Microarchitectural Implications of **Event-driven** Server-side Web Applications

Yuhao Zhu
UT Austin
with Daniel Richins, Matthew Halpern, Vijay Janapa Reddi

MICRO 2015

Instruction Supply is a Critical Aspect of Microarchitecture Design

Exploit Instruction Locality

a.k.a., Common Case Design

Instruction Supply

Cache

Branch Predictor

TLB

Exploit Instruction Locality

a.k.a., Common Case Design

Instruction Supply

Exploit Instruction Locality

a.k.a., Common Case Design

Instruction Supply

Cache — Hot instructions

Branch Predictor — Hot branch history patterns

TLB

Exploit Instruction Locality

a.k.a., Common Case Design

Instruction Supply

Cache	Hot instructions
Branch Predictor	Hot branch history patterns
TLB	Hot pages

Exploit Instruction Locality

a.k.a., Common Case Design

		SPEC CPU (mostly)	
Instruction Supply	Cache	Hot instructions	✓
	Branch Predictor	Hot branch history patterns	✓
	TLB	Hot pages	✓

Exploit Instruction Locality

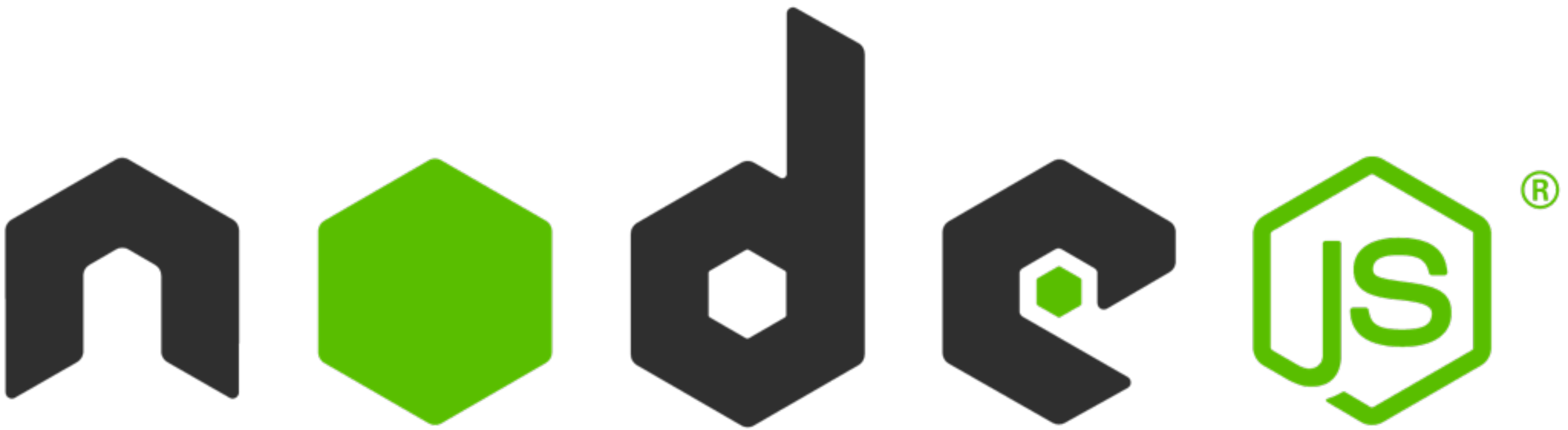
a.k.a., Common Case Design

		SPEC CPU (mostly)	Event-driven Applications
Instruction Supply	Cache	Hot instructions	✓
	Branch Predictor	Hot branch history patterns	✓
	TLB	Hot pages	✓

Locality Lost

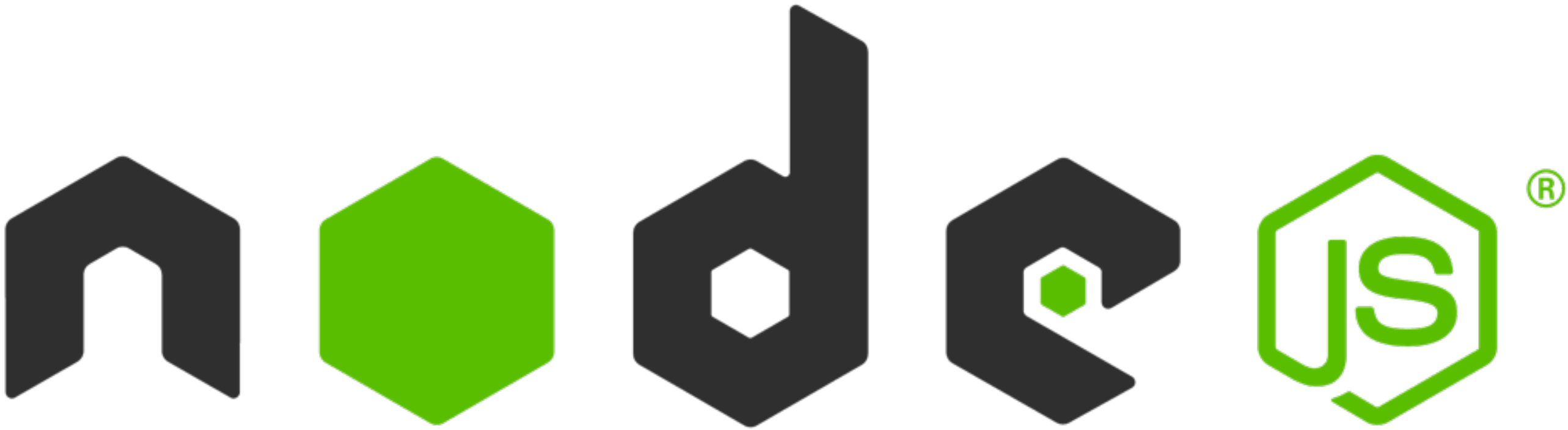
		SPEC CPU (mostly)	Event-driven Applications
Instruction Supply	Cache	Hot instructions	✓
	Branch Predictor	Hot branch history pattern	✓
	TLB	Hot pages	✓

Locality Lost



TLB

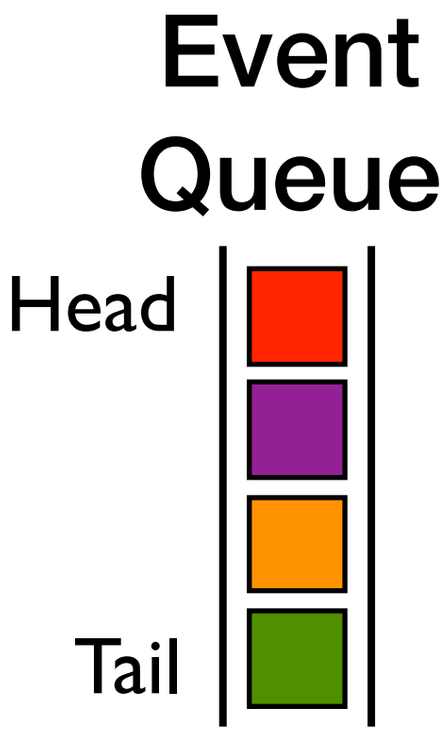
Hot pages



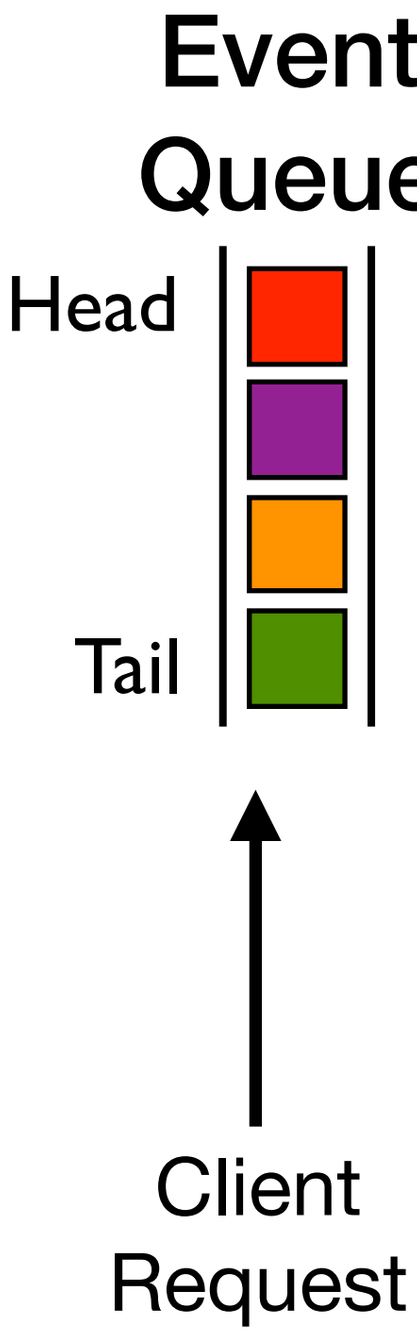
TLB

Hot pages

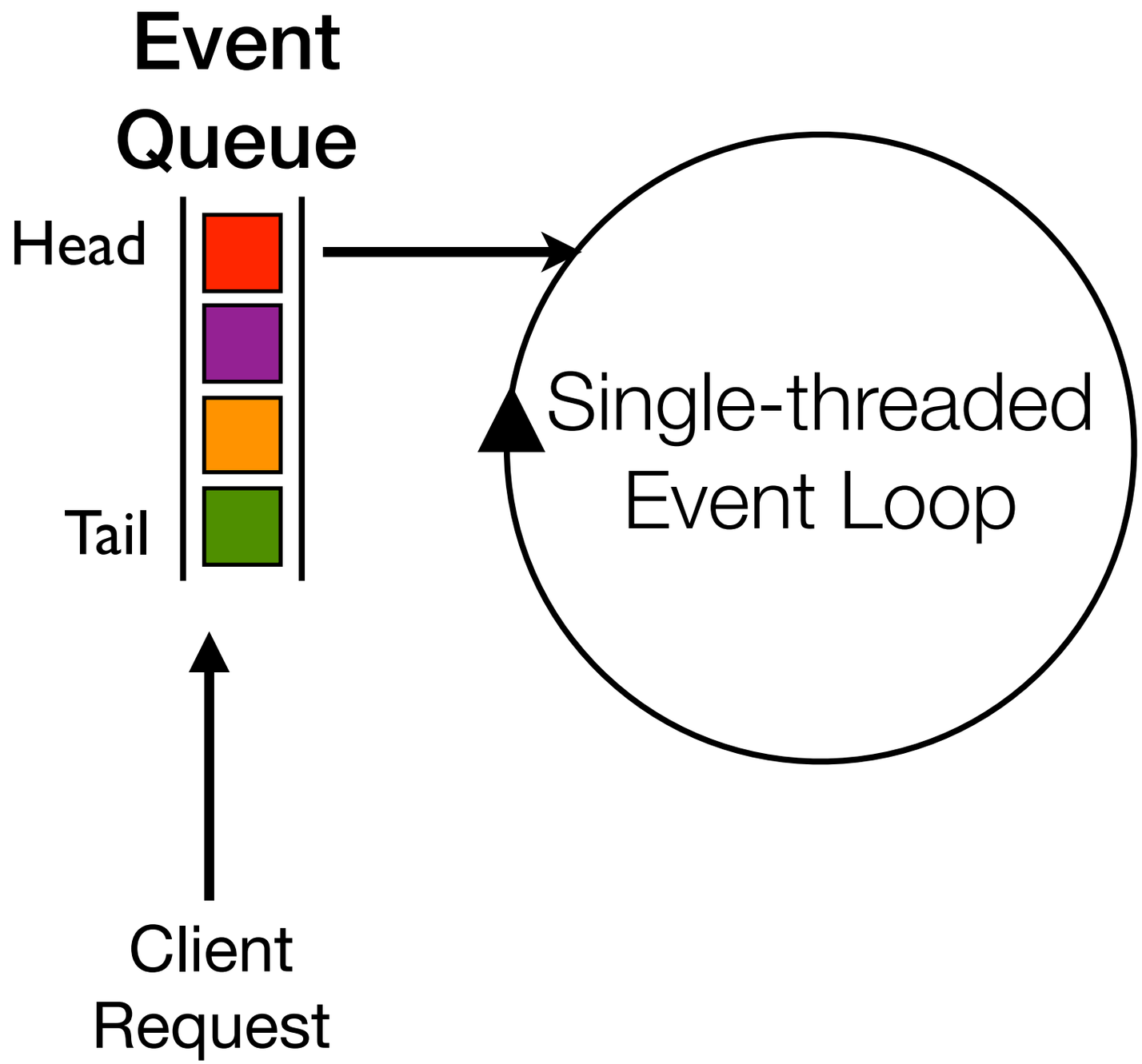
Event-driven Execution Model



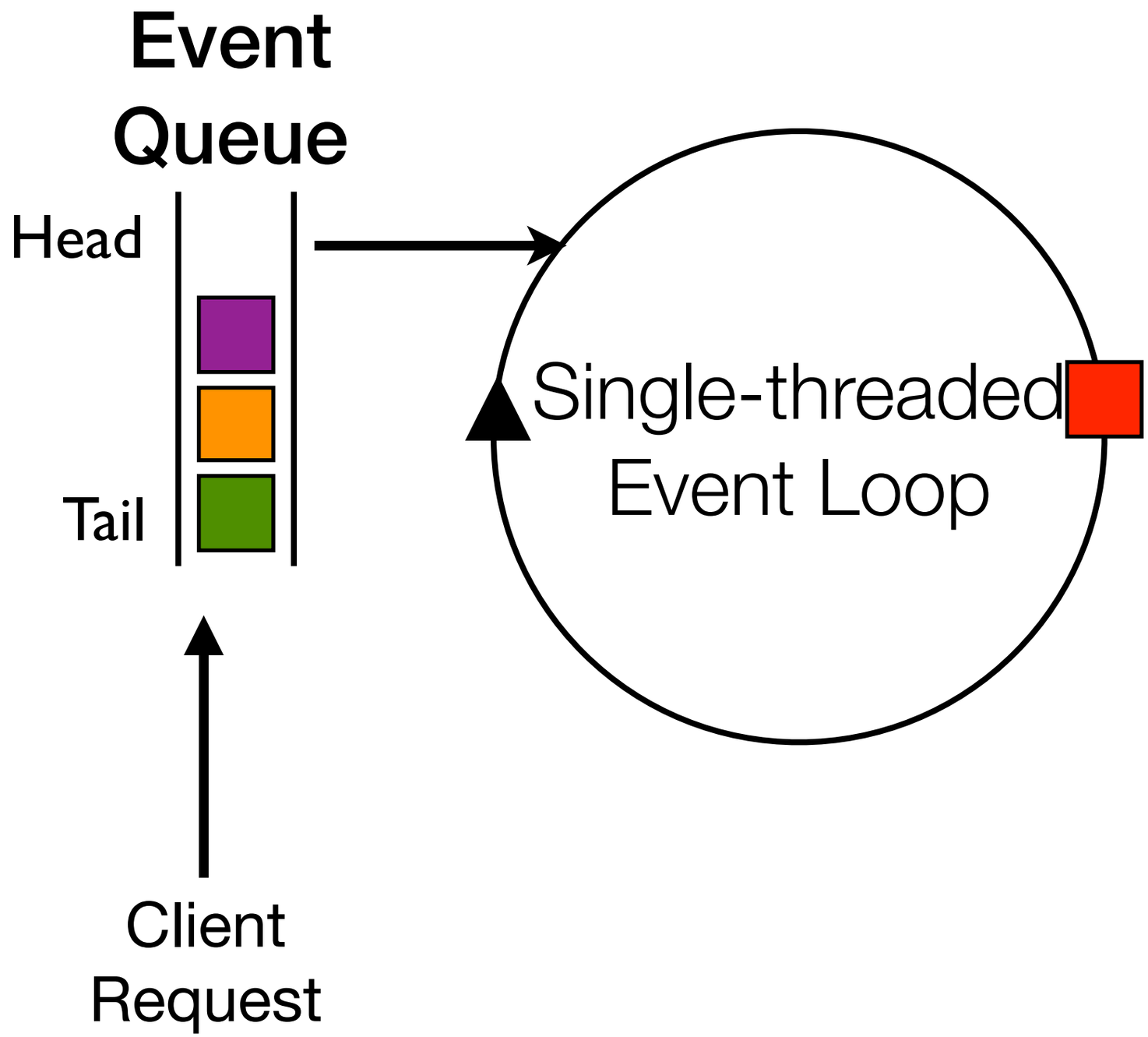
Event-driven Execution Model



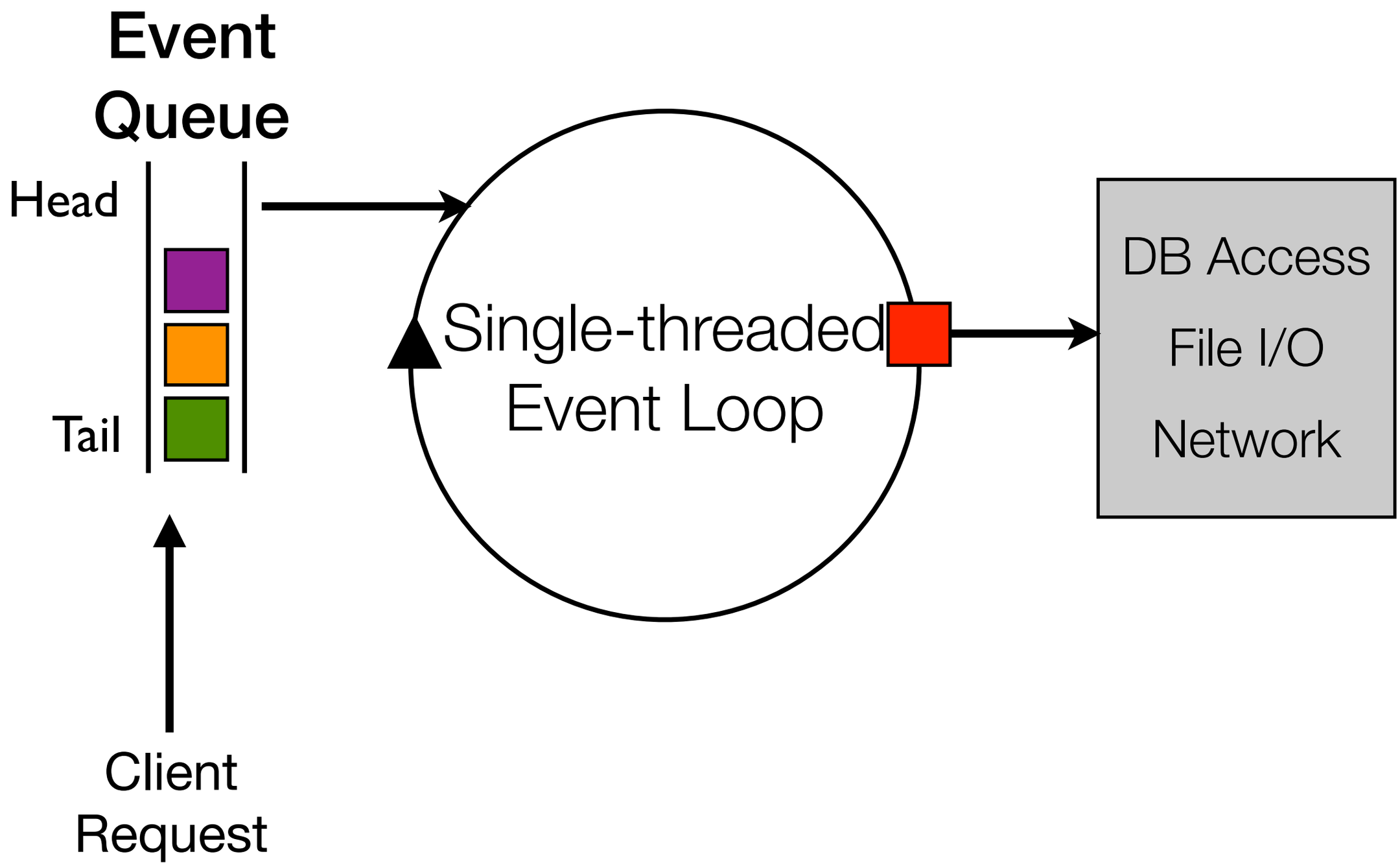
Event-driven Execution Model



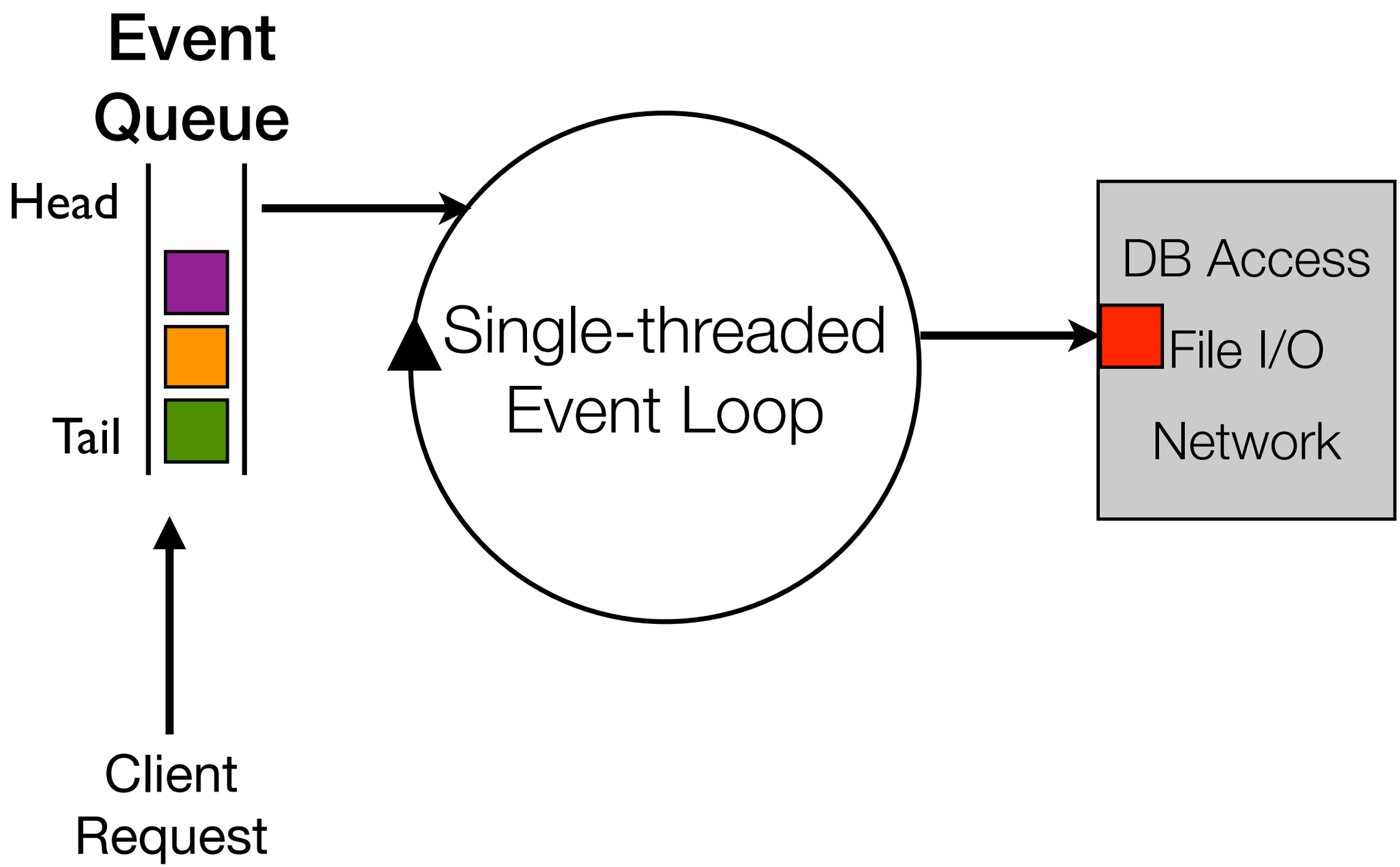
Event-driven Execution Model



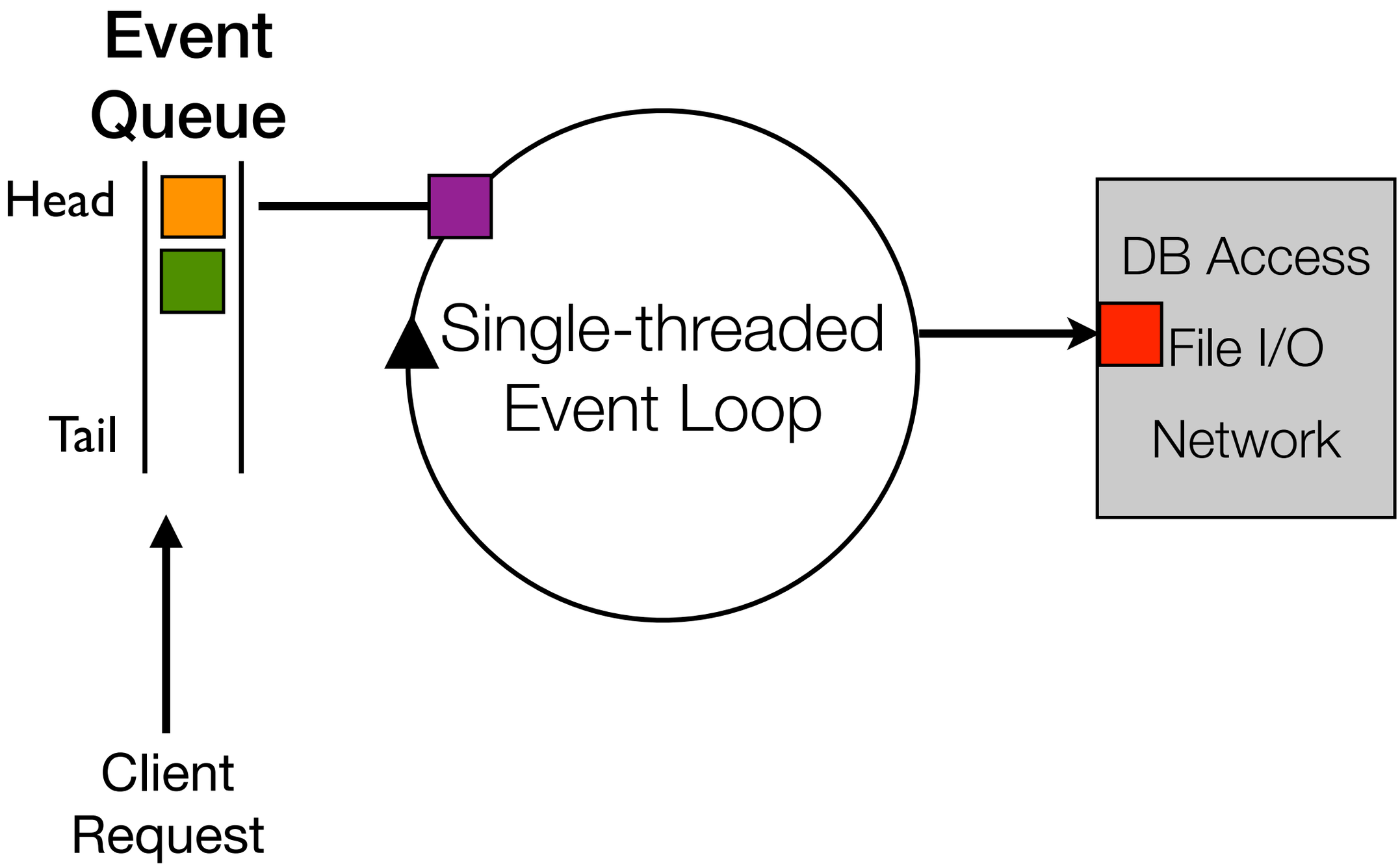
Event-driven Execution Model



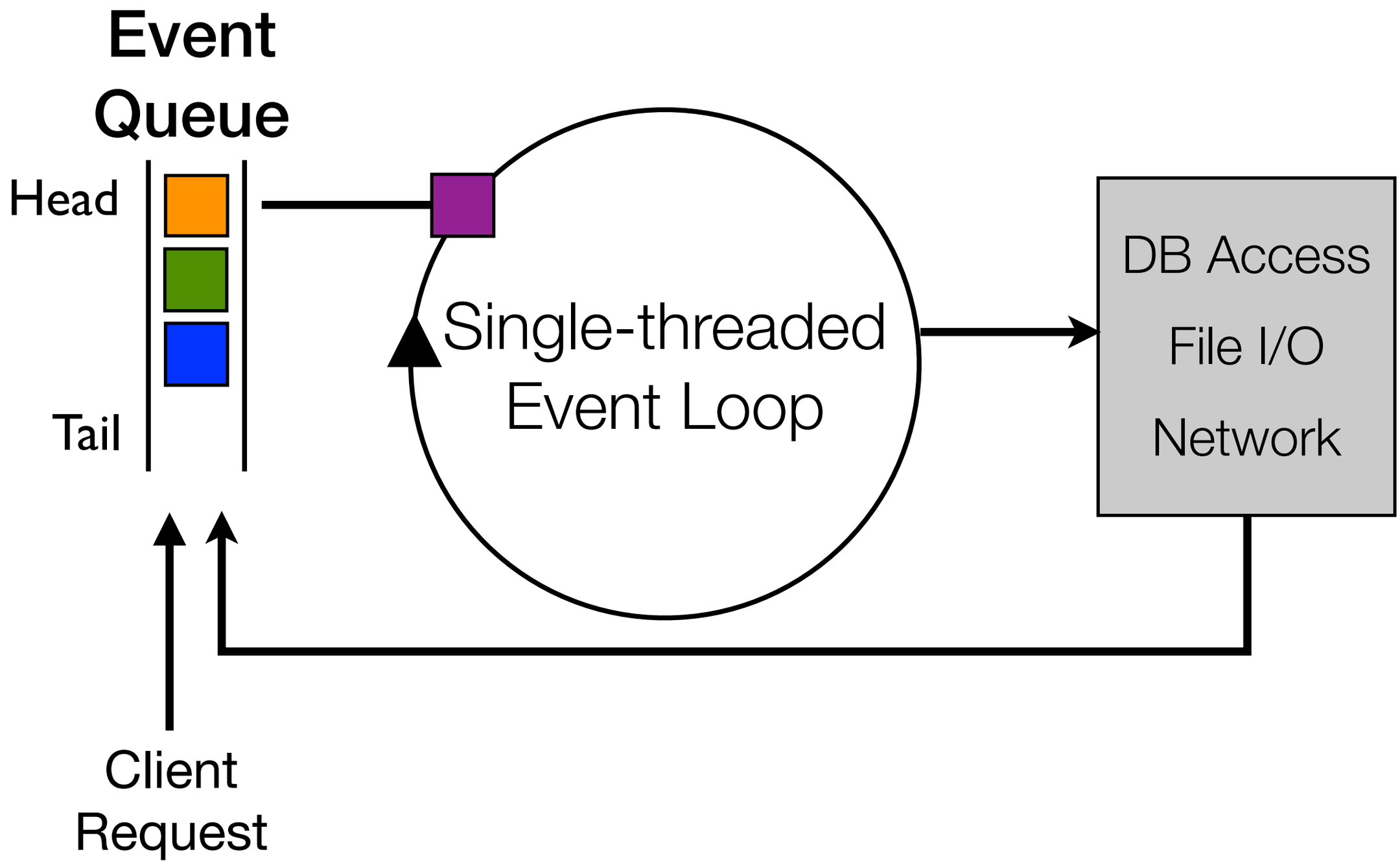
Event-driven Execution Model



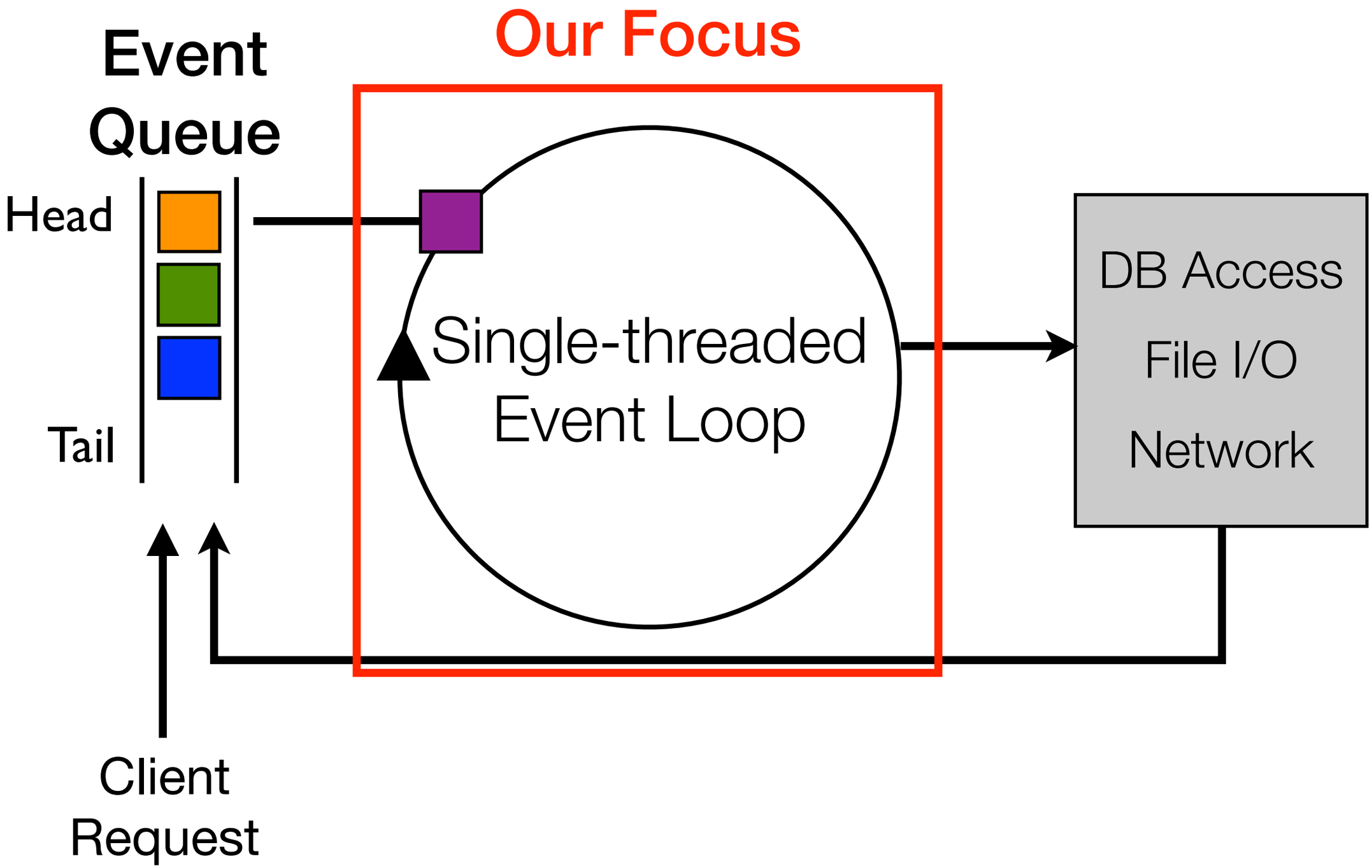
Event-driven Execution Model



Event-driven Execution Model



Event-driven Execution Model



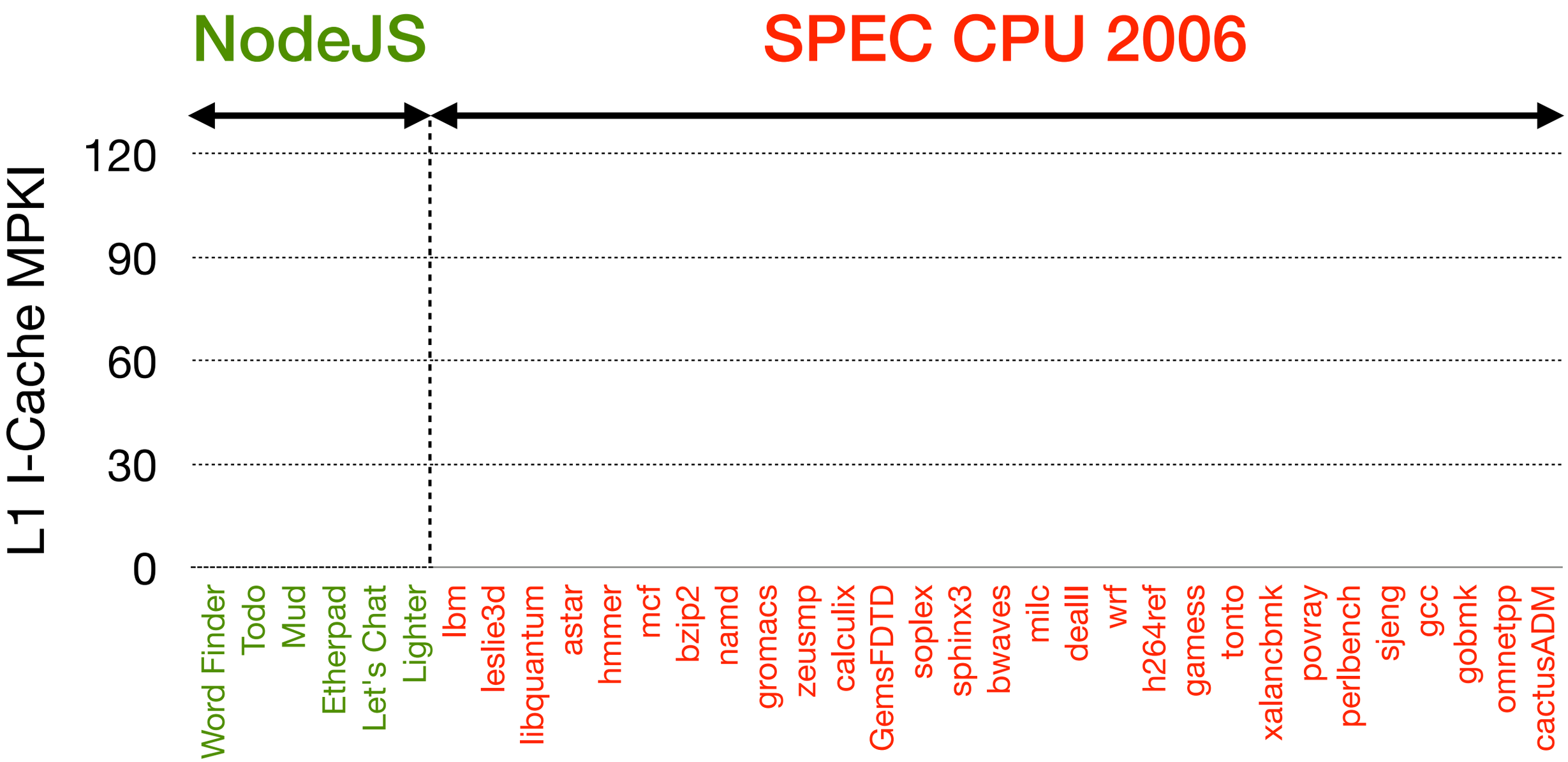
Applications

Application	Domain
Etherpad Lite	Document Collaboration
Let's Chat	Messaging
Lighter	Content Management
Mud	Gaming
Todo	Task Management
Word Finder	API Services

https://github.com/nodebenchmark/

Locality Lost

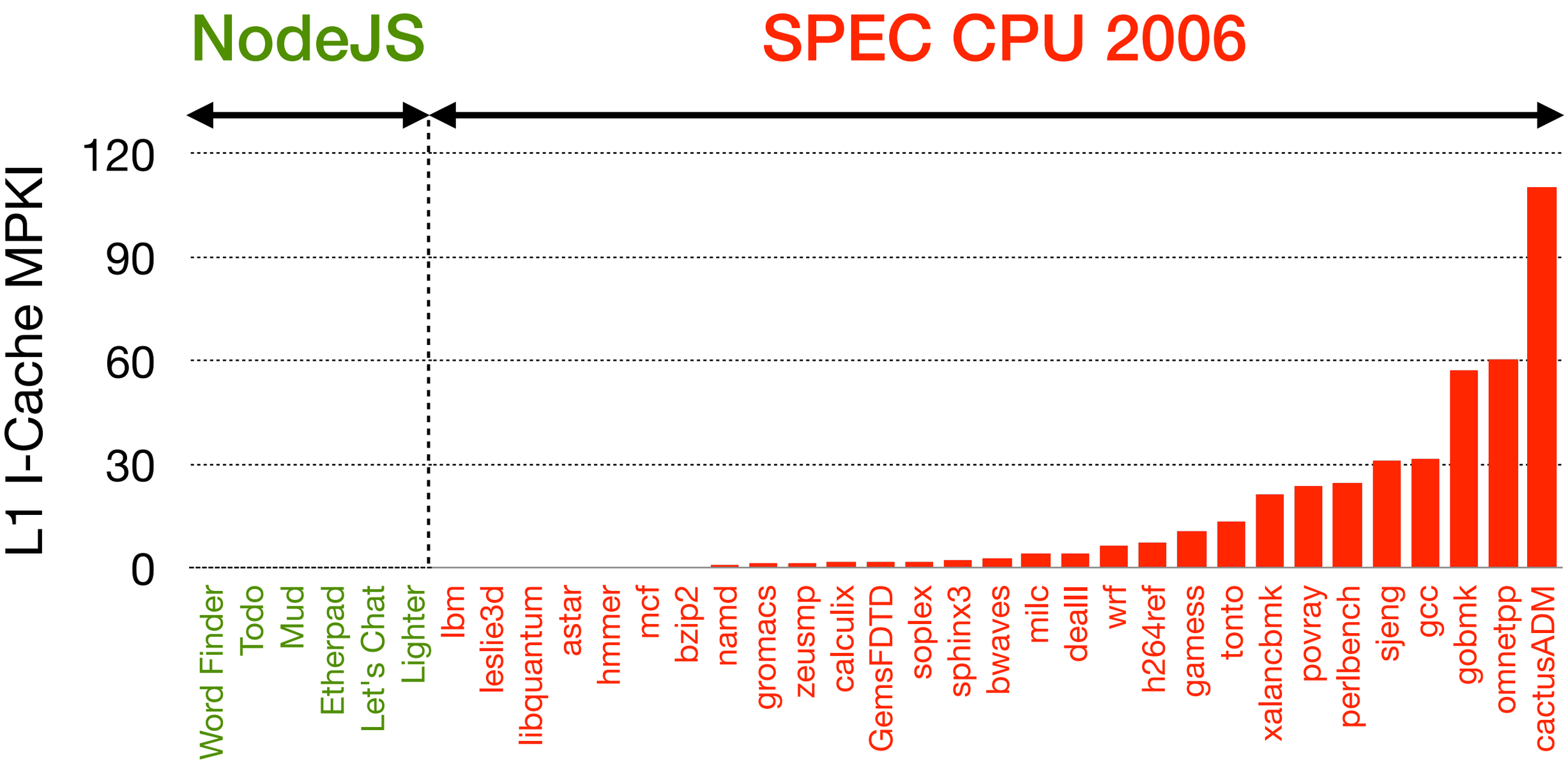
Locality Lost



I-Cache Parameters

32 KB, 64 B cache line, 8-way

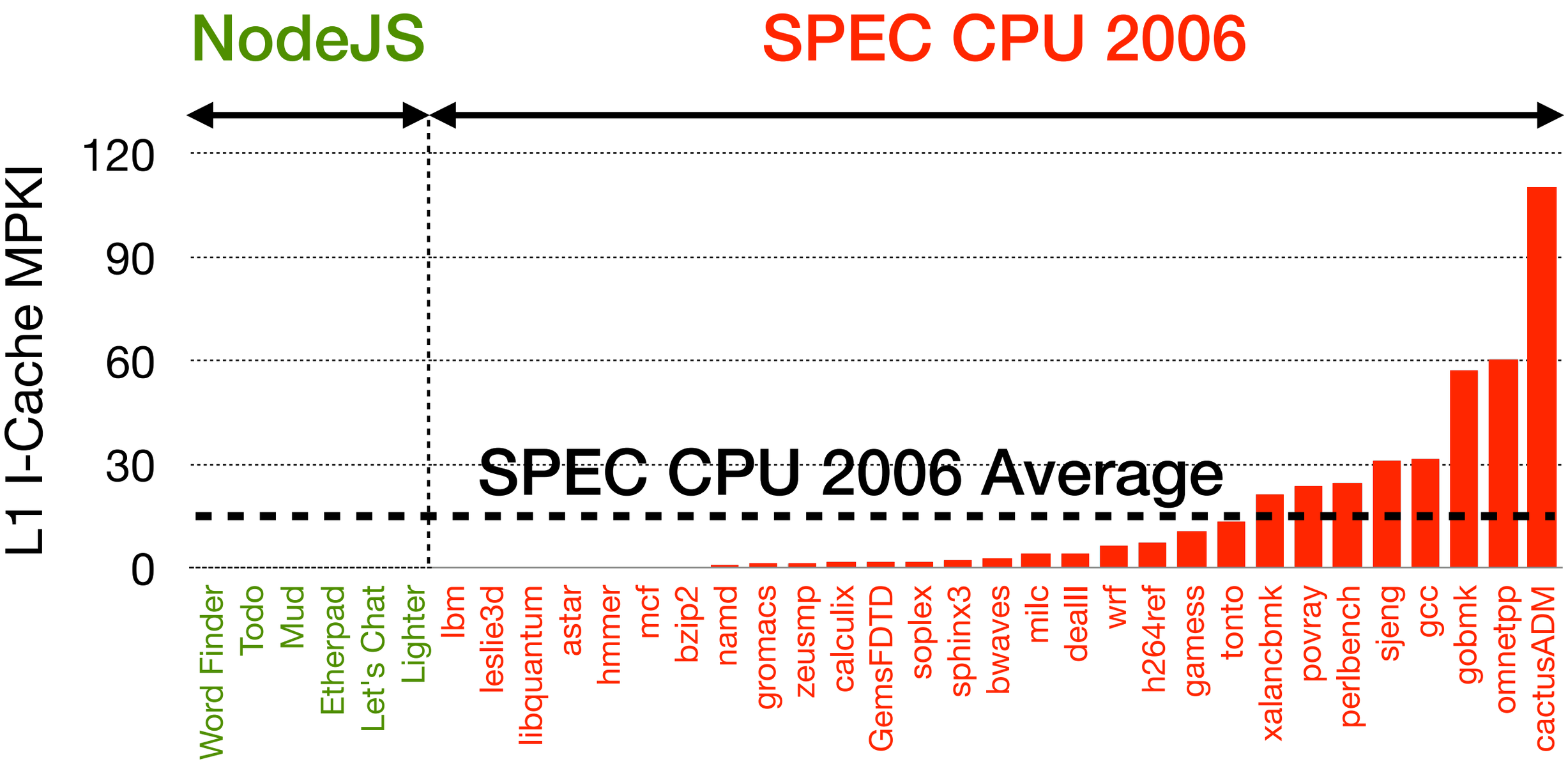
Locality Lost



I-Cache Parameters

32 KB, 64 B cache line, 8-way

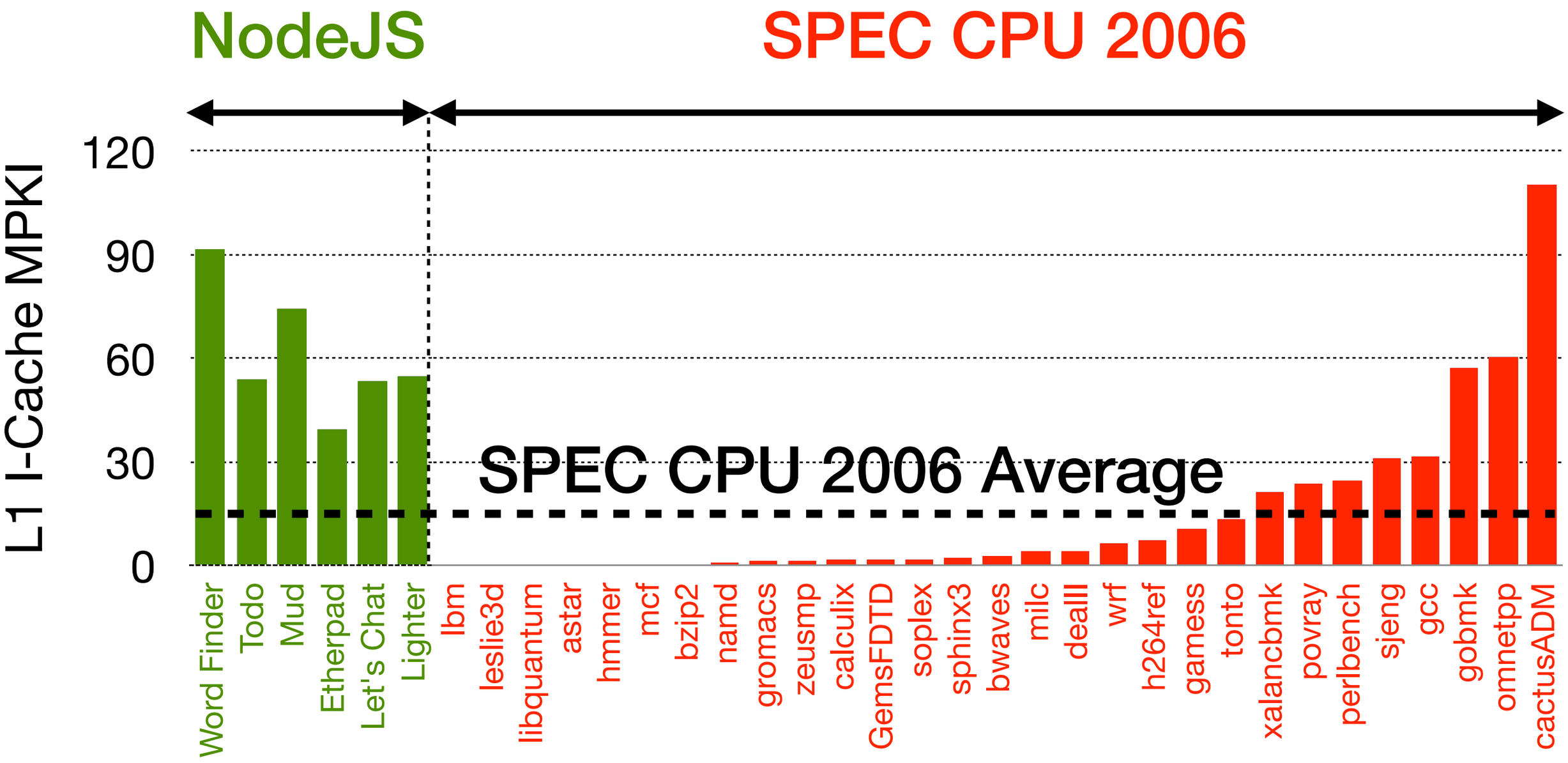
Locality Lost



I-Cache Parameters

32 KB, 64 B cache line, 8-way

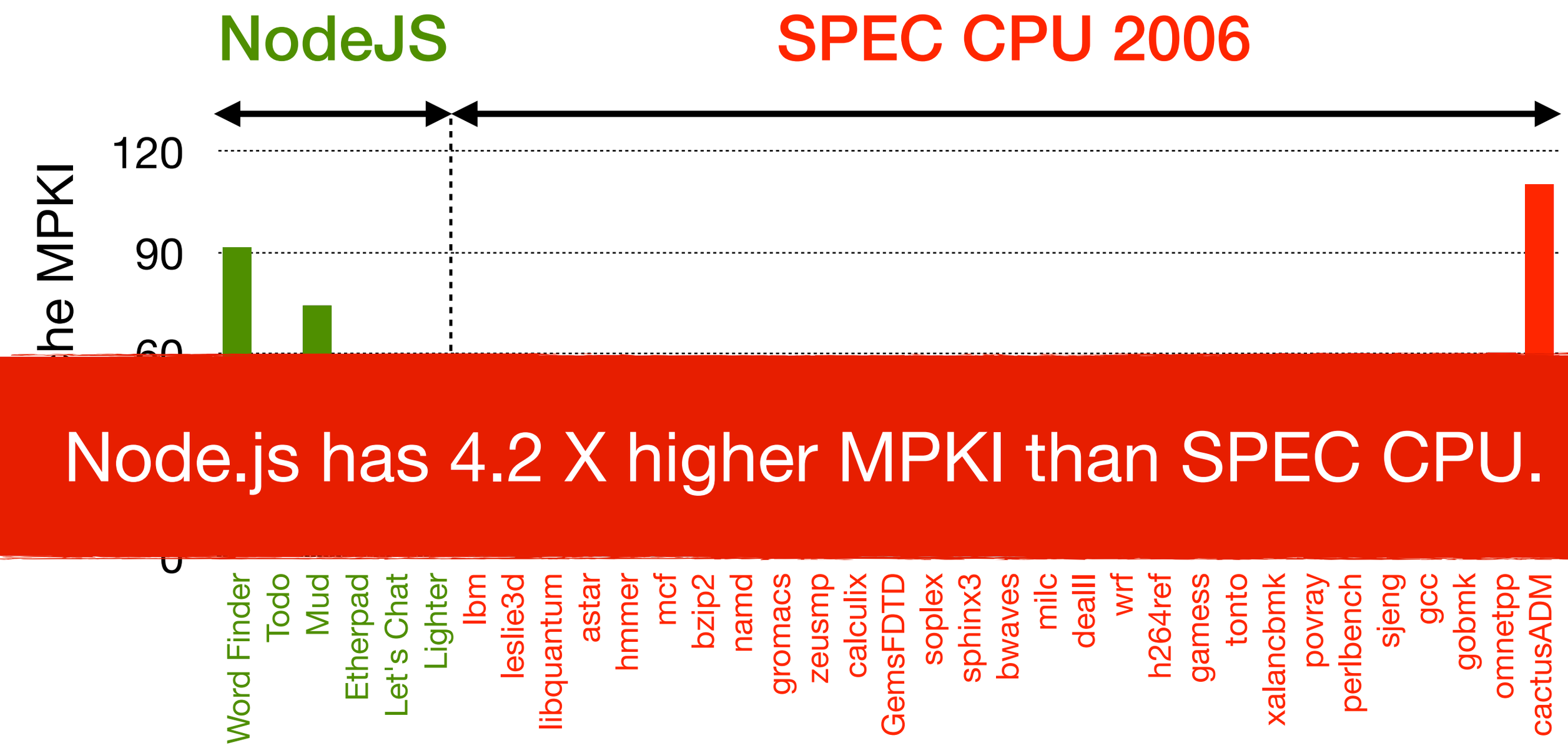
Locality Lost



I-Cache Parameters

32 KB, 64 B cache line, 8-way

Locality Lost



Node.js has 4.2 X higher MPKI than SPEC CPU.

I-Cache Parameters

32 KB, 64 B cache line, 8-way

Root Cause Analysis

Root Cause Analysis

High I-$ Miss Ratio

Root Cause Analysis

High I-$ Miss Ratio

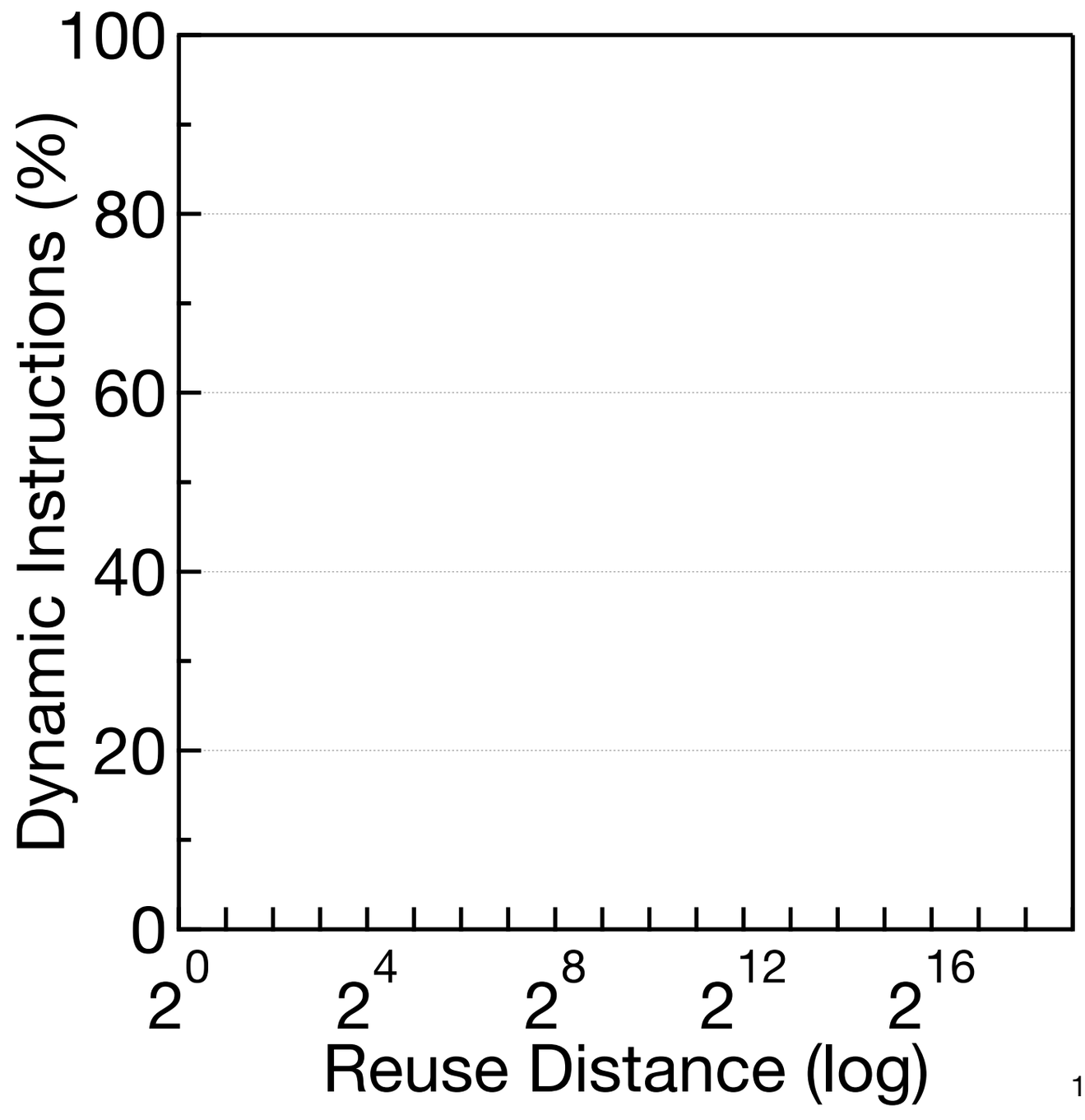
Large Instruction Reuse-distance

Root Cause Analysis

High I-$ Miss Ratio

↑

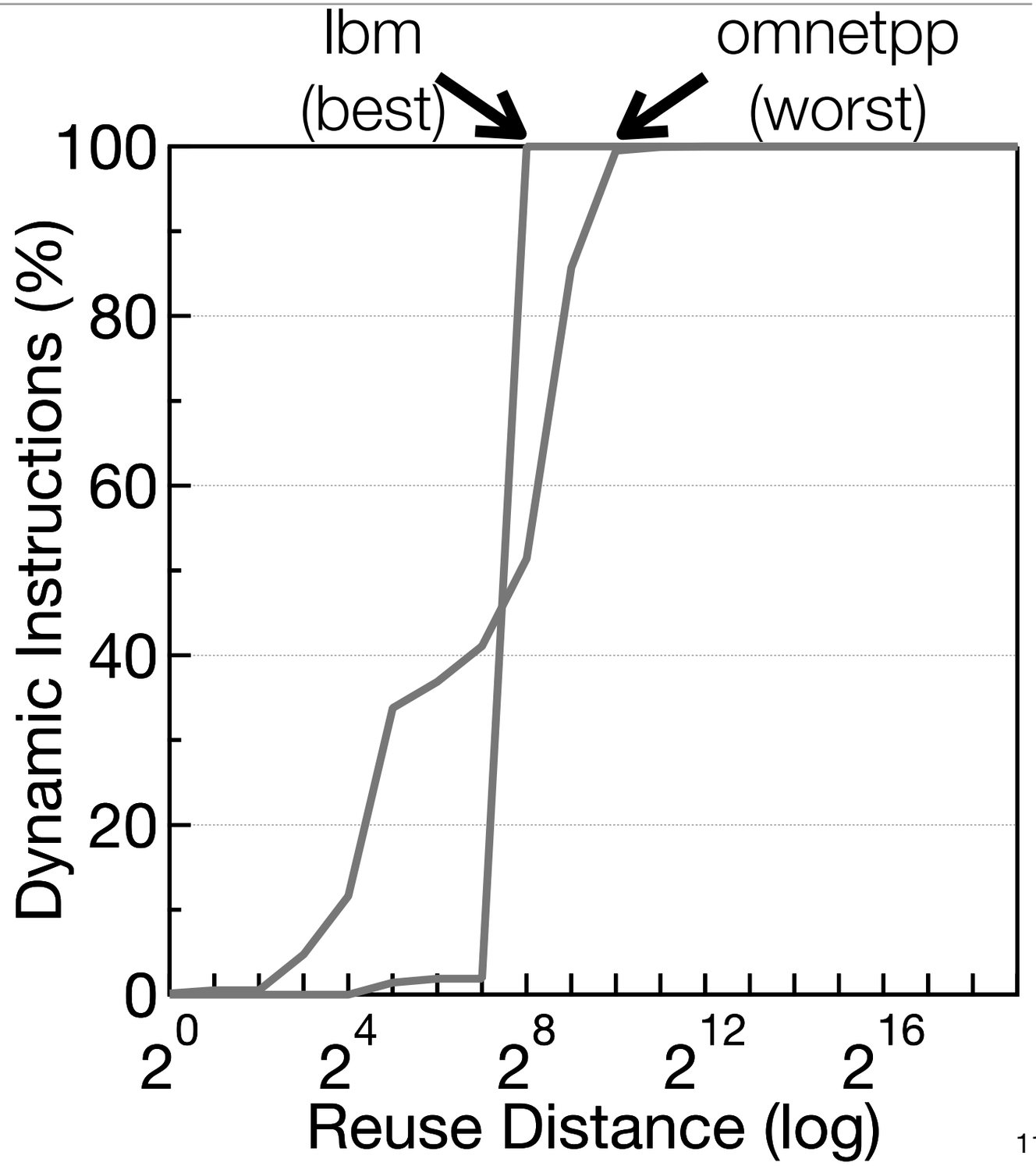
Large Instruction Reuse-distance



Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-distance

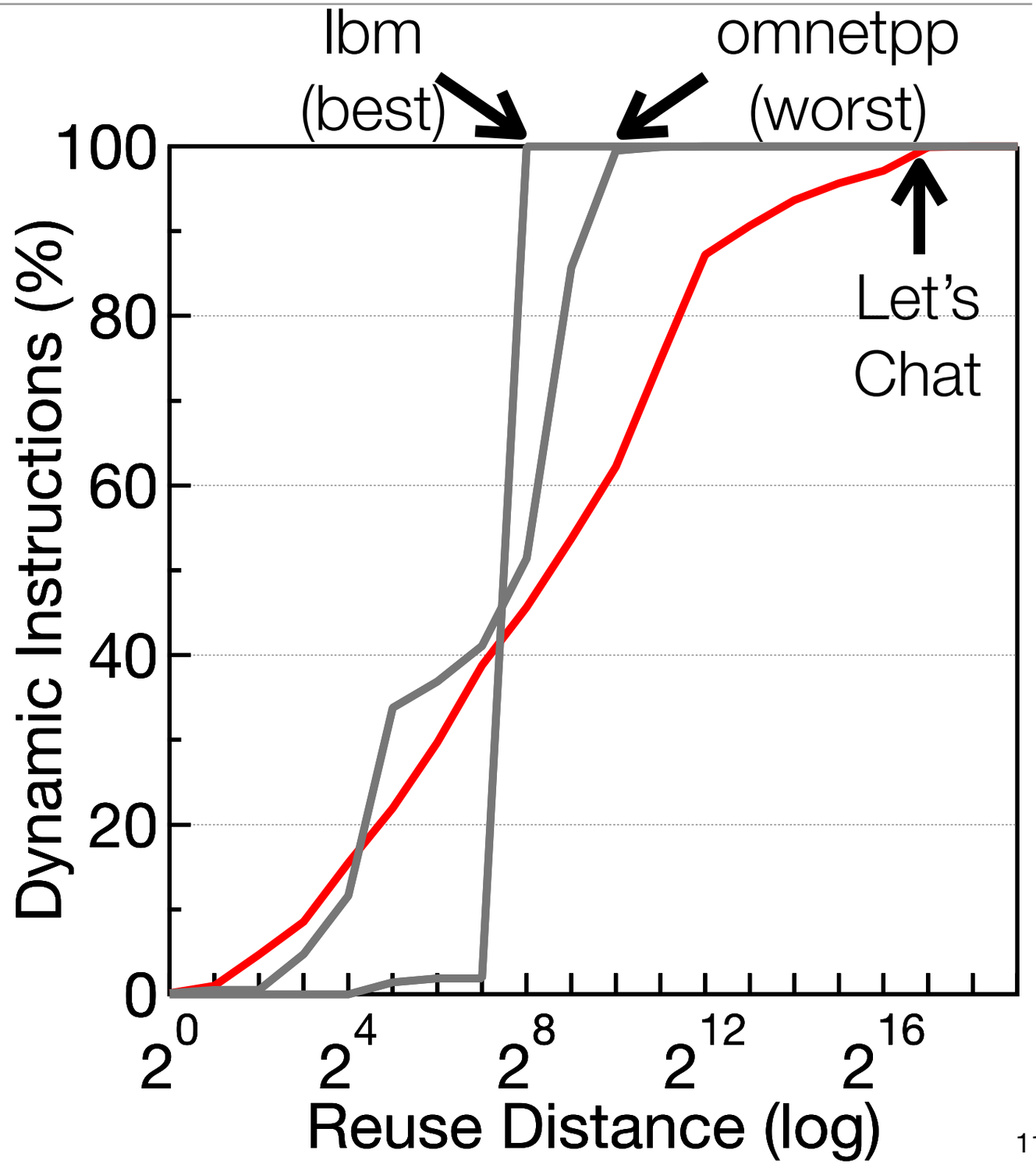


Root Cause Analysis

High I-$ Miss Ratio

↑

Large Instruction Reuse-distance

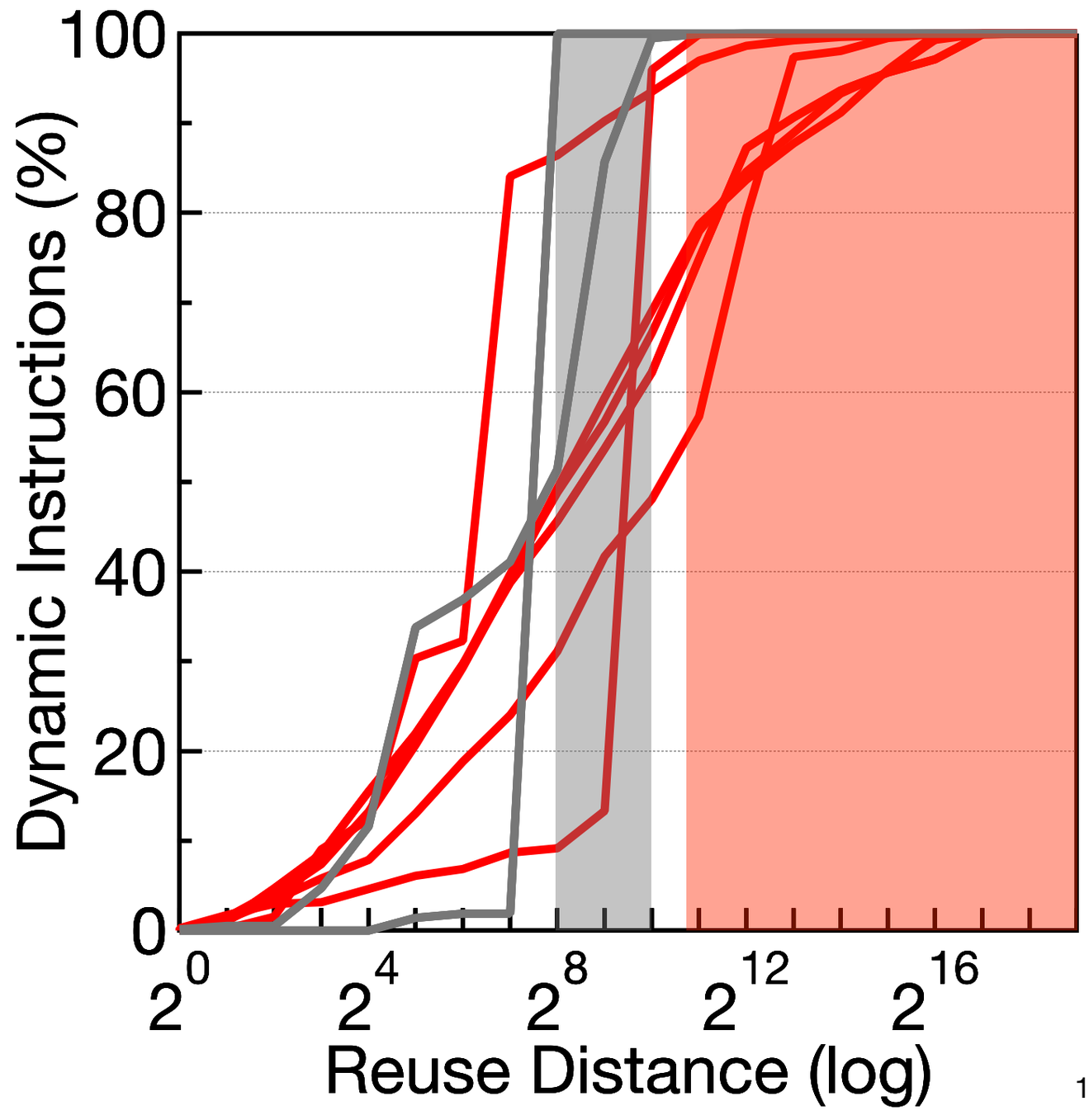


Root Cause Analysis

High I-$ Miss Ratio

↑

Large Instruction Reuse-distance



Root Cause Analysis

High I-$ Miss Ratio

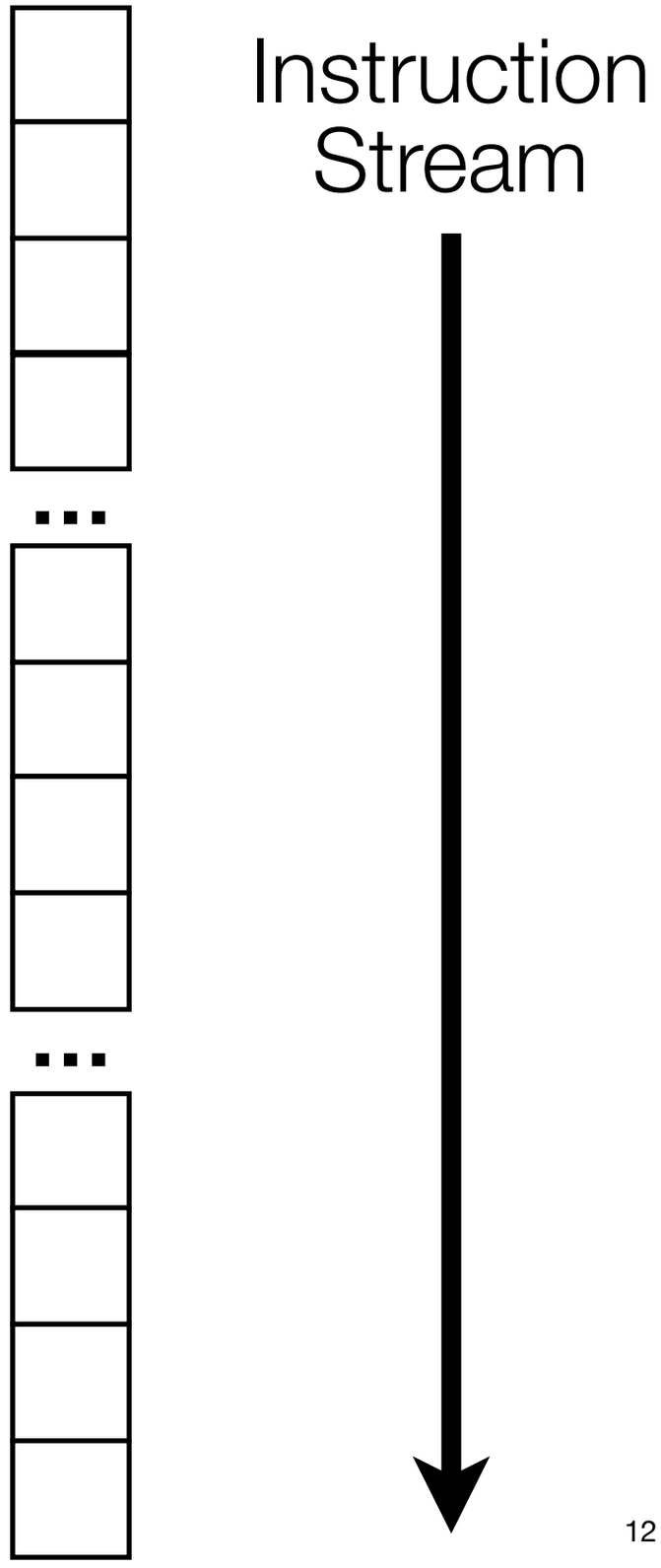
Large Instruction Reuse-distance

Root Cause Analysis

High I-$ Miss Ratio

↑

Large Instruction Reuse-distance

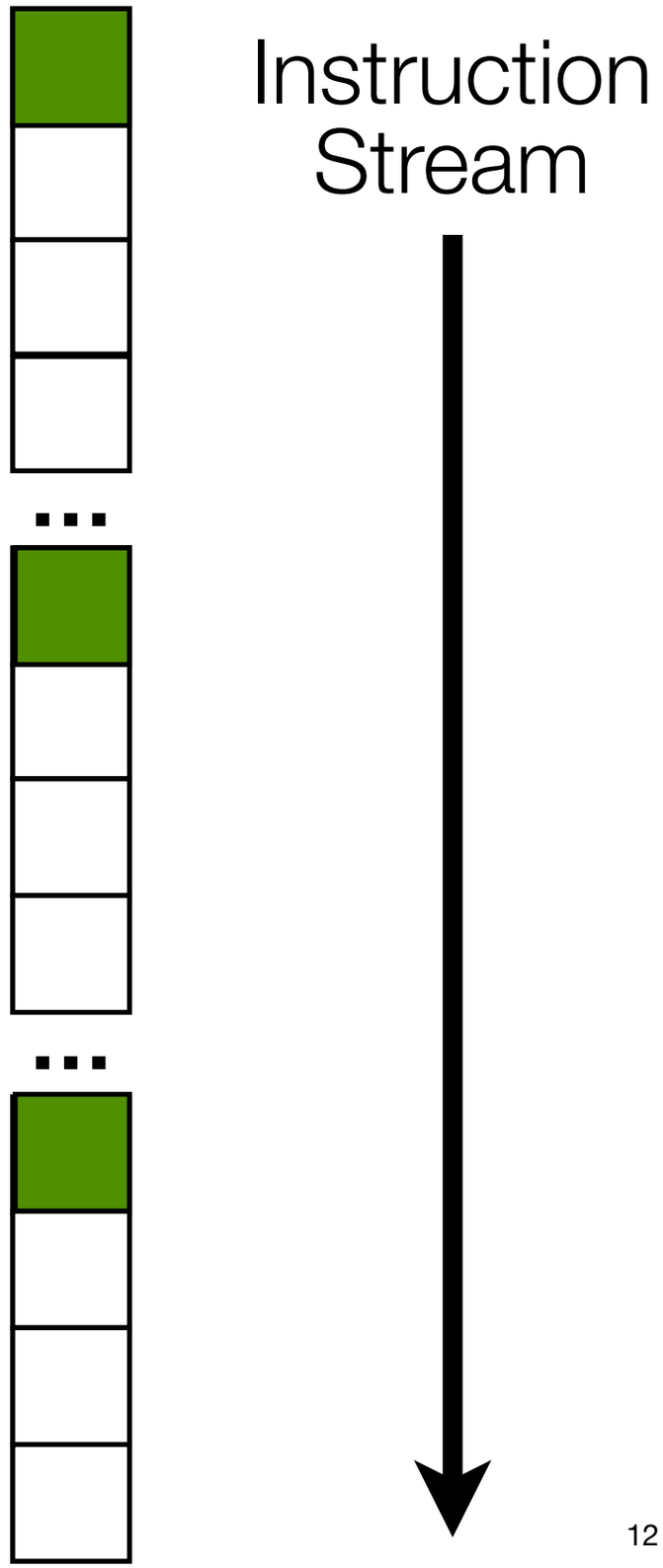


Root Cause Analysis

High I-$ Miss Ratio

↑

Large Instruction Reuse-distance



Root Cause Analysis

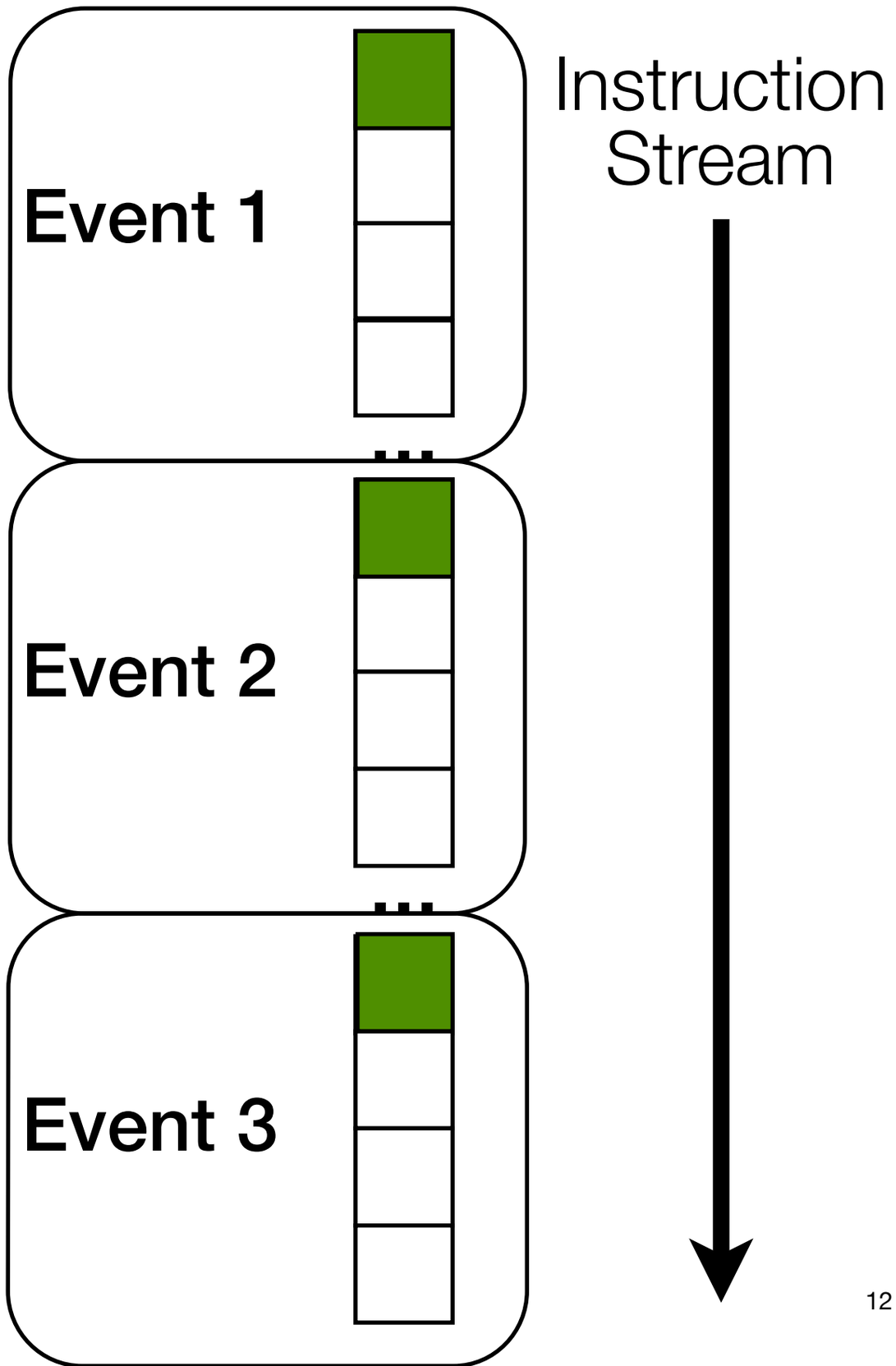
High I-$ Miss Ratio

↑

Large Instruction Reuse-distance

↗

Inter-event Code Reuse



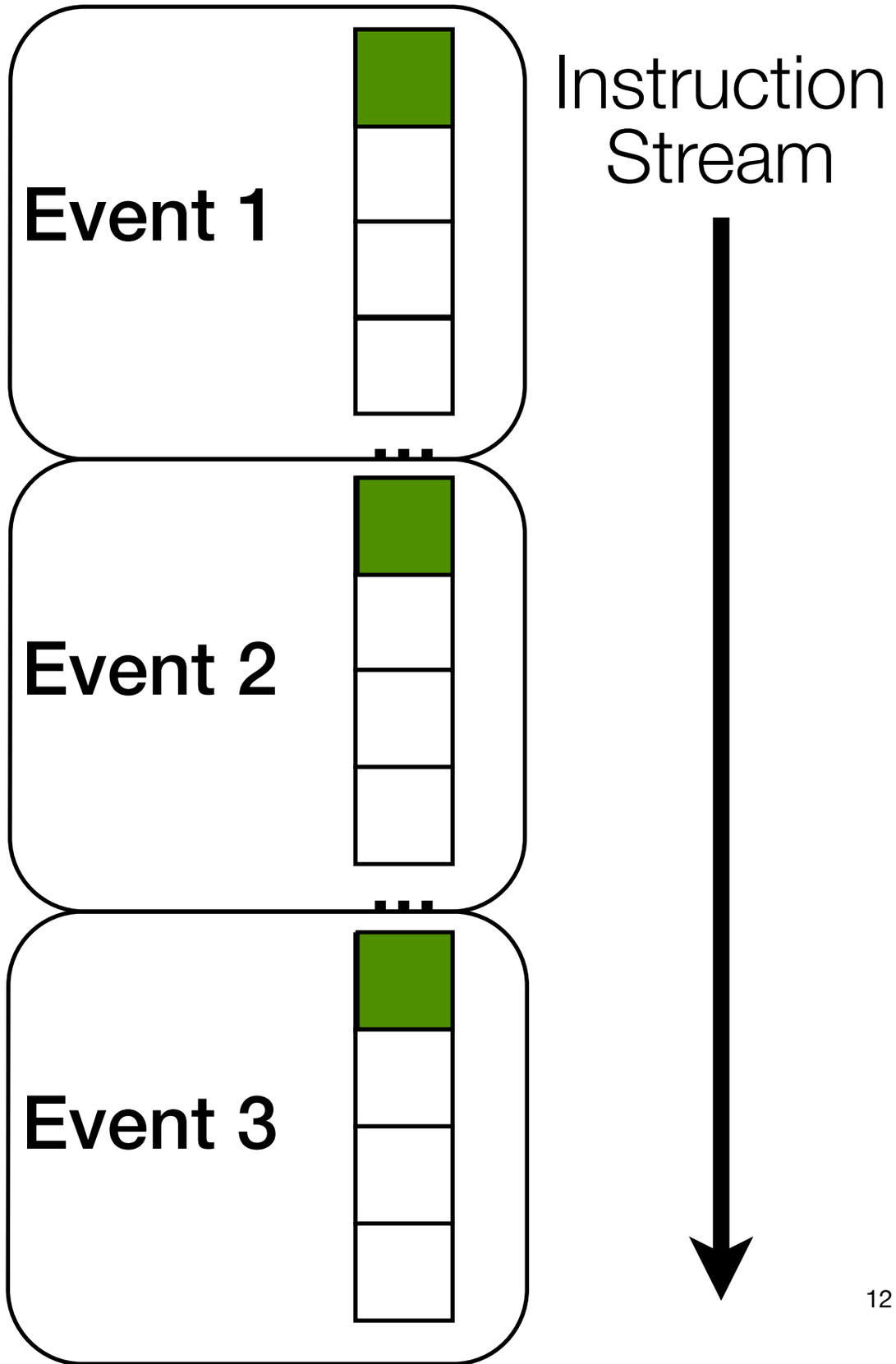
Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-distance

Inter-event Code Reuse

Large Event Footprint



Root Cause Analysis

High I-$ Miss Ratio

↑

Large Instruction Reuse-distance

↗ ↖

Inter-event Code Reuse

Large Event Footprint

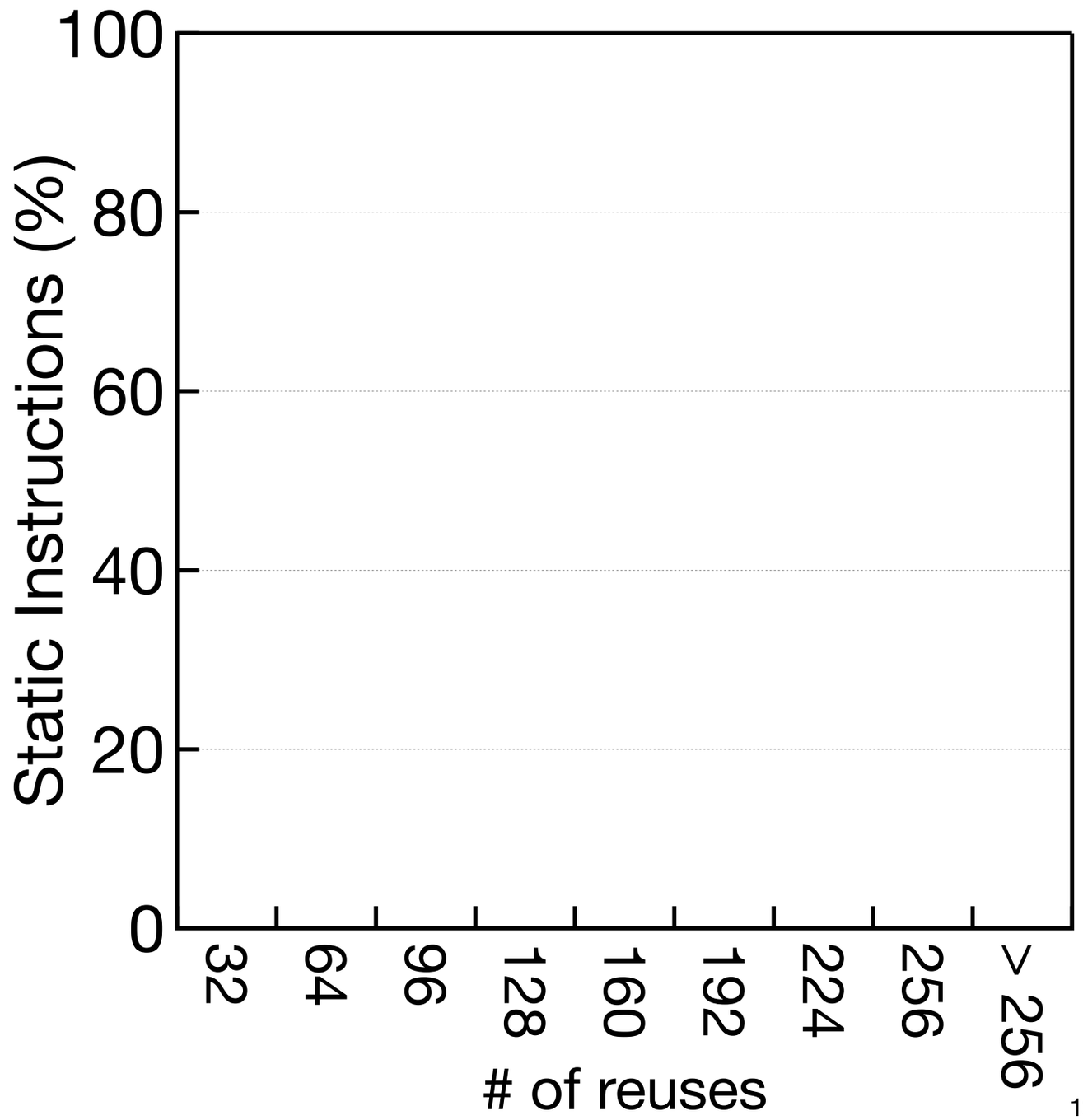
Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-distance

Inter-event Code Reuse

Large Event Footprint



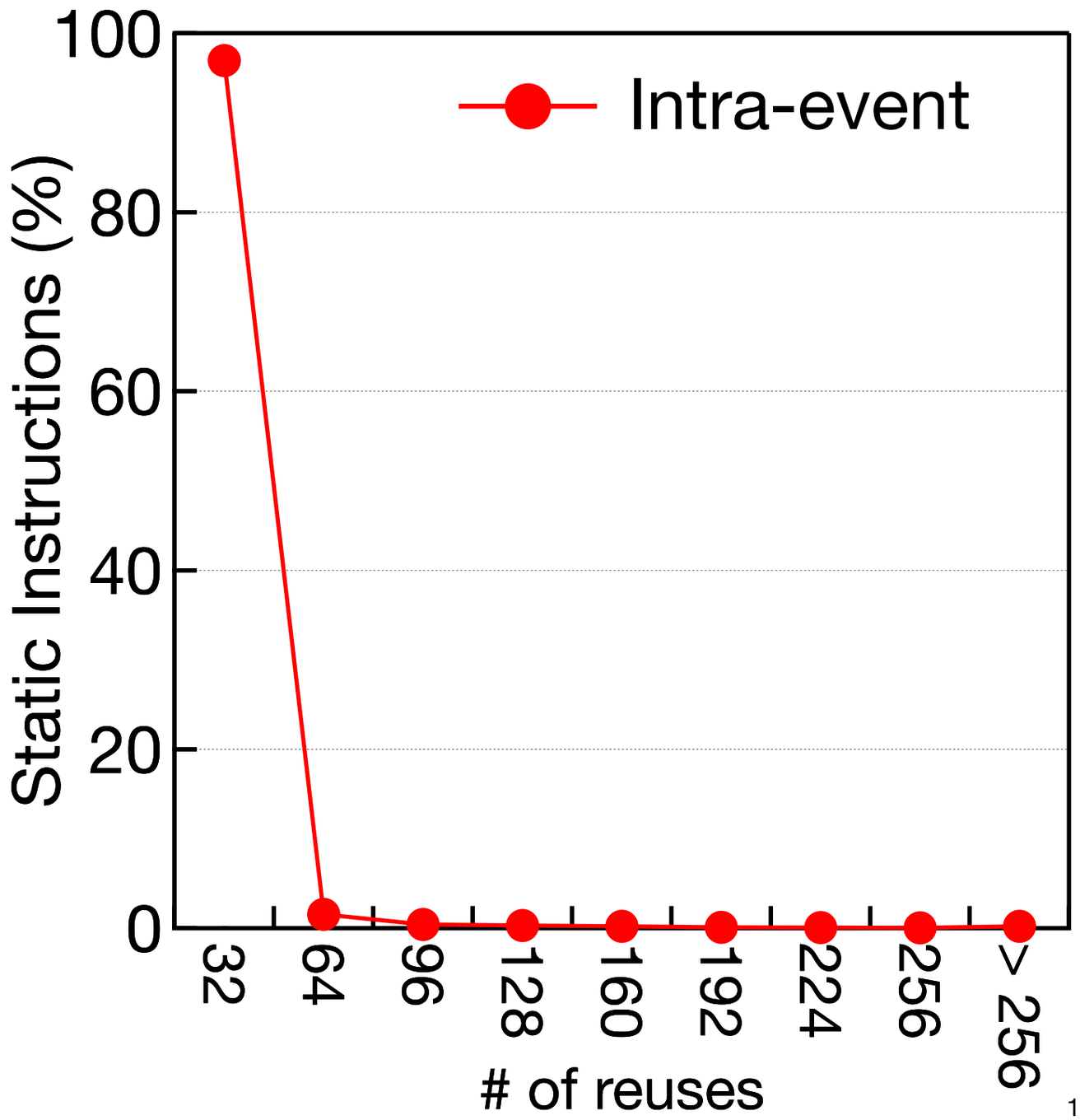
Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-distance

Inter-event Code Reuse

Large Event Footprint



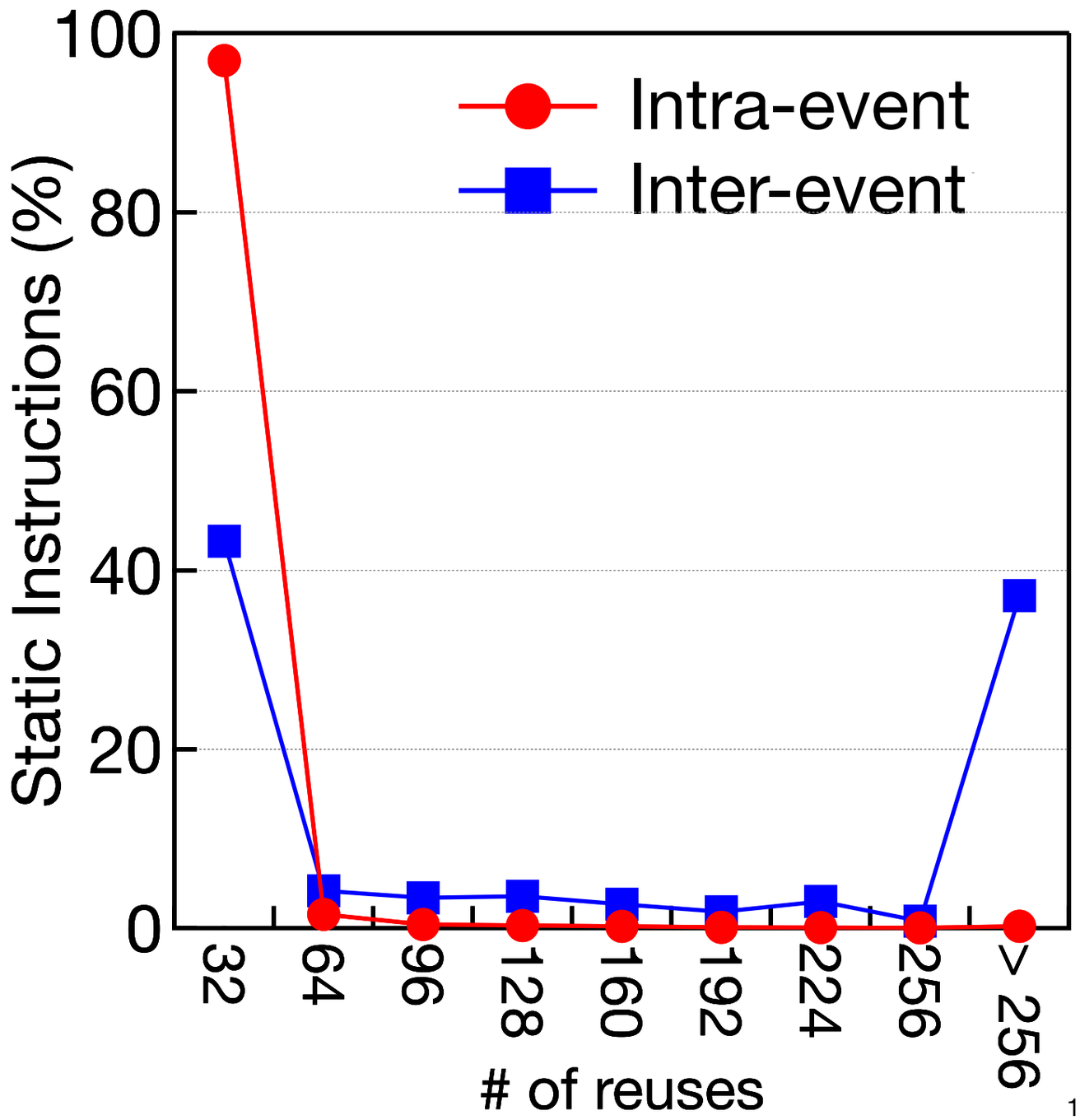
Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-distance

Inter-event Code Reuse

Large Event Footprint



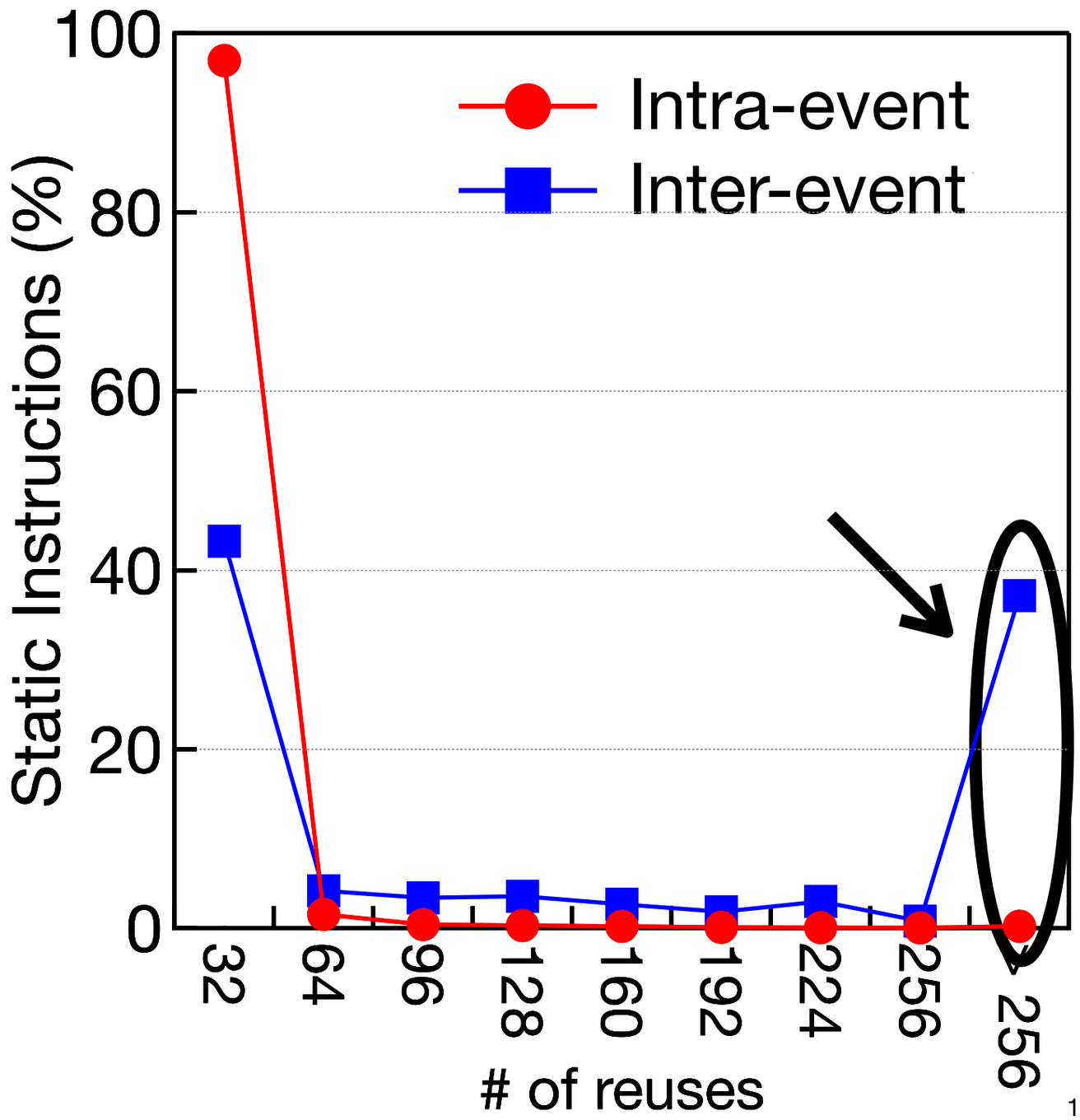
Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-distance

Inter-event Code Reuse

Large Event Footprint



Root Cause Analysis

High I-$ Miss Ratio

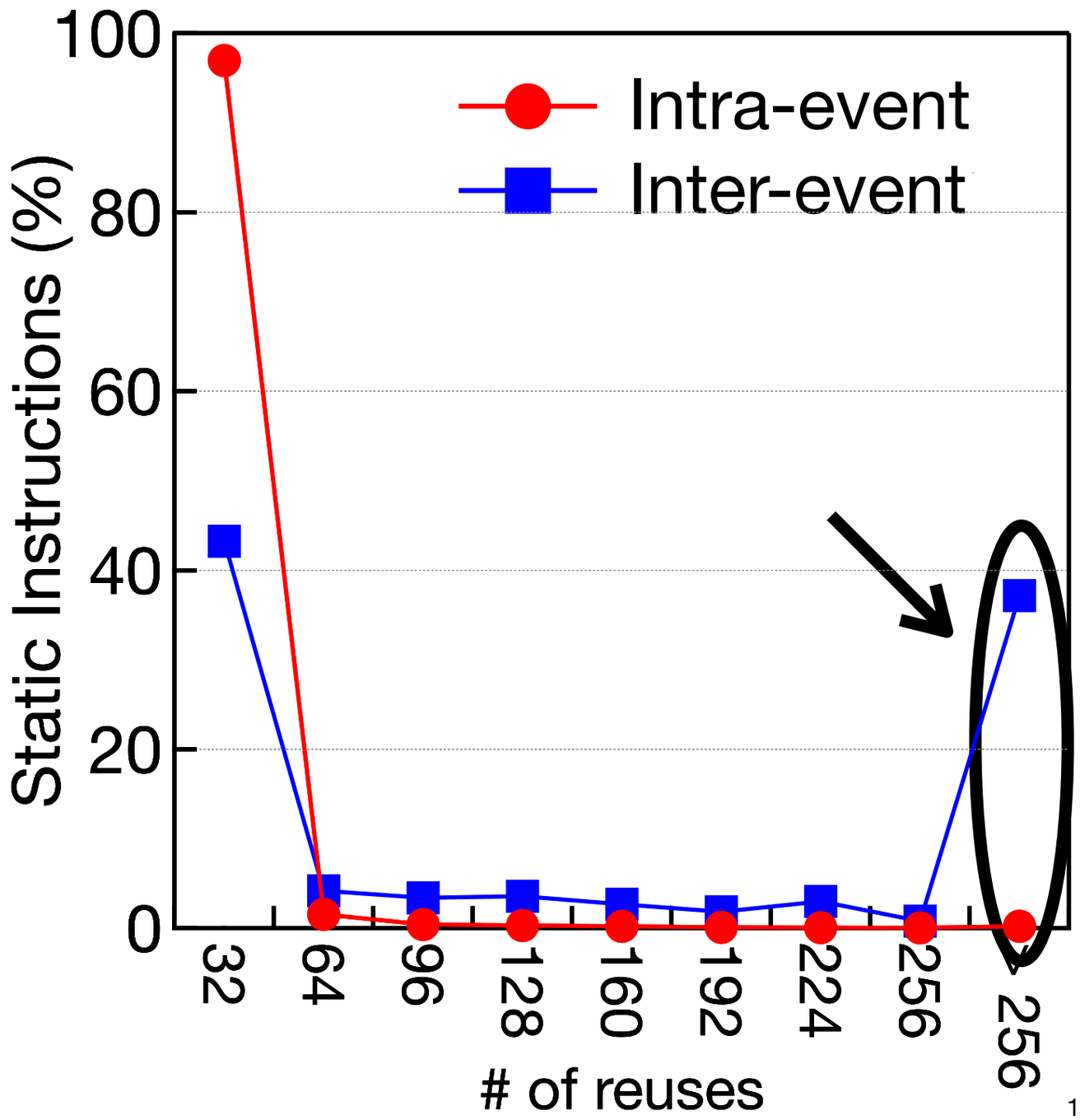
↑

Large Instruction Reuse-distance

↗ Inter-event Code Reuse

↖ Large Event Footprint

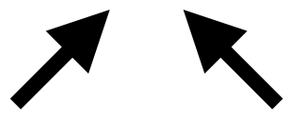
Most instruction reuses are inter-event.



Root Cause Analysis

High I-$ Miss Ratio

Large Instruction
Reuse-distance



Inter-event
Code Reuse

Large Event
Footprint

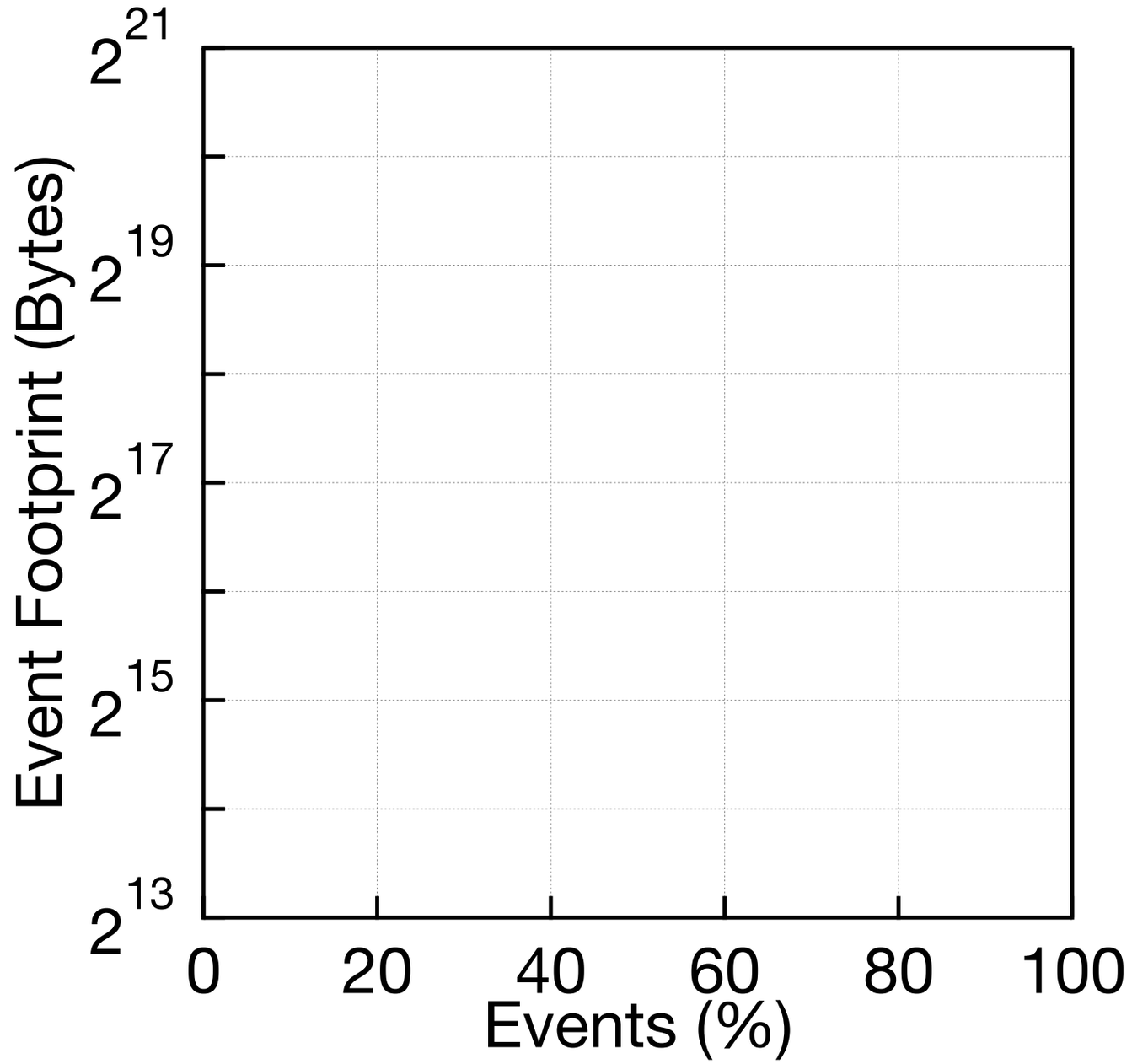
Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-distance

Inter-event Code Reuse

Large Event Footprint



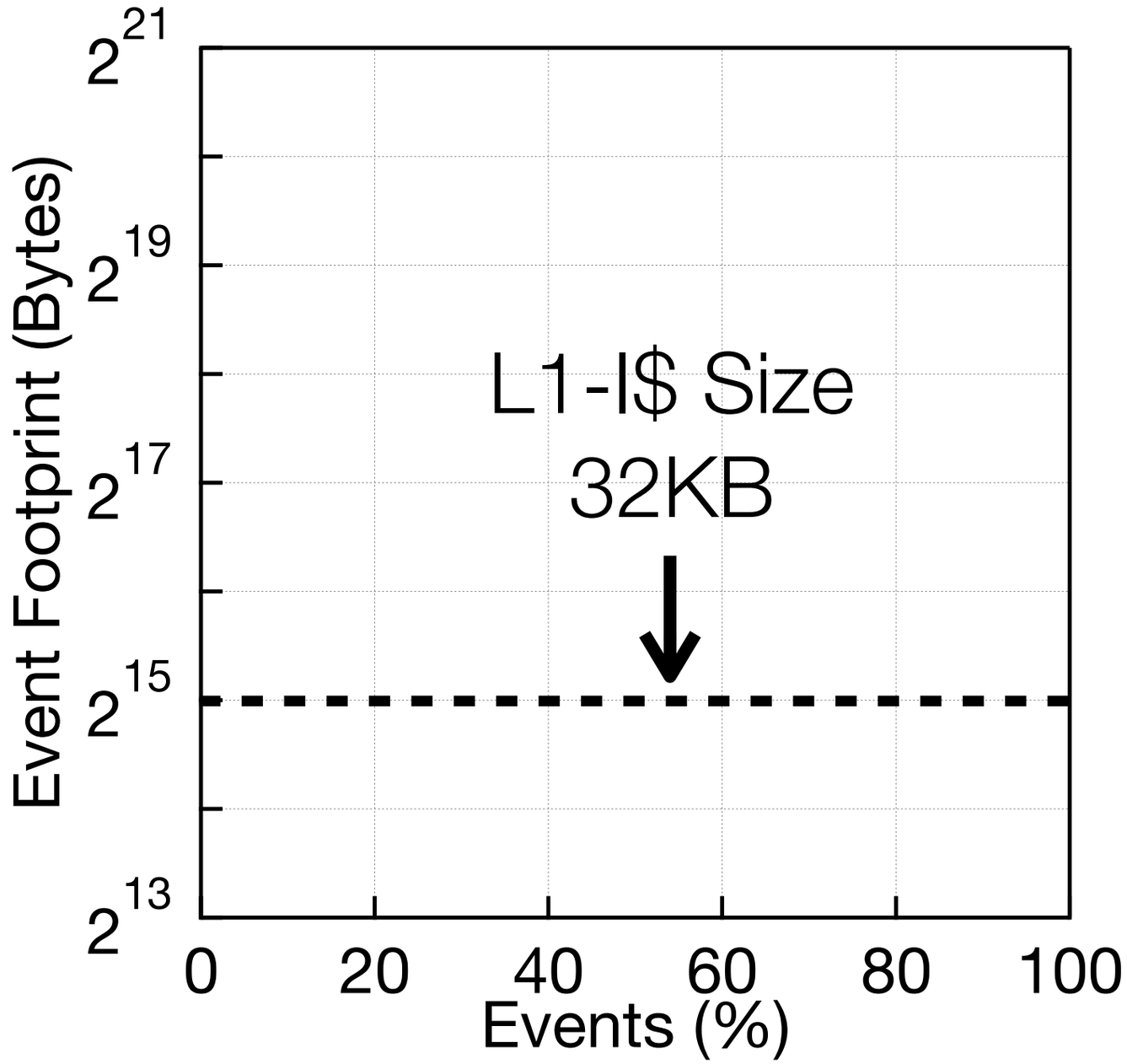
Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-distance

Inter-event Code Reuse

Large Event Footprint



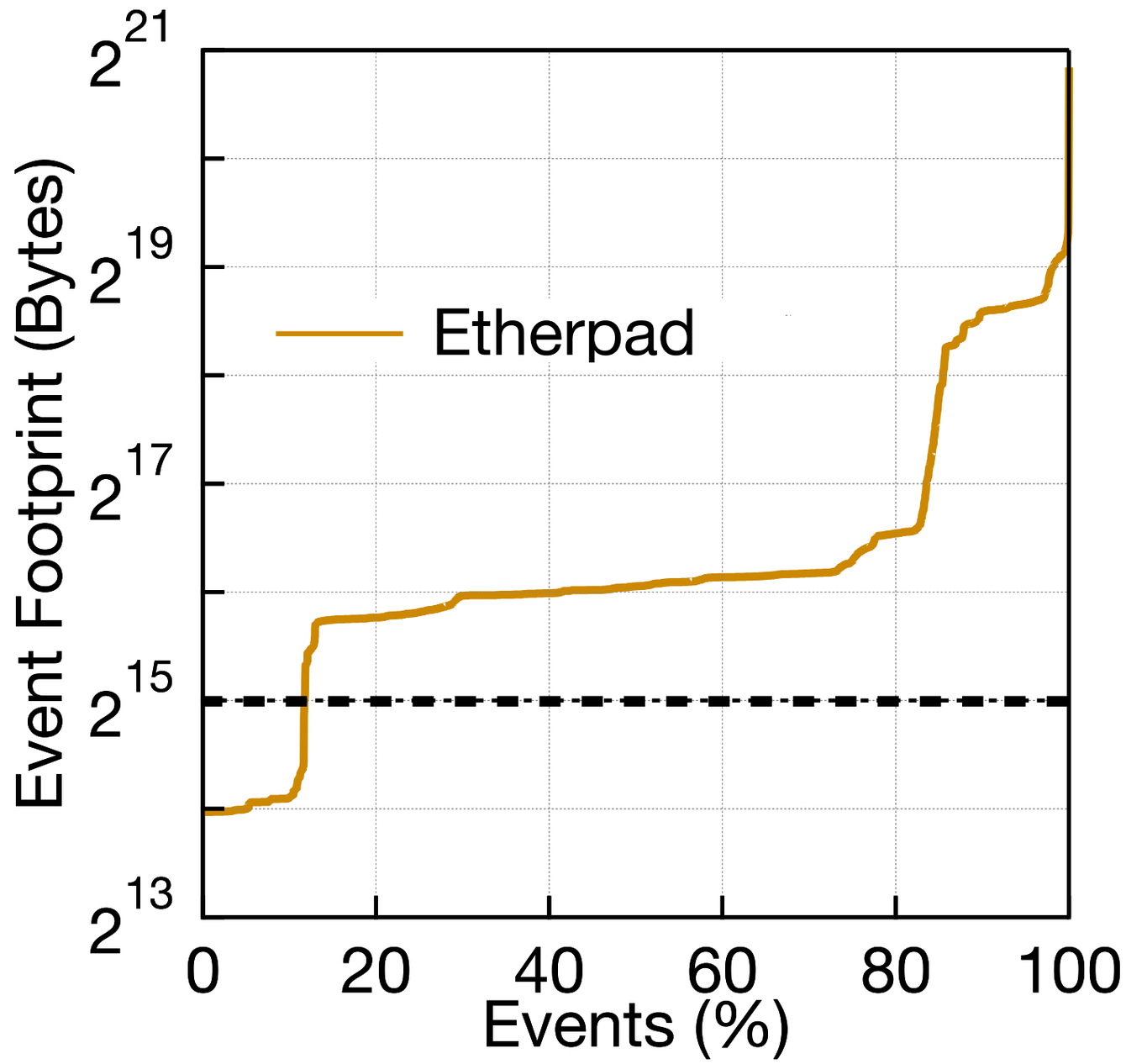
Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-distance

Inter-event Code Reuse

Large Event Footprint



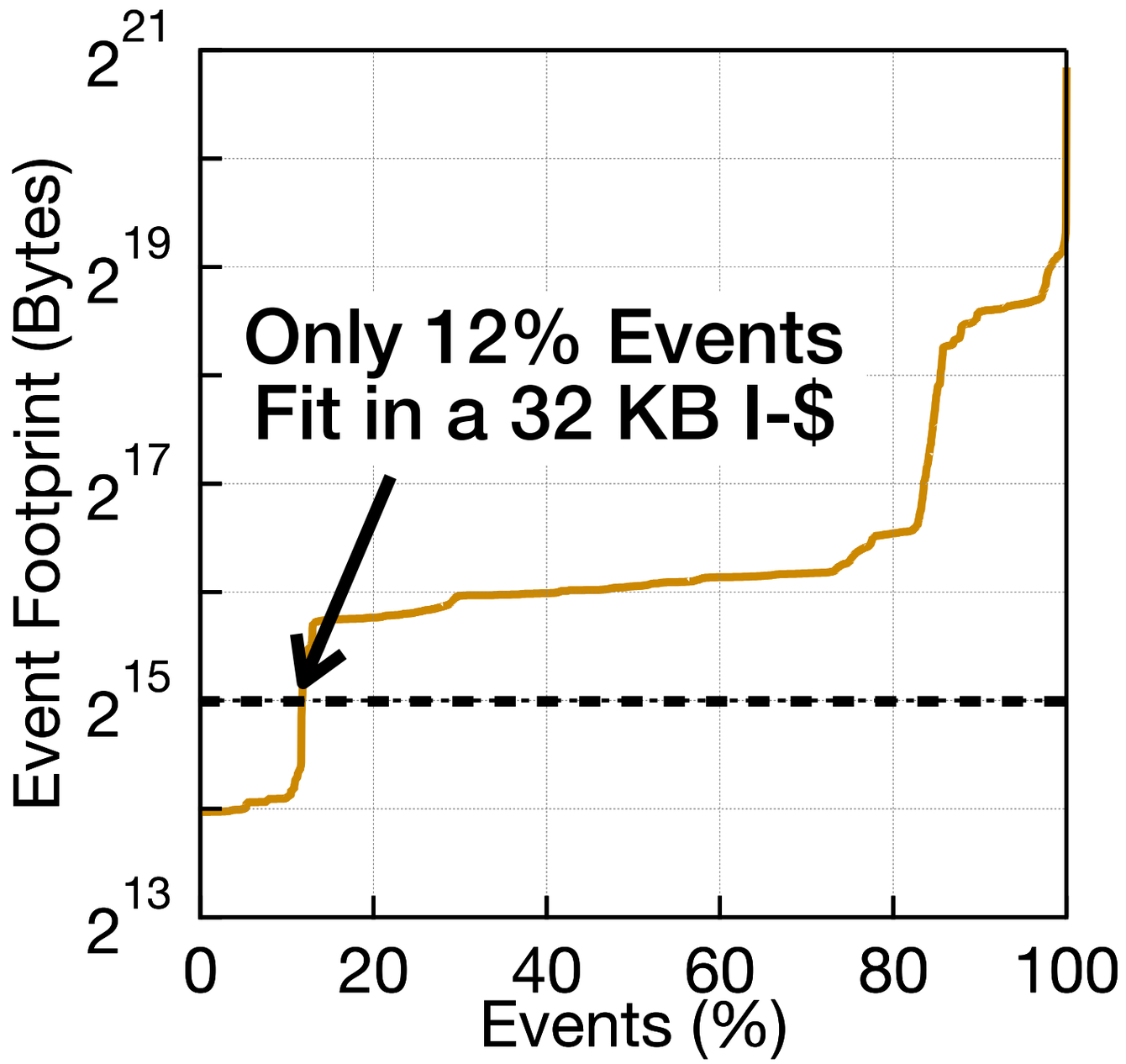
Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-distance

Inter-event Code Reuse

Large Event Footprint



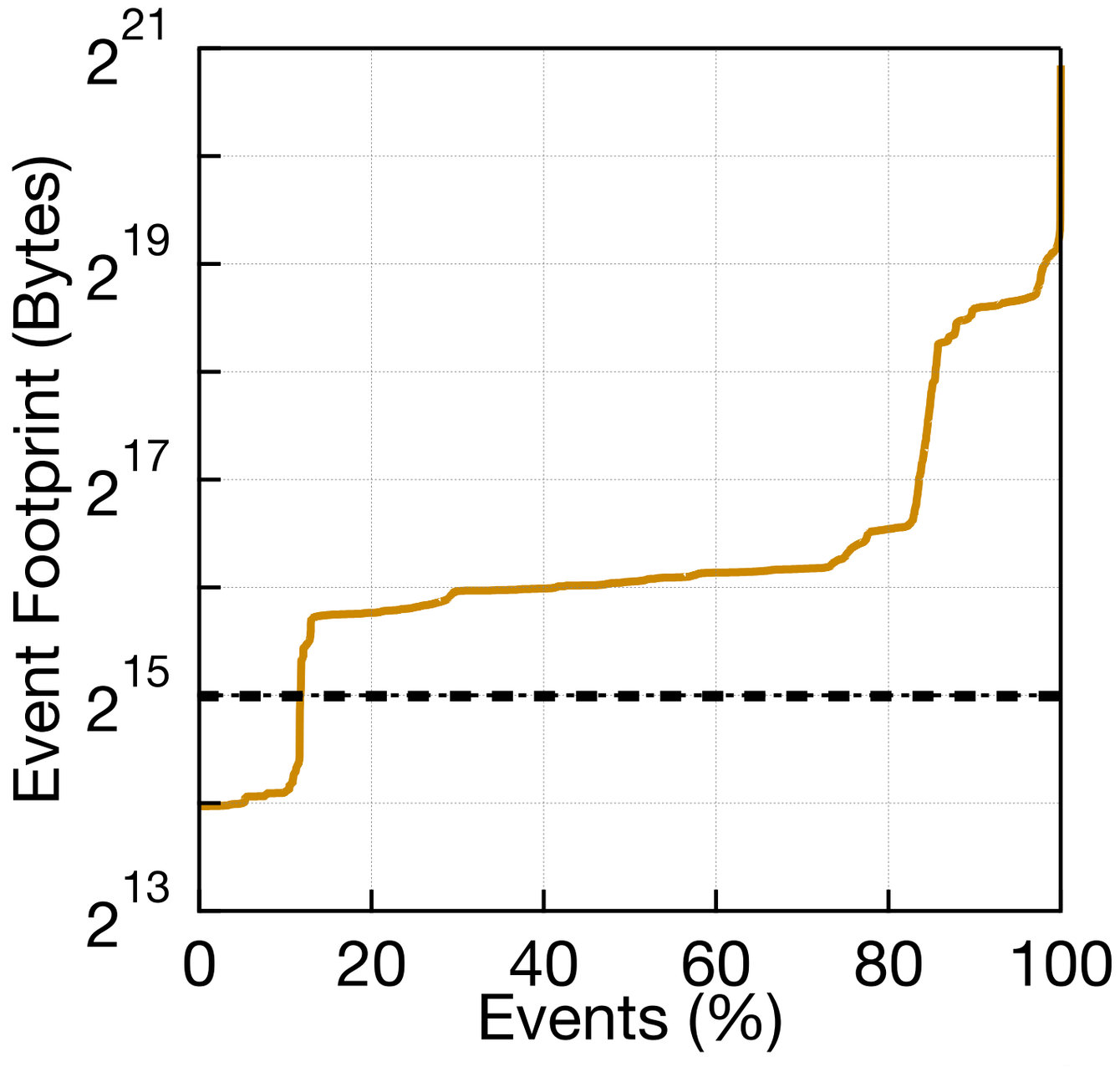
Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-distance

Inter-event Code Reuse

Large Event Footprint



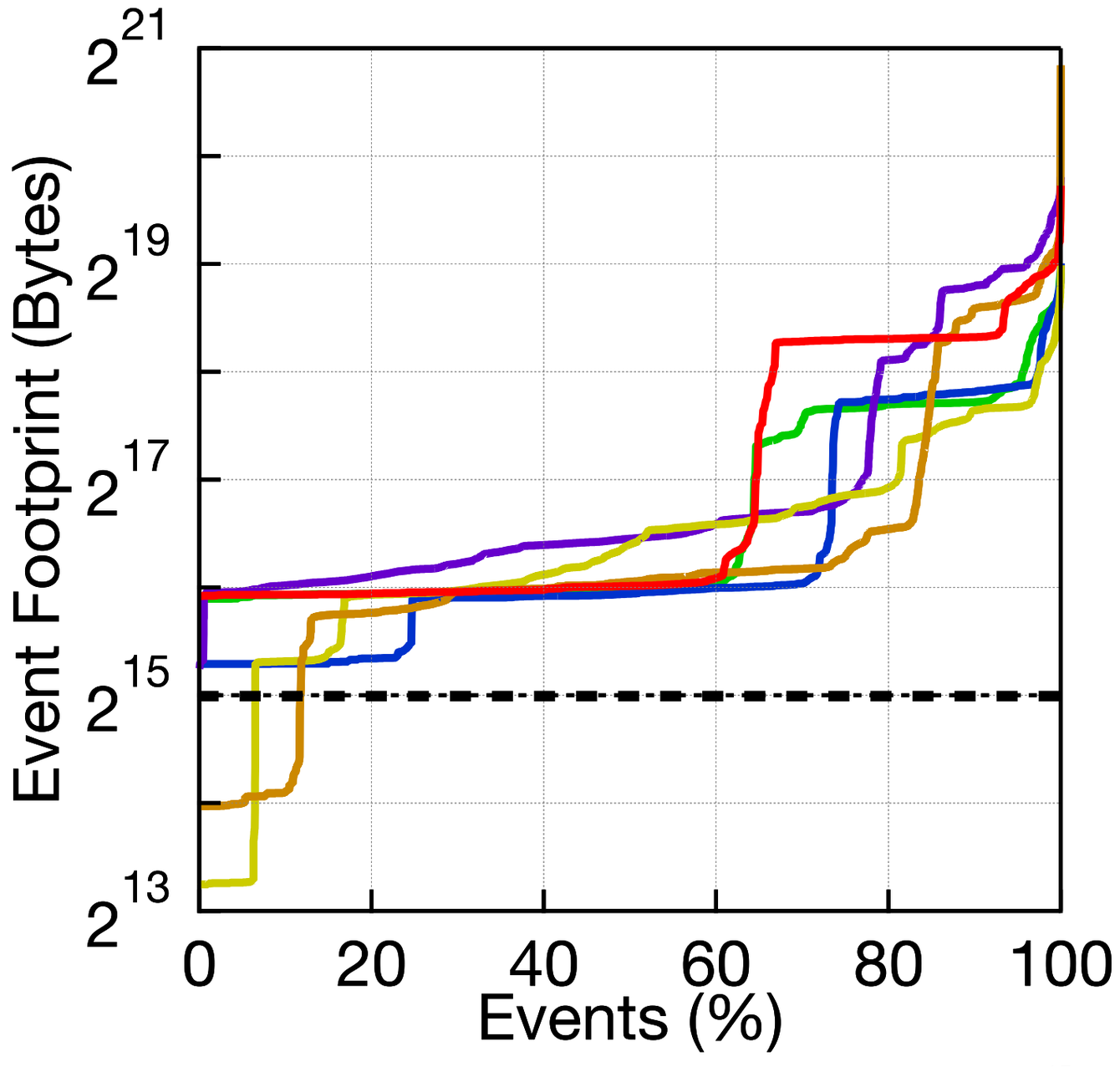
Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-distance

Inter-event Code Reuse

Large Event Footprint



Root Cause Analysis

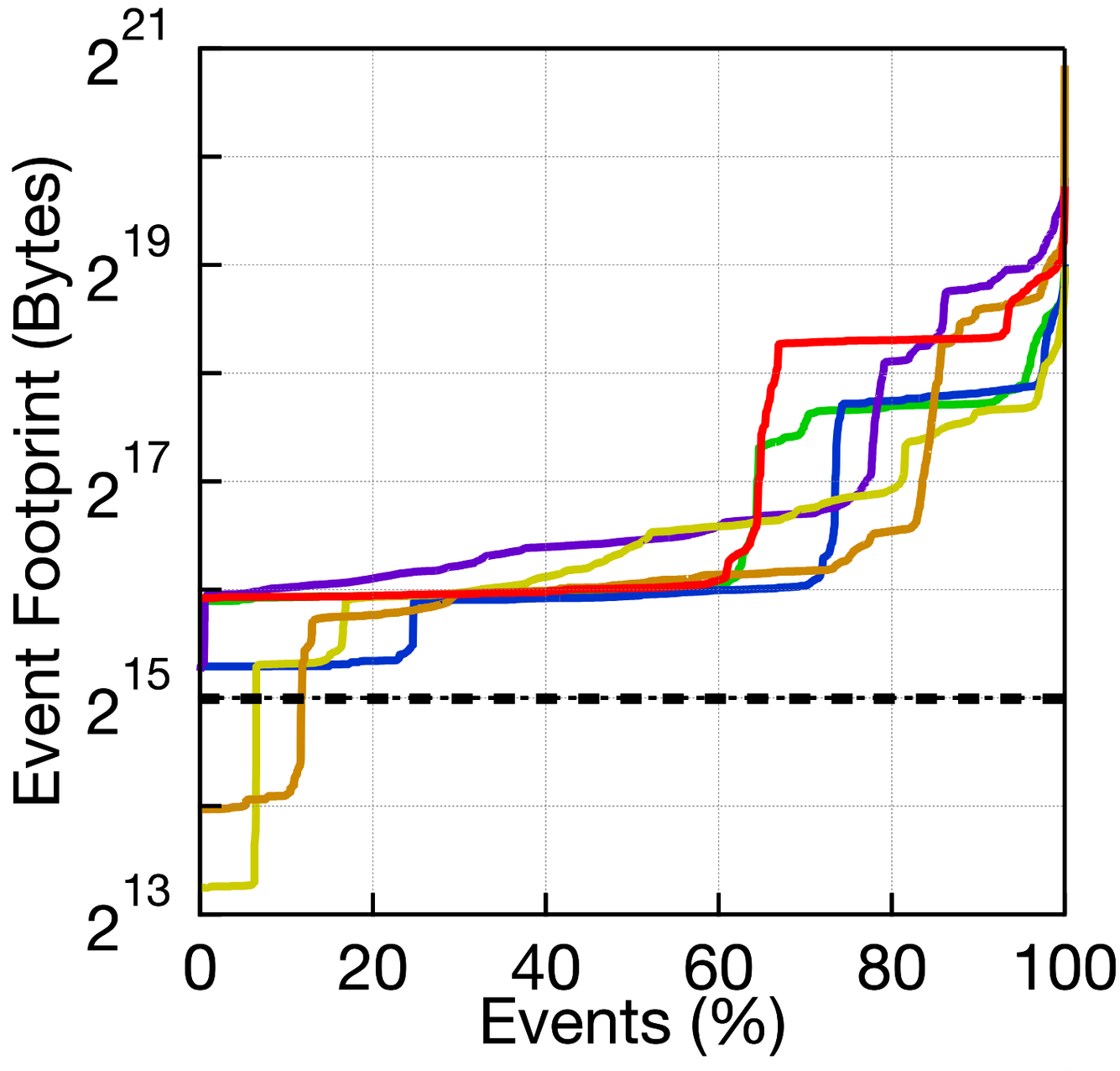
High I-$ Miss Ratio

Large Instruction Reuse-distance

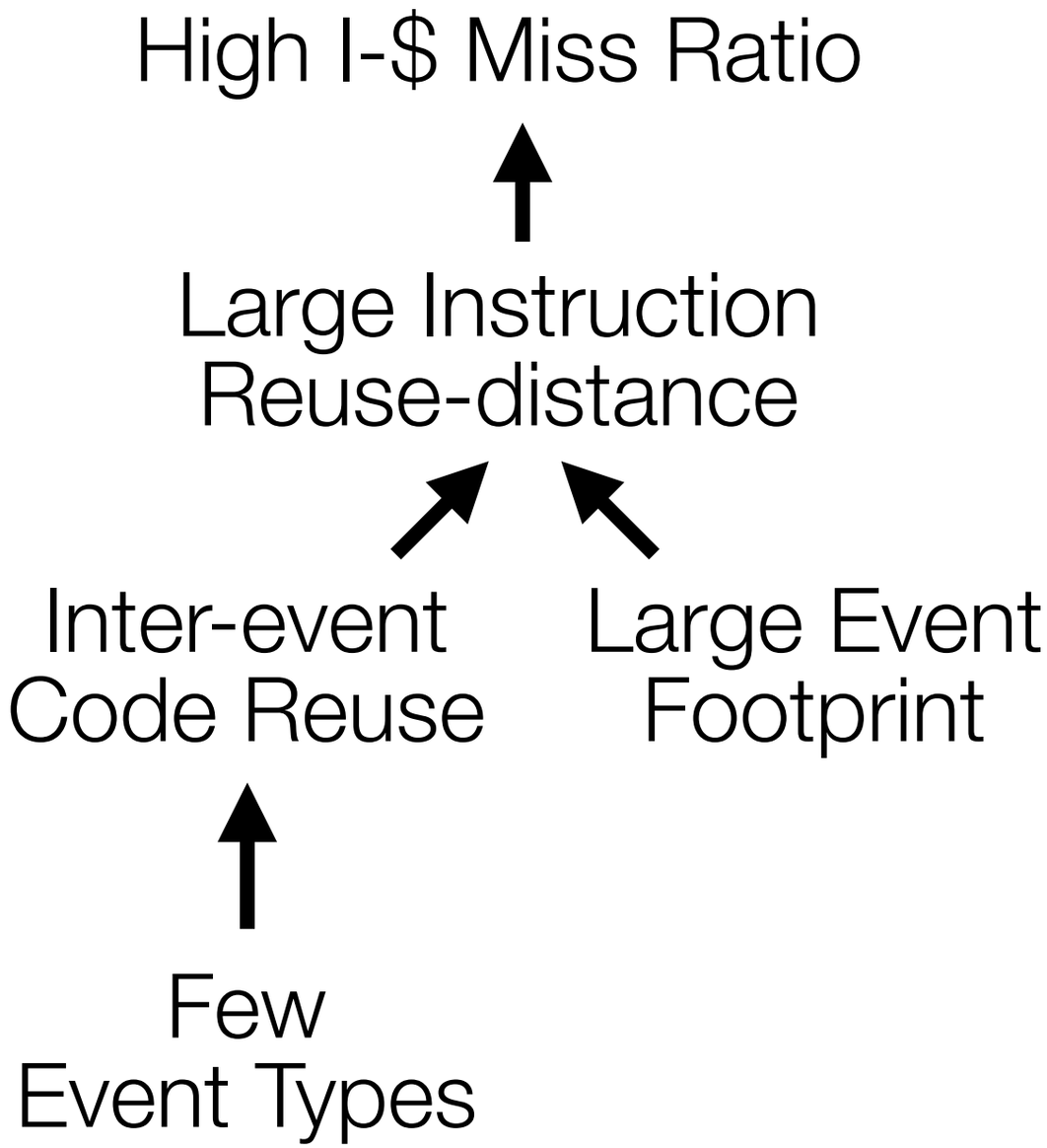
Inter-event Code Reuse

Large Event Footprint

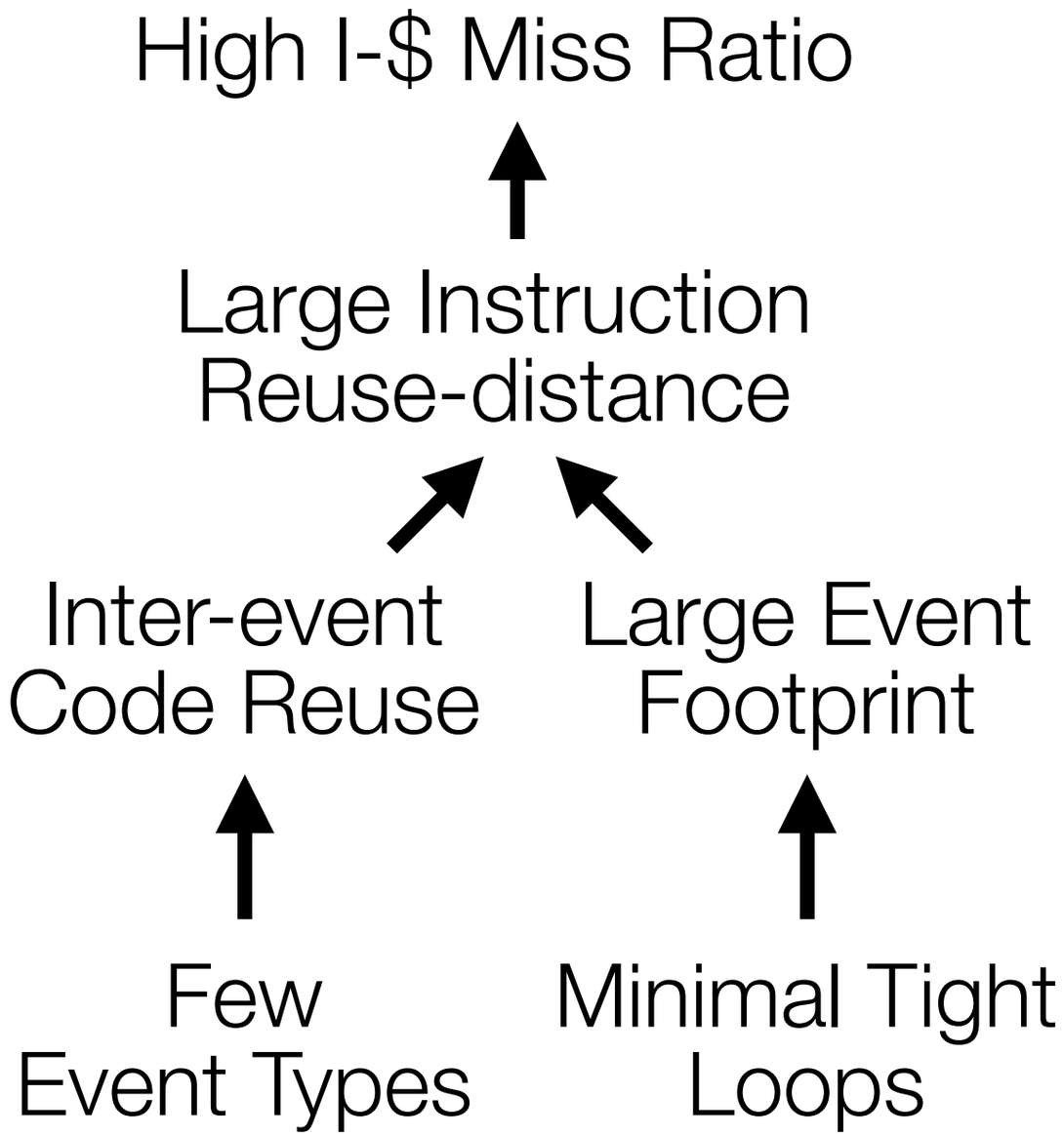
Most events' footprints do not fit in a typical I-$.



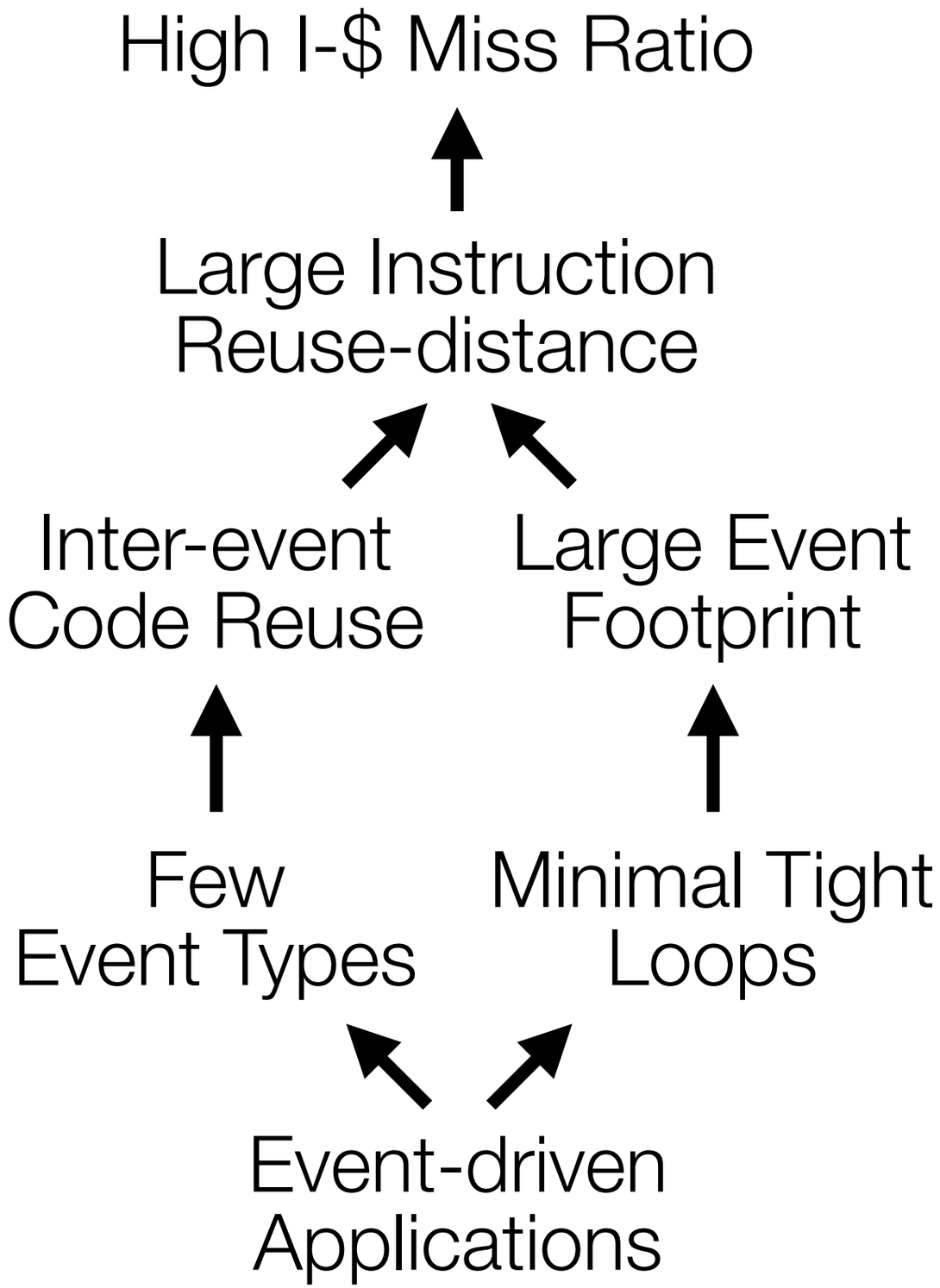
Root Cause Analysis



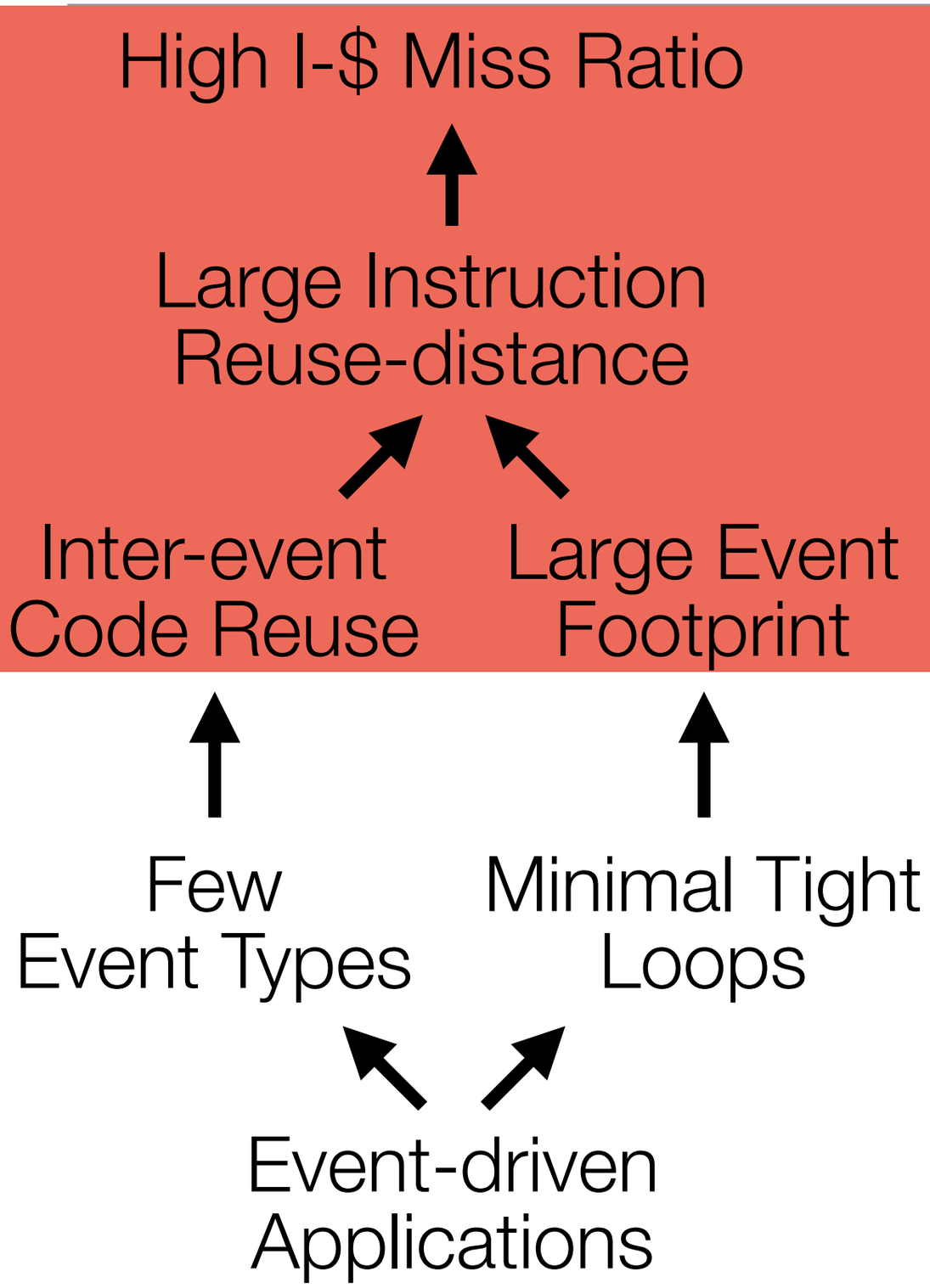
Root Cause Analysis



Root Cause Analysis

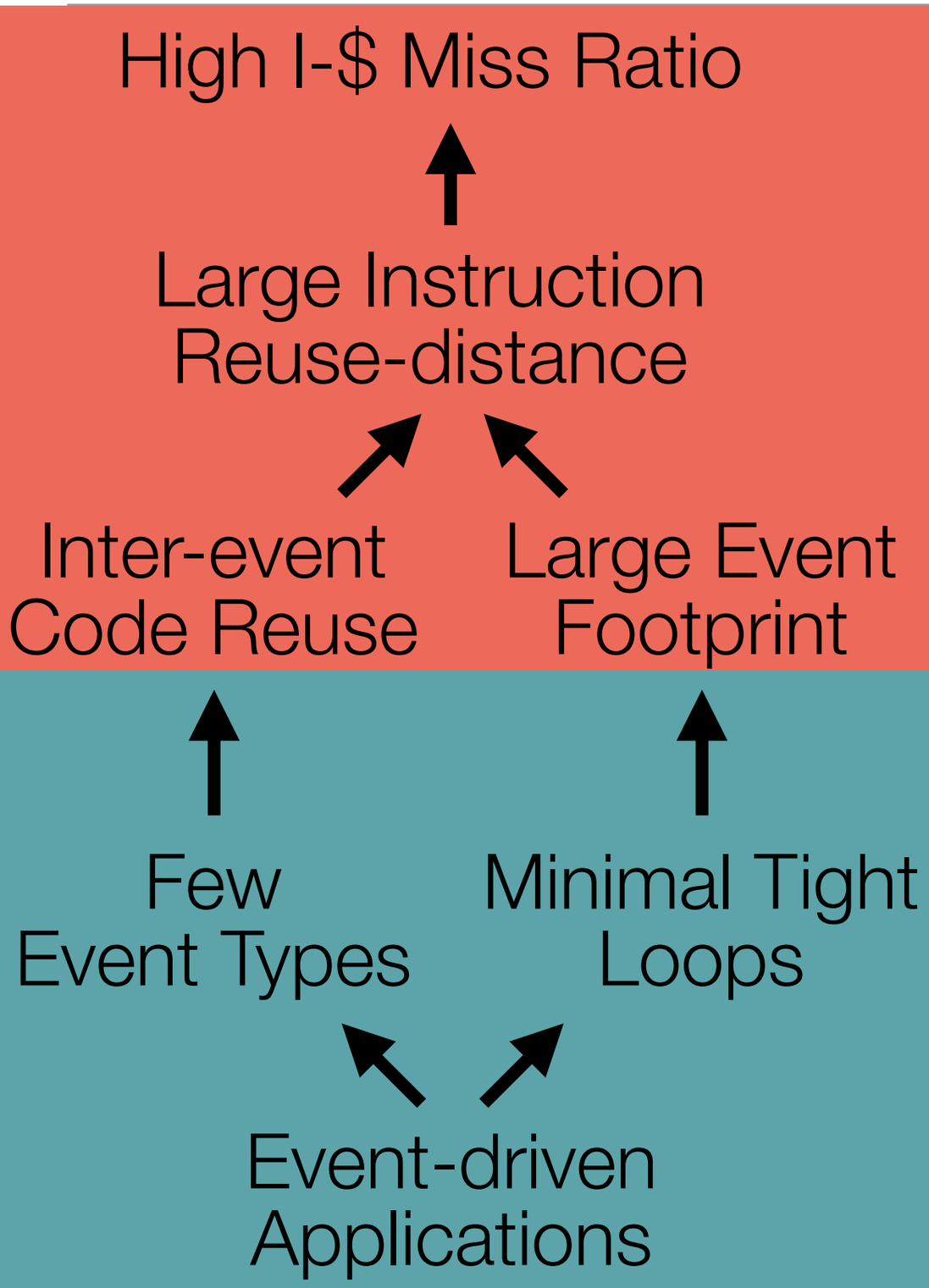


Root Cause Analysis



Microarchitecture Behaviors

Root Cause Analysis



Microarchitecture Behaviors

Application Characteristics

Root Cause Analysis

High I-$ Miss Ratio

Large Instruction Reuse-Distance

Inter-event Code Reuse

Large Event Footprint

Few Event Types

Minimal Tight Loops

Event-driven Applications

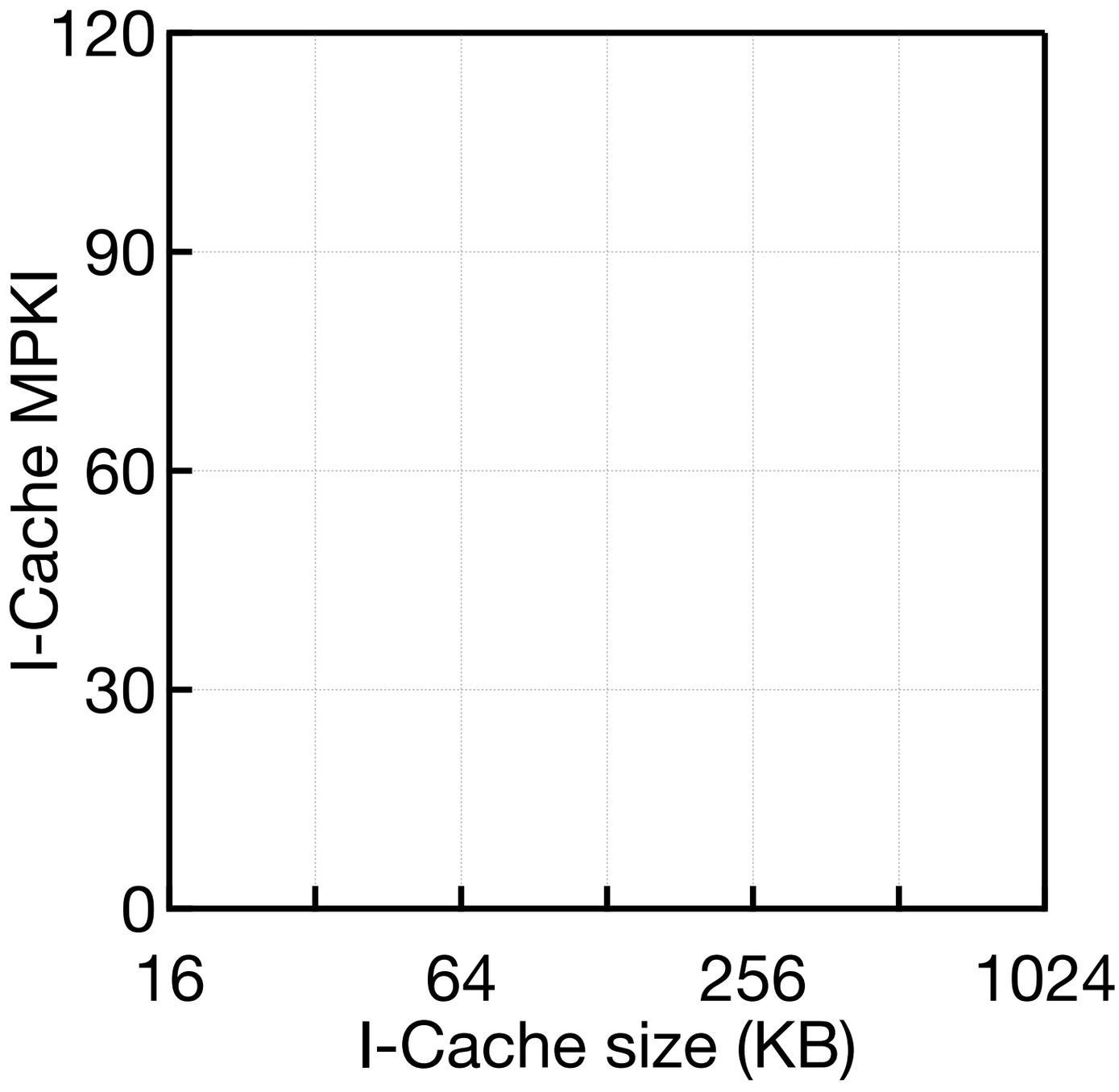
Microarchitecture Behaviors

Application Characteristics

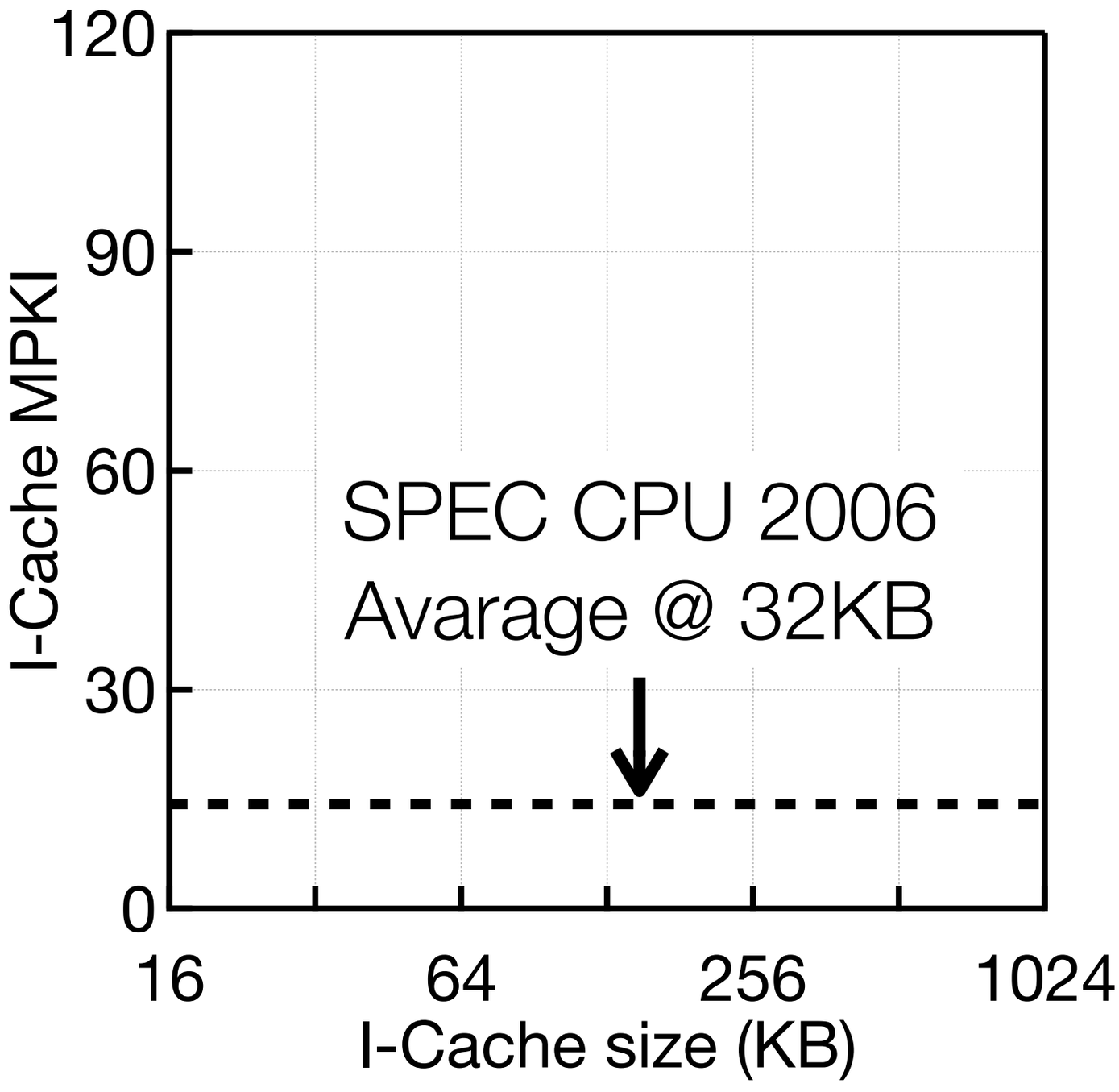
Can we better capture **instruction locality** to improve instruction supply efficiency?

Scale Up Hardware Resources?

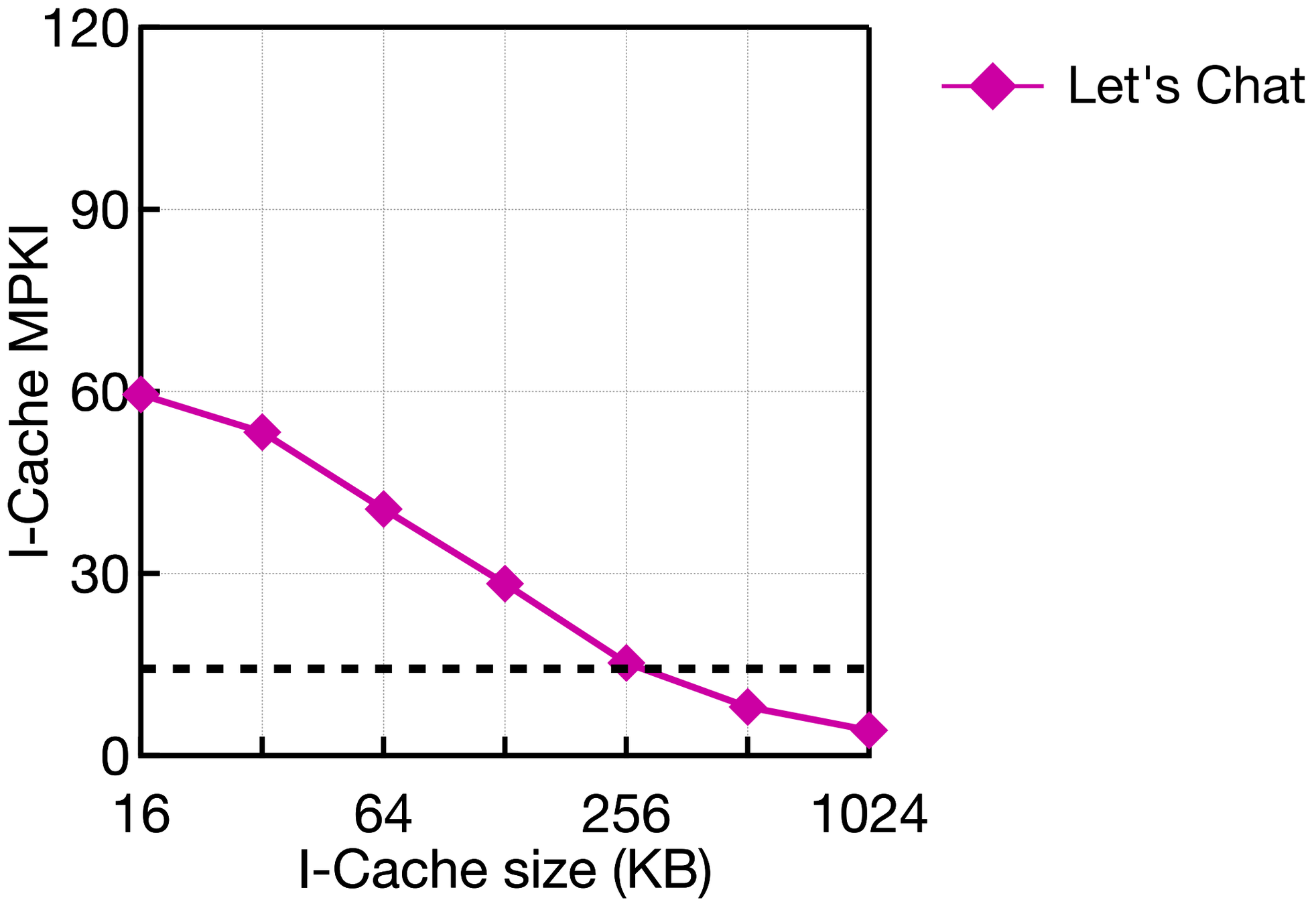
Scale Up Hardware Resources?



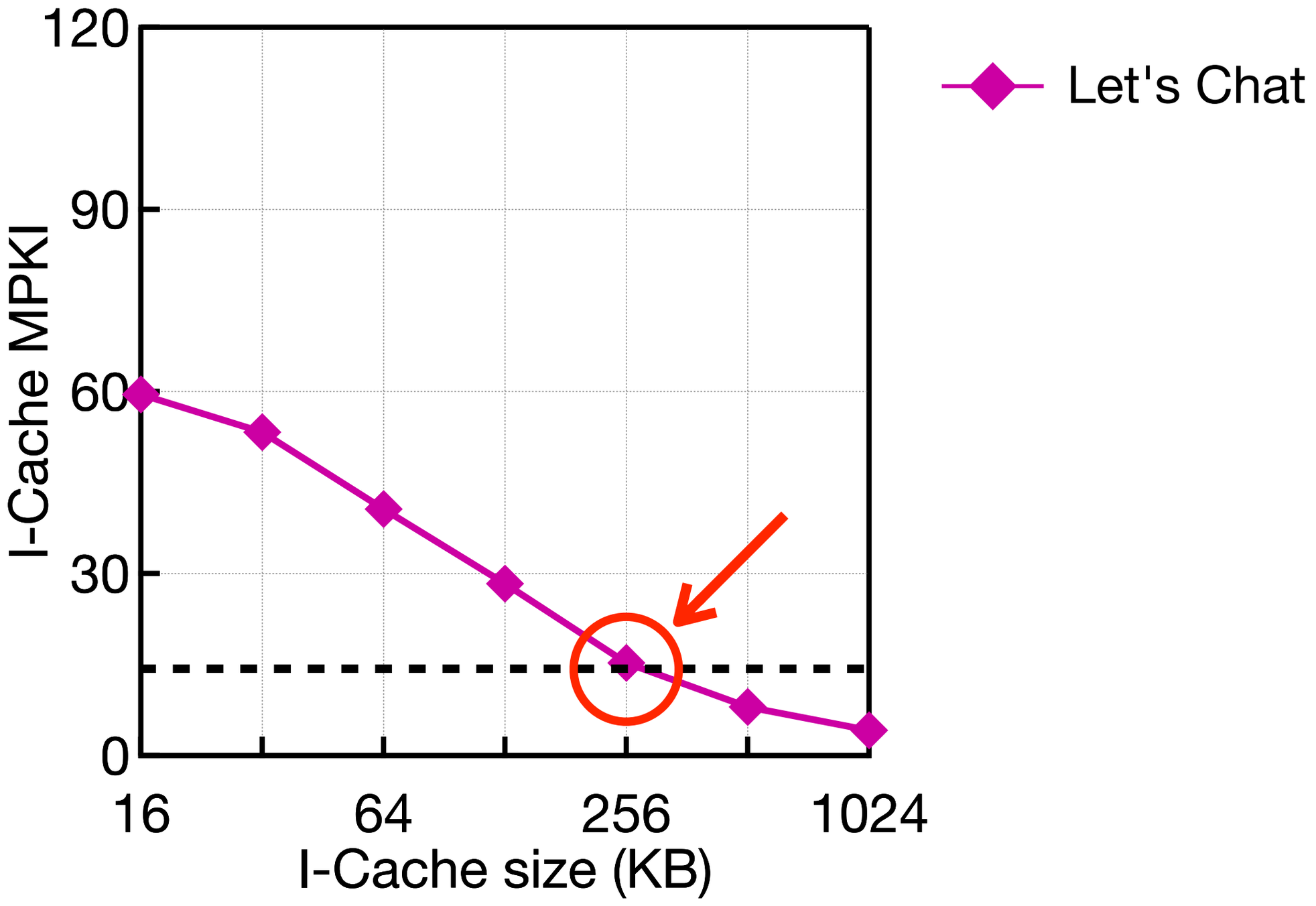
Scale Up Hardware Resources?



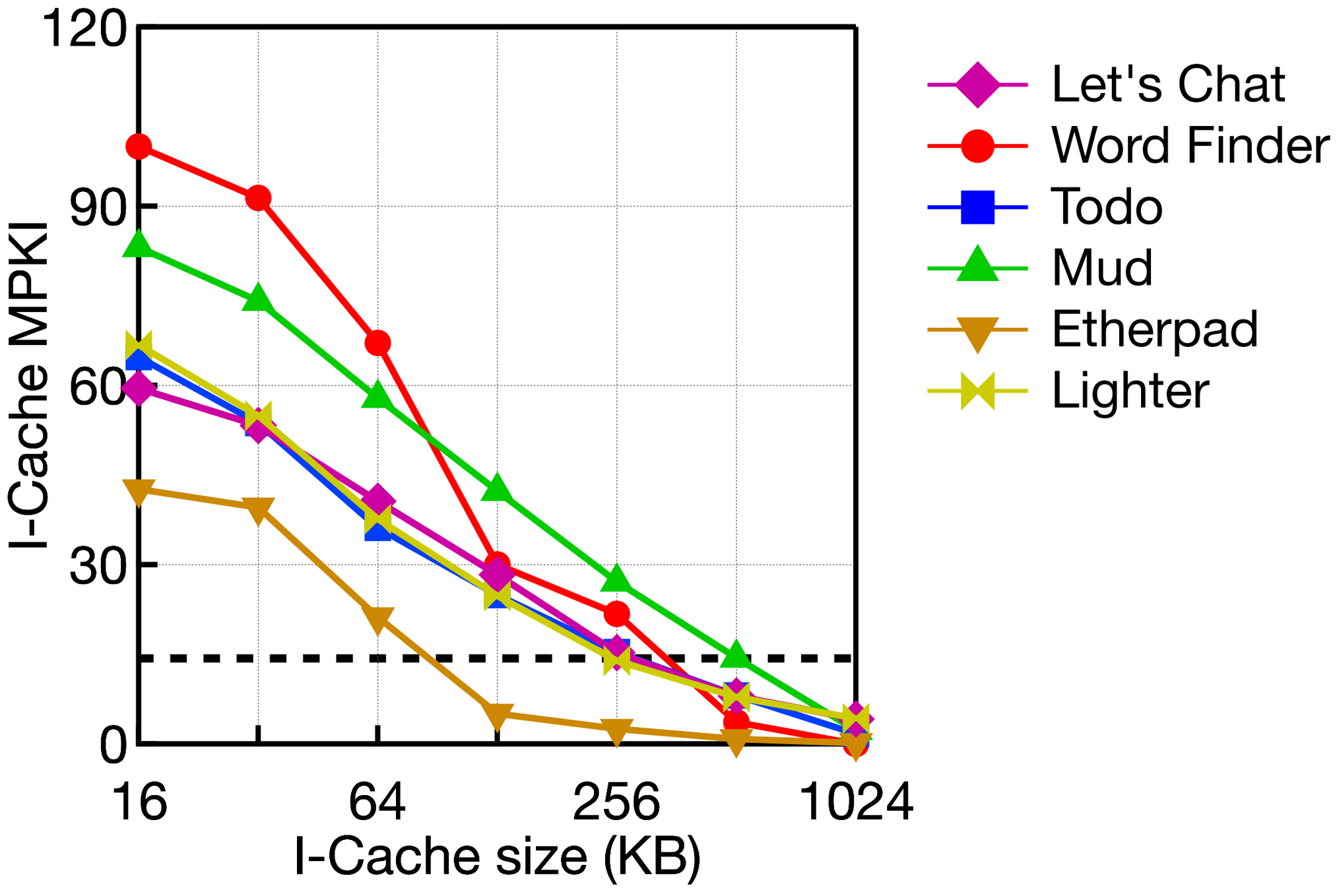
Scale Up Hardware Resources?



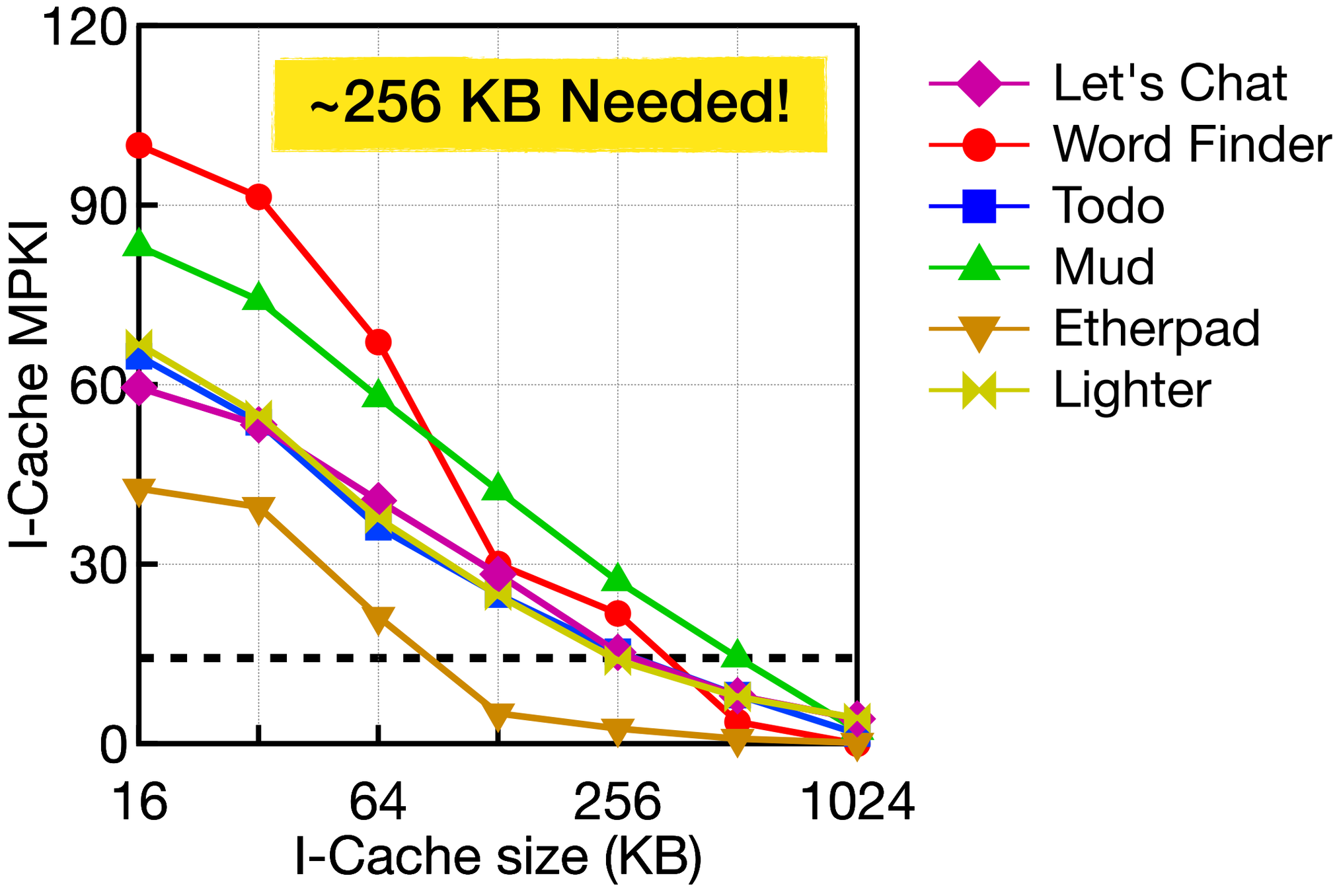
Scale Up Hardware Resources?



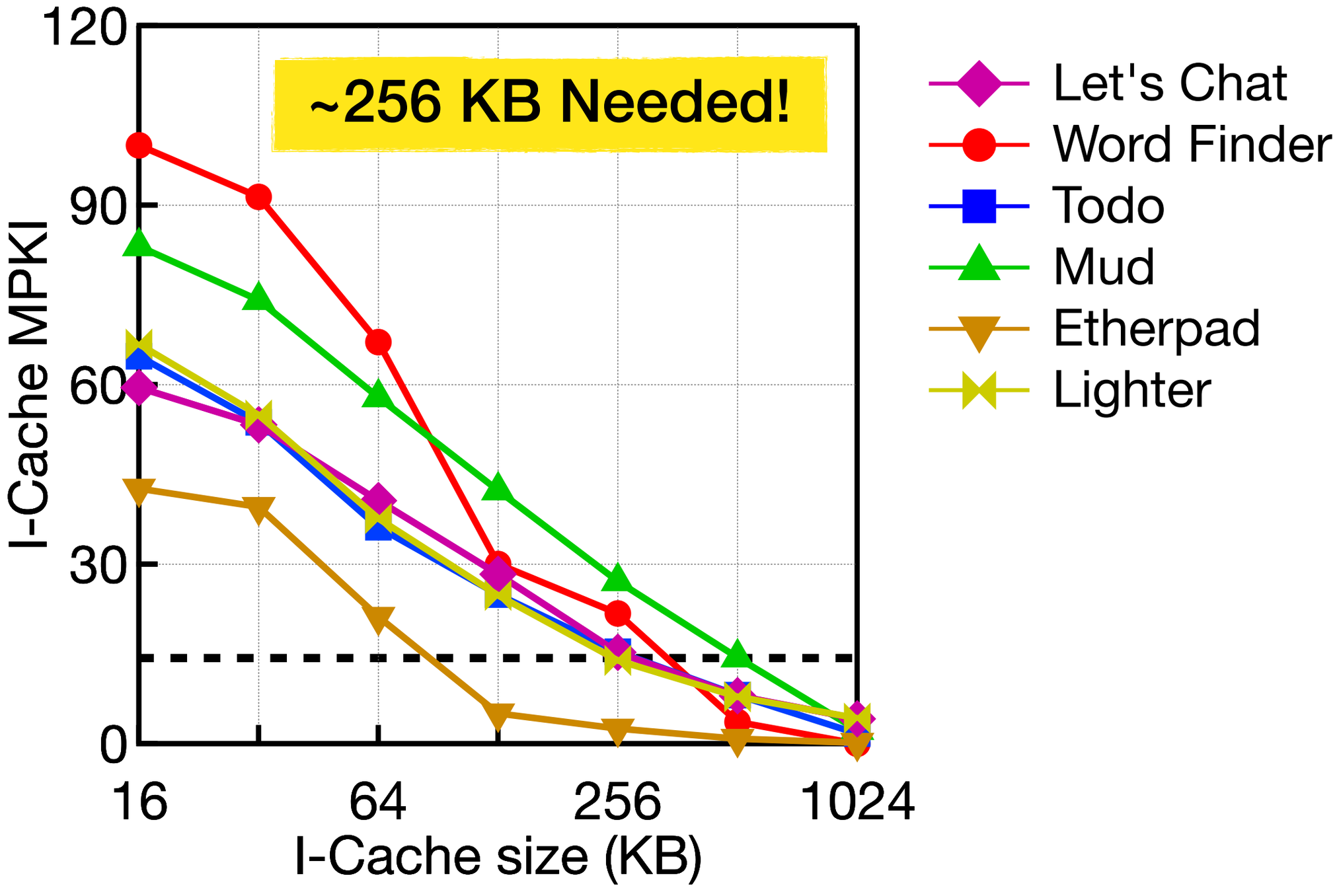
Scale Up Hardware Resources?



Scale Up Hardware Resources?

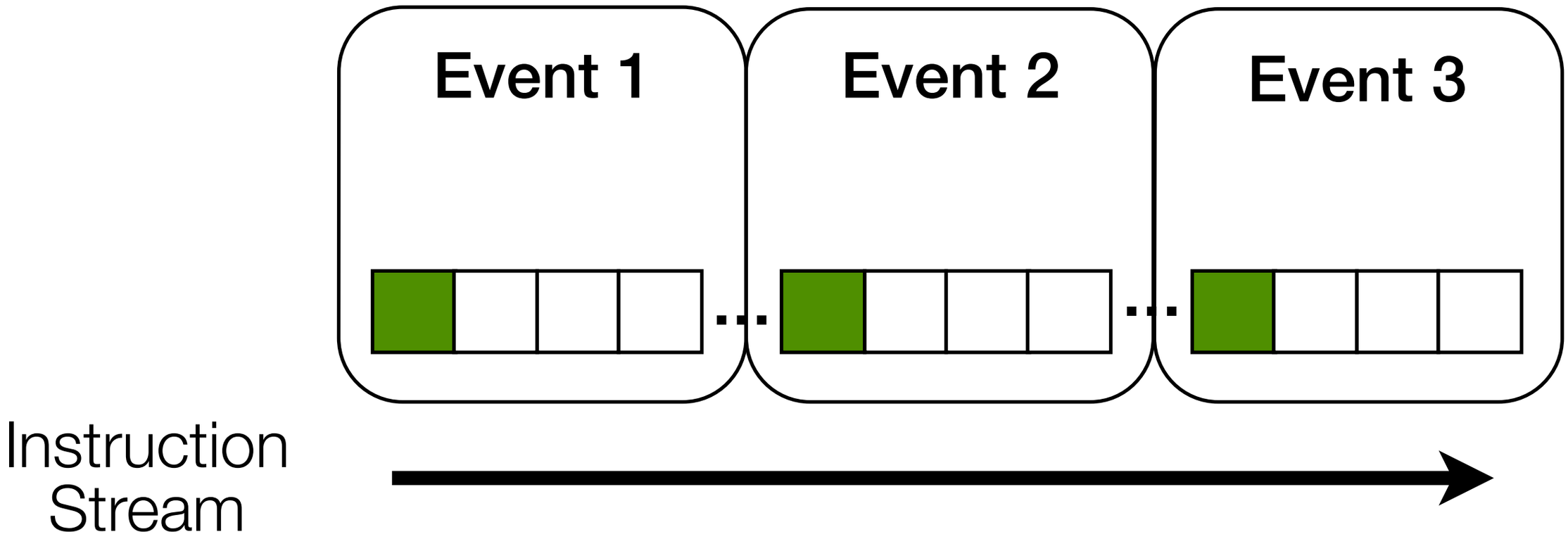


Scale Up Hardware Resources?



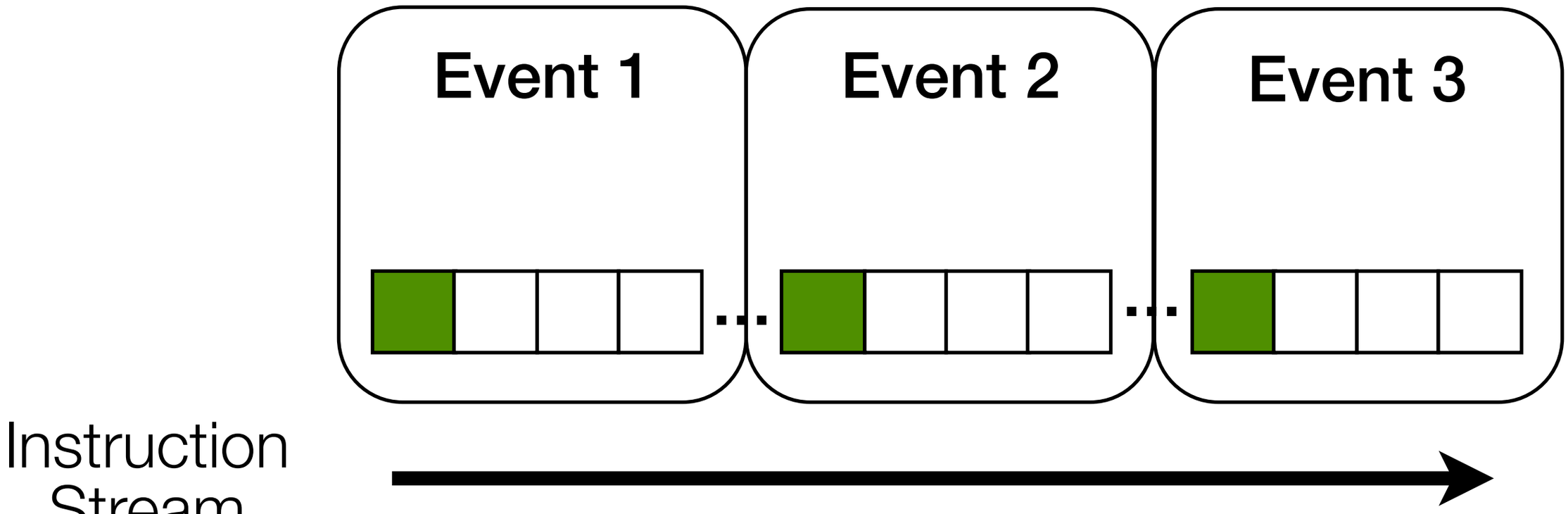
Exploit Inter-Event Locality

Exploit Inter-Event Locality



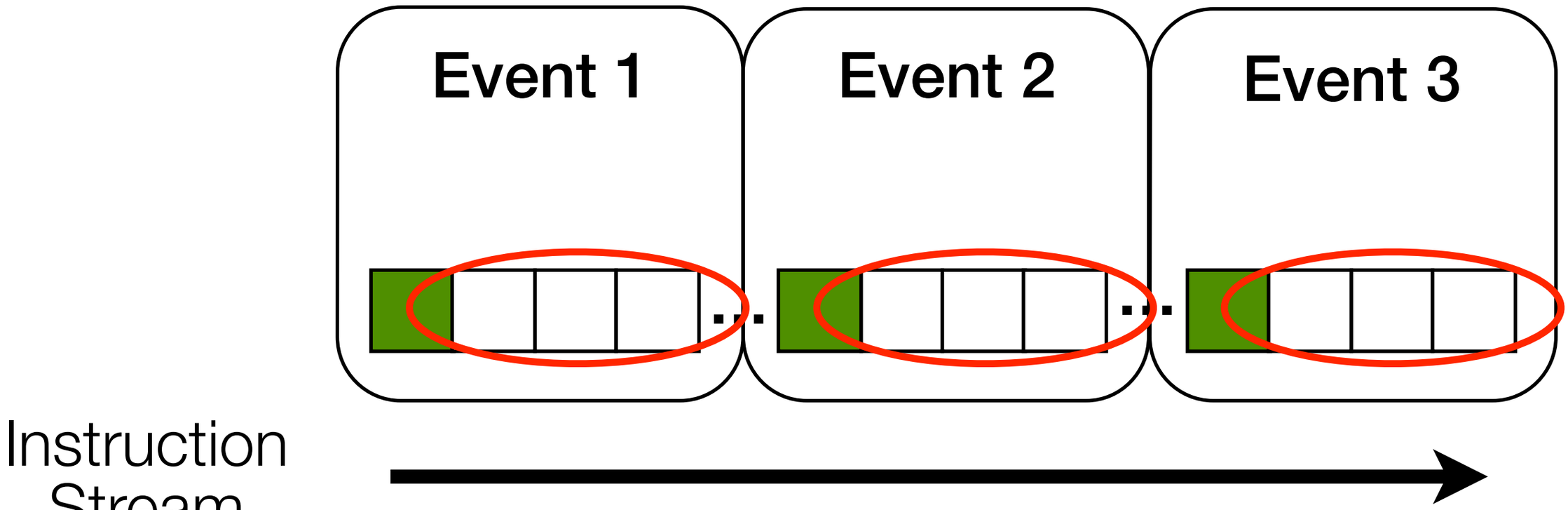
Exploit Inter-Event Locality

1. ***Retain*** *the reused portion of an event's footprint in the cache*



Exploit Inter-Event Locality

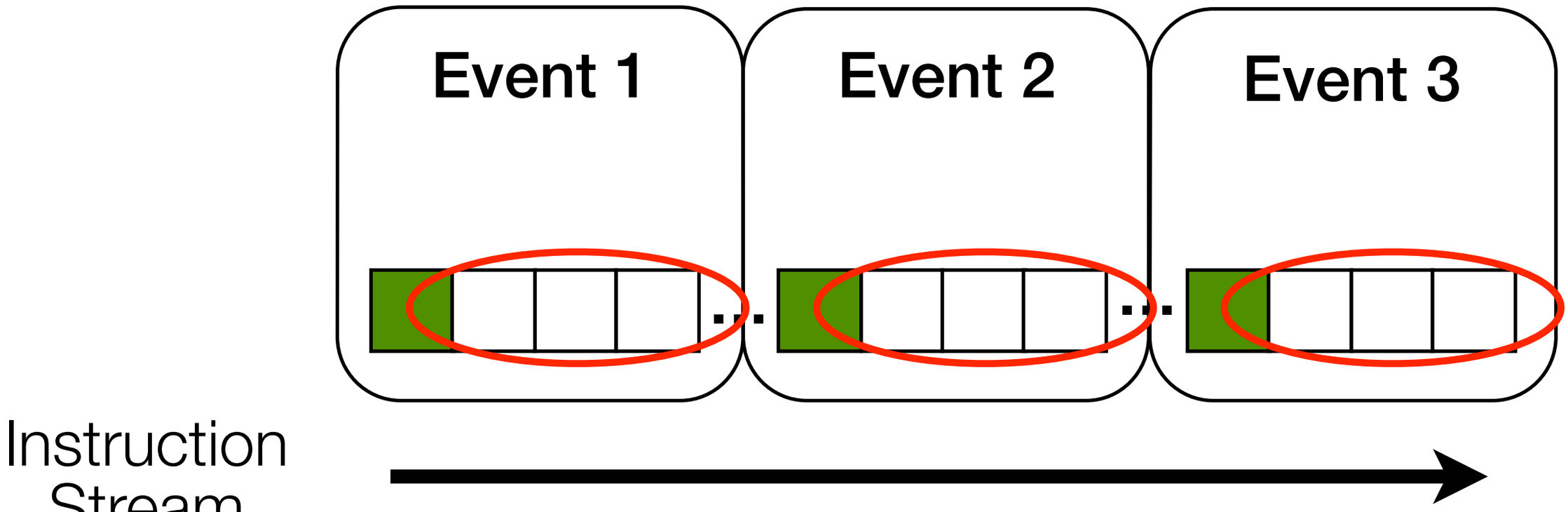
1. ***Retain*** *the reused portion of an event's footprint in the cache*



Exploit Inter-Event Locality

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*



Exploit Inter-Event Locality

PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

Exploit Inter-Event Locality

PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

Exploit Inter-Event Locality

PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

Exploit Inter-Event Locality

PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position

Exploit Inter-Event Locality

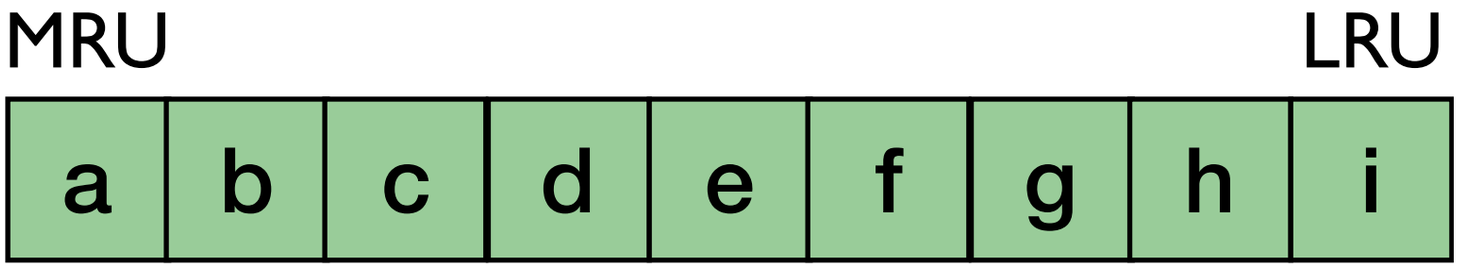
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

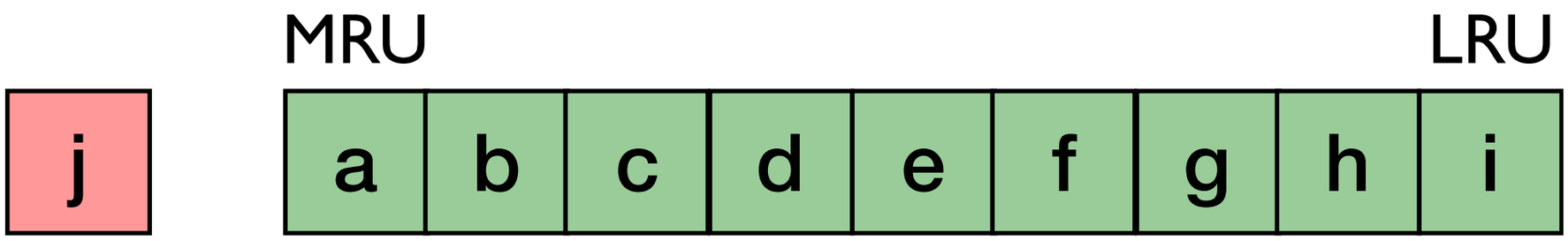
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

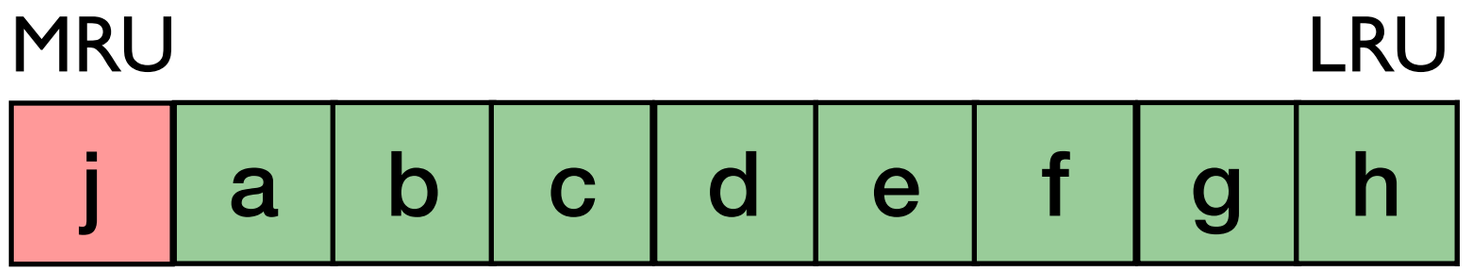
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

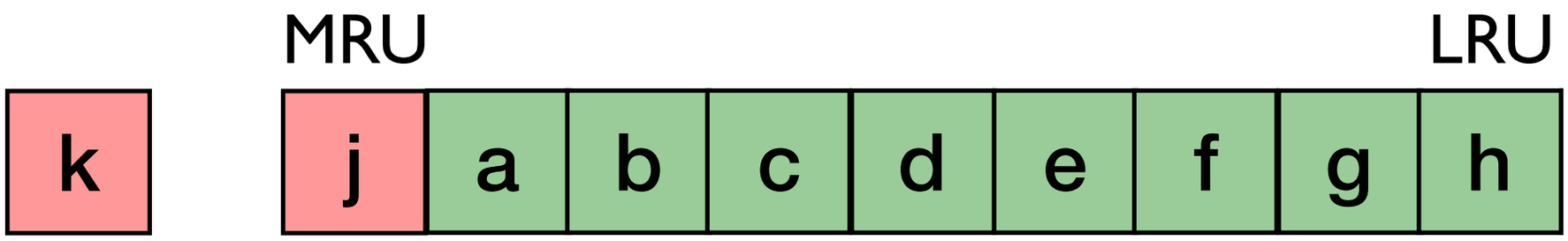
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

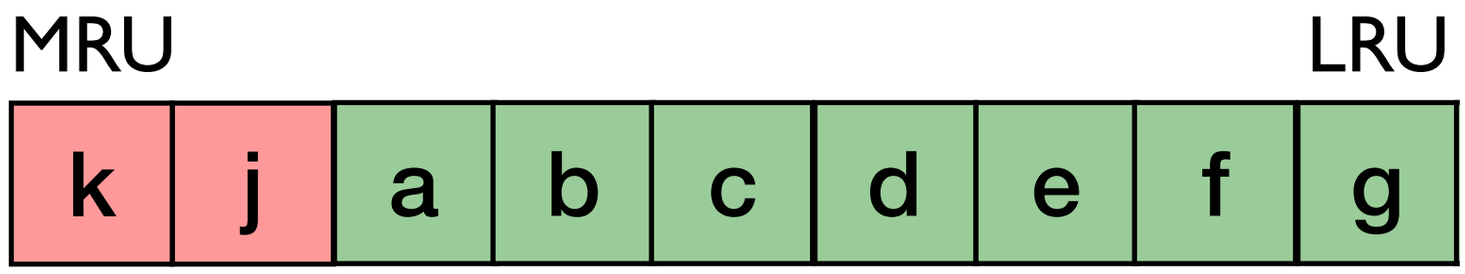
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

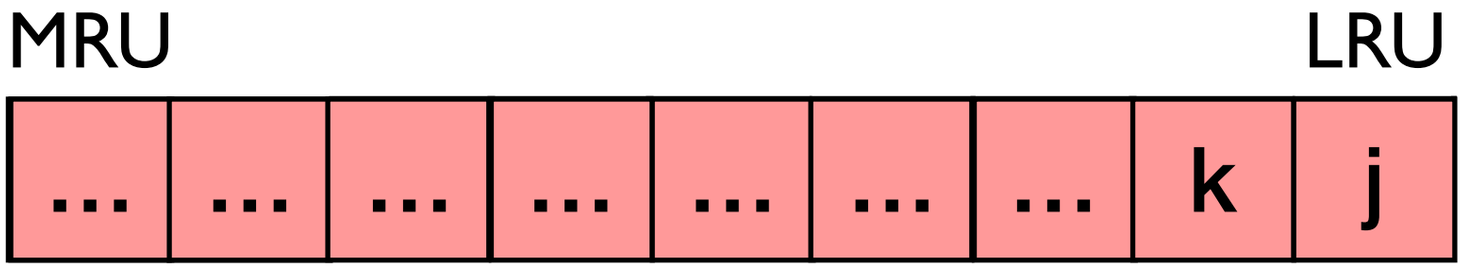
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

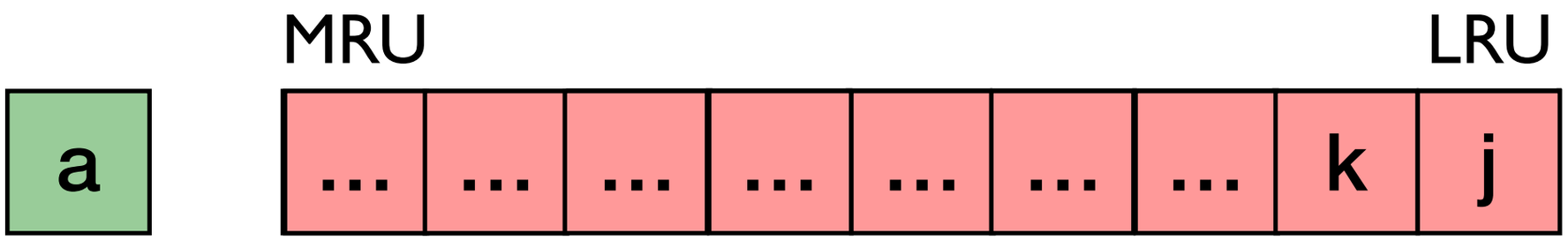
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

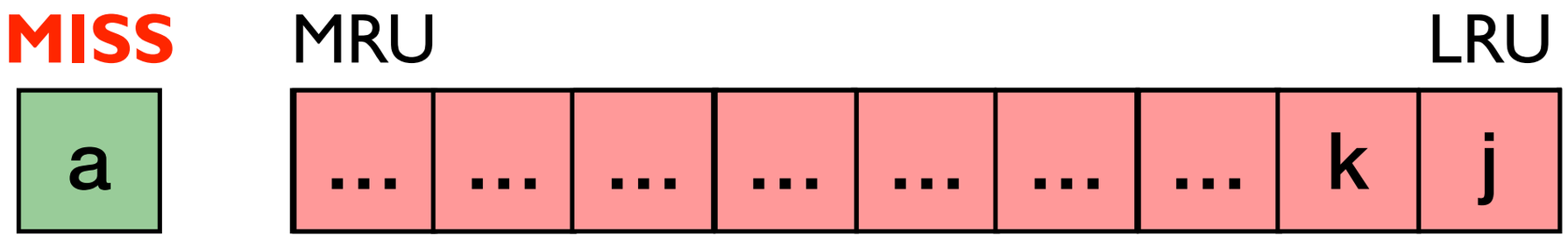
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

PRINCIPLES

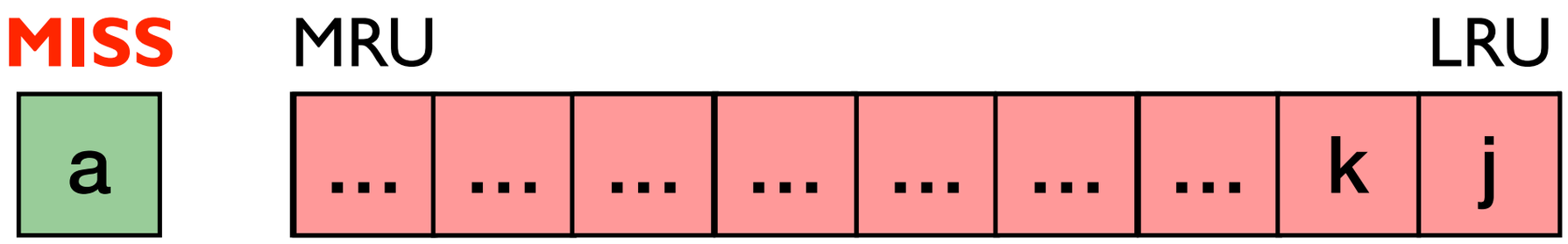
1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position

Inter-event Locality Lost!



Exploit Inter-Event Locality

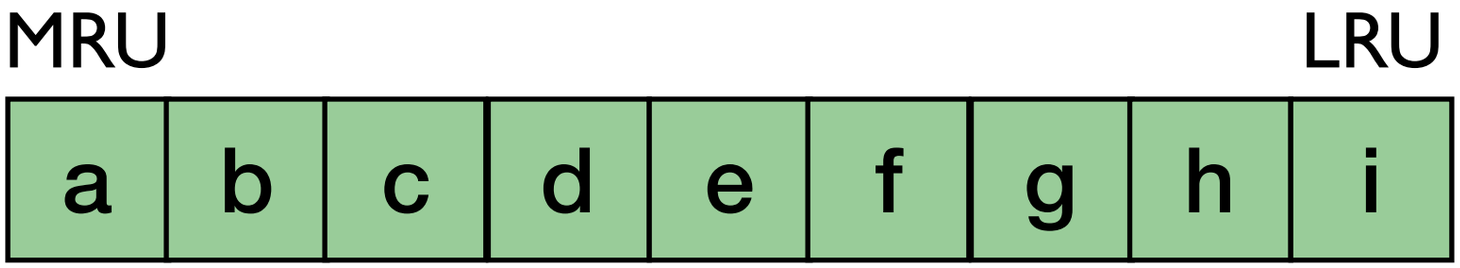
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

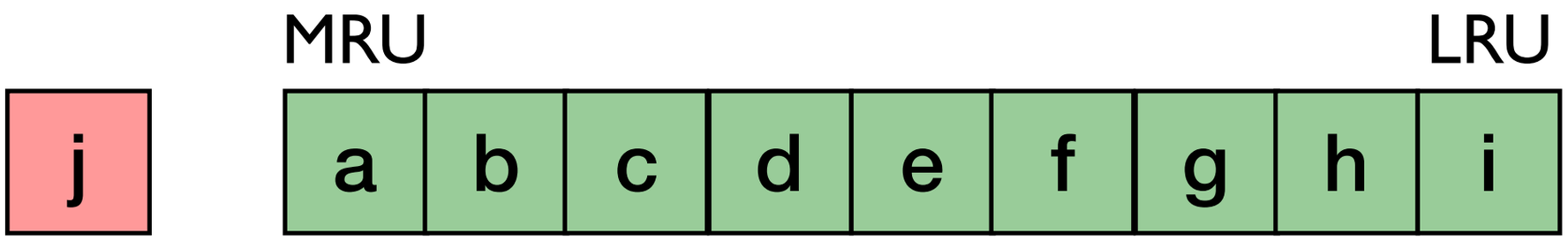
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

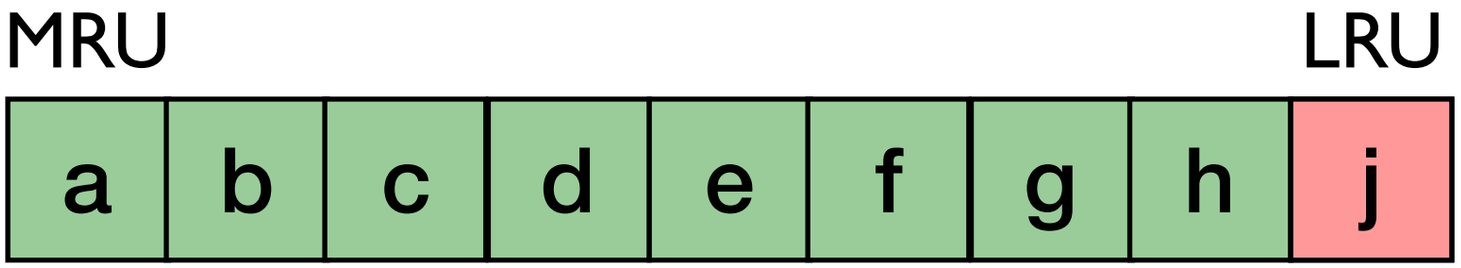
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

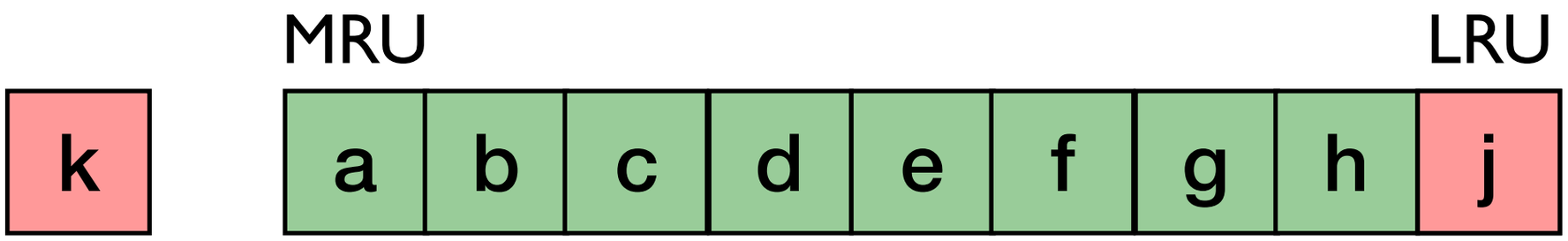
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

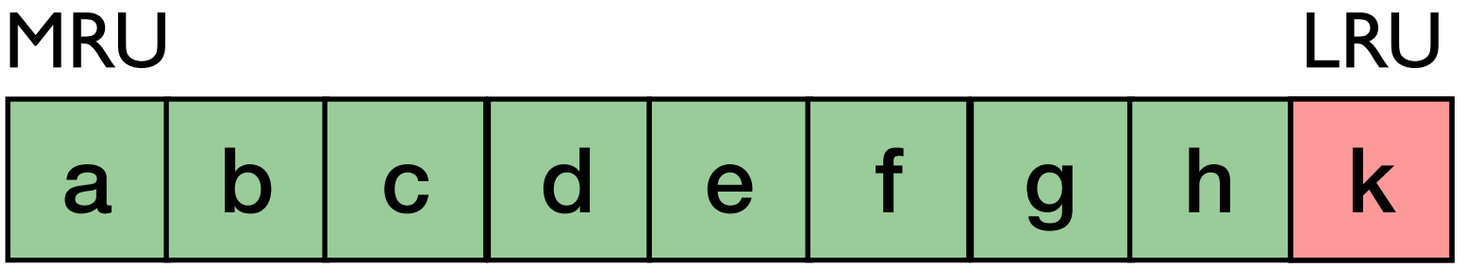
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

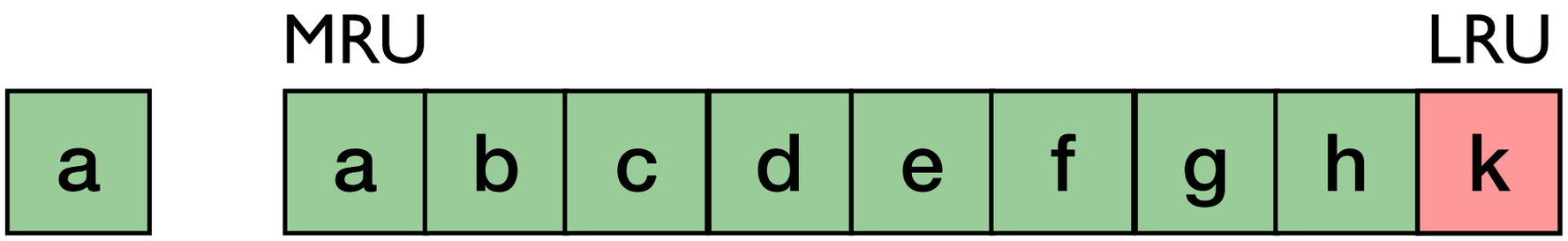
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

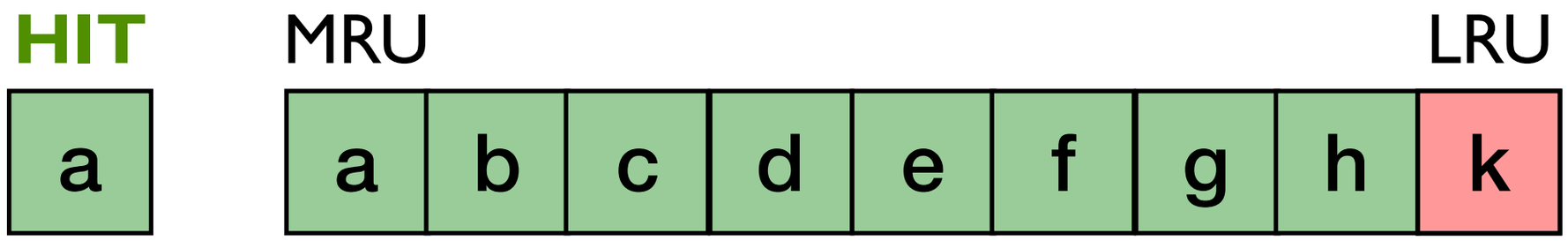
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position



Exploit Inter-Event Locality

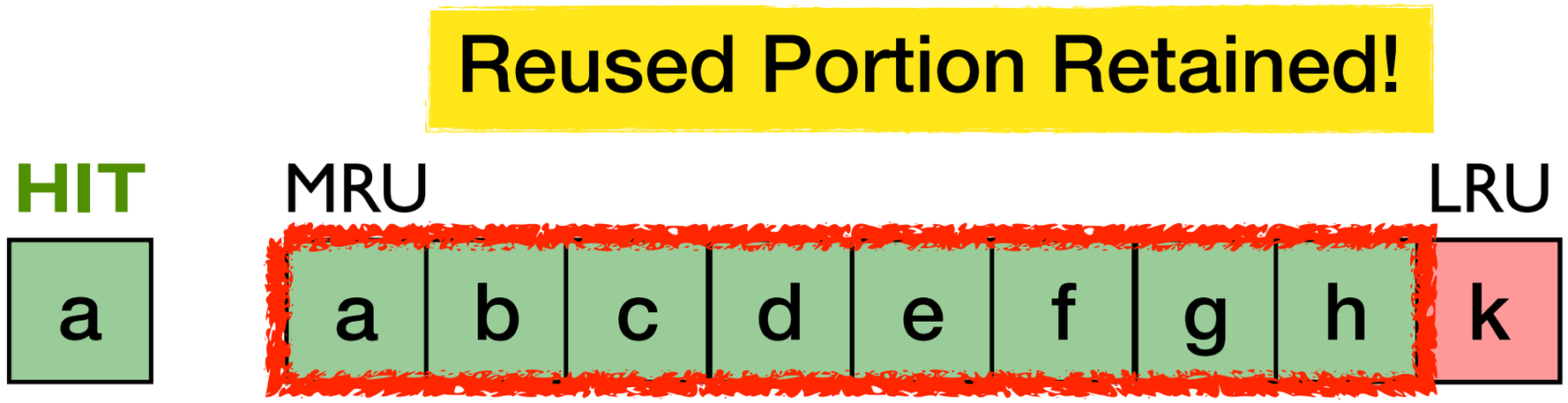
PRINCIPLES

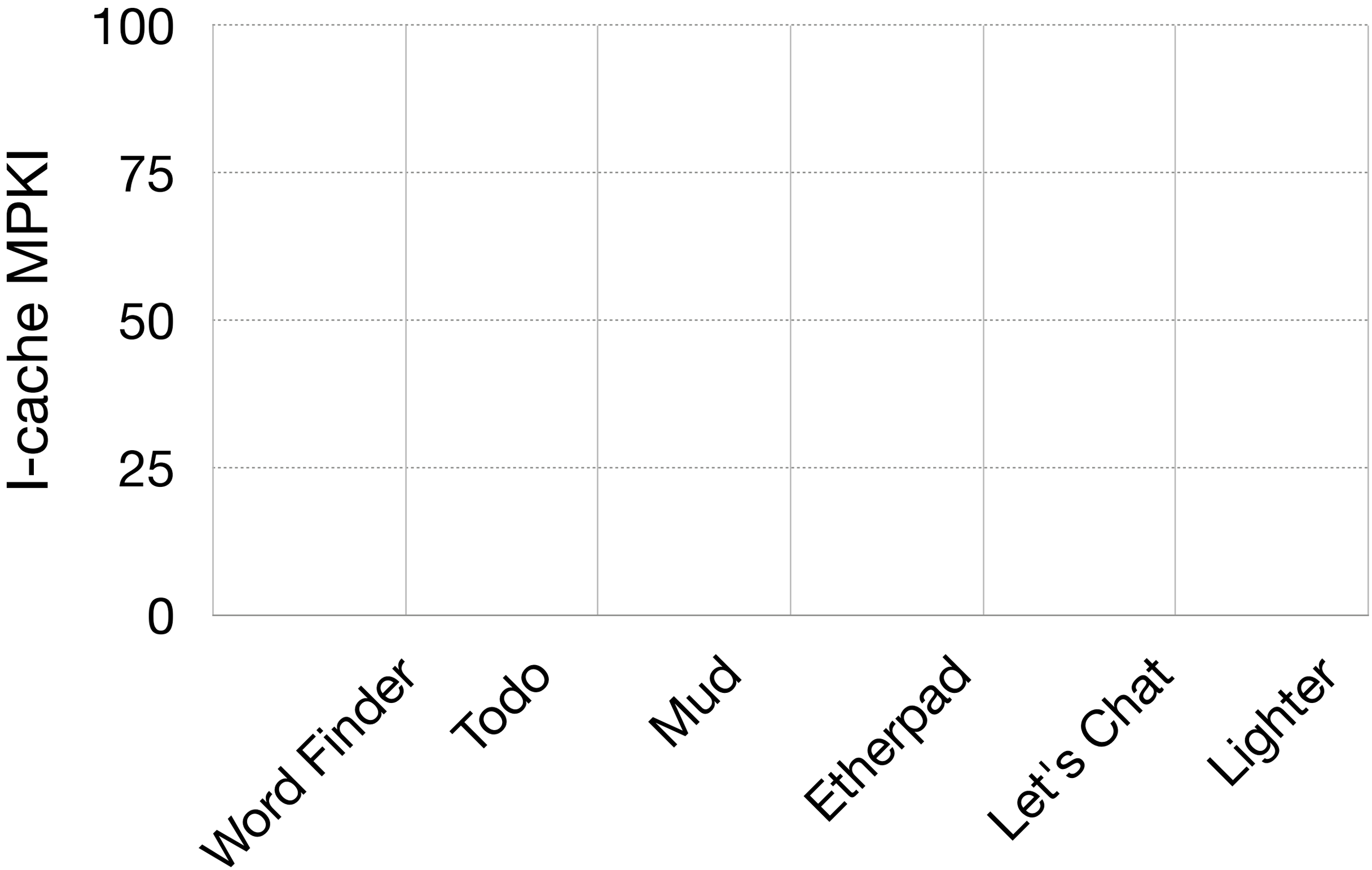
1. ***Retain*** *the reused portion of an event's footprint in the cache*

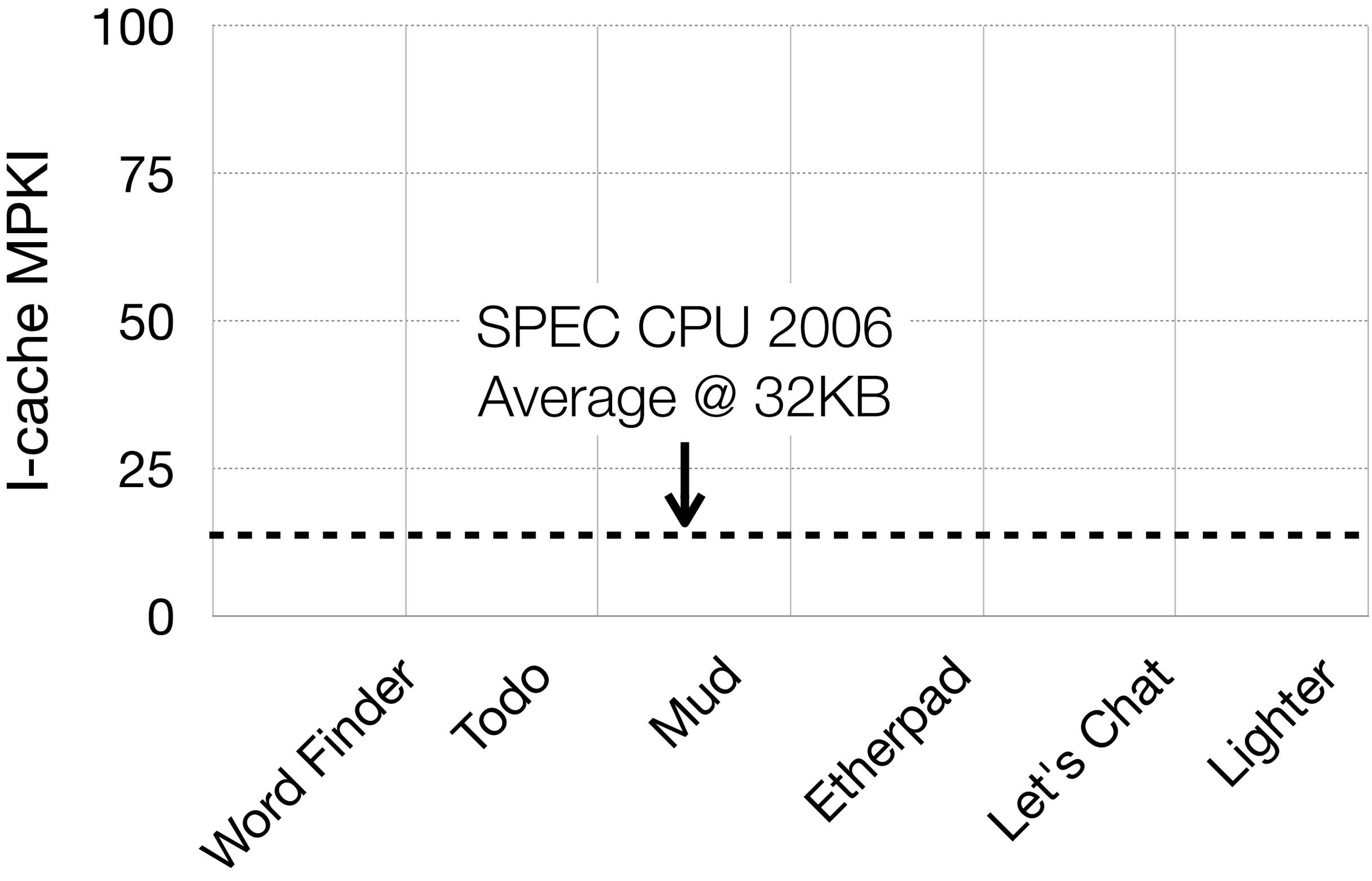
2. ***Prefetch*** *the unretained part*

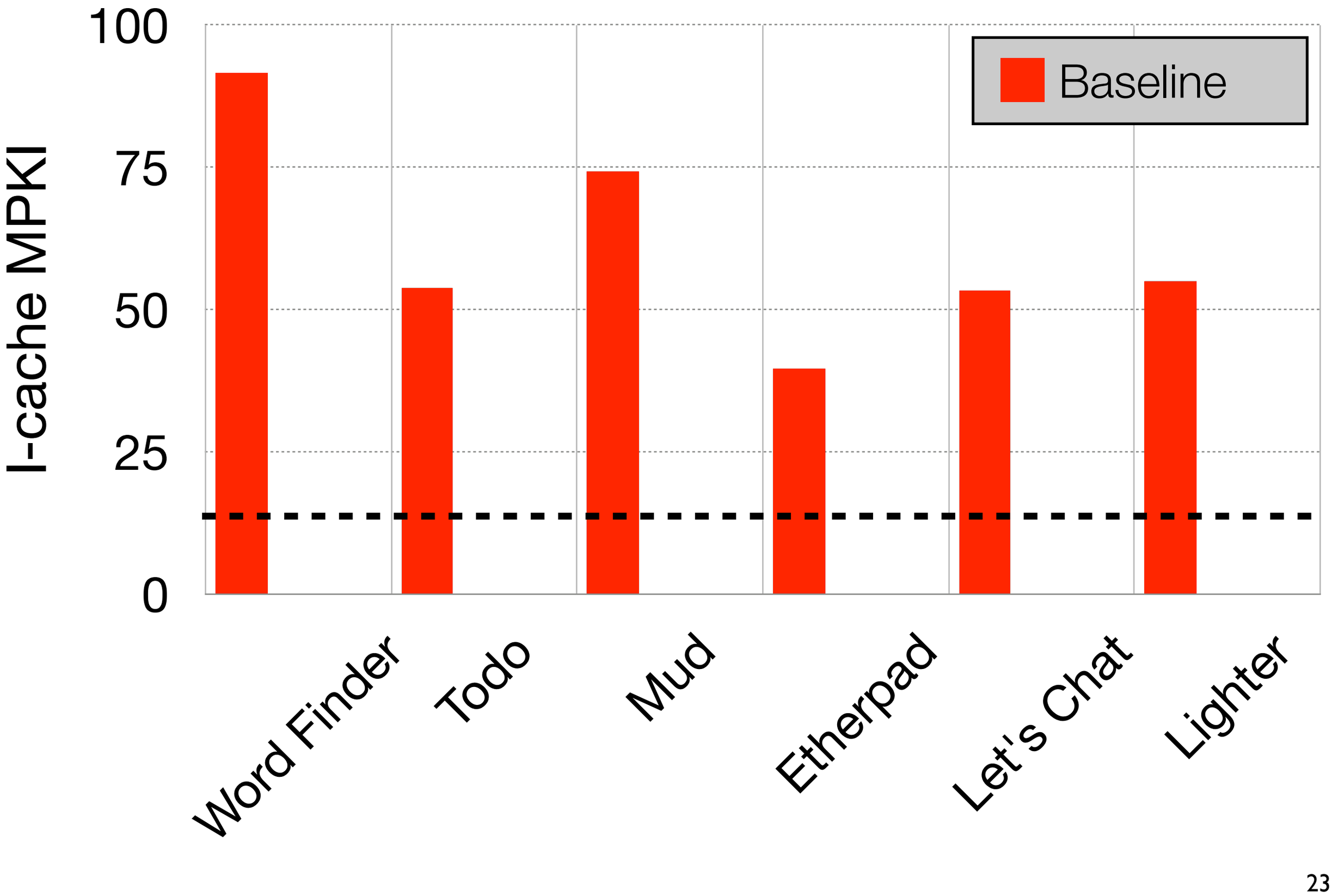
PRACTICES

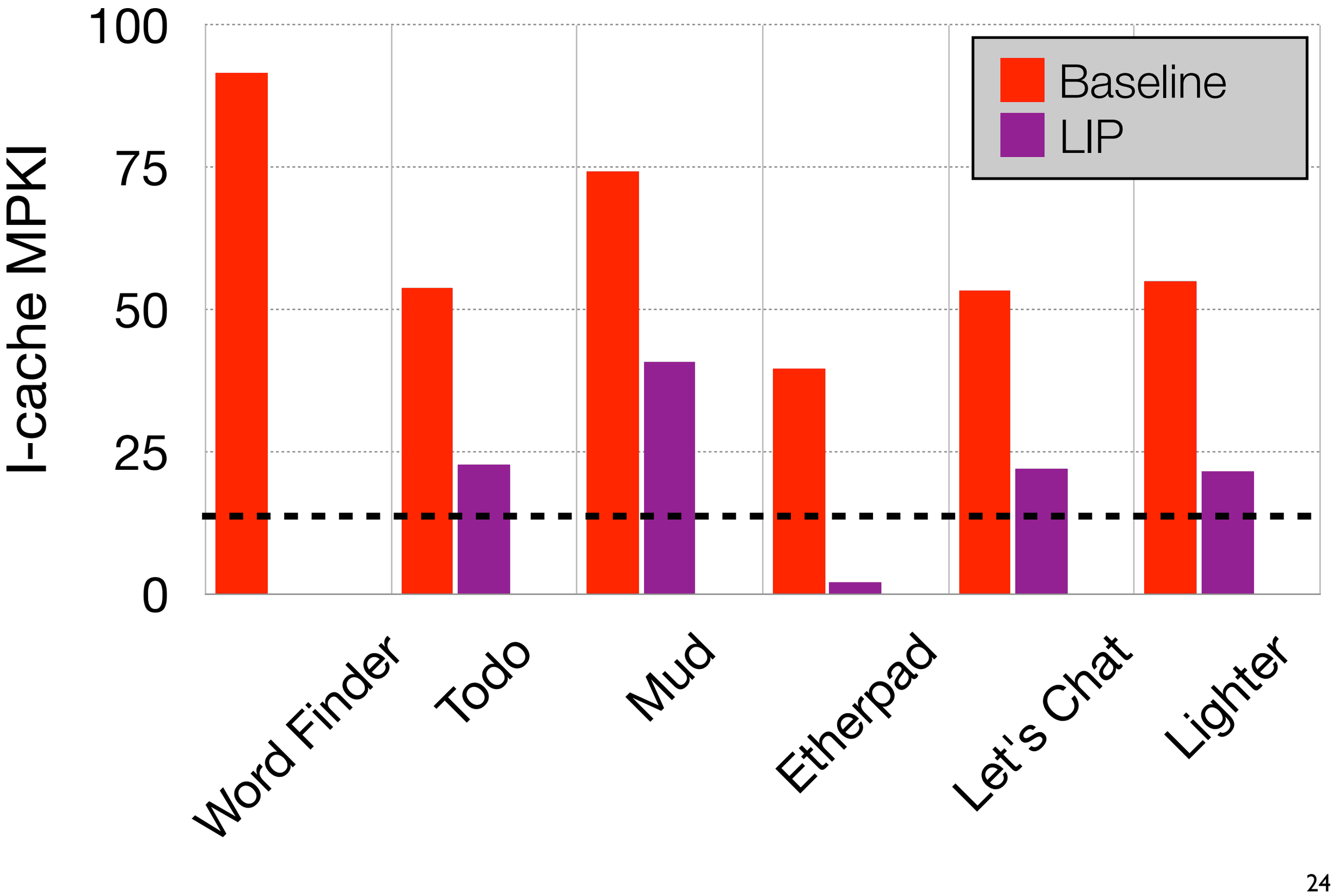
- LRU Cache Insertion Policy (**LIP**) (Qureshi et al., [ISCA'07])
  - Insert incoming line into LRU position, not MRU position




I-cache MPKI
100
75
50
25
0
Word Finder
Todo
Mud
Etherpad
Let's Chat
Lighter


I-cache MPKI
100
75
50
25
0
SPEC CPU 2006
Average @ 32KB
Word Finder
Todo
Mud
Etherpad
Let's Chat
Lighter


I-cache MPKI
100
75
50
25
0
Baseline
Word Finder
Todo
Mud
Etherpad
Let's Chat
Lighter


I-cache MPKI
100
75
50
25
0
Baseline
LIP
Word Finder
Todo
Mud
Etherpad
Let's Chat
Lighter

Exploit Inter-Event Locality

PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

Exploit Inter-Event Locality

PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- Temporal Instruction Fetch Streaming (**TIFS**) (Ferdman et al., [MICRO'08])
  - Find patterns in miss sequences

Exploit Inter-Event Locality

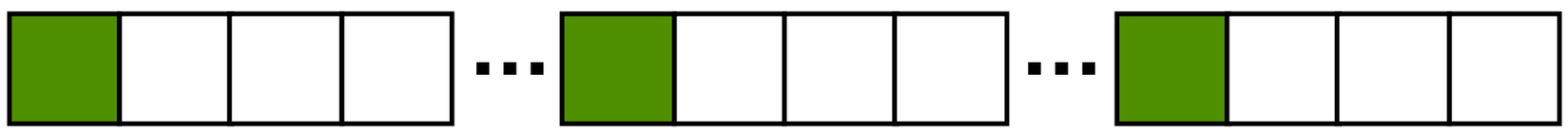
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- Temporal Instruction Fetch Streaming (**TIFS**) (Ferdman et al., [MICRO'08])
  - Find patterns in miss sequences



Exploit Inter-Event Locality

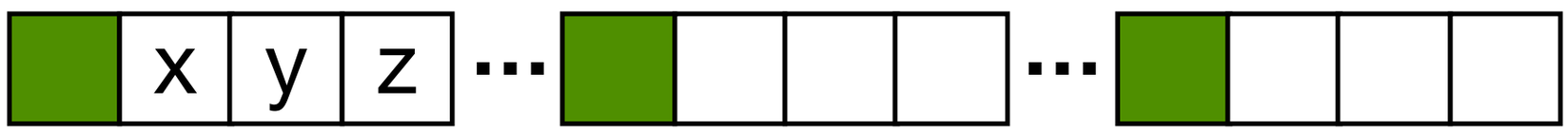
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- Temporal Instruction Fetch Streaming (**TIFS**) (Ferdman et al., [MICRO'08])
  - Find patterns in miss sequences



Exploit Inter-Event Locality

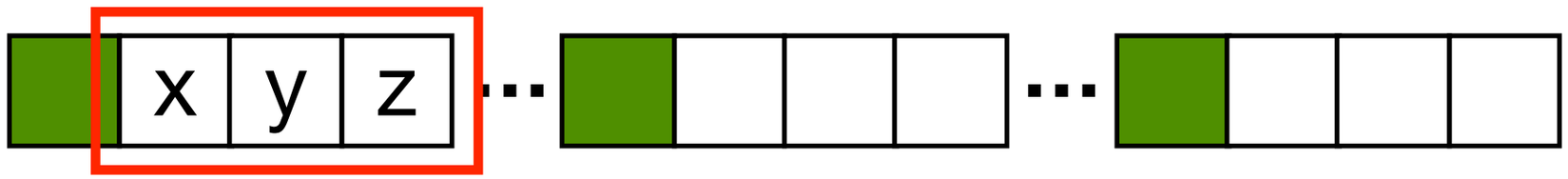
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- Temporal Instruction Fetch Streaming (**TIFS**) (Ferdman et al., [MICRO'08])
  - Find patterns in miss sequences



Exploit Inter-Event Locality

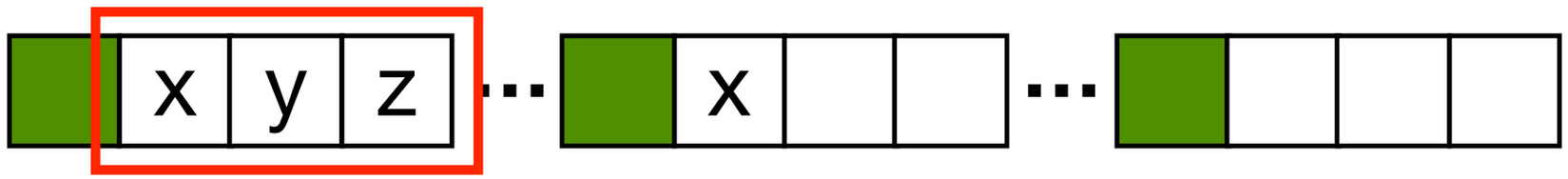
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- Temporal Instruction Fetch Streaming (**TIFS**) (Ferdman et al., [MICRO'08])
  - Find patterns in miss sequences



Exploit Inter-Event Locality

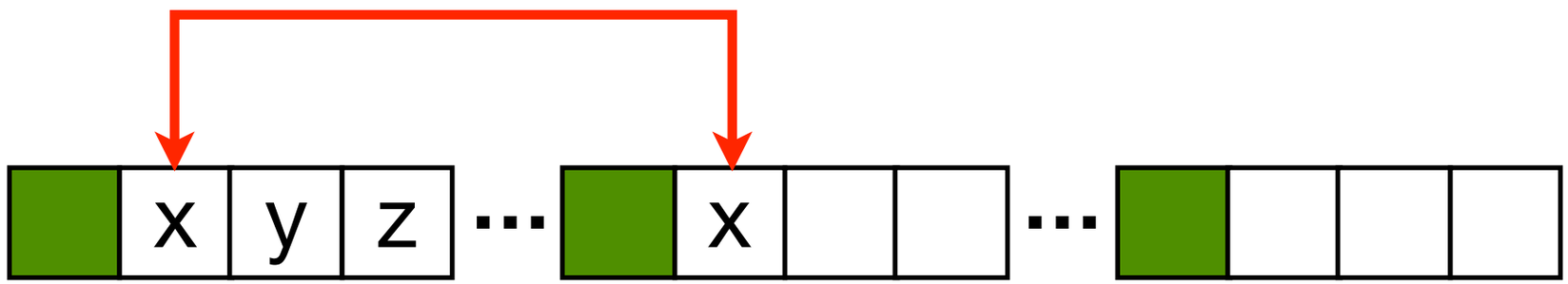
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- Temporal Instruction Fetch Streaming (**TIFS**) (Ferdman et al., [MICRO'08])
  - Find patterns in miss sequences



Exploit Inter-Event Locality

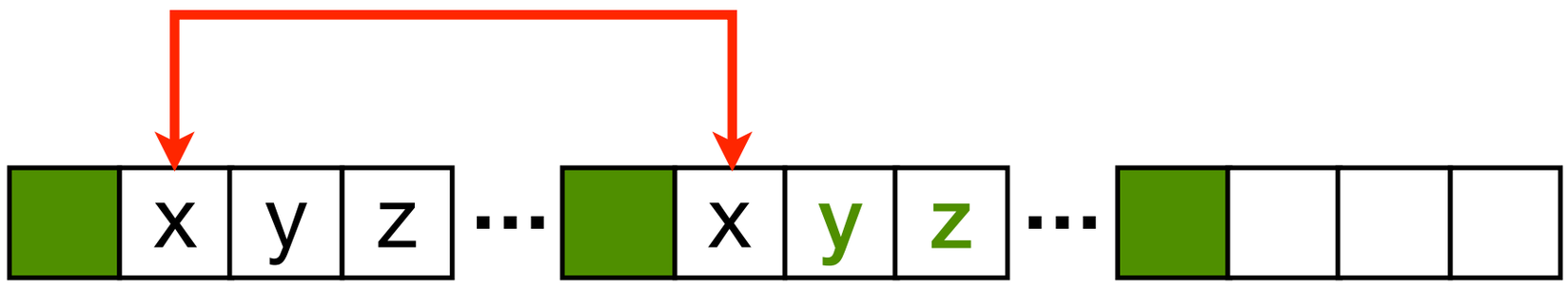
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- Temporal Instruction Fetch Streaming (**TIFS**) (Ferdman et al., [MICRO'08])
  - Find patterns in miss sequences



Exploit Inter-Event Locality

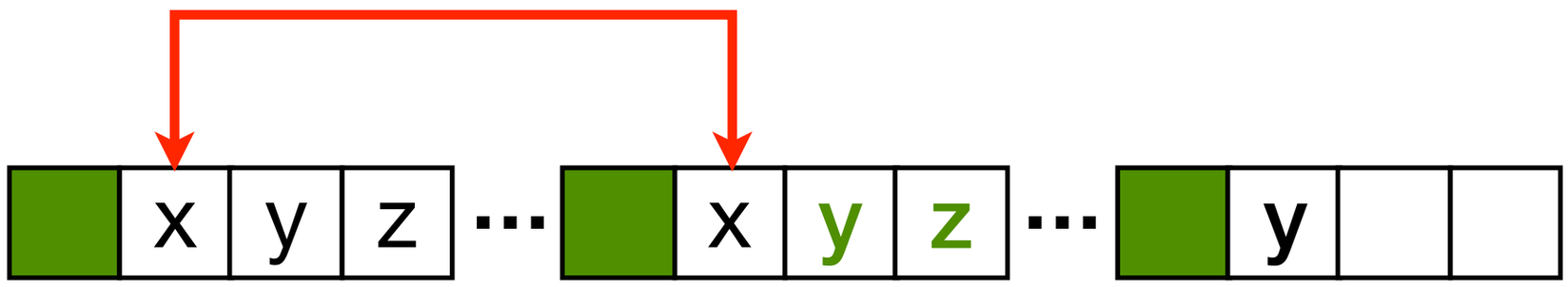
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- Temporal Instruction Fetch Streaming (**TIFS**) (Ferdman et al., [MICRO'08])
  - Find patterns in miss sequences



Exploit Inter-Event Locality

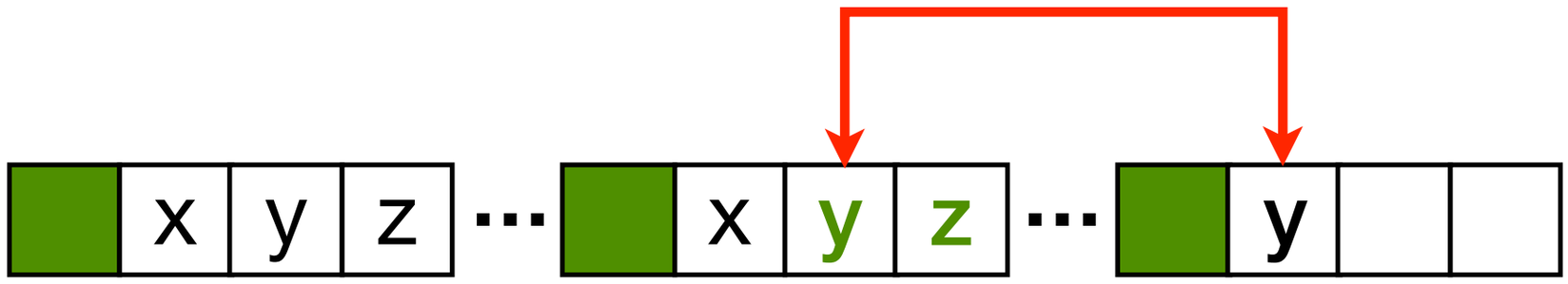
PRINCIPLES

1. ***Retain*** *the reused portion of an event's footprint in the cache*

2. ***Prefetch*** *the unretained part*

PRACTICES

- Temporal Instruction Fetch Streaming (**TIFS**) (Ferdman et al., [MICRO'08])
  - Find patterns in miss sequences



Exploit Inter-Event Locality

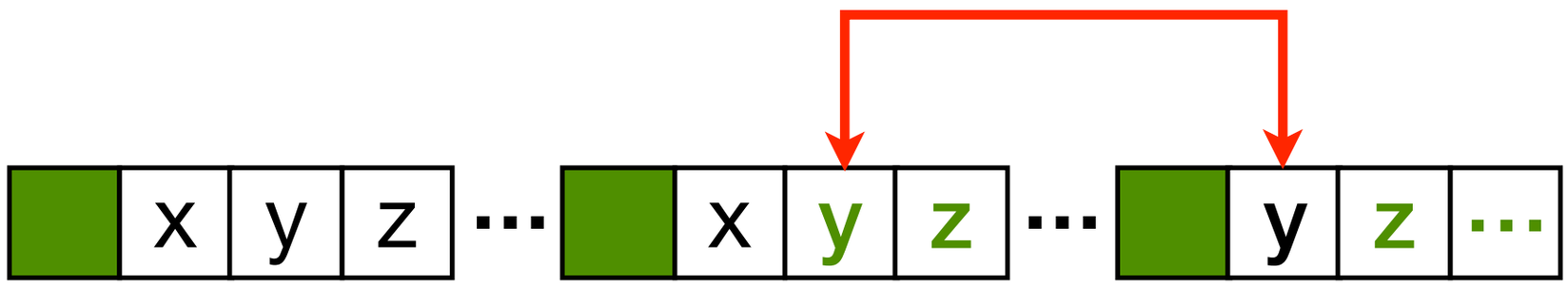
PRINCIPLES

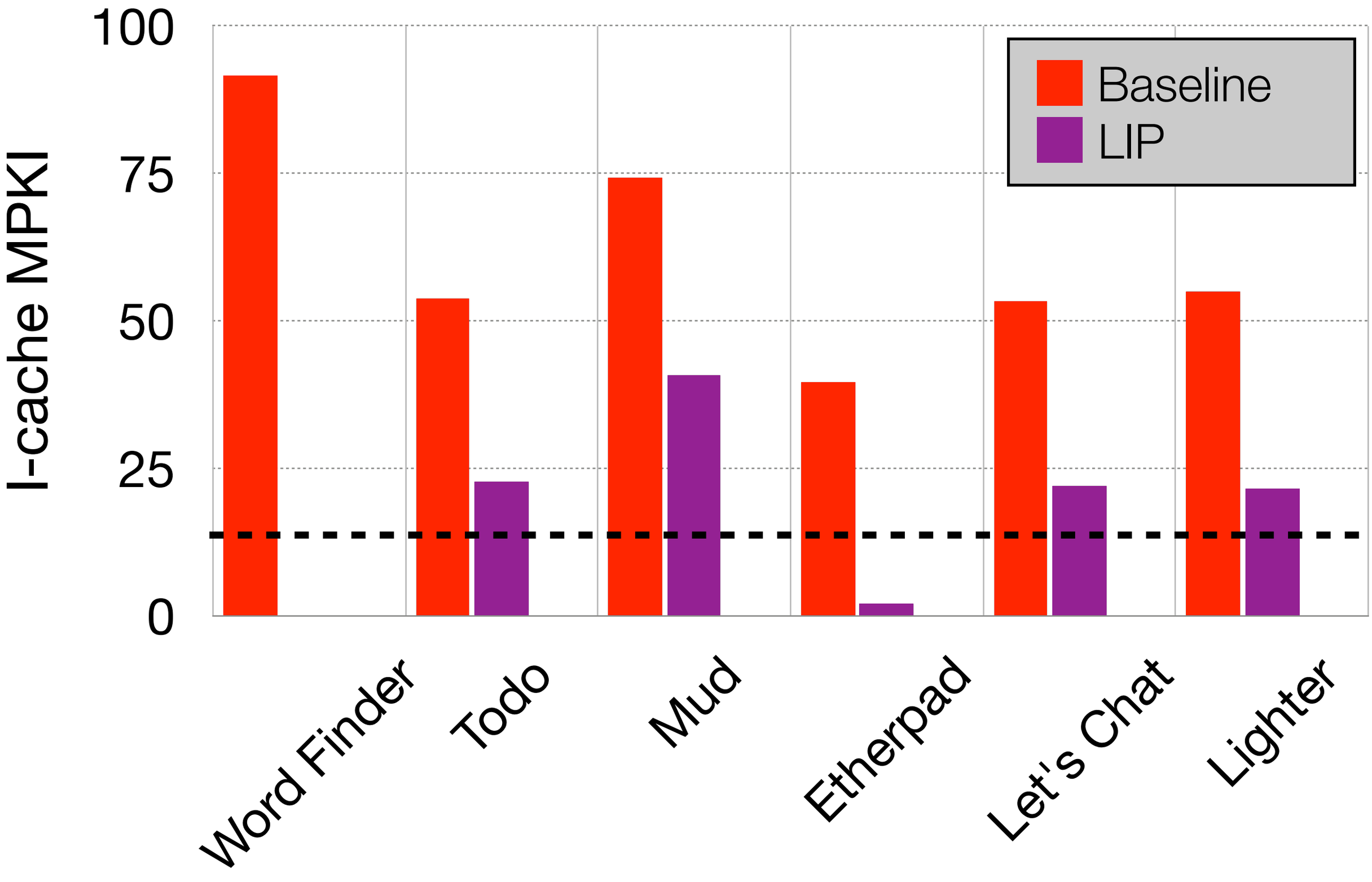
1. ***Retain*** *the reused portion of an event's footprint in the cache*

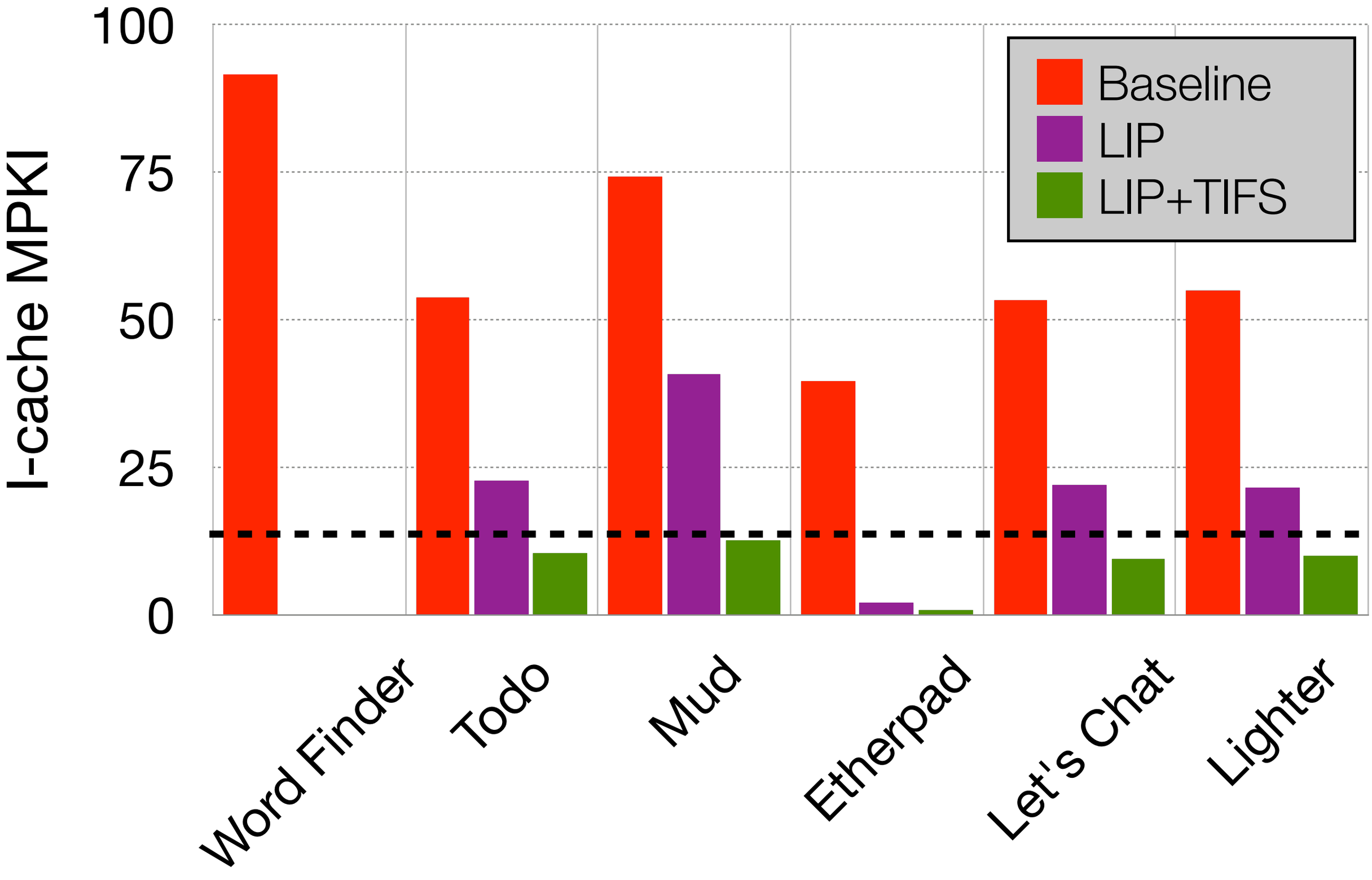
2. ***Prefetch*** *the unretained part*

PRACTICES

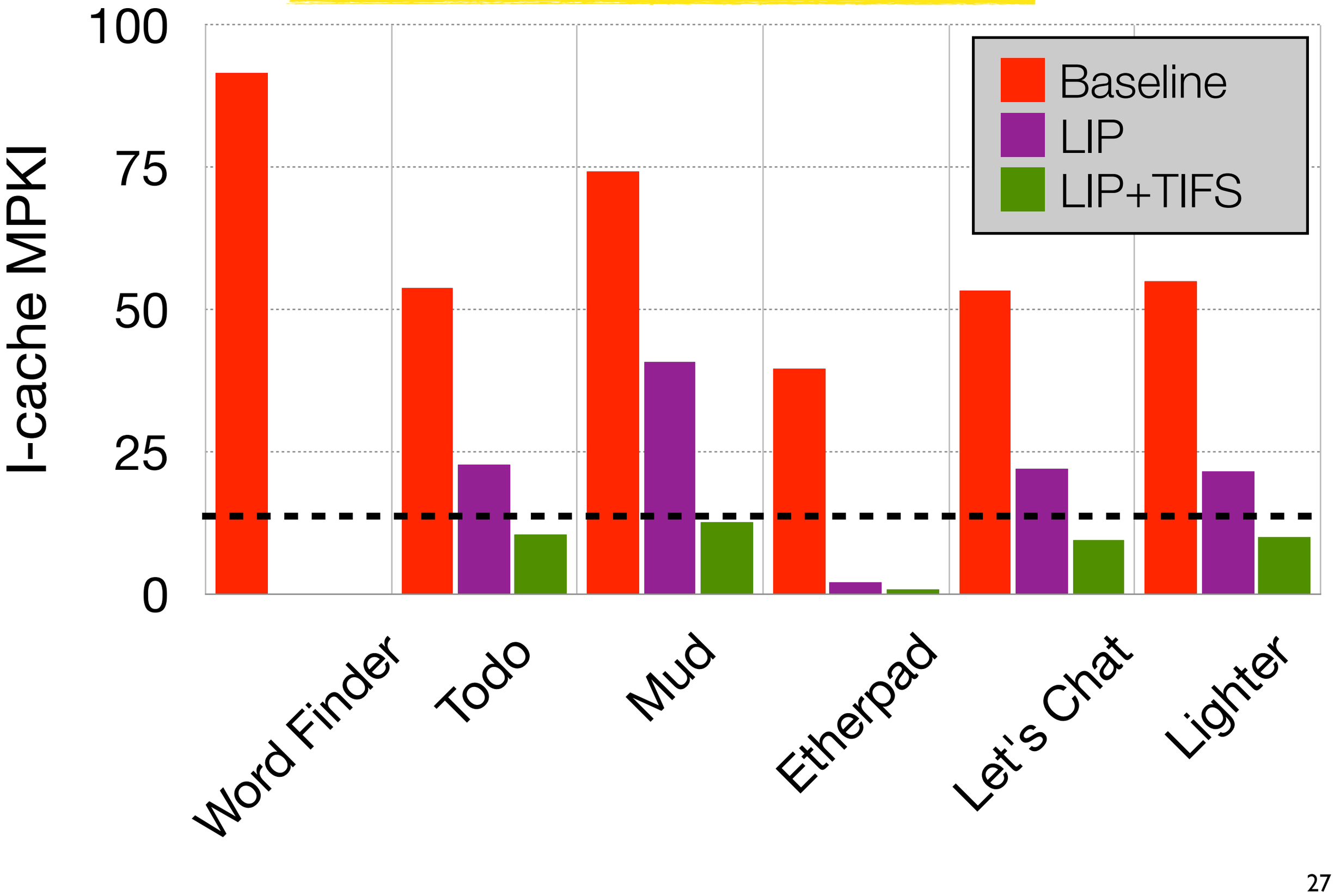
- Temporal Instruction Fetch Streaming (**TIFS**) (Ferdman et al., [MICRO'08])
  - Find patterns in miss sequences




I-cache MPKI
100
75
50
25
0
Baseline
LIP
Word Finder
Todo
Mud
Etherpad
Let's Chat
Lighter


I-cache MPKI
100
75
50
25
0
Baseline
LIP
LIP+TIFS
Word Finder
Todo
Mud
Etherpad
Let's Chat
Lighter

88% Average MPKI Reduction



Exploit Instruction Locality

a.k.a., Common Case Design

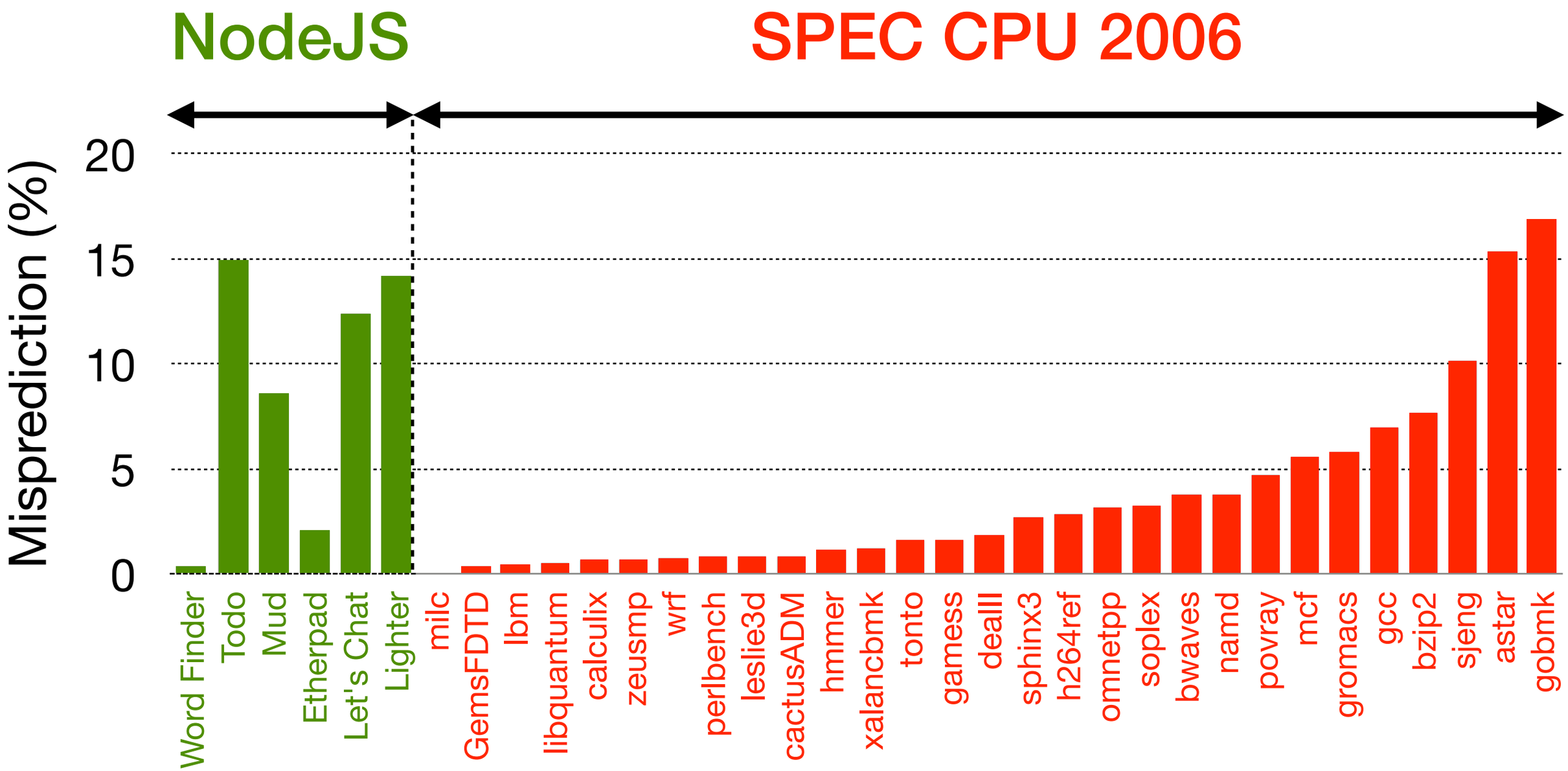
		SPEC CPU (mostly)	Event-driven Applications
	Cache	Hot instructions	✓
Instruction Supply	Branch Predictor	Hot branch history patterns	✓
	TLB	Hot pages	✓

Exploit Instruction Locality

a.k.a., Common Case Design

		SPEC CPU (mostly)	Event-driven Applications
Instruction Supply	Cache	Hot instructions	✓
	Branch Predictor	Hot branch history patterns	✓
	TLB	Hot pages	✓

Beyond Instruction Cache — Branch Predictor

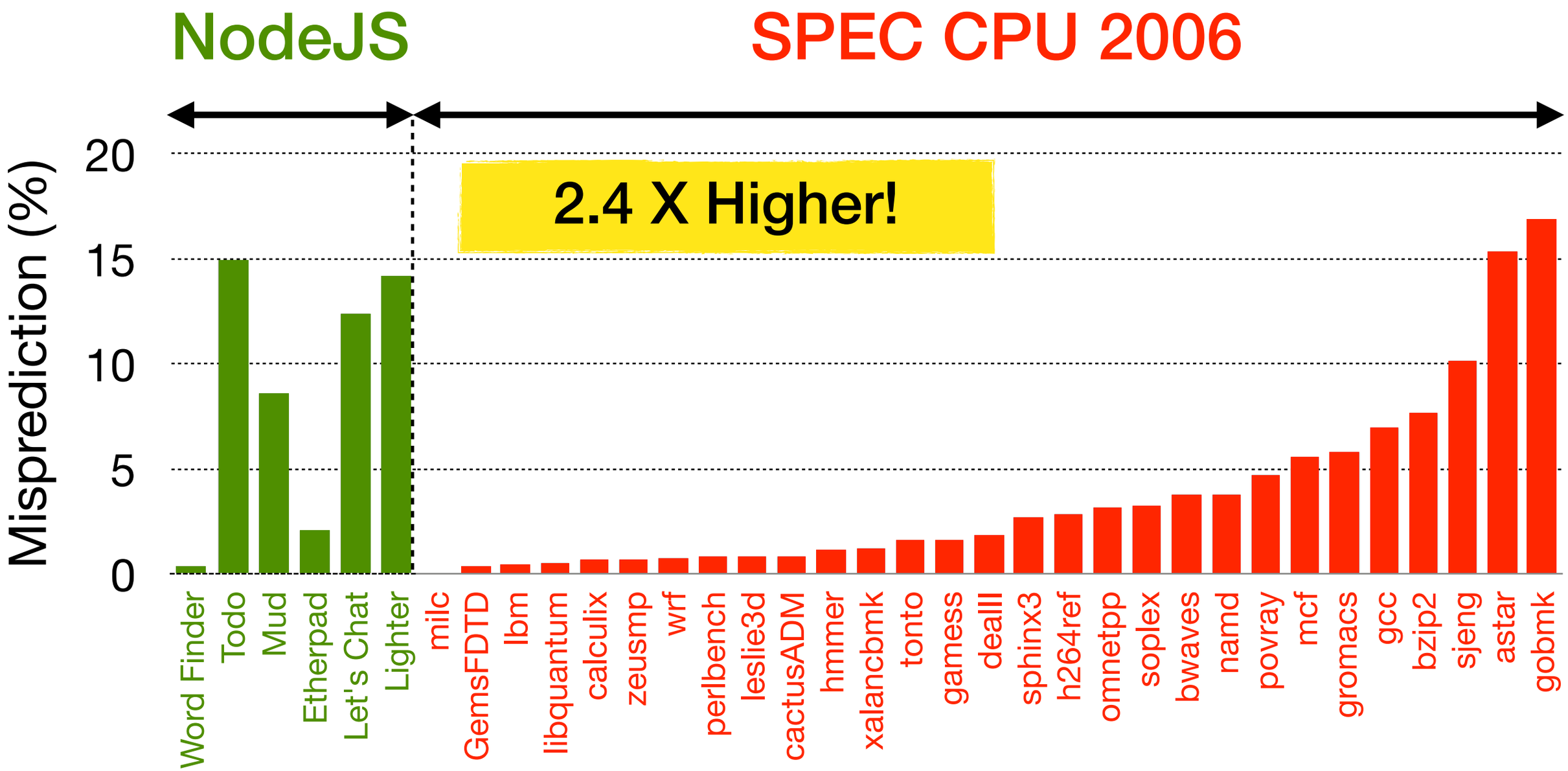


Tournament Predictor

12-bit history register

256 local branch histories

Beyond Instruction Cache — Branch Predictor

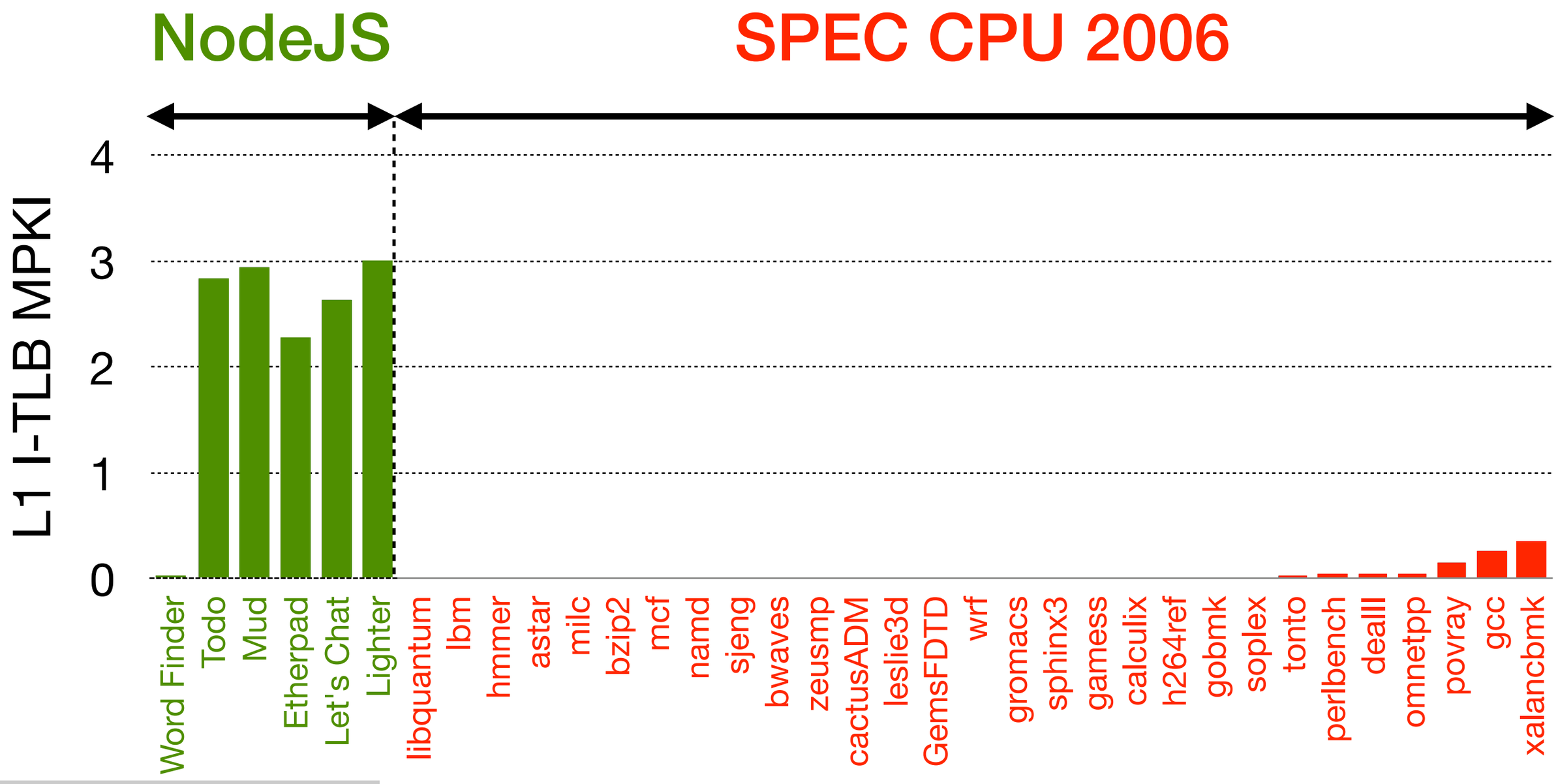


Tournament Predictor

12-bit history register

256 local branch histories

Beyond Instruction Cache — TLB

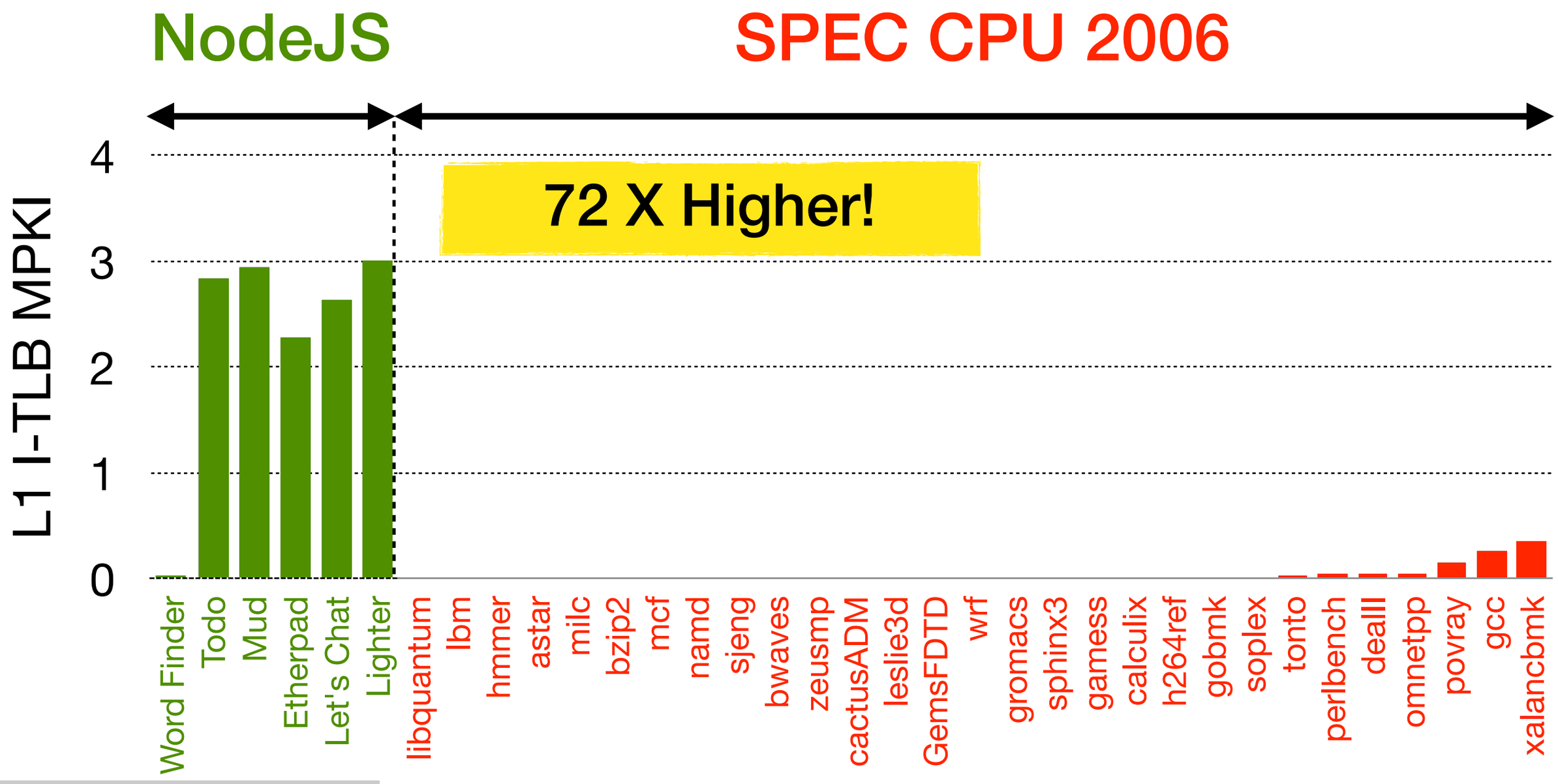


I-TLB Parameters

64 KB, 4-way

4 KB page size

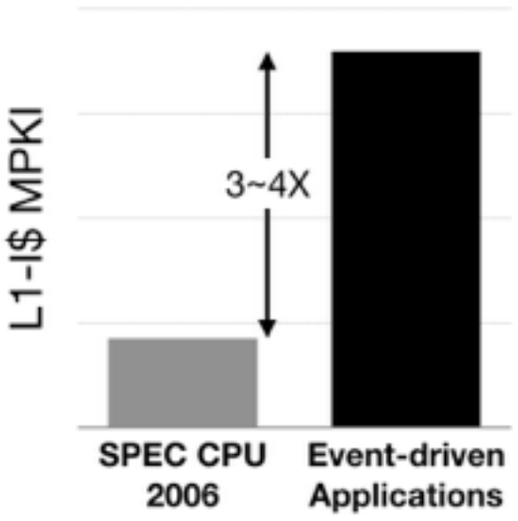
Beyond Instruction Cache — TLB

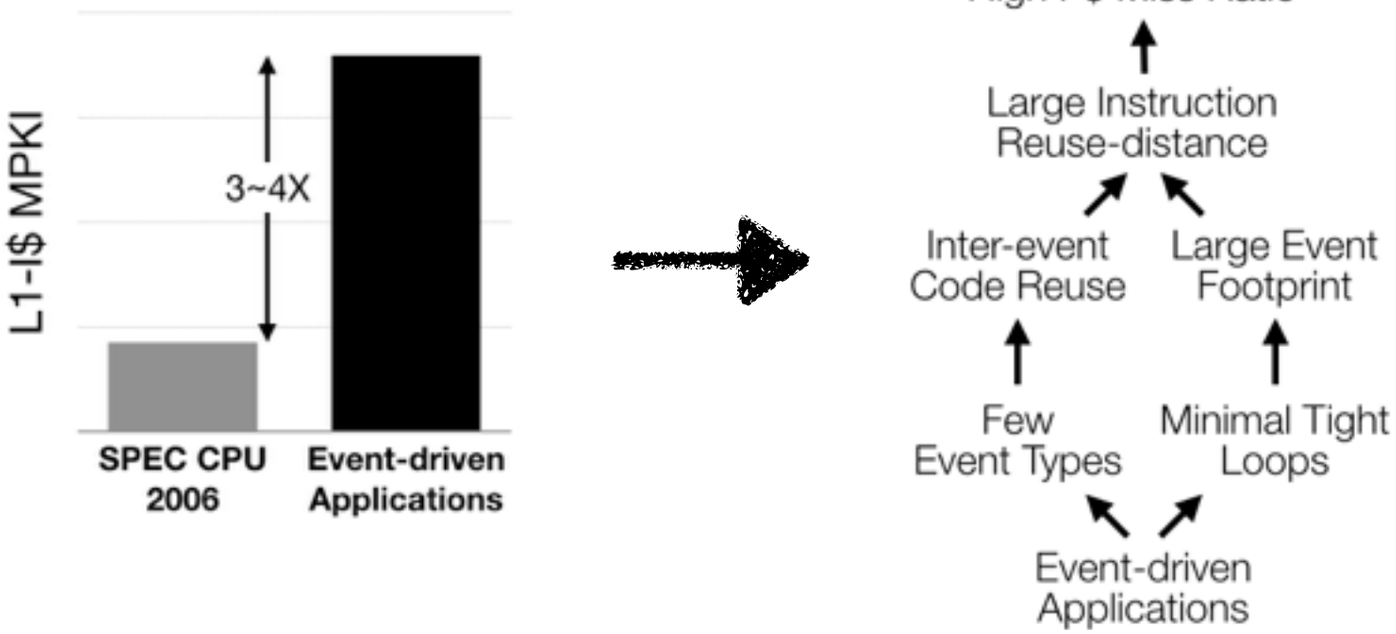


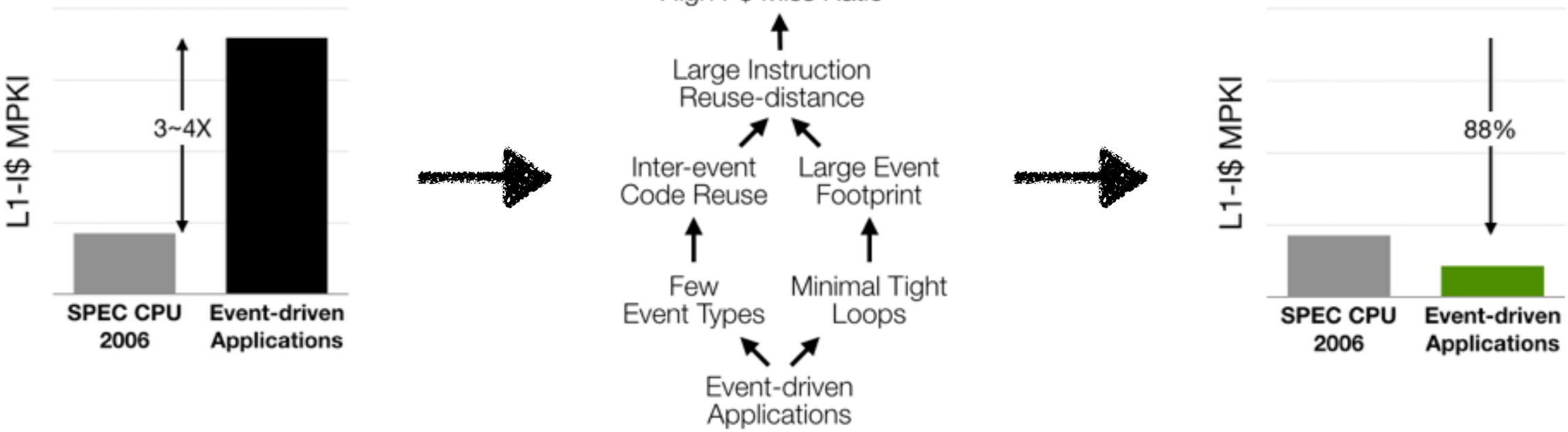
I-TLB Parameters

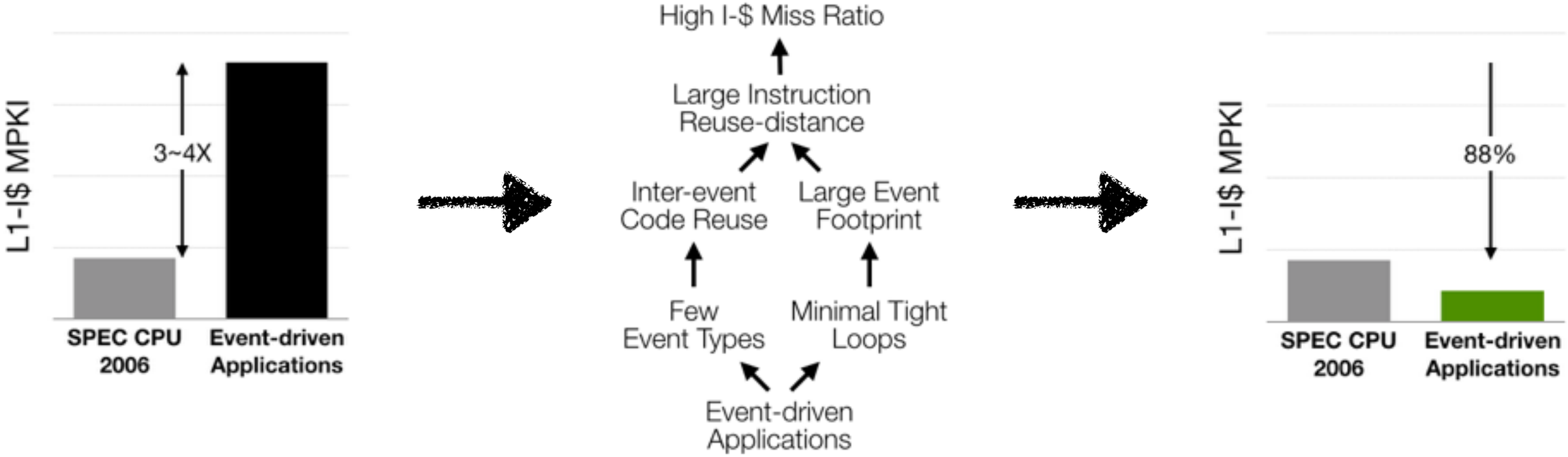
64 KB, 4-way

4 KB page size

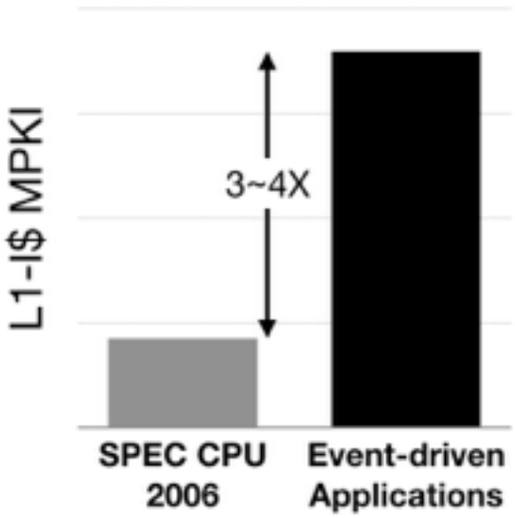

L1-I$ MPKI
3~4X
SPEC CPU 2006
Event-driven Applications

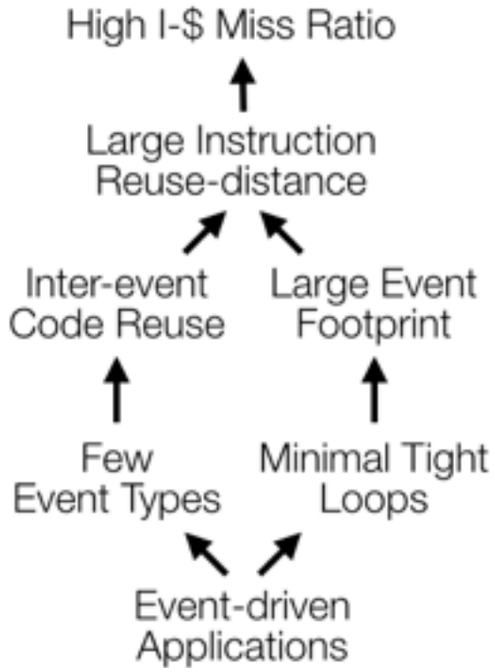

L1-I$ MPKI
3~4X
SPEC CPU 2006
Event-driven Applications
High I-$ Miss Ratio
Large Instruction Reuse-distance
Inter-event Code Reuse
Large Event Footprint
Few Event Types
Minimal Tight Loops
Event-driven Applications

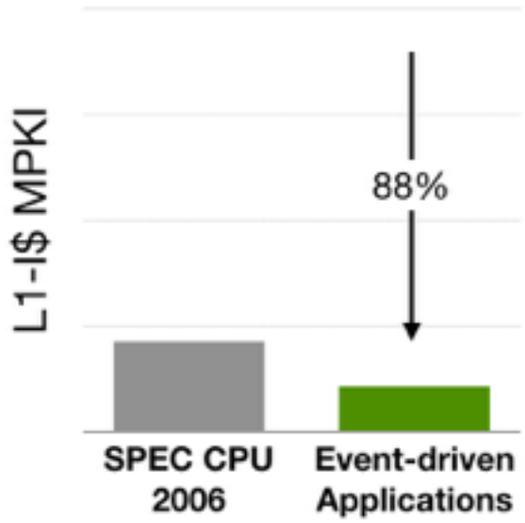

L1-I$ MPKI
3~4X
SPEC CPU 2006
Event-driven Applications
High I-$ Miss Ratio
Large Instruction Reuse-distance
Inter-event Code Reuse
Large Event Footprint
Few Event Types
Minimal Tight Loops
Event-driven Applications
L1-I$ MPKI
88%
SPEC CPU 2006
Event-driven Applications



Event-based processing is a fundamental computation pattern.







Event-based processing is a fundamental computation pattern.

Web

Mobile

Sensor networks

Cloud

Internet-of-Things

Microarchitectural Implications of **Event-driven** Server-side Web Applications

Yuhao Zhu
UT Austin
with Daniel Richins, Matthew Halpern, Vijay Janapa Reddi

MICRO 2015