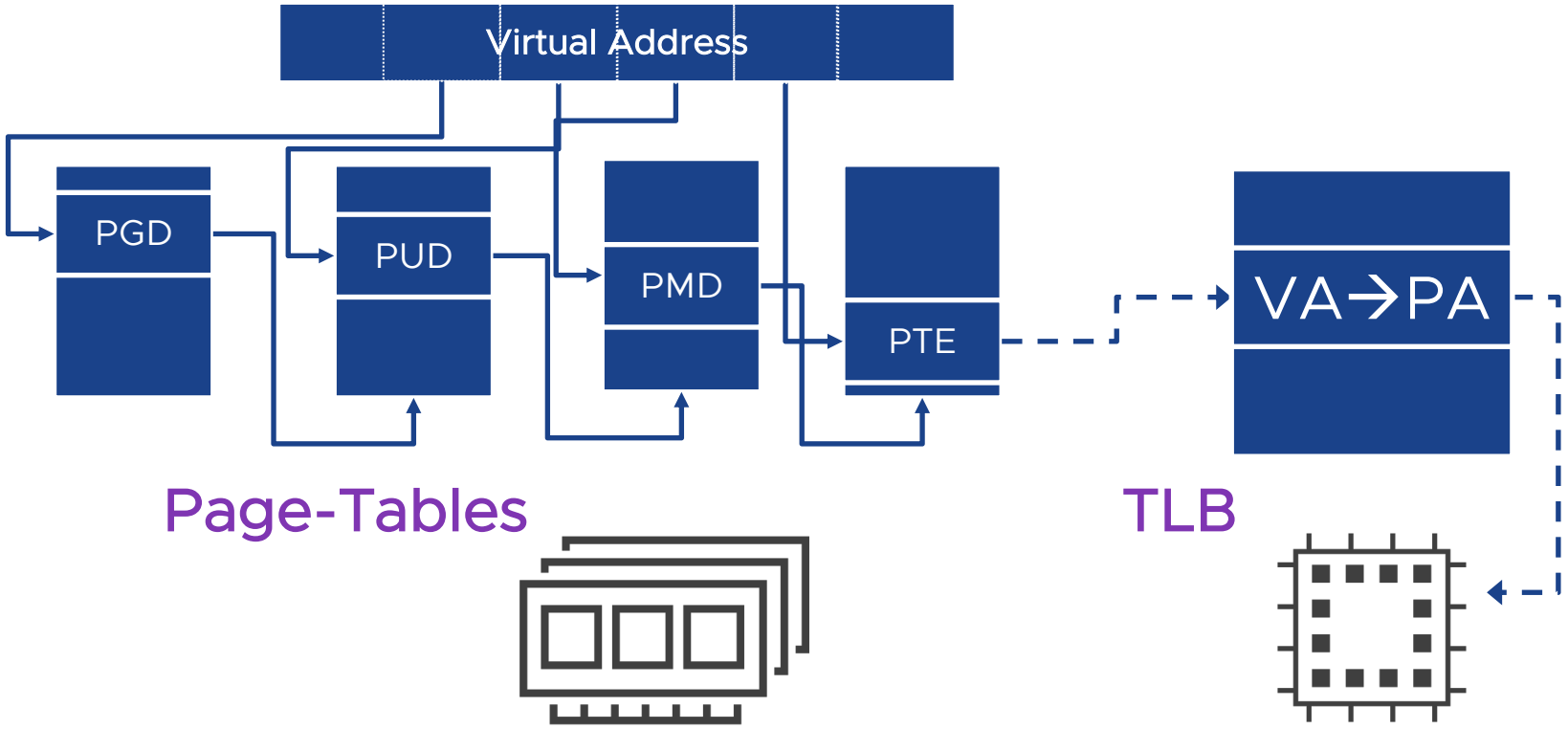
Don't shoot down TLB shootdowns!

Nadav Amit, Amy Tai,
Michael Wei

April 2020

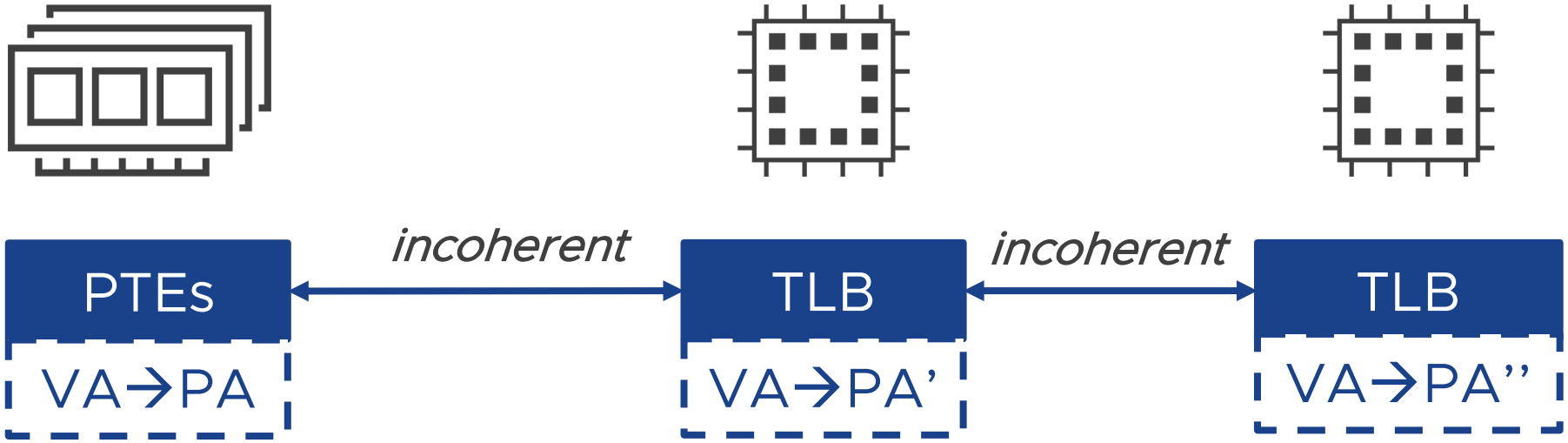
vmware

Translation Lookaside Buffer (TLB)



TLB = cache for virtual to physical address translations

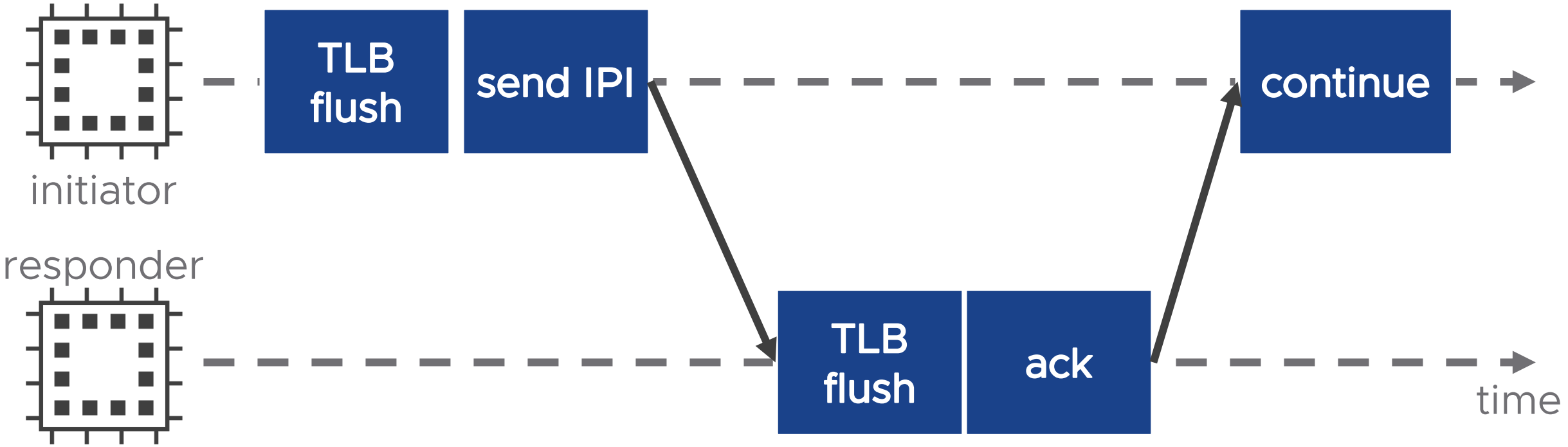
TLB Coherency



Hardware does not maintain TLBs coherent

The problem is left for software (OS)

TLB Shootdown (in Linux)



Challenge

TLB shootdowns are expensive.

How can we further optimize them?

This work focus on:

- Linux/x86 – common lessons
- Userspace mappings – common case

Lessons are relevant to other environments

Existing Solutions

Hardware based TLB invalidations

- Not available on all architectures

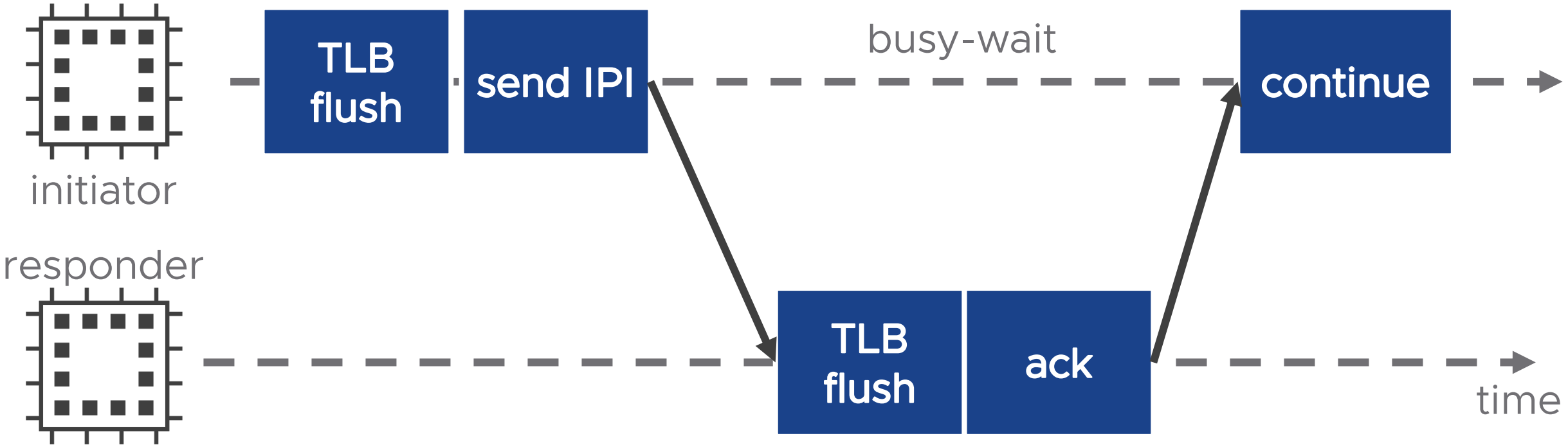
- Does not coexist (yet) with software techniques:
  - No selective target cores for TLB invalidation

Software solutions

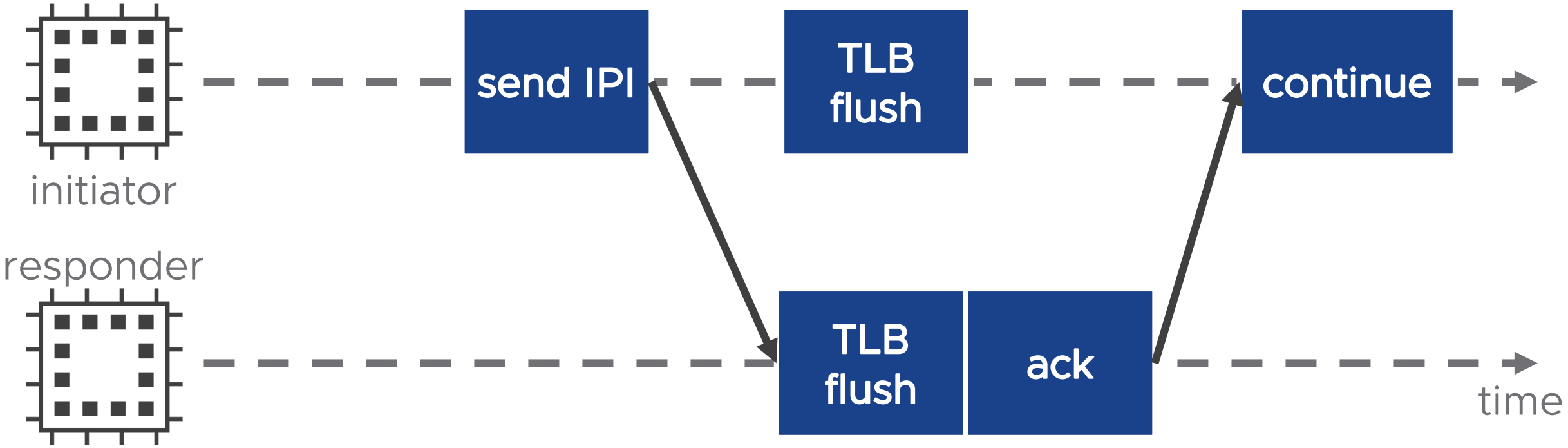
- Replicating page-tables [RadixVM, Clements'13]
  - Can increase overhead with low-latency IPIs

- Aggressive batching [LATR, Kumar'18]
  - Breaks POSIX semantics

TLB Flushes in Linux and FreeBSD

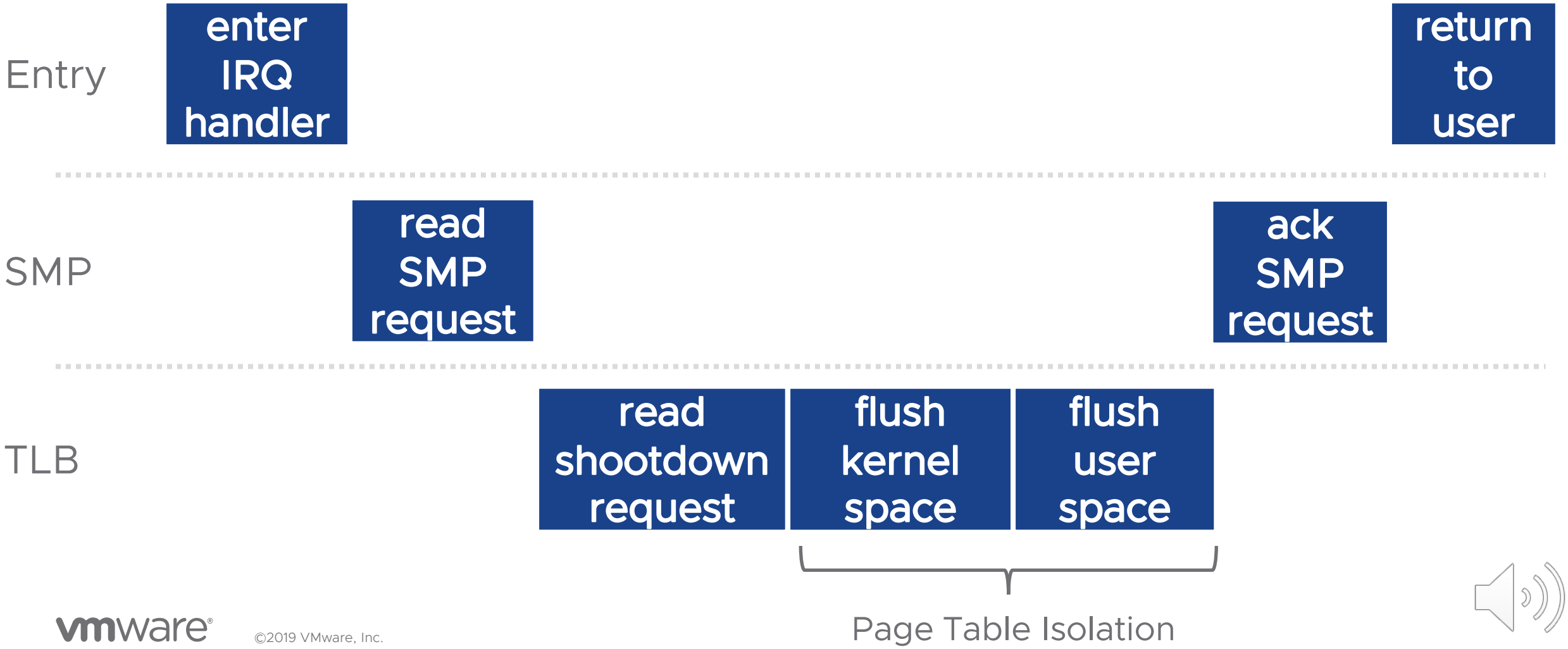


Optimization 1: Concurrent Flushes (forgotten lesson)

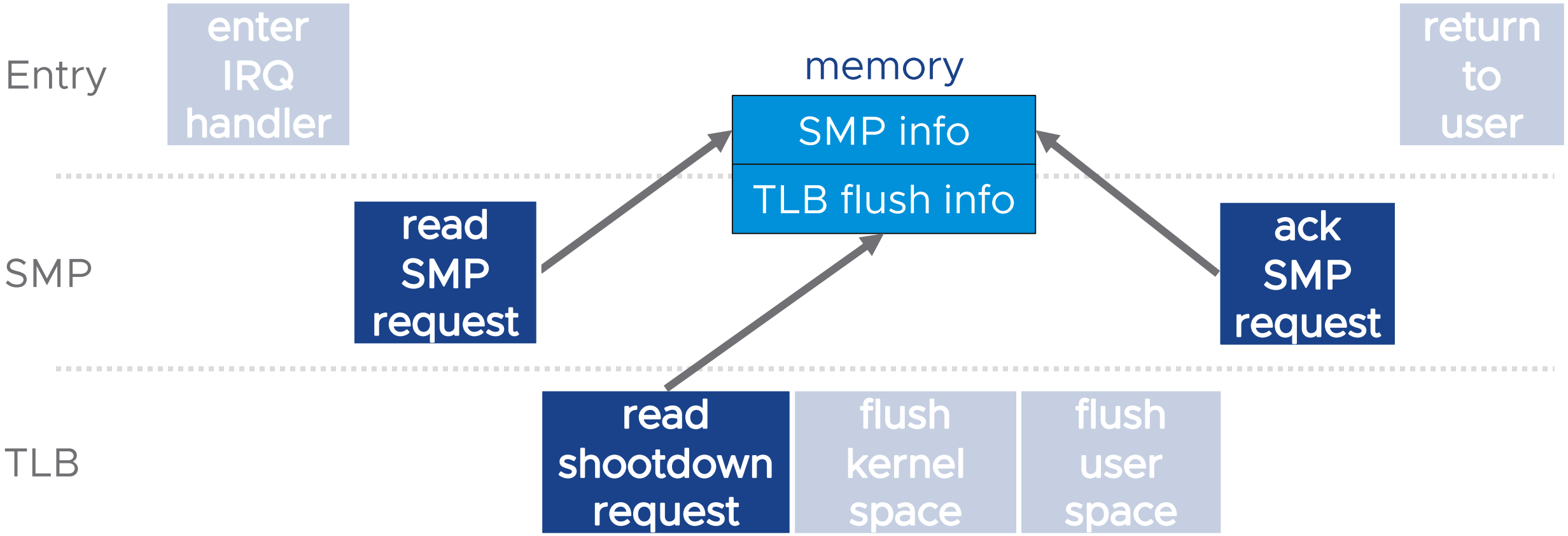


RP3 TLB consistency algorithm [Rosenburg'89]

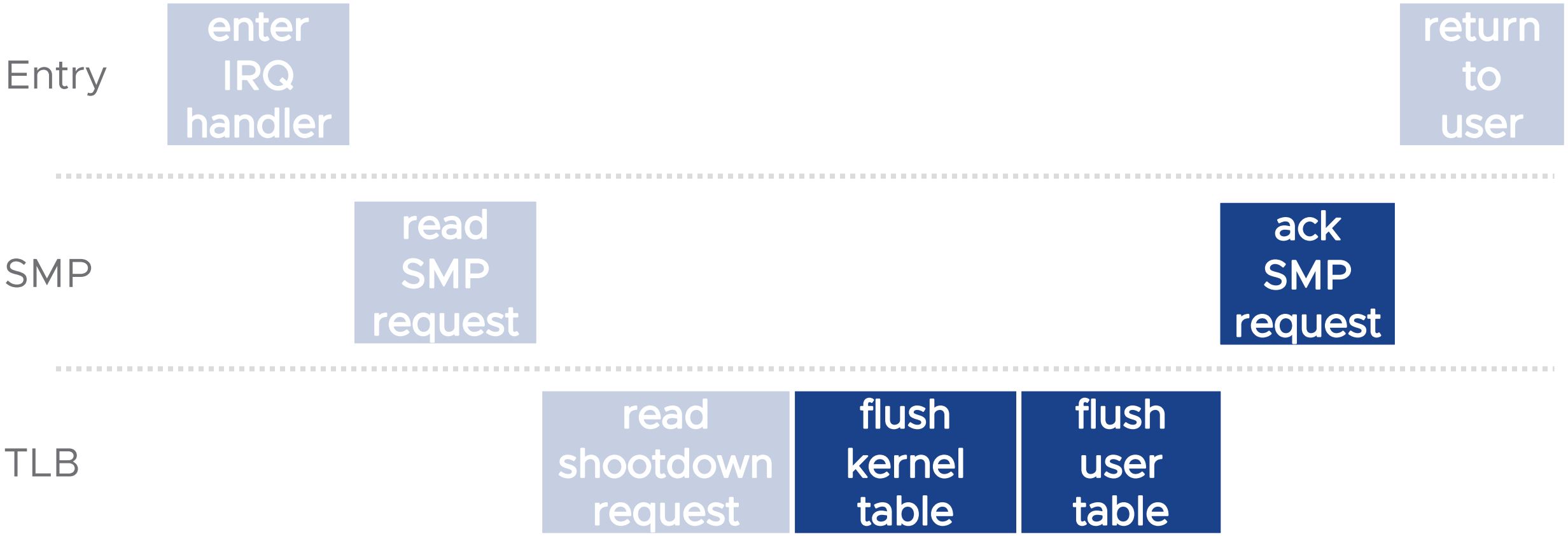
TLB Shootdown Responder



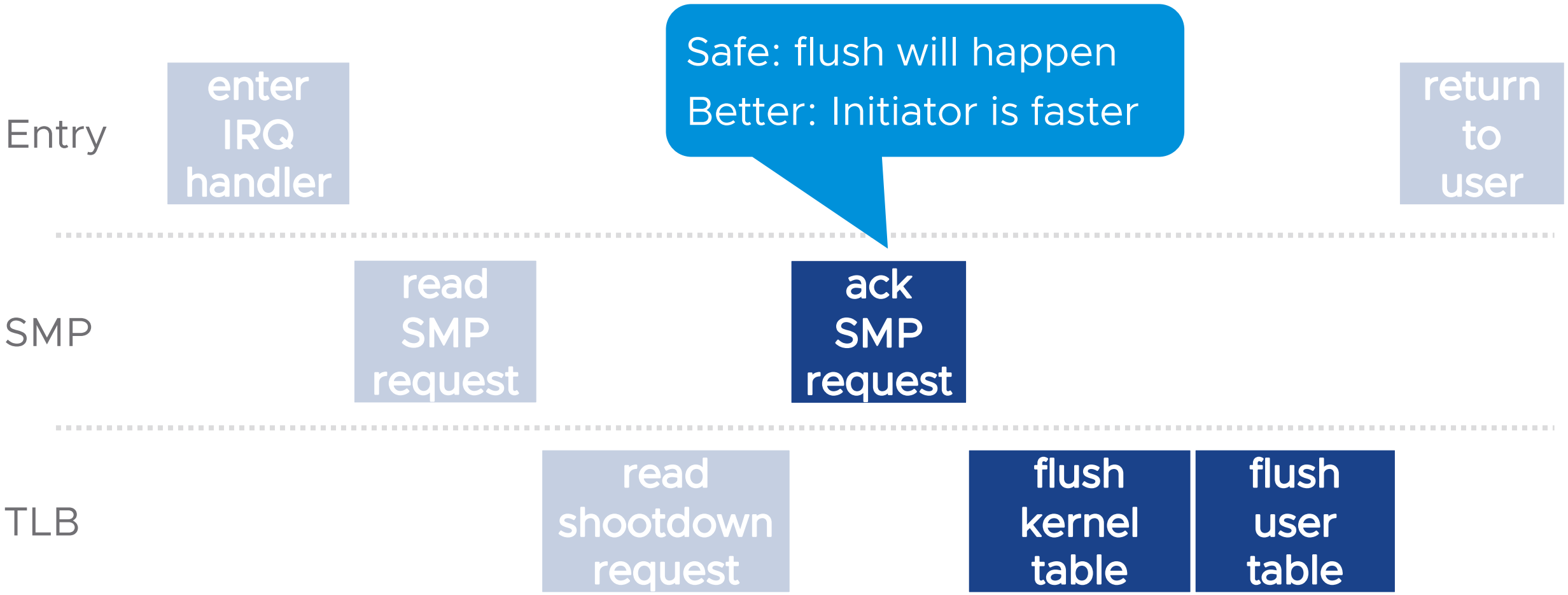
Optimization 2: Cacheline Consolidation



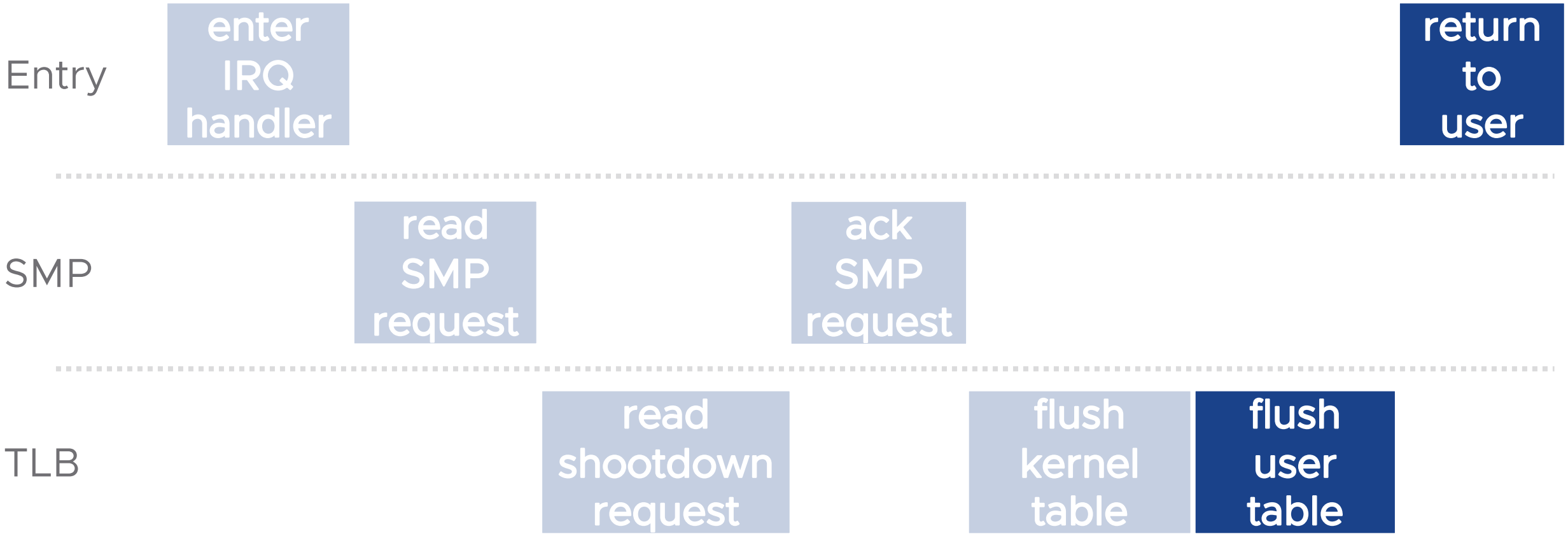
Optimization 3: Early Acknowledgment



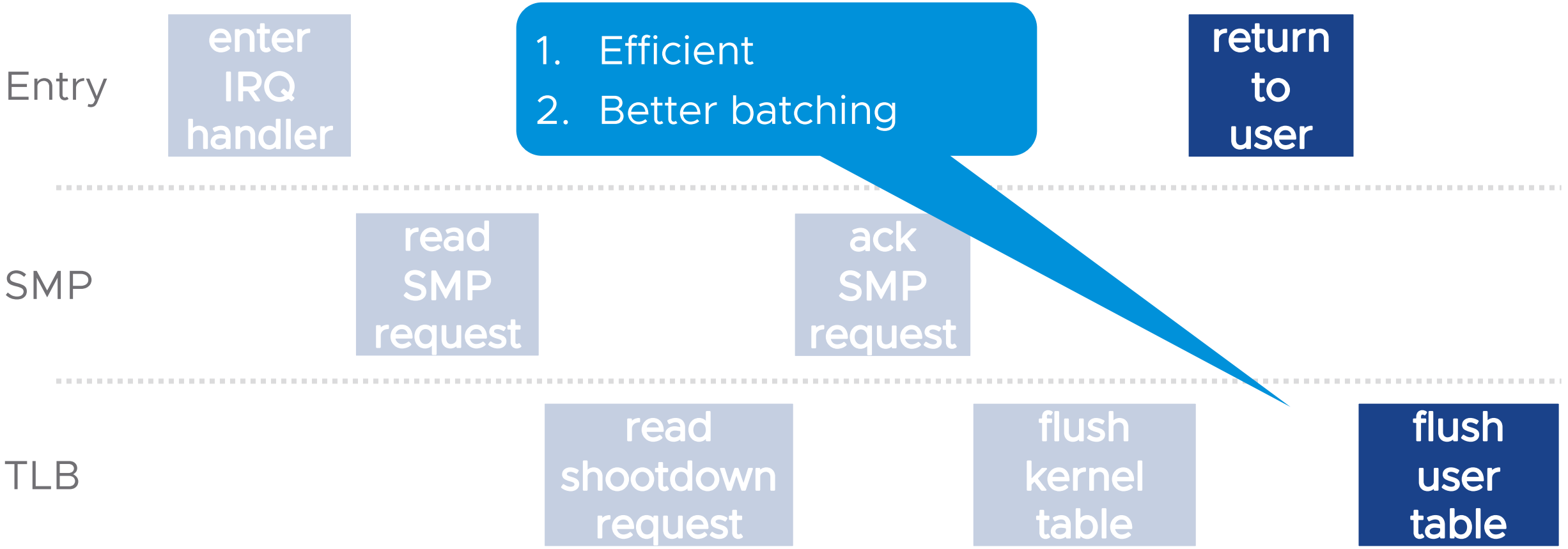
Optimization 3: Early Acknowledgment



Optimization 4: In-Context Flushes



Optimization 4: In-Context Flushes



In the Paper

Userspace-safe batching

- Deferring TLB shootdowns while the kernel runs

Avoiding TLB flushes on **Copy-on-Write**

- Special case we can optimize

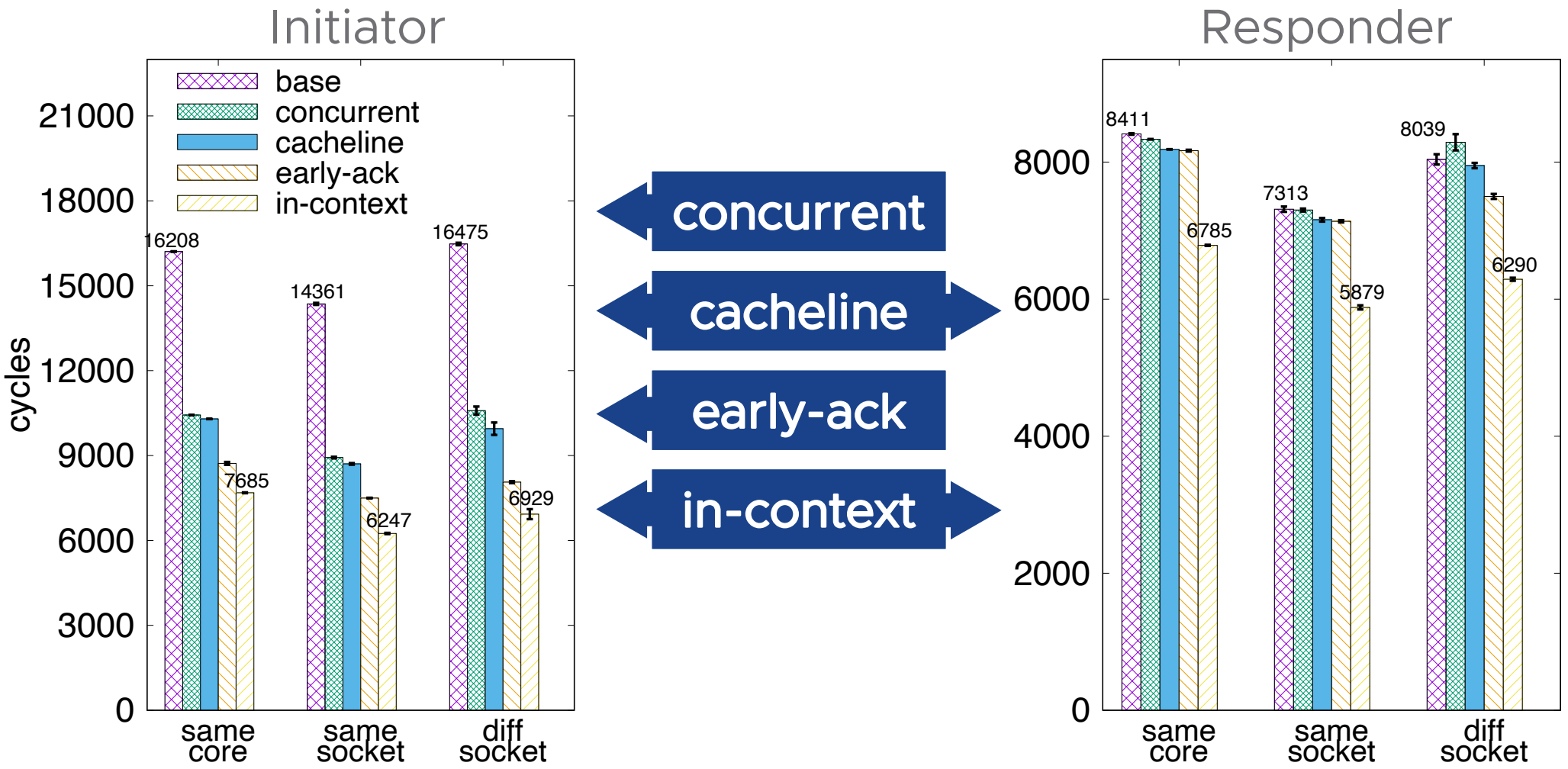
TLB flushes in **virtualization**

- The effect of page size mismatch

Many important and subtle details

Evaluation: Unmapping and Flushing 10 PTEs

madvise(MADV_DONTNEED)



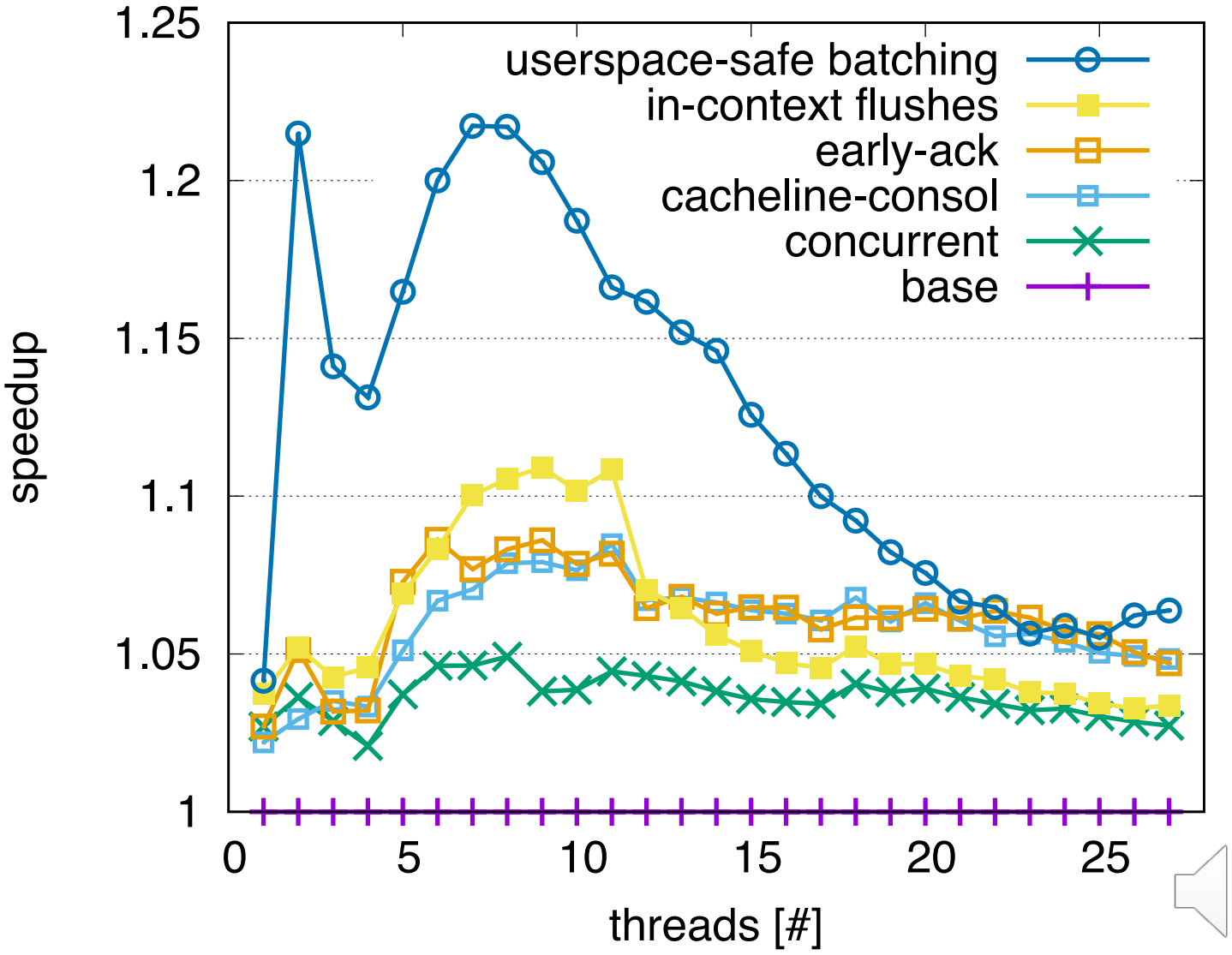
Evaluation: SysBench – Random Writes

Random writes

Periodic flushes

Memory-mapped file

Emulated persistent memory, no write-cache



Conclusions

TLB shootdown can be improved

Doing it well in software ➔ better hardware interfaces

We are working to push these enhancements to Linux