GCLab

Mimir: MapReduce over MPI

Tao Gao, Yanfei Guo, Boyu Zhang, Pietro Cicotti, Yutong Lu, Pavan Balaji, Michela Taufer

Project Overview

Data analytics is an integral part of large-scale scientific computing. MapReduce has gained the most traction. Efforts have been made to enable efficient MapReduce for supercomputing systems, but they are limited to homogeneous workloads. Mimir, a novel MapReduce over MPI framework, tackles skewed data, imbalance in memory usage, and loss in data scalability with (a) combiner optimizations to minimize and balance memory usage; (b) dynamic repartitions to balance memory usage across processes and reduce execution time; and (c) a split method to handle datasets with superkeys. Results show that Mimir can scale to at least 3,072 processes on the Tianhe-2 supercomputer.

Scalability Study

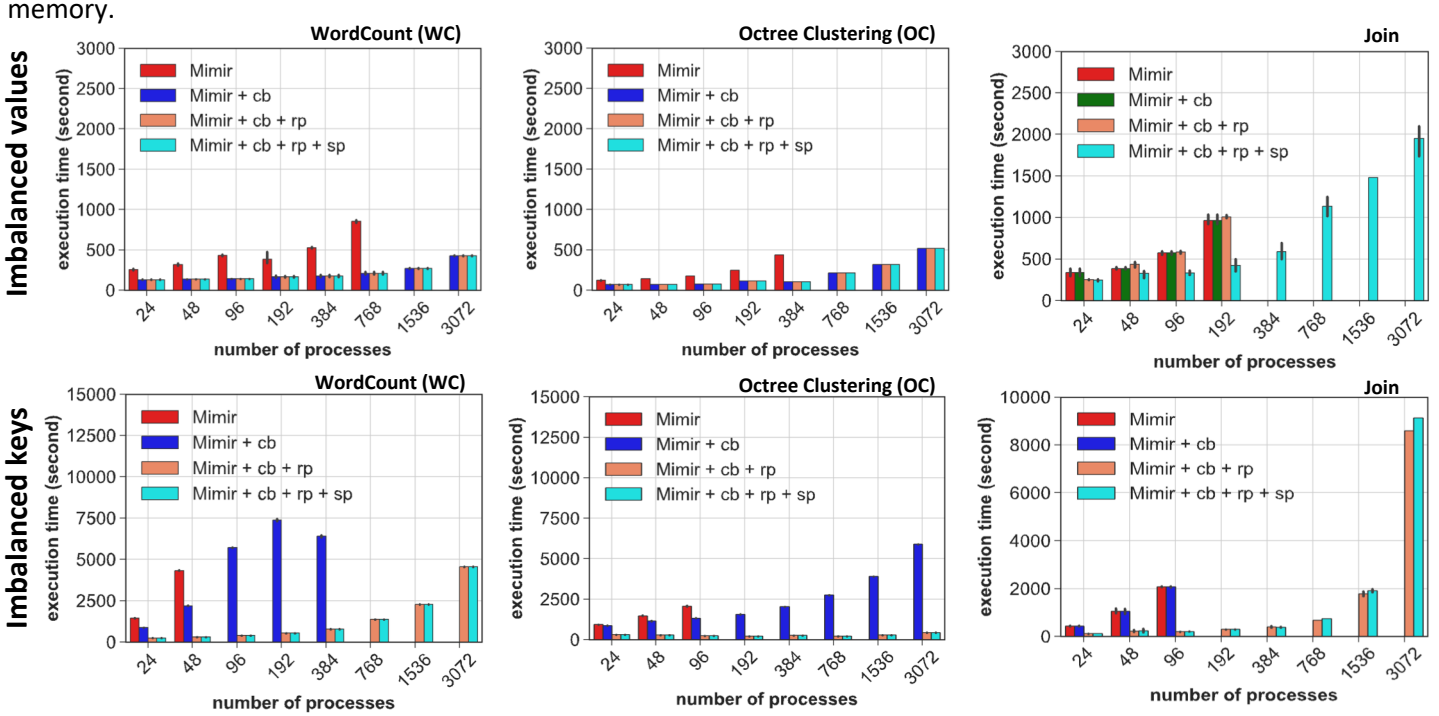
Benchmarks:

- WordCount (WC) – single-pass MapReduce application with associative and commutative reduce function
- Octree Clustering (OC) – iterative chain of MR jobs
- Join – single-pass MR application that merges two imbalance datasets

Tianhe-2: compute node with two Intel Xeon E2-2692v2 CPUs (12 cores each, 24 cores total) running at 2.2 GHz. Each node has 64 GB of memory.

Scalability of Mimir in terms of number of processes for the in-memory workflow, the combiner workflow (cb), the dynamic repartition workflow (rp), and splitting approach (sp). Numbers in red represent the configuration with best performance after which further optimizations do not increase the performance.

	Value			Key mapping		
	WC	OC	Join	WC	OC	Join
Mimir	798	384	192	48	96	96
Mimir+cb	3072	3072	192	384	3072	96
Mimir+cb+rb	3072	3072	192	3072	3072	3072
Mimir+cb+rb+sp	3072	3072	3072	3072	3072	3072



Publications: T. Gao, Y. Guo, B. Zhang, P. Cicotti, Y. Lu, P. Balaji, and M. Taufer: **Mimir: Memory-Efficient and Scalable MapReduce for Large Supercomputing Systems.** IPDPS 2017: 1098-1108.
T. Gao, Y. Guo, B. Zhang, P. Cicotti, Y. Lu, P. Balaji, and M. Taufer. **Scalable Data Skew Mitigation in MapReduce over MPI for Supercomputing Systems.** Journal of IEEE Transactions on Parallel and Distributed Systems (IEEE TPDS), 2019.

THE UNIVERSITY OF TENNESSEE KNOXVILLE