

José
Duato

Sudhakar
Yalamanchili

Lionel
Ni

INTERCONNECTION NETWORKS

an Engineering Approach

IEEE
COMPUTER
SOCIETY



Duato
Yalamanchili
Ni

INTERCONNECTION NETWORKS



IEEE
COMPUTER
SOCIETY
PRESS

Interconnection Networks

An Engineering Approach



IEEE Computer Society Press

Mohamed E. Fayad

Editor-in-Chief, Practices for Computer Science and Engineering

SELECTED TITLES

Distributed Objects: Methodologies for Customizing Operating Systems

Nayeem Islam

Software Engineering Risk Management

Dale Karolak

Digital Design and Modeling with VHDL and Synthesis

K.C. Chang

Industrial Strength Software: Effective Management Using Measurement

Lawrence Putnam and Ware Myers

Unified Objects: Object-Oriented Programming Using C++

Babak Sadr

Interconnection Networks: An Engineering Approach

José Duato, Sudhakar Yalamanchili, and Lionel Ni

Meeting Deadlines in Real-Time Systems: The Rate Monotonic Analysis

Daniel Roy and Loic Briand

Executive Briefings

Controlling Software Development

Lawrence Putnam and Ware Myers

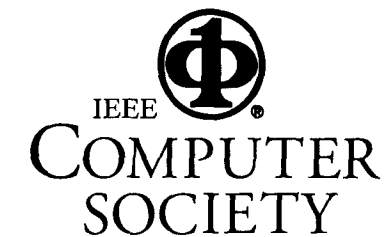
Interconnection Networks

An Engineering Approach

José Duato

Sudhakar Yalamanchili

Lionel Ni



Los Alamitos, California

Washington • Brussels • Tokyo

Library of Congress Cataloging-in-Publication Data

Duato, José

Interconnection networks: an engineering approach / José Duato, Sudhakar

Yalamanchili, Lionel Ni.

p. cm.

Includes bibliographical references.

ISBN 0-8186-7800-3

1. Computer networks. 2. Multiprocessors. I. Yalamanchili.

II. Ni, Lionel M. III. Title.

TK5105.5.D88 1997

004.6—dc21

97-20502

CIP

To Josep, Albert, and Joana

Copyright © 1997 by The Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy isolated pages beyond the limits of US copyright law, for private use of their patrons. Other copying, reprint, or republication requests should be addressed to: IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.

IEEE Computer Society Press Order Number BP07800

Library of Congress Number 97-20502

ISBN 0-8186-7800-3

Additional copies may be ordered from:

IEEE Computer Society Press
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314
Tel: +1-714-821-8380
Fax: +1-714-821-4641
Email: cs.books@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: +1-908-981-1393
Fax: +1-908-981-9667
mis.custserv@computer.org

IEEE Computer Society
13, Avenue de l'Aquilon
B-1200 Brussels
BELGIUM
Tel: +32-2-770-2198
Fax: +32-2-770-8505
euro.ofc@computer.org

IEEE Computer Society
Ooshima Building
2-19-1 Minami-Aoyama
Minato-ku, Tokyo 107
JAPAN
Tel: +81-3-3408-3118
Fax: +81-3-3408-3553
tokyo.ofc@computer.org

Editor-in-Chief: Mohamed Fayad

Publisher: Matt Loeb

Acquisitions Editor: Bill Sanders

Developmental Editor: Cheryl Baltes

Advertising/Promotions: Tom Fink

Production Editor: Lisa O'Conner

Cover Artist: John Marshall

About the cover: *Interconnection Networks* by John Marshall, oil on canvass, 1997, is based on a concept by José Duato. The painting includes symbols representing important elements in the field. To find out more about the symbols in the painting, visit the book's website at <http://computer.org/books>.

Printed in the United States of America

To Ching, Wayland, and Elaine



Peace comes within the source of the men, when they realize the wonders of the Universe, when they realize that it is really everywhere, and it is within each of us.

— In the beginning (Swahili song), Beautiful World

Foreword

The performance of most digital systems today is limited by their communication or interconnection, not by their logic or memory. In a high-end system today, most of the power is used to drive wires and most of the clock cycle is spent on wire delay, not gate delay. As technology improves, memories and processors become small, fast, and inexpensive. The speed of light, however, remains unchanged. The pin density and wiring density that govern interconnections between system components are scaling at a slower rate than the components themselves. Also, the frequency of communication between components is lagging far beyond the clock rates of modern processors. These factors combine to make interconnection the key factor in the success of future digital systems.

As designers strive to make more efficient use of scarce interconnection bandwidth, interconnection networks are emerging as a nearly universal solution to the system-level communication problems for modern digital systems. Originally developed for the demanding communication requirements of multicomputers, interconnection networks are beginning to replace buses as the standard system-level interconnection. They are also replacing dedicated wiring in special-purpose systems as designers discover that routing packets is both faster and more economical than routing wires.

Interconnection networks have a long history. Circuit switched networks have long been used in telephony. The 1950s saw many proposals for interconnecting computers and cellular automata, but few prototypes. These came in the 1960s. Solomon in 1962 was the first of a long line of mesh-connected, bit-serial multicomputers. Staran with its flip network, C.mmp with a crossbar, and Illiac-IV with a wider 2-D mesh followed in the early 1970s. This period also saw several indirect networks used in vector and array processors to connect multiple processors to multiple memory banks. To address this problem the academic community developed several variants of multistage interconnection networks. The BBN Butterfly in 1982 was one of the first multiprocessors to use an indirect network. The binary n-cube or hypercube network was proposed in 1978 and first implemented in the Caltech Cosmic Cube in 1981. In the early 1980s, the academic community focused on mathematical properties of these networks and became increasingly separated from the practical problems of interconnecting real systems.

The last decade has been a golden-age for interconnection network research. Driven by the demanding communication problems of multicomputers and enabled by the ability to construct single-chip VLSI routers, the research community has made a series of breakthroughs that have revolutionized communication in digital systems. The Torus Routing Chip, in 1985, was one of the first steps along this path. The first of a series of single-chip routing components, it introduced wormhole routing and the use of virtual channels for deadlock avoidance. This chip, and others like it, laid the framework for analysis of routing, flow-control, deadlock, and livelock issues in modern direct networks. A flurry of research followed with new theories of deadlock and livelock, new adaptive routing algorithms, new methods for performance analysis, new approaches to collective communication, and new network architectures arriving on a regular basis. By the early 1990s, low-dimensional direct networks based on these concepts had largely replaced the indirect networks of the 1970s and the hypercubes of the 1980s and could be found in machines from Cray, Intel, Mercury, and others.

The applicability of interconnection networks to the general communication problem in digital systems was clear by 1995 with the appearance of Myrinet. These point-to-point networks were the technology of choice to replace buses which were running into a performance ceiling due to electrical limits. As evidenced by their use in the barrier network in the Cray T3E, they also represented an economical alternative to dedicated wiring. However, one barrier remained to the widespread use of interconnection network technology. The information on the design and analysis of these networks was not readily available to practicing engineers and designers. Descriptions of the technology were

scattered across hundreds of conference papers and journal articles with overlapping treatment and often inconsistent notation.

This book, for the first time, makes the technology of interconnection networks accessible to the engineering student and the practicing engineer. It gives a comprehensive treatment of interconnection networks from topology, through routing and flow control, to a discussion of the software layer. The authors of the book are three key members of the research community and are responsible for developing much of the technology described. Their unique knowledge and rare insight into the material make for a technically rich treatment of the material that brings together the best of many research papers and fills in the gaps by putting the work in context.

This book improves upon the original research papers in several ways. It presents the material in a logical order, as opposed to the chronological order of the original papers. The text eliminates overlap between related papers, introduces a consistent notation, and fills in gaps. Perhaps most importantly, the book places each piece of work in context painting a “big picture” of interconnection network technology throughout the text. Finally, unlike many research papers, the presentation is at a level appropriate for a first-year engineering graduate student or a practicing engineer. By demystifying interconnection networks and making the technology accessible and available for wide application, this book will empower a generation of digital designers.

I trust that you will find this book as enjoyable and insightful as I have.

William J. Dally
Cambridge, MA
June 1997

Preface

Interconnection networks are becoming increasingly pervasive in many different applications, with the operational costs and characteristics of these networks considerably depending on the application. For some applications, interconnection networks have been studied in depth for decades. This is the case for telephone networks, computer networks (telecommunications), and backplane buses. The design and operation of these networks are covered in many excellent books. However, in the last 10 years we have seen a rapid evolution of the interconnection network technology that is currently being infused into a new generation of multiprocessor systems. The technology is mature enough to find its way into commercial products, while constantly presenting new avenues for growth and application. This is the case for the interconnection networks used in multicomputers and distributed shared-memory multiprocessors. At the current time, the basic design issues governing this new generation of multiprocessor network technology are diffused over a large number of technical papers. This book is an attempt to present a coherent description of this technology in a form that emphasizes the engineering aspects of their construction and use.

The lack of standards and the need for very high performance and reliability have been primarily responsible for pushing the development of interconnection networks for multicomputers. This technology was transferred to distributed shared-memory multiprocessors. More recently, this network technology began to be transferred to local area networks (LANs). Also, it has been proposed as a replacement for backplane buses, creating the concept of system area networks. Hence, the advances in interconnection networks for multicomputers are becoming the basis for the development of interconnection networks for other architectures and environments. Therefore, there is a need for formally stating the basic concepts, the alternative design choices, and the design tradeoffs for most of those networks. In this book, we address this challenge and present in an structured way the basic underlying concepts of most interconnection networks, and representative solutions that have currently been implemented or proposed in the literature.

Style of the Book

The literature on routing algorithms and architectures for multiprocessors is extensive. It is not our intent to serve as a resource for a comprehensive review of this area, although we hope that the bibliography will serve as a good starting point for the interested reader. With the fast pace of development in this area, any such attempt would have a rather short lifetime. Rather we hope to bring together the basic issues and results, and present them from a practical perspective. We hope such a presentation will have a more lasting value. As a result, selections from the literature for inclusion have more been guided by the nature of the problems they address, and the authors’ familiarity with the work. We hope to be as comprehensive as possible in the treatment of fundamental issues rather than extent.

This book follows an engineering approach. It only considers those issues that should be taken into account by the designer. However, for each of those issues, we present a broad set of solutions proposed in the literature. Most of these solutions have been proposed very recently (1991–1996). These solutions are described and compared in an informal but rigorous way. We explicitly avoided including complex mathematical descriptions, making the text clear and easier to understand. The reader can certainly find the detailed mathematical treatment in the associated journal and conference papers. We have striven to ensure that the vast majority of material is from refereed journal and conference proceedings, presented from an engineering perspective. However, descriptions are precise. Moreover, we define some new concepts and structure the knowledge on interconnection networks. A considerable effort has been made to establish new and more accurate classifications for