Why Multicast Protocols (Don't) Scale:
An Analysis
of Multipoint Algorithms for Scalable Group Communication

Thesis by

Eve M. Schooler

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

Also published as Caltech Computer Science Technical Report caltechCSTR/2001.003

California Institute of Technology

Pasadena, California

2001

(Defended September 19, 2000)

Acknowledgments

First and foremost, I would like to thank my committee members for their participation and feedback: Jehoshua (Shuki) Bruck, K. Mani Chandy, Deborah Estrin, Jason Hickey, and Alain Martin. I also would like to thank several other individuals who read my thesis and provided valuable comments: Bob Felderman, Joe Kiniry, Rajit Manohar, and Ruth Sivilotti. I extend special thanks to my committee chair and thesis advisor K. Mani Chandy, who has provided generous support and friendship throughout my stay at Caltech. I am also grateful to him for providing the freedom to work on whatever research presented itself as interesting. I consider myself lucky to have been part of his research group, where I came into contact with many gifted researchers: Michel Charpentier, Roman Ginis, Peter Hofstee, Joe Kiniry, K. Rustan Leino, Berna Massingill, Adam Rifkin, Paul Sivilotti, John Thornley, and Dan Zimmerman. I thank each of these individuals not only for contributing to my education, but also for enriching my experience at Caltech. Where else could one resort to the humor in using "the magic 8 ball" to make admissions or coding decisions? I have also been blessed to collaborate with Rajit Manohar, whose advice, guidance and insights have been invaluable. It is a rare friend who is willing to critique half-formed ideas.

There have been many others who keep this department running and whose help has always been greatly appreciated: Cynthia Brady, Jeri Chittum, Betta Dawson, Cindy Ferrini, Louise Foucher, Diane Goodfellow, Cici Koenig, Yvonne Recendez, and Gail Stowers. To the army of people who have maintained the systems on which I rely, I bow down to you: Dave Felt, Roman Ginis, Joe Kiniry, Dave LeBlanc, Chris Lee, Rajit Manohar, Mika Nyström, and Dan Zimmerman. To my friends at Myricom and Cornell, where most of my simulations were run, I could not have finished without the use of your fast machines (may it be a long time until I have to port the network simulator to yet another OS). I am deeply grateful. I am also thankful for the breadth of knowledge of my friends at the Fairchild Engineering Library: Kimberly Douglas and Hema Ramachandran. The resources they have assembled and have made available for on-line researchers have made my life immensely more productive.

There have been many others within the Computer Science Department who have provided both comraderie and intellectual companionship: Cindy Ball, Al Barr, Eric Bax, Cynthia Brady, David Breen, Mathieu Desbrun, Boris Dimitrov, Ilja Friedel, Eitan Grinspun, Rohit Khare, Cici

Koenig, David Laidlaw, Alain Martin, Daniel Maskit, Rajit Manohar, Mark Meyer, Mika Nyström, Marc Reiffel, Steve Taylor, Jerrell Watts, and Zöe Wood. I thank them for making this journey a memorable one. Of course, I am also especially grateful to certain unnamed individuals for never making me feel too "old."

I owe a debt of gratitude to my office mates: Rajit Manohar for sharing his exquisite taste in chocolate and music, Zöe Wood whose artistry and spirit I prize, and Eitan Grinspun for his willingness to launch into song whenever the mood struck us!

I thank Ed Perry and his group at HP Labs Broadband Systems Lab for a rewarding summer internship in 1996. I extend heartfelt thanks to my friends and colleagues at Microsoft's Bay Area Research Center: Jim Gemmell and Jim Gray, as well as a host of other researchers who were there during my internship during the Summer of 1997. It was my honor to work with so accomplished a group of individuals, who were willing to entertain the notion of a telecommuting internship! I have thoroughly enjoyed our collaborations.

There are many individuals from the Internet Engineering Task Force (IETF) community who I thank for providing an extra backdrop against which to do research in Networks: Mark Handley, Ruth Lang, Jöerg Ott, Allison Mankin, Scott Shenker, Abel Weinrib, and many others from the Multiparty Multimedia Session Control (MMusic) working group. In addition, I extend a huge thank you to Steve Coya for making it possible for me to attend IETF meetings on a student budget.

I thank the outstanding community of people who comprise the High Speed Networking Division of USC's Information Science Institute (ISI), where I was employed prior to Caltech and where I continued to work part-time during my first two years in graduate school. Several colleagues and friends from ISI deserve special thanks for having encouraged me over the years and for serving as constant role models: Celeste Anderson, Yigal Arens, Bob Braden, Steve Casner, Danny Cohen, Deborah Estrin, Mike Gorman, Mary Hall, and the late Jon Postel.

Other individuals whose encouragement and friendship I have valued immensely, and whose advice about staying in school I probably should have heeded back when I was completing my Masters at UCLA: Thelma and Jerry Estrin, Len and Stella Kleinrock, and Verra Morgan.

Despite appearances, there are women at Caltech! I have been fortunate to have met many phenomenal women scientists and other Caltech affiliates who have been supportive throughout: Cindy Ball, Andrea Belz, Melanie Bennett Brewer, Zehra Cataltepe, Min Chen, Christina Cohen, Roian Egnor, Carmit Eliyahu, Caroline Fohlin, Jen Linden, Berna Massingill, Helen Parker, Susan Pelletier, Cathy Wong, and Zöe Wood.

Friends I cherish for their wisdom and perspective and without whom I could not have survived: Ruth Ballenger, Suzanne Biegel, Nan Boden, Christina Cohen, Herb Donaldson, Hyewon Hyun, Rajit Manohar, Leanne Lung Nemeth, Ruth Sivilotti, Beverly Stein, and Sara Tucker.

I also wish to thank my friends who have looked after me in times of need: Dr. Lynda Roman and Nurse Charlotte Haravey of the USC Norris Cancer Center, and Nurses Divina Bautista and Alice Sogomonian at the Caltech Student Health Center.

Most importantly, I thank my family: my sisters, who have helped me to balance life and work and to understand what is truly important; my father whose intellect and curiosity know no bounds and whose fascination with invention has always inspired me; my mother who taught me my first algorithms when she taught me to read music (the for-loop of musical repeats, the go-to of skipping to a coda), who has always served as the model for the person I strive to be, and whose ability to create beauty wherever she goes awes me still.

I treasure my son Sean, who has good-naturedly marked the passage of time while I have been at Caltech! He certainly deserves his own doctorate in something: Zoology, for he can navigate the Caltech campus to find and identify most resident wildlife; Patience, for enduring my demanding schedule this past year so I could finish "the book I have been writing"; Frogology, the art of catching frogs from ponds on the Caltech campus? In addition, I could not have completed my degree were it not for Leanne Lung Nemeth, Cathy Saewert, and the wonderful teachers at the Childrens Center at Caltech, who have allowed me to do my research knowing that Sean was in good hands when he was in their care.

Above all, I am indebted to Bob Felderman, the most exceptional husband and father I know. I am grateful to him for his ceaseless encouragement, as well as his sense of humor. I thank him for happily providing advice and feedback (when solicited!), and for having the ingenuity to ask just the right questions when I was trying to debug my simulations, work out a problem, or turn a phrase. To say Bob has been infinitely supportive – not only during the normal course of everyday grad student life, but also during a particularly difficult year of health challenges – would be an understatement. We will always joke that Caltech considers spouses "dependents," for we know who has depended on whom! For the many sacrifices he has made, and for his generosity of spirit, I am eternally grateful. It is to him that I dedicate this thesis.

The research described in this thesis was funded in part by an Earl C. Anthony Graduate Fellowship, a Career Development Grant from the American Association of University Women, a Microsoft Graduate Fellowship, as well as the Air Force Office of Scientific Research and the National Science Foundation. I thank all of them for their generous support.

Abstract

With the exponential growth of the Internet, there is a critical need to design efficient, scalable and robust protocols to support the network infrastructure. A new class of protocols has emerged to address these challenges, and these protocols rely on a few key techniques, or micro-algorithms, to achieve scalability. By scalability, we mean the ability of groups of communicating processes to grow very large in size. We study the behavior of several of these fundamental techniques that appear in many deployed and emerging Internet standards: Suppression, Announce-Listen, and Leader Election.

These algorithms are based on the principle of efficient multipoint communication, often in combination with periodic messaging. We assume a loosely-coupled communication model, where acknowledged messaging among groups of processes is not required. Thus, processes infer information from the periodic receipt or loss of messages from other processes.

We present an analysis, validated by simulation, of the performance tradeoffs of each of these techniques. Toward this end, we derive a series of performance metrics that help us to evaluate these algorithms under lossy conditions: expected response time, network usage, memory overhead, consistency attainable, and convergence time. In addition, we study the impact of both correlated and uncorrelated loss on groups of communicating processes.

As a result, this thesis provides insights into the scalability of multicast protocols that rely upon these techniques. We provide a systematic framework for calibrating as well as predicting protocol behavior over a range of operating conditions. In the process, we establish a general methodology for the analysis of these and other scalability techniques. Finally, we explore a theory of composition; if we understand the behavior of these micro-algorithms, then we can bound analytically the performance of the more complex algorithms that rely upon them.

Contents

List of Figures

List of Tables

List of Programs