Acknowledgements

I would like to thank Dr. Srinath Srinivasa, my advisor for making this research possible. Sanket Patil, Mandar Mutalikdesai and Aditya Ramana Rachakonda helped all along with numerous discussions, reviews and comments. Dr. Krithi Ramamritham for providing valuable inputs and Sabishaw B. and Ravi Madipadaga for their critique and encouragement. Finally, I would also like to thank Corporate Technology, Siemens Information Systems Ltd. for funding the research.
Abstract

Stream grids are wide-area grid computing environments that are fed by a set of stream data sources. The two main contributors to stream data generation are (1) the large scale deployment of sensor networks for a wide range of applications from monitoring geophysical activities to supply chain management, and (2) internet applications like financial tickers, transaction logs and network monitoring. Queries arrive at the grid from users and applications external to the system, and seek to tap into one or more data streams. In the literature, queries on stream data are usually classified into two distinct types; a lifetime query which is never revoked and a one-shot query which is transient in nature. The kind of queries considered in this work is of a third type called long-running continuous (LRC) queries, that we also term as “open-world” queries. These queries are neither short-lived nor infinitely long-lived. They live long enough to make the prospect of multi-query optimization meaningful. But queries may also terminate at any time, requiring re-optimization of the query plans. The queries are “open” from the grid perspective as the grid cannot control or predict: (1) arrival of a query with time, location, required data, and (2) query revocation. Query optimization in such an environment has two major challenges: (a) optimizing in a multi-query environment and (b) continuous optimization due to new query arrivals and revocations.

Generating a globally optimal query plan is shown to be an intractable problem. This is compounded by the fact that the query plan would need to be recomputed frequently, with every new query arrival and existing query revocation. Hence, this work explores the idea of emergent optimization, where globally optimal query plans emerge as a result of local autonomous decisions taken by the grid nodes. Drawing concepts from evolutionary game theory, grid nodes are modeled as autonomous agents that seek to maximize a self-interest function using one of a set of different strategies. A node autonomously changes strategies based on: (1) variations in query arrival and revocation patterns in the grid, and (2) how its strategy is faring compared to the strategies of other nodes in the grid. The ex-
experiments performed show that by adopting emergent optimization, the grid as a whole is able to not only select the correct optimization strategy, but also adapt to varying query conditions.

While the work presented in this thesis illustrates the complexities involved in building a solution for open-world query processing, it may be noted without loss of generality that open-world systems, per se, by virtue of their uncertain environment, offer challenges significantly different from their algorithmic counterparts, where the environment is completely described apriori. In general, most real-world systems can be classified as open-world systems involving a large number of actors. Some explorations into possible theoretical foundations of open-world systems was also performed as part of this work.