BALM: QoS-Aware Memory Bandwidth Partitioning for Multi-Socket Cloud Nodes

David Gureya[1,2], Vladimir Vlassov[2] and João Barreto[1]

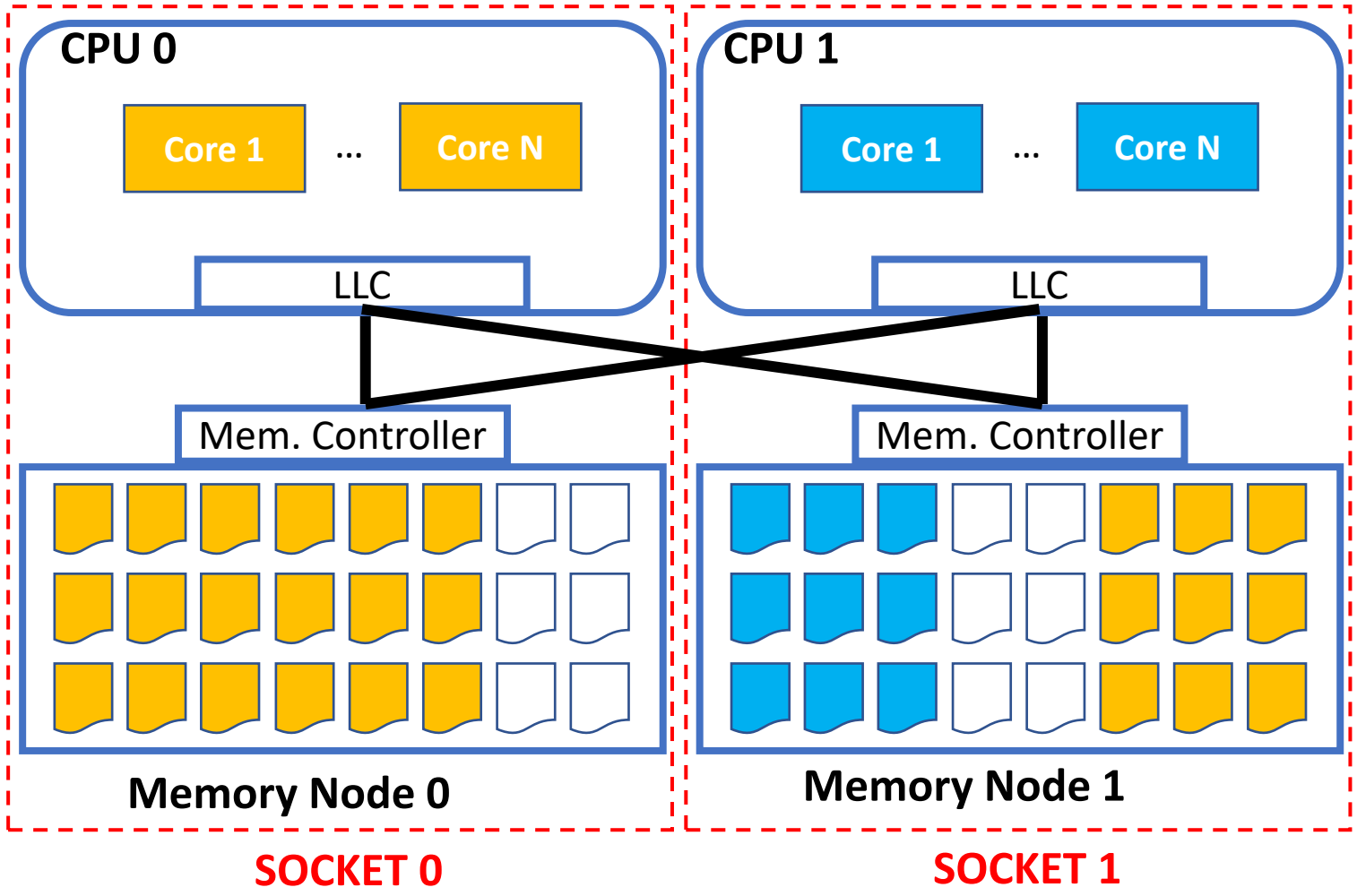
July 6, 2021 – SPAA 2021

[1] INESC-ID, IST, University of Lisbon, Portugal | [2] KTH Royal Institute of Technology, Sweden

Workload Consolidation (1/2)

- Widely used to improve resource utilization
  - Multiple workloads are consolidated on the same physical servers

Workload Consolidation (2/2)



Best-Effort Application (BEA)

- Throughput-oriented

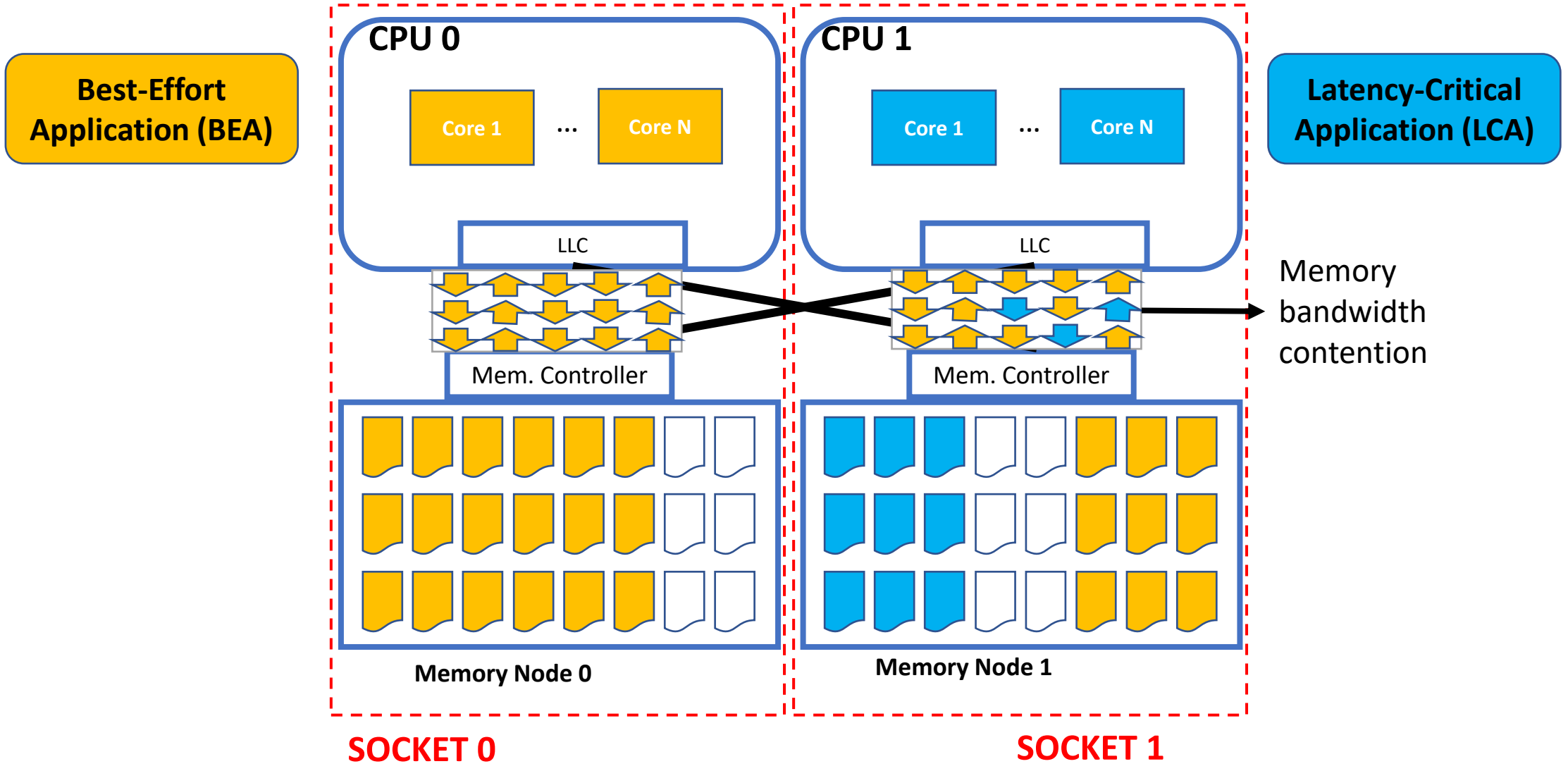
Bandwidth-intensive BEAs

- Place pages across Memory nodes

Latency-Critical Application (LCA)

- QoS requirements

Challenge: Contention for Memory Bandwidth



Problem: QoS-Aware Resource Allocation

- Safeguard the SLO of LCAs while maximizing the throughput of the BEAs

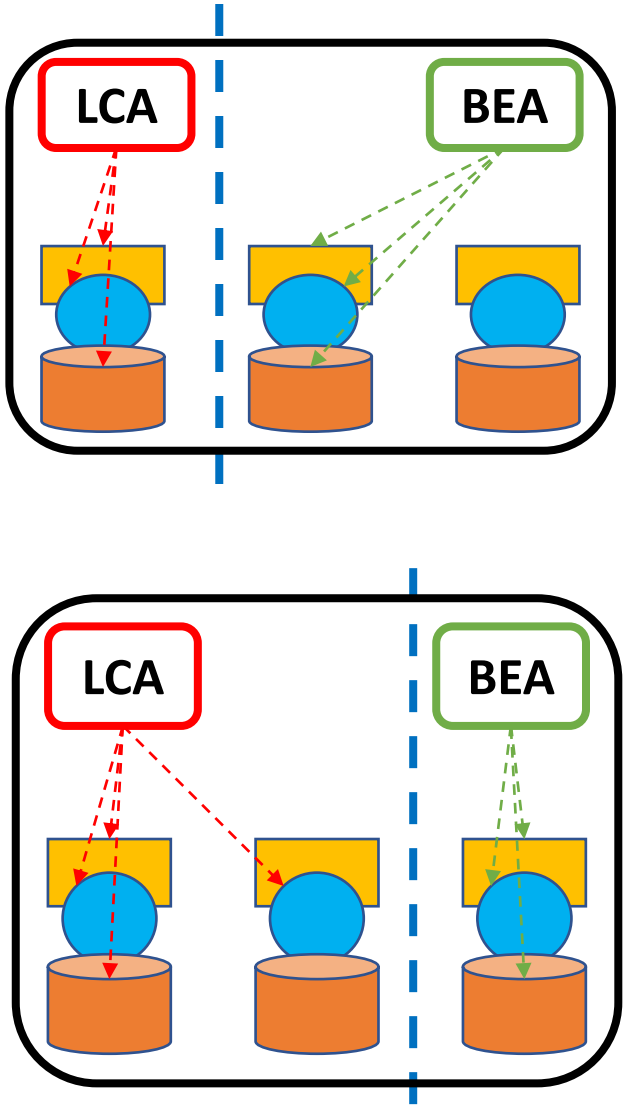
- **Dynamic problem**: resource usage by each application can change at any time

Existing Solutions (1/2)

- Solve the problem by partitioning resources

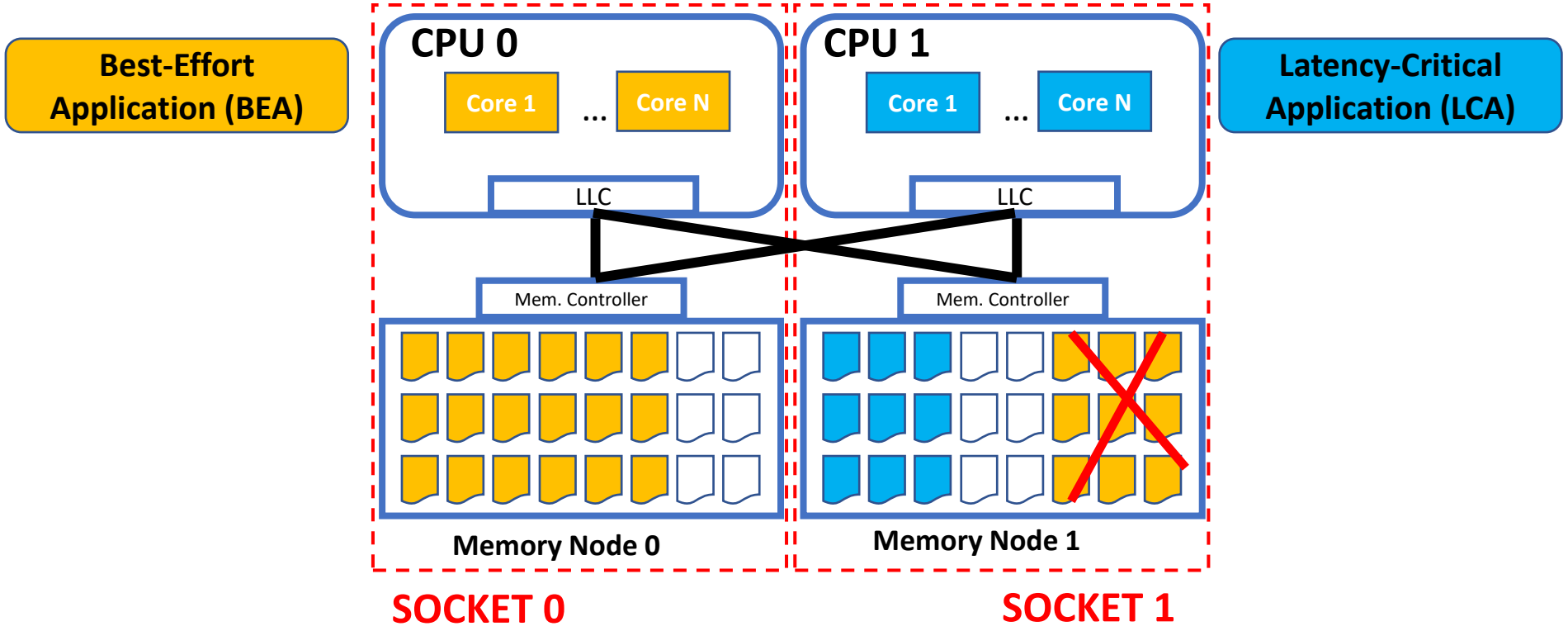
- Recent systems dynamically adjust partitions
  - Heracles [ISCA'15]
  - Parties [ASPLOS'19]
  - Clite [HPCA'20]
  - Caladan [OSDI'20]

- However, existing solutions are tailored to **single-socket architectures** only

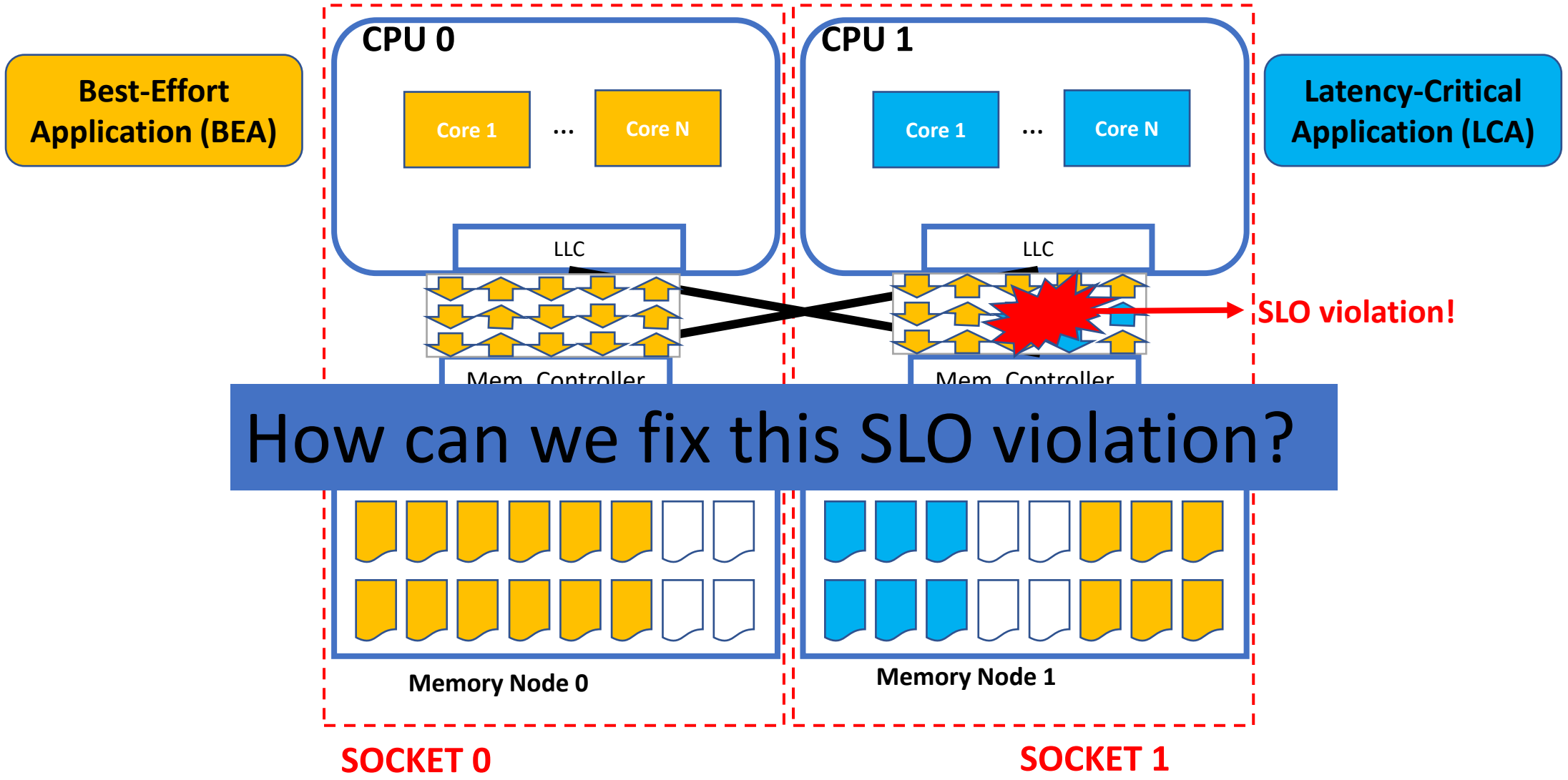


Existing Solutions (2/2)

- Disallows cross-socket sharing of memory

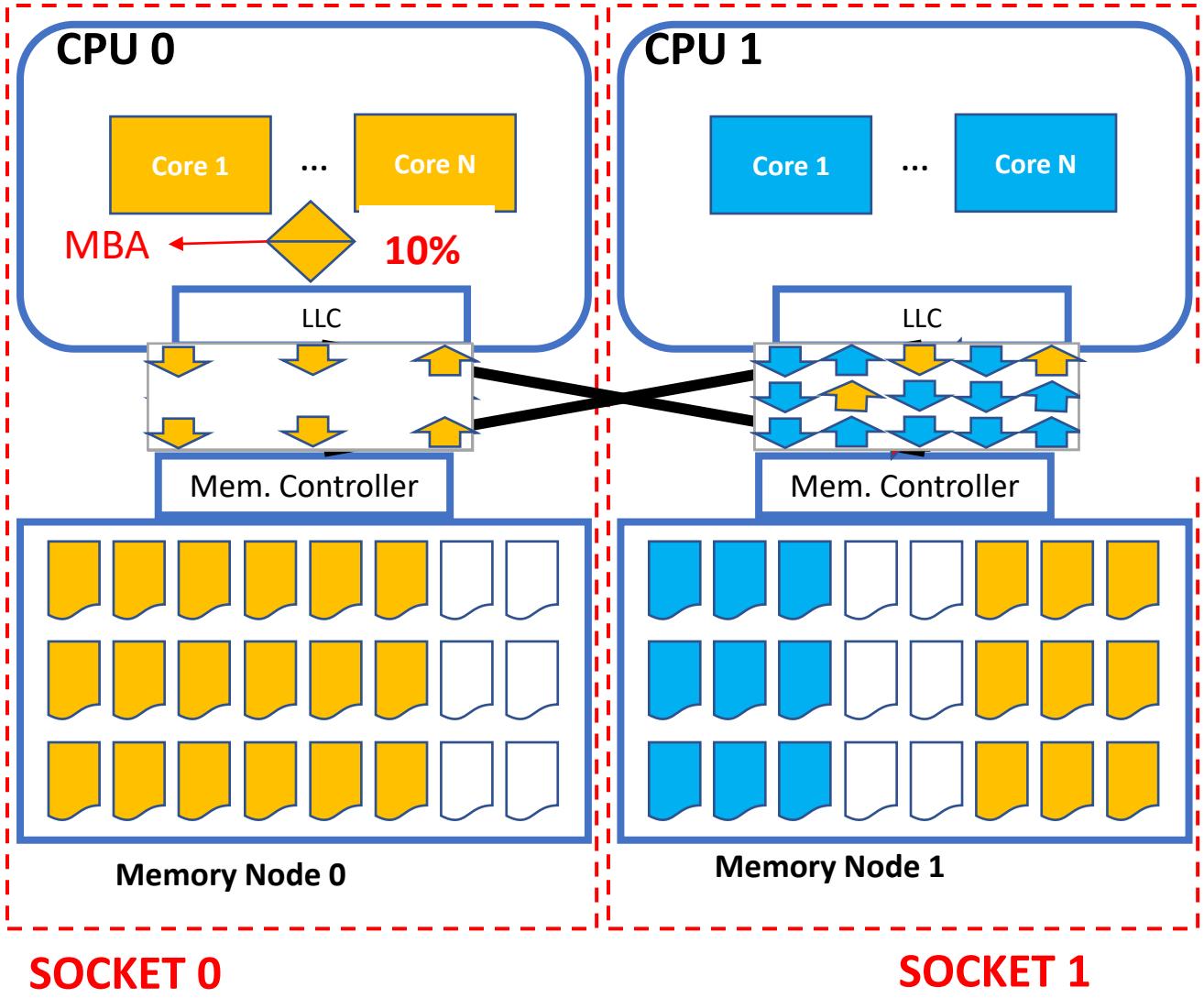


Cross-socket QoS-Aware Memory Bandwidth Allocation



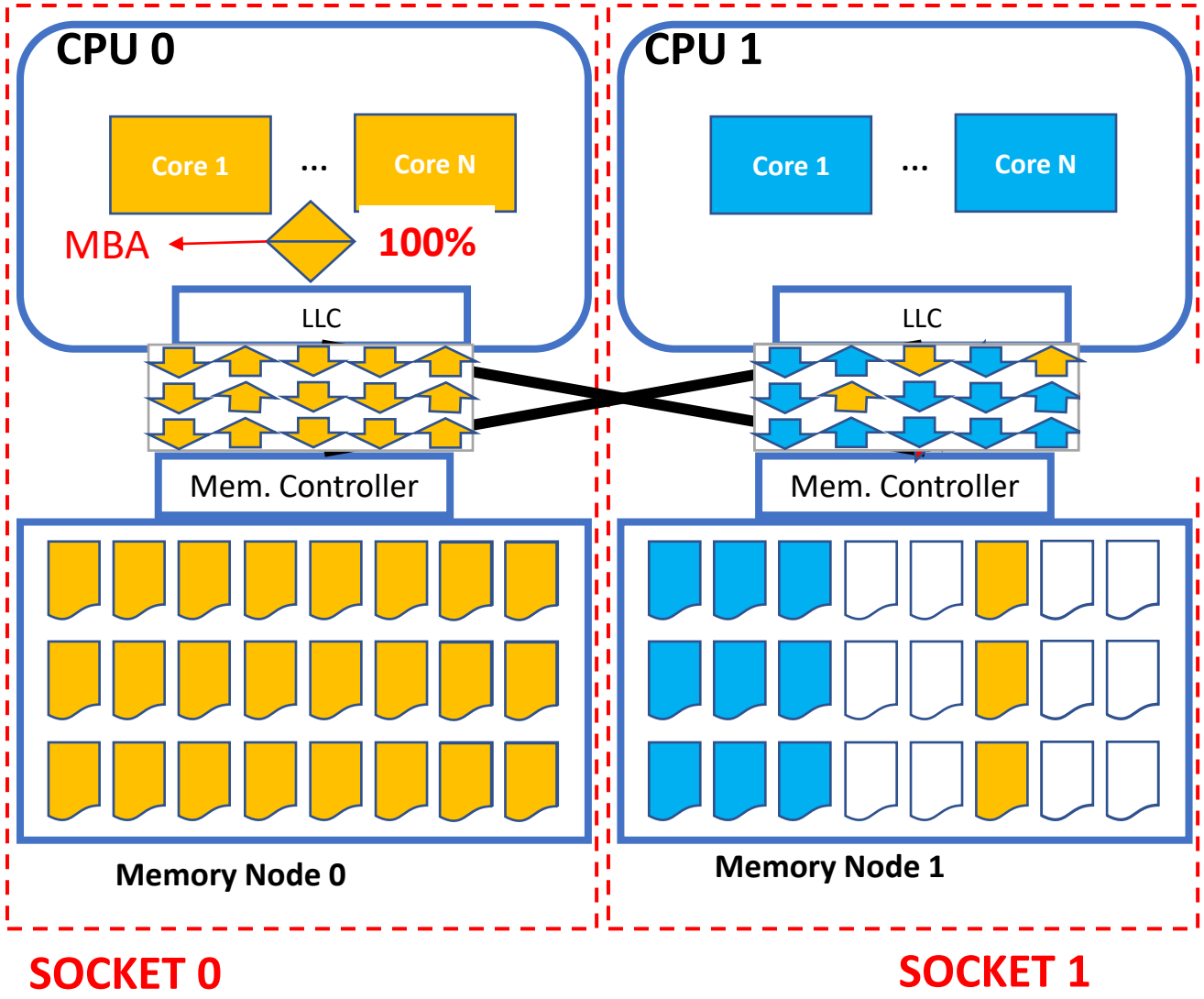
Fixing SLO Violation (1/2)

- **Possibility 1:** Intel **Memory Bandwidth Allocation (MBA)**
  - Can fix SLO violations almost instantaneously
  - Has a relevant cost on the performance of BEAs



Fixing SLO Violation (2/2)

- **Possibility 2: Page Migration (pgm)**
  - Can adjust memory bandwidth on a **per-memory node granularity**
  - Page migration is slow but more efficient for BEA



Contributions

- Study the hardware mechanism of Intel for memory bandwidth allocation (MBA) in a multi-socket scenario.
  - MBA unnecessarily reduces the throughput of BEAs by considerable margins

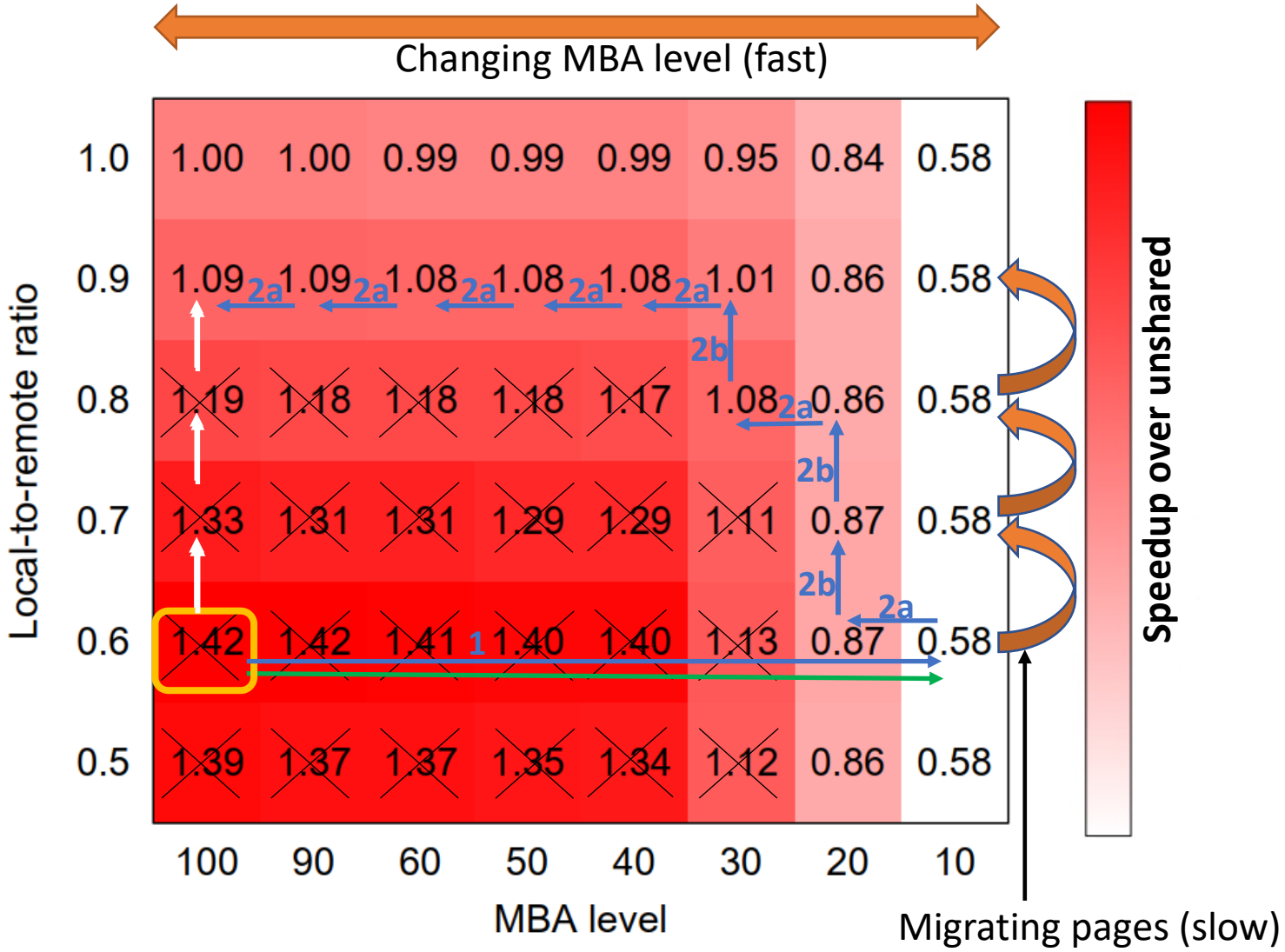
- We propose BALM, a novel QoS-aware memory bandwidth allocation technique for cross-socket sharing of memory in multi-socket architectures.
  - MBA with a novel cross-socket page migration scheme to obtain the best of both mechanisms

Mitigating SLO Violations with BALM

- BALM uses MBA and page migration together as a **2-dimensional allocation mechanism**

- BALM fixes SLO violations in 2 steps:
  - Set MBA to its most restrictive level
  - Apply incremental page migration
    - Gradually release MBA throttling

- BALM is expected to be:
  i. As quick as MBA in fixing SLO violations
  ii. Converge to valid optimal configuration as page migration



Evaluation: Questions (1/4)

1. What performance advantage does BALM bring to memory-intensive BEAs on dual-socket NUMA systems?
2. How effective is BALM in fixing SLO violations?

- Compared BALM with:
  - MBA
  - Page migration (pgm)
  - Unshared

Evaluation: Methodology (2/4)

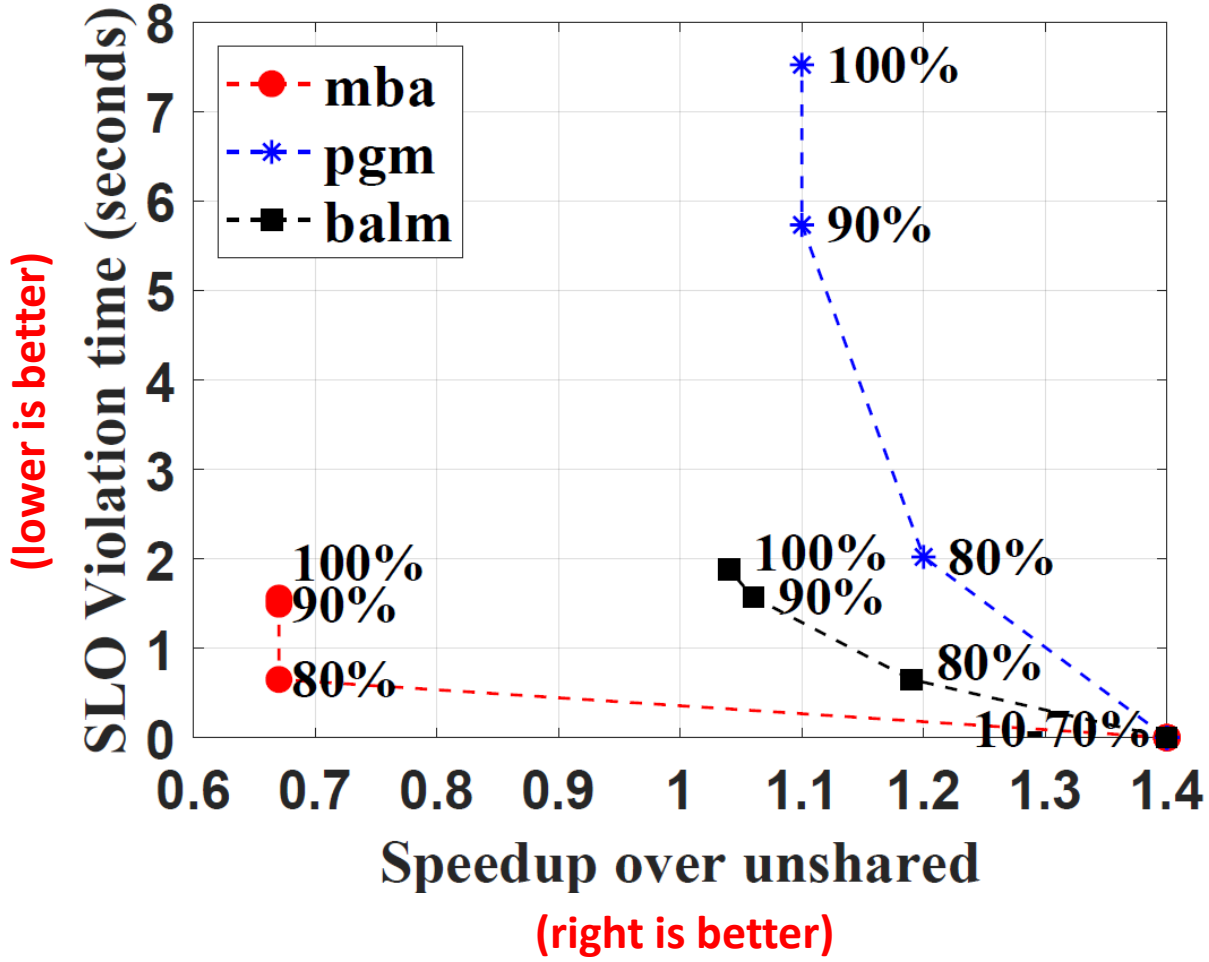
- BEAs: multithreaded benchmarks from PARSEC, SPLASH, NAS
- LCAs: Memcached

- Machines: dual-socket system
  - Intel Xeon Silver 4114 CPU, 2 NUMA nodes, 10 cores per node, supports MBA

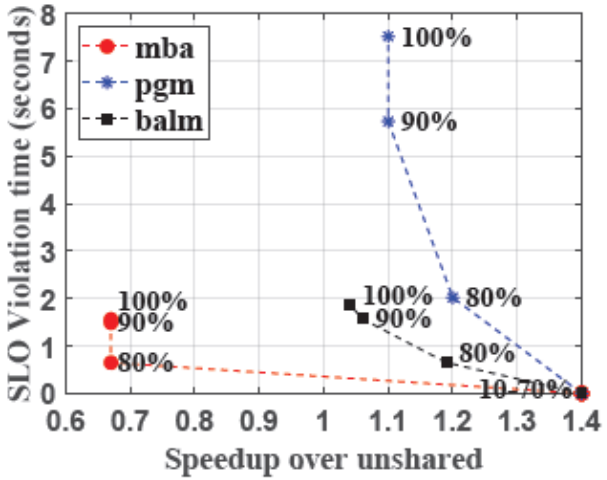
- Execution Scenario

Socket 0	Socket 1
Memcached (4 threads)	OC/MG/SP/UA (8 threads)

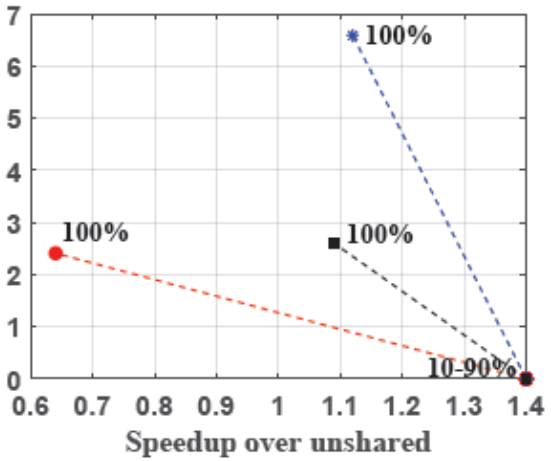
Evaluation: Results (Memcached vs. OC) (3/4)



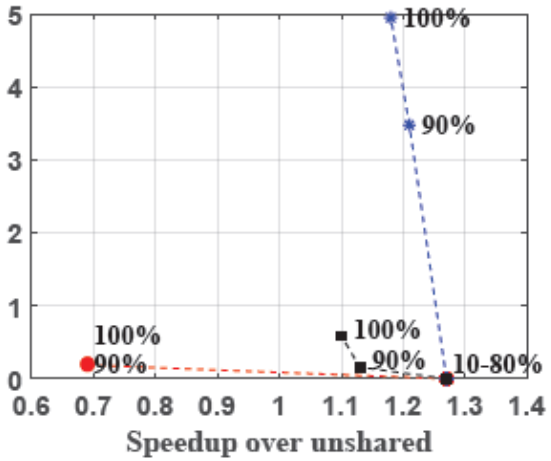
Evaluation: Results (Memcached vs. all) (4/4)



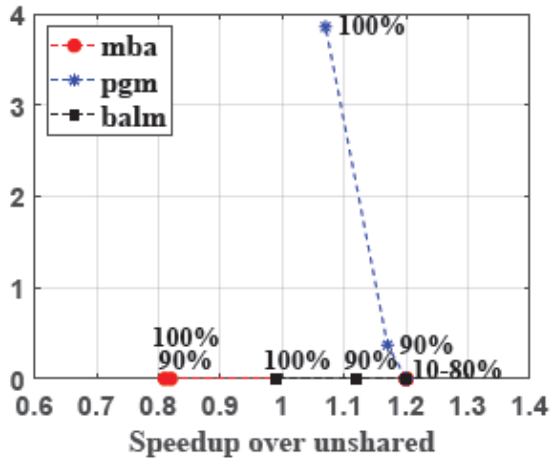
(a) Ocean_cp



(b) MG.C



(c) SP.C



(d) UA.C

Conclusion

- SoTA QoS-aware resource allocation systems need to be generalized to allow cross-socket sharing of memory bandwidth

- BALM can safeguard the LCAs with **marginal SLO violation windows**, while delivering up to **87% throughput gains** to bandwidth-intensive BEAs