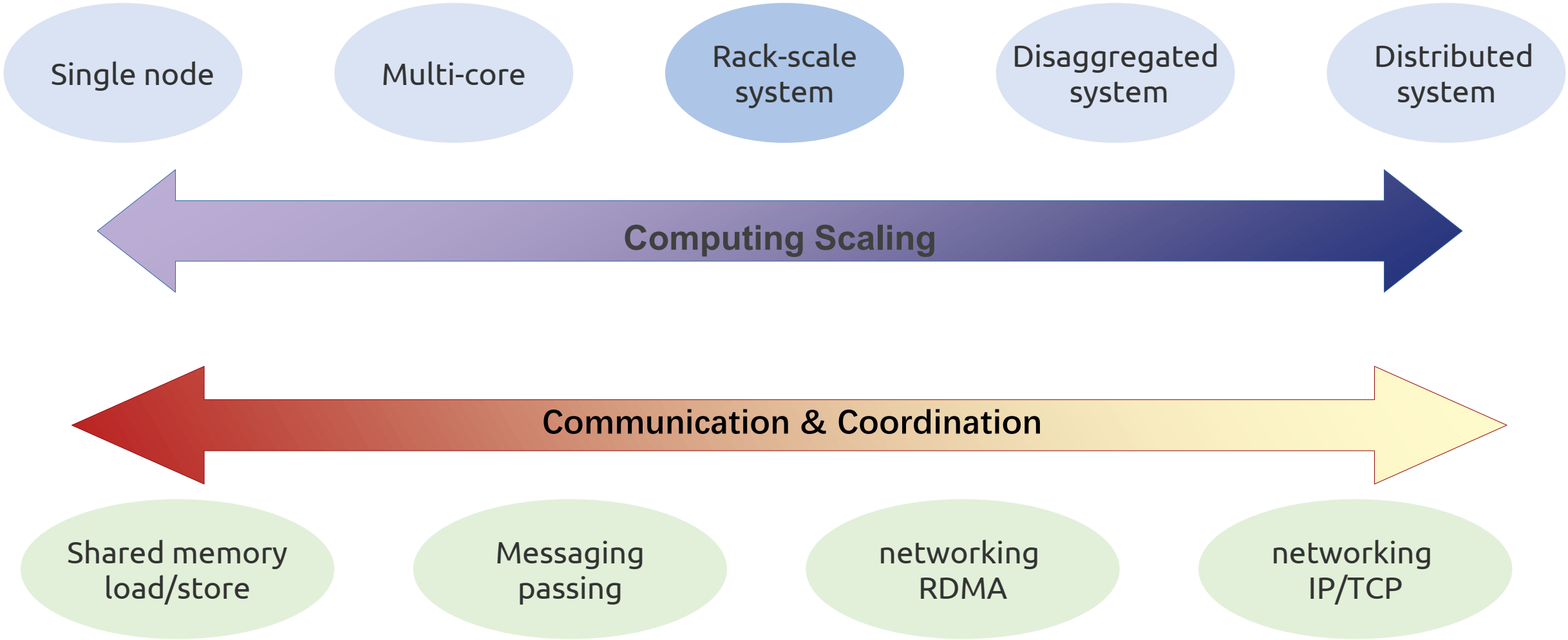
Towards Rack-as-a-Computer in Memory Interconnect Era with Coordinated Operating System Sharing

Yuxin Ren, Mingrui Liu, Hongbo Li, Chang Liao, Xiaojia Huang,
Jianhua Zhang, Hanjun Guo, Yubo Liu, Ning Jia

Huawei

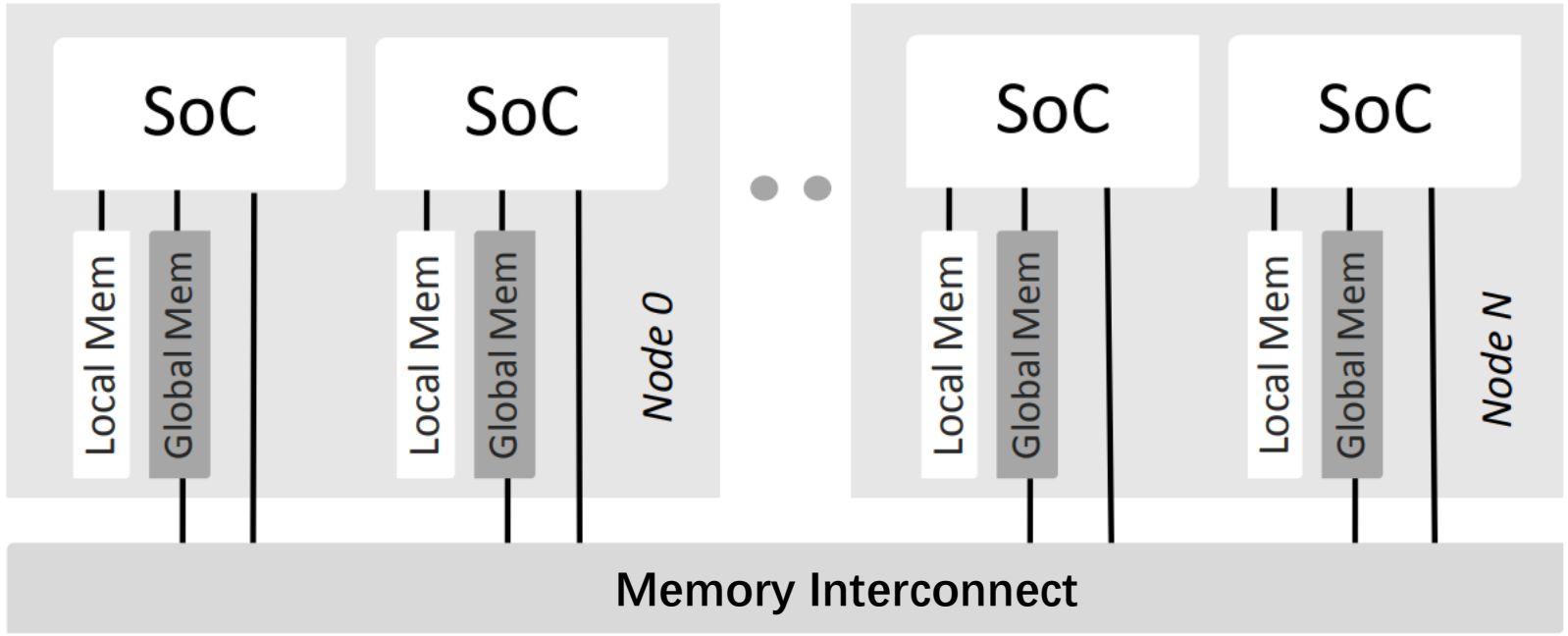
开放原子开源基金会 OPENATOM FOUNDATION | OpenEuler

Ever-Increasing Scale

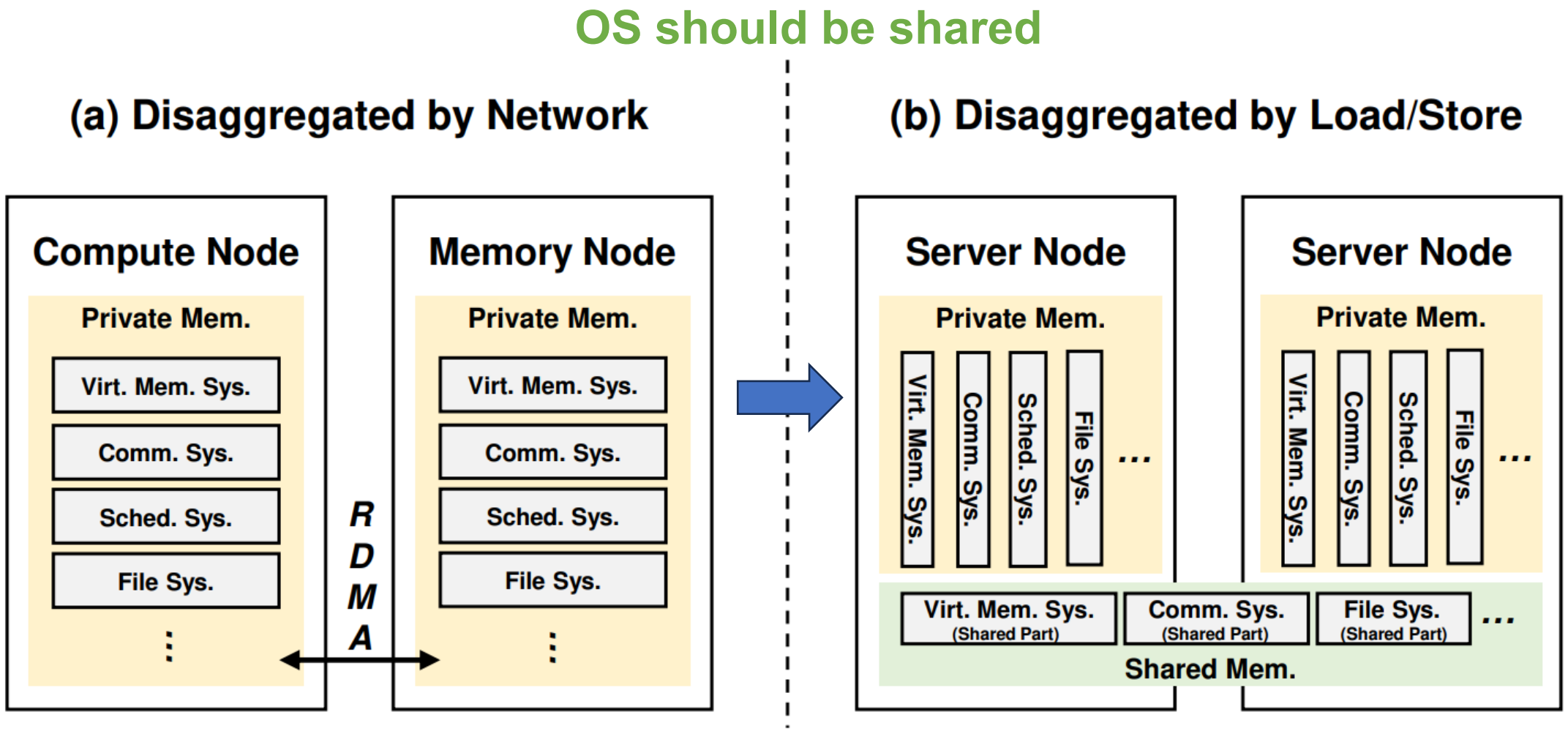


Rack-scale machine: memory interconnect

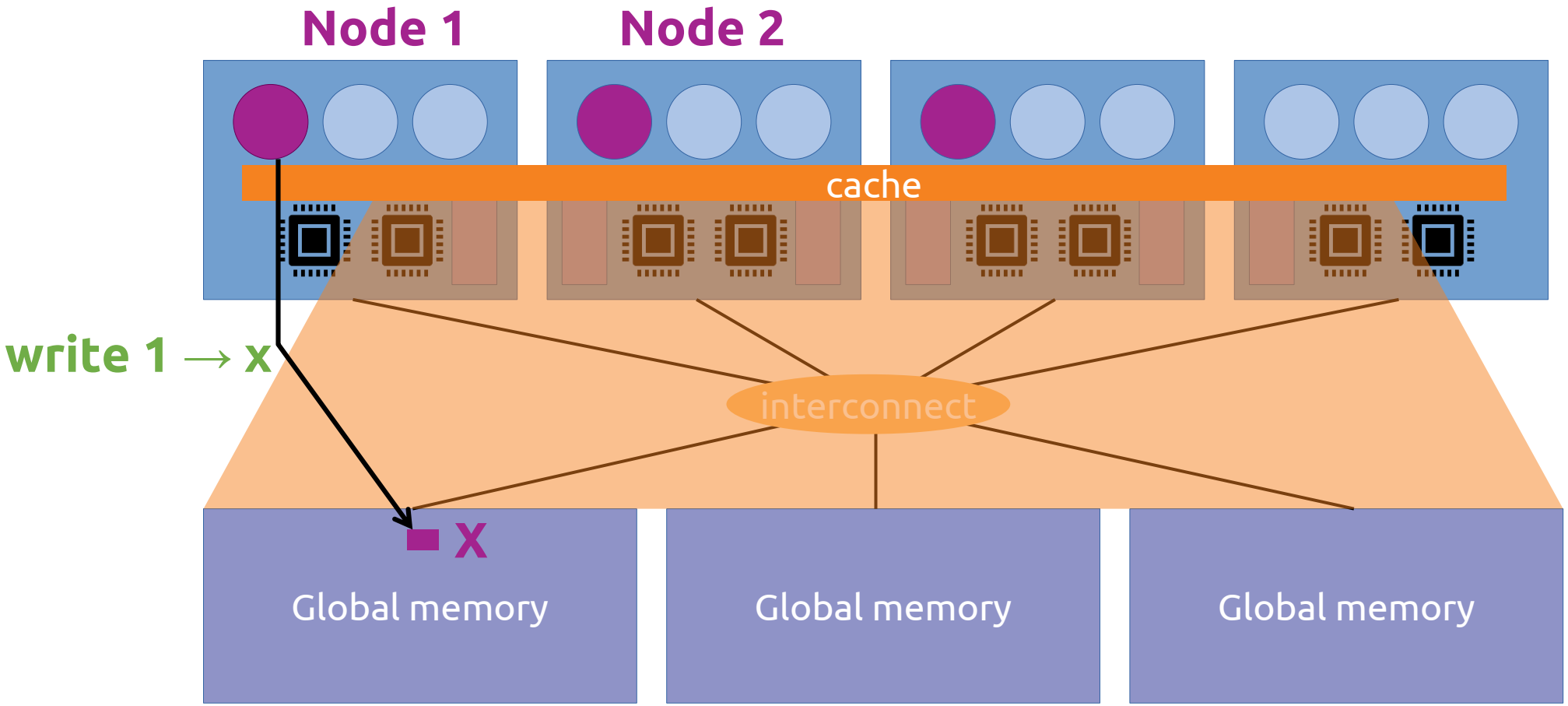
- **Rack-scale**: all nodes are interconnected with *load/store* *to access memory*
  - Each node access both local and global memory
  - Petabytes of byte-addressable memory shared across thousands of cores.



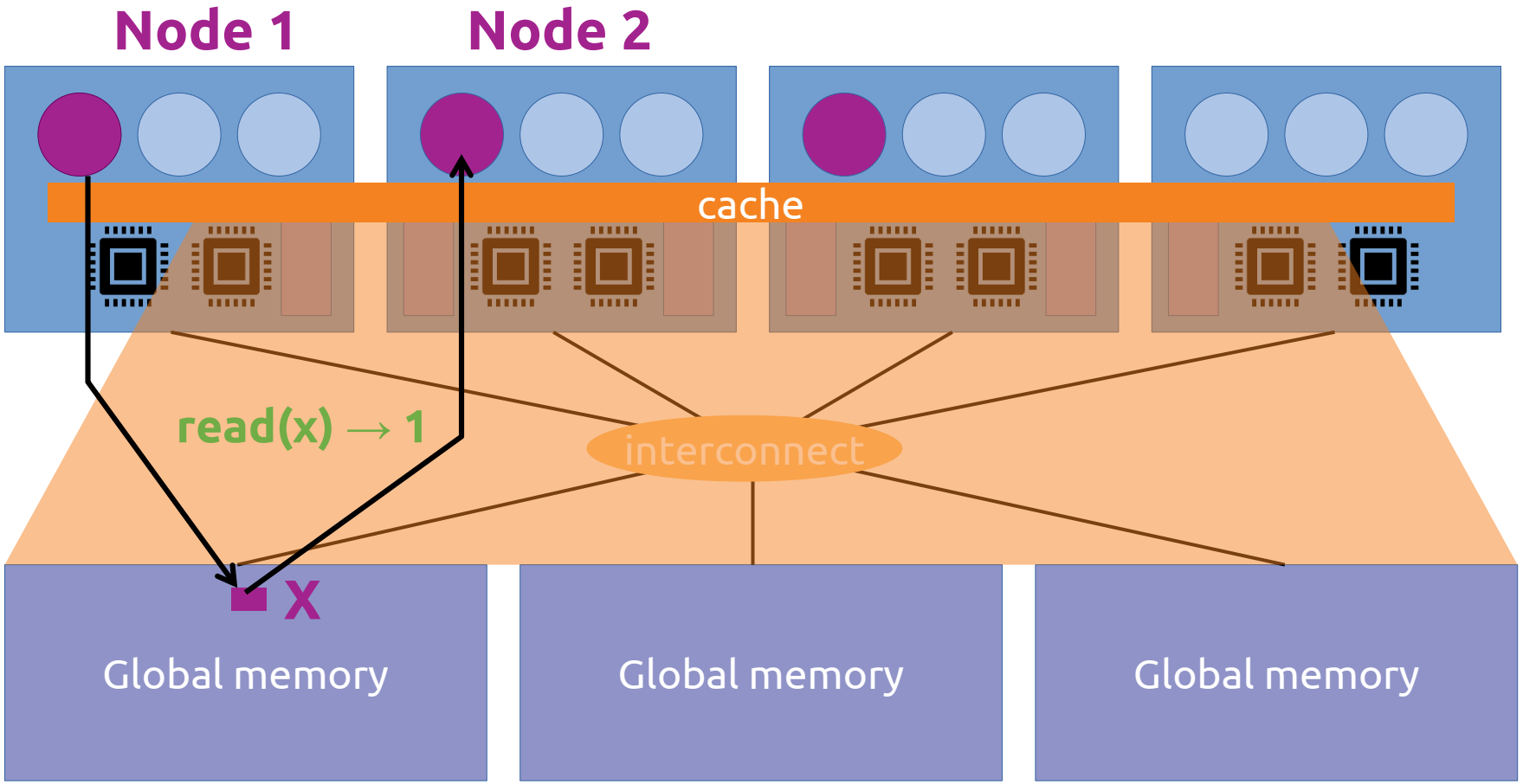
Rack-scale machine: sharing capability



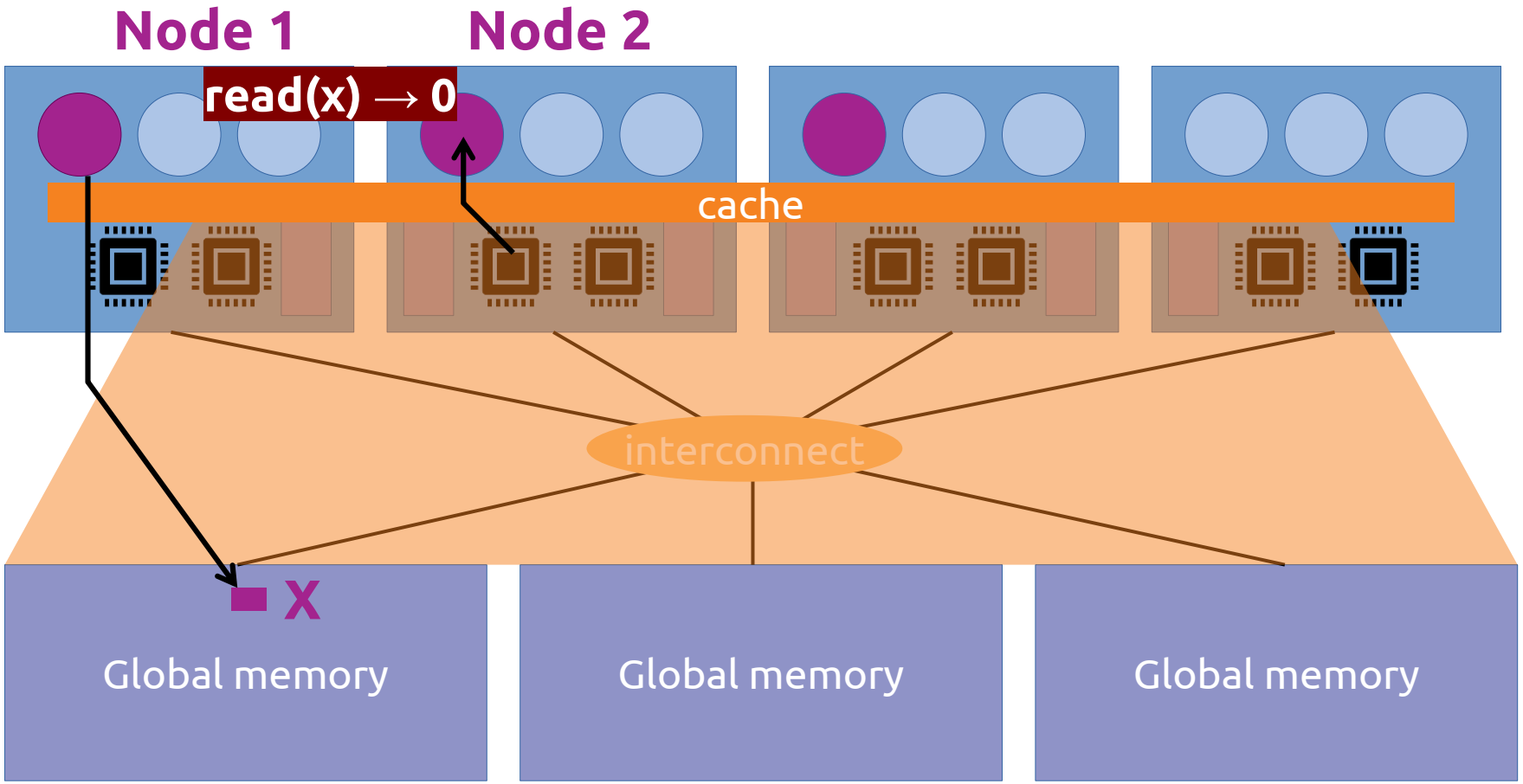
Challenge: non-cache-coherence



Challenge: non-cache-coherence



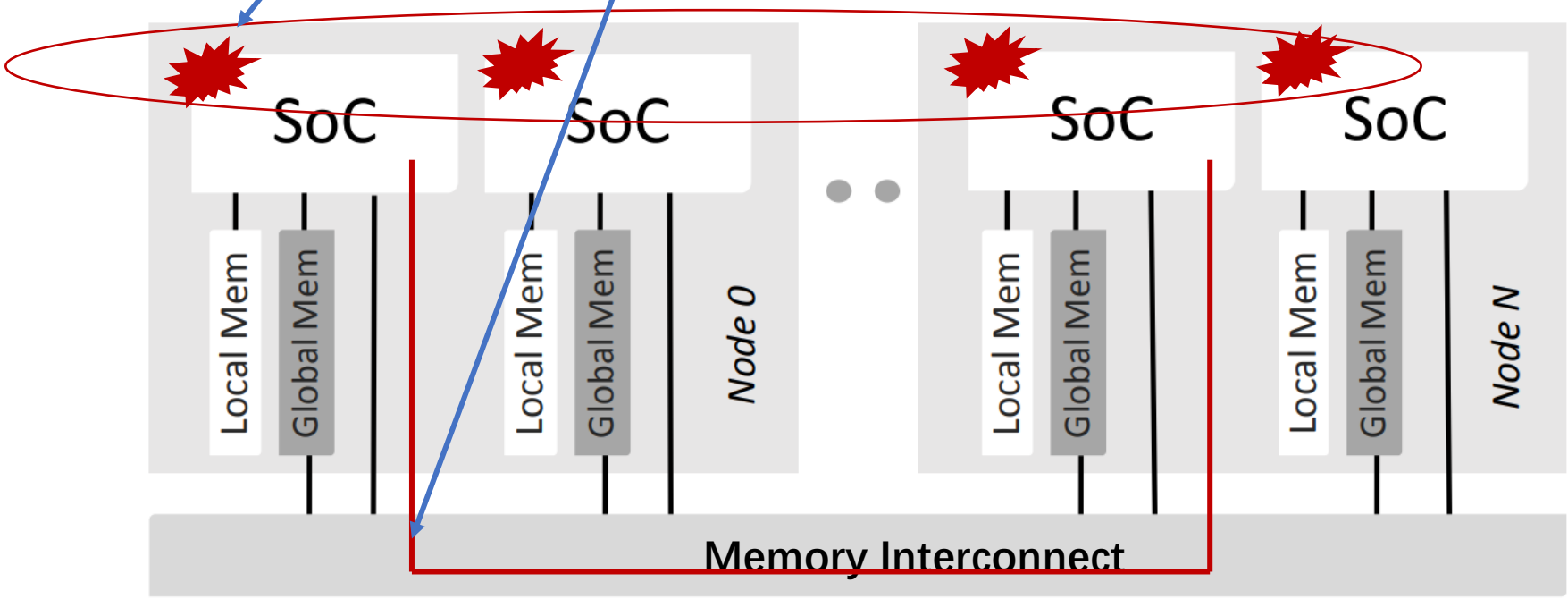
Challenge: non-cache-coherence



- Dangling references
- null pointer
- use after free

Challenge: memory fault

- **Increased fault frequency:** multi-hop and interconnect switch
- **Expanded fault radius:** propagation among multi-node



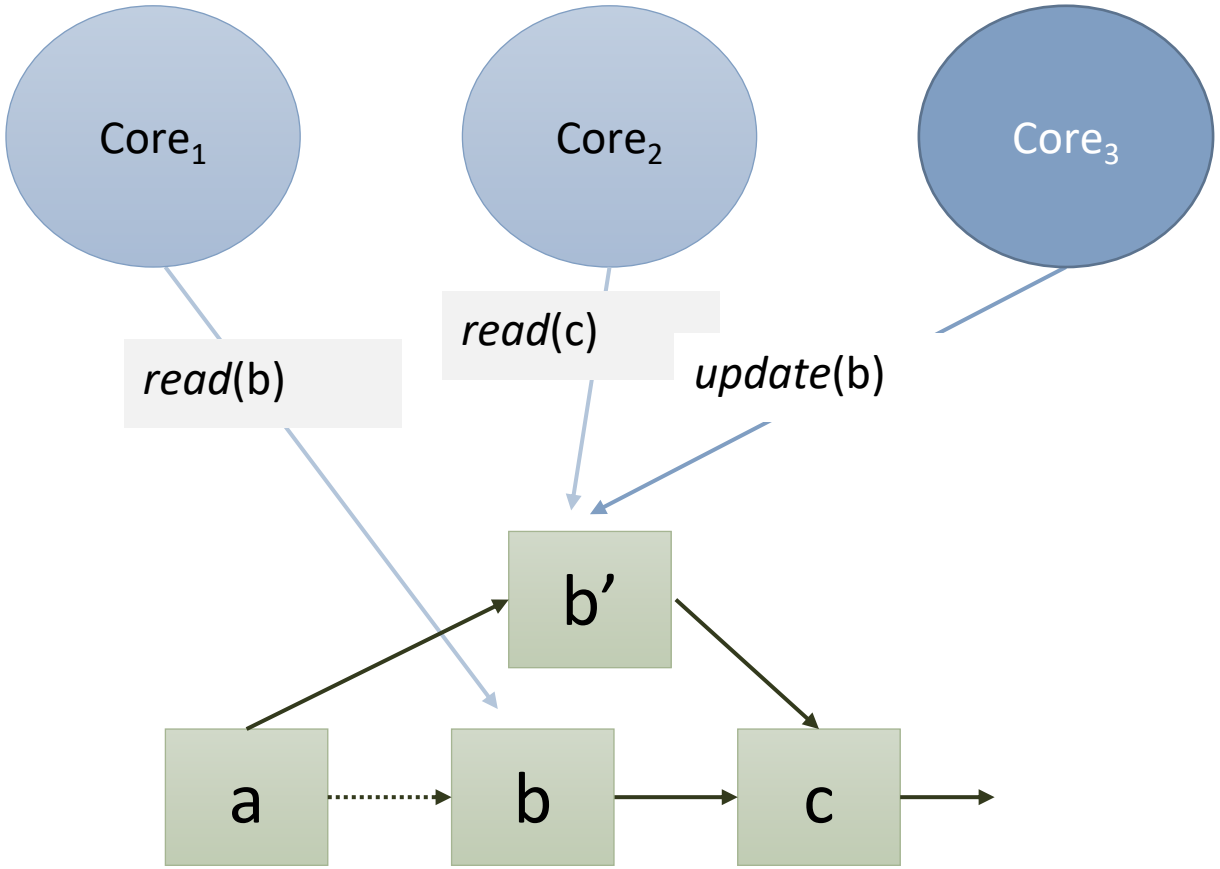
FlacOS: development kit

Common mechanism and primitives to develop on top of rack-scale shared memory

- **Synchronization**
  - Uncacheable memory
  - Cache flushing
  - Partition
  - Quiescence-based methods
- **Reliability**
- **Memory management**

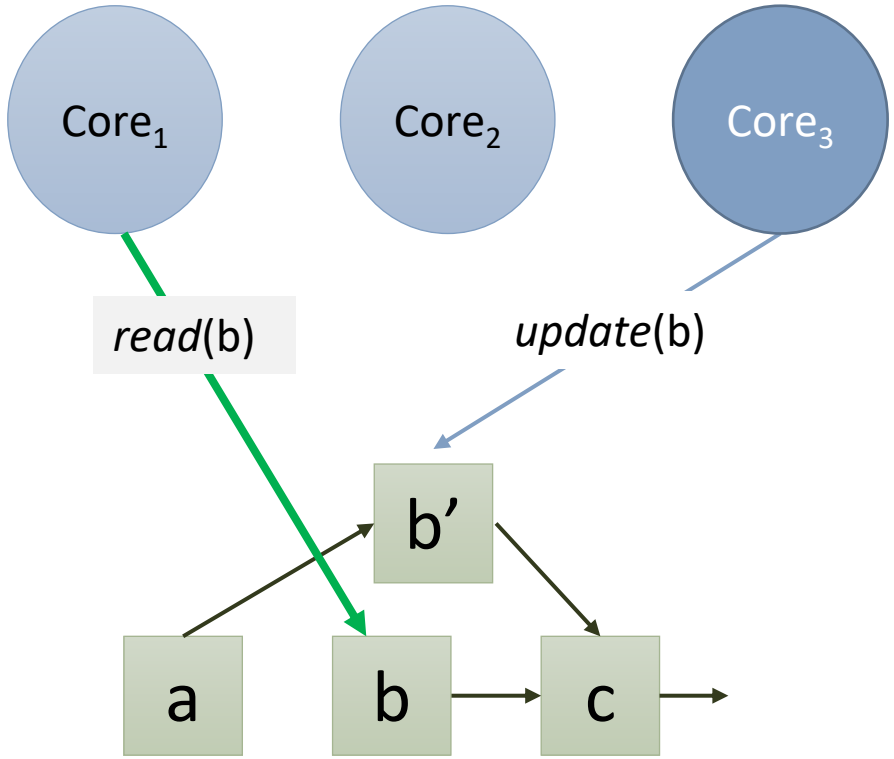
Quiescence-based cache coherence

Data staleness in Read-copy-update (RCU)

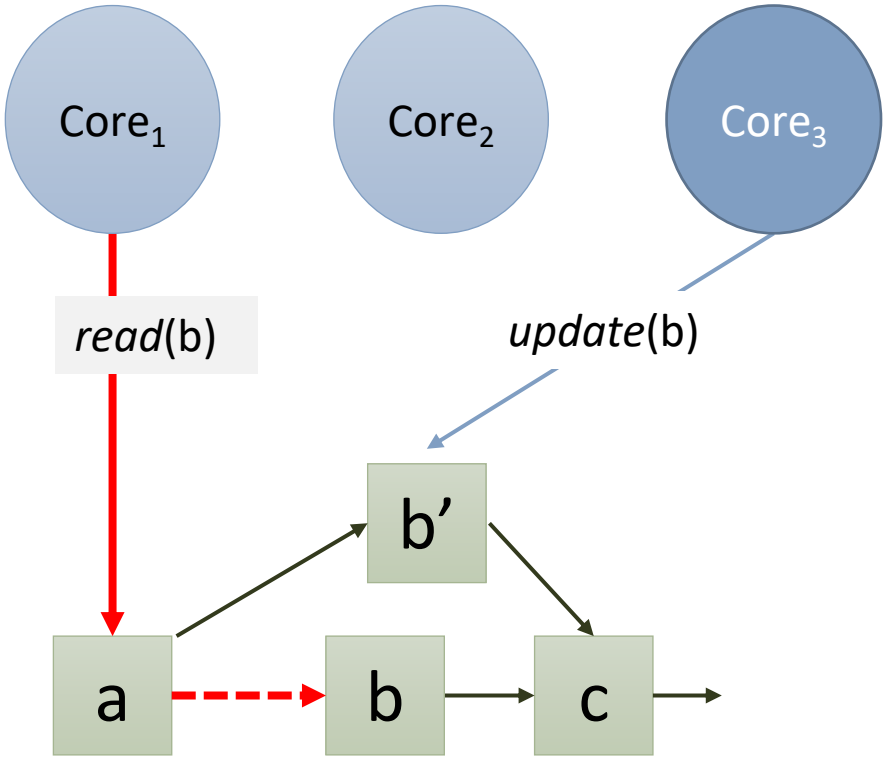


Quiescence-based cache coherence

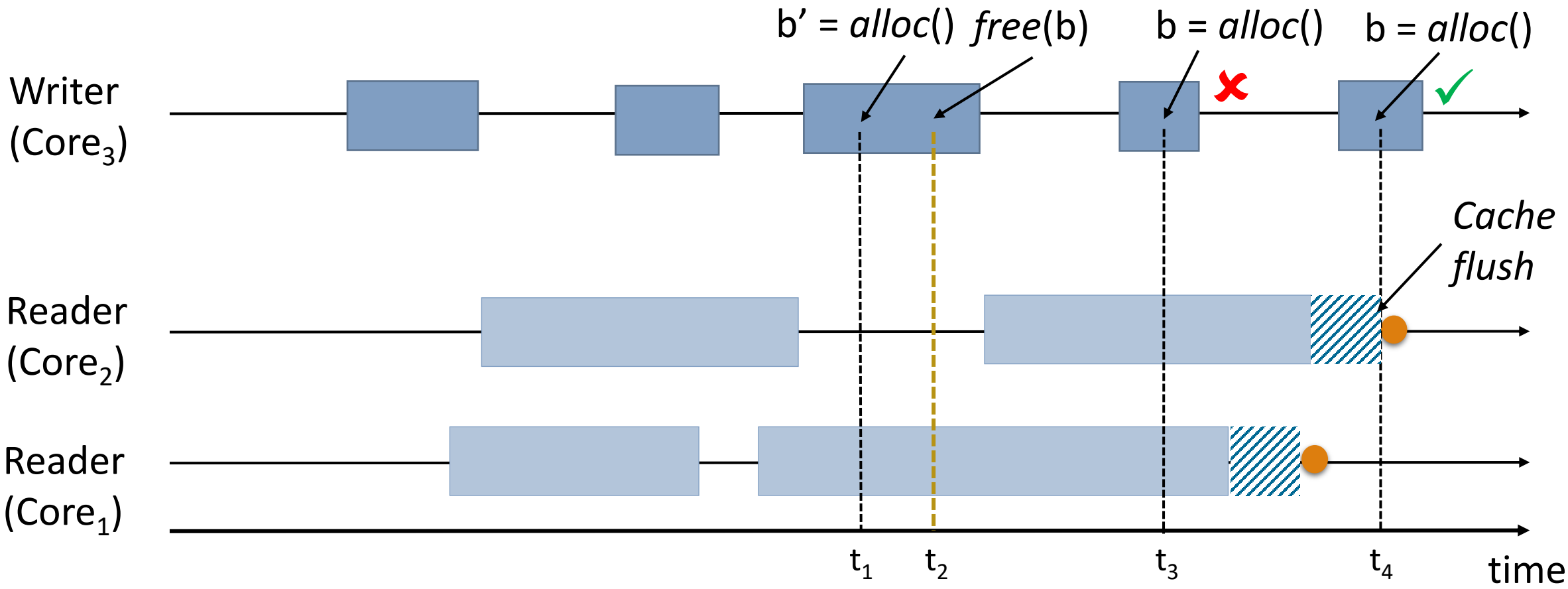
Parallel reference in RCU



Stale cache-line access in Rack



Quiescence-based cache coherence



When to reuse freed memory? Quiescence state [1, 2]

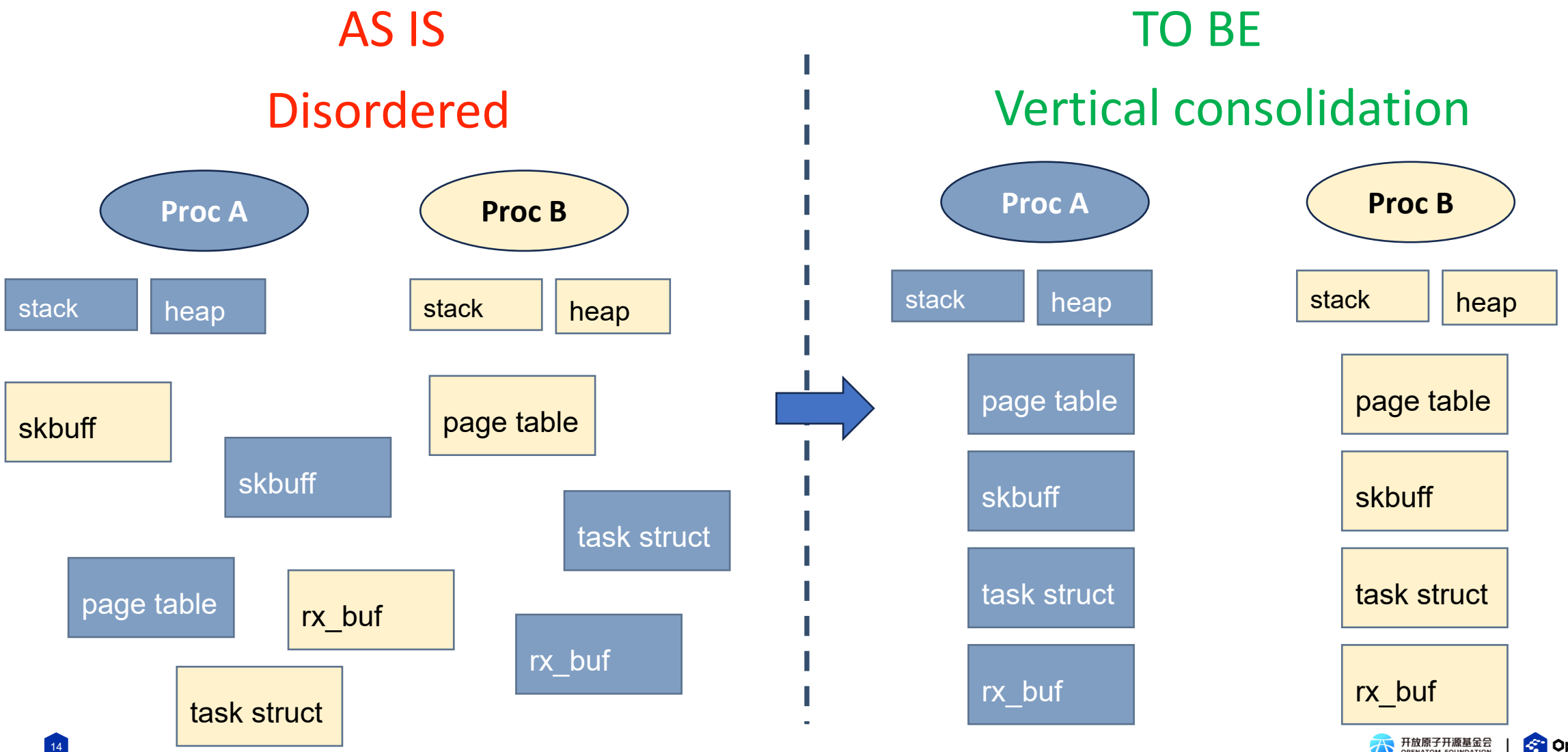
[1] Yuxin Ren, Gabriel Parmer, Dejan Milojicic Bounded Incoherence: A Programming Model for Non-Cache-Coherent Shared Memory Architectures (PMAM'20)
[2] Yuxin Ren, Gabriel Parmer, Dejan Milojicic Ch'i: Scaling Microkernel Capabilities in Cache-Incoherent Systems, (ROSS'20)

FlacOS: development kit

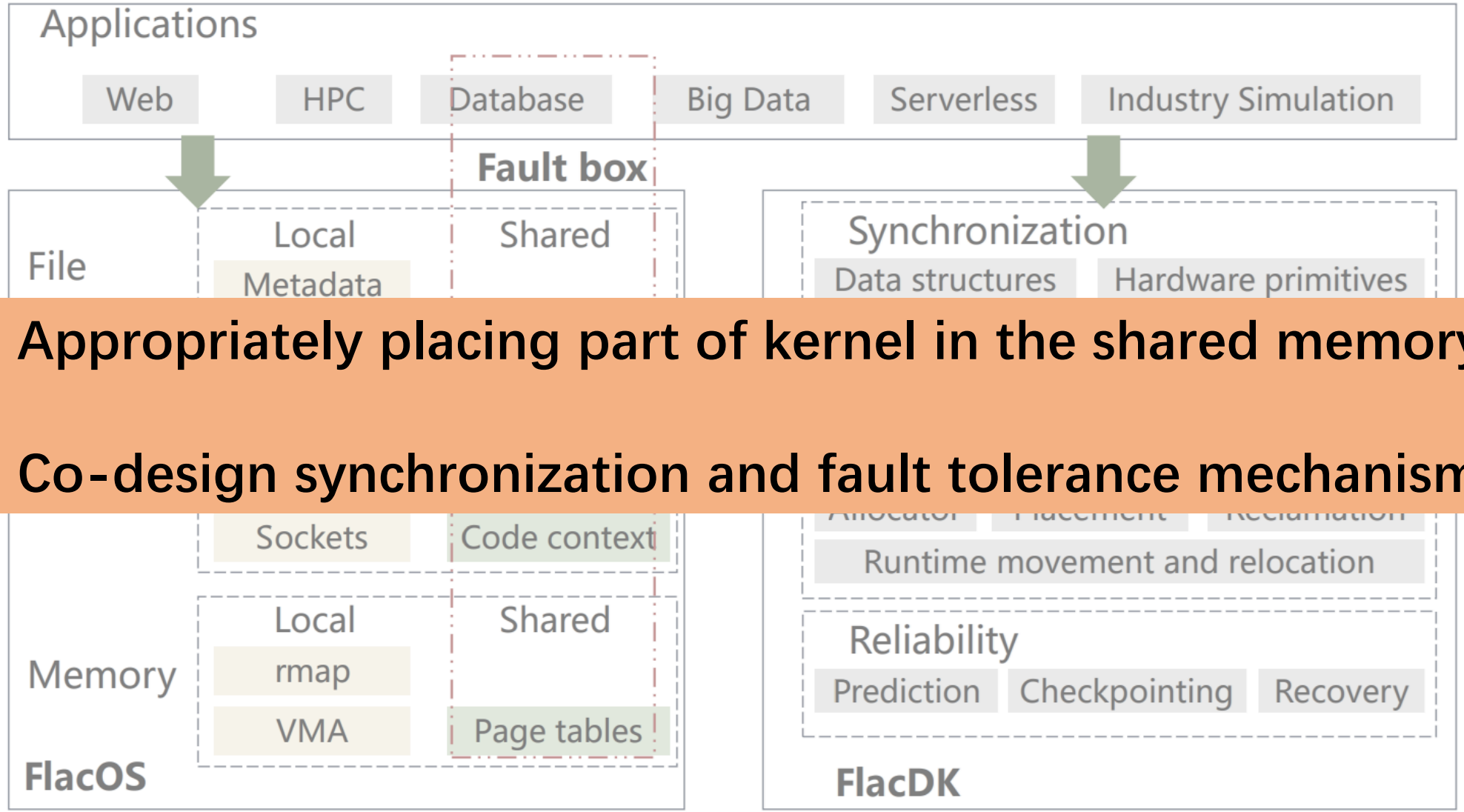
Common mechanism and primitives to develop on top of rack-scale shared memory

- Synchronization
- **Reliability**
  - Checkpointing
  - N-modular redundancy
  - **Fault box**
- Memory management

FlacOS: fault box

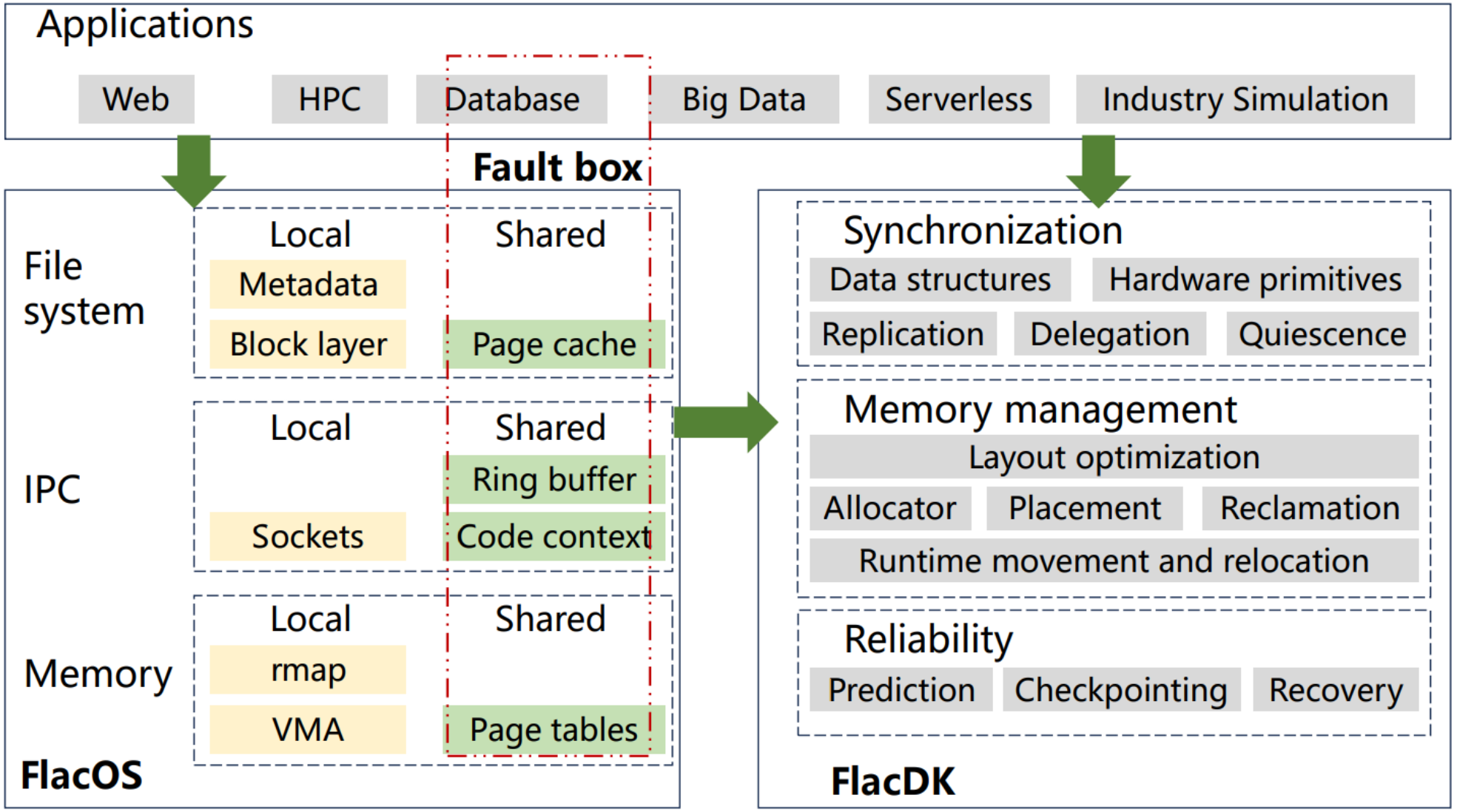


FlacOS: principles



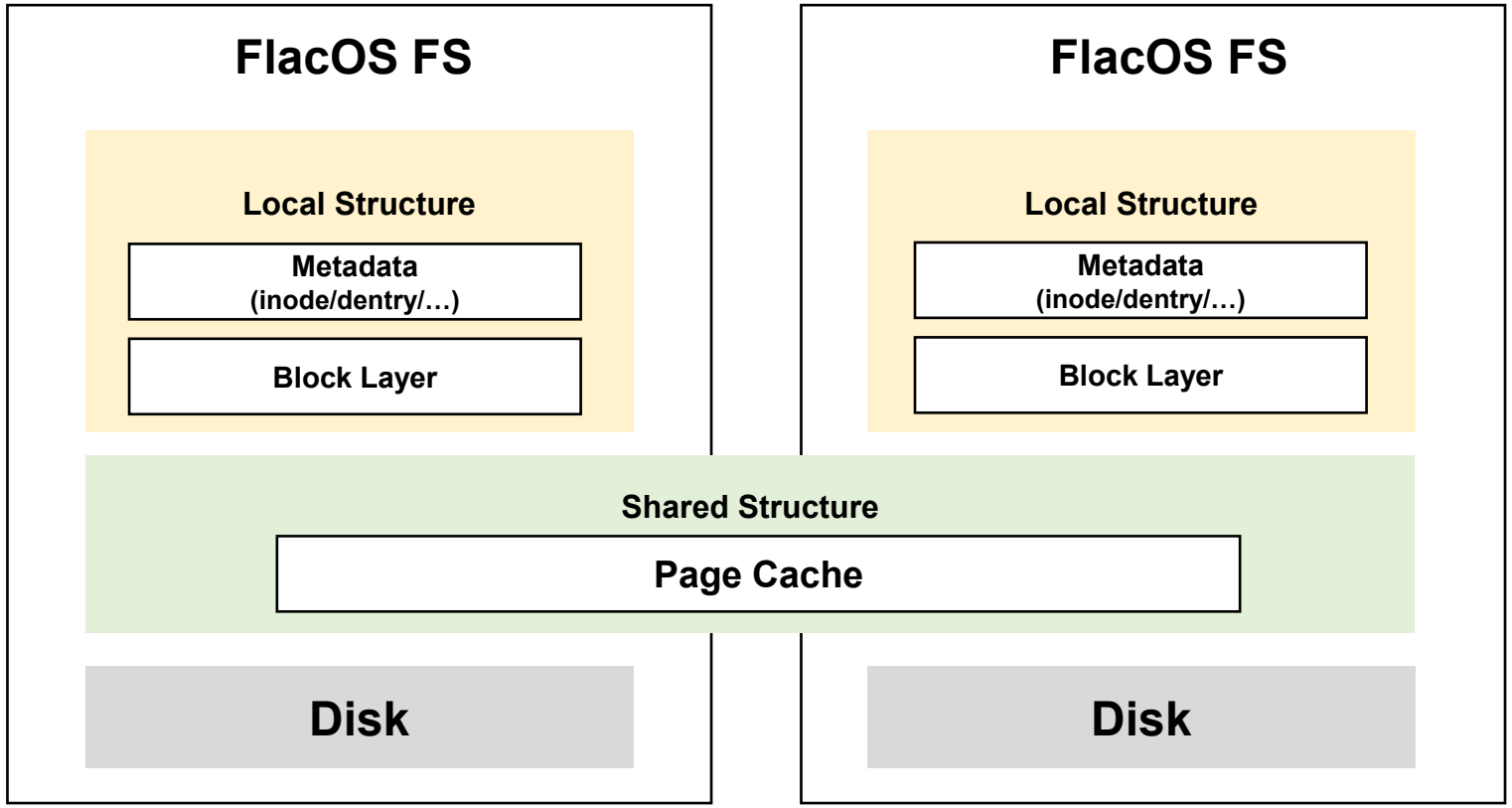
- Appropriately placing part of kernel in the shared memory
- Co-design synchronization and fault tolerance mechanisms

FlacOS: architecture



FlacOS: file system

- **Shared page cache**
  - Reduced redundancy
  - Accelerate shared file access
  - Elastic cache
- **Local meta data**
  - Complicated structures
  - Compatibility



Prototype evaluation

- Two Kunpeng node, each has 4 NUMA, 320 cores
- Interconnect：HCCS
- Simulation: virtual machine plus persistent memory
- 4GB Pytorch container image

3.81X speedup

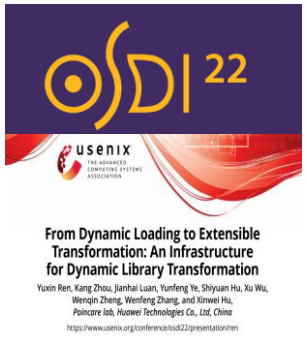
Default	FlacOS
21.067s	5.526s

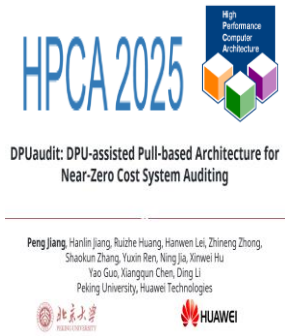
Conclusion

- Rack-as-a-Computer
- Rack-scale OS sharing
- Rack-scale reliability

Open-source OS research in Huawei

- **Broadly covers system research**
  - Operating system
  - Networking and storage
  - System security
  - Real-time system
- **Many papers in top conference: OSDI, HPCA, FAST, USENIX Security…**





THANKS