Cooperative Computing in Dynamic Environments MIT9904-12

Progress Report: January 1, 2002—June 30, 2002

Nancy Lynch and Idit Keidar

Project Overview

The Theory of Distributed Systems group at MIT pursues research on the development of models, analysis and verification methods, and algorithms for distributed systems. Most of its recent work focuses on systems that are highly dynamic, where participants may join and leave the system and may change location. Moreover, the network topology may change, and components may fail and recover. We address the problems brought about by such difficult environments

(1) by developing formal modeling and analysis techniques based on interacting state machines; and
(2) by developing useful "building blocks" for dynamic systems---definitions of global services and efficient algorithms to implement them.

The work on formal modeling and analysis involves extending the basic I/O automaton model to support new features such as dynamic process creation, mobility, timing, and continuous behavior. It includes the development of new methods for analyzing performance and fault-tolerance properties. We are also developing the high-level distributed programming language IOA and a toolkit to facilitate the design, analysis, and verification of systems using the techniques developed as a result of our theoretical work. The work on services and algorithms focuses on high-level communication and data-management services, including services that support dynamic reconfiguration.

Progress Through June 2002

(1) Formal modeling and analysis

The last six months have been particularly productive for the IOA project in terms of publishing our research results and passing some milestones in our implementation work. The first paper accepted for publication in this period is by Bogdanov, Garland, and Lynch [BGL02]. It discusses issues related to the mechanical translation of nondeterministic specifications into first-order logic using the ioa2lsl tool developed by Bogdanov. This tool translates IOA programs into Larch Shared Language specifications in a style that is suitable for formal reasoning. A second paper, by Kirli, Chefter, Dean, Garland, Lynch, Ne Win, and Ramirez [KCDGLNR02], focuses on the IOA simulator. It includes a case study in which a simulation relation between a mutual exclusion algorithm and its specification is checked with the simulator and subsequently proved correct with LP. This case study shows how the paired simulator, the ioa2lsl tool, and the dynamic invariant detection tool Daikon may be used together to simplify the verification process. A third publication of the IOA project, which draws on the above-mentioned two papers and several MEng theses [Ch98,De01,Ra00], is a comprehensive manual for the IOA simulator. It is intended to serve as a guide for both users and developers of the IOA simulator.

In terms of implementation work, we have completed a preliminary release of the IOA toolkit and made the software available on our Web pages (http://theory.lcs.mit.edu/tds/ioa/). This release does not contain the code generator, because we are still working on the detailed design of the composer, which is a part of the code generator. The composer will automatically transform a composite automaton into an equivalent primitive

automaton. Tauber and Garland are writing a comprehensive document that describes all the steps in this transformation.

We are interested in constructing interfaces between IOA and a number of theorem provers, so as to make IOA accessible to a larger user community and to benefit from ongoing work in automated deduction outside of MIT. This has motivated us to start developing a tool that is similar to Bogdanov's but targets Isabelle, an interactive theorem prover developed at Cambridge University. Luhrs is currently writing a specification for the design of this translation tool.

Kawabe, a visitor to our group from NTT, has recently completed a project in which he used the IOA toolkit to verify safety properties of the implementation of the Nepi2 network programming system, developed by Mano and Kawabe at NTT. Kawabe coded the abstract specification and the distributed implementation of Nepi2 in IOA, and used LP to prove that the distributed implementation is correct with respect to the abstract specification. Kawabe's proof is the largest that has been carried out so far using the IOA toolkit. A paper summarizing his results will be presented at the *Workshop on Foundations of Software Engineering (FOSE '02)*.

Mitra has developed a language called HIOA for specifying hybrid systems, based on the Hybrid Input/Output Automaton [LSV02] modeling framework. In this language, the discrete transitions of an automaton are specified in the precondition-effect style of the IOA language. Trajectories are specified using ``state models'', a natural representation derived from work in the field of dynamical systems. Work has continued also on DIOA, aimed at a final journal publication. The DIOA model supports trace pasting and substitutivity results, which will support compositional design and levels of refinement.

(**2) Algorithms for dynamic distributed systems**:

Bar-Joseph, Keidar, and Lynch completed a technical report [BKL02a] and conference paper [BKL02b] on dynamic atomic broadcast; this work will be presented in DISC '02. In this work, they introduce a new problem of atomic broadcast in a dynamic setting where processes may join, leave voluntarily, or fail (by stopping) during the course of computation. They provide a formal definition of the Dynamic Atomic Broadcast problem and present and analyze a new algorithm for its solution. The algorithm exhibits constant message delivery latency in the absence of failures, even during periods when participants join or leave. When failures occur, the latency bound is linear in the number of actual failures. These bounds improve upon previously suggested algorithms solving similar problems in the context of view-oriented group communication. Their algorithm uses a solution to a variation on the standard distributed consensus problem, in which participants do not know a priori who the other participants are. They define the new problem, which they call Consensus with Unknown Participants, and give an early-stopping algorithm to solve it.

Lynch and Shvartsman have completed a conference paper [LS02a], and have nearly completed a technical report [LS02b], on a new algorithm that emulates atomic read/write shared objects in a dynamic network setting. The algorithm is called RAMBO: Reconfigurable Atomic Memory for Basic Objects. To ensure that the data is highly available and long-lived, each object is replicated at several network locations. To ensure atomicity, reads and writes are performed using "quorum configurations", each of which consists of a set of members plus sets of read-quorums and write-quorums. The algorithm is reconfigurable: the quorum configuration is allowed to change during computation, and such changes do not cause violations of atomicity. Any quorum configuration may be installed at any time---no intersection requirement is imposed on the sets of members or on the quorums of distinct configurations. The algorithm tolerates processor and link failures.

The algorithm performs three major activities, all concurrently:

(1) reading and writing the objects,
(2) choosing new configurations and notifying members, and
(3) identifying and removing ("garbage-collecting") obsolete configurations.

The algorithm is composed of two sub-algorithms: a main algorithm, which handles reading, writing, and garbage-collection, and a reconfiguration algorithm, which handles the selection and dissemination of new configurations.

The main safety property, atomicity, holds for arbitrary patterns of asynchrony. Performance properties depend on particular failure and timing assumptions. In particular, if participants gossip periodically in the background, if garbage-collection is scheduled periodically, if reconfiguration is not requested too frequently, and if quorums of active configurations do not fail, then read and write operations complete within time that is proportional to the maximum message latency.

In related work, Seth Gilbert is trying to improve the performance of the RAMBO algorithm in several ways, most significantly, by increasing the concurrency of garbage-collection operations. Two LAN implementations of RAMBO are also in progress. Gilbert and Lynch have also written a short technical note [GL02] formalizing and proving a conjecture made by Brewer at *PODC 2000*, about the impossibility of implementing atomic data in a partionable network setting. Also, Rui Fan is working on an MS thesis [Fa02] which contains a network implementation of atomic objects that separates the handling of the data from the quorum management.

In July 2002, Keidar and Rajsbaum presented a tutorial based on [KR01], at *PODC '02*. The tutorial discusses the performance of fault tolerant consensus algorithms in synchronous failure-free runs. It also includes a new lower bound proof, which appears in the recently accepted paper [KR02].

At the same conference, Keidar and Bakr presented their work on performance evaluation of distributed algorithms deployed in a widely distributed setting over the Internet using TCP [BK02]. In this work they consider a simple primitive that corresponds to a communication round in in which every host sends information to every other host. They present the results of experiments with four algorithms that are typically used to implement this primitive. They observe that message loss has a large impact on algorithm running times, which causes leader-based algorithms to usually outperform decentralized ones.

In June 2002, Keidar presented a white paper about challenges in evaluating distributed algorithms [Ke02].

Our work on reliable multicast continued with Livadas finalizing his formal specifications of the reliable multicast service, the Scalable Reliable Multicast (SRM) protocol, and a caching-enhanced version of SRM (CESRM). He proved that the model of SRM is a faithful implementation of the reliable multicast service specification with no timeliness guarantees. Subsequently, Livadas defined a set of constrained executions of SRM; namely, the admissible timed executions in which: i) hosts neither leave the reliable multicast group nor crash, ii) inter-host transmission latencies and their estimates are bounded, and iii) the time to detect the loss of any packet is bounded. Livadas has shown that these executions of SRM are ``live'', in the sense that in any such execution, SRM guarantees the timely delivery of the appropriate packets to the appropriate members of the reliable multicast group. The recent results of his work have been accepted to FORTE '02 [LL02].

Khazan completed his work on the performance analysis of his group communication service (GCS) [KK00], which provides Virtual Synchrony (VS) semantics. He derived an upper bound on the time from when the final network event occurs until all clients of the network component receive the final view. An important contribution of this work is that it uses a compositional approach; the performance properties of GCS are obtained by combining those of the system components. Khazan also illustrated the utility of his GCS system by defining and analyzing a simple data-management application that can be built using GCS [Kh02]. The application allows a dynamically-changing group of clients to access and modify the data. It guarantees ``interim atomicity'' which means that, while the underlying network component is stable, clients perceive the data object as atomic. The details of this work, along with the results on performance analysis, appear in Khazan's recently-completed PhD thesis [Kh02].

Research Plan for the Next Six Months

(1) Formal modeling and analysis

We expect the detailed design of the IOA composer to be completed soon. This will enable Garland to finish the static checks related to composition in the IOA front end, and Tauber to finish the implementation of the IOA code generator tool.

By the end of this summer, Ne Win will start coding the tool for IOA to Isabelle translation based on Luhrs' design. We are planning to continue our analysis work by conducting experiments in reasoning about distributed algorithms. Our aims are to determine which features of which theorem provers work best for supporting which types of reasoning, and to enhance the interfaces to those tools to support the use of these features. We may add ACL2 to the collection of theorem provers that are interfaced with IOA.

As future work, Kawabe is interested in designing and implementating a verification tool tailored for Nepi2 programs. Garland and Kirli are prepared to collaborate with him on this project.

We will evaluate Mitra's HIOA language by specifying hybrid system examples that are of interest to our group, for example, automated transportation systems and mobile and embedded systems. Also, Mitra plans to adapt and extend stability and convergence results from control theory to hybrid systems and to the HIOA framework. Other plans for the next six months include enhancing the IOA language with constructs for specifying timing behavior. We also intend to finish up a paper on DIOA for journal submission.

(**2) Algorithms for dynamic distributed systems**:

We will continue our work on RAMBO by analyzing its performance in a larger number of situations, including situations in which the timing and behavior stabilizes from some point onward. We will work on optimizations, including concurrent garbage collection and pre-releasing values to speed up reads. Pre-releasing values will change the user-viewable behavior to something slightly weaker than atomicity; what are the weaker guarantees? We will continue our work on implementations. Other related work includes studying and analyzing methods for choosing new configurations. Fan will continue his work on separating the handling of data from quorum management; his algorithm may also be extended to allow reconfiguration. Finally, after we have acquired enough understanding of the possible optimizations, it would be interesting to prove lower bound results that say that the remaining costs are inherent.

In the next six months, Livadas intends to remove, from the liveness analysis of SRM, the constraint that hosts neither leave the reliable multicast group nor crash. This constraint relaxation involves the analysis of the performance of SRM in group membership environments, which is becoming increasingly important due to the increase in host mobility. Livadas also intends to analyze the correctness and the performance of the caching-enhanced algorithm CESRM and to compare its performance to that of SRM.

Lynch and Stoica are beginning a project on designing a fault-tolerant overlay network algorithm for a dynamically-changing wide-area network. Our algorithm is based on the Chord algorithm of Karger et al., but adds extra redundancy features for fault-tolerance. We plan to simulate and analyze the resulting design, and compare the experimental and theoretical results.

We are interested in designing new algorithms for resource allocation in highly dynamic networks. These include algorithms that bound the number of participants that can have simultaneous access to a resource, or that ensure mutual exclusion between groups of participants. We believe that such algorithms would be beneficial for systems with limited resources and for systems where users may be classified into different groups in terms of their access rights to resources. We hope to collaborate with the Networks and Mobile Systems Group at MIT to identify the challenges in this research direction.

Citations:

[BGL02] Andrej Bogdanov, Stephen Garland, and Nancy Lynch. Mechanical Translation of I/O Automaton Specifications into First-Order Logic. To appear in *the 22nd IFIP WG 6.1 International Conference on Formal Techniques for Networked and Distributed Systems (FORTE 2002),* Houston, Texas, November 2002.

[BK02] Omar Bakr and Idit Keidar. Evaluating the Running Time of a Communication Round over the Internet. In *Proceedings of the 21st ACM Symposium on Principles of Distributed Computing (PODC)*, July 2002, pages 243--252.

[BKL02a] Ziv Bar-Joseph and Idit Keidar and and Nancy Lynch. Real-Time Dynamic Atomic Broadcast. MIT Laboratory for Computer Science Technical Report, LCS-TR-840, 2002.

[BKL02b] Ziv Bar-Joseph and Idit Keidar and and Nancy Lynch. Real-Time Dynamic Atomic Broadcast. To appear in *Proceedings of the 16th International Symposium on DIStributed Computing (DISC),* October,2002, Toulouse, France.

[Ch98] Anna E. Chefter. A Simulator for the IOA Language Master of Engineering and Bachelor of Science in Computer Science and Engineering Thesis, Massachusetts Institute of Technology, Cambridge, MA, May 1998.

[De01] Laura G. Dean. Improved Simulation of Input/Output Automata. Master of Engineering Thesis, Massachusetts Institute ofTechnology, Cambridge, MA, September 2001.

[Fa02] Rui Fan. MS thesis in progress.

[GL02] Seth Gilbert and Nancy Lynch. Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services. In *Sigact News* 33(2), June, 2002.

[Ke02] Idit Keidar. Challenges in Evaluating Distributed Algorithms. In *Proceedings of International Workshop on Future Directions in Distributed Computing (FuDiCo),* Pages 22--25, Bertinoro, Italy, June, 2002.

[Kh02] Roger Khazan. *A One-Round Algorithm for Virtually Synchronous Group Communication in Wide Area Networks*. Ph.D. dissertation. Department of Electrical Engineering and Computer Science. MIT. May 22, 2002.

[KCDGLNR02] Dilsun Kirli, Anna Chefter, Laura Dean, Stephen Garland,
Nancy Lynch, Toh Ne Win, and Antonio Ramirez. Simulating Nondeterministic Systems at Multiple Levels of Abstraction. To be presented at *Tools Day held in conjunction with CONCUR '02*, 2002.

[KCDGLNR02] Dilsun Kirli, Anna Chefter, Laura Dean, Stephen Garland, Nancy Lynch, Toh Ne Win, and Antonio Ramirez. The IOA Simulator. Technical Report MIT-LCS-TR-843, MIT Laboratory for Computer Science, Cambridge, MA, 2002.

[KK00] Idit Keidar and Roger Khazan. A Virtually Synchronous Group Multicast Algorithm for WANs: Formal Approach. To appear *SIAM Journal on Computing*.

[KR01] Idit Keidar and Sergio Rajsbaum, On the Cost of Fault-Tolerant Consensus
When There Are No Faults -- A Tutorial, 2001, Also published as MIT-LCS-TR-821, Preliminary version in SIGACT News 32(2), pages 45--63, June 2001 (published May 15th 2001).

[KR02] Idit Keidar and Sergio Rajsbaum, A Simple Proof of the Uniform Consensus Synchronous Lower Bound. To appear in *Information Processing Letters*, 2002.

[LL02] Carolos Livadas and Nancy A. Lynch, A Formal Venture into Reliable Multicast Territory, Formal Techniques for Networked and Distributed Systems. To appear in *the 22nd IFIP WG 6.1 International Conference on Formal Techniques for Networked and Distributed Systems (FORTE 2002),* Houston, Texas, November 2002.

[LS02a] Nancy Lynch and Alex Shvartsman. RAMBO: A Reconfigurable Atomic Memory Service for Dynamic Networks. To appear in *Proceedings of the 16th International Symposium on DIStributed Computing (DISC)*, Toulouse, France, October, 2002.

[LS02b] Nancy Lynch and Alex Shvartsman. RAMBO: A Reconfigurable Atomic Memory Service for Dynamic Networks. Technical Report MIT-LCS-TR-856, MIT Laboratory for Computer Science Technical Report, 2002.

[LSV02] Nancy Lynch and Roberto Segala and Frits Vaandraager. Hybrid I/O Automata. Submitted for journal publication.

[Ra00] Antonio Ramirez-Robredo. Paired Simulation of I/O Automata. Masters thesis. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, September, 2000.