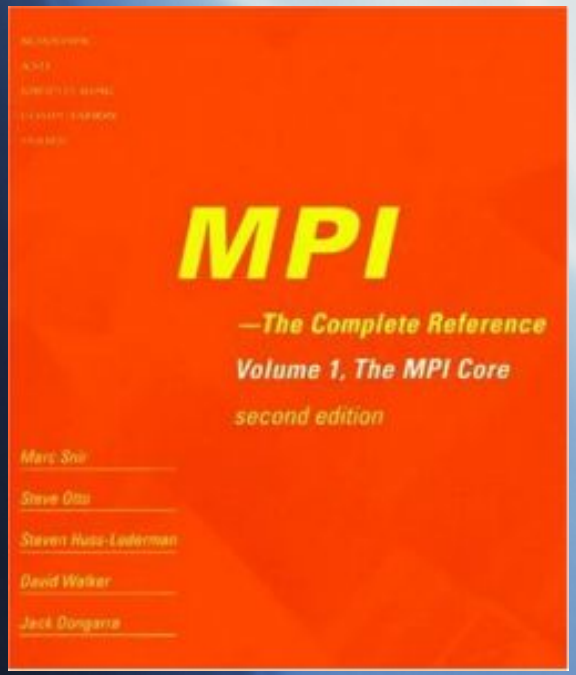
RICE George R. Brown School of Engineering
Computer Science

What MPI and Hadoop can learn from each other!



Vivek Sarkar
E.D. Butcher Chair in Engineering
Professor of Computer Science
Rice University
vsarkar@rice.edu

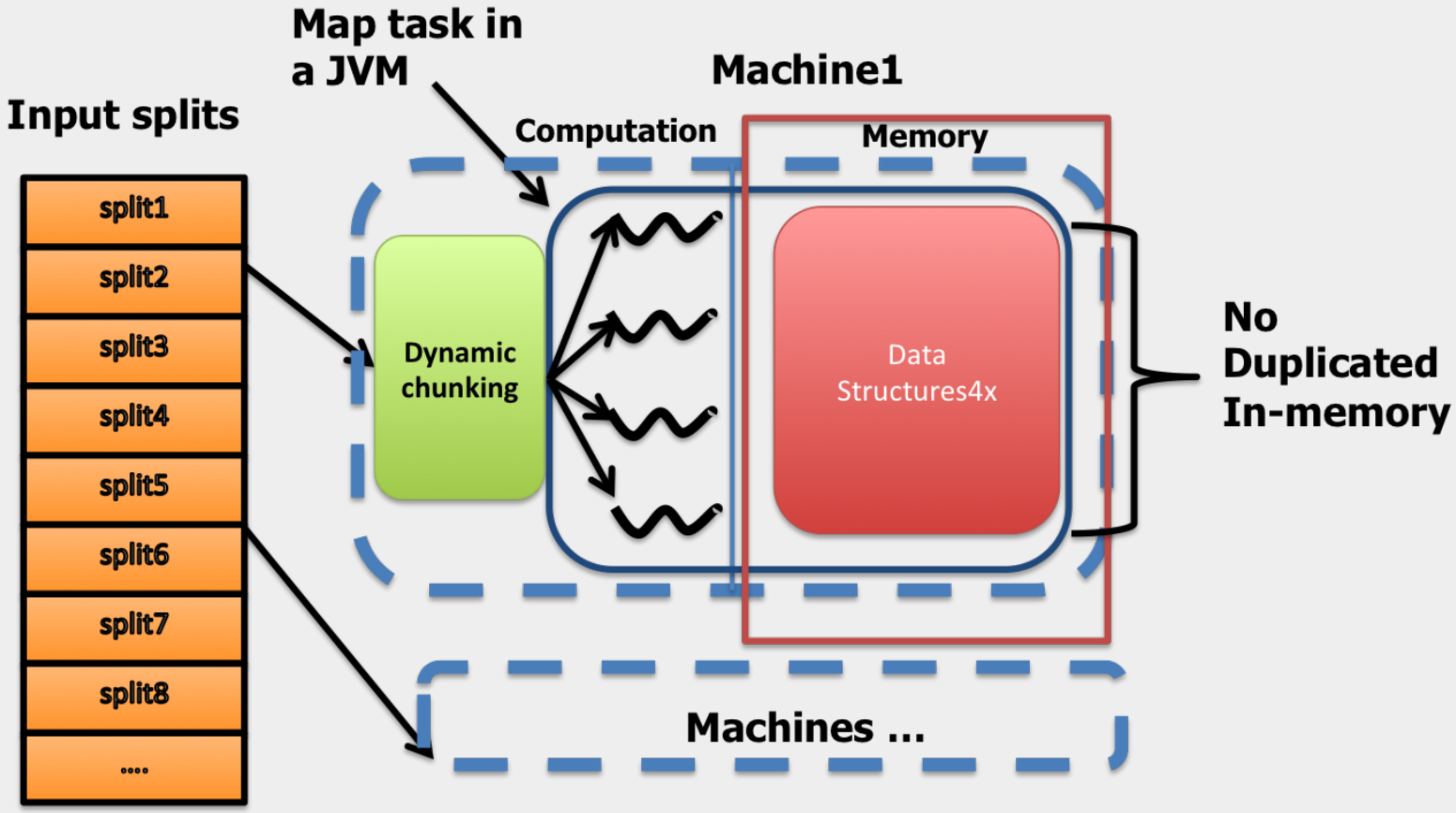
Similarities between Hadoop MapReduce and MPI

- Both provide a simplified programming system for distributed-memory parallelism
- Both support bulk-synchronous parallelism
  - A MapReduce execution consists of independent Map tasks followed by collective Reduce tasks
  - An MPI program (typically) consists of independent phases interleaved with collective calls

Hadoop MapReduce and MPI face complementary and shared challenges

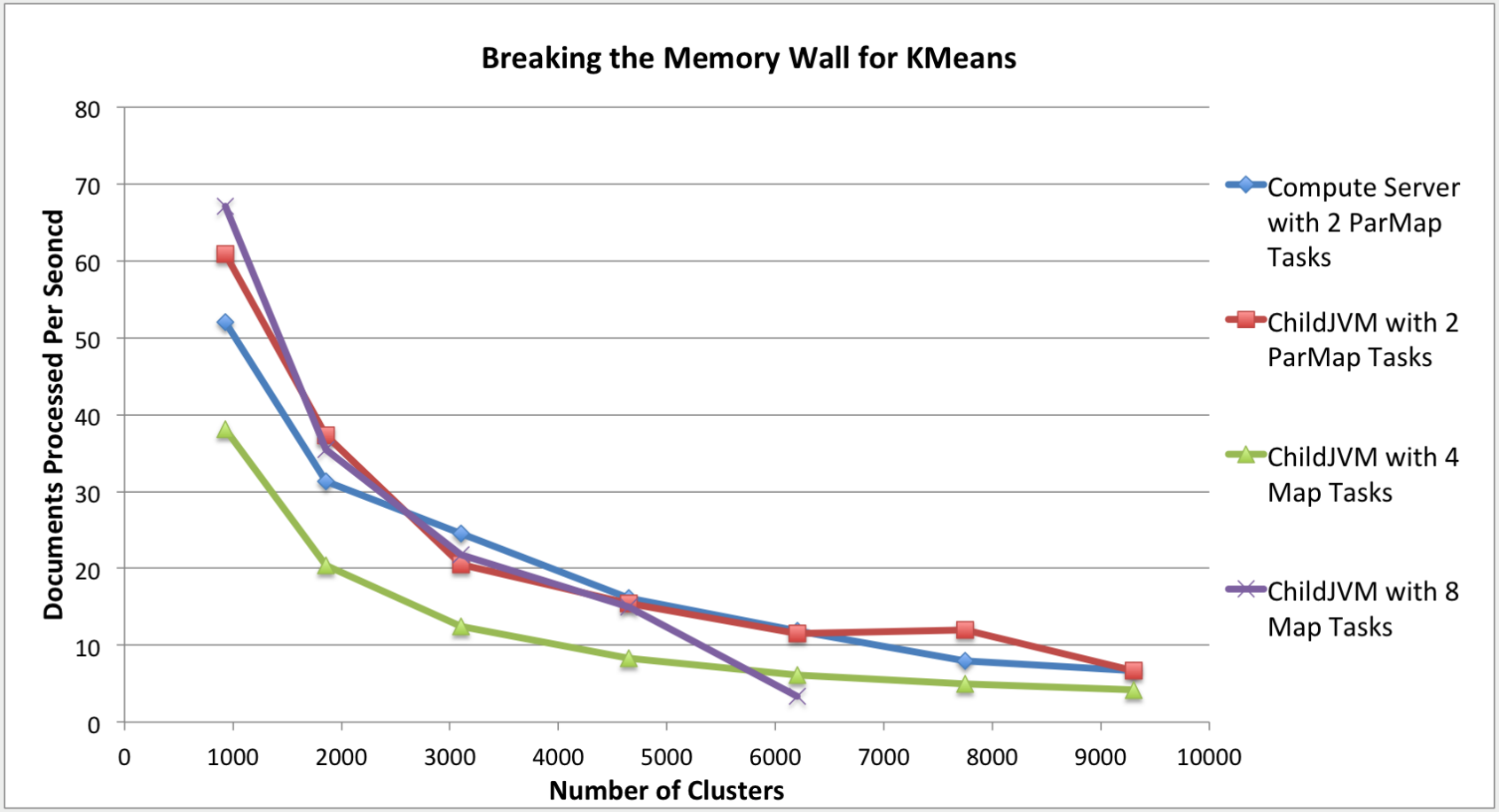
- MapReduce
  - Large overheads for Iterative MapReduce
  - Use of I/O instead of high-performance communication
- MPI
  - Programmability challenges e.g., MPI rank instead of key-value pairs
  - Fault tolerance burdens on programmer
- Both
  - Intra-node strong scaling

Parallel Mapper approach to addressing Strong Scaling Challenge in MapReduce



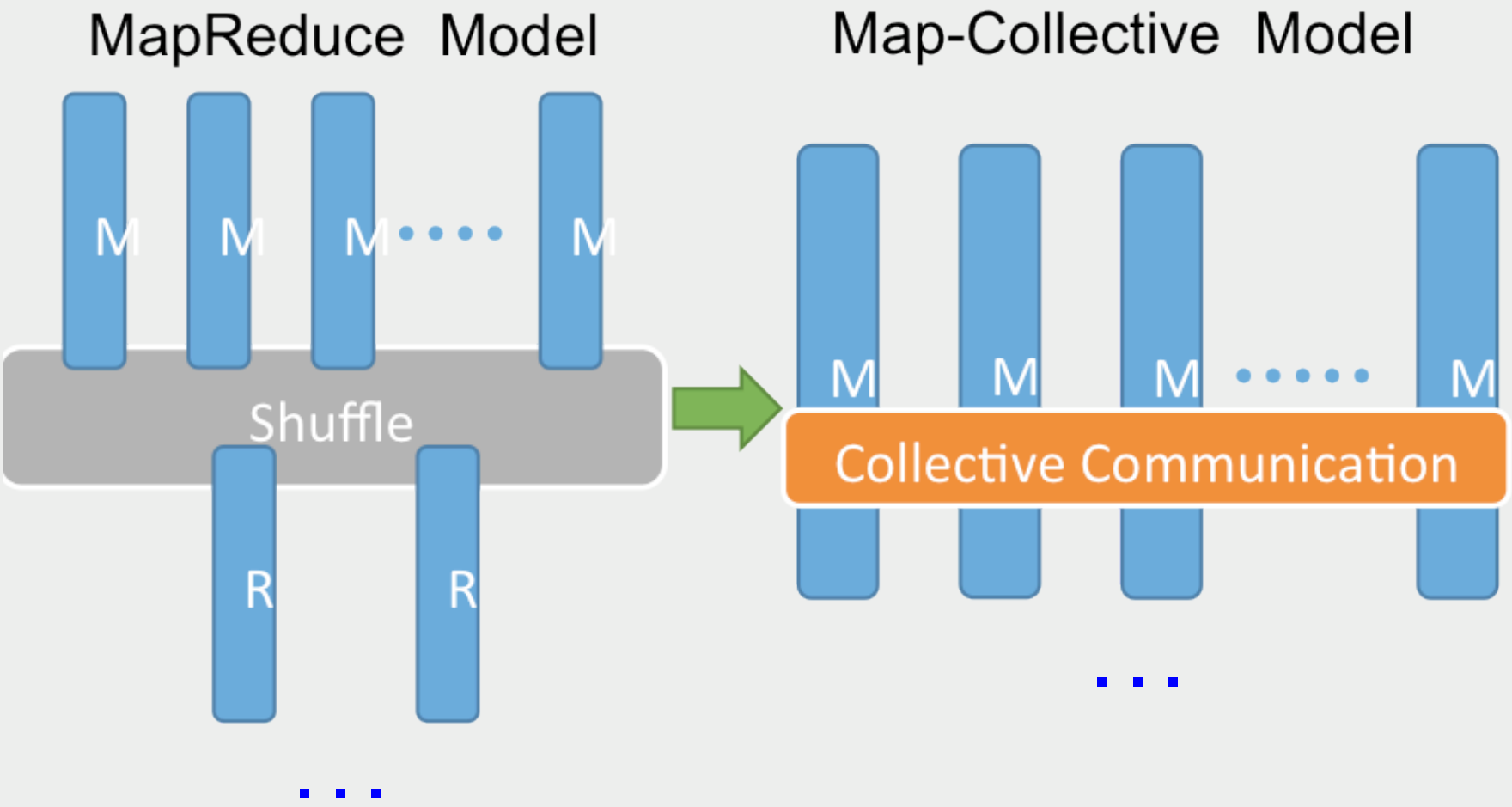
"Optimized runtime systems for MapReduce applications in multi-core clusters"
MS Thesis, Yunming Zhang, Map 2014.

Results using Parallel Mapper and Compute Server extensions



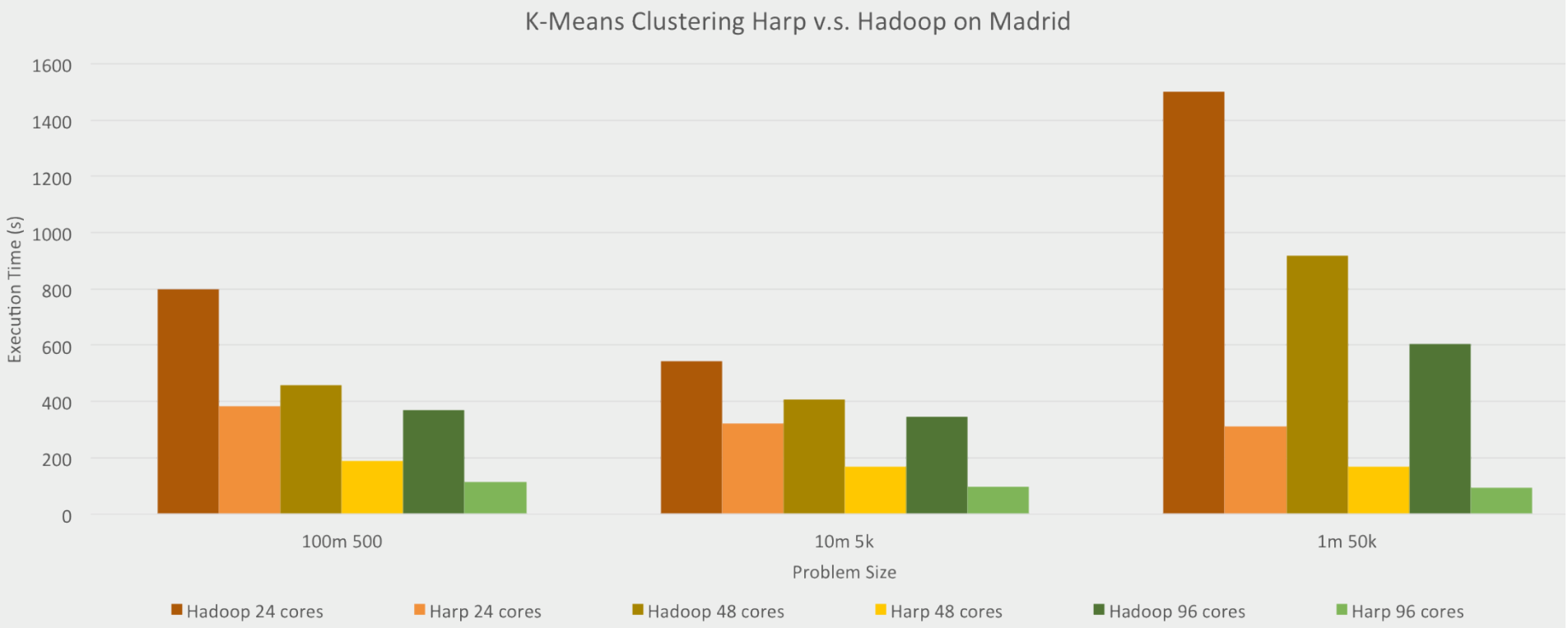
"Optimized runtime systems for MapReduce applications in multi-core clusters"
MS Thesis, Yunming Zhang, Map 2014.

Replacing Shuffle by Collective Communications for Iterative MapReduce



"Harp Collective Collection", Bingjing Zhang, Judy Qiu, Geoffrey Fox.

K-Means Clustering Performance on 8-node Cluster



"Harp Collective Collection", Bingjing Zhang, Judy Qiu, Geoffrey Fox.

Summary

- Significant opportunity for HPC cluster software to influence commercial cloud software and vice versa
- HPC techniques can significantly improve performance of Iterative MapReduce
- MapReduce techniques can improve programmability of HPC software
  - Iterative MapReduce is similar to patterns found in scientific codes
  - Imagine MPI with key-value pairs instead of MPI ranks!