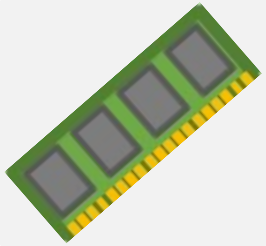
NrOS: Effective Replication and Sharing in an Operating System

Ankit Bhardwaj, Chinmay Kulkarni, Reto Achermann, Irina Calciu,
Sanidhya Kashyap, Ryan Stutsman, Amy Tai, and Gerd Zellweger

THE UNIVERSITY OF UTAH®
THE UNIVERSITY OF BRITISH COLUMBIA
vmware®
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

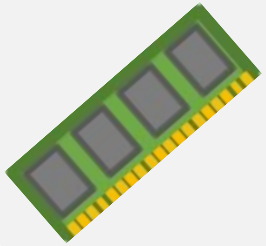
Kernel development is hard

Kernel development is hard



Manual memory management

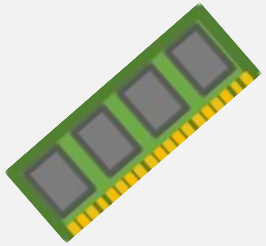
Kernel development is hard



Manual memory management

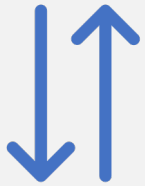
Type-unsafe code

Kernel development is hard



Manual memory management

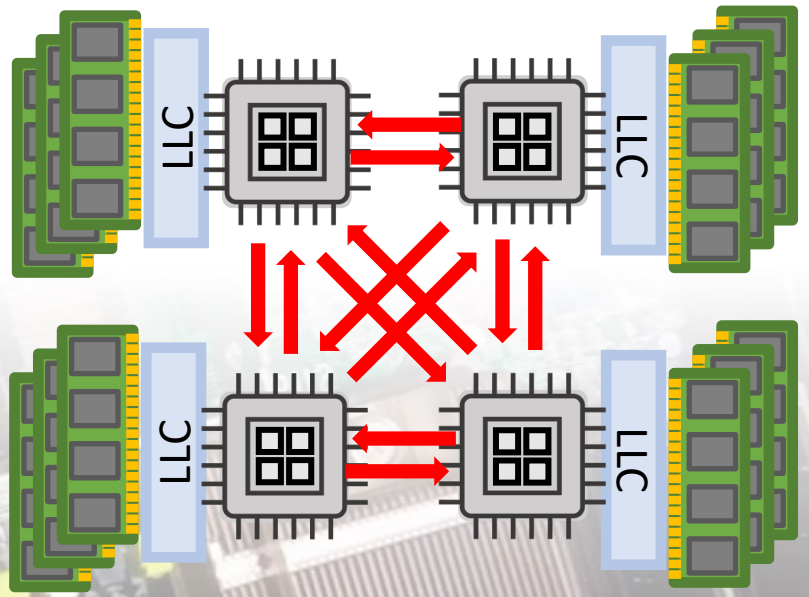
Type-unsafe code



Complex and asynchronous event handling

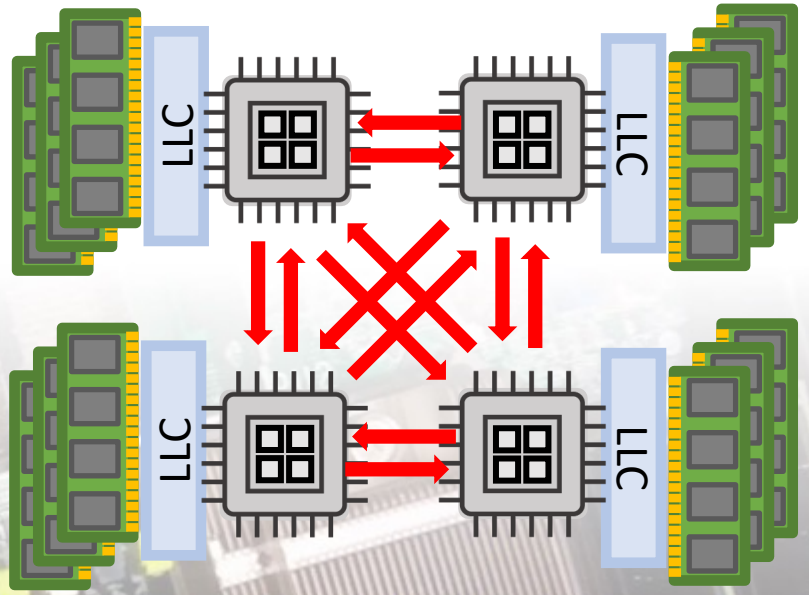
Concurrency makes it harder

- Increasing core counts
- NUMA



Concurrency makes it harder

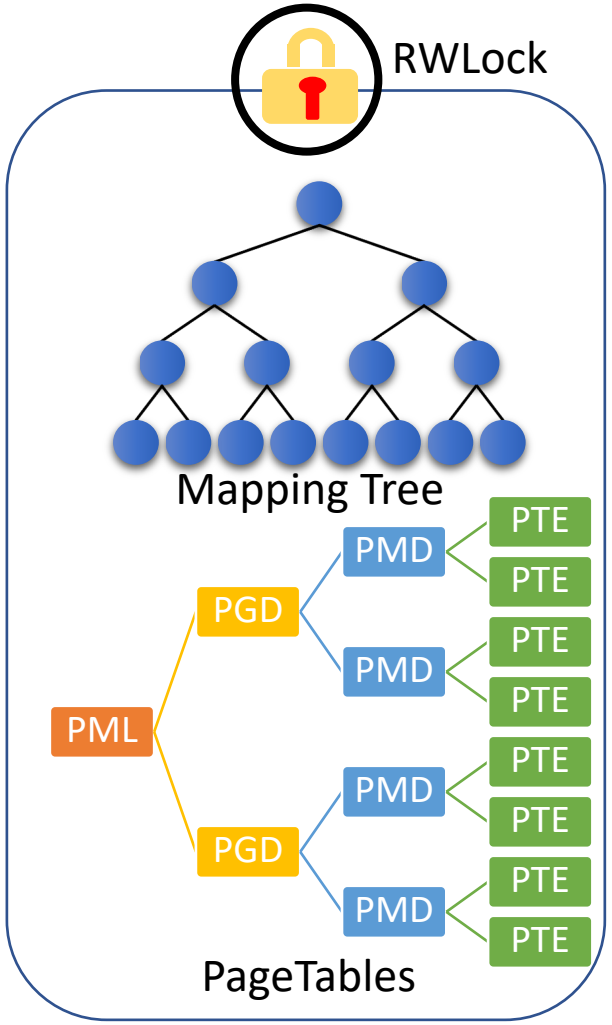
- Increasing core counts
- NUMA

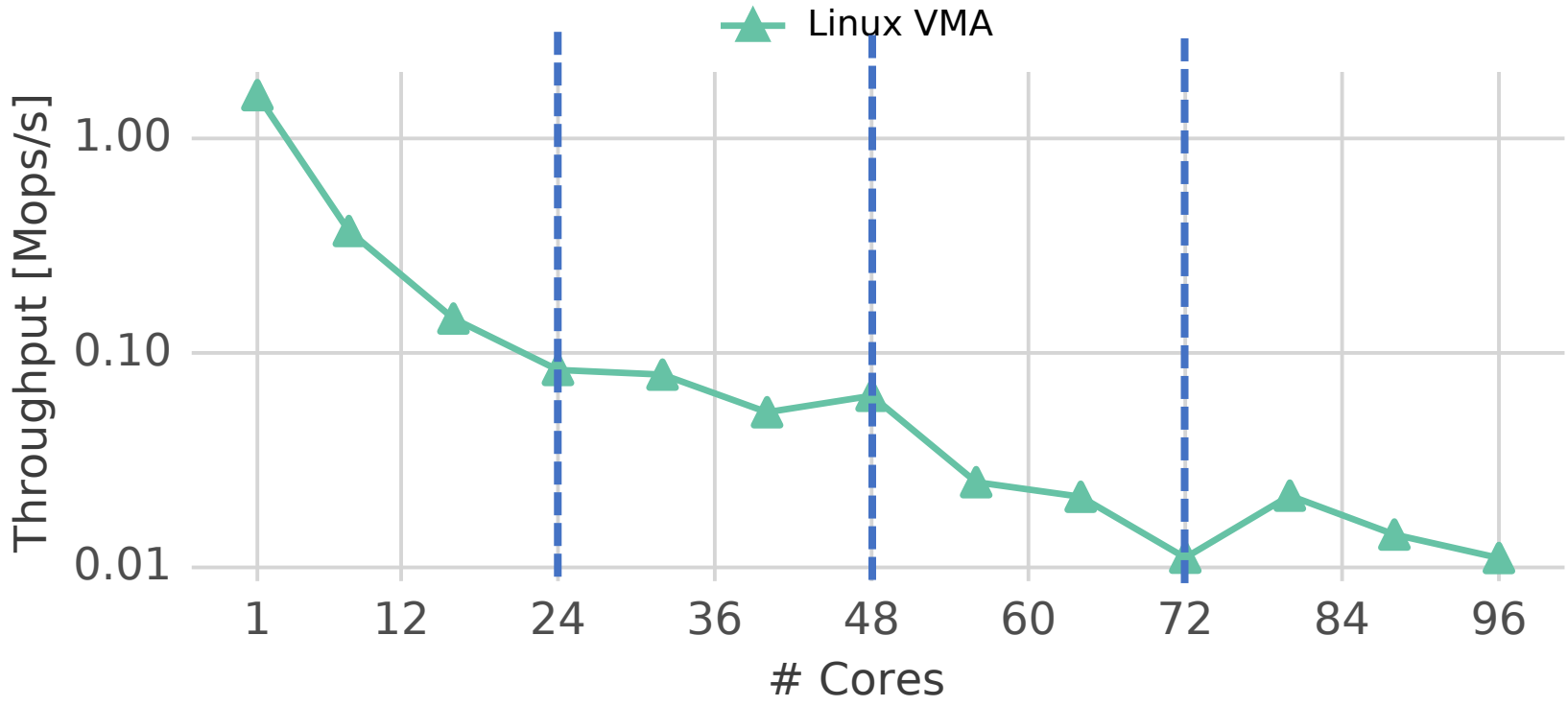


Solution: Use coarse locks around sequential data structures

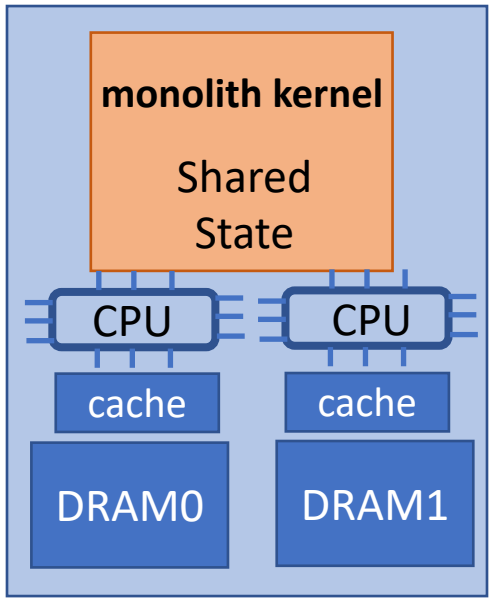
Simple & black-box approach

Big locks are bad for performance





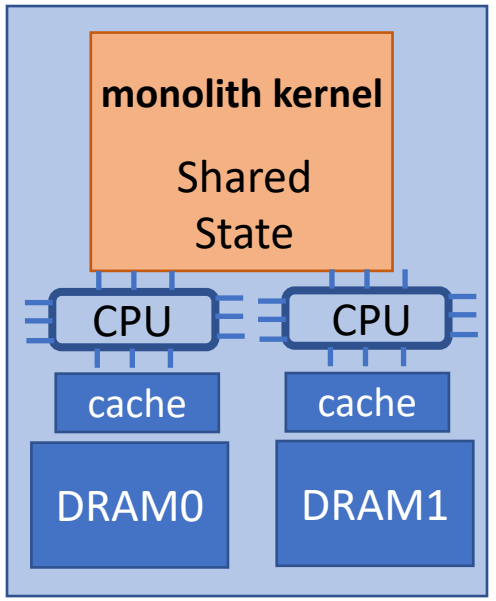
Mitigations in monolithic kernels



Mitigations in monolithic kernels

Monolithic kernels share state across cores

- Use fine-grained locking
- Lock-free data-structures
- Relativistic programming (RCU etc.)

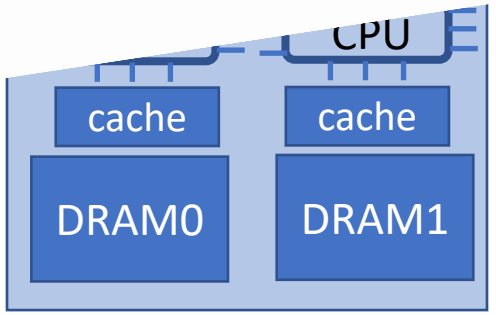


Lots of additional complexity

Mitigations in monolithic kernels

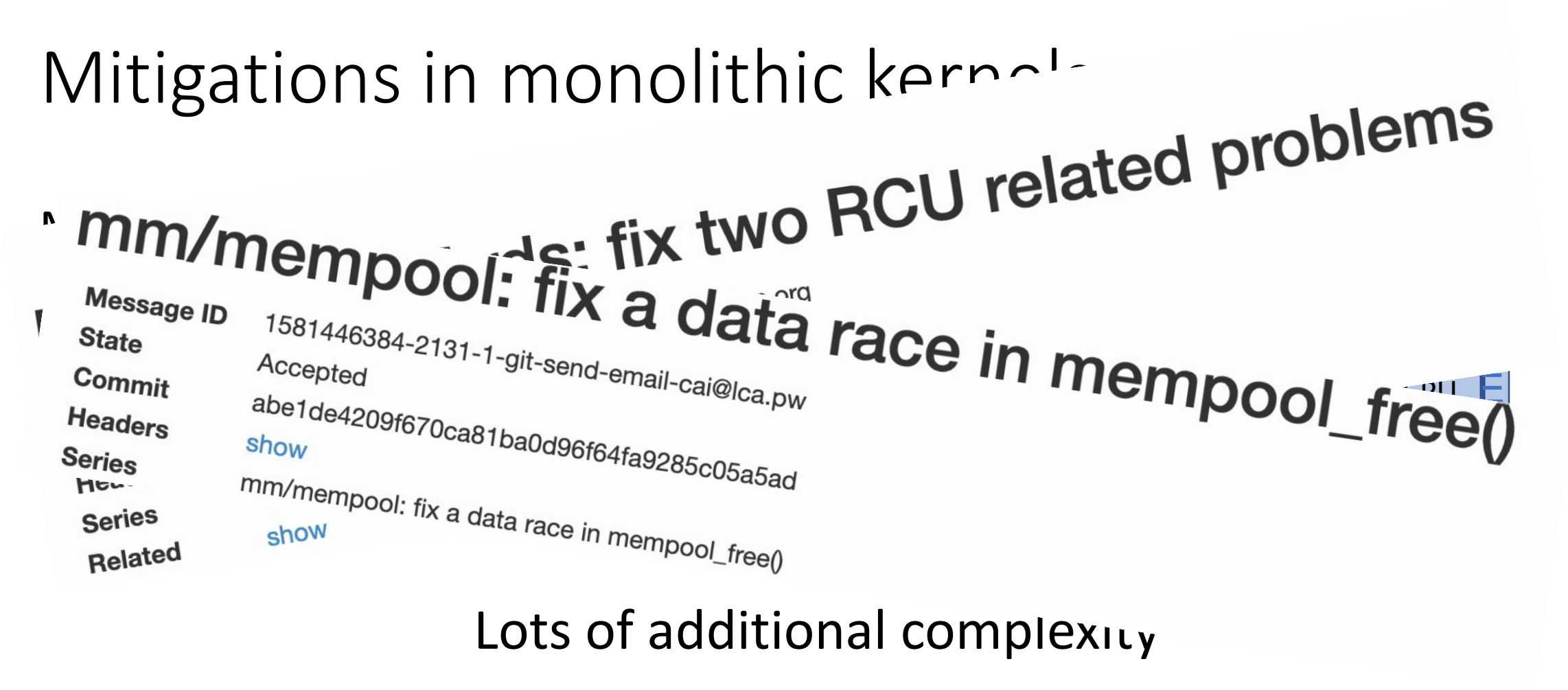
[4.18,004/235] rds: fix two RCU related problems

Message ID	20180924113104.696412478@linuxfoundation.org
State	New, archived
Headers	show
Series	4.18.10-stable review
Related	show



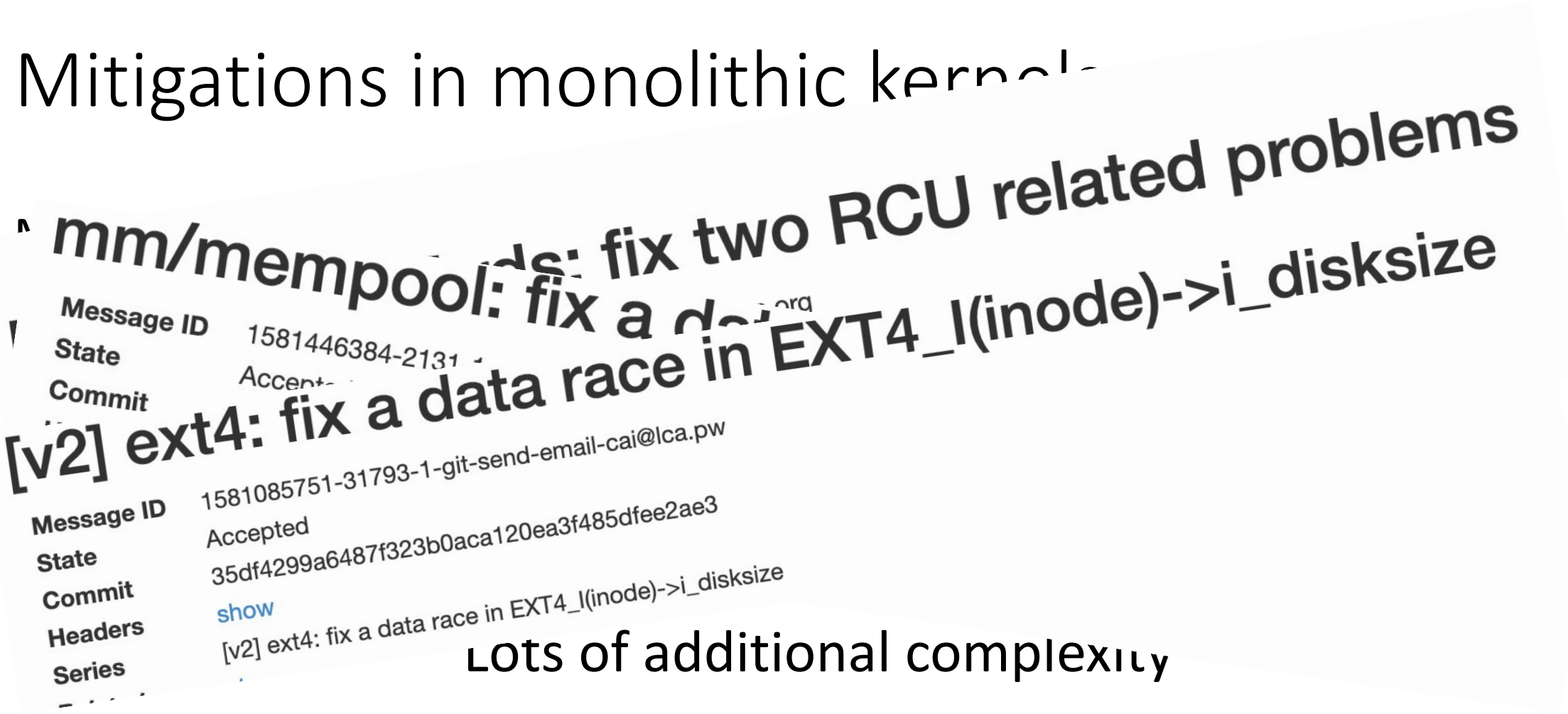
Lots of additional complexity

Mitigations in monolithic kernels



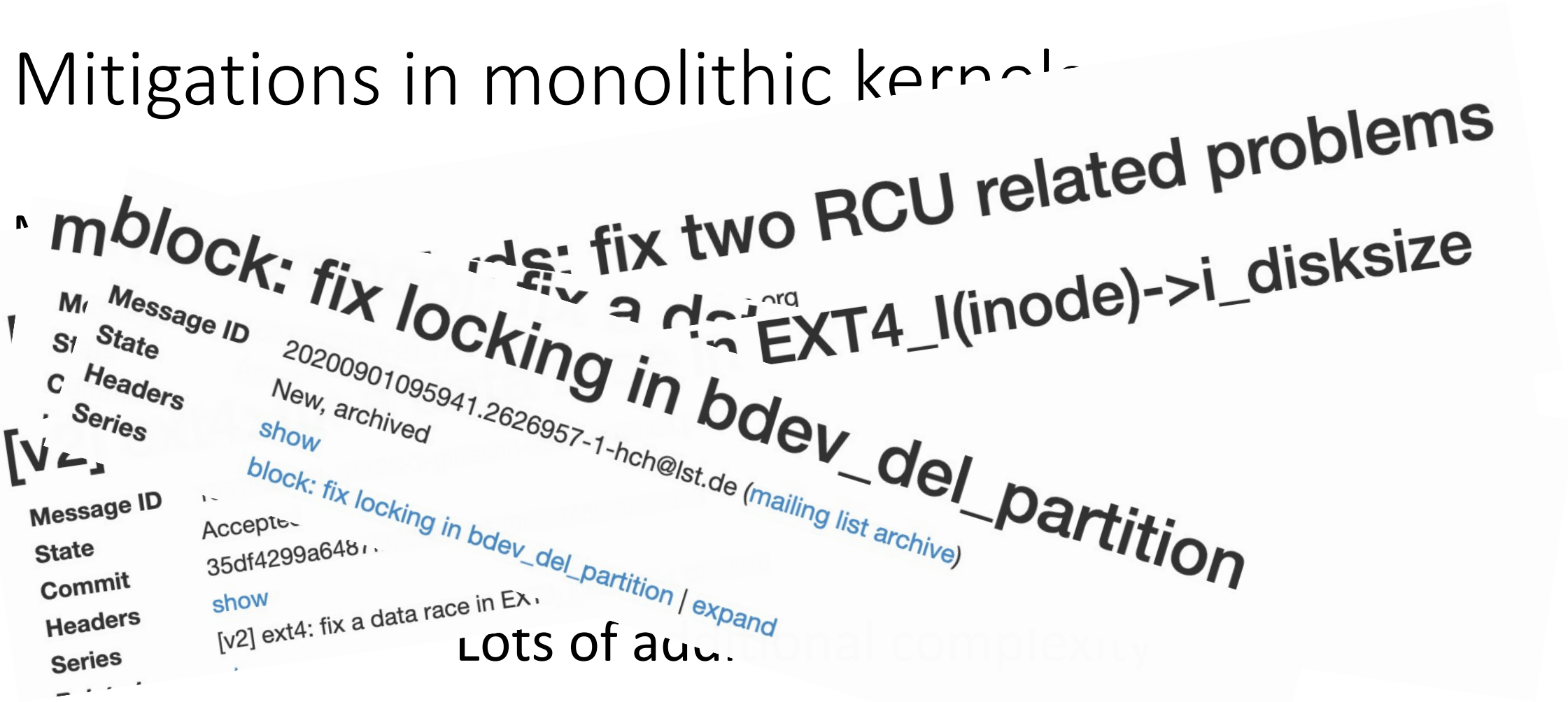
Lots of additional complexity

Mitigations in monolithic kernels



Lots of additional complexity

Mitigations in monolithic kernels



An error has occured.We dont even know what is it.So cant fix it ,
and you have restart your computer.By the way if you restart your ,
computer you will lose any unsaved information in all open applications.
In the other hand you dont have any other option :)

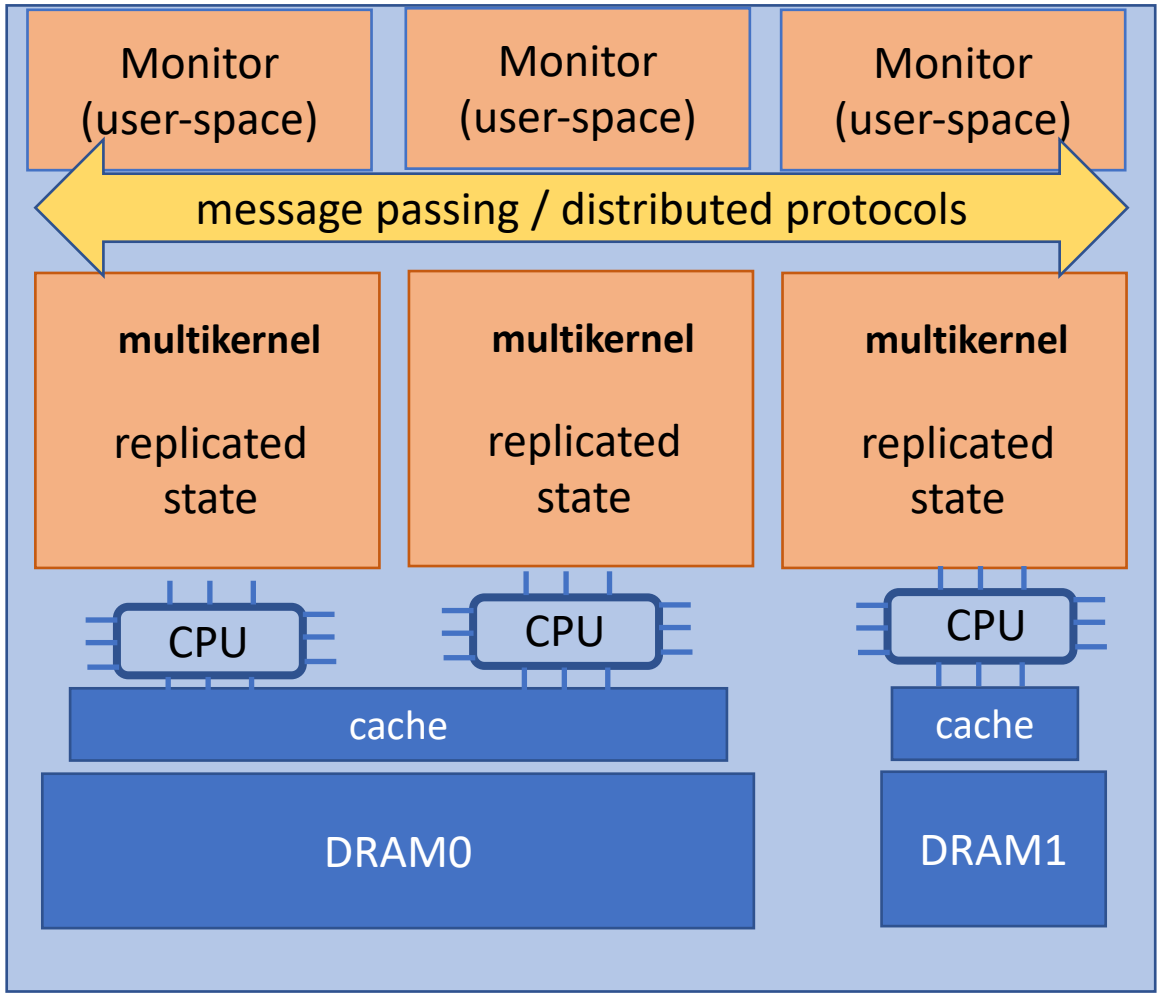
Press Enter to return to Windows(It wont work), or

Press CTRL+ALT+DEL to restart your computer.

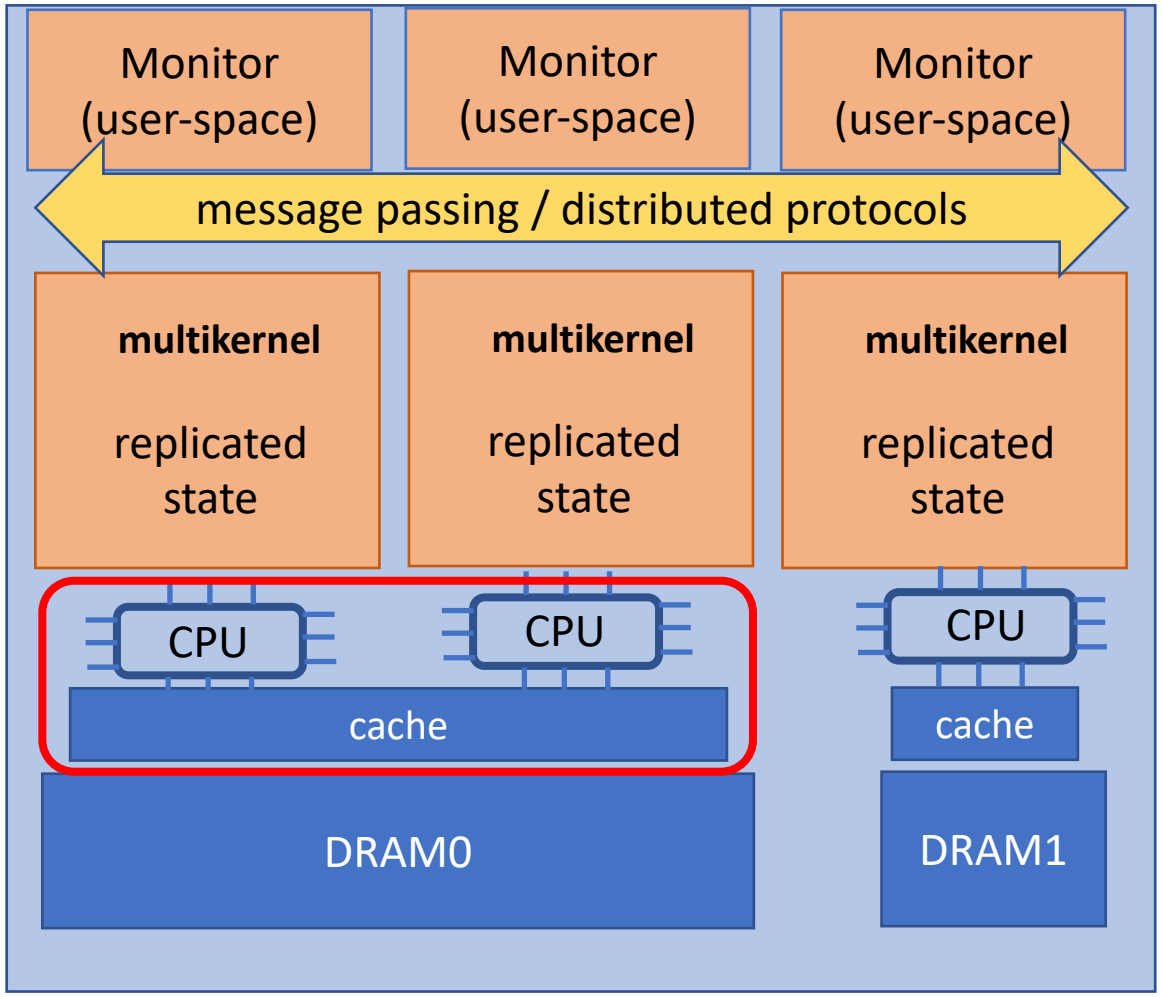
Error : 0E : 016F : BFF9B3D4

Uni.Q Design

Multikernel: Barrelfish's approach

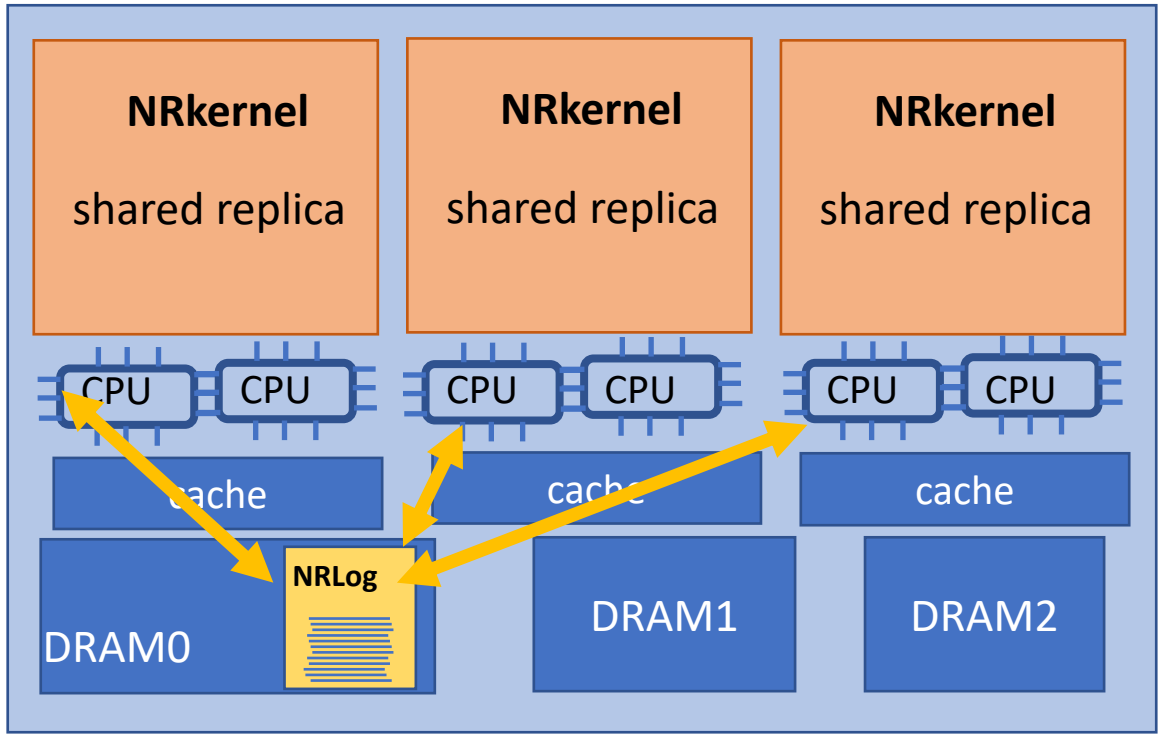


Multikernel: Barrelfish's approach

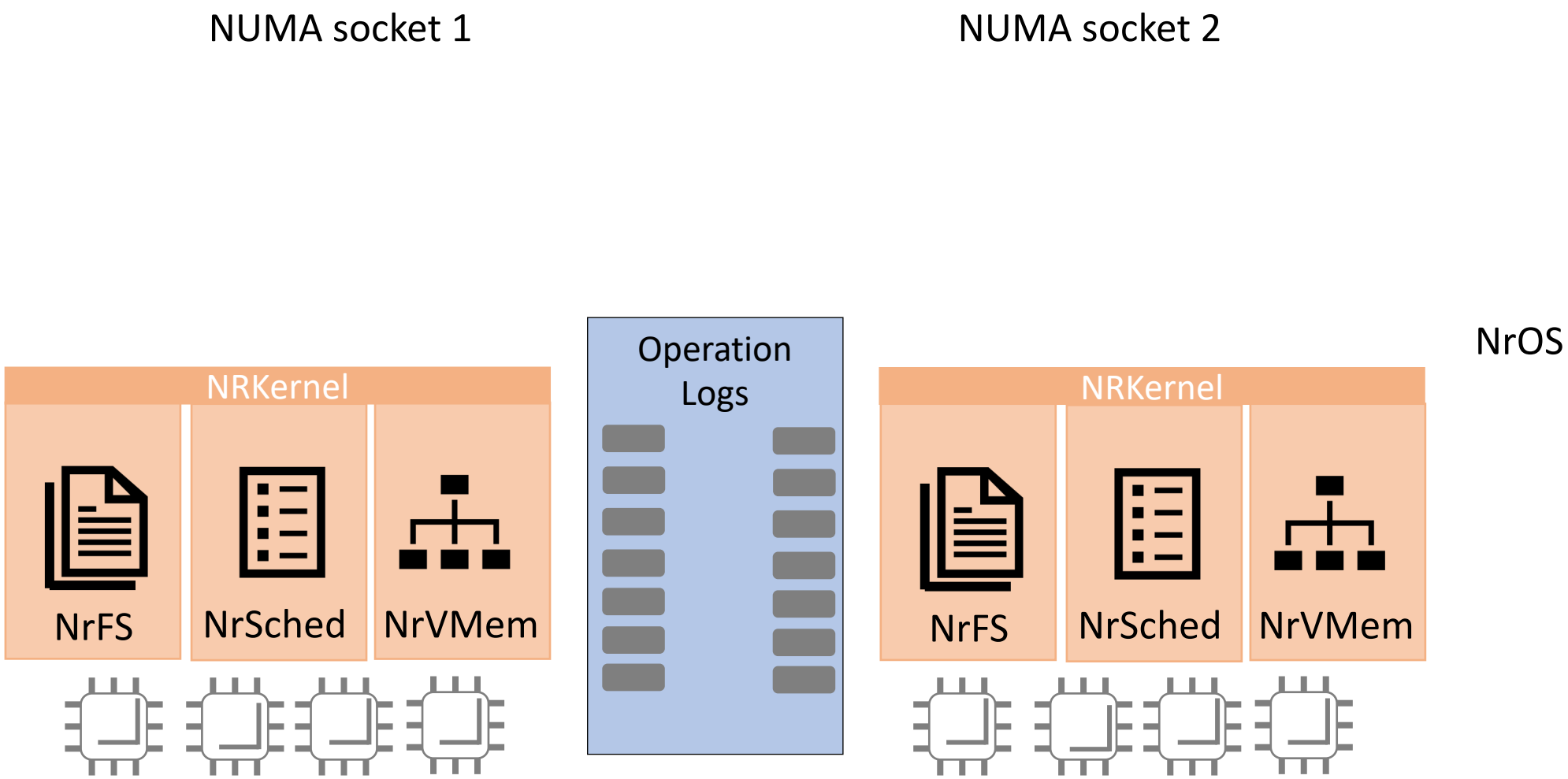


NRkernel: A log-based multikernel

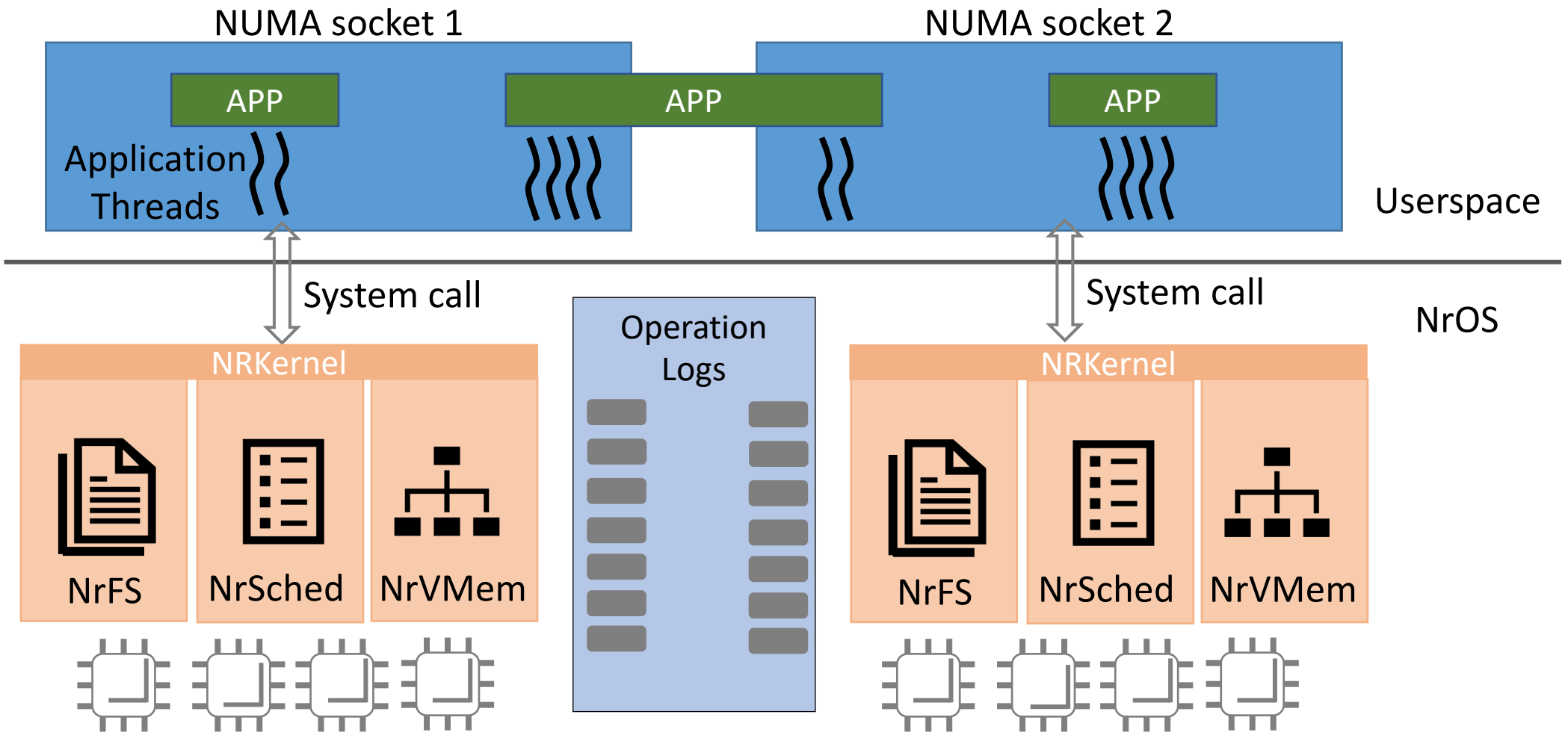
- Sequential kernel data structures
- Unified black-box approach to concurrency
- Replicates NRkernel structures
  - NUMA-local accesses
- Benefits of both message passing and shared memory



NrOS Overview



NrOS Overview

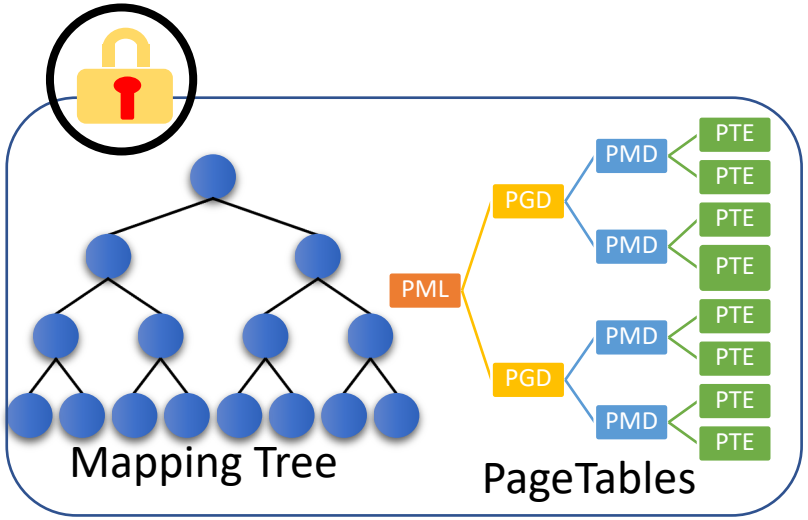


Node Replication: Log based replication

NUMA Node 1

NUMA Node 2

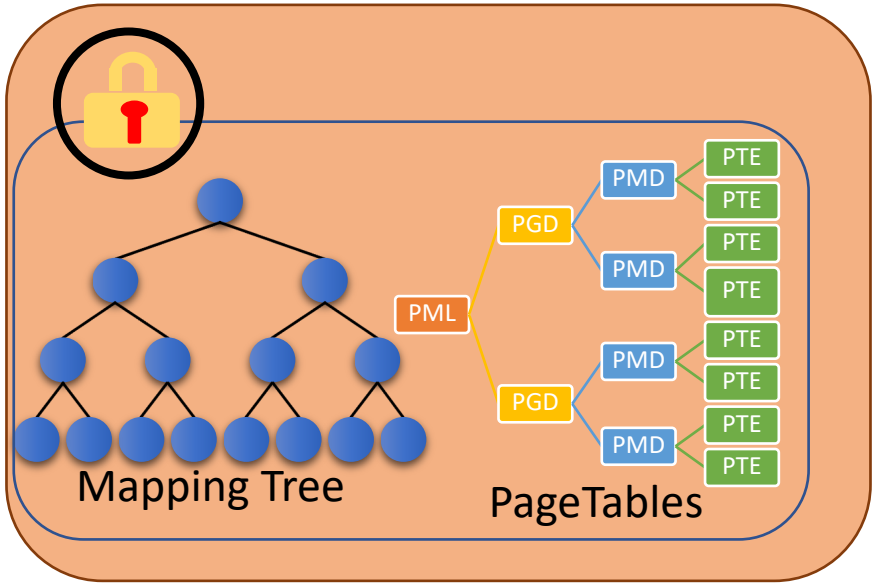
Node Replication: Log based replication



NUMA Node 1

NUMA Node 2

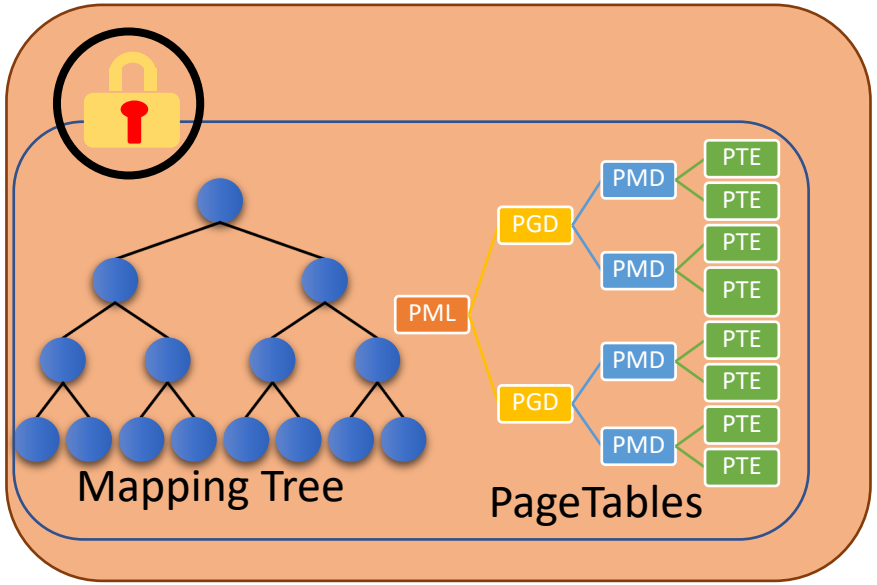
Node Replication: Log based replication



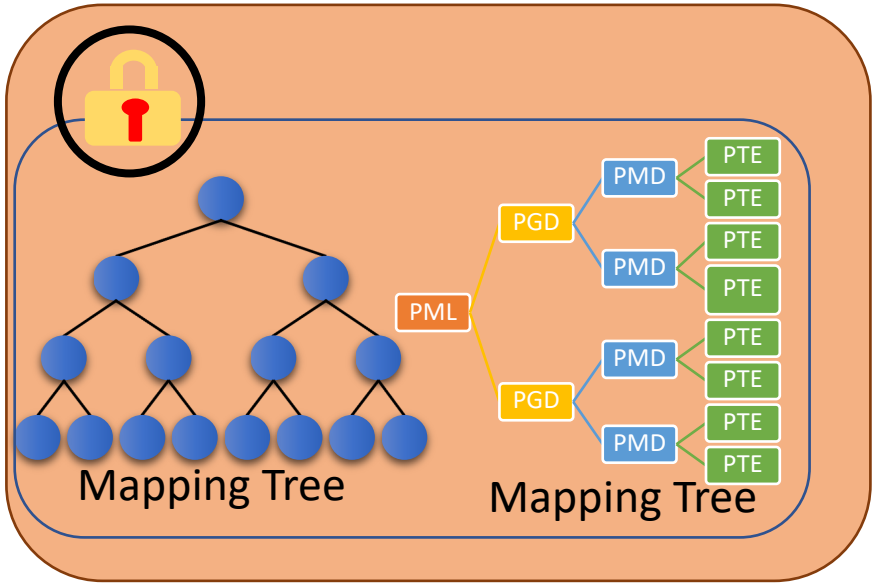
NUMA Node 1

NUMA Node 2

Node Replication: Log based replication

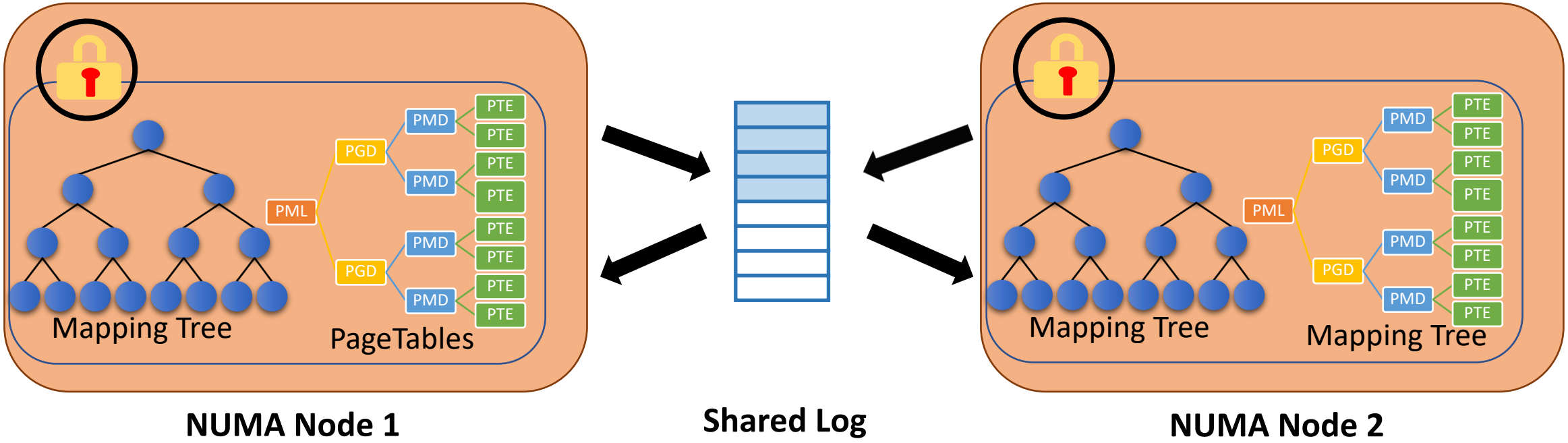


NUMA Node 1

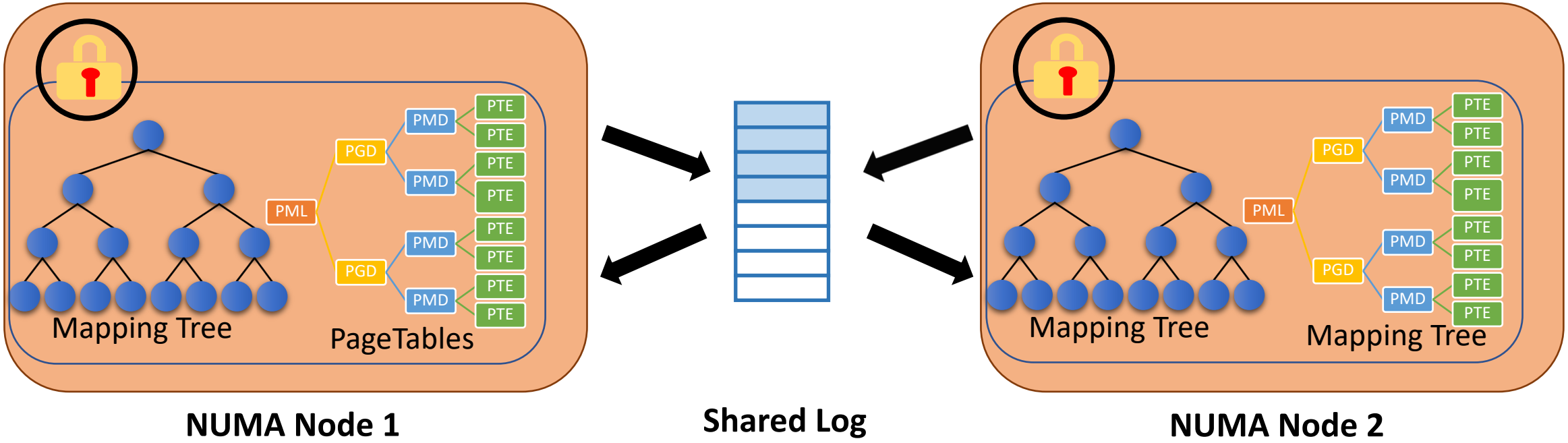


NUMA Node 2

Node Replication: Log based replication

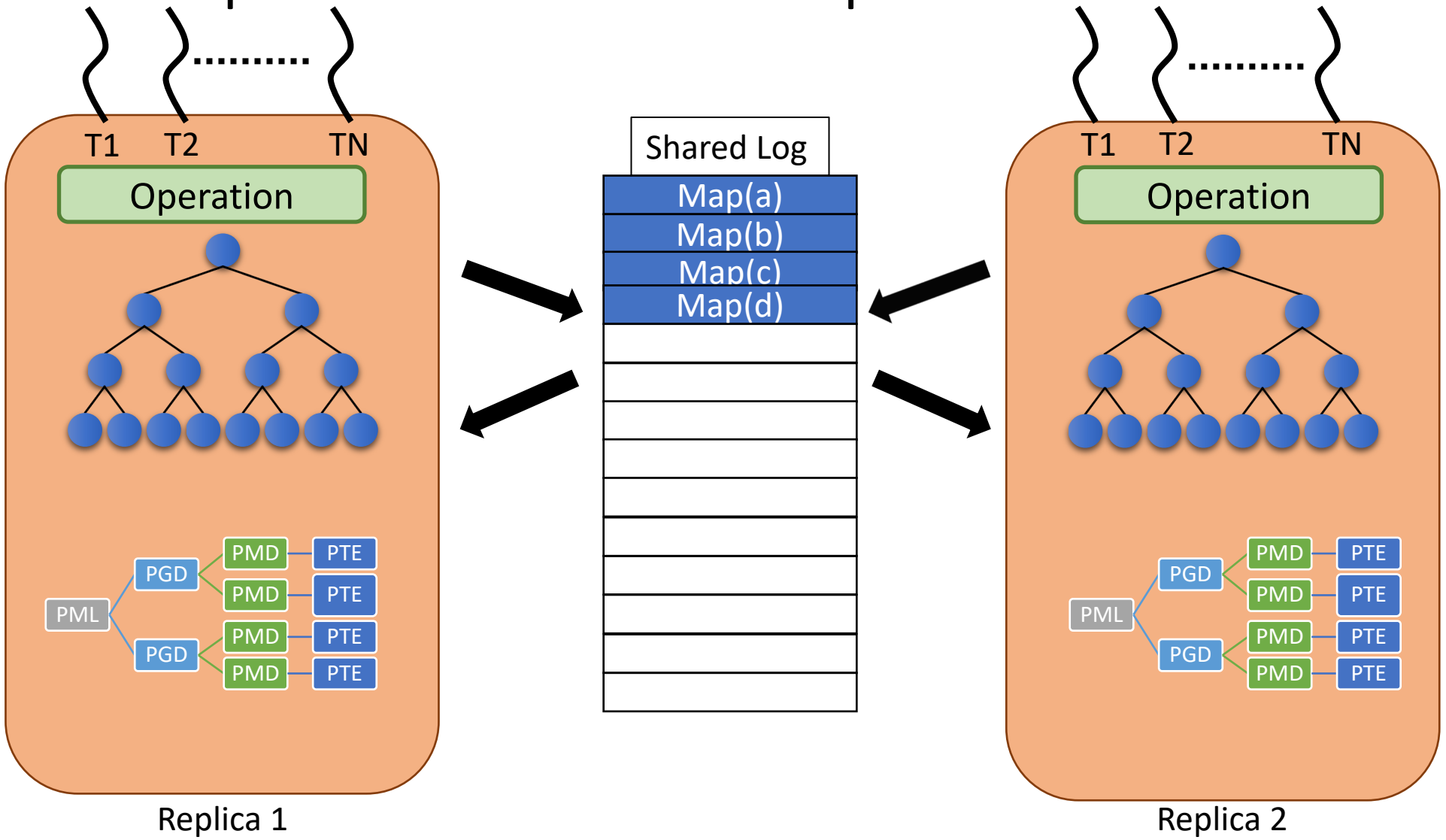


Node Replication: Log based replication

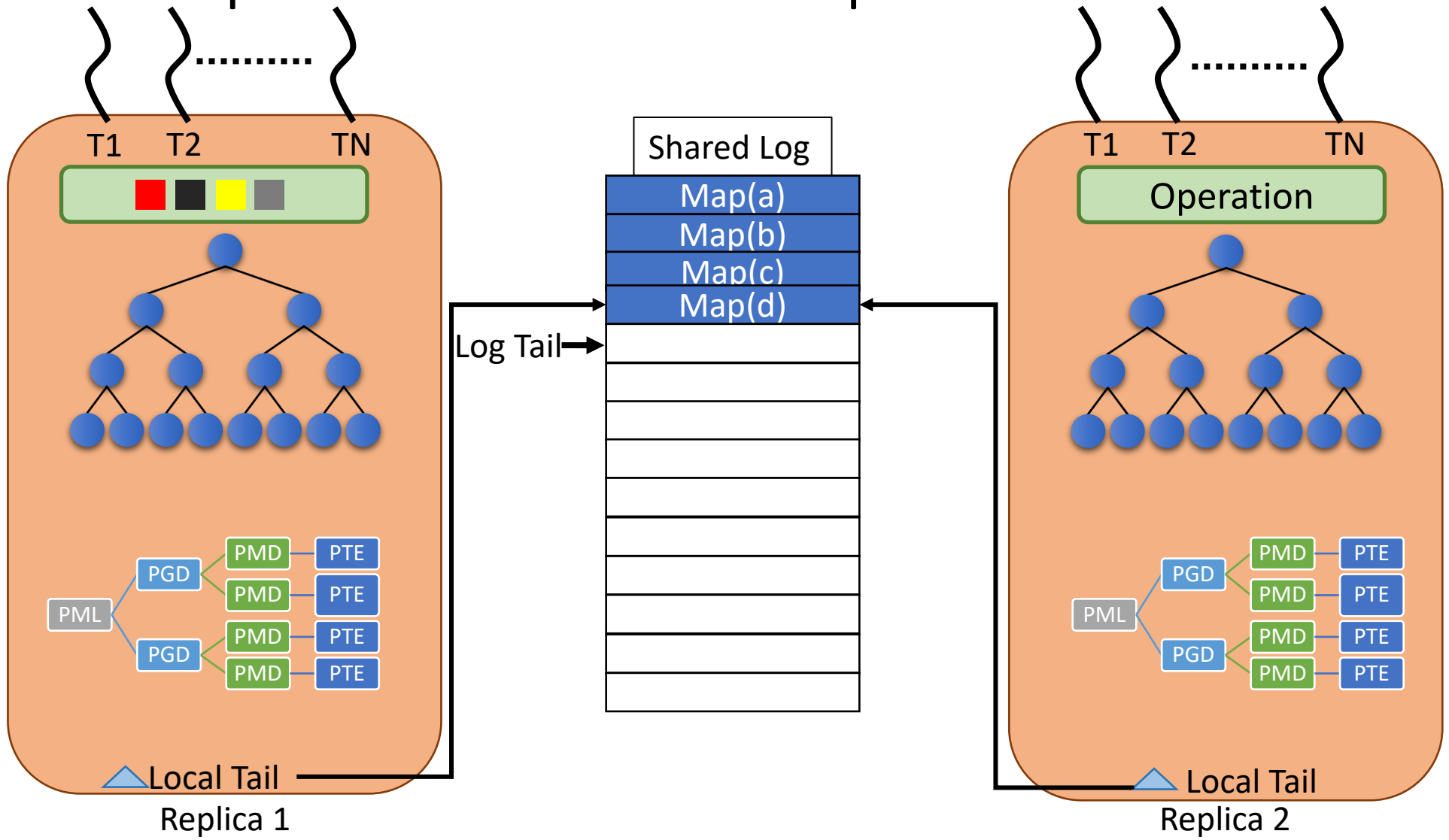


- Write operations replicate through the shared log
- Read operations access the local replica

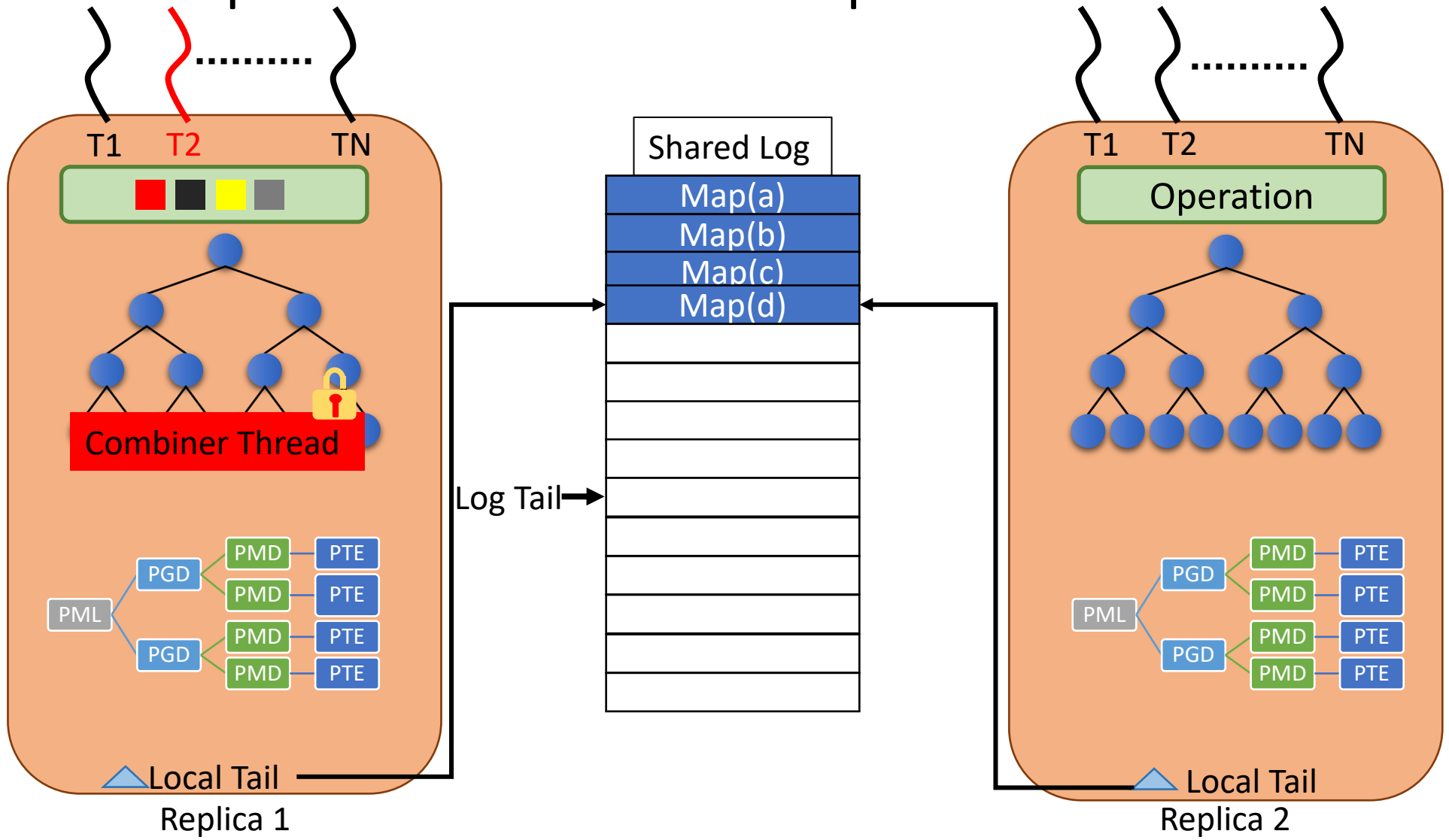
Node Replication: Write Operation



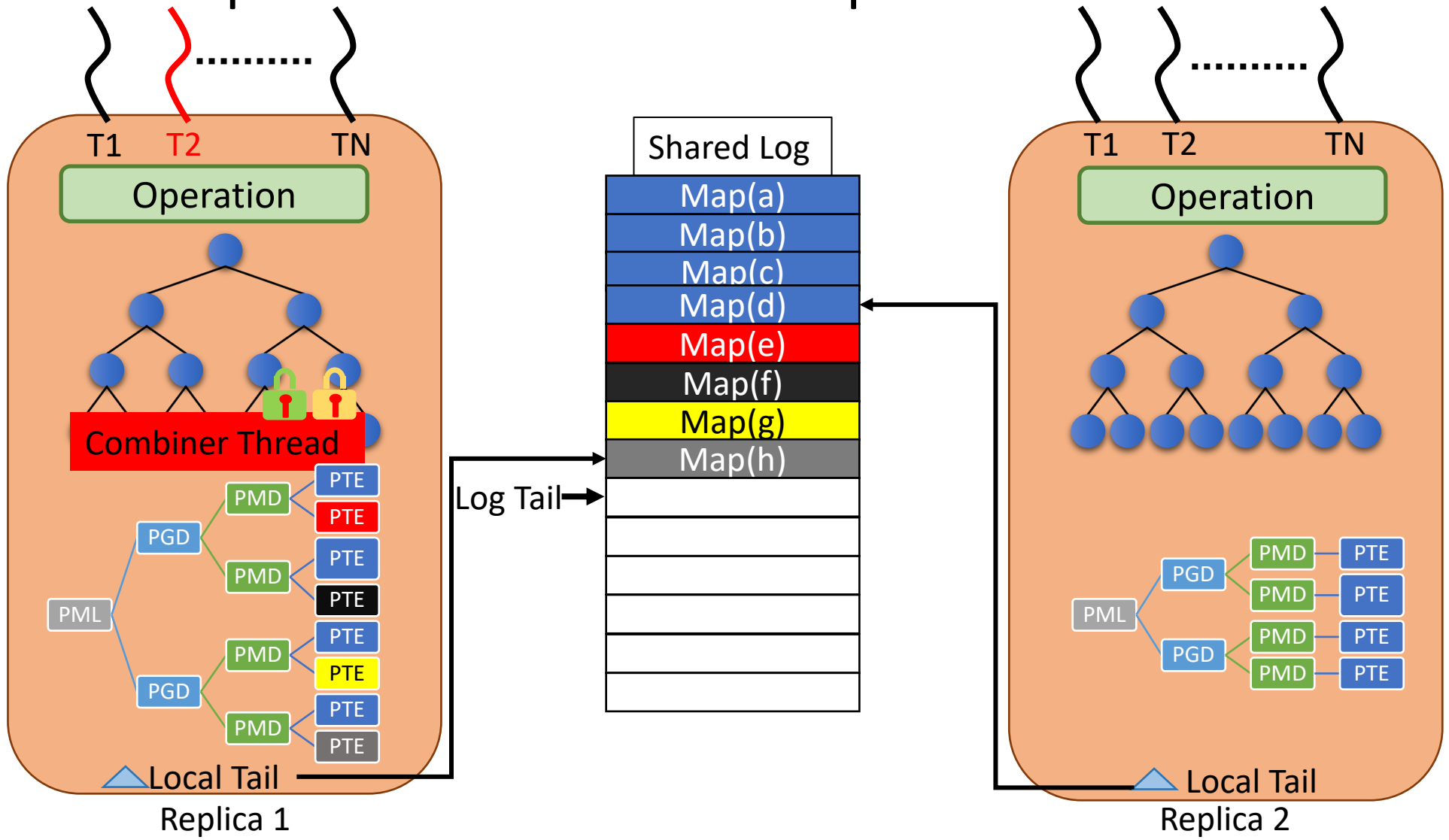
Node Replication: Write Operation



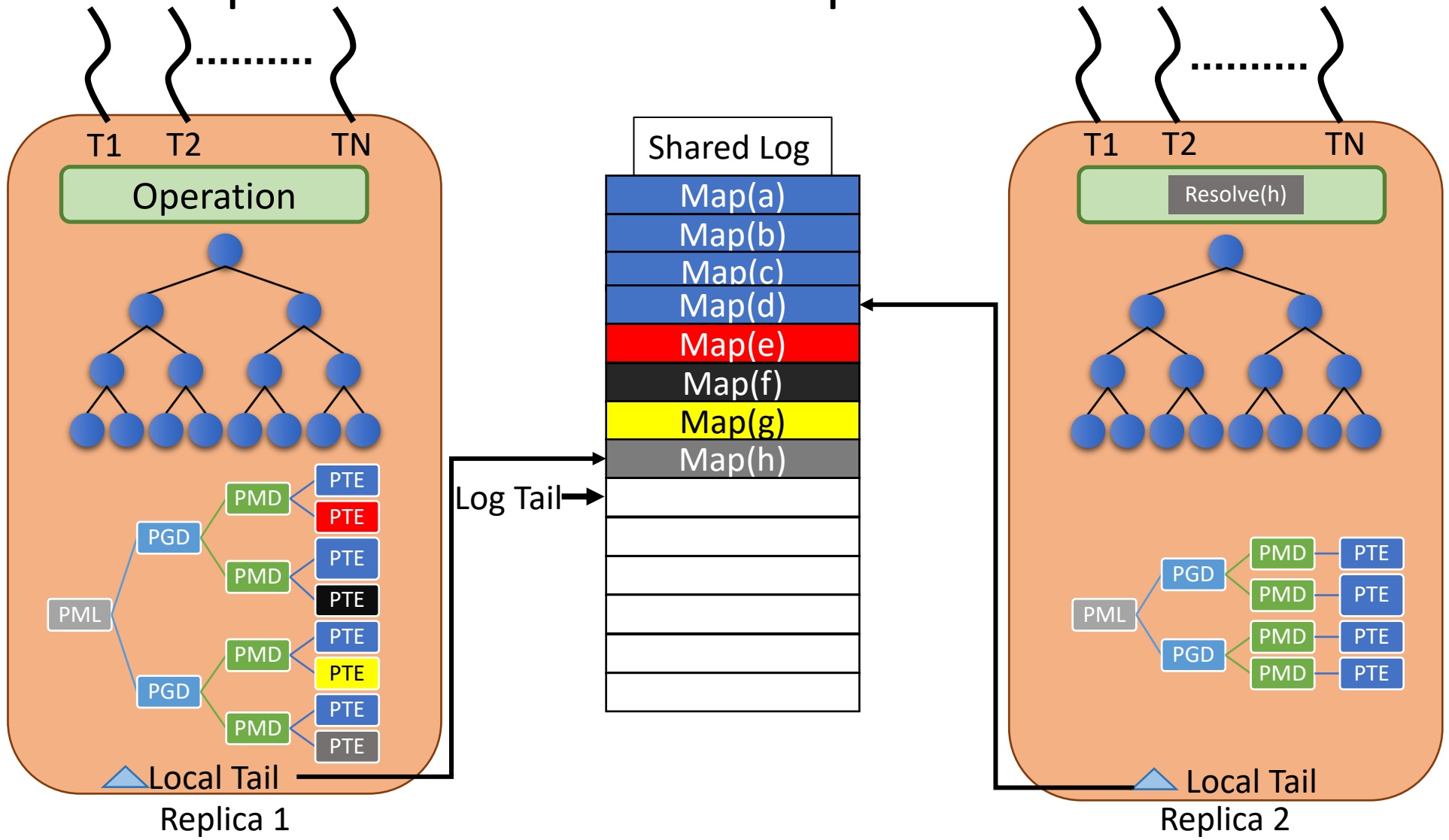
Node Replication: Write Operation



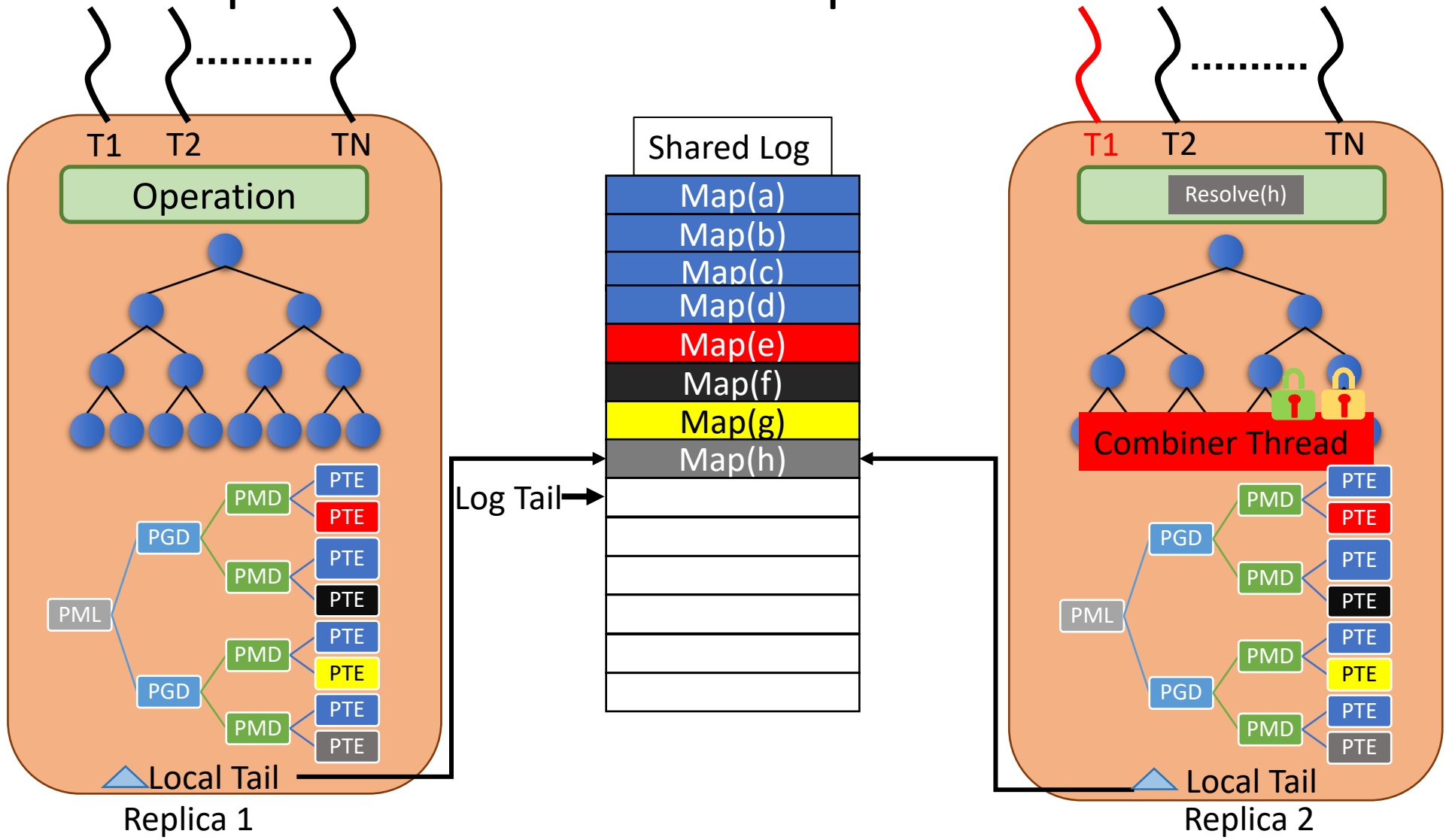
Node Replication: Write Operation



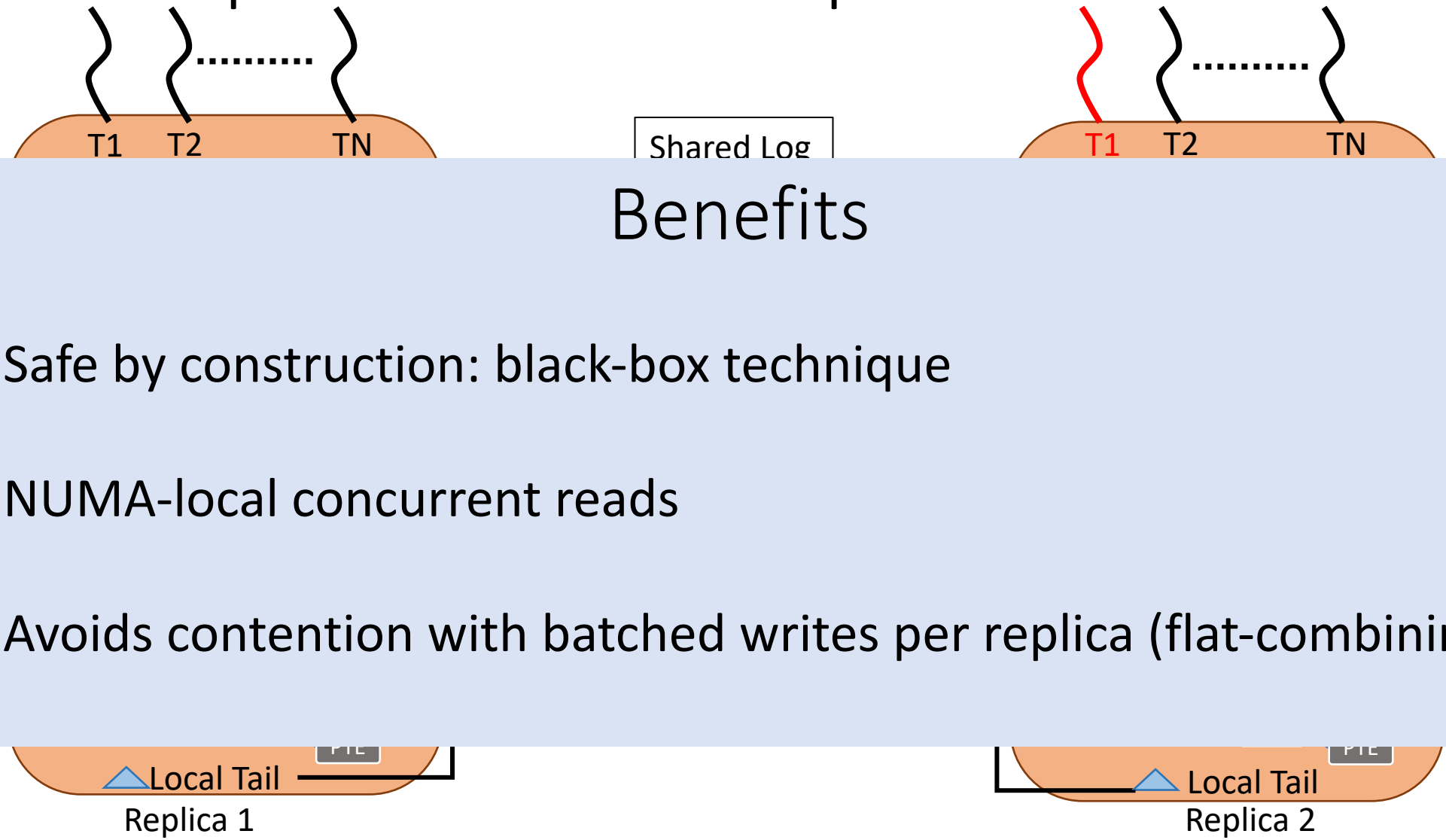
Node Replication: Read Operation



Node Replication: Read Operation



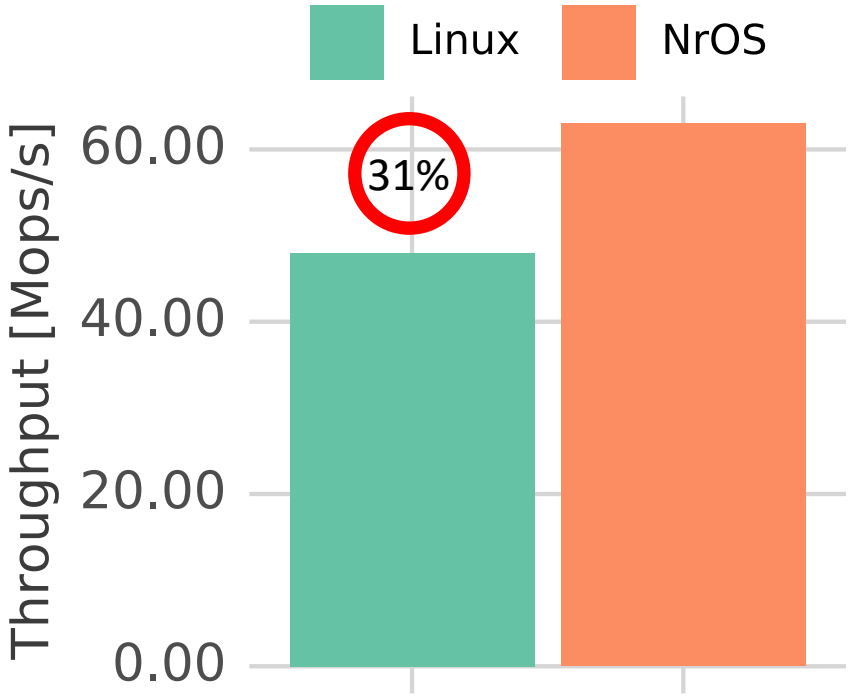
Node Replication: Read Operation



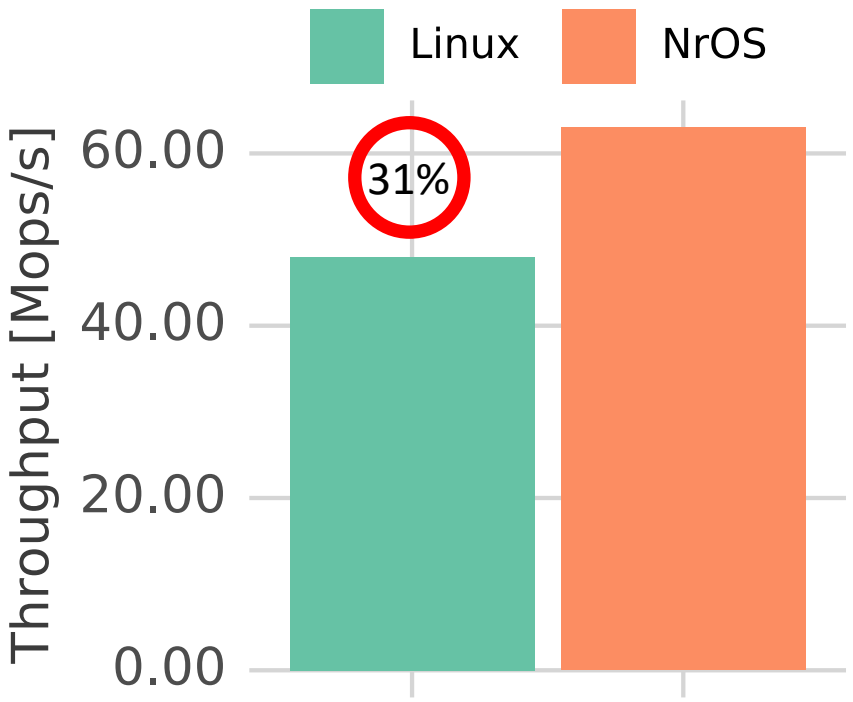
Benefits

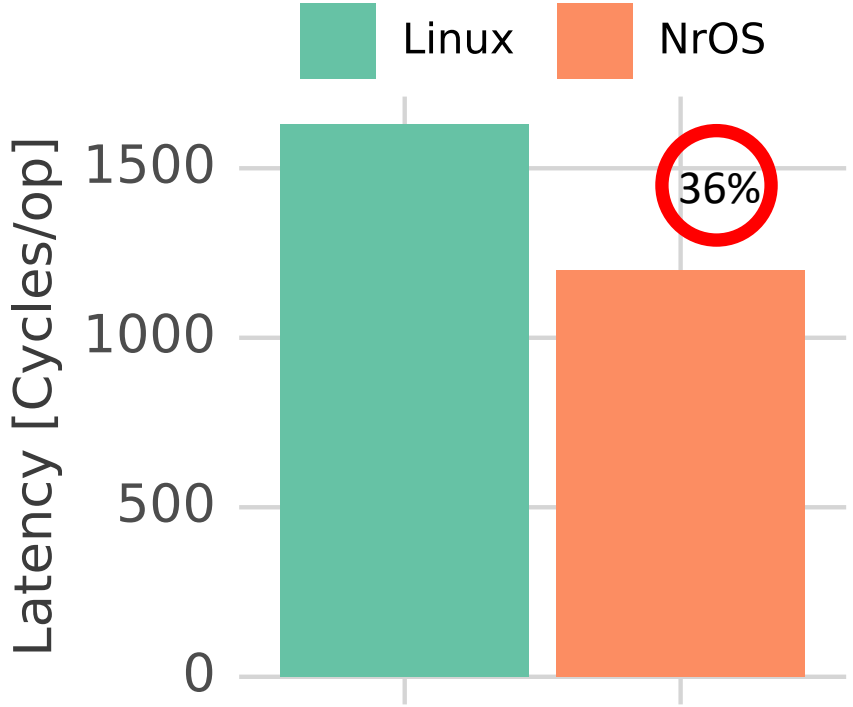
- Safe by construction: black-box technique
- NUMA-local concurrent reads
- Avoids contention with batched writes per replica (flat-combining)

Memcached: Page-table replication benefits

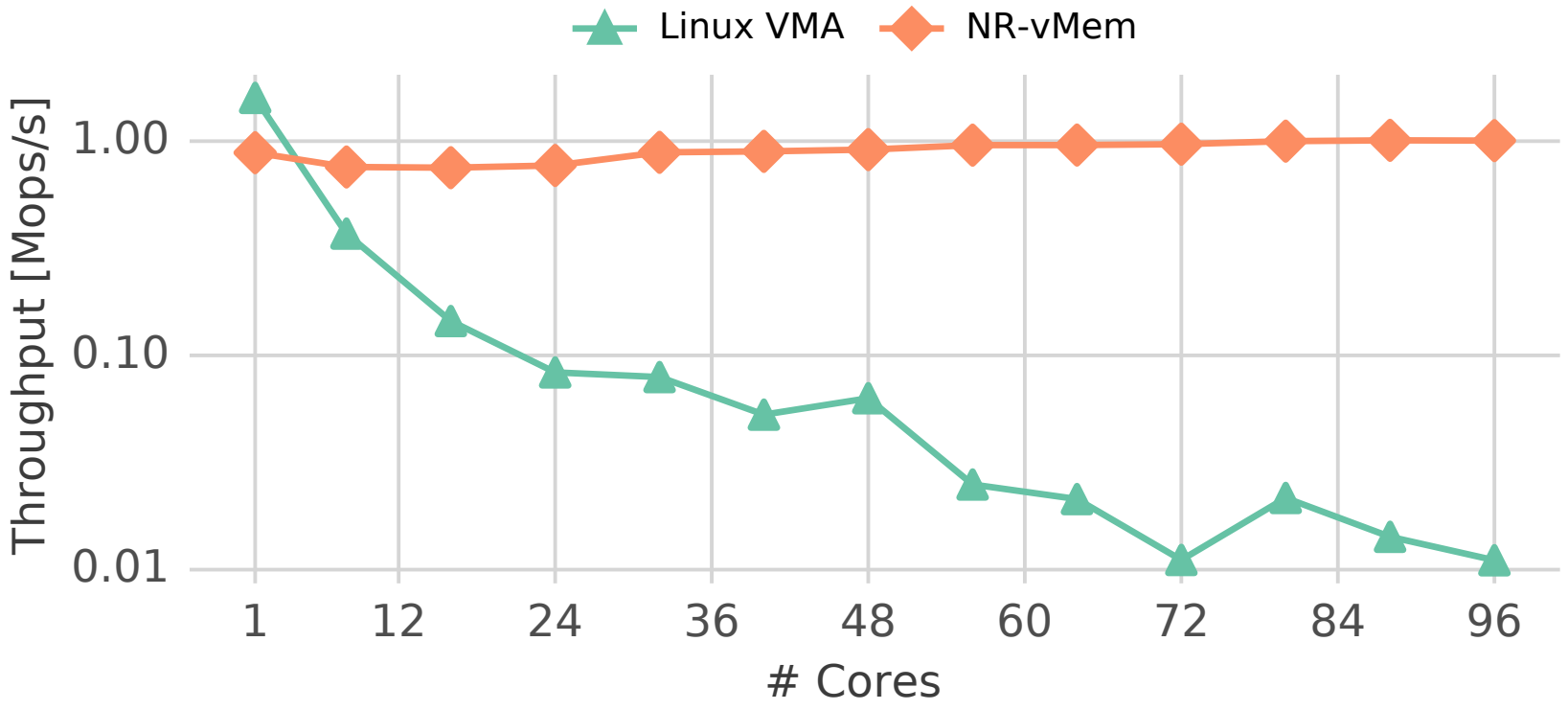


Memcached: Page-table replication benefits

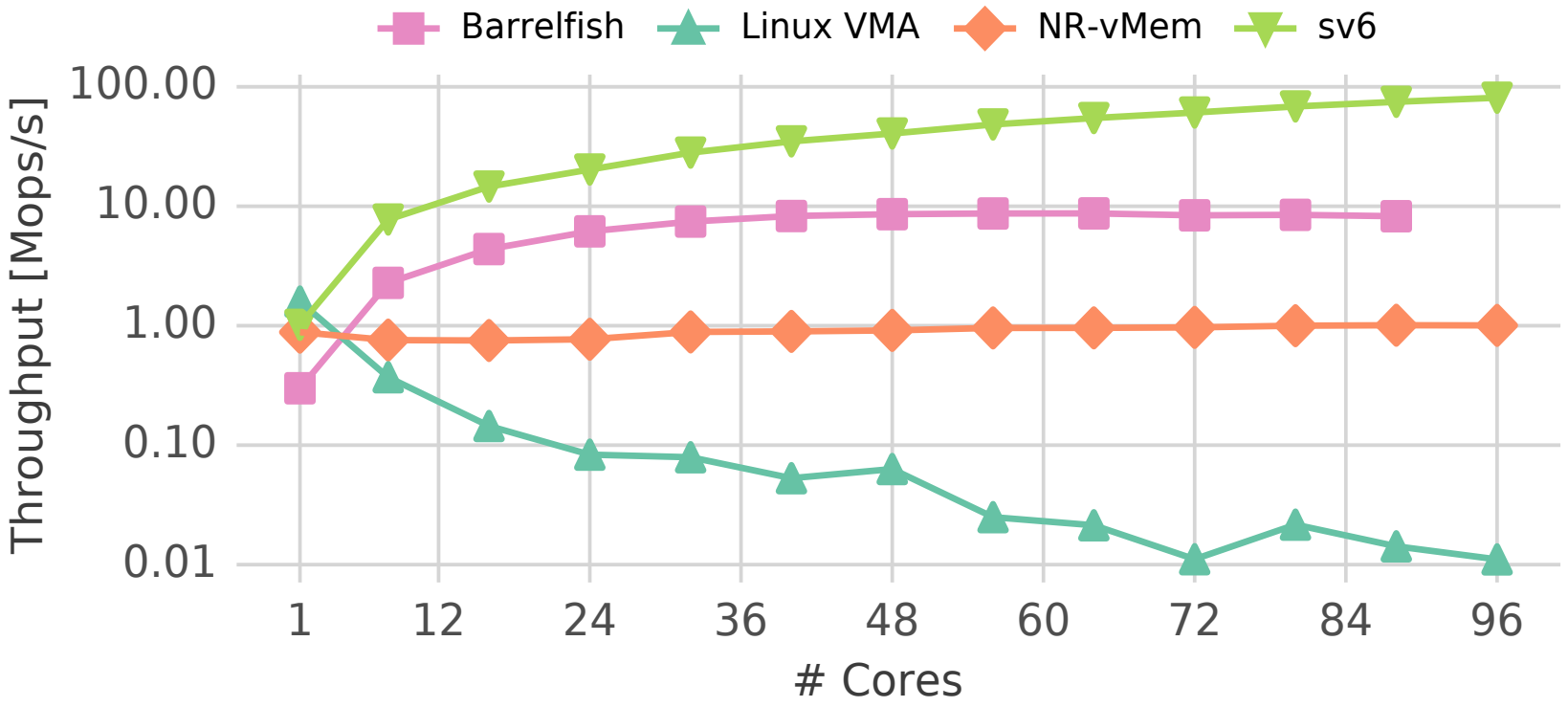




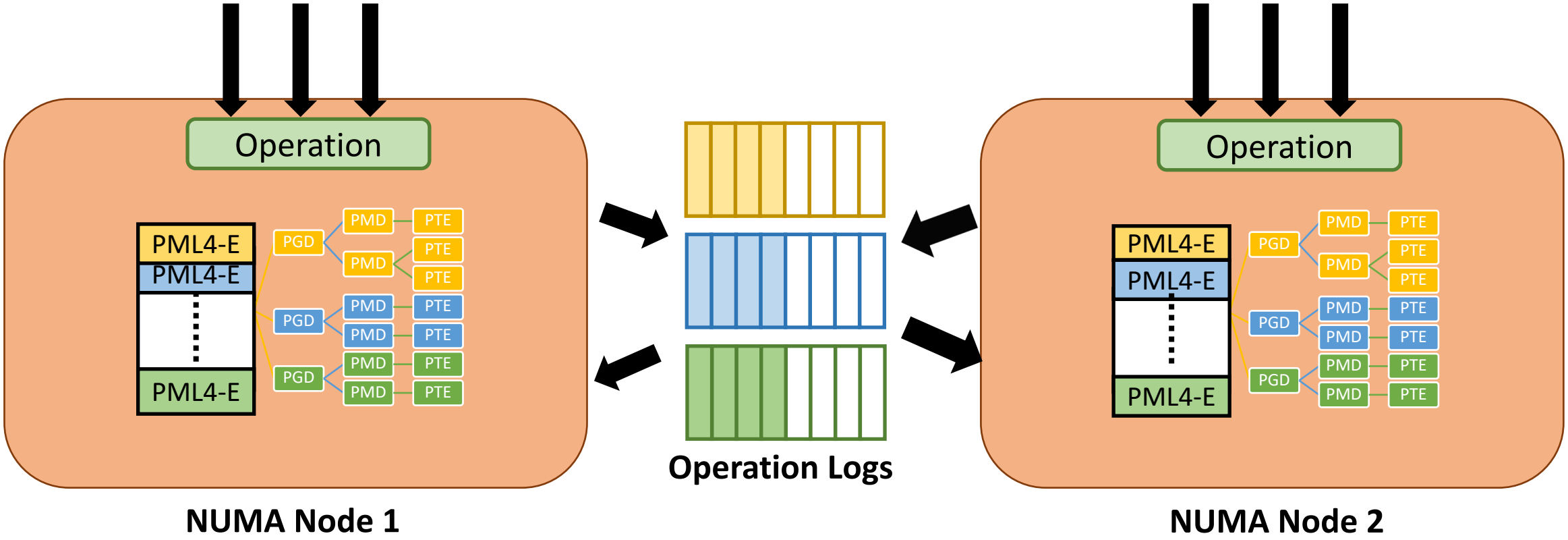
NR-VMem: Combining helps



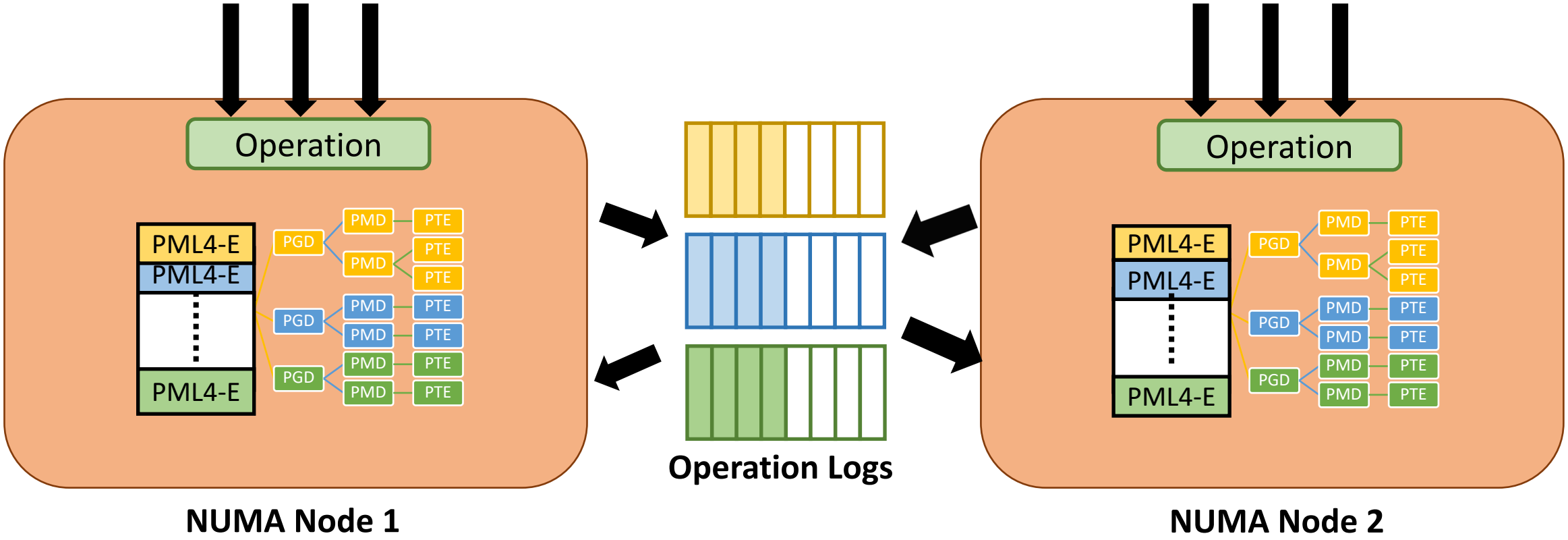
NR-VMem: Combining helps



Concurrent Node Replication (CNR)

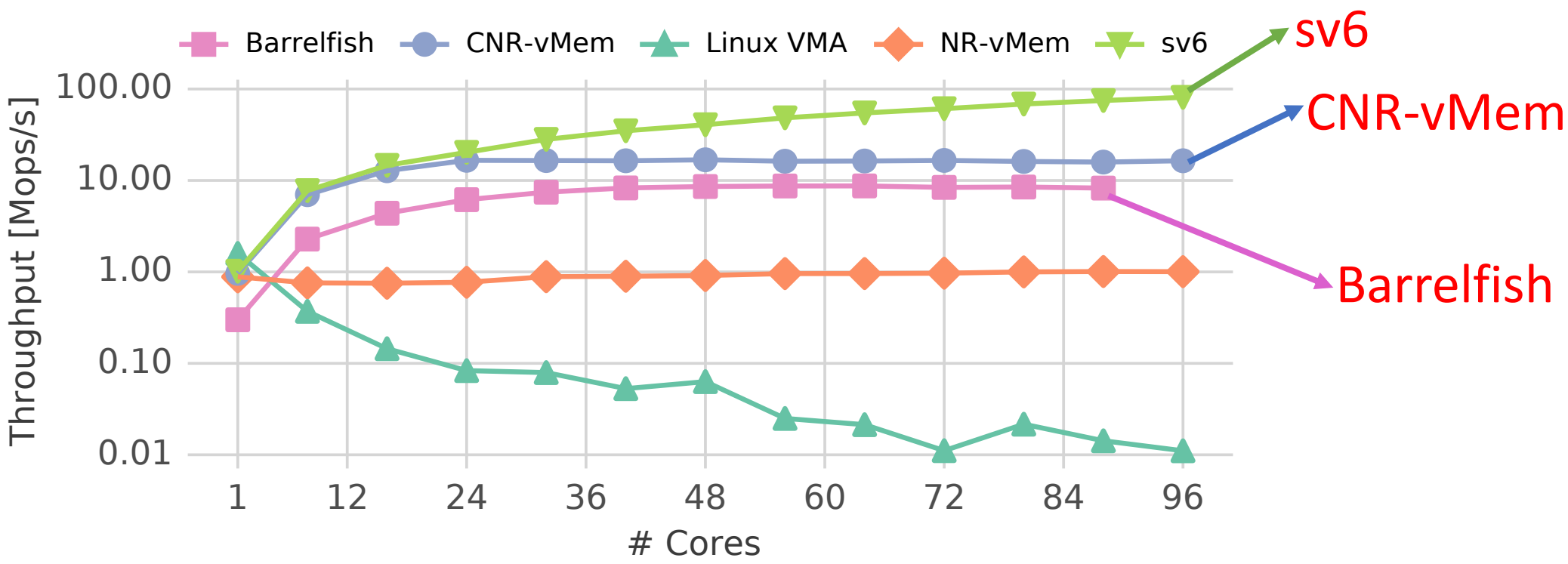


Concurrent Node Replication (CNR)

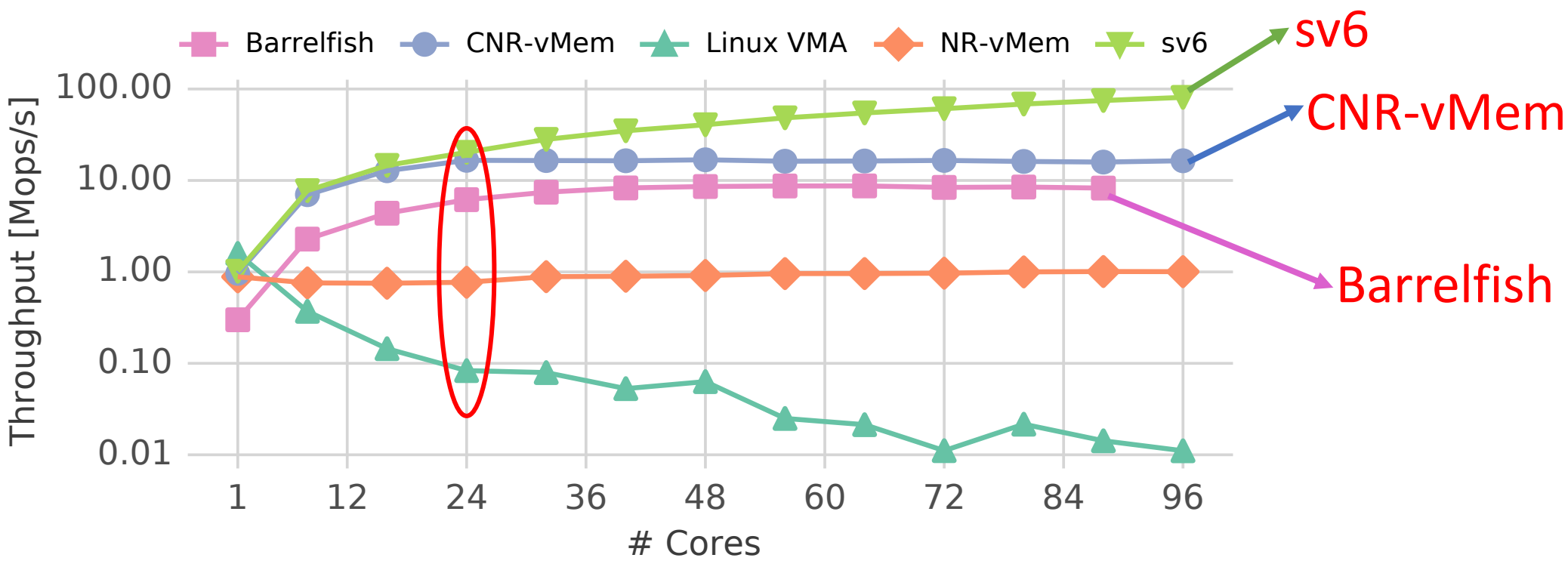


- Commutative operations can map to different logs
- Conflicting operations map to the same log

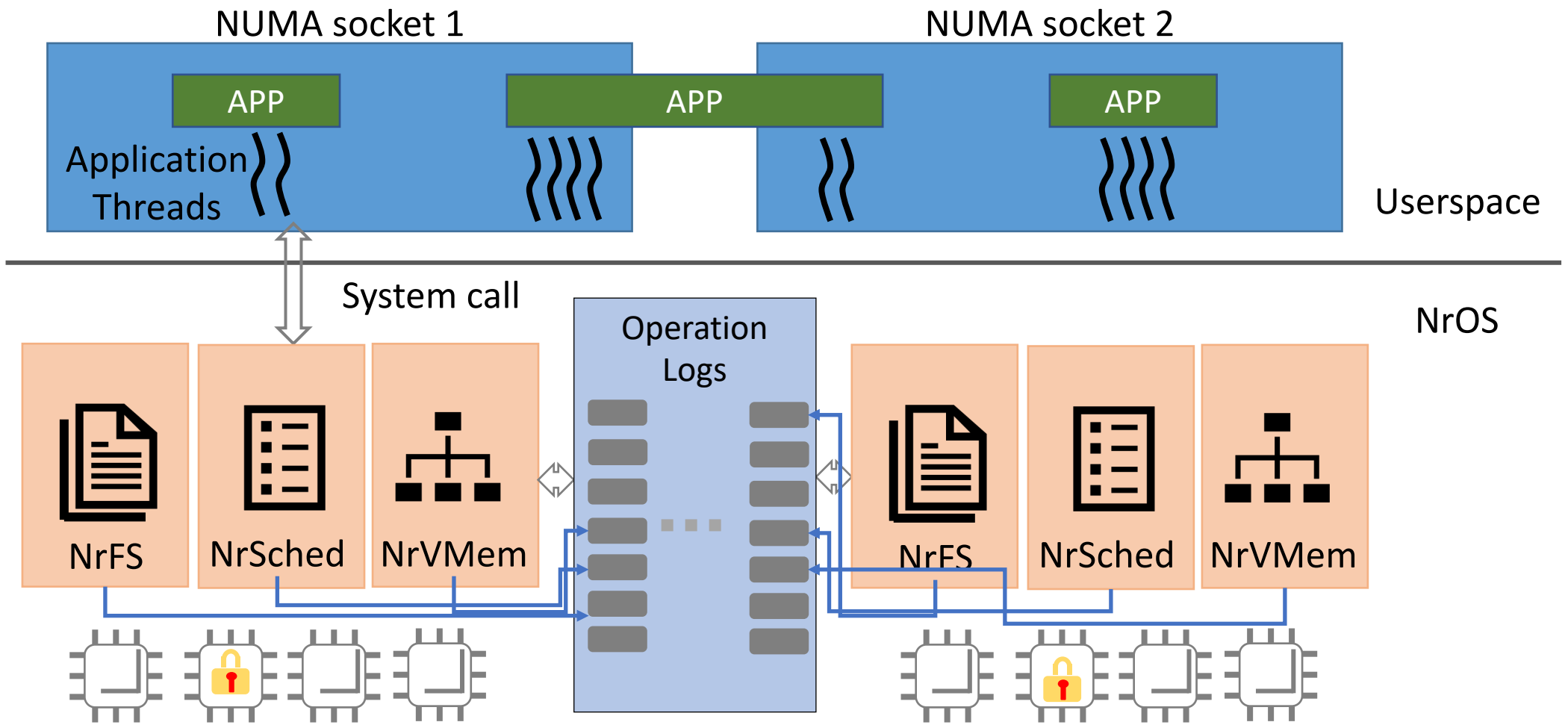
CNR-VMem: Scalable, NUMA-local PageTables



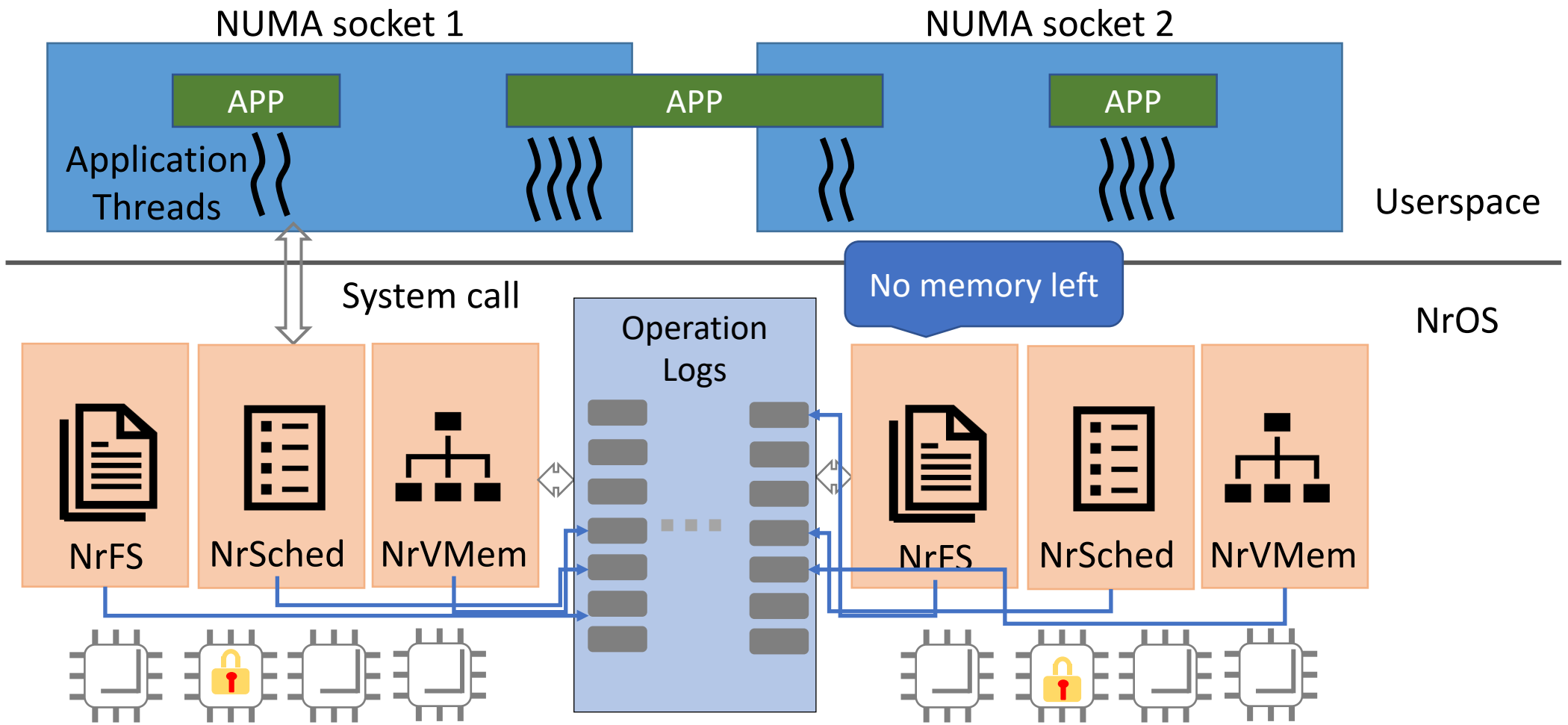
CNR-VMem: Scalable, NUMA-local PageTables



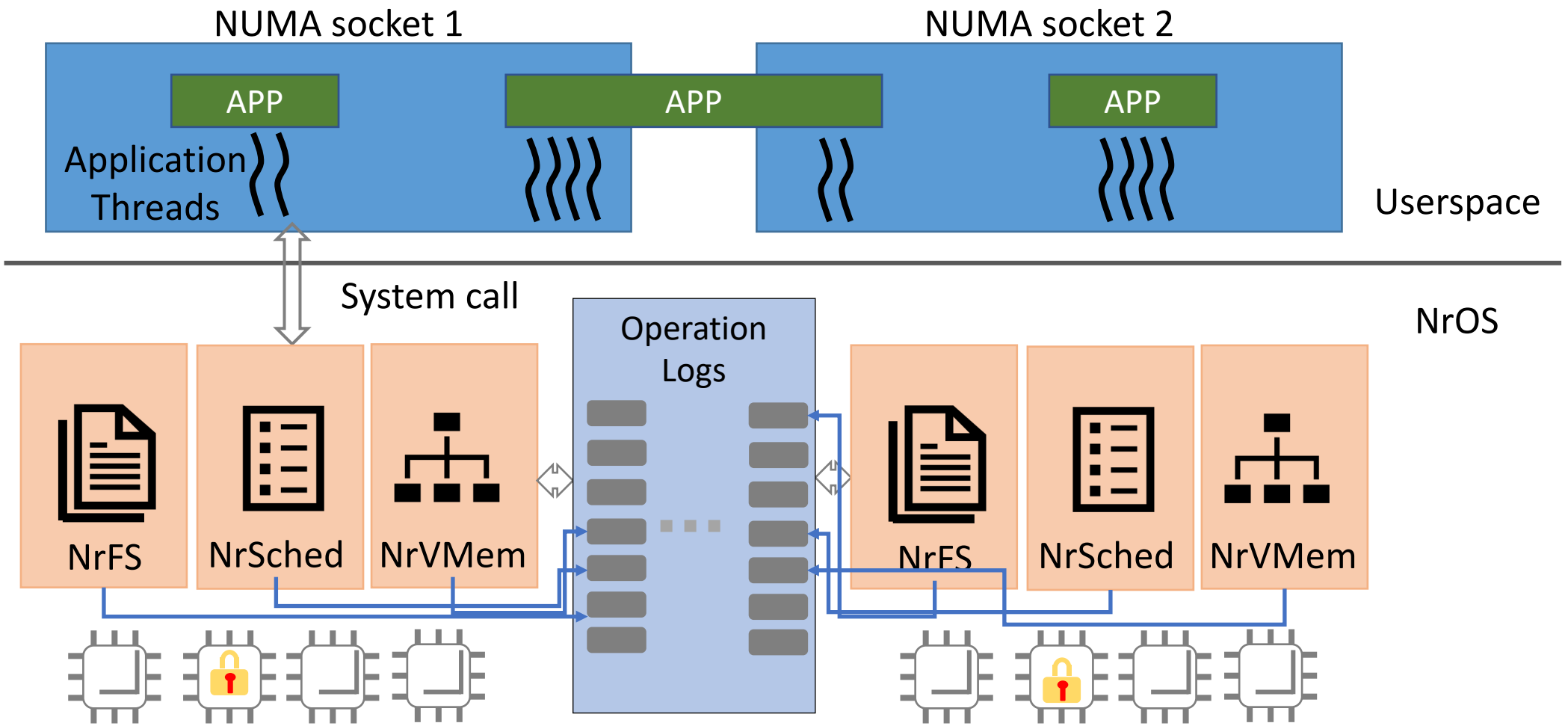
NrOS Summary



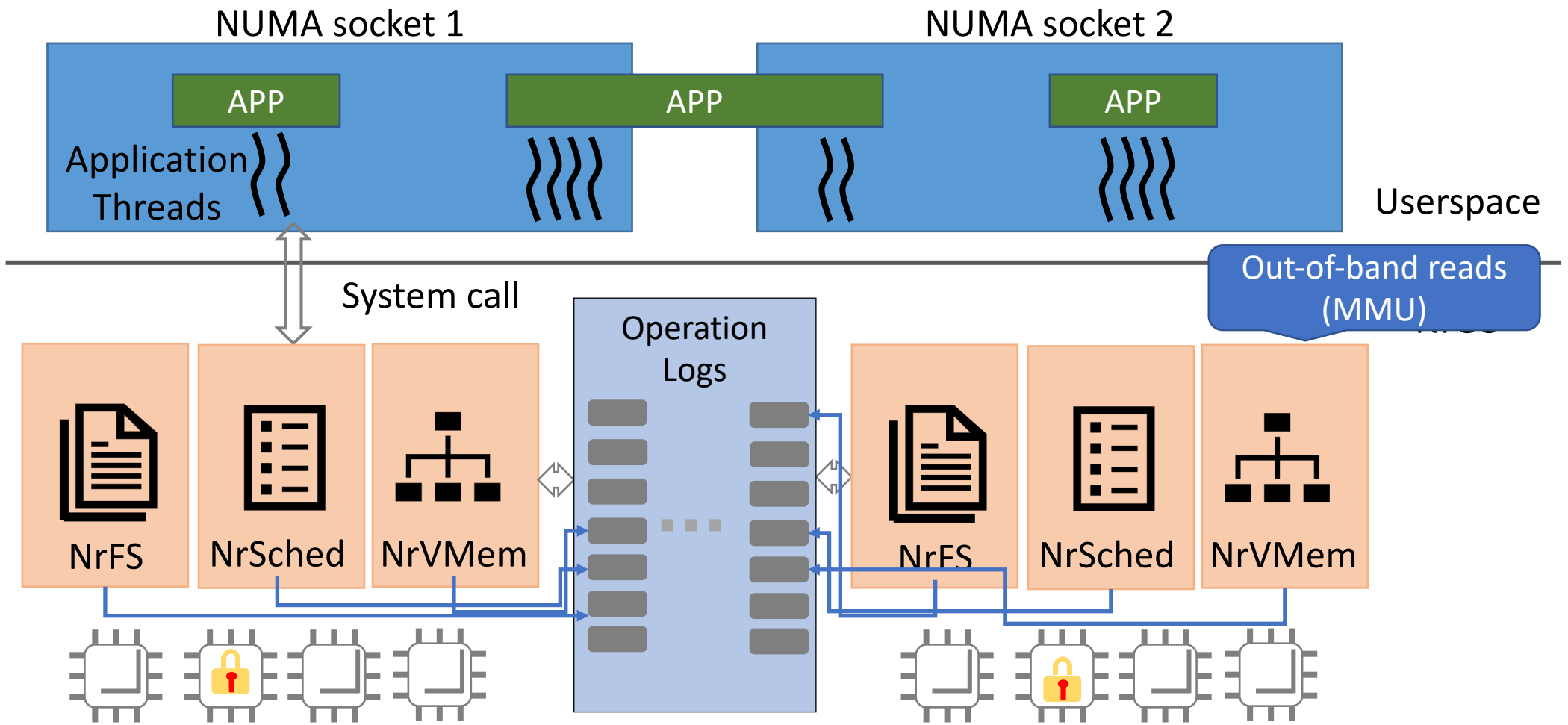
NrOS Summary



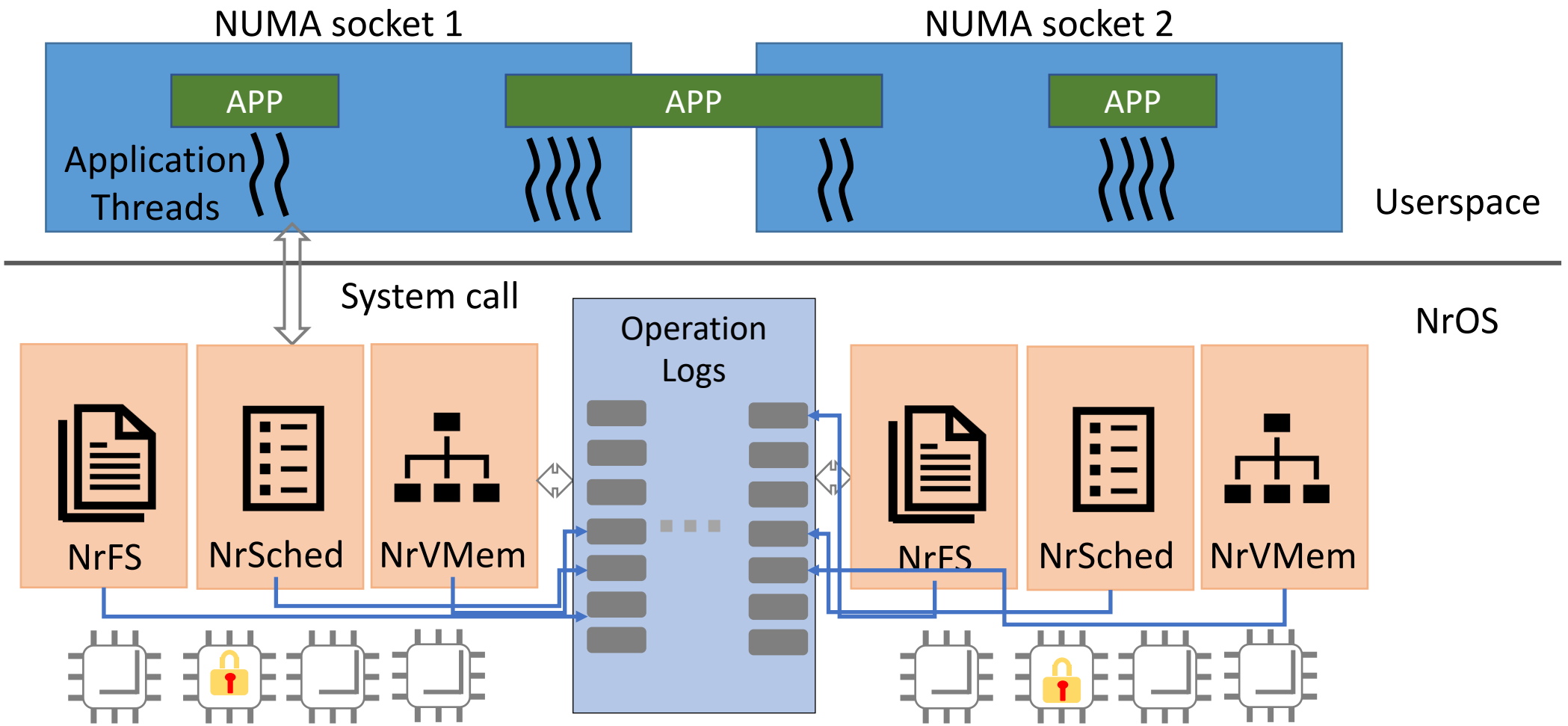
NrOS Summary



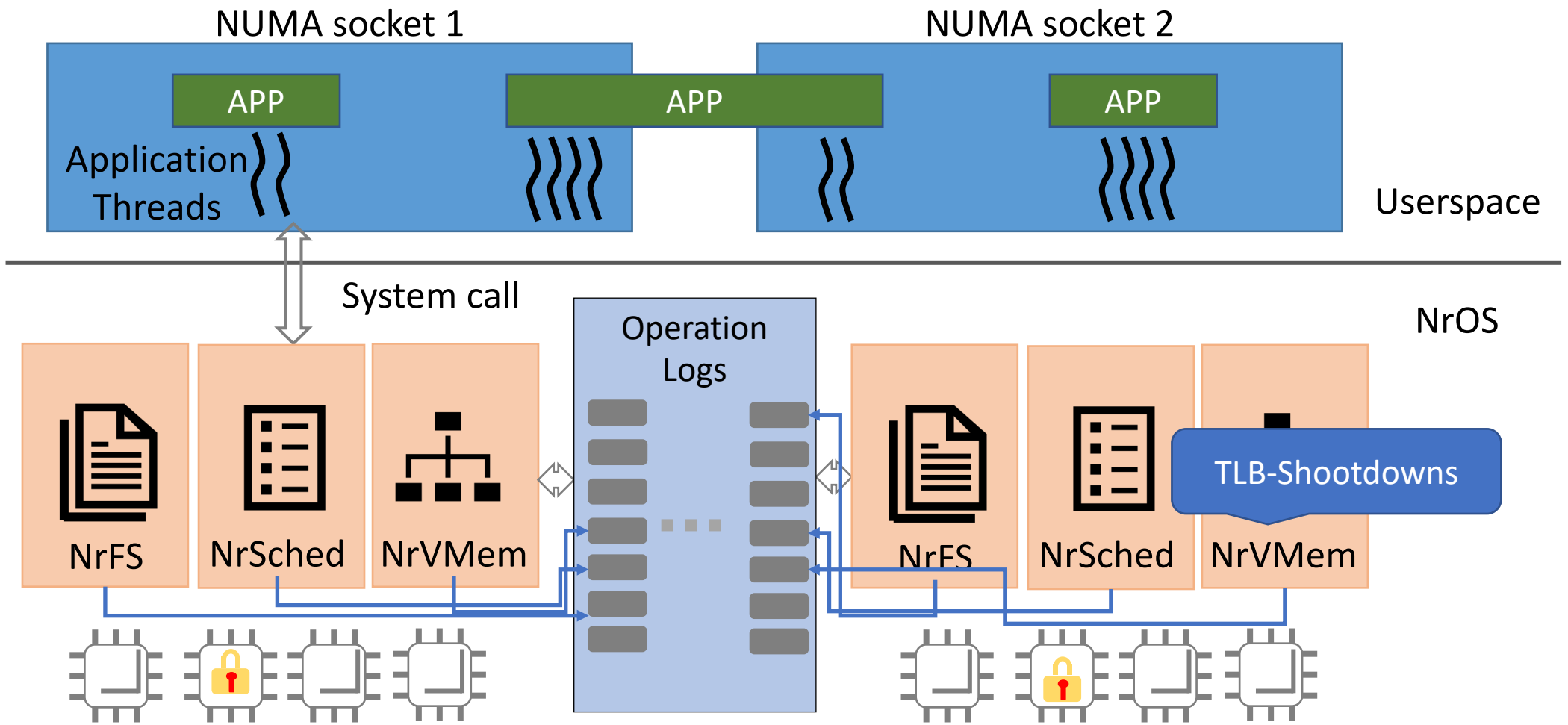
NrOS Summary



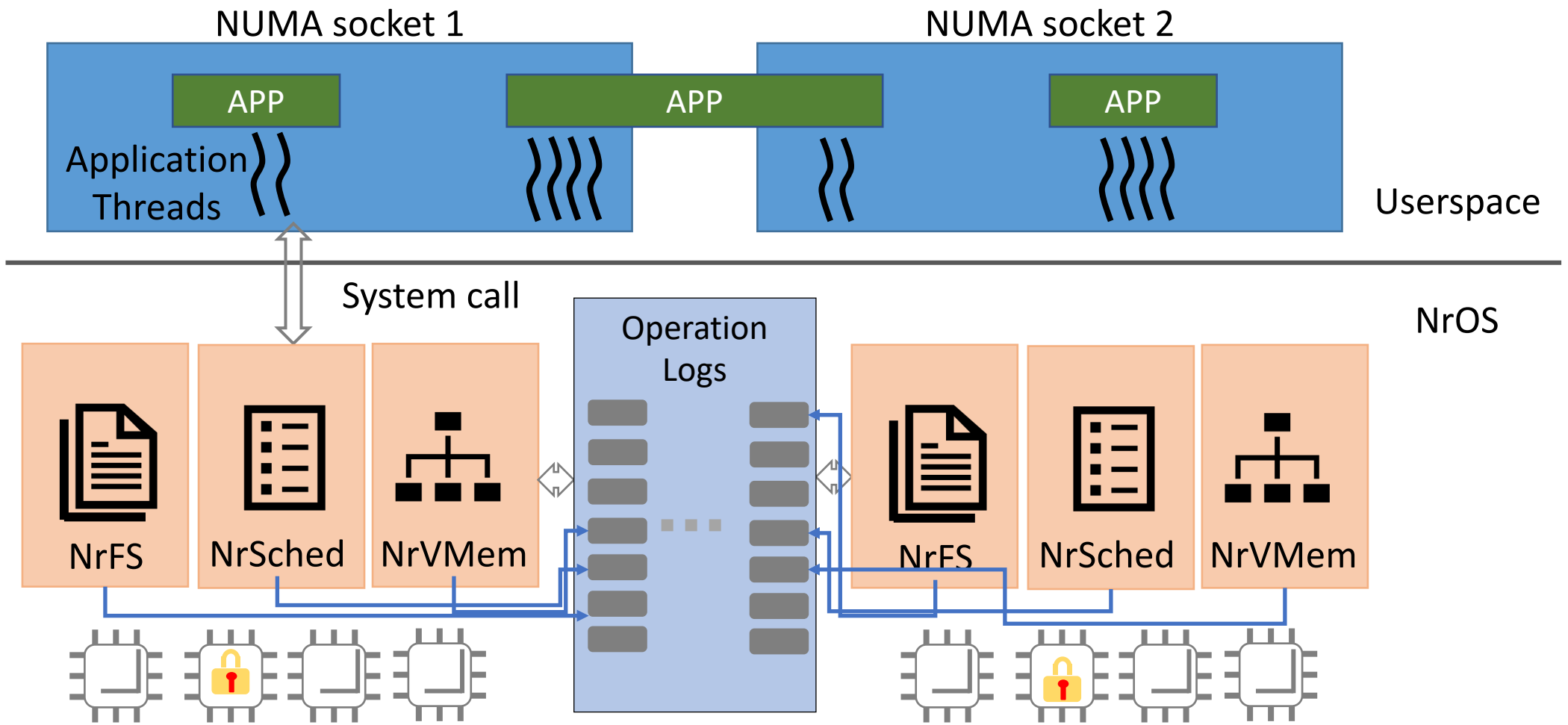
NrOS Summary



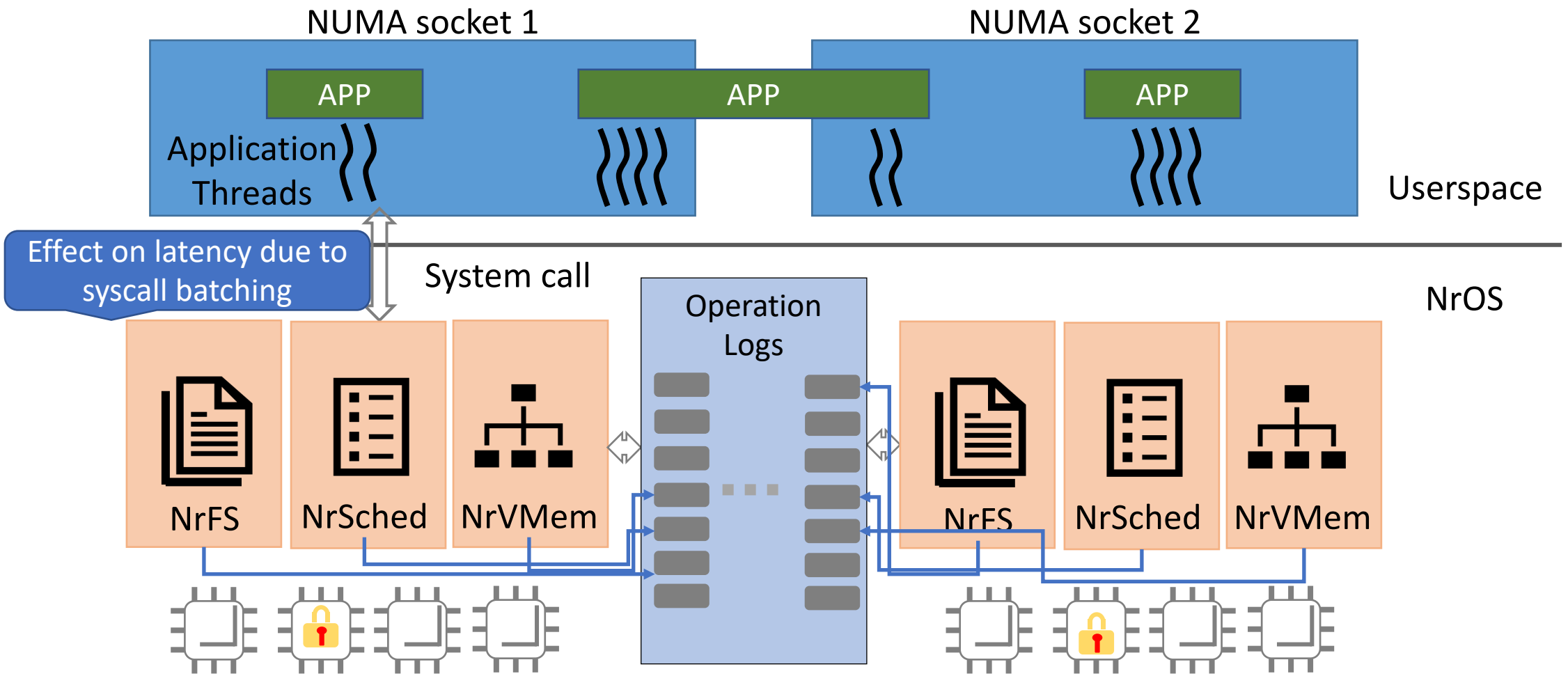
NrOS Summary



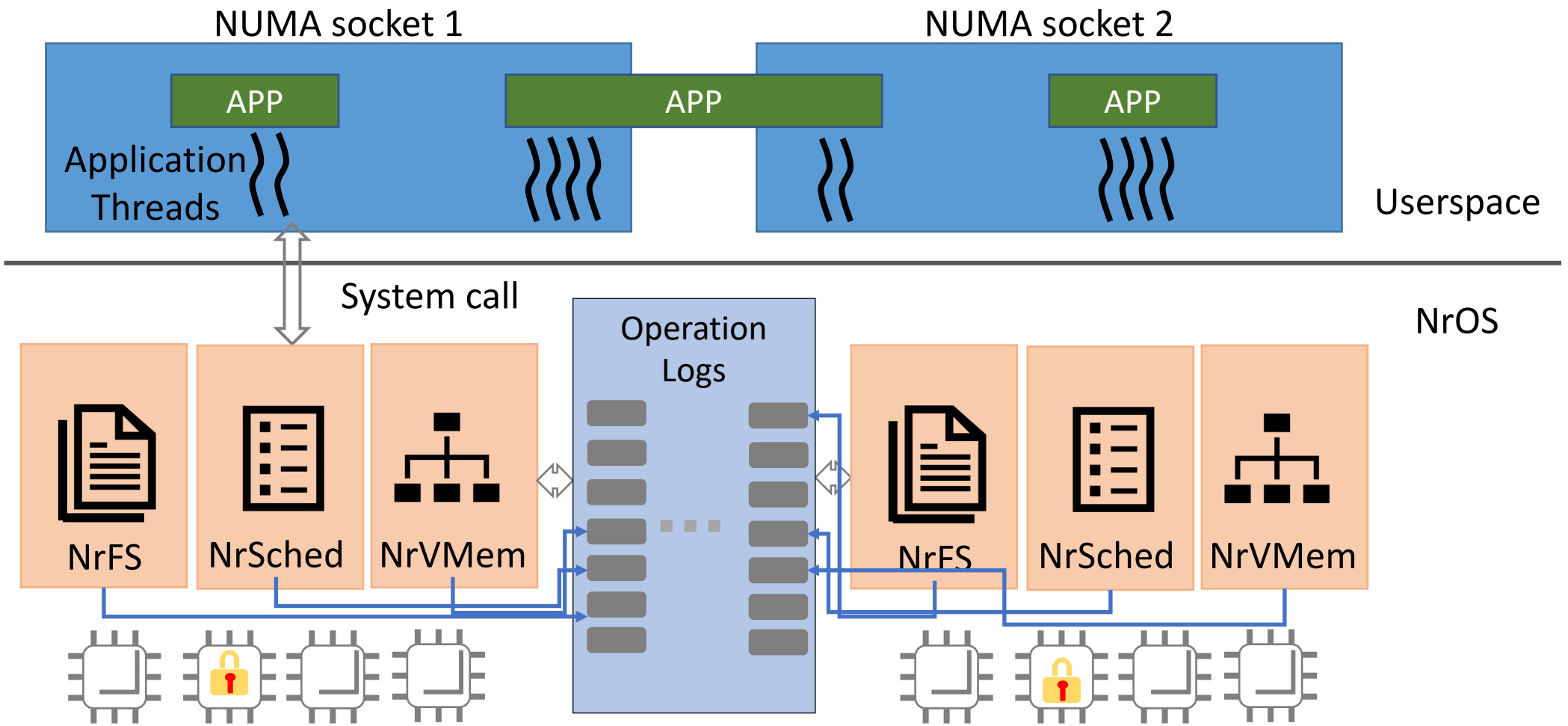
NrOS Summary



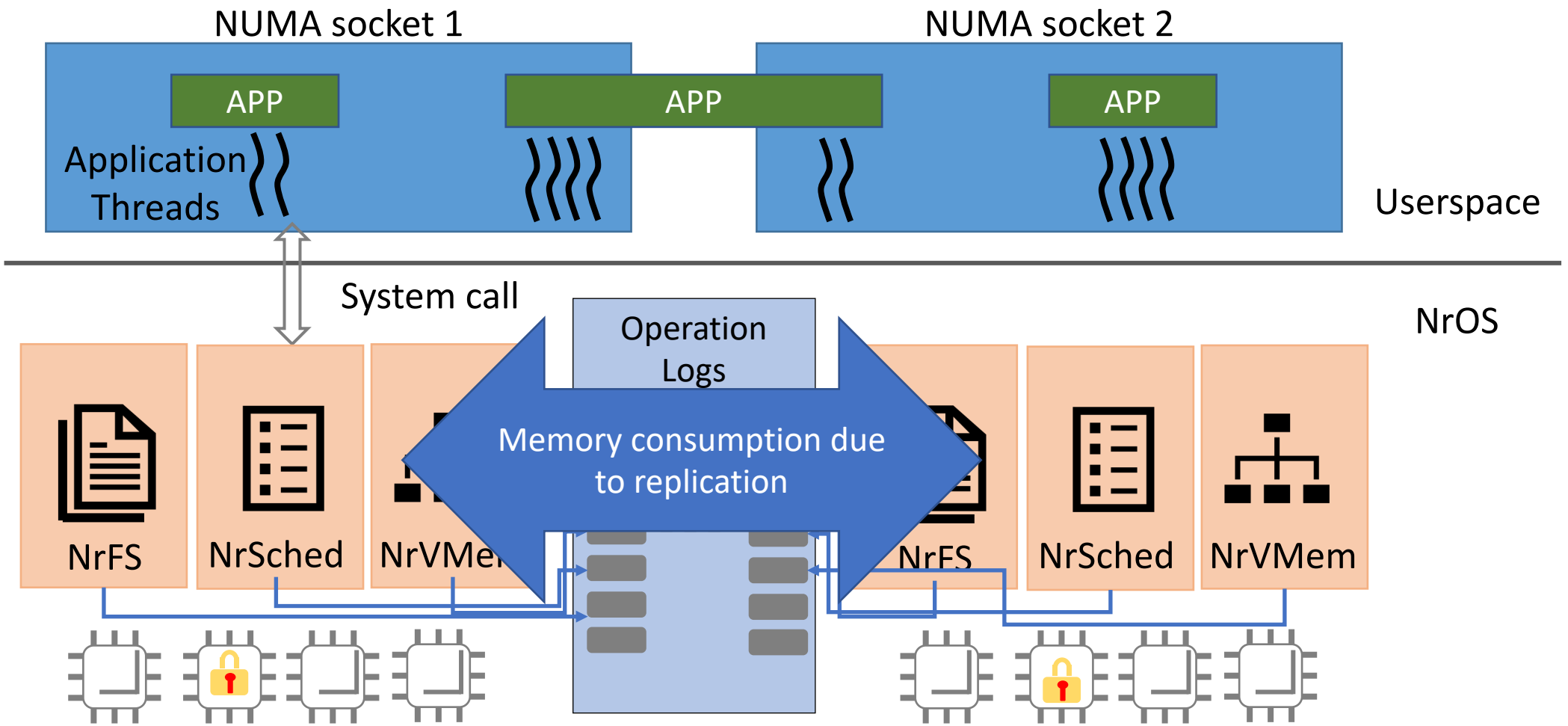
NrOS Summary



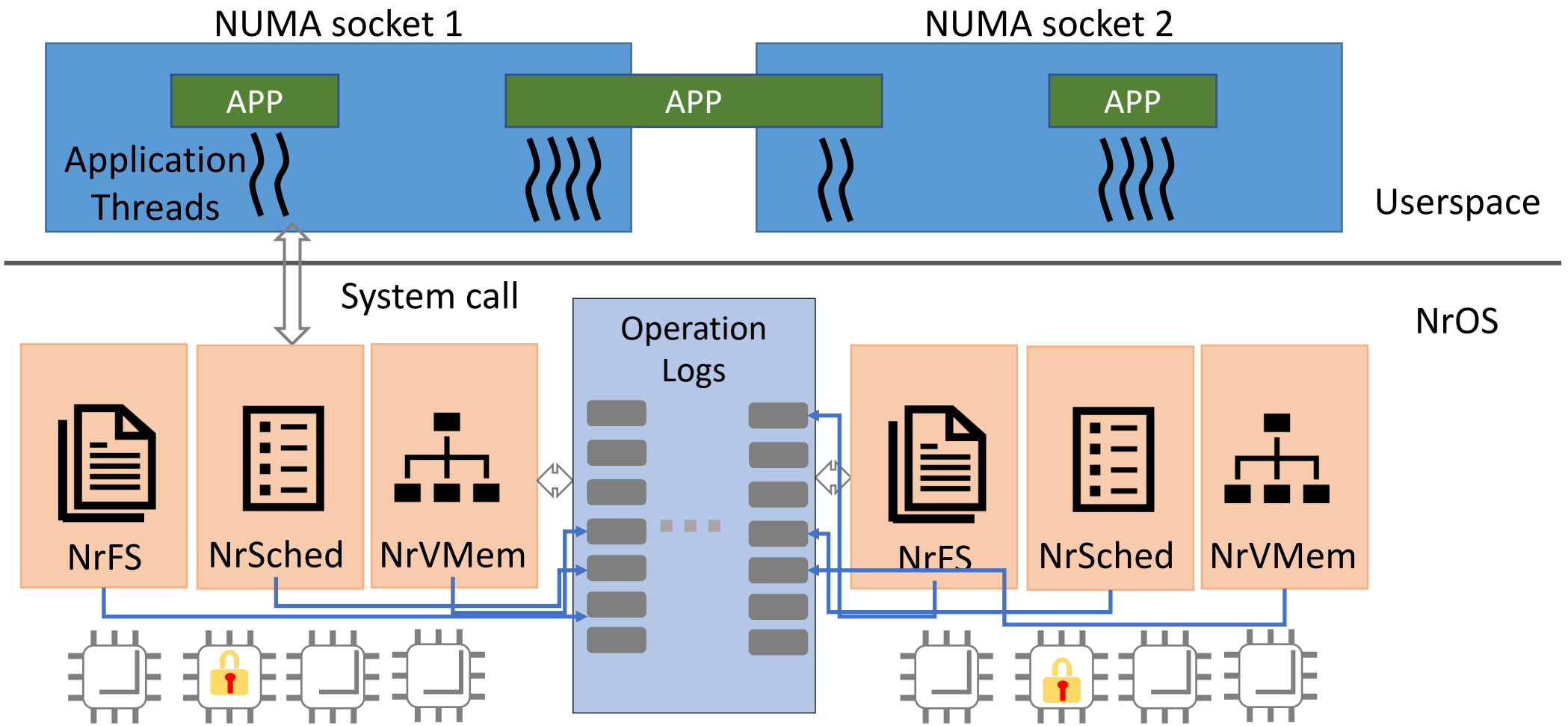
NrOS Summary



NrOS Summary



NrOS Summary



NRkernel: Key takeaways

- Simple, black-box approach (one size fits all) for concurrency
- Decomposes the problem of building scalable kernel data-structures for OS kernels
- Trade-off memory/CPU for performance: Replicate, make all reads NUMA local

https://nrkernel.systems