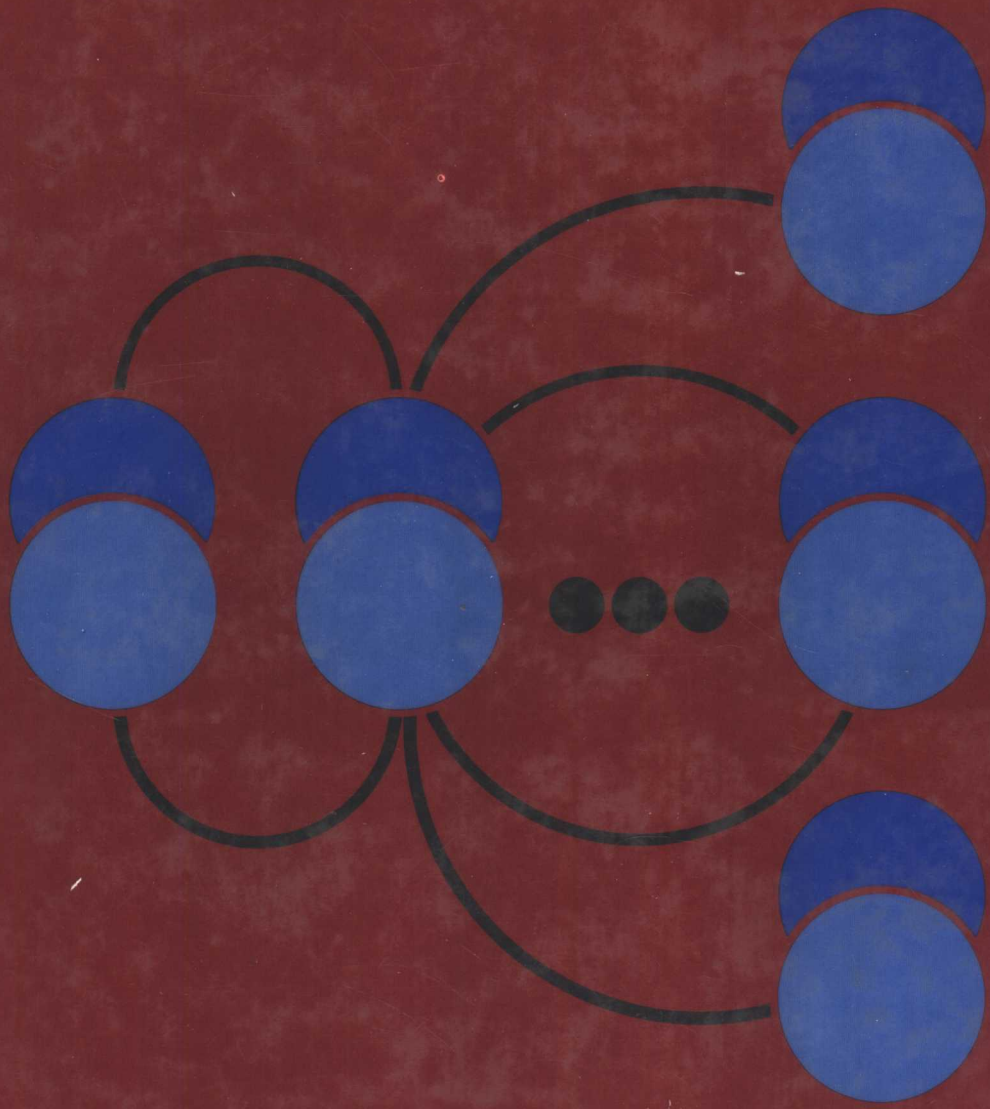
DATA NETWORKS

DIMITRI BERTSEKAS / ROBERT GALLAGER



DATA NETWORKS

Dimitri Bertsekas

Massachusetts Institute of Technology

Robert Gallager

Massachusetts Institute of Technology

PRENTICE-HALL, INC., Englewood Cliffs, New Jersey 07632

To Joanna and Marie

Editorial/production supervision by Margaret Rizzi
Cover design by Wanda Lubelska
Manufacturing buyer: Rhett Conklin

A Division of Simon & Schuster
Englewood Cliffs, New Jersey 07632

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 0-13-196825-4 025

PRENTICE-HALL INTERNATIONAL (UK) LIMITED, London
PRENTICE-HALL OF AUSTRALIA PTY. LIMITED, Sydney
PRENTICE-HALL CANADA INC., Toronto
PRENTICE-HALL HISPANOAMERICANA, S.A., Mexico
PRENTICE-HALL OF INDIA PRIVATE LIMITED, New Delhi
PRENTICE-HALL OF JAPAN, INC., Tokyo
PRENTICE-HALL OF SOUTHEAST ASIA PTE. LTD., Singapore
EDITORA PRENTICE-HALL DO BRASIL, LTD., Rio de Janeiro

Preface

The field of data networks has evolved over the last fifteen years from a stage where networks were designed in a very ad hoc and technology-dependent manner to a stage where some broad conceptual understanding of many underlying issues now exists. The major purpose of this book is to convey that conceptual understanding to the reader.

Previous books in this field broadly separate into two major categories. The first, exemplified by Tannenbaum [Tan81] and Stallings [Sta85], are primarily descriptive in nature, focusing on current practice and selected details of the operation of various existing networks. The second, exemplified by Kleinrock [Kle76], Hayes [Hay84], and Stuck and Arthurs [StA85], deal primarily with performance analysis. This book, in contrast, is balanced between description and analysis. The descriptive material, however, is used to illustrate the underlying concepts, and the analytical material is used to provide a deeper and more precise understanding of the concepts. We feel that a continuing separation between description and analysis is unwise in a field after the underlying concepts have been developed; understanding is then best enhanced by focusing on the concepts.

The book is designed to be used at a number of levels, varying from a senior undergraduate elective, to a first year graduate course, to a more advanced graduate course, to a reference work for designers and researchers in the field. The material has been tested in a number of graduate courses at M.I.T. and in a number of short courses at varying levels. The book assumes some

background in elementary probability and some background in either electrical engineering or computer science, but aside from this, the material is self-contained.

Throughout the book, major concepts and principles are first explained in a simple non-mathematical way. This is followed by careful descriptions of modelling issues and then by mathematical analysis. Finally, the insights to be gained from the analysis are explained and examples are given to clarify the more subtle issues. Figures are liberally used throughout to illustrate the ideas. For lower-level courses, the analysis can be glossed over; this allows the beginning and intermediate-level to grasp the basic ideas, while enabling the more advanced student to acquire deeper understanding and the ability to do research in the field.

Chapter 1 provides a broad introduction to the subject and also develops the layering concept. This layering allows the various issues of data networks to be developed in a largely independent fashion, thus making it possible to read the subsequent chapters in any desired depth (including omission) without seriously hindering the ability to understand other chapters.

Chapter 2 treats the two lowest layers of the above layering. The lowest, or physical, layer is concerned with transmitting a sequence of bits over a physical communication medium. We provide a brief introduction to the subject which will be helpful but not necessary in understanding the rest of the text. The next layer, data link control, deals with transmitting packets reliably over a communication link. Section 2.4, treating retransmission strategies, should probably be covered in any course, since it brings out the subtleties, in the simplest context, of understanding distributed algorithms, or protocols.

Chapter 3 develops the queueing theory used for performance analysis of multiaccess schemes (Chapter 4) and, to a lesser extent, routing algorithms (Chapter 5). Less analytical courses will probably omit most of this chapter, simply adopting the results on faith. Little's theorem and the Poisson process should be covered however, since they are simple and greatly enhance understanding of the subsequent chapters. This chapter is rich in results, often developed in a far simpler way than found in the queueing literature. This simplicity is achieved by considering only steady-state behavior and by sometimes sacrificing rigor for clarity and insight. Mathematically sophisticated readers will be able to supply the extra details for rigor by themselves, while for most readers the extra details would obscure the line of argument.

Chapter 4 develops the topic of multiaccess communication, including local area networks, satellite networks, and radio networks. Less theoretical courses will probably skip the last half of section 4.2, all of section 4.3, and most of section 4.4, getting quickly to local area networks and satellite networks in section 4.5. Conceptually, one gains a great deal of insight into the nature of distributed algorithms in this chapter.

Chapter 5 develops the subject of routing. The material is graduated in order of increasing difficulty and depth, so readers can go as far as they are

comfortable. Along with routing itself, which is treated in greater depth than elsewhere in the literature, further insights are gained into distributed algorithms. There is also a treatment of topological design and a section on recovery from link failures.

Chapter 6 deals with flow control (or congestion control as it is sometimes called). The first three sections are primarily descriptive, describing first the objectives and the problems in achieving these objectives, second, some general approaches, and finally, the ways that flow control is handled in several existing networks. The last section is more advanced and analytical, treating recent work in the area.

A topic that is not treated in any depth in the book is that of higher-layer protocols, namely the various processes required in the computers and devices using the network to communicate meaningfully with each other given the capability of reliable transport of packets through the network provided by the lower layers. This topic is different in nature than the other topics covered and would have doubled the size of the book if treated in depth.

We apologize in advance for the amount of acronyms and jargon in the book. We felt it was necessary to include at least the most commonly used acronyms in the field, both to allow readers to converse with other workers in the field and also for the reference value of being able to find out what these acronyms mean.

An extensive set of problems are given at the end of each chapter except the first. They range from simple exercises to gain familiarity with the basic concepts and techniques to advanced problems extending the results in the text. Solutions of the problems are given in a manual available to instructors from Prentice-Hall.

Each chapter contains also a brief section of sources and suggestions for further reading. Again, we apologize in advance to the many authors whose contributions have not been mentioned. The literature in the data network field is vast, and we limited ourselves to references that we found most useful, or that contain material supplementing the text.

The stimulating teaching and research environment at M.I.T. has been an ideal setting for the development of this book. In particular we are indebted to the many students who have used this material in courses. Their comments have helped greatly in clarifying the topics. We are equally indebted to the many colleagues and advanced graduate students who have provided detailed critiques of the various chapters. Special thanks go to our colleague Pierre Humblet whose advice, knowledge, and deep insight have been invaluable. In addition, Erdal Arikan, David Castanon, Robert Cooper, Tony Ephremides, Eli Gafni, Marianne Gardner, Paul Green, Ellen Hahne, Bruce Hajek, Robert Kennedy, John Spinelli, and John Tsitsiklis have all been very helpful. We are also grateful to Nancy Young for typing the many revisions and to Amy Hendrikson for computer typesetting the book using the TEX system. Our editors at Prentice-

Hall have also been very helpful and cooperative in producing the final text under a very tight schedule. Finally we wish to acknowledge the research support of DARPA under grant ONR-N00014-84-K-0357, NSF under grants ECS-8310698, and ECS-8217668, and ARO under grant DAAG 29-84-K-000.

Dimitri Bertsekas

Robert Gallager

Contents

PREFACE xiii

Chapter 1
INTRODUCTION AND LAYERED NETWORK ARCHITECTURE 1

1.1 Historical Overview, 1

1.1.1 Technological and Economic Background, 4
1.1.2 Communication Technology, 5
1.1.3 Applications of Data Networks, 6

1.2 Messages and Switching, 8

1.2.1 Messages and Packets, 8
1.2.2 Sessions, 9
1.2.3 Circuit Switching and Store and Forward Switching, 11

1.3 Layering, 14

1.3.1 The Physical Layer, 17
1.3.2 The Data Link Control (DLC) Layer, 20
1.3.3 The Network Layer, 22
1.3.4 The Transport Layer, 24
1.3.5 The Session Layer, 25

1.3.6 The Presentation Layer, 25
1.3.7 The Application Layer, 26

1.4 A Simple Distributed Algorithm Problem, 26

1.5 Notes and Suggested Reading, 29

Chapter 2
DATA LINK CONTROL AND COMMUNICATION CHANNELS 31

2.1 Overview, 31

2.2 The Physical Layer: Channels and Modems, 34

2.2.1 Filtering, 35
2.2.2 Frequency Response, 37
2.2.3 The Sampling Theorem, 40
2.2.4 Bandpass Channels, 42
2.2.5 Modulation, 43
2.2.6 Frequency- and Time-Division Multiplexing, 47
2.2.7 Other Channel Impairments, 48
2.2.8 Digital Channels, 48
2.2.9 Propagation Media for Physical Channels, 49

2.3 Error Detection, 50

2.3.1 Single Parity Checks, 50
2.3.2 Horizontal and Vertical Parity Checks, 51
2.3.3 Parity Check Codes, 52
2.3.4 Cyclic Redundancy Checks, 54

2.4 ARQ—Retransmission Strategies, 58

2.4.1 Stop-and-Wait ARQ, 59
2.4.2 ARPANET ARQ, 62
2.4.3 Go Back n ARQ, 63

Rules Followed by Sending DLC, 66
Rules Followed by Receiving DLC, 67
Go Back n with Modulus $m > n$, 68
Efficiency of Go Back n Implementations, 70

2.4.4 Selective Repeat ARQ, 71

2.5 Framing, 73

2.5.1 Character-Based Framing, 75
2.5.2 Bit-Oriented Framing—Flags, 76
2.5.3 Length Fields, 79
2.5.4 Framing with Errors, 80
2.5.5 Maximum Frame Size, 82

2.6 Standard DLCs, 85

2.7 Session Identification and Addressing, 91

2.7.1 Session Identification in TYMNET, 92
2.7.2 Session Identification in the Codex Networks, 94

2.8 Error Recovery at the Network and Transport Layer, 95

2.8.1 End-to-End acks, Flow Control, and Permits, 96
2.8.2 Using End-to-End acks for Error Recovery, 98
2.8.3 The X.25 Network Layer Standard, 99

2.9 Summary, 101

2.10 Notes, Sources, and Suggested Reading, 102

PROBLEMS, 103

Chapter 3
DELAY MODELS IN DATA NETWORKS 111

3.1 Introduction, 111

3.1.1 Multiplexing of Traffic on a Communication Link, 112

3.2 Queueing Models—Little's Theorem, 114

3.3 The *M*/*M*/1 Queueing System, 122

3.3.1 Main Results, 124
3.3.2 Occupancy Distribution Upon Arrival, 132
3.3.3 Occupancy Distribution Upon Departure, 134

3.4 The *M*/*M*/*m*, *M*/*M*/∞, and *M*/*M*/*m*/*m* Systems, 134

3.4.1 $M/M/m$: The m-Server Case, 135
3.4.2 $M/M/\infty$: Infinite-Server Case, 138
3.4.3 $M/M/m/m$: The m-Server Loss System, 140

3.5 The $M/G/1$ System, 140

3.5.1 $M/G/1$ Queues with Vacations, 147
3.5.2 Reservations and Polling, 150

Single-User System, 152
Multiuser System, 154
Limited Service Systems, 157

3.5.3 Priority Queueing, 159

Nonpreemptive Priority, 159
Preemptive Resume Priority, 161

3.6 Networks of Transmission Lines, 163

3.7 Time Reversibility—Burke's Theorem, 167

3.8 Networks of Queues—Jackson's Theorem, 174

3.9 Summary, 180

3.10 Notes, Sources, and Suggested Reading, 180

PROBLEMS, 182

APPENDIX A: Review of Markov Chain Theory, 194

3.A.1 Discrete-Time Markov Chains, 194
3.A.2 Detailed Balance Equations, 196
3.A.3 Partial Balance Equations, 197
3.A.4 Continuous-Time Markov Chains, 197

APPENDIX B: Summary of Results, 199

Chapter 4
MULTIACCESS COMMUNICATION 205

4.1 Introduction, 205

4.1.1 Satellite Channels, 207
4.1.2 Multidrop Telephone Lines, 208
4.1.3 Multitapped Bus, 208
4.1.4 Packet Radio Networks, 209

4.2 Slotted Multiaccess and the Aloha System, 209

4.2.1 Idealized Slotted Multiaccess Model, 209
Discussion of Assumptions, 210
4.2.2 Slotted Aloha, 211
4.2.3 Stabilized Slotted Aloha, 216
Stability and Maximum Throughput, 216
Pseudo-Bayesian Algorithm, 217
Approximate Delay Analysis, 219
Binary Exponential Backoff, 221
4.2.4 Unslotted Aloha, 222

4.3 Splitting Algorithms, 224

4.3.1 Tree Algorithms, 225
Improvements to the Tree Algorithm, 227
Variants of the Tree Algorithm, 229
4.3.2 First-Come First-Serve Splitting Algorithms, 229
Analysis of FCFS Splitting Algorithm, 233
Improvements in the FCFS Splitting Algorithm, 237
Practical Details, 238
The Last-Come First-Serve (LCFS) Splitting Algorithm, 238
Delayed Feedback, 240
Round Robin Splitting, 240

4.4 Carrier Sensing, 240

4.4.1 CSMA Slotted Aloha, 241
4.4.2 Pseudo-Bayesian Stabilization for CSMA Aloha, 244
4.4.3 CSMA Unslotted Aloha, 246
4.4.4 FCFS Splitting Algorithm for CSMA, 247

4.5 Multiaccess Reservations, 249

4.5.1 Satellite Reservation Systems, 250
4.5.2 Local Area Networks: CSMA/CD and Ethernet, 254
Slotted CSMA/CD, 255

Unslotted CSMA/CD, 256
The IEEE 802 Standards, 257

4.5.3 Local Area Networks: Token Rings, 258

IEEE 802.5 Token Ring Standard, 261
Expected Delay for Token Rings, 262
Slotted Rings and Register Insertion Rings, 263

4.5.4 Local Area Networks: Token Buses and Polling, 265

IEEE 802.4 Token Bus Standard, 266
Implicit Tokens: CSMA/CA, 267

4.5.5 Higher-Speed Local Area Networks, 267

Expressnet, 269
Homenets, 270

4.5.6 Generalized Polling and Splitting Algorithms, 272

4.6 Packet Radio Networks, 274

4.6.1 TDM for Packet Radio Nets, 276
4.6.2 Collision Resolution for Packet Radio Nets, 277
4.6.3 Transmission Radii for Packet Radio, 280
4.6.4 Carrier Sensing and Busy Tones, 281

4.7 Summary, 282

4.8 Notes, Sources, and Suggested Reading, 283

PROBLEMS, 283

Chapter 5
ROUTING IN DATA NETWORKS 297

5.1 Introduction, 297

5.1.1 Main Issues in Routing, 299
5.1.2 An Overview of Routing in Practice, 302

Routing in the ARPANET, 303
Routing in the TYMNET, 305
Routing in SNA, 307

5.2 Network Algorithms and Shortest Path Routing, 308

5.2.1 Undirected Graphs, 308
5.2.2 Minimum Weight Spanning Trees, 312

5.2.3 Shortest Path Algorithms, 315

The Bellman-Ford Algorithm, 318
Dijkstra's Algorithm, 322
The Floyd-Warshall Algorithm, 323

5.2.4 Distributed Asynchronous Bellman-Ford Algorithm, 325
5.2.5 Adaptive Routing Based on Shortest Paths, 333

Stability Issues in Datagram Networks, 333
Stability Issues in Virtual Circuit Networks, 336

5.3 Broadcasting Routing Information—Coping with Link Failures, 340

5.3.1 Flooding—The ARPANET Algorithm, 343
5.3.2 Flooding without Periodic Updates, 345
5.3.3 Topology Broadcast without Sequence Numbers, 347

5.4 Flow Models, Optimal Routing, and Topological Design, 355

5.4.1 An Overview of Topological Design Problems, 360
5.4.2 The Subnet Design Problem, 362

Capacity Assignment Problem, 362
Heuristic Methods for Capacity Assignment, 364
Network Reliability Issues, 367
Spanning Tree Topology Design, 370

5.4.3 The Local Access Network Design Problem, 371

5.5 Characterization of Optimal Routing, 374

5.6 Feasible Direction Methods for Optimal Routing, 382

5.6.1 The Frank-Wolfe (Flow Deviation) Method, 385

5.7 Projection Methods for Optimal Routing, 392

Unconstrained Nonlinear Optimization, 392
Nonlinear Optimization Over the Positive Orthant, 394
Application to Optimal Routing, 396

5.8 Routing in the Codex Network, 403

5.9 Summary, 405

5.10 Notes, Sources, and Suggested Reading, 406

PROBLEMS, 407

Chapter 6
FLOW CONTROL 423

6.1 Introduction, 423

6.1.1 Main Objectives of Flow Control, 424

Keeping Delay Small within the Subnet, 424
Fairness, 425
Buffer Overflow, 427

6.2 Window Flow Control, 429

6.2.1 End-to-End Windows, 430

Limitations of End-to-End Windows, 432

6.2.2 Node-by-Node Windows for Virtual Circuits, 435
6.2.3 The Isarithmic Method, 437
6.2.4 Window Flow Control at the User Level, 438

6.3 Overview of Flow Control in Practice, 439

Flow Control in the ARPANET, 439
Flow Control in the TYMNET, 440
Flow Control in SNA, 440
Flow Control in the Codex Network, 441
Flow Control in X.25, 442

6.4 Flow Control Schemes Based on Input Rate Adjustment, 442

6.4.1 Combined Optimal Routing and Flow Control, 443
6.4.2 Max-Min Flow Control, 448
6.4.3 Implementation of Input Rates in a Dynamic Environment, 453

6.5 Summary, 455

6.6 Notes, Sources, and Suggested Reading, 455

PROBLEMS, 456

REFERENCES 463

INDEX 477

1

Introduction and Layered Network Architecture

1.1 HISTORICAL OVERVIEW

Primitive forms of data networks have a long history, including the smoke signals used by primitive societies, and certainly including nineteenth century telegraphy. The messages in these systems were first manually encoded into strings of essentially binary symbols, and then manually transmitted and received. Where necessary, the messages were manually relayed at intermediate points.

A major development, in the early 1950s, was the use of communication links to connect central computers to remote terminals and other peripheral devices, such as printers and remote job entry points (RJE) (see Fig. 1.1). The number of such peripheral devices expanded rapidly in the 1960s with the development of time-shared computer systems and with the increasing power of central computers. With the proliferation of remote peripheral devices, it became uneconomical to provide a separate long-distance communication link to each peripheral. Remote multiplexers or concentrators were developed to collect all the traffic from a set of peripherals in the same area and to send it on a single link to the central processor. Finally, to free the central processor from handling all this communication, special processors called *front ends* were developed to control the communication to and from all the peripherals. This led to the more complex structure shown in Fig. 1.2. The communication is automated in such systems, in contrast to telegraphy, for example, but the control of the communication is centrally exercised at the computer. While it is perfectly appropriate and widely accepted to refer to such a system as a data network, or computer communication network, it is simpler to view it as a computer

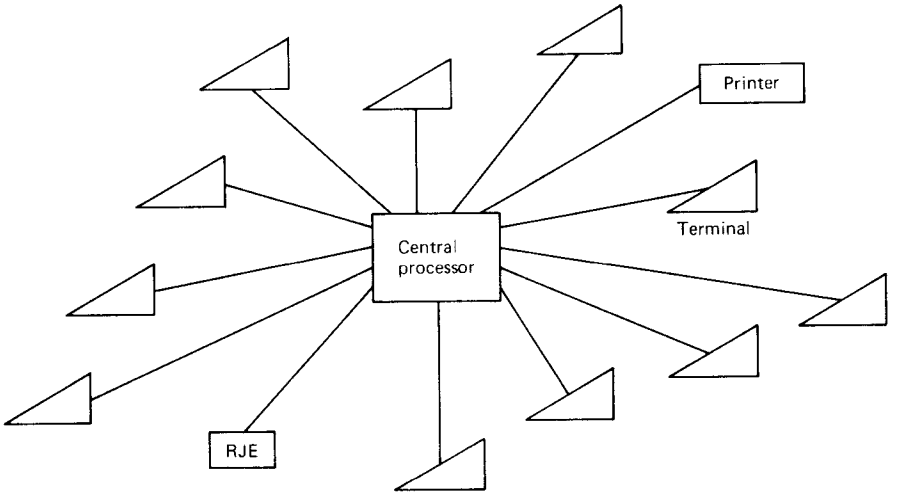


Figure 1.1 A network with one central processor and a separate communication link to each device.

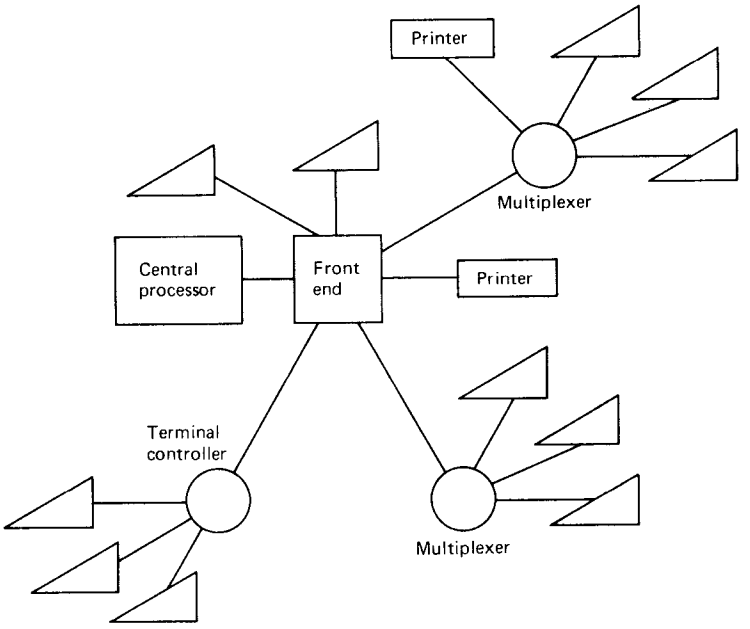


Figure 1.2 A network with one central processor but with shared communication links to devices.