Dr. Ananth Grama
Director of the Computational Science & Engineering Program, Purdue University
Assoc. Director of the Center for Science of Information, A Science & Technology Center of the National Science Foundation

TransMR: Data-Centric Programming Beyond Data Parallelism

MapReduce and related data-centric programming models have proven to be effective for a variety of large-scale distributed computations, in particular, those that manifest data parallelism. Independent execution of tasks in MapReduce, however, limits its applicability to data-parallel applications. This application scope can be increased by allowing data-sharing across concurrent computations. The fault-tolerance model of MapReduce relies on deterministic replay, which makes data-sharing (side-effects) across computations harder to support. This talk: (i) motivates the use of distributed key-value stores (as opposed to distributed file systems) as the underlying storage substrate; (ii) investigates data sharing (side-effects) in programming models operating on distributed key-value stores, specifically, the inconsistencies between the fault recovery mechanisms in execution and storage layers; (iii) defines semantics for a novel programming model, TransMR (Transactional MapReduce), which addresses these inconsistencies; and (iv) demonstrates broad application scope and enhanced performance through data-sharing across computations for a prototype implementation of the proposed semantics. It proposes efficient solutions to a number of problems encountered in supporting transactional execution of maps and broader applicability of these optimizations.

Work in collaboration with Karthik Kambatla, Naresh Rapolu, and Adnan Hassan, supported by the US National Science Foundation.