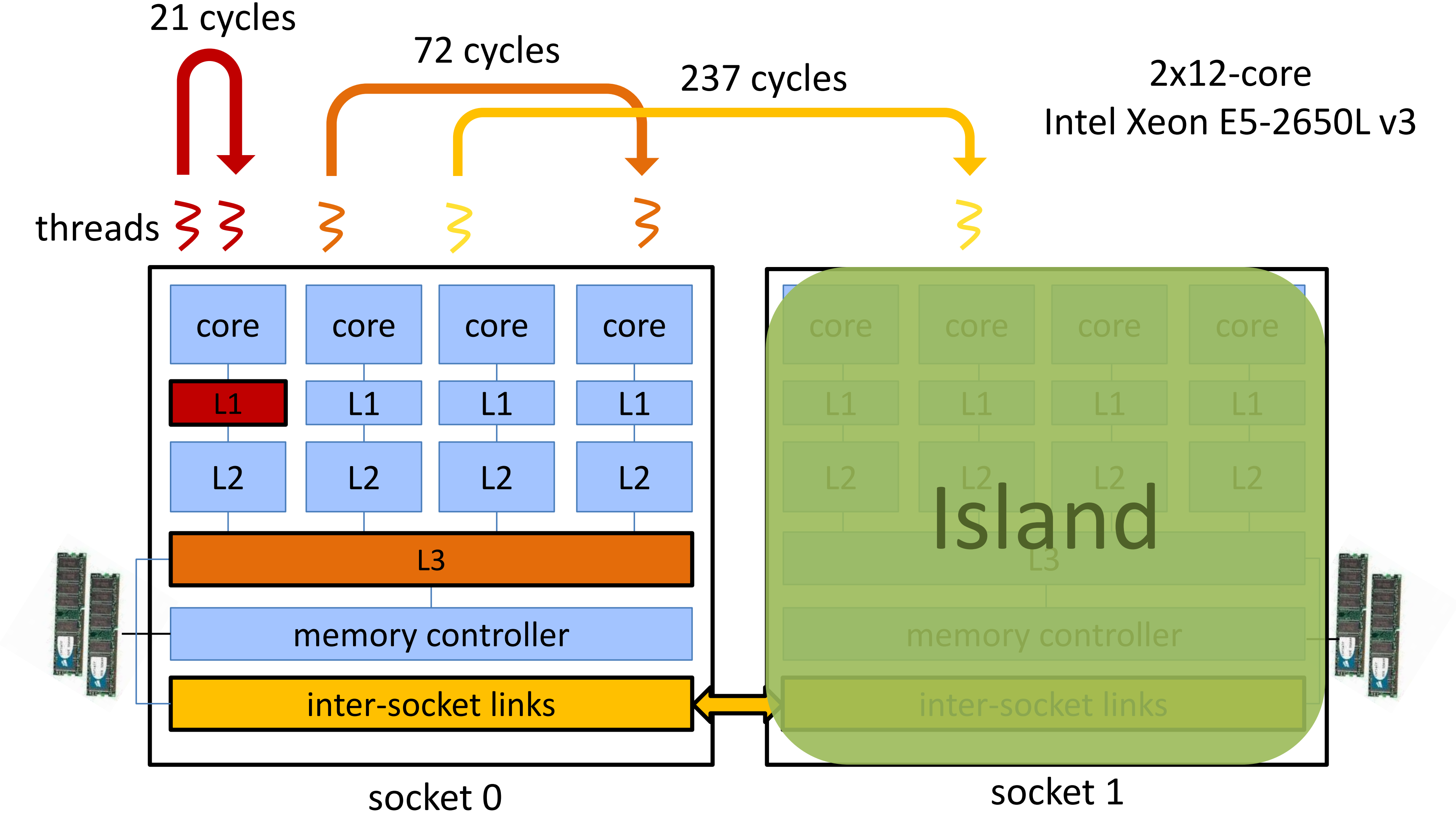
More Than A Network: Distributed OLTP on Clusters of Hardware Islands

Danica Porobic, Pınar Tözün, Raja Appuswamy, Anastasia Ailamaki

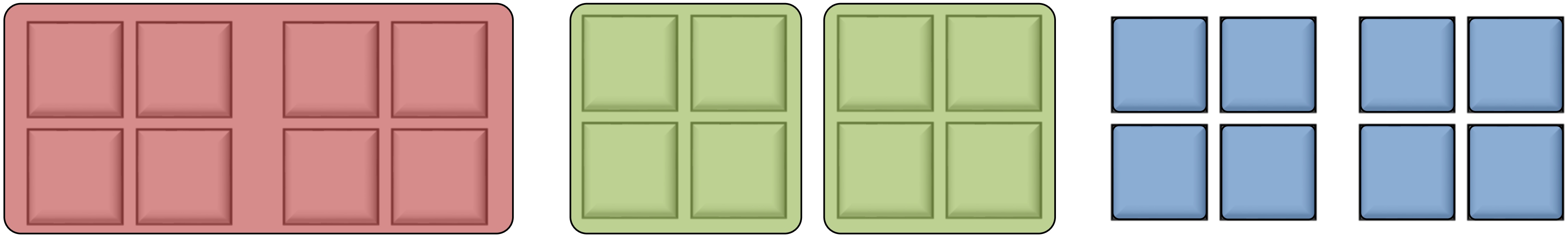
Data-Intensive Applications and Systems

Multisocket multicores



Challenge: non-uniform communication

OLTP on Hardware Islands



Shared-everything

✓ stable

✗ not optimal

Island shared-nothing

✓ robust middle ground

Shared-nothing

✓ fast

✗ sensitive to workload

Optimal configuration depends on workload and hw

Rack-scale hardware platforms

- Abundant non-uniform parallelism
  - Need to scale across many cores
- Large main memories
  - Datasets are memory resident
- Network & DRAM converge
  - Need to scale across multiple nodes

Complex hierarchy of Hardware Islands

How different are clusters of Islands?

- Does Island topology still matter in the cluster environment?

- Does faster communication always improve throughput?

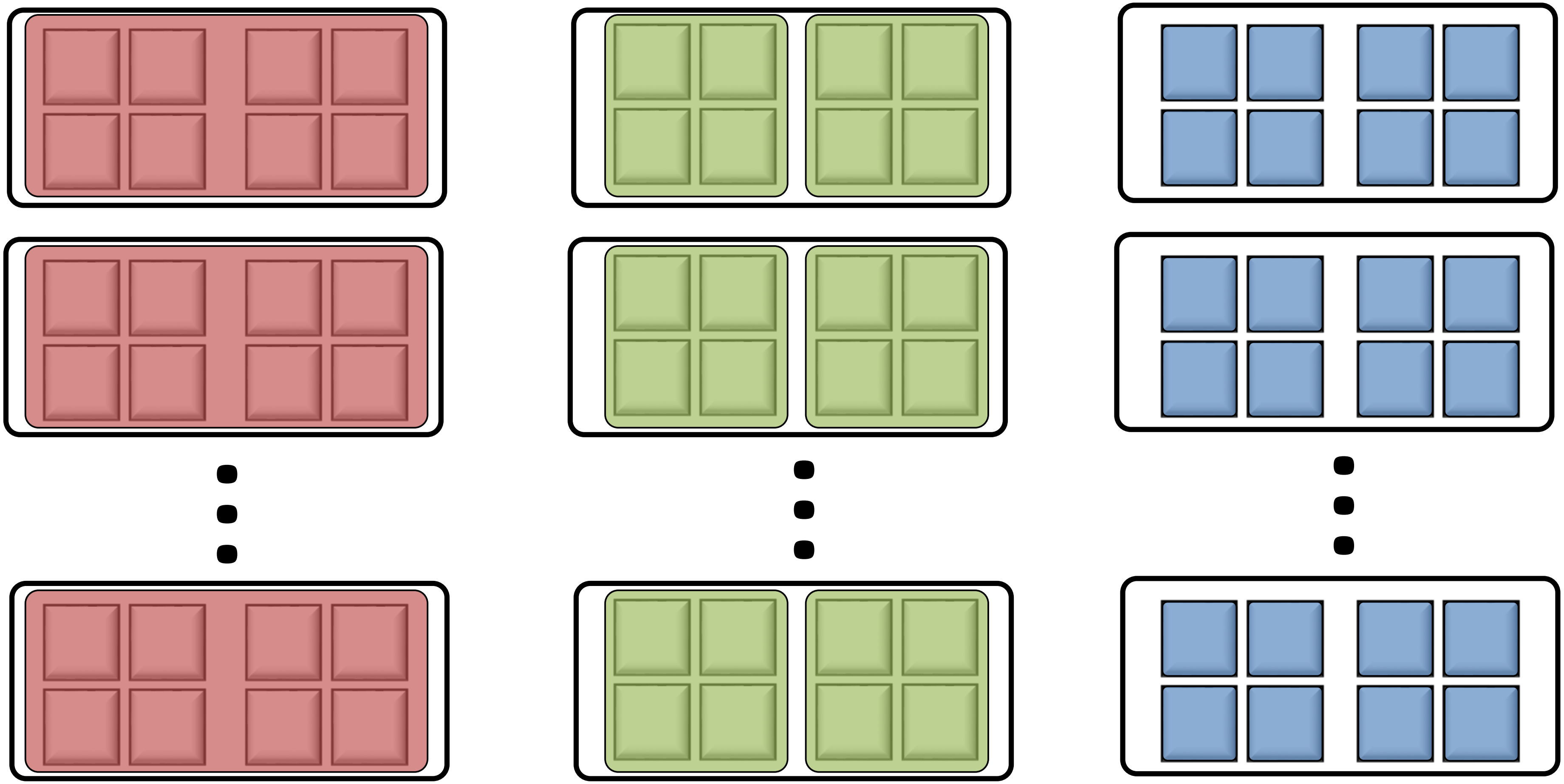
- How do scale-up designs perform when used in distributed deployments?

Experimental setup

- Shore-MT
- 2 x 6-core Intel Xeon X5660
- 10 Gbps Ethernet
- TCP/IP and shared memory communication
- TPC-C and partition-sensitive microbenchmark

- Silo
- 8 x 10-core Intel Xeon E7-L8867
- Unix sockets and shared memory communication
- Partition-sensitive microbenchmark

Distributed deployments



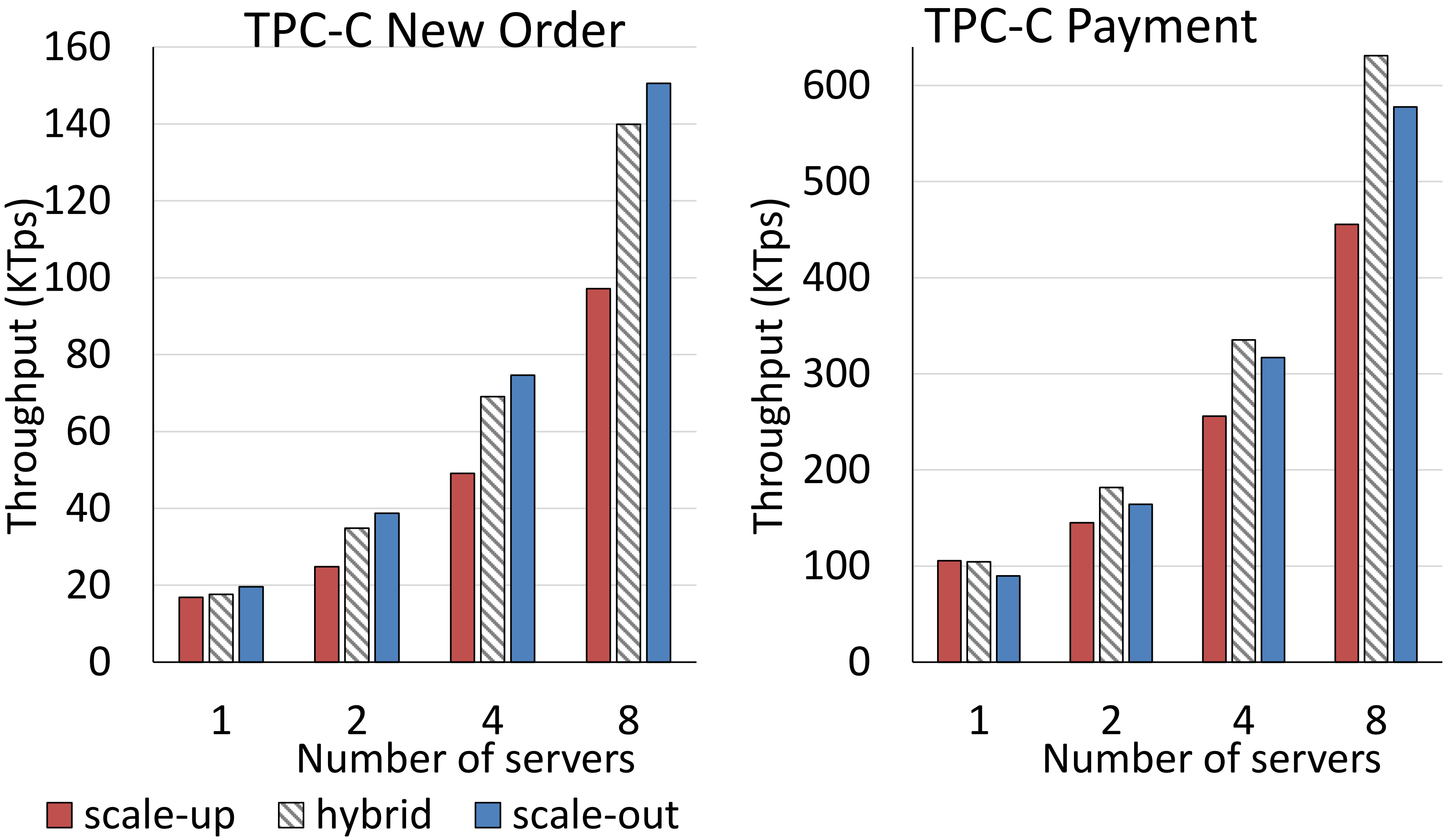
scale-up

hybrid

scale-out

Scaling out across the cluster

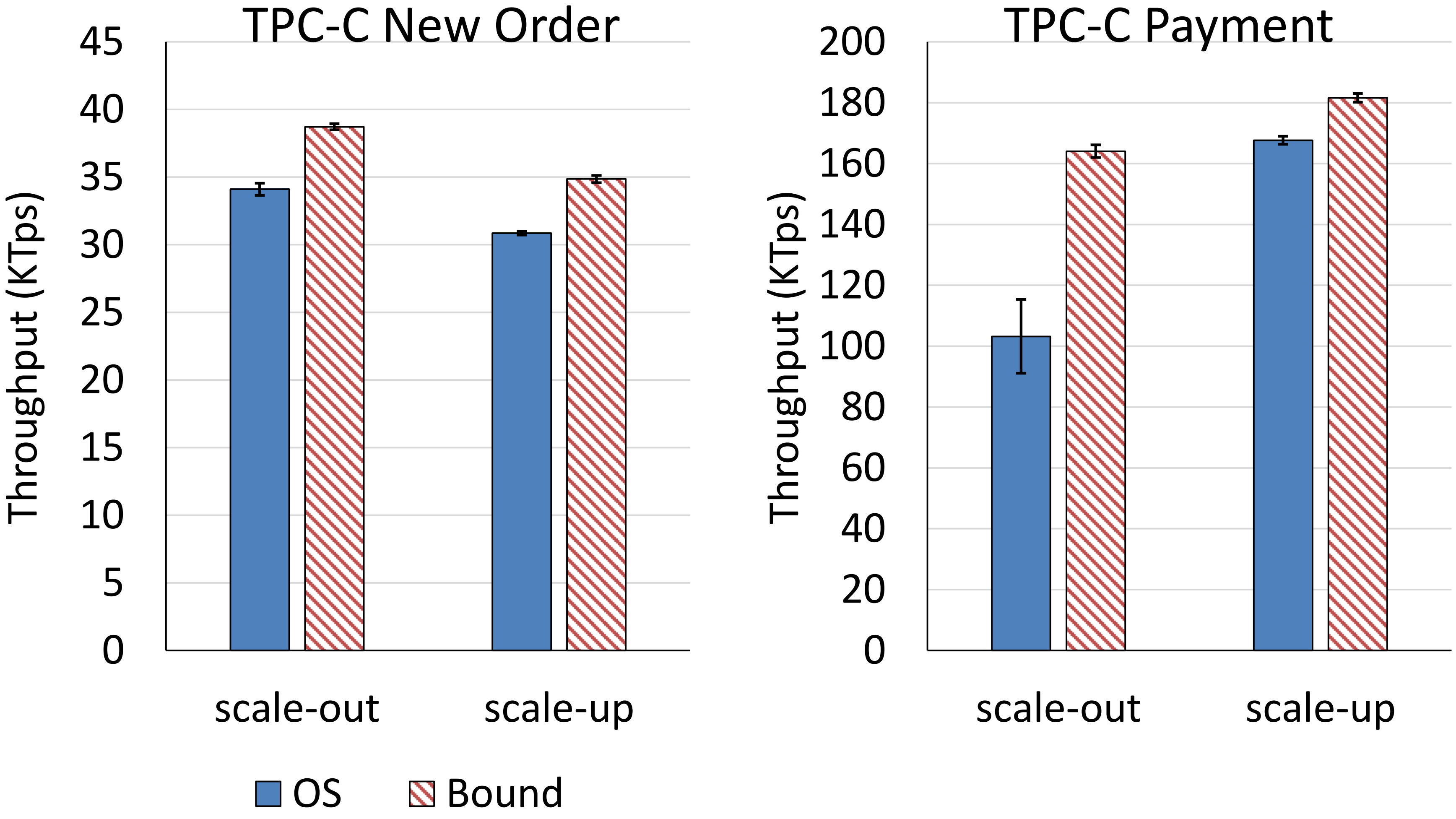
Shore, TCP/IP



No configuration is optimal for every cluster

Impact of placement

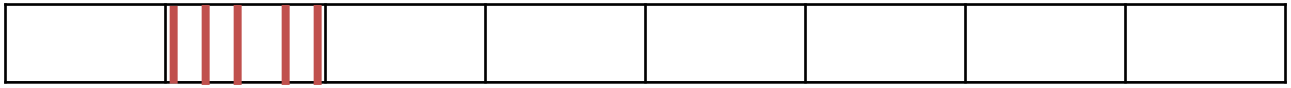
Shore, TCP/IP



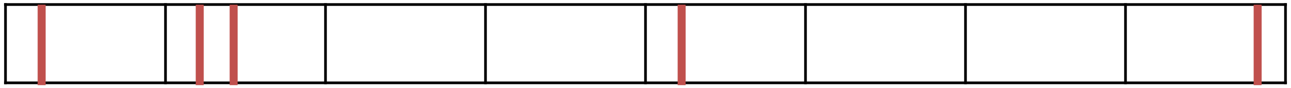
Thread migrations hurt performance & predictability

Partition sensitive microbenchmark

- Single site version
  - probe/update N rows from the local site

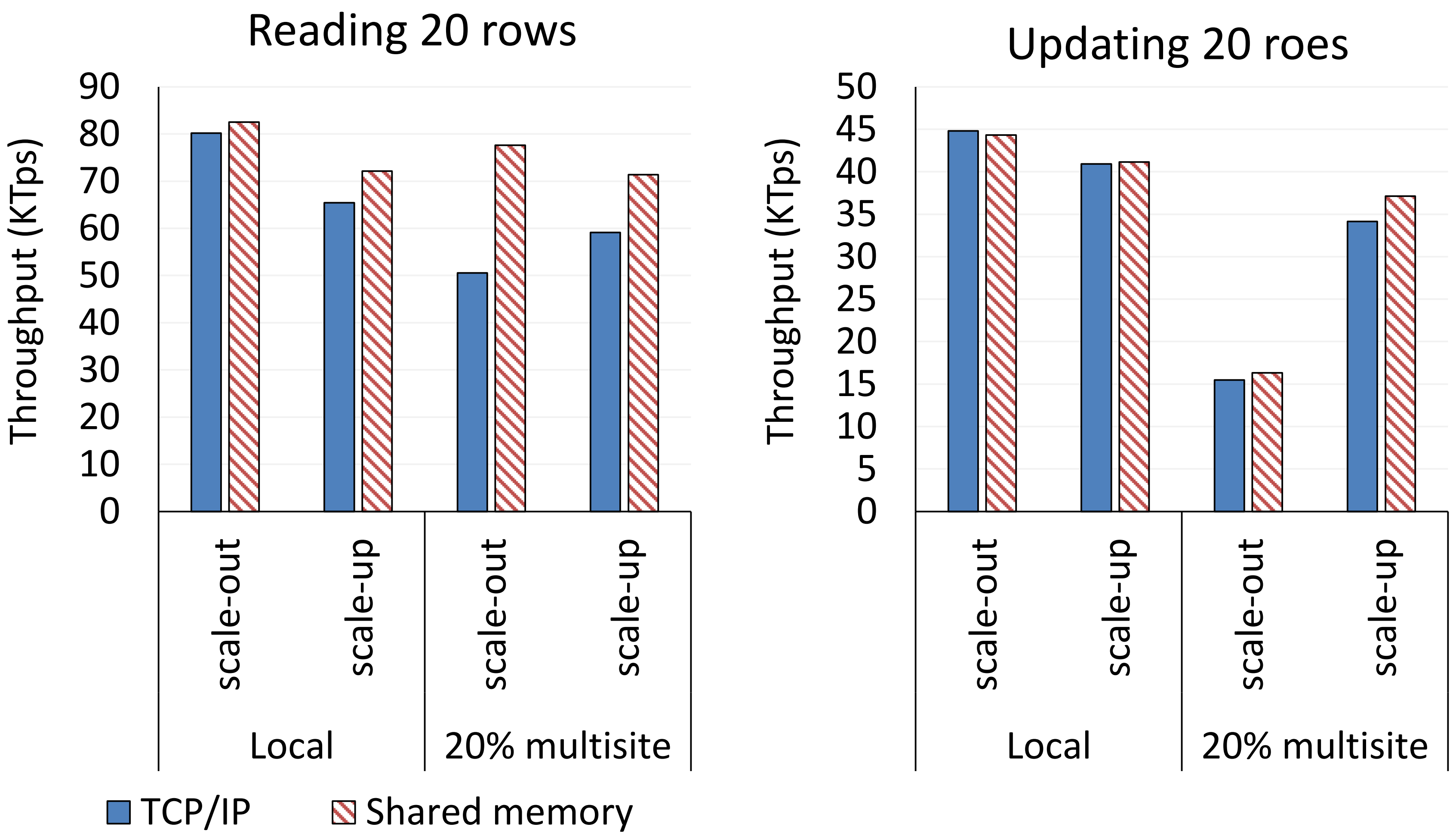


- Multisite version
  - probe/update 1 row from the local site
  - probe/update N-1 rows uniformly from any site
  - sites may reside on the same instance



Impact of fast communication

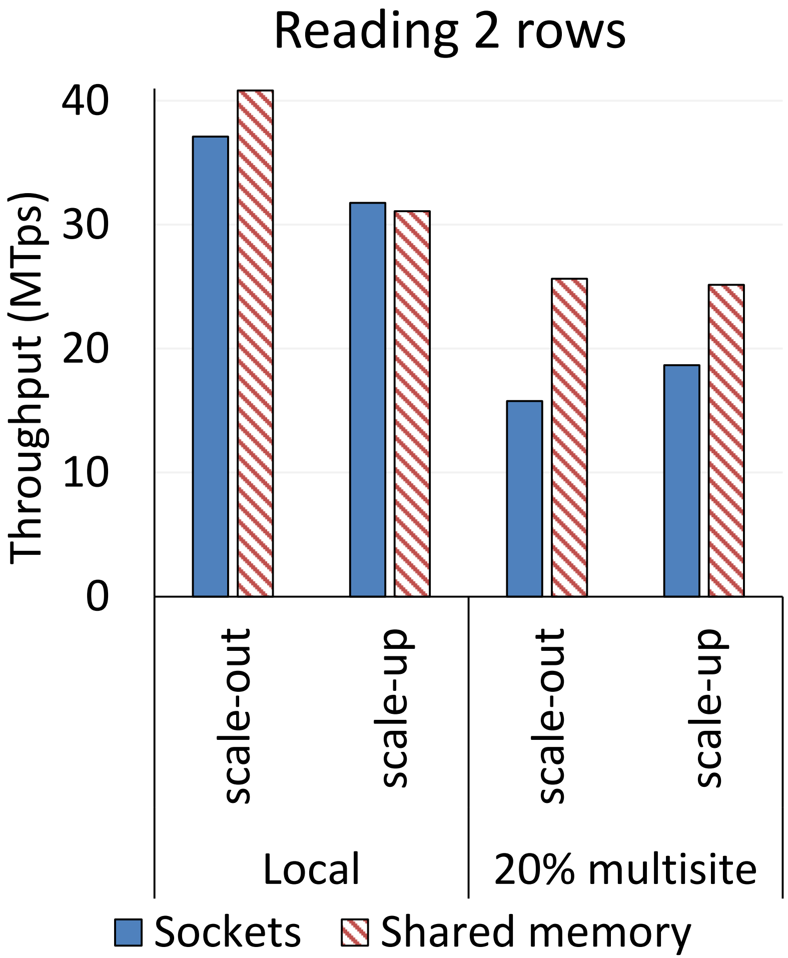
Shore-MT

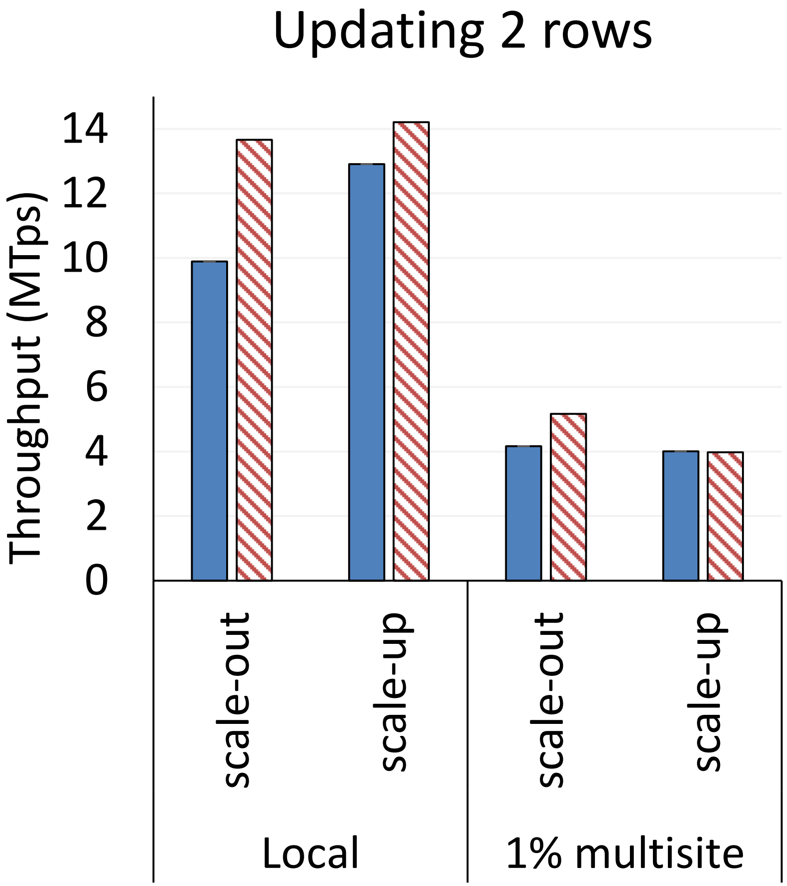


Read-only: helps, Updates: little impact

Scaling out a scale-up system

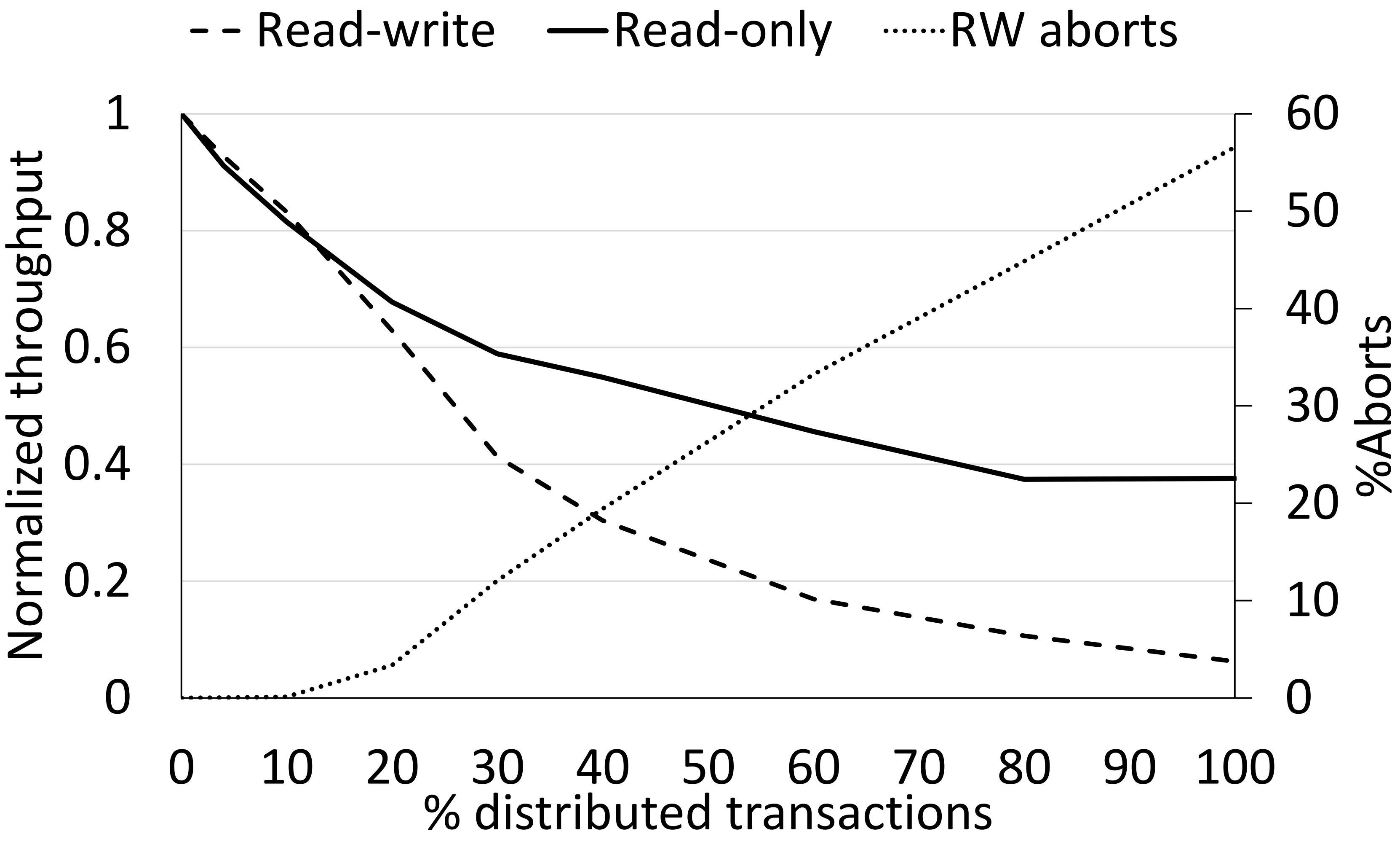
Silo





Distributed updates cause severe throughput drop

Why don't updates scale out?



Multicore-optimized OCC very sensitive to delays

OLTP on a Cluster of Islands

- Scale-up designs sensitive to scale-out delays
- Islands-awareness required, but insufficient for optimal cluster deployments
- Fast communication can improve throughput, but does not guarantee improvement

Thank you!