the L1i prefetcher in the L2C
- **Our approach**
  - Extend the sTLB with a buffer dedicated to translations brought by L1i prefetchers
  - Adapt the L2C replacement to accommodate lines brought by L1i prefetchers

Front-End Bottlenecks

- Contemporary server workloads exhibit large instruction footprints
  - Footprint growing by ~30% per year
  - Front-end structures (*i.e.*, L1i, TLBs) lag behind the demand of these workloads
- More PTEs required to map the instruction footprint

L1i and sTLB misses dominate front-end stalls. Around 10% of total execution cycles in server workloads.

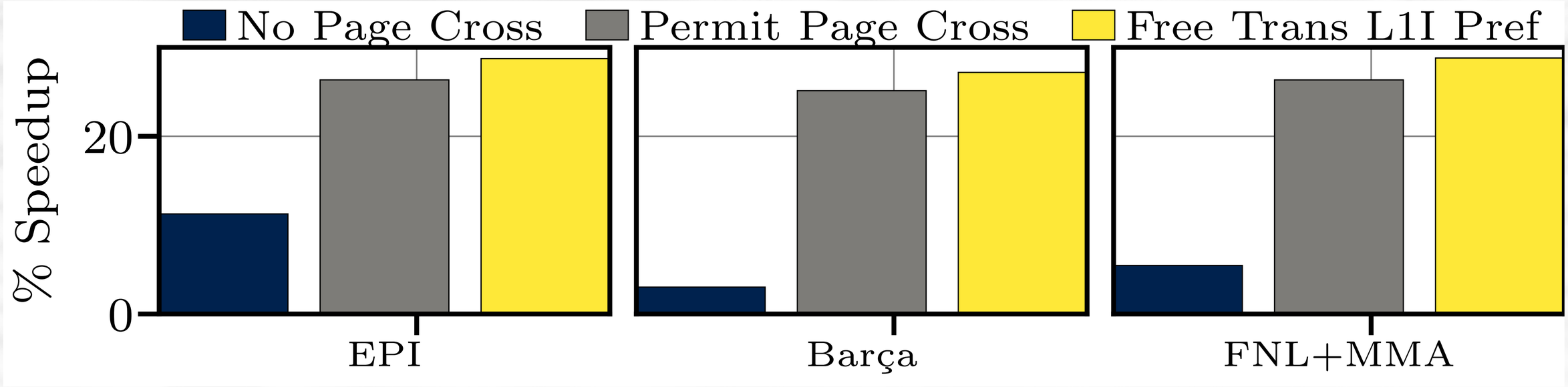
Impact of L1i Prefetching on TLBs



[HPCA2025]: To Cross, or Not to Cross Pages for Prefetching?

- Cross-page prefetching consists in prefetching beyond page boundaries:
  - An address translation might be needed.

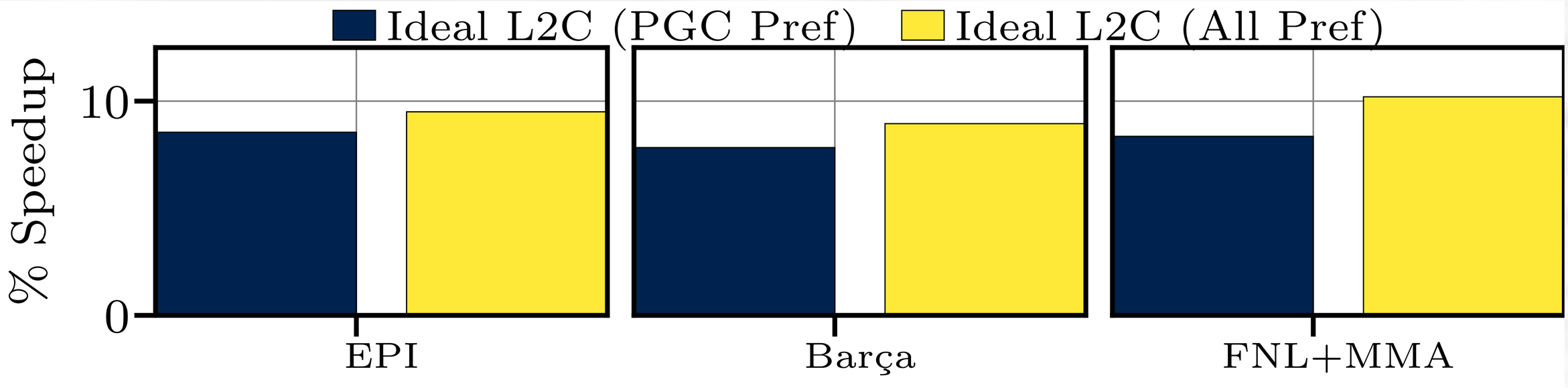
Impact of L1i Prefetching on TLBs



- Allowing instruction prefetchs to cross page boundaries delivers higher performance
- The ideal scenario deliver even higher performance

Addressing the cost of address translation for these prefetches can deliver additional performance.

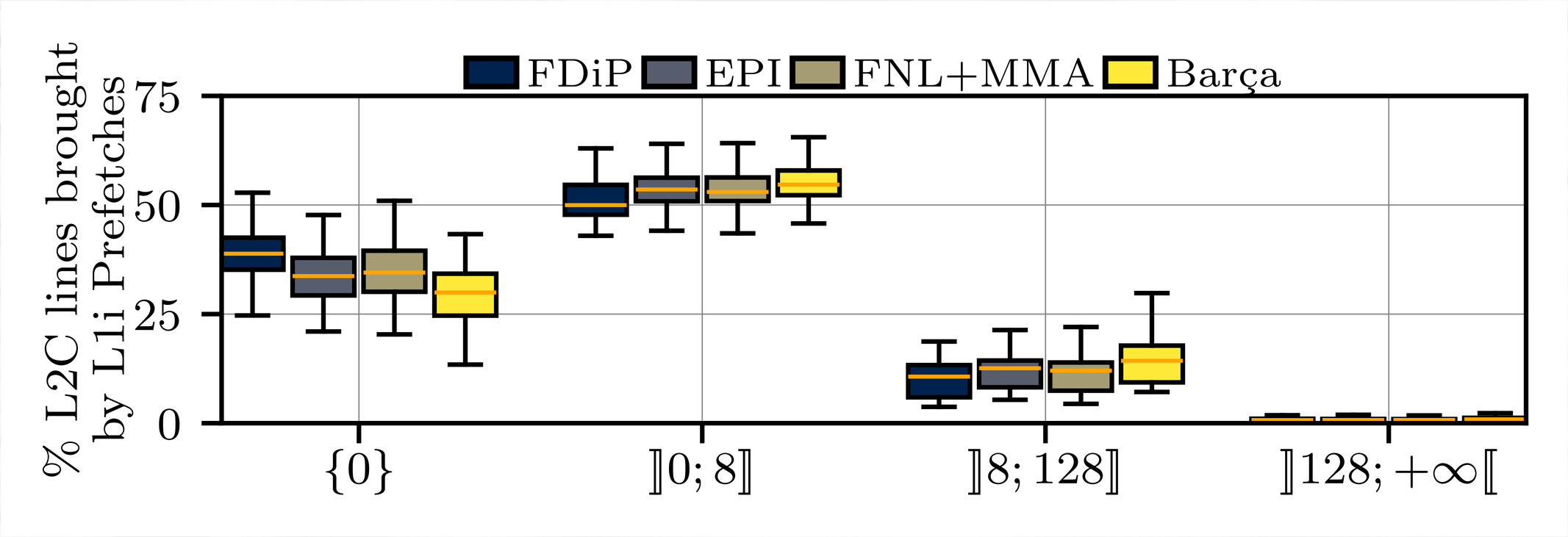
Impact of L1i Prefetching on L2C Management



- We show that code lines brought by both in-page and cross-page prefetch requests incur pressure on the L2C;
- Appropriate management of code lines has significant performance potential.

Adequate management of all lines brought by L1i prefetch requests shows higher benefits.

Reuse of Prefetched Code Lines



- We analyze the reuse of code lines in the L2C;
- A significant portion of codes lines brought by L1i Prefetch are dead on arrival;
- Other lines experience a significant amount of reuse before eviction.

Lines brought by L1i prefetches experience variable reuse patterns and call for a smart policy.

IP-CaT: A Framework to coordinate instruction prefetching

tPB

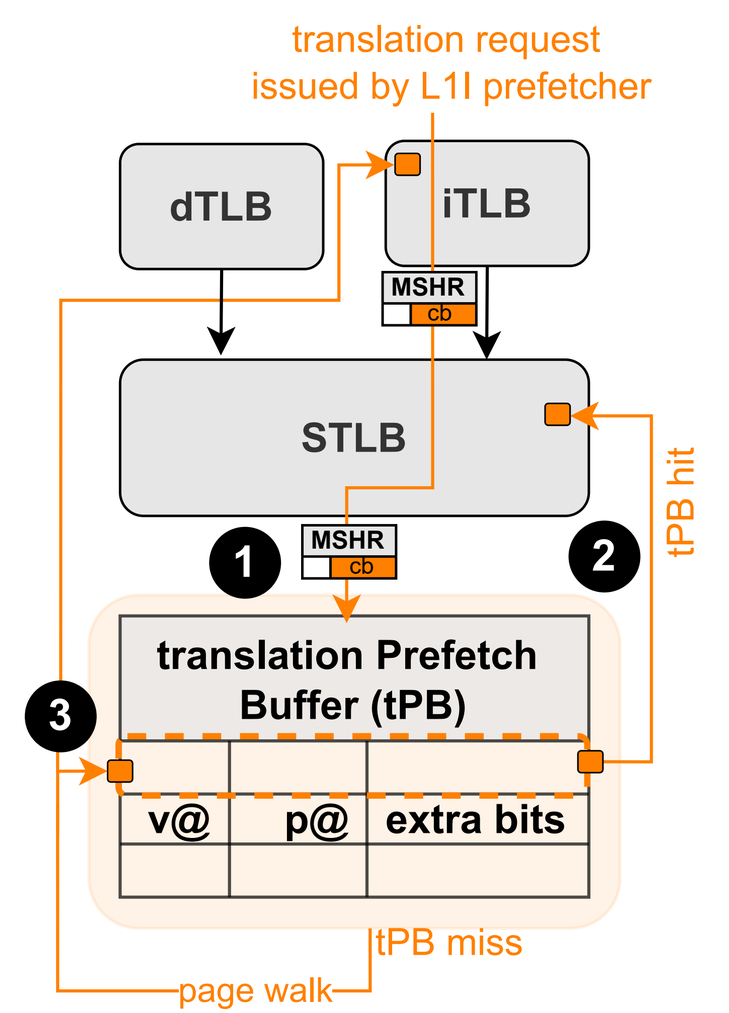
- Mitigating prefetch address translation costs yields higher performance:
  - We introduce the **tPB** to optimize cross-page translations.

TIPRP

- L1i prefetches exhibit variable reuse, requiring an intelligent policy:
  - We propose **TIPRP** to manage L2C code lines via a variation of set-dueling

We combine tPB and TIPRP in a single proposal: IP-CaT.

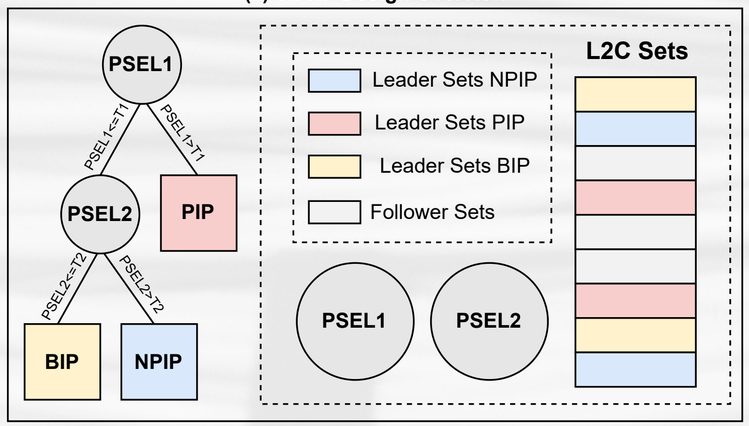
tPB: Accelerating Address Translation for Cross-Page Instruction Prefetches



- The tPB is a 64-entry buffer for cross-page prefetch translations;
- **(1)** tPB is looked up upon a miss in iTLB and STLB:
  - **(2)** tPB **hit**, the entry is inserted in the sTLB;
  - **(3)** tPB **miss**, a page table walk fills the tPB, which then updates the iTLB.

TIPRP: Smart Management of L2C Instruction Lines

(a) TIPRP Design Overview



(b) TIPRP Operation

```
if (current_set in Leader Sets NPIP)
     enable --> NPIP
elif (current_set in Leader Sets PIP)
     enable --> PIP
elif (current_set in Leader Sets BIP)
     enable --> BIP
else{ // follower sets
     if (PSEL1 > T1) enable --> PIP
     else{
          if (PSEL2 > T2) enable --> NPIP
          else            enable --> BIP }
```

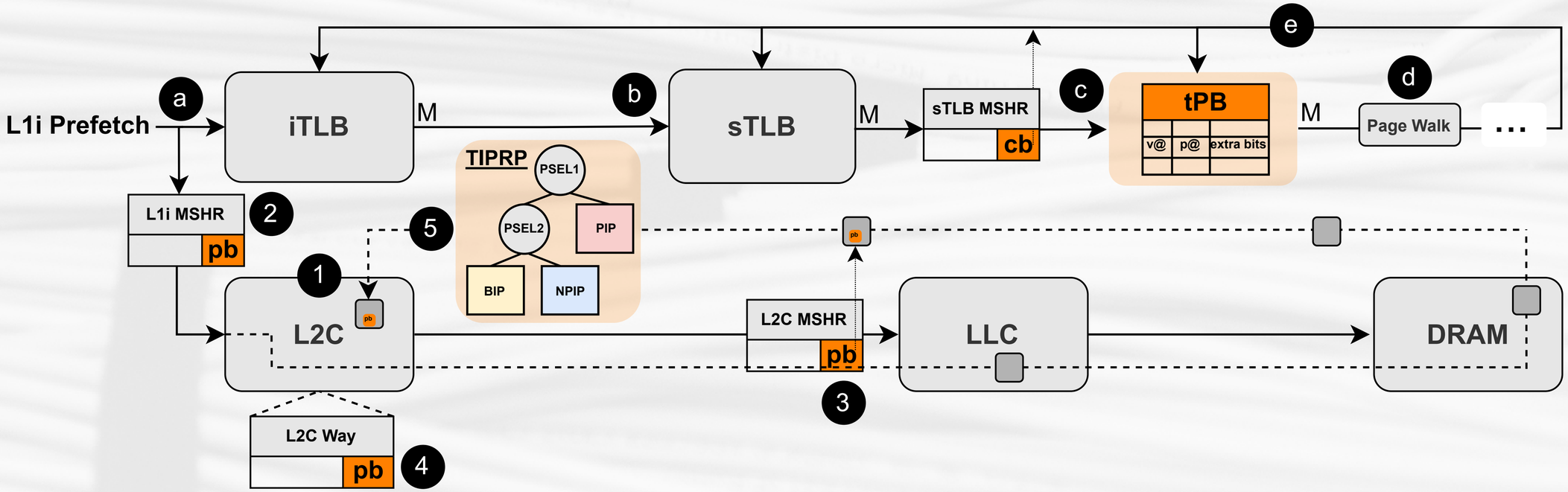
(c) TIPRP Training

Leader Set Type	Demand L2C Hit	L2C Eviction
Leader Sets PIP	if(pb==0) no update if(pb==1) PSEL1++, PSEL2--	if(pb==0) no update if(pb==1) PSEL1--, PSEL2--
Leader Sets NPIP	if(pb==0) PSEL1--, PSEL2++ if(pb==1) no update	if(pb==0) PSEL1++, PSEL2-- if(pb==1) no update
Leader Sets BIP	if(pb==0) PSEL1--, PSEL2-- if(pb==1) no update	if(pb==0) PSEL1++, PSEL2++ if(pb==1) no update

- TIPRP is made of three sub-policies:
  1. **PIP** —> Prioritizes lines brought by L1i prefetch requests
  2. **NPIP** —> Deprioritizes lines brought by L1i prefetch requests;
  3. **BIP** —> Lines brought by L1i prefetch requests bypass the L2C;

- The L2C is managed using a **variation** of set-dueling:
  - Each sub-policy is applied in separate sets and competes to determine which is the best at a given time
  - Events signaling a positive/negative behaviour for lines brought by L1i prefetch requests update PSEL1 and PSEL2
  - Decision tree is designed to provide a continuum of policies

IP-CaT: Synergistic TLB and Instruction Cache Management



- The iTLB is looked up upon an L1i prefetch request
  - iTLB miss —> STLB miss —> tPB look up
- Upon a tPB miss, a page walk is initiated
  - Address translation stored in iTLB and tPB when linked to an L1i cross-page prefetch request
  - Otherwise, stored in iTLB and STLB
- Upon a tPB hit, the entry is inserted in the STLB

- TIPRP decides which sub-policy to enable in the following sets of the L2C
  - c.f., TIPRP component

Experimental Setup

- ChampSim simulator (GitHub)
- Considered μ-architectures:
  - **Cascade Lake**
  - **Zen 4**
- Workloads
  - Qualcomm & Server (105 workloads)
  - SimPoints methodology
- **Single**-core evaluation
- **Multi**-page size evaluation
- **Multi**-core evaluation

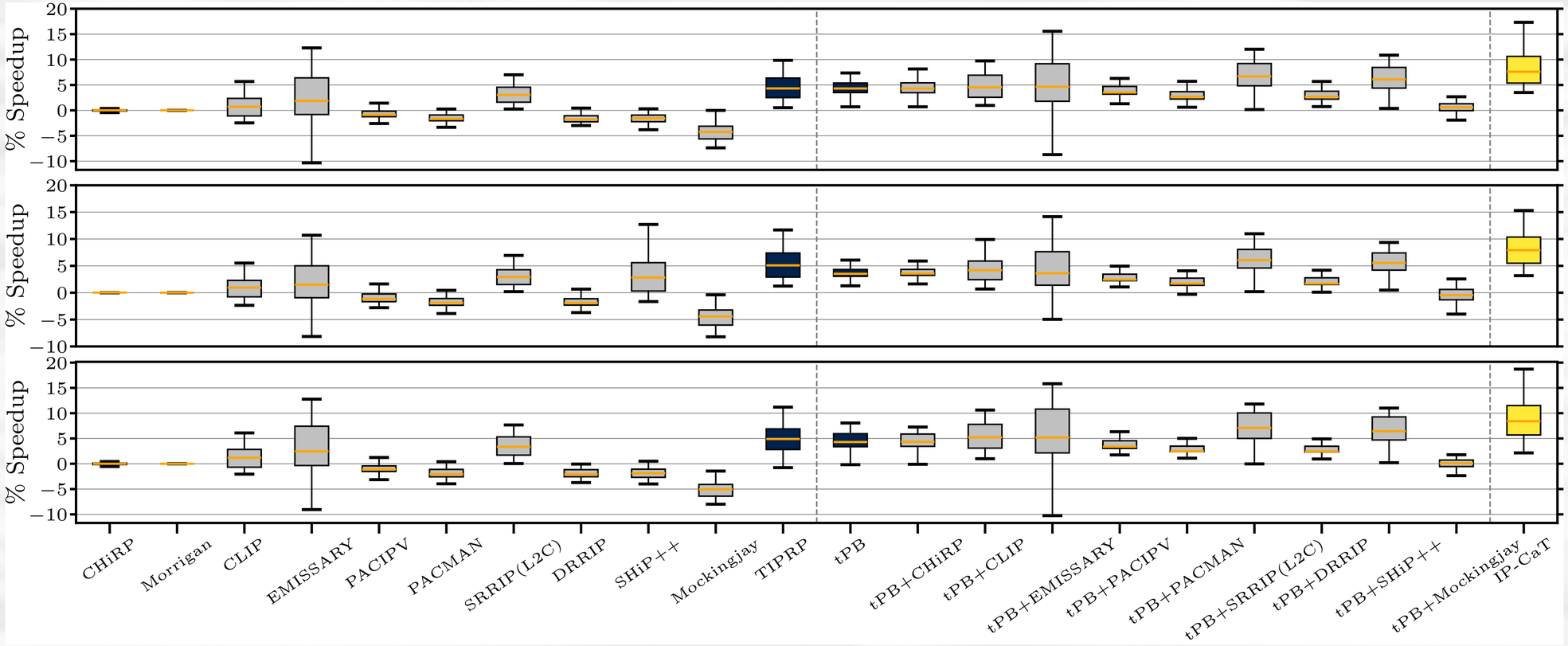
Component	Parameters	Latency
ITLB	64-entry, 4-way	1cc
DTLB	64-entry, 4-way	1cc
STLB	3584-entry, 14-way	8cc
L1 I-Cache	32KB, 8-way	4cc
L1 D-Cache	32KB, 8-way, Berti Prefetcher	4cc
L2 Cache	1MB, 8-way	10cc
LLC	4MB/core, 8-way	20cc
DRAM	4GB, tRP=tRCD=tCAS=24cc	variable

**Representative of a Zen4 μ-arch.*

Alternative Techniques

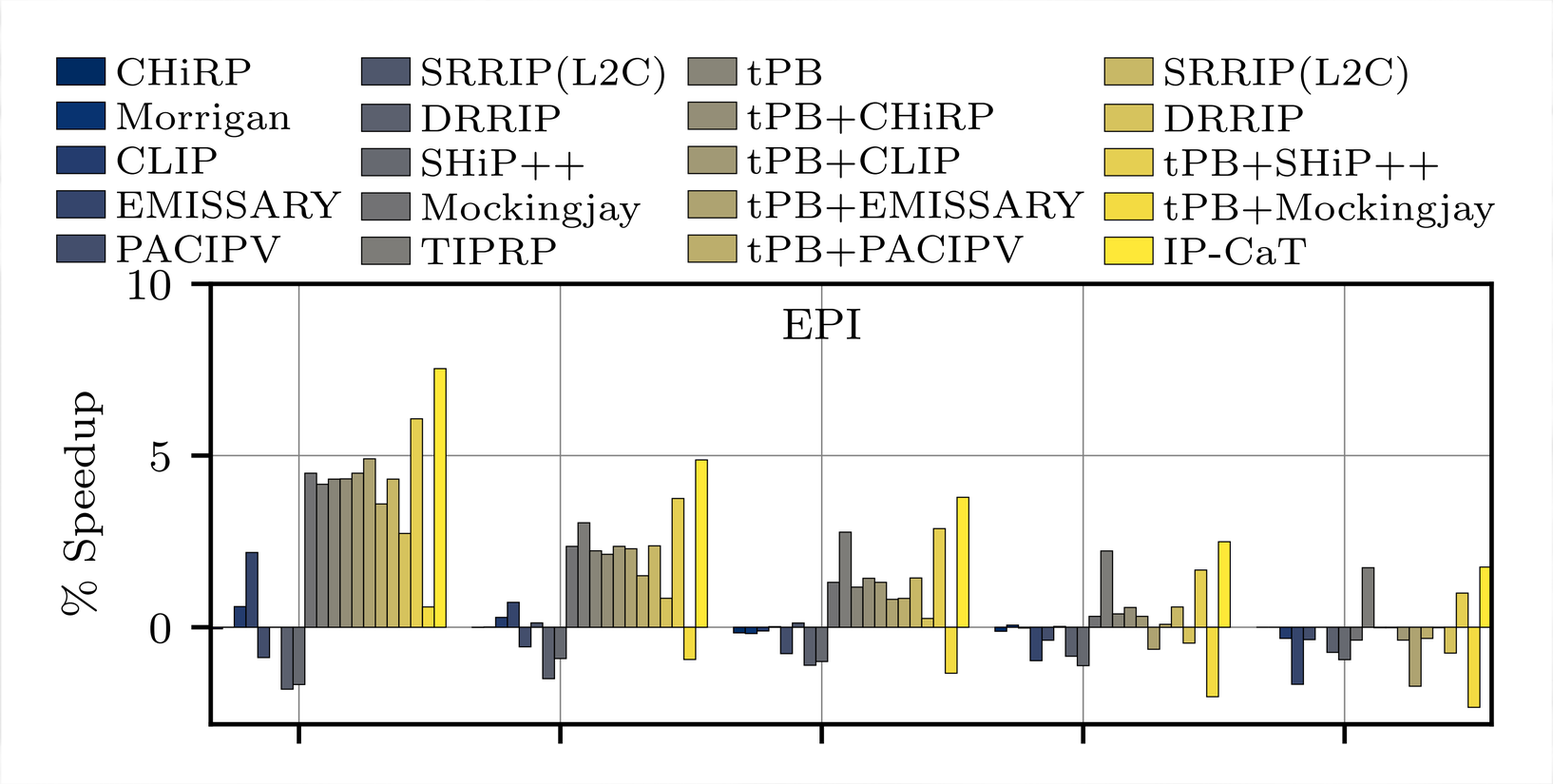
- Designs evaluated:
  - CHiRP
  - Morrigan
  - CLIP
  - EMISSARY
  - PACIPV
  - PACMAN
- SRRIP (L2C)
- DRRIP
- SHiP++
- Mockingjay
- IP-CaT (our design)
- Considered L1i Prefetchers:
  - EPI
  - FNL+MMA
  - Barça

Single-Core Evaluation | Performance



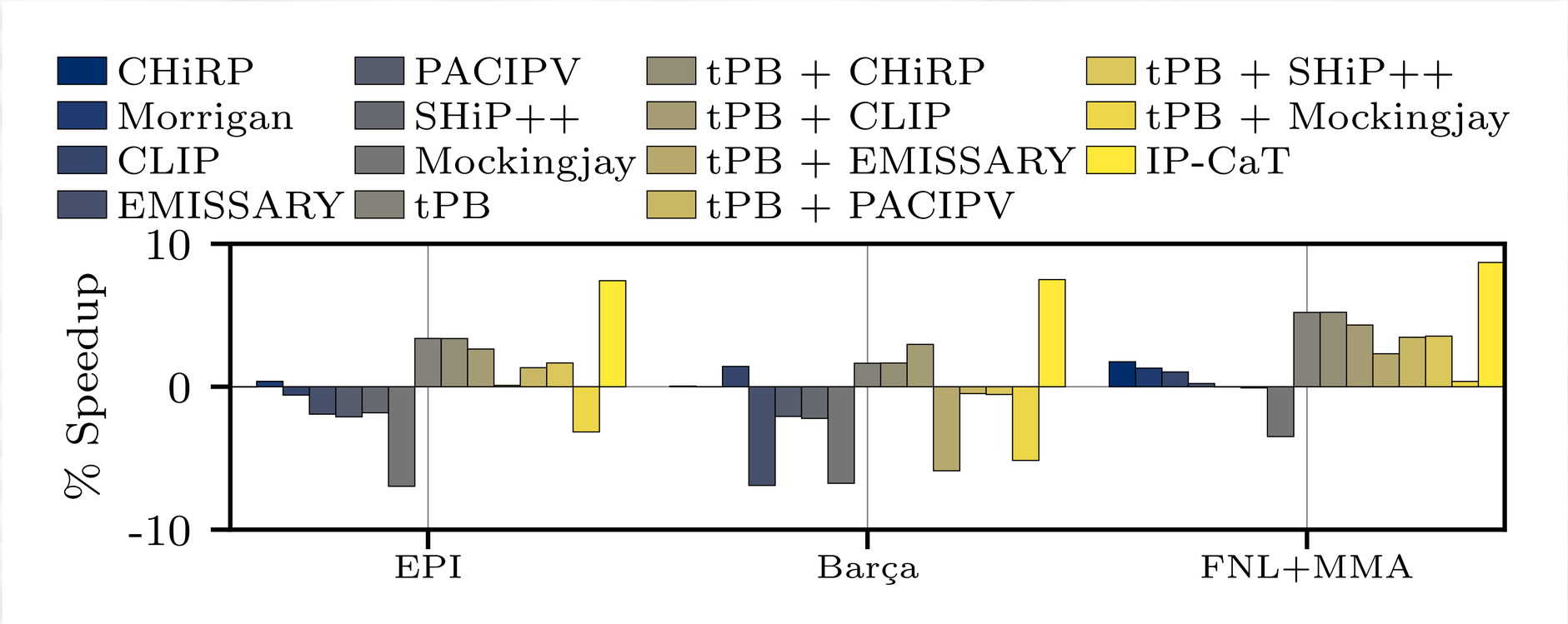
- Considering EPI in a single-core scenario, IP-CaT outperforms all evaluated designs
  - **TIPRP** leverages **2.9%** geomean speed-up
  - **IP-CaT (tPB+TIPRP)** leverages **6.1%** geomean speed-up

Multi-Page Size Evaluation | Performance



- We evaluate scenarios where **0%**, **70%**, **80%**, **90%**, and **100%** of the memory footprint is mapped to large pages;
- In all considered scenarios, **IP-CaT** outperforms all considered designs:
  - When using more large pages, **TIPRP** provides most of the performance improvements;
  - Trends are similar considering other L1i prefetcher (*i.e.*, Barça, and FNL+MMA).

Multi-Core Evaluation | Performance



- We evaluate **IP-CaT** against other techniques in a 4-core scenario:
  - In all considered scenarios, **IP-CaT** outperforms all considered designs

Conclusions

We show that the address translation latency of L1i cross-page prefetch requests undermines the benefits of these prefetchers.

We propose the first work to orchestrate TLB and cache management to maximize the benefits of L1i prefetchers.

This work generalises to any L1i prefetchers.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you for your attention!

Paper available here:

Acknoledgments/ Fundings

- Alexandre Valentin Jamet acknowledges his AI4S fellowship within the “Generación D” initiative by Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by NextGenerationEU through PRTR.