Multicasting and Routing Services in the NII and Networked Environments with Imbedded Computing
Final Technical Report
Grant Number F19628-96-C-0038

J.J. Garcia-Luna-Aceves, Anujan Varma
University of California, Santa Cruz
Santa Cruz, California 95064

Contents

1 INTRODUCTION

The Internet is experiencing the widespread adoption of applications that support real-time reliable or unreliable distribution of multimedia information to multiple destinations (e.g., *nv, vat, NeVot, wb*, and *shdr*). In addition, with the continuing decrease in hardware costs, very simple or special-purpose devices will soon have computing and communications capabilities of their own and enjoy internetwork addressability traditionally assigned to mainframes, workstations, and more recently personal computers. As a result, very simple devices will become reachable through the Internet; furthermore, traditional computing, storage, and media-distribution resources will be implementable by aggregating such simpler devices (e.g., a traditional Internet host becomes a network of devices). This vision is already being addressed in a number of projects, such as ISI's Netstation project. We refer to this next architectural step in the evolution of the Internet as the *networked imbedded-computing environment* (NICE).

NICE implies the interconnection to the national information infrastructure of vast numbers of internetwork-addressable devices and traditional networked resources. The emergence of NICE and new Internet real-time multimedia applications presents many new challenges on multicasting and routing. First, there are many more destinations that must be reached and many more routing nodes that have to be involved in the delivery of information to destinations. Second, multipoint communication plays a key role in communicating with very small devices and sensors. Third information sent from many sources to many destinations must be delivered reliably. Fourth, emerging distributed applications require more end-to-end support than simply multicasting of data packets.

The University of California at Santa Cruz (UCSC) addressed these challenges by focusing on the following topics:

- Scalable routing architectures and protocols for enabling routing for any number of destinations while maintaining relatively small routing tables.
- Routing architectures and protocols for routing over multiple paths without incurring permanent lopping of packets.
- Architectures and protocols that support quality-of-service guarantees in a scalable manner.
- Scalable architectures and protocols for one-to-many and many-to-many communication, including secure multicasting.
- Scalable architectures and protocols for end-to-end reliable multicasting.
- Scalable architectures and protocols for supporting group collaboration on an end-to-end basis.
- Congestion control solutions applicable to one-to-one and one-to-many communication.

The research work in this project resulted in 37 refereed papers published in journals and conferences, seven Ph.D. theses, and four M.S. theses.

The theses completed with support from this project are the following:

1. Srinivas Vutukury, "Multipath Routing Mechanisms for Traffic Engineering and Quality of Service in The Internet," PhD Thesis, Computer Science, University of California, Santa Cruz, March 2001.

2. Hans-Peter Dommel, "Group Coordination Support in Networked Multimedia Systems," Ph.D. Thesis, Computer Engineering, University of California, Santa Cruz, December 1999.

3. Brian Levine, "Supporting Large-Scale Group Communication Applications on The Internet," Ph.D. Thesis, Computer Engineering, University of California, Santa Cruz, June 1999.

4. Clay Shields, "Secure Hierarchical Multicast Routing and Multicast Internet Anonymity," Ph.D. Thesis, Computer Engineering, University of California, Santa Cruz, June 1999.

5. Mehrdad Parsa, "Multicast Routing in Computer Networks," Ph.D. Thesis, Computer Engineering, University of California, Santa Cruz, June 1998.

6. Jochen Behrens, "Distributed Routing for Very Large Networks Based on Link Vectors," Ph.D. Thesis, Computer Engineering, University of California, Santa Cruz, 1997.

7. Lampros Kalampoukas, "Congestion Management in High Speed Networks," Ph.D. Thesis, Computer Engineering, University of California, Santa Cruz, August 1997.

8. Bradley Smith, "Securing Distance-Vector Routing Protocols," M.S. Thesis, Computer Engineering, University of California, Santa Cruz, June 1997.

9. Diamatis Kourkouzelis, "Multipath Routing Using Diffusing Computations," M.S. Thesis, Computer Engineering, University of California, Santa Cruz, March 1997.

10. Brian Levine, "A Comparison of Known Classes of Reliable Multicast Protocols," M.S. Thesis, Computer Engineering, University of California, Santa Cruz, June 1996.

11. Clay Shields, "Ordered Core Based Trees," M.S. Thesis, Computer Engineering, University of California, Santa Cruz, June 1996.

The published articles describing the results of our research in this project are the following:

1. S. Vutukury and J.J. Garcia-Luna-Aceves, "MPATH: A Loop-free Multipath Routing Algorithm," to appear in *Microprocessors and Microsystems Journal*, Elsevier, 2001.

2. H.P. Dommel and J.J. Garcia-Luna-Aceves, "Multisite Coordination in Shared Multicast Trees," selected for publication in Special Issue of *The Journal of Supercomputing* (Kluwer), 2001.

3. H.-P. Dommel and J.J. Garcia-Luna-Aceves, "A Coordination Framework and Architecture for Internet Groupwork," *Journal of Network and Computer Applications (JNCA)*, Special Issue on Support for Open and Distance Learning on the WWW, Academic Press pub, Fall 2000.

4. C. Shields and J.J. Garcia-Luna-Aceves, "HIP–A Protocol for Hierarchical Multicast Routing," *Computer Communications*, Vol. 23, No. 7, pp. 628-641, Elsevier, 2000.

5. S. Vutukury and J.J. Garcia-Luna-Aceves, "A Practical Framework for Minimum-Delay Routing in Computer Networks," *Journal of High Speed Networks*, Vol. 8, No. 4, pp. 241-263, Wiley, 1999.

6. H.-P. Dommel and J.J. Garcia-Luna-Aceves, "Efficacy of Floor Control Protocols in Distributed Multimedia Collaboration," *Cluster Computing* (Baltzer Sci. Pub.), Special Issue on Multimedia Collaborative Environments, Vol. 2, No. 1, 1999.

7. H. P. Dommel and J. J. Garcia-Luna-Aceves "Group Coordination Support for Internet Collaboration," *IEEE Internet Computing Magazine*, Special Issue on Multimedia and Collaborative Computing over the Internet, Vol. 3, No. 2, pp. 74-80, March/April 1999.

8. M. Parsa, Q. Zhu, and J.J. Garcia-Luna-Aceves, "An Iterative Algorithm for Delay-Constrained Minimum-Cost Multicasting," *IEEE/ACM Transactions on Networking,* Vol. 6, No. 4, August 1998.

9. B.N. Levine and J.J. Garcia-Luna-Aceves, "A Comparison of Reliable Multicast Protocols," *ACM Multimedia Systems Journal,* Vol. 6, No. 5, August 1998.

10. S. Murthy and J.J. Garcia-Luna-Aceves, "A "A Loop-Free Routing Protocol for Large-Scale Internets Using Distance Vectors," *Computer Communications,* Vol. 21, No. 2, 1998, pp. 147–161.

11. B. Smith and J.J. Garcia-Luna-Aceves, "Efficient Security Mechanisms for The Border Gateway Routing Protocol," *Computer Communications Journal,* Vol. 21, No. 3, pp. 203–210, 1998.

12. M. Parsa and J.J. Garcia-Luna-Aceves, "A Protocol for Scalable Loop-Free Multicast Routing," *IEEE Journal on Selected Areas in Communications,* Vol. 15, No. 3, pp. 316–331, April 1997.

13. H.-P. Dommel and J.J. Garcia-Luna-Aceves, "Networking Foundations for Collaborative Computing at Internet Scope," *International ICSC Congress on Intelligent Systems and Applications*, Symposium on Interactive and Collaborative Computing (ICC'2000), Wollongong, Australia, December 12 - 15, 2000.

14. S. Vutukury and J.J. Garcia-Luna-Aceves, "A Multipath Framework Architecture for Integrated Services," *Proc. IEEE Globecom 2000*, San Francisco, California, November 27–December 1, 2000.

15. S. Vutukury and J.J. Garcia-Luna-Aceves, "A Traffic Engineering Approach based on Minimum-delay Routing," *Proc. IEEE IC3N 2000*, Las Vegas, Nevada, October 16–18, 2000.

16. H.P. Dommel and J.J. Garcia-Luna-Aceves, "A Coordination Architecture for Internet Groupwork," *Proc. 26th EUROMICRO Conference*, Workshop on Multimedia and Telecommunications, Maastricht, The Netherlands September 4–7, 2000.

17. B. Levine, J. Crowcroft, C. Diot, J.J. Garcia-Luna-Aceves, and J. Kurose, "Consideration of Receiver Interest for IP Multicast Delivery," *Proc. Infocom 2000*, Tel-Aviv, Israel, March 26–30, 2000.

18. H.P. Dommel and J.J. Garcia-Luna-Aceves, "Ordered End-to-End Multicast for Distributed Multimedia Systems," *Proc. Hawaii International Conference on System Sciences (HICS-33)*, Maui, Hawaii, January 4–7, 2000.

19. S. Vutukury and J.J. Garcia-Luna-Aceves, "A Distributed Algorithm for Multipath Computation," *Proc. IEEE Globecom '99*, Rio de Janeiro, Brazil, December 5–9, 1999.

20. S. Vutukury and J.J. Garcia-Luna-Aceves, "A Scalable Architecture for Providing Deterministic Guarantees," *Proc. IEEE IC3N 99*, Boston, Massachusetts, October 11–13, 1999.

21. S. Vutukury and J.J. Garcia-Luna-Aceves, "An Algorithm for Multipath Computation using Distance-Vectors with Predecessor Information," *Proc. IEEE IC3N 99*, Boston, Massachusetts, October 11–13, 1999.

22. C. Shields and J.J. Garcia-Luna-Aceves, "A Scalable Protocol for Secure Multicast Routing," *Proc. ACM SIGCOMM 99*, Cambridge, Massachusetts, September 1–3, 1999.

23. S. Vutukury and J.J. Garcia-Luna-Aceves, "A Simple Approximation to Minimum-Delay Routing," *Proc. ACM SIGCOMM 99*, Cambridge, Massachusetts, September 1–3, 1999.

24. H.P. Dommel and J.J. Garcia-Luna-Aceves, "Multisite Coordination in Shared Multicast Trees," *Proc. International Conference on Parallel and Distributed Processing Techniques and Applications* (PDPTA'99), Las Vegas, NV, June 28–July 1, 1999.

25. J.J. Garcia-Luna-Aceves, S. Vutukury, and W.T. Zaumen, "A Practical Approach to Minimizing Delays in Internet Routing Protocols," *Proc. IEEE ICC '99*, Vancouver, Canada, June 6–10, 1999.

26. H.-P. Dommel and J.J. Garcia-Luna-Aceves, "Comparison of Floor Control Protocols for Collaborative Multimedia Environments," *Proc. SPIE Symposium on Voice, Video, and Data Communications*, Boston, MA., November 2–5, 1998,

27. H.-P. Dommel and J.J. Garcia-Luna-Aceves, "A Novel Group Coordination Protocol for Distributed Multimedia Group Collaboration", *Proc. 1998 IEEE International Conference on Systems, Man, and Cybernetics,* San Diego, California, October 11-14, 1998.
(**WINNER OF BEST STUDENT PAPER AWARD**)

28. B.N. Levine, S. Paul, and J.J. Garcia-Luna-Aceves, "Organizing Multicast Receivers Deterministically According to Packet-Loss Correlation," *Proc. ACM Multimedia 98: Sixth ACM International Multimedia Conference,* Bristol, UK, September 1998.

29. C. Shields and J.J. Garcia-Luna-Aceves, "The HIP Protocol for Hierarchical Multicast Routing," *Proc. ACM PODC '98: Seventeenth ACM Symposium on Principles of Distributed Computing,* Puerto Vallarta, Mexico, 28–July 2, 1998.

30. J. Behrens and J.J. Garcia-Luna-Aceves, "Hierarchical Routing Using Link Vectors," *Proc. IEEE INFOCOM '98,* San Francisco, California, March 29–April 2, 1998.

31. L. Kalampoukas, A. Varma and K. K. Ramakrishnan, "Improving TCP Throughput over Two-Way Asymmetric Links: Analysis and Solutions," *Proc. ACM Sigmetrics '98*, June 1998.

32. B.N. Levine and J.J. Garcia-Luna-Aceves, "Improving Internet Multicast Using Routing Labels," *Proc. IEEE ICNP '97: Fifth IEEE International Conference on Network Protocols,* Atlanta, GA, October 1997.

33. J. Behrens and J.J. Garcia-Luna-Aceves, "Fast Dissemination of Link State Updates Using Bounded Sequence Numbers with no Periodic Updates or Age Fields," *Proc. IEEE ICDCS '97: IEEE 17th International Conference on Distributed Computing Systems,* Baltimore, Maryland, May 27–30, 1997.

34. C. Shields and J.J. Garcia-Luna-Aceves, "The Ordered Core Based Tree Protocol," *Proc. IEEE INFOCOM '97,* Kobe, Japan, April 7–11, 1997.

35. L. Kalampoukas, A. Varma and K. K. Ramakrishnan, "Explicit Window Adaptation: A Method to Enhance TCP Performance," *Proc. IEEE INFOCOM '98,* San Francisco, California, March 29–April 2, 1998.

36. B. Smith and J.J. Garcia-Luna-Aceves, "Securing the Border Gateway Routing Protocol," *Proc. Global Internet'96*, London, UK, 20-21 November 1996.

37. B.N. Levine, D. Lavo, and J.J. Garcia-Luna-Aceves, "The Case for Concurrent Reliable Multicasting Using Shared ACK Trees," *Proc. ACM Multimedia '96: Fourth ACM International Multimedia Conference*, Boston, MA, November 1996.

There were 83 student papers considered for the 1998 Best Student Paper Award at the 1998 IEEE International Conference on Systems, Man, and Cybernetics. Each paper was reviewed by three reviewers. Each paper was reviewed by three people and the results were tabulated to determine the five papers selected as finalists for the competition. Each of the five finalists were required to present their paper in a special Student Paper Competition session as well as their regularly scheduled paper session. H-P. Dommel received the Best Student Paper Award for the paper "A Novel Group Coordination Protocol for Distributed Multimedia Group Collaboration."

This final report is organized as follows. Section 2 presents the results of our work on scalable routing architectures based on path-finding techniques. Section 3 presents our work on routing over multiple paths, minimum-delay routing, and new approaches for providing performance guarantees in a scalable manner in computer networks. Section 4 presents our work on multicast routing architectures and protocols. Section 5 presents our results on end-to-end reliable multicasting. Section 6 presents our work on group collaboration support based on multipoint communication support. Section 7 presents our results on congestion control. In each of these sections, we summarize the main results of our work, followed by the main papers describing the technical details of our research. Section 8 summarizes directions for future research.

2 SCALABLE ROUTING USING PATH-FINDING MECHANISMS

We developed two new hierarchical routing algorithms for computer networks to enable routing to any number of destinations while maintaining small routing tables at each router. Both algorithms accommodate an arbitrary number of aggregation levels; nodes and networks are organized into level-1 areas, and multiple areas at a given level are aggregated into a higher-level area.

The first algorithm is called the hierarchical information path-based routing (HIPR) algorithm. HIPR extends our prior work on routing algorithms based on source tracing or path finding techniques[1]. HIPR is based on the incremental exchange of hierarchical routing trees, and can be viewed as the first distributed version of Dijkstra's SPF algorithm running over a hierarchical graph.

Simulation results show that, for the case in which only two levels of aggregation are used, HIPR is far superior than OSPF in terms of time to converge, control messages exchanged, and computational overhead. HIPR is also much more scalable than OSPF, because it supports any number of aggregation levels and in contrast to OSPF requires no backbone. The following paper describes this algorithm in detail:

- S. Murthy and J.J. Garcia-Luna-Aceves, "A "A Loop-Free Routing Protocol for Large-Scale Internets Using Distance Vectors," *Computer Communications,* Vol. 21, No. 2, 1998, pp. 147–161.

The second algorithm is called area-based link vector algorithm. It extends our previous work on routing using partial link-state information.[2]. and allows each router to maintain only a partial topology that includes those links needed to reach destinations in its own area, and remote areas.

Simulation results show that ALVA is superior to OSPF for the case of two-level aggregation. As is also the case with HIPR, ALVA does not require backbones and supports any number of aggregation levels.

- J. Behrens and J.J. Garcia-Luna-Aceves, "Hierarchical Routing Using Link Vectors," *Proc. IEEE INFOCOM '98,* San Francisco, California, March 29–April 2, 1998.

We also analyzed different ways in which link-state updates can be validated using sequence numbers, without the need for periodic updates, which may become undesirable in networks where nodes should not allow other entities to determine when they will be transmitting control information, or when battery life is an issue. These mechanisms are applicable to LVA and ALVA. These techniques and their performance are described in the following paper:

- J. Behrens and J.J. Garcia-Luna-Aceves, "Fast Dissemination of Link States Using Bounded Sequence Numbers with no Periodic Updates or Age Fields," *Proc. ICDCS'97 International Conference on Distributed Computing Systems*, Baltimore, Maryland, May 27-30, 1997.

There are several mechanisms defined to protect the communication across links between neighboring nodes. However, today's Internet routing and multicasting protocols provide few mechanisms, if any, to protect the exchange of control information. We developed the only known method to secure both classes of information in distance-vector routing protocols in constant space, which means that the overhead of the security mechanism is only linear with respect to the number of destinations. This method is based on the path-traversal mechanisms developed for HIPR and ALVA, and was applied to BGP and intra-domain routing protocol. The following papers describe the details of this work:

- B. Smith and J.J. Garcia-Luna-Aceves, "Efficient Security Mechanisms for The Border Gateway Routing Protocol," *Computer Communications Journal,* Vol. 21, No. 3, pp. 203–210, 1998.
- B. Smith and J.J. Garcia-Luna-Aceves, "Securing the Border Gateway Routing Protocol," *Proc. Global Internet'96*, London, UK, 20-21 November 1996.

3 MULTIPATH ROUTING AND QUALITY OF SERVICE

The ability to route packets over multiple paths becomes essential when network delays must be minimized, which has been proven by Gallager. In essence, minimum-delay routing can be achieved or approximated only by using multiple available paths to reach any one destination. In addition, using multiple paths is critical for providing fault-tolerant routing in very large networks or internetworks.

Unfortunately, there are many limitations to today's Internet routing protocols. The widely deployed routing protocol RIP[3] provides only one next-hop choice for each destination and does not prevent temporary loops from forming. Cisco's EIGRP[4] ensures loop-freedom but can guarantee only a single loop-free path to each destination at any given router. The link-state protocol OSPF[5] offers a router multiple choices for packet-forwarding only when those choices offer the minimum distance. When there is fine granularity in link costs metric, perhaps for accuracy, there is less likelihood that multiple paths with equal distance exist between each source-destination pair, which means the full connectivity of the network is still not used for load-balancing. Also, OSPF and other algorithms based on topology-broadcast (e.g., [6, 7]) incur too much communication overhead, which forces the network administrators to partition the network into areas connected by a backbone. This makes OSPF complex in terms of router configuration required.

To address the limitations of today's Internet routing protocols, we developed several novel algorithms for routing of packets over multiple paths that need not be of equal cost. Formally, let a computer network be represented as a graph $G = (N, L)$, where N is set of nodes (routers) and L is the set of edges (links), and let N^i be the set of neighbors of node i. The problem consists of finding the successor set at each router i for each destination j, denoted by $S_j^i \subseteq N^i$, so that when router i receives a packet for destination j, it can forward the packet to one of the neighbor routers in the successor set S_j^i. By repeating this process at every router, the packet is expected to reach the destination. If the routing graph SG_j, a directed subgraph of G, is defined by the link set $\{(m, n) | n \in S_j^m,\ m \in N\}$, a packet destined for j follows a path in SG_j. Two criteria determine the efficiency of the routing graph constructed by the protocol: *loop-freedom* and *connectivity*. It is required that SG_j be free of loops, at least when the network is stable, because routing loops degrade network performance. In a dynamic environment, a stricter requirement is that SG_j be loop-free at *every instant*, i.e., if S_j^i and SG_j are parameterized by time t, then $SG_j(t)$ should be free of loops at any time t. If there is at most one element in each S_j^i, then SG_j is a tree and there is only one path from any node to j. On the other hand, if S_j^i's have more than one element, then SG_j is a directed acyclic graph (DAG) and has greater connectivity than a simple tree enabling traffic load balancing.

We developed fault-tolerant and self-organizing routing algorithms that provide multiple loop-free paths to each destination using only distances to destinations, the distance and second-to-last hop of the path to each destination, or partial link-state information corresponding to those links in the paths used to reach destinations.

We also introduced a generalization of loop-freedom conditions for routing algorithms based on any type of information, and applied multipath routing algorithms to a load-balancing routing framework to obtain "near-optimal" delays. A key component of this framework is a fast responsive routing protocol that determines multiple successor choices for packet forwarding, such that the routing graphs implied by the routing tables are free of loops even during network transitions. By load-balancing traffic over the multiple next-hop choices, congestion and delays are reduced significantly.

The following papers describe our results on multipath routing and minimum-delay routing:

- S. Vutukury and J.J. Garcia-Luna-Aceves, "MPATH: A Loop-free Multipath Routing Algorithm," to appear in *Microprocessors and Microsystems Journal*, Elsevier, 2001.
- S. Vutukury and J.J. Garcia-Luna-Aceves, "A Distributed Algorithm for Multipath Computation," *Proc. IEEE Globecom '99*, Rio de Janeiro, Brazil, December 5–9, 1999.
- S. Vutukury and J.J. Garcia-Luna-Aceves, "An Algorithm for Multipath Computation using Distance-Vectors with Predecessor Information," *Proc. IEEE IC3N 99*, Boston, Massachusetts, October 11–13, 1999.
- S. Vutukury and J.J. Garcia-Luna-Aceves, "A Practical Framework for Minimum-Delay Routing in Computer Networks," *Journal of High Speed Networks*, Vol. 8, No. 4, pp. 241-263, Wiley, 1999.
- S. Vutukury and J.J. Garcia-Luna-Aceves, "A Simple Approximation to Minimum-Delay Routing," *Proc. ACM SIGCOMM 99*, Cambridge, Massachusetts, September 1–3, 1999.
- J.J. Garcia-Luna-Aceves, S. Vutukury, and W.T. Zaumen, "A Practical Approach to Minimizing Delays in Internet Routing Protocols," *Proc. IEEE ICC '99*, Vancouver, Canada, June 6–10, 1999.

When multiple paths to destinations are provided at the routing layer, such end-to-end protocols as TCP may suffer performance degradations due to packets being delivered out of order. To solve this problem without having to establish virtual circuits in routers or tags in packets, we developed a traffic engineering approach that allows routers to forward packets of a given TCP connection over the same path, while distributing packets of different TCP flows over different paths. This work is described in the following paper:

- S. Vutukury and J.J. Garcia-Luna-Aceves, "A Traffic Engineering Approach based on Minimum-delay Routing," *Proc. IEEE IC3N 2000*, Las Vegas, Nevada, October 16–18, 2000.

It is now widely accepted that explicit resource reservations must be made in the Internet to provide the kind of guarantees (bandwidth, delay and delay-jitter) new application

demand. There are two QoS architectures being proposed in the Internet today. The Integrated Services (Intserv) [8, 9] architecture provides deterministic guarantees to individual flows by reserving resources on a single route from the source to the destination using a signaling protocol (e.g., RSVP [10]); however, it cannot scale well because of the excessive state maintained in routers. The Differential Services (Diffserv) architecture [12, 11] aggregates routing and reservation state in the routers to achieve scalability, but cannot provide deterministic guarantees. Both approaches also suffer from the inherent limitation of relying on single-path routing and single-path signaling for resource reservations.

In this project, we introduced a multipath routing framework for the provision of QoS guarantees in computer networks, without the need to maintain per-flow state at routers. This is the first routing architecture capable of providing deterministic guarantees in wired networks using the same amount of state as the Diffserv architecture. This work is described in the following papers:

- S. Vutukury and J.J. Garcia-Luna-Aceves, "A Multipath Framework Architecture for Integrated Services," *Proc. IEEE Globecom 2000*, San Francisco, California, November 27–December 1, 2000.

- S. Vutukury and J.J. Garcia-Luna-Aceves, "A Scalable Architecture for Providing Deterministic Guarantees," *Proc. IEEE IC3N 99*, Boston, Massachusetts, October 11–13, 1999.

4 MULTICAST ARCHITECTURES AND PROTOCOLS

Today's Internet multicast routing protocols can be classified as link-state and distance-vector protocols. MOSPF is a link-state multicast routing protocol; it scales poorly with the size of the Internet because it requires each router to know which multicast group is present on which link. Accordingly, whenever a change is made to the membership of the multicast group, all routers must be notified by flooding. The distance-vector multicast routing protocol (DVMRP) adopted in MBONE uses variants of reverse-path forwarding (RPF) and assumes that routers participate in a multicast group unless they state otherwise. Because routers are assumed to participate in a multicast group and to some extent engage in the maintenance of trees for which they have no members in their attached networks, DVMRP does not scale well and is useful only for relatively small dense sessions. The scaling problems of MOSPF and DVMRP have led to the development of the core based tree (CBT) protocol [13] and the protocol independent multicast (PIM) architecture [14].

CBT and PIM have three scaling problems: they require the preselection of special nodes (cores in CBT or rendezvous points in PIM); the receivers' requests to join a multicast group result in shortest paths from the receivers to the source; and they are able to scale with the number of multicast groups and not the sources only by using a single multicast tree per group. Having to locate one or multiple cores or one or multiple rendezvous points in a large internet adds unnecessary network configuration tasks, and is not as fault-tolerant as providing a fully distributed algorithm in all routers perform the same functions. Building multicast trees using shortest paths from the receivers to the source is not a problem when unicast routes are symmetric, (i.e, the shortest path from a source to receiver is the same as the shortest path from the receiver to the source); however, unicast routes in a large internet can have asymmetric characteristics, resulting in suboptimum or unfeasible multicast trees. This last issue is a concern in multimedia applications that require quality of service (QoS) guarantees from source to the destination set. Using a multicast tree per group results in suboptimum paths when there is more than one source per group.

As part of this project, we completed our work on delay-constrained minimum-cost multicasting. This work has been used by several other authors to propose distributed heuristics for achieving multicast routing that adheres to delay constraints.

- M. Parsa, Q. Zhu, and J.J. Garcia-Luna-Aceves, "An Iterative Algorithm for Delay-Constrained Minimum-Cost Multicasting," *IEEE/ACM Transactions on Networking,* Vol. 6, No. 4, August 1998.

We also completed the design, verification and analysis of the Multicast Internet Protocol (MIP), which is the first multicast routing protocol that is free of loops at every instant. Instantaneous loop freedom is very important in multicast routing, because multicast packets multiply themselves as they traverse multicast routing loops. MIP is based on "diffusing computations," similar to those used for unicast routing in DUAL, the routing algorithm

used in Cisco's EIGRP. We showed by simulations that MIP incurs far less overhead than PIM. This work is described in the following paper:

- M. Parsa and J.J. Garcia-Luna-Aceves, "A Protocol for Scalable Loop-Free Multicast Routing," *IEEE Journal on Selected Areas in Communications,* Vol. 15, No. 3, pp. 316–331, April 1997.

We showed that prior specifications of the Core Based Tree (CBT) protocol produce undetected loops and fail to construct multicast routing trees. To address the limitations of CBT, we developed the Ordered Core based Tree (OCBT) protocol. To make OCBT more scalable, we developed a hierarchical version of OCBT, which we call HIP.

HIP allows the use of shared tree protocols such as PIM-SM or CBT as the inter-domain routing protocol in a hierarchy that can include any routing protocol at the lowest level. The architecture consists of two protocols: one that encapsulates an entire routing domain to allow it to appear as a *virtual router* on a higher-level shared tree; and a second protocol that provides mechanisms for rendezvous point or core distribution and recursively applies the first protocol to produce trees of domains that contain trees of domains. HIP differs from prior hierarchical multicast protocols in many ways. It is the first architecture to allow any multicast protocol at the lowest level while using a shared tree for higher-level routing. It provides a simple, efficient mechanism for RP or core location dissemination. HIP aligns easily with existing unicast domains and it does not require explicit assignment of levels except at the highest level. HIP is suitable for any shared tree protocol, such as PIM-SM or CBT, that forms a tree by sending join messages to some central router. It can also provide additional robustness for the shared tree protocol through the ability to replace a single rendezvous point or core with several routers that operate together in a distributed fashion and can tolerate members failing.

Using our work on HIP as our baseline, we developed Keyed HIP (KHIP), a secure multicast routing protocol. We showed that other shared-tree multicast routing protocols are subject to attacks against the multicast routing infrastructure that can isolate receivers or domains or introduce loops into the structure of the multicast routing tree. KHIP changes the multicast routing model so that only trusted members are able to join the multicast tree. This protects the multicast routing against attacks that could form branches to unauthorized receivers, prevents replay attacks and limits the effects of flooding attacks. Untrusted routers that are present on the path between trusted routers cannot change the routing and can mount no denial-of-service attack stronger than simply dropping control messages. KHIP also provides a simple mechanism for distributing data encryption keys while adding little overhead to the protocol.

The work on OCBT, HIP, and KHIP is described in the following papers:

- C. Shields and J.J. Garcia-Luna-Aceves, "The Ordered Core Based Tree Protocol," *Proc. IEEE INFOCOM '97,* Kobe, Japan, April 7–11, 1997.

- C. Shields and J.J. Garcia-Luna-Aceves, "HIP–A Protocol for Hierarchical Multicast Routing," *Computer Communications*, Vol. 23, No. 7, pp. 628-641, Elsevier, 2000.

- C. Shields and J.J. Garcia-Luna-Aceves, "A Scalable Protocol for Secure Multicast Routing," *Proc. ACM SIGCOMM 99*, Cambridge, Massachusetts, September 1–3, 1999.

Lastly, we developed architectural improvements to IP multicast to encourage communication and cooperation between IP and higher-layer protocols. The first, called AIM, is a generalization of internetwork multicasting. AIM provides an extended set of semantics so that sources can restrict the delivery of packets to a subset of information. AIM can be implemented using CBT or PIM. The second architecture, called Streams, allows higher-layer protocols place packets into application-defined logical streams, so that hosts may prune the routing of packets based on meaningful contexts. The third architecture, called RMA, introduces new routing services designed to support reliable multicast protocols. The cooperation between RMA and reliable multicast protocols provides very scalable and efficient reliable communication among hosts, and allows any set of reliable multicast protocols to cooperate across an internet by providing a common interface.

We also analyzed the role of locality of reference or receiver interest in multicast routing. This work is presented in the following papers:

- B.N. Levine and J.J. Garcia-Luna-Aceves, "Improving Internet Multicast Using Routing Labels," *Proc. IEEE ICNP '97: Fifth IEEE International Conference on Network Protocols,* Atlanta, GA, October 1997.

- B. Levine, J. Crowcroft, C. Diot, J.J. Garcia-Luna-Aceves, and J. Kurose, "Consideration of Receiver Interest for IP Multicast Delivery," *Proc. Infocom 2000*, Tel-Aviv, Israel, March 26–30, 2000.

5 RELIABLE MULTICASTING

The increasing popularity of real-time applications supporting either group collaboration or the broadcast of multimedia information over the Internet (e.g., nv, vat, wb, and shdr) are making the provision of reliable and unreliable end-to-end multicast services an integral part of its architecture. In addition to these emerging multimedia applications, there are many reasons why multicasting should be an integral part of network control. First, the availability of mixed media permits a given router to have very many neighbors, and reliable multicast becomes an effective tool to make the routing protocol more efficient by allowing a router to send a single control message to entire multicast groups, rather than addressing the same message individually to each neighbor. Second, most communication with simple devices in the NICE is likely to involve multiple devices at a time. Third, there are many special-purpose networks and applications that need reliable multicast as an inherent control function; one typical example consist of disseminating notifications of critical simulation events in a mobile distributed simulation network in which a real participant of the simulation (e.g., an aircraft) has to notify the rest of the simulation participants of a critical simulation event (e.g., a simulated target is killed or engaged). Fourth, trusting all network nodes and links is a luxury that cannot be afforded in the NII, because of its size and multitude of administrative authorities; consequently, routing protocols need to be designed with Byzantine robustness to guard against both faulty or malicious behavior of links and nodes, and intra-network reliable multicast becomes useful for distributing group keys used to implement Byzantine robustness efficiently.

Clearly, reliable multicast within the network or on an end-to-end basis is needed in the future Internet architecture. Although reliable multicast protocols have existed for quite some time (e.g., see [20]) the multicast problem facing the future Internet is compounded by its current size and continuing growth, which makes the handling of acknowledgments a major technological problem (known as the "acknowledgment implosion problem").

The two basic approaches to reliable multicast proposed to date are called *sender initiated* and *receiver initiated*. In the sender-initiated approach, the sender maintains the state of all the receivers to whom it has to send information and from whom it has to receive acknowledgments (ACKs). Each sender's transmission or retransmission is multicast to all receivers; for each packet that each receiver obtains correctly, it sends a unicast ACK to the receiver. Most recent proposals for reliable multicast protocols to date are sender initiated (e.g., see [15], [16]).

In contrast, in the receiver-initiated approach, the responsibility of receiving information reliably lies on the receivers, each of whom must ask the sender for the information that is in error or missing. The sender multicasts all packets, giving priority to retransmissions. A receiver sends a negative acknowledgment (NACK) when it detects an error or a lost packet. Ramakrishnan and Jain [17] and more recently Jacobson [18] have proposed this type of protocol for reliable multicast. To help with ACK implosion, each receiver is asked to wait for a random period of time when it detects an error or lost packet before it multicasts its NACK.

Pingali, Towsley, and Kurose [19] have shown that, when host processing is the main

constraining resource, receiver-initiated multicast protocols have better throughput than the traditional sender-initiated multicast protocol, which requires the sender to receive all the acknowledgments from the receivers. This is not surprising, because the traditional sender-initiated approach was not designed to support large multicast groups.

However, the receiver-initiated approach has several limitations for its application in the Internet. First, it is easy to show by example that, when NACKs can get lost in the Internet, the source may not receive the few NACKs flowing back to it and the receiver-initiated approach, as proposed by Jacobson, is not safe (i.e., it cannot guarantee that *all* packets will be delivered correctly to all destinations).

Another problem with the receiver-initiated approach is that it is difficult to determine the times that different receivers have to wait before sending NACKs in order to reduce the flow of NACKs to the source; this is very similar to 'hidden-terminal' problems in packet radio, in which senders of packets (NACKs in our case) cannot detect (or detect too late) other transmissions (in the case of the NII, this is due to the inherent latencies of a large network). Therefore, although the goal of using a NACK-only retransmission strategy is to limit ACK implosion, the scheme provides no guarantees that only a few NACKs will flow back to the source.

It is also difficult to use the feedback from the receivers to implement multicast-oriented congestion avoidance techniques, because of the random nature of the the NACKs flowing back to the source. Furthermore, this scheme is very difficult to extend to enforce complete ordering of packets originated from different sources in the same group (e.g., in a collaborative visualization session), because it relies on tuning of changing operational parameters.

A critical design issue for the adoption in the Internet of any reliable multicast protocol that has been overlooked in the discussions of receiver-initiated approaches is the need to provide proper initialization and termination of the multicast connections. A NACK-based scheme cannot be used for such purposes, because the source (or sources) can never tell when all the intended receivers are ready to start accepting packets without confusing them with older packets. On the other hand, assuming an off-line mechanism for initialization (e.g., using another protocol to initialize one or multiple nodes of the reliable multicast group) is no different than relying on a sender-initiated approach in which the source receives an ACK for each receiver that is initialized properly.

In this project, we provided the first comparative analysis of sender-initiated, receiver-initiated, tree-based and ring-based protocols for end-to-end reliable multicasting. The basic conclusion from this analysis is that reliable multicast protocols that organize receivers along receiver trees that correlate with the underlying multicast routing tree used for packet forwarding constitute the best approach. Such protocols were shown to be far more scalable and support higher maximum throughput than any other type of reliable multicast protocol proposed to date, such as sender initiated, receiver initiated, or ring-based protocols.

With the results of our comparative analysis, we focused on tree-based reliable multicast protocols, and developed the Lorax protocol. Lorax is the first known protocol that constructs and maintains a single shared acknowledgment (ACK) tree for concurrent reliable multicasting.

Lorax eliminates the need to maintain an ACK tree for each source of a reliable multicast group, and can be used in combination with any of several tree-based reliable multicast protocols proposed to date. Lorax provides solutions to several open questions concerning the implementation of shared ACK trees. Preserving reliability during restructuring of the ACK tree is easily guaranteed using aggregated acknowledgments that propagate from each leaf towards the source.

Lorax constructs and maintains a shared ACK tree. Overhead traffic is contained during initial ACK-tree construction by growing the ACK tree from a known root. Impatient nodes are quieted down and allowed to join the ACK tree by means of expanded ring searches that are narrow in scope. Hierarchical labeling of each node makes implicit routing of acknowledgments simple and preserves loop-free routing of such acknowledgments over the ACK tree at all times.

Lastly, we developed the Tracer protocol, which is the first protocol that organizes the receivers of a reliable multicast group deterministically along a logical tree according to the packet-loss correlation. Tracer's implementation relies only on multicast trace packets in IGMP.

The following papers describe the details of this work:

- B.N. Levine and J.J. Garcia-Luna-Aceves, "A Comparison of Reliable Multicast Protocols," *ACM Multimedia Systems Journal,* Vol. 6, No. 5, August 1998.

- B.N. Levine, S. Paul, and J.J. Garcia-Luna-Aceves, "Organizing Multicast Receivers Deterministically According to Packet-Loss Correlation," *Proc. ACM Multimedia 98: Sixth ACM International Multimedia Conference,* Bristol, UK, September 1998.

- B.N. Levine, D. Lavo, and J.J. Garcia-Luna-Aceves, "The Case for Concurrent Reliable Multicasting Using Shared ACK Trees," *Proc. ACM Multimedia '96: Fourth ACM International Multimedia Conference*, Boston, MA, November 1996.

6 GROUP COORDINATION BASED ON MULTICAST SUPPORT

With the advent of multicast support in networks and internetworks, more powerful collaborative multimedia applications (CMA) can enter mainstream computing. Users of such applications can overcome their separation in time and space and share work efforts in real-time, with the goal of approximating the quality of face-to-face interactions. However, compared to advances in reliable multicasting and multicast routing, there has been little progress with regard to group coordination support for such systems.

We addressed two key problems in group coordination above and beyond reliable multicasting of individual packets in a group:

- Coordinated access to shared resources.
- Message ordering for reliable multicast communication.

We addressed coordinated access to resources as a floor-control problem. Floor control is an access discipline for CMA that may solve flawed tele-presence and coordination problems. It lets users attain exclusive control over a shared resource by attaining a *floor*, which is a short-lived synchronization primitive for multimedia objects. The floor semantics is generalized to multimedia from its traditional notion as the "right to speak." Floor allocation establishes clear rules for turn-taking, and prevents race conditions, indefinite resource holding, and unfairness in resource access patterns. Participating stations agree on a specific floor policy, concerning service order and priorities. The spectrum of control can range from lenient to strict, reflecting characteristics of tasks and interaction styles, such as user roles, usage quotas, or resource contention periods. As a component within a general coordination architecture for many-to-many groupwork, floor control coexists with protocols for reliable ordered multicast and media synchronization at a sub-application level.

Deployment of floor control over the Internet or a very large network may be complex due to system capabilities, in terms of the network, host, and communication subsystem, as well as user behavior. Surprisingly, a detailed analysis on the operational principles and performance of floor control protocols had been missing. In this project, we carried out the first taxonomy and efficacy analysis of floor control protocols based on their operational principles.

We have shown that floor control over a shared propagation tree, corresponding to the underlying end-to-end reliable multicast tree, represents the most scalable and efficient way to store and forward control information. In particular, it allows to exploit the hierarchical nature of multicast groups, supports selective control packet dissemination to multicast subgroups, and distributes load about floor state keeping across the multicast tree, without compromising ease of implementation. Based on our analysis, we developed the Hierarchical Group Coordination Protocol (HGCP), a novel approach for floor control in shared propagation trees. HGCP enables nodes to coordinate among one another over a multicast tree to determine which node holds the floor for a given resource at any given time. The protocol

makes use of the shared trees built by such reliable multicast protocols as Lorax or the AIM and RMA architectures we present in the previous section.

End-to-end multicast ordering is useful for ensuring the collective integrity and consistency of distributed operations. It is applicable for distributed multi-party collaboration or other multipoint applications, where the ordered reception of messages at all hosts is critical.

Existing reliable multicast protocols largely lack support for ordering. Our novel mechanism can be added to existing reliable multicast services at low cost by performing cascaded total ordering of messages among on-tree hosts en route from senders to receivers. The protocol operates directly on a given end-to-end multicast tree, contrasting other tree-based approaches requiring a separate propagation graph to be built to compute ordering information. For better load distribution, resilience, and ordered subcasting of messages within multicast groups, sequencer nodes are elected dynamically based on address extensions to hosts in the multicast tree.

We developed the Tree-based Ordered Multicast (TOM) protocol, which relies on an underlying reliable multicast tree for propagation of ordering information besides acknowledgments and retransmissions. This tree is assumed to approximate the underlying multicast routing tree, which for the Internet is built using various protocols such as DVMRP, CBT or PIM-SM for a general overview). For the following description, we assume that hosts do not fail and network partitions do not occur. Trees can be constructed per source, which amortizes itself only for long-lived or large-volume transmissions, or dissemination can be based on a shared tree, across which (negative) acknowledgments are relayed between hosts. In such a tree, sources may change frequently, only one collective infrastructure must be maintained, and a source need not know the identity of all receivers in the multicast group. However, the paths from sources to receivers may be suboptimal.

The key idea in TOM is to multicast a message from a source to a receiver set combined with sending ordering information for the message (sequence numbers or time stamps) to a common node on the tree elected as ordering node for this receiver set (or multicast group). The ordering node sequences messages assigned to it and multicasts binding sequence numbers for final delivery to the receiver set, where pending messages are to be delivered. TOM can be deployed in the form of an API accessible to applications with ordering needs.

The following papers describe the details of this work:

- H.-P. Dommel and J.J. Garcia-Luna-Aceves, "A Novel Group Coordination Protocol for for Distributed Multimedia Group Collaboration", *Proc. 1998 IEEE International Conference on Systems, Man, and Cybernetics,* San Diego, California, October 11-14, 1998.
  (**WINNER OF BEST STUDENT PAPER AWARD**)

- H. P. Dommel and J. J. Garcia-Luna-Aceves "Group Coordination Support for Internet Collaboration," *IEEE Internet Computing Magazine*, Special Issue on Multimedia and Collaborative Computing over the Internet, Vol. 3, No. 2, pp. 74-80, March/April 1999.

- H.P. Dommel and J.J. Garcia-Luna-Aceves, "Ordered End-to-End Multicast for Distributed Multimedia Systems," *Proc. Hawaii International Conference on System Sci-*

ences (HICS-33), Maui, Hawaii, January 4–7, 2000.

- H.-P. Dommel and J.J. Garcia-Luna-Aceves, "A Coordination Framework and Architecture for Internet Groupwork," *Journal of Network and Computer Applications (JNCA)*, Special Issue on Support for Open and Distance Learning on the WWW, Academic Press pub, Fall 2000.

7 CONGESTION CONTROL

We studied mechanisms to control congestion at edge routers of a network or internetwork in ways that help the performance of TCP. The key objective is to execute packet discard in ways that signal TCP sources early about the existence of congestion. Packet discard policies have been studied extensively in recent years. The random early discard (RED) and Drop-from-front policies are examples of such schemes.

We developed the explicit window adaptation (EWA) scheme, which controls the end-to-end window size used for flow control to correspond to the round-trip delay-bandwidth product. EWA achieves very low packet loss, a high degree of fairness, and almost perfect bandwidth utilization. EWA does not require any modifications to TCP. The following paper describes the details of this work:

- L. Kalampoukas, A. Varma and K. K. Ramakrishnan, "Explicit Window Adaptation: A Method to Enhance TCP Performance," *Proc. IEEE INFOCOM '98,* San Francisco, California, March 29–April 2, 1998.

8 CONCLUSIONS

The NICE Routing and Internetworking project made many contributions to advance the state of the state of the art in internetworking. Key contributions of this project include new solutions for hierarchical routing, routing over multiple loop-free paths of unequal cost, loop-free multicasting, secure multicasting, scalable solutions for reliable and total ordered multicasting, and scalable solutions for resource sharing using multicast services.

The results obtained in this project open up a number of opportunities for future research.

The progress made in multipath routing and scalable routing protocols enables new research on fault-tolerant routing in very large networks, internetworks and sensor networks. Of particular interest is the provision of minimum-delay routing with added constraints that may result from the environment or the applications that the network supports. In general, a new architecture and protocols for fault-tolerant internetworking can be developed such that: (a) routers can *protect* efficiently against attacks and faults, and *detect and respond* to them in a timely manner; (b) no routing and multicasting function has single points of failure; and (c) QoS guarantees are provided in a scalable and fault-tolerant manner.

Our results on multicasting can be used to develop new approaches for the establishment of virtual private networks as end-to-end secure multicast groups, rather than aggregations of tunnels among routers.

Finally, our work on coordination of processes based on multicast supports opens new avenues of research for the support of fault-tolerant multimedia applications.

References

[1] J.J. Garcia-Luna-Aceves and S. Murthy, "A Path Finding Algorithm for Loop-Free Routing", *IEEE/ACM Trans. Networking*, February 1997.

[2] J.J. Garcia-Luna-Aceves and J. Behrens, "Distributed, Scalable Routing Based on Vectors of Link States," *IEEE Journal on Selected Areas in Communications*, October 1995.

[3] C. Hendrick. Routing Information Protocol. *RFC*, 1058, june 1988.

[4] R. Albrightson, J.J. Garcia-Luna-Aceves, and J. Boyle. EIGRP-A Fast Routing Protocol Based on Distance Vectors. *Proc. Networld/Interop 94*, May 1994.

[5] J. Moy. OSPF Version 2. *RFC*, 1247, August 1991.

[6] J. Spinelli and R. Gallager. Event Driven Topology Broadcast without Sequence Numbers. *IEEE Trans. Commun.*, 37:468–474, 1989.

[7] R. Perlman. Fault-tolerant broadcast of routing information. *Computer Networks and ISDN*, 7, 1983.

[8] S. Shenker, D. Clark, and L. Zhang. A service model for an integrated services internet. *Internet Draft*, Oct. 1993.

[9] E. Crawley et al. A framework for qos-based routing in the internet. *Internet Draft*, April 1998.

[10] L. Zhang et al. RSVP: A New Resource Reservation Protocol. *IEEE Communications Magazine*, 31(9):8–18, 1993.

[11] Y. Bernet et al. An Framework for Differentiated Services. *Internet Draft*, May 1998.

[12] D. Black et al. An Achitecture for Differentiated Services. *Internet Draft*, May 1998.

[13] T. Ballardie et al., "Core Based Trees (CBT)–An Architecture for Scalable Inter-Domain Multicast Routing," *Proc. ACM SIGCOMM 93*, San Francisco, CA, 1993.

[14] S. Deering, et al., "An Architecture for Wide-Area Multicast Routing," *Proc. ACM SIGCOMM 94*, London, UK, August 1994.

[15] W.T. Strayer, et al., *XTP: The Xpress Transfer Protocol*, Addison-Wesley, 1992.

[16] J. Crowcroft and K. Paliwoda, "A Multicast Transport Protocol," *Proc. ACM SIGCOMM 88*, Stanford, California, 1988, pp. 247-256.

[17] S. Ramakrishnan and B.N. Jain, "A Negative Acknowledgment with Periodic Polling Protocol for Multicast over LANs," *Proc. IEEE INFOCOM 87*, San Francisco, California, 1987.

[18] V. Jacobson, "Lightweight Session–A New Architecture for Realtime Applications and Protocols," *Proc. Networking '93: 3th Annual PI Meeting*, ARPA, August 1993.

[19] S. Pingali, D. Towsley, J. Kurose, "A Comparison of Sender-Initiated and Receiver-Initiated Reliable Multicast Protocols," *Proc. SIGMETRICS 94*, Santa Clara, CA, 1994.

[20] J.M. Chang and N. Maxemchuck, "Reliable Broadcast Protocols," *ACM Trans. Comp. Systems*, Vol. 2, No. 3, April 1984.