Computer Science Department Technical Report
University of California
Los Angeles, CA 90024-1596

THE WAVELENGTH-DIVISION OPTICAL NETWORK: ARCHITECTURES, TOPOLOGIES, AND PROTOCOLS

Joseph Anthony Bannister

March 1990
CSD-900007

UNIVERSITY OF CALIFORNIA

Los Angeles

The Wavelength-Division Optical Network: Architectures, Topologies, and Protocols

A dissertation
submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Joseph Anthony Bannister

1990

The dissertation of Joseph Anthony Bannister is approved.

Kirby A. Baker

Jack W. Carlyle

Richard R. Muntz

Izhak Rubin

Mario Gerla, *Committee Chair*

University of California, Los Angeles

1990

Sto scrito xe dedicado in memoria de mia mamma,
Mariuccia Antonia Visintin Bannister.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

Much credit for this dissertation should go to my advisor and committee chair, Mario Gerla. Working with Mario has been a most rewarding and enjoyable experience because of his enthusiasm, helpfulness, fairness, and broad knowledge of the field of computer communications. I feel especially fortunate to have had such a positive professional and personal relationship.

I would also like to express my gratitude to my committee members, Kirby Baker, Jack Carlyle, Dick Muntz, and Izhak Rubin. I sincerely appreciate their efforts to improve the quality of this research and am thankful for the good working relationship that we maintained throughout my research.

Other faculty and staff members at UCLA have generously provided advice and assistance during my tenure in the Computer Science Department. A partial list of faculty members to whom I owe a debt of gratitude is Al Avižienis, Wes Chu, Miloš Ercegovac, Walter Karplus, Dave Rennels, and Cavour Yeh. The staff members of the department are some of the most dedicated, helpful, and friendly people that one could wish for; special thanks go to Verra Morgan, June Myers, and Doris Sublette.

The most enjoyable aspect of my time in the department has been its students. My research group, in particular, has provided support and stimulation that have directly benefited my own work. I would like to acknowledge the contributions and friendship of my fellow graduate students Beto Avritzer, José Suruagy Monteiro, Frank Schaffa, Phil Schmidt, Tsung-Yuan Tai, and Unni Warrier. In addition to my research group, several other graduate students have been good friends and worthy colleagues, including Mark Joseph, Carl Kesselman, Maria Pozzo, and T. M. Ravi.

Several people outside the UCLA community should also be singled out for thanks. In particular, Mel Cutler, my manager in the Computer Science Laboratory of The Aerospace Corporation, has been very encouraging and generous with respect to my research. Besides providing me with the substantial amount of computer resources needed in my research, Mel has always been able to offer excellent guidance and render sound judgement on matters both technical and nontechnical. I would also like to thank the following people for their support, encouragement, and friendship: George Gilley and Jon King of The Aerospace Corporation, and Jacob Abraham of the University of Texas at Austin. Specifically, I would like to acknowledge my collaboration with Luigi Fratta of the Politecnico di Milano—his numerous contributions and critical review have significantly improved the caliber of my research.

Finally, the understanding and support of my mate, Pamela Billings, have been instrumental in the maintenance of my sanity during the completion of this dissertation. My nocturnal companions, Zhinky and Perch, have also made dissertation writing a bit more tolerable.

The work represented in this dissertation has been partially supported by Pacific Bell-State of California MICRO Grant 4-541121-19907 KC7F.

VITA

1977	B.A. (Mathematics) University of Virginia Charlottesville, Virginia
1979–1980	Member of the Engineering Staff Xerox Corporation El Segundo, California
1980	M.S. (Engineering) University of California Los Angeles, California
1980–1982	Research Engineer Research Triangle Institute Research Triangle Park, North Carolina
1982–1984	Teaching and Research Assistant Computer Science Department University of California Los Angeles, California
1984	M.S. (Computer Science) University of California Los Angeles, California
1984–1985	Senior Engineer Sytek, Incorporated Culver City, California
1985–1988	Research Scientist Unisys Corporation Santa Monica, California
1988–present	Member of the Technical Staff The Aerospace Corporation El Segundo, California

Publications and Presentations

Bannister, J.A., K.S. Trivedi, V. Adlakha and T.A. Alspaugh, Jr., "Problems Related to the Integration of Fault-Tolerant Aircraft Electronic Systems," Final Technical Report for NASA Contract NAS1-16489, Research Triangle Institute, Research Triangle Park, North Carolina, August 1981.

Bannister, J.A. and J.B. Clary, "Performance Analysis of Systolic Array Architectures," *Proceedings of the SPIE Symposium, Real Time Signal Processing V*, Alexandria, Virginia, May 1982.

Bannister, J.A. and K.S. Trivedi, "Task Allocation and Load Balancing in Fault-Tolerant Distributed Systems," *Proceedings of the AMSE Conference on Modelling and Simulation*, Vallèe de Chevreuse, France, July 1982.

Bannister, J.A. and K.S. Trivedi, "Task and File Allocation in Fault-Tolerant Distributed Systems," *Proceedings of the Second Symposium on Reliability in Distributed Software and Database Systems*, Pittsburgh, Pennsylvania, July 1982.

Bannister, J.A. and K.S. Trivedi, "Task Allocation in Fault-Tolerant Distributed Systems," *Acta Informatica*, vol. 20, fasc. 3, December 1983.

Bannister, J.A. and U. Warrier, "Design and Analysis of Network Down Line Load Protocols," *Proceedings of IEEE INFOCOM '88*, New Orleans, Louisiana, March 1988.

Bannister, J.A., "Product Form Queueing Networks: State Dependence Revisited," Technical Report CSD-880021, Computer Science Department, University of California, Los Angeles, March 1988.

Bannister, J.A. and M. Gerla, "Design of the Wavelength-Division Optical Network," Technical Report CSD-890022, Computer Science Department, University of California, Los Angeles, May 1989.

Bannister, J.A., L. Fratta, and M. Gerla, "Designing Metropolitan Area Networks for High-Performance Applications," *Proceedings of the Seventh International Teletraffic Congress Specialists' Seminar*, Adelaide, Australia, September 1989; also to appear in *Computer Networks and ISDN Systems*.

Bannister, J.A. and M. Gerla, "Design of the Wavelength-Division Optical Network," To appear in the *Proceedings of the International Conference on Communications*, Atlanta, Georgia, April 1990.

Bannister, J.A., L. Fratta, and M. Gerla, "Topological Design of the Wavelength-Division Optical Network," To appear in the *Proceedings of IEEE INFOCOM '90*, San Francisco, California, June 1990.

Kesselman, C.F., M.M. Gorlick and J.A. Bannister, "Integrated Evaluation of Parallel Systems," Technical Report SD-TR-88-109, USAF Space Division, El Segundo, California, December 1988.

Smith, F.M. and J.A. Bannister, "A Preliminary Testability Analysis of the NEBULA Architecture," Technical Reports RTI/1822/04-02F and CORADCOM-80-0780-F, Research Triangle Institute, Research Triangle Park, North Carolina, December 1980.

Trivedi, K.S. and J.A. Bannister, "An Algorithm with Applications to Two Problems in the Design and Operation of Fault-Tolerant Distributed Systems," Technical Report 1982-5, Computer Science Department, Duke University, Durham, North Carolina, March 1982.

Warrier, U., A. Relan, O. Berry and J.A. Bannister, "A Network Management Language for OSI Networks," *Proceedings of ACM SIGCOMM '88*, Stanford, California, August 1988.

Bannister, J.A., "Design Optimization in Wavelength-Division Optical MANs," Presented at the Third IEEE Workshop on Metropolitan Area Networks, Dana Point, California, March 1989.

ABSTRACT OF THE DISSERTATION

The Wavelength-Division Optical Network: Architectures, Topologies, and Protocols

by

Joseph Anthony Bannister

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 1990

Professor Mario Gerla, *Chair*

Although optical-fiber waveguides can transport information at rates of several terabits per second (Tbps), this raw bandwidth can not be easily harnessed, because digital electronic circuits are not capable of transmitting or receiving information at such high rates. The *Wavelength-Division Optical Network (WON)* is a multichannel, multihop, store-and-forward, packet-switching network that takes advantage of the enormous bandwidth of lightwave technology by multiplexing several different, noninterfering wavelengths of light onto a single optical fiber. Such wavelength-division multiplexing (WDM) allows the creation of a large number of high-speed channels on an optical fiber. The WON, which is constructed from multitransceiver stations quasistatically tuned to the WDM channels, provides excellent throughput at a low cost, compared to other proposed networks. The WON is unique among networks because it possesses both a *physical topology* and a *vir-*

tual topology, and its virtual topology, which defines the logical interconnection of stations, may be specified independently of its physical topology. Furthermore, when wavelength-agile transceivers are used, the WON's virtual topology can be redefined at any point during its lifetime, which makes it possible to adapt the network to evolving conditions such as changes in the traffic load.

This work addresses the problems of designing the WON. The implementation of a WON requires the solution of several interrelated problems, including the design of its cable plant, its physical topology, its virtual topology, and its routing procedures. The goal of cable-plant and physical-topology design is to provide a power-efficient optical-signal distribution system with the lowest possible cost. We develop and demonstrate techniques for designing minimum-cost physical topologies for the WON. Given a specific physical topology and traffic requirements, the goal of virtual-topology and routing-protocol design is to select a routing procedure and assign stations to WDM channels so that the WON's performance is optimized when the routing protocol is used with the virtual topology. We adapt the well-known simulated annealing and genetic algorithms for use in optimizing the virtual topology of the WON to achieve optimum performance. We also develop a technique, called *detour routing*, that efficiently delivers packets when stations have as few as one packet buffer per transmitter.

Chapter 1

Introduction

This introductory chapter defines the scope of our research, describes its underlying technological assumptions, discusses new network architectures that are the target of our study, and outlines the problems to be solved in this dissertation.

1.1 Lightwave Technology and High-Performance Communication Networks

The optical spectrum, which can be defined as electromagnetic radiation with frequencies between 150 and 1500 terahertz, represents an enormous bandwidth that could be employed to transmit information at a very high rate. In contrast, the radiowave and microwave spectra, which range from about 100 kilohertz to 10 gigahertz, correspond to a much smaller bandwidth and rate of information transfer. Therefore, optical signaling is the technique most likely to be used in high-bandwidth communication systems.

Lightwave technology encompasses those components and techniques that may be used to transmit digital information by means of visible or nearly visible light that travels over optical-fiber waveguides. The application of this basic technology has resulted in the development of several new devices, including low-loss optical fibers, light sources, light detectors, connectors, couplers, optical-signal processors, and lenses. Over the previous two decades lightwave technology has steadily improved in performance, reliability, and cost. The speed–distance product, which measures the fundamental performance capabilities of a lightwave transmission system, has grown dramatically and is expected to continue growing at a similar pace. To see that the cost and reliability of lightwave components have improved, we need only to observe the widespread deployment of optical fibers for long-distance telephone communication.

Today new systems are being designed and developed to take advantage of advances in lightwave technology. The immediate payoff has come from the use of high-speed point-to-point links in voice and data communication, such as the Synchronous Optical Network (SONET) [SON88a, SON88b, SON88c], the Fiber-Distributed Data Interface (FDDI) [FDD87], and the Distributed Queue Dual Bus (DQDB) [DQD88]. Such networks are, however, inherently limited by their use of electronic switching equipment, which can not transfer signals at optical rates. For example, because of its geometric scaling properties, the metal-oxide-silicon integrated circuit appears to have a fundamental limit on the rate at which it can be clocked, and this limit implies that the rate of serial data transfer can not exceed an inherent upper bound. The exact value of the upper bound is a matter of debate, but it is generally set at about 1 gigabit per second (Gbps). To overcome this electro-optical bottleneck, we must look to different network architectures that can provide full connectivity among all users and achieve aggregate throughput in excess of the bottleneck. Ideally, such a network should provide an aggregate

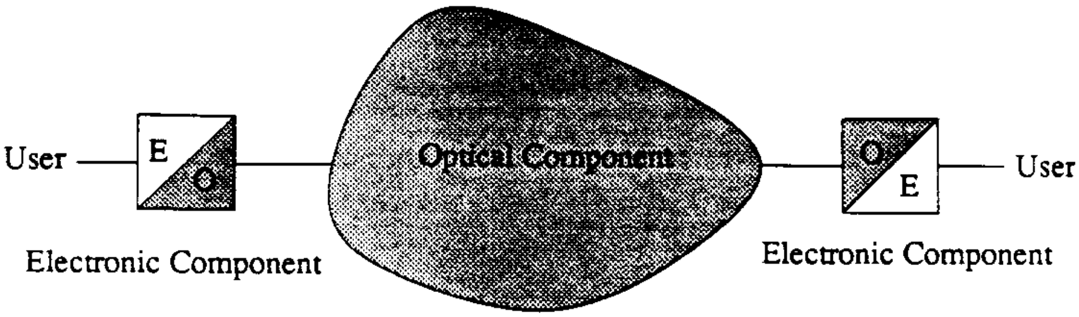


Figure 1.1: The Lightwave Communication Network.

throughput that scales linearly with the number of users, and whose maximum aggregate throughput approaches the product of the number of individual users and the maximum rate of data transfer across the electro-optical interface (i.e., about 1 Gbps).

In Figure 1.1 we show a basic model of the lightwave communication network. The users of such a network can be viewed as digital electronic devices with an electro-optical interface to the optical-fiber cable plant. The scope of this research is to consider the quantitative and qualitative design of such a network for use as a metropolitan area network (MAN).

1.2 Future Networking Requirements and Trends

In light of people's desire and need to communicate, we expect that networks will continue to evolve in the future much as they have in the past. The obvious trend in networks has been toward networks with higher transmission speeds, larger user populations, and higher throughput per user. Of course, other trends can also be discerned, such as the increasing functionality of communication services, but we now focus primarily on system performance, cost, and reliability.

Several requirements for the MAN of the future can be postulated in the following areas:

- Geographical dispersion
- Large network population
- High throughput
- Low delay
- Low cost
- Dependable service

Given that the scope of this research is MANs, it is reasonable to consider the requirements that such networks will satisfy. Although networks such as DQDB and FDDI have been proposed as MANs, we view them as first-generation design efforts and hold forth the expectation that future generations of MANs will exceed the current capabilities on many fronts. The transmission speed and user population of the future MAN will certainly surpass those of today's MAN in much the same way that the transmission speed and user population of today's local area network (LAN) has surpassed those of the first LAN. Using the original Xerox Ethernet or the Cambridge Ring as examples of the proto-LAN and FDDI or DQDB as their potential replacements, we see that transmission speed has grown from 3 megabits per second (Mbps) to at least 100 Mbps, which is better than a 30-fold increase. The growth in user populations is less straightforward to track and is dependent upon the types of problems to which the network is being applied; but the proto-LANs rarely served more than 100 users, and the emerging generation of LANs set design limits of 1000 users per network. Thus, we could conservatively expect to see at least a 10-fold increase in transmission speed and user population of a typical MAN of the future.

An extremely important consideration in MANs is geographical span, i.e., the radius of coverage of the network. Again, using the LAN as an analogy, we have seen the LAN's physical dimensions grow from about five kilometers (km) to about 100 km. As the size of the LAN is strongly influenced by the type of multiaccess protocol used, we estimate a 10-fold growth in physical dimensions of the MAN. At any rate, MANs will be required to serve sprawling metropolitan areas such as Los Angeles–Long Beach or Tokyo–Yokohama, cities which currently cover geographical regions with spans of up to 60 km. A design goal is to be able to support a region measuring 100 km in radius.

Hand in hand with the growth in network population comes a need for corresponding growth in performance. As computer performance increases and new applications arise to fill the "bandwidth vacuum", the load offered by each user will also increase. Thus, the MAN must achieve significantly higher levels of throughput in the future. The increase in throughput must be accomplished in a manner that does not create a negative impact on delay.

As more and more people come to rely upon the MAN, the economic motivation to ensure that service is delivered reliably becomes paramount. As a basic infrastructure, the MAN will actually provide a service that many individuals will depend upon for their livelihood, safety, and general well-being. The issue of security, which is a *sine qua non* for a large number of users, is perhaps best dealt with as another aspect of dependability [Jos88]. Users (and providers) of a service have a tacit expectation that the service provides integrity and privacy. Thus, a violation of security, whether through malicious intent or not, should be viewed as a failure of the service, and the mechanisms to guarantee secure service are best incorporated into the design of the service's fault tolerance.

One of the benefits that one expects from a large-scale communication system of the scope of a MAN is the realization of definite economies of scale. The user's cost per unit of service should be advantageously competitive with that of other forms of service. Since the development and operating costs of the network will be indirectly transferred to the user, it is essential to consider the use of structures that minimize these costs.

1.3 New Directions in Communication Architectures

Although the performance characteristics of optical-fiber and digital electronic technology continue to advance at a rapid pace, we are quickly arriving at a point where the information-handling capacity of the optical component of Figure 1.1 far exceeds that of the electronic component of the network. While it is possible to transmit and receive lightwave signals at rates of terabits per second (Tbps), cost-effective electro-optical interfaces that convert electrical signals to lightwave signals (and vice versa) at these rates are unavailable. It would be extremely difficult to build the electronics needed to write and read data to and from a communication medium operating in the Tbps range. Since no individual user in Figure 1.1 can utilize the full bandwidth of the optical component, each user represents a potential bottleneck in the network.

A user needing to communicate with another user is thus presented with a torrential "river of bandwidth" which beckons the users with a promise of high-speed information transfer but is inaccessible because each user is incapable of driving the electro-optical interface at the necessary speed. One option is to package the bandwidth of the communication subsystem so that the slower electro-optical interfaces of the user may effectively tap in. Although the performance of single-channel lightwave communication systems, as gauged by the speed–distance product, is growing phenomenally, the more useful alternative of multiplexing several high-speed lightwave signals onto a single optical fiber promises a multiplicative improvement in the potential bandwidth of computer-communication systems. This approach, called *wavelength-division multiplexing (WDM)*, makes possible the use, on a single optical fiber, of several independent, noninterfering signaling channels that operate at rates compatible with the user's electro-optical interface. This is, at least in principle, a means of providing each user with a manageable portion of an enormous aggregate bandwidth.

This fundamental shift in technologies and their tradeoffs forces the computer communication network architect to consider new architectures for connecting computers. Realizing that one can use fast packet-switching based on high-speed digital electronics made possible by very large-scale integrated circuits, researchers have proposed LAN and MAN architectures that operate by transmitting a packet from source to destination with switching via intermediate stations, e.g., the Manhattan Street Network (MSN) of Maxemchuk [Max85]. This approach, termed *multihop*, stands in contrast to conventional LAN and MAN architectures, which typically allow delivery in a single hop by means of complete sharing of the transmission medium. The multihop approach can result in networks that have, in comparison with conventional LANs and MANs, much higher total throughput, wider geographical dispersion, and larger populations of stations. One of the most attractive and exciting proposals for large, high-speed LANs and MANs, originally known as the Multichannel Multihop Lightwave Network and later renamed ShuffleNet [Aca87, AKH87, HK88, AK89], is based on the premise that optical fibers of the future will offer essentially infinite aggregate bandwidth that can be unbundled and repackaged for individual network users in bite-size chunks in the form of WDM channels.

The target of study in this research is the *Wavelength-Division Optical Network (WON)*. The WON is a new class of network architecture that can be used as a MAN for high-speed communication between a large population of users. The WON achieves very high throughput by using the following four principles of operation:

- Message parallelism: several messages can be in transit in the WON at any given instant of time.
- Multichannel WDM: a number of WDM channels are available on the optical medium.
- Wavelength selectivity: each station can be programmed to receive and send on specific WDM channels.
- Malleable topology: although every WON has a *physical* topology, which is static over the network's operational lifetime, its *virtual* topology, which specifies the logical interconnection between stations, can be changed with modest effort.

The WON, based on the combination of the multihop approach, multiple WDM channels, and wavelength selectivity, generalizes networks such as the MSN and

ShuffleNet by permitting a wide range of station interconnections. Although the WON clearly enables one to construct networks with plentiful aggregate bandwidth, it is a nontrivial problem to structure these networks in a way that yields acceptable cost, performance, and dependability. The richness of possible interconnections in the WON also imposes on the designer a number of complex design decisions to be resolved. We intend to undertake in this research a careful study of these design decisions and their impact on the operating characteristics of the WON. The principal issue addressed in this paper is the design of WONs with the goal of meeting specific cost and performance requirements. Although the original architects of the multihop approach have considered certain fundamental design problems, we believe that our research goes considerably beyond these pioneering steps and deals with the design and analysis of such networks in a realistic and original way.

1.4 An Overview of Related Work

Several networks have been recently proposed to take advantage of the high bandwidth of optical fibers. We next review some of these proposals and place them into perspective with the WON, which is the network to be studied in our research.

The MSN was one of the earliest high-performance MANs to be based on lightwave technology [Max85]. In the MSN stations are arranged on a grid and joined according to a two-dimensional toroidal interconnection graph. Every station has north, south, east, and west neighbors to which it is connected by a unidirectional link. Thus, each station has two transmitters and two receivers, and these are arranged so that no two transmitters (or receivers) are on links with opposite polarity (e.g., north and south, or east and west). The columns (streets) and rows (avenues) of the grid alternate direction as do the roads in the borough of Manhattan. The MSN makes good use of message parallelism by allowing the concurrent transmission of packets on the network's numerous point-to-point links. The stations can operate as store-and-forward packet switches, but Maxemchuk has advocated the use of deflection (hot-potato) routing and slot-oriented packets of fixed size, which eliminates or reduces the need for buffering in the stations. The MSN is essentially a conventional store-and-forward network with point-to-point links between its simple switches (stations) but with the constraints that all stations have exactly two transceivers and are connected to form a two-dimensional toroidal network. The WON, by contrast, has no such constraints on its physical topology and could not use point-to-point links because of its requirement for a completely shared medium with complete access to the full complement of WDM channels.

The work of Chlamtac *et al.* on lightwave networks has produced designs known as the Store-and-Forward With Integrated Frequency-Time (SWIFT) network [CG87] and Lightnet [CGK88, CGK89]. SWIFT is a multichannel network using a combination of frequency- and time-division multiplexing to switch a packet over one or more channels on its journey from source to destination. In the SWIFT architecture all stations are store-and-forward and share a common medium partitioned into a small number of channels, but each station dynamically tunes its transmitter and receiver according to a fixed schedule which provides multihop paths between all source–destination pairs; the WON, with its abundance of channels, can use a fixed tuning of stations to channels and thus can avoid the need to coordinate the dynamic tuning of transmitters and receivers, which would anyways be difficult since light sources and detectors typically require tuning times that are orders of magnitude greater than message delivery times. Lightnet, on the other hand, takes the more radical approach of using all-optical components, thus avoiding the problems inherent in electro-optical conversion (except at the end users). Lightnet establishes so-called light pipes over WDM channels from a source to a destination. This approach, although promising, will require technological advances that might not materialize in the near or medium term. Thus, we view it as a possible long-term solution.

Blazenet and Blazelan [HC87, HC] are a network architecture proposed for use as LANs, MANs, and wide area networks (WANs). It, like Lightnet, is an all-optical network that uses optical fiber as a storage medium, analogous to an acoustic delay line. As an all-optical network, we do not consider Blazenet to be implementable for some time. Moreover, Blazelan uses point-to-point links, similar to conventional WANs.

Acampora and his colleagues at AT&T and Columbia University have recently proposed and studied ShuffleNet [Aca87, AKH87, HK88, AK89], which is a lightwave network based on multiplexing a large number of WDM channels on a single optical fiber shared by all the networks stations. ShuffleNet has been proposed as a LAN/MAN capable of providing very high throughput to a large number of network users by taking advantage of the message parallelism inherent in WDM. An advantage of ShuffleNet is that it uses a great deal less optical fiber than point-to-point store-and-forward networks such as the MSN. Thus, it would have a lower cost to install and maintain than such networks, yet achieve the same level of performance. The logical interconnection of stations in ShuffleNet is based upon the recirculating perfect shuffle, which provides very short paths between all of its stations. The WON can be viewed as a direct generalization of ShuffleNet that permits the use of wavelength agility and a designer-specified virtual topology.

We also mention the influence of an early broadband network, Sytek's LocalNet system [Bib81, EF83], which was based on the analog technology of community access television (CATV). LocalNet used frequency-division multiplexing (FDM) to divide the 300-megahertz bandwidth of a CATV distribution system into over 100 channels, each operating at a speed of 128 kilobits per second. The FDM channels were shared by stations with radio-frequency (RF) modems that used the carrier-sense–multiple-access protocol with collision detection (CSMA/CD), and channels were connected by means of router or bridge stations attached to multiple channels. The multichannel, multihop character of LocalNet makes it one of the early forerunners of the WON. The major difference between networks like LocalNet and the WON is quantitative rather than qualitative. The analog CATV technology of LocalNet is inherently bandwidth-limited to less than 1 Gbps, and the number of high-speed FDM channels that the cable can support is small. A single-mode optical fiber, on the other hand, has a much higher intrinsic bandwidth and, as we shall see, could support over *one thousand* 1-Gbps WDM channels. Because both the WON and LocalNet use digital packet switches, it is largely the vast difference in communication bandwidth that distinguishes them; it is also this difference that allows richer network structures to be built in the WON.

1.5 The Contributions and Organization of this Research

The goal of our research is to conduct an in-depth study of the WON. The plan for achieving this goal relies on the decomposition of this task into three coherent problem areas:

- How to design cost-effective physical topologies for the WON
- How to design high-performance virtual topologies for the WON
- How to design high-performance, robust routing protocols for the WON

We have addressed each problem, examining a large number of design alternatives and environmental conditions, and many subproblems have arisen in our investigation of each problem area.

Our research into the design of the physical topology shows that it is possible to design minimum-cost physical topologies for the WON, given a simple cost model of the WON and the class of topology to be designed. We apply the algorithms to a number of different network geographies and topological classes and draw conclusions as to the relative merits of the different classes. Depending on the requirements levied upon the WON (e.g., optical power budget, reliability, performance, and security), the most cost-effective topological class in all cases studied is the Tree-Net topology.

After showing the difficulty of designing optimal virtual topologies for the WON, we propose algorithms for finding virtual topologies with nearly optimal performance, given the traffic requirements and an existing physical topology. We apply the algorithms to a wide range of problems and compare the quality of solutions to that achieved by existing proposals based on structured virtual topologies that can be algorithmically constructed. We are, without exception, able to find better solutions using our algorithms, identifying virtual topologies that excel in both delay and throughput. We observe that, given a fixed network geography, the class of physical topology chosen for the WON can have an effect upon the performance that can be achieved by optimizing the virtual topology. There is a basic tradeoff in which more costly physical topologies allow us to achieve better performance by optimizing the virtual topology than do the less costly physical topologies. We also observe the phenomenon—called distributed cut-through—in which network geographies that cluster stations near the headend of the WON provide sneak paths that can be used to establish minimum-distance paths between source–destination station pairs. In large WONs with excessive propagation delays, distributed cut-through can be exploited by adding near the headend auxiliary stations whose function is to encourage the use of sneak paths which reduce propagation delay.

To prevent the occurrence of performance-degrading congestion in the WON, we propose a routing strategy that rapidly responds to fluctuations in load and permits the admission of high levels of traffic into the WON. The routing algorithm, called detour routing, generalizes and improves upon the well-known deflection routing scheme, which has been proposed for use in high-speed optical networks. We demonstrate by extensive simulation the superiority of detour routing over deflection routing and provide an approach whereby the virtual topology of the WON can be tuned to optimize its performance under the detour routing protocol.

The remainder of the dissertation is organized as follows. A reading of Chapters 1–3 is required to follow any of the succeeding chapters. Having read the three preliminary chapters, however, the reader may proceed to any of Chapters 4–6, as they can be read more or less independently. Reading of the concluding chapter presupposes familiarity with the rest of the dissertation.

In Chapter 2 we describe the architecture of the WON and, since it is a new proposal, give extensive examples illustrating its characteristics and different options. We give basic definitions that will be used later in the dissertation. The chapter culminates with a taxonomy of the WON that illustrates the basic design alternatives available to the network architect.

Chapter 3 provides an overview of the WON design process, which we argue is fundamentally different from the traditional network design paradigm. We develop a new paradigm that is based upon a decomposition of the overall WON design problem into the subproblems of cable-plant design, physical-topology design, virtual-topology design, and routing and congestion control. Of these four subproblems, we factor cable-plant design out of the scope of our research (as it is treated by other authors) and concentrate on the other three areas of design.

Chapter 4 focuses on the problem of finding the physical topology with the least cost. This problem, which applies to the tree and star topological classes, is decomposed into a clustering and a location subproblem. We develop efficient procedures for the solution of these problems and apply them in a series of case studies.

The virtual-topology design problem is the subject of Chapter 5. We discuss the complexity of the problem and propose the use of the simulated annealing algorithm for the dedicated-channel variant and the genetic algorithm for the shared-channel variant. The solution techniques are applied to a large number of cases and the performance of the algorithms is compared against existing solutions.

In Chapter 6 we study the use of the detour routing protocol in the WON. We compare its performance to deflection routing and demonstrate that performance can be greatly enhanced by using a virtual topology that is specially tuned for use with the detour routing algorithm.

A review of our research results and suggestions for further work are given in Chapter 7. In particular, we recommend continued efforts in developing heuristics that can be applied to the optimization of very large WONs, in developing strategies for adaptively and incrementally reconfiguring the WON's virtual topology to respond to failures or fluctuations in traffic, and in defining approaches to use the WON in an integrated traffic environment that requires the transmission of video, voice, and data.

Appendices A–E are mathematical proofs of results used in the body of the dissertation and are not essential to the continuity of the main subject matter. They are included for the sake of completeness.

Chapter 2

The Wavelength-Division Optical Network

The *Wavelength-Division Optical Network (WON)* is a new class of lightwave network and is the object of study in our research. We therefore devote this chapter to the description of the WON and discussions of some basic issues in its design. analysis, implementation, and operation.

2.1 Description of the WON

2.1.1 The Basic Architecture

The WON is a new lightwave network architecture suitable for use as a packet-switching MAN. It is intended to support the integrated transmission of both bursty (e.g., computer communications) and stream (e.g., voice) traffic.

Recalling Figure 1.1, we consider the WON to be divided into an optical and an electronic component. The WON is a hybrid network that uses a combination of optical and electronic technologies to transmit information.

The WON's optical component consists of an optical-fiber cable plant, which is responsible for the transport of optical signals through the network and includes all optical fibers, couplers, taps, connectors, light sources, light detectors, amplifiers, and other necessary optical devices. The physical medium is capable of supporting a large number of WDM channels, which consist of K distinct, noninterfering wavelengths of light that act as carriers for modulating signals and operate independently of each other. We further assume that all signals in the network are essentially digital in nature (including digitized analog signals).

The WON's electronic component consists of a collection of N stations, which perform packet transmission, packet reception, and limited packet switching. Stations also act as points of attachment for the network's users; all transmitted information is offered to and obtained from the network via its stations. A station will generally have a number of independently tunable transmitters and receivers that write to and read from the physical medium. To exchange information, each transmitter or receiver is tuned to a specific wavelength which it modulates or demodulates. The tuning of transceivers, while subject to change over time, is assumed to be relatively stable, and a station will retain its assigned wavelength tuning for an extended period. The station has one input port, one output port, p transmitters, and p receivers. The input and output ports serve as points of attachments for users, which can be computers, subnetworks, or other electronic equipment requiring data-communication services. The transceivers are responsible for transferring data over the optical fiber at WDM-channel rates and must also perform conversions between electronic and optical signals.

The WON, which is a generalization of ShuffleNet [Aca87, AKH87, HK88,

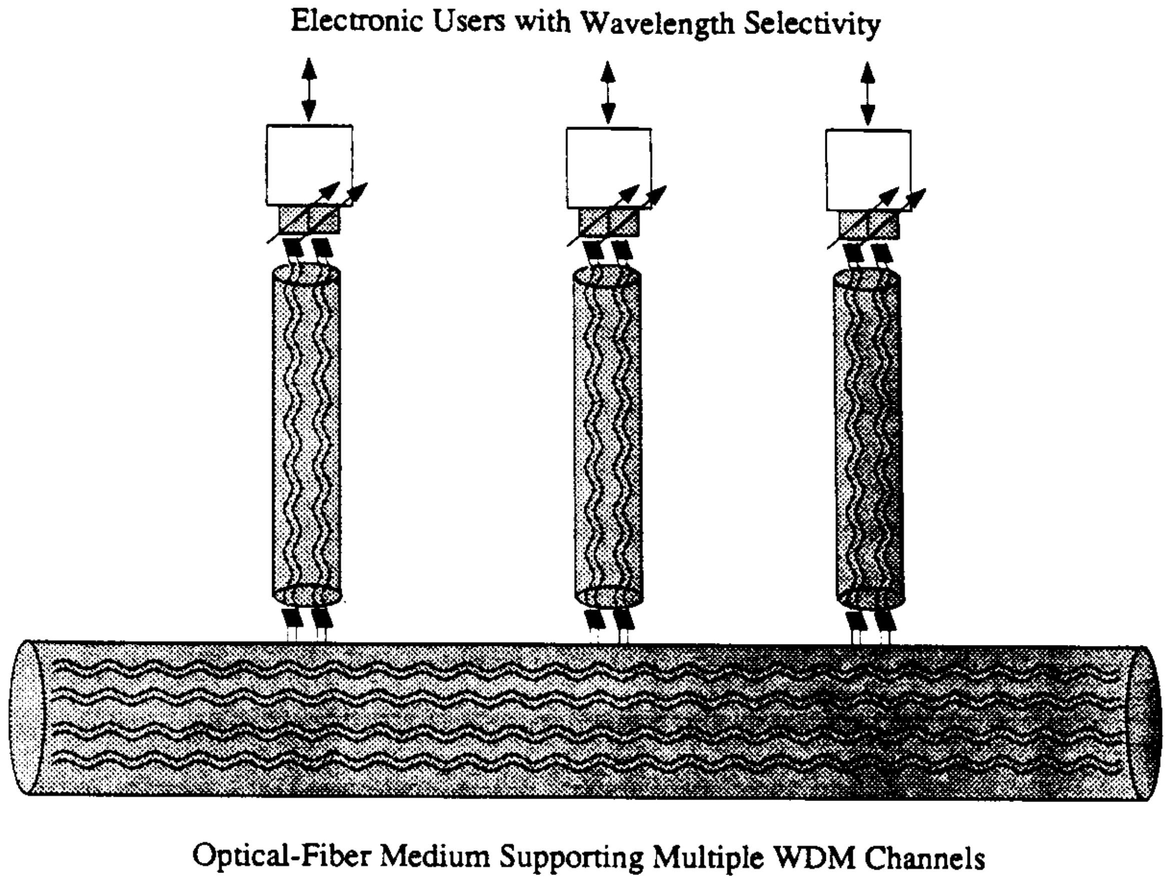

Electronic Users with Wavelength Selectivity
Optical-Fiber Medium Supporting Multiple WDM Channels

Figure 2.1: A Generic WON.

AK89], is a multichannel, multihop network that uses a single optical-fiber medium to provide packet-switching communication services to a population of users. The users—which can be computers, subnetworks, or other devices—are connected to the WON through multiport stations that have a small set (typically two) of transmitters and receivers tuned to particular wavelengths. The stations may use tunable transceivers, but it is not intended that tuning occur dynamically during the transmission and reception of packets, as such tuning procedures are relatively slow and would usually require the station to be taken offline.

The basic elements of the WON are shown in Figure 2.1. The WON consists of

N stations, each accommodating p wavelength-tunable transceivers and one user port. As optical transceivers represent a significant part of the station's cost, it is typical to choose p to be small—generally no more than two. Each station behaves as a simple $(p+1) \times (p+1)$ electronic switch and has the ability to buffer packets temporarily. The stations are connected to a common, shared medium capable of carrying K wavelengths, each of which can be independently modulated and demodulated by the stations' transceivers. The WON's full complement of WDM channels is accessible to all stations, but, since each station is tuned to at most p different WDM channels, communicating with another station requires either retuning one of the transmitters of the sending station or else forwarding a message via one of the station's neighbors. Not only would the former alternative require rapidly tunable transmitters, which are not expected to be available in the near future and which add to the station's cost, but the problems of coordinating the transmissions of tunable transmitters present a formidable challenge [CG87]. The latter alternative thus offers a much more reasonable solution.

Each station may be viewed as a $(p+1) \times (p+1)$ packet switch, as shown in Figure 2.2 (with $p = 2$). Packets arriving via the electronic input port or one of the optical receivers are examined by the station's switching element and routed to the appropriate optical transmitter or electronic output port. Although any kind of routing is allowed in principle, the station must switch packets at link speeds, and thus a simple, streamlined routing procedure—such as hardware-assisted decoding of a source-specified route in the header of the packet—would be used in practice. Each transmitter is supplied with packet buffers to accommodate the queueing of packets contending for service by the same transmitter.

Cell-oriented, fixed-size packets with synchronized arrivals and variable-size

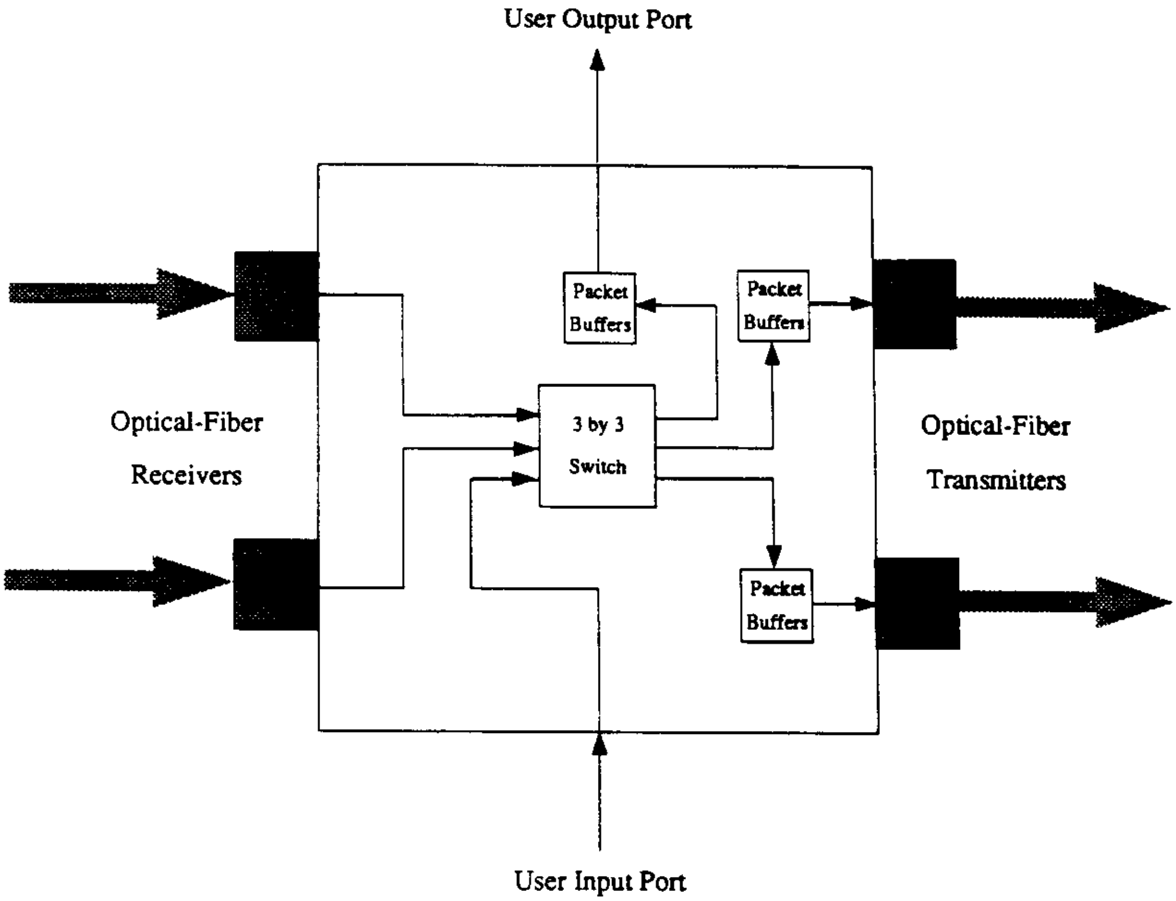


Figure 2.2: The WON Station.

packets with asynchronous arrivals are allowed in the WON. The advantage of using a cell-oriented station is that we can guarantee that buffer overflow does not occur, even with as few as one buffer per transmitter. If packets are synchronized (or artificially aligned) to arrive at the station simultaneously, then a conflict, which occurs when different packets are to use the same transmitter, can be resolved by granting access to one of the packets and forcing the other packet to be sent out via another transmitter. Of course, this disrupts the route of the diverted packet, but it can be rerouted over this new path. In general, cell-oriented communication can only be used when there is at most one transmitter per channel, otherwise buffer overflow is sure to occur, since the transmitter might not get to transmit immediately because of contention for the channel.

In the eight-station WON shown in Figure 2.3, all stations have two transmitters and two receivers, and the WON uses a total of 16 WDM channels. The tuning of this WON corresponds to the ShuffleNet interconnection originally studied in [Aca87, AKH87, HK88] and is based upon the recirculating perfect shuffle. The logical interconnection between stations is more intuitively drawn in Figure 2.4.

The WON delivers each packet from its source to its destination by multihopping, a technique in which a packet starts at its source station and travels to its destination station via a sequence of intermediate stations, undergoing switching and electro-optical conversion at each point along the way. To illustrate, a packet traveling from station 1 to station 5 of Figures 2.3 and 2.4 follows the (shortest possible) sequence of hops: $1 \xrightarrow{\lambda_{12}} 6 \xrightarrow{\lambda_{1}} 0 \xrightarrow{\lambda_{10}} 5$. Multihopping permits the concurrent transmission of many packets on different channels, as opposed to traditional LANs and MANs which permit only a single packet to be in transit at any given time.[1]

[1] This is a simplification, since even traditional LANs and MANs, such as FDDI and DQDB,

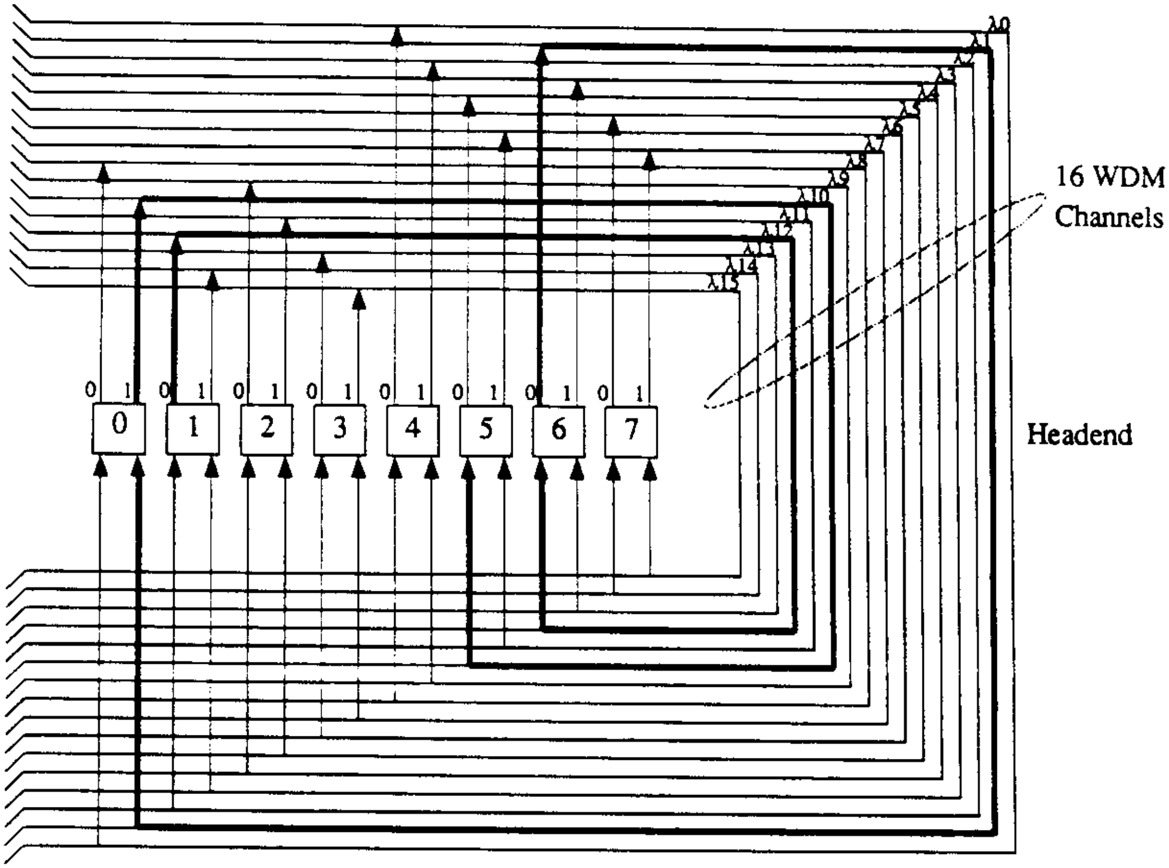


Figure 2.3: An Eight-Station WON.

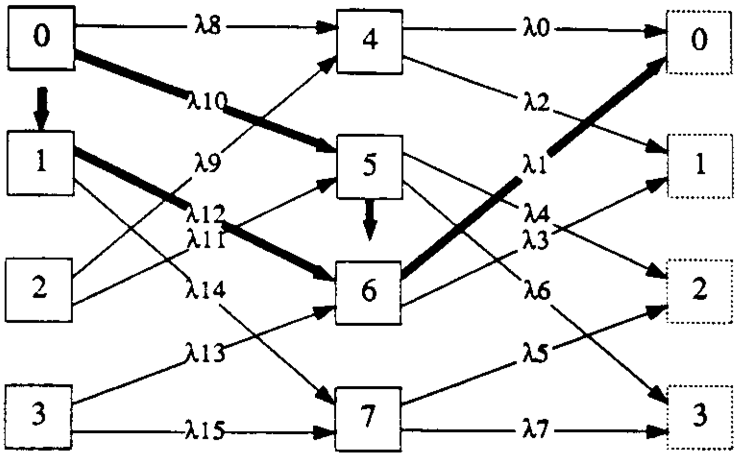


Figure 2.4: The ShuffleNet Interconnection Graph.

The use of multihopping allows the WON to achieve a very high level of message parallelism, i.e., a large number of messages can be concurrently in transit on different channels. It is through message parallelism that the WON attains high throughput; similar throughput levels can not be attained in conventional LANs and MANs because no more than a few messages at a time can be transmitted at any given point in time. In the WON, however, it is theoretically possible to have a message being concurrently transmitted on each WDM channel. Since each message can be repeated on several WDM channels according to the multihop scheme, not all the message parallelism can be realized. Even though the multihopping of messages represents a penalty that can be said to "waste" usable throughput, if the level of multihopping can be kept low enough (i.e., there is not an excessive number of hops per message) then the throughput would remain high. After all, it has been observed [Aca87], bandwidth now represents one of the most plentiful of resources in the lightwave network, and we may therefore afford to be relatively

allow more than one packet to be in transit at a time. The degree of concurrency in packet transmission is, however, very limited in these networks.

less concerned about its waste.

In the WON each station transmits directly to other stations, the receivers of which are tuned to the same wavelengths as the first station's transmitters. Thus, each station can be viewed as directly transmitting to a set of receiving stations. Likewise, each station can also be viewed as directly receiving from a set of transmitting stations. The directed graph with N nodes in which node i has an arc directed to node j if and only if station i directly transmits to station j in the WON is called the *virtual topology* of the WON.

When each WDM channel of a WON has exactly one transmitter and one receiver, it is said to be a *dedicated-channel WON (DCWON)*. The virtual topology of any DCWON can be represented as a *p-regular directed graph*, in which each node has exactly p incoming arcs and p outgoing arcs. In a DCWON each WDM channel behaves like a point-to-point link, so that there is no contention for the channel since its single transmitter enjoys exclusive access. It is nevertheless possible to conceive of configurations in which several stations share a given WDM channel. When a WDM channel has more than one transmitter or receiver, it is said to be a *shared-channel WON (SCWON)*. In the SCWON transmitters contend for access to their shared WDM channel, and it is therefore necessary to use a multiaccess protocol to ensure the orderly scheduling of transmissions. In contrast to the DCWON, the virtual topology of the SCWON can have nodes with an arbitrary number of incoming and outgoing arcs.

2.1.2 Physical Topology

From the most elementary point of view, the WON exists as a physical topology, which encompasses all devices and waveguides required to distribute photonic signals to the WON's users. This includes optical-fiber cables, couplers, connectors, taps, amplifiers, etc. The interconnection of these devices, as well as their actual locations or layout, is included in this definition.

A principal function of the WON's physical topology is to distribute signals to all stations. It is necessary to choose carefully the WON's physical topology in order to ensure that the transmitter-to-receiver power for every pair of stations is in the specified range, i.e., neither too weak nor too strong.

The design of a physical topology must meet the transmitter-to-receiver power budget, i.e., the optical signal launched from a given transmitter must arrive to any other receiver with a specified minimum power. The construction of such physical topologies is challenging because the design must limit the number of points at which losses occur yet provide the required station connectivity. A few special topologies for lightwave networks have been analyzed for their ability to transport optical signals in large networks with acceptable power loss [NTM85, GF88].

Three principal topologies have been proposed for the WON [Aca87]: they are the star, tree, and bus. We show in Figure 2.5 the general form of each of these three topologies and discuss each one individually below.

A particularly promising physical topology is Tree-Net, which was proposed for use as a MAN serving several hundreds of stations [GF88]. The Tree-Net physical topology is—not surprisingly—organized as a binary tree with bus segments or station clusters at the leaves of the tree, as shown in Figure 2.6. The nonleaf nodes of the Tree-Net are simple 3-decibel (dB) couplers, with the exception of the root node, which is a regenerating headend capable of reamplifying all the WDM channels. Each station attaches to its bus segment (which is allowed to be a trivial bus with only one station) by means of a tap. We assume that the power

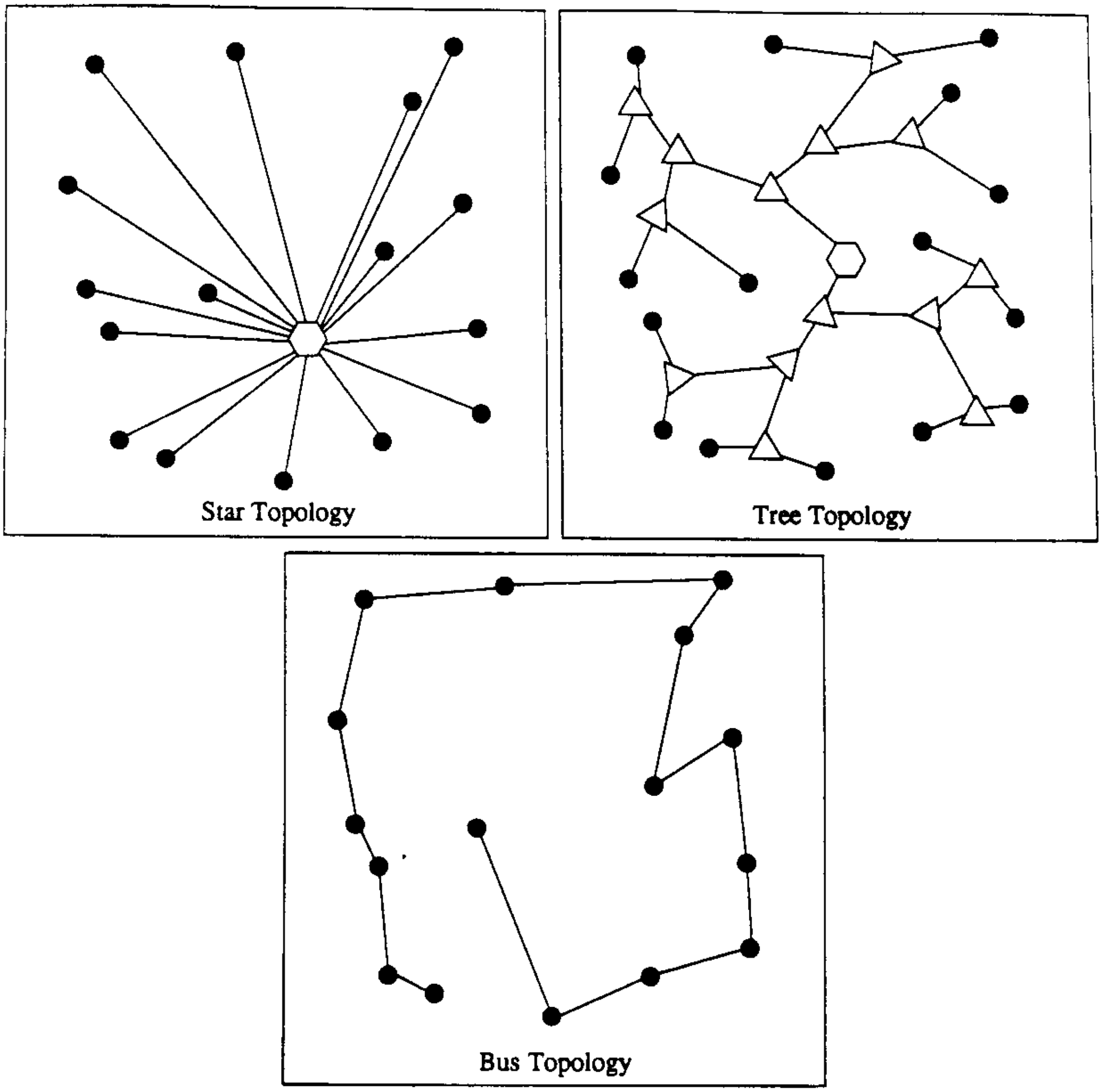


Figure 2.5: Three Principal Physical Topologies for the WON.

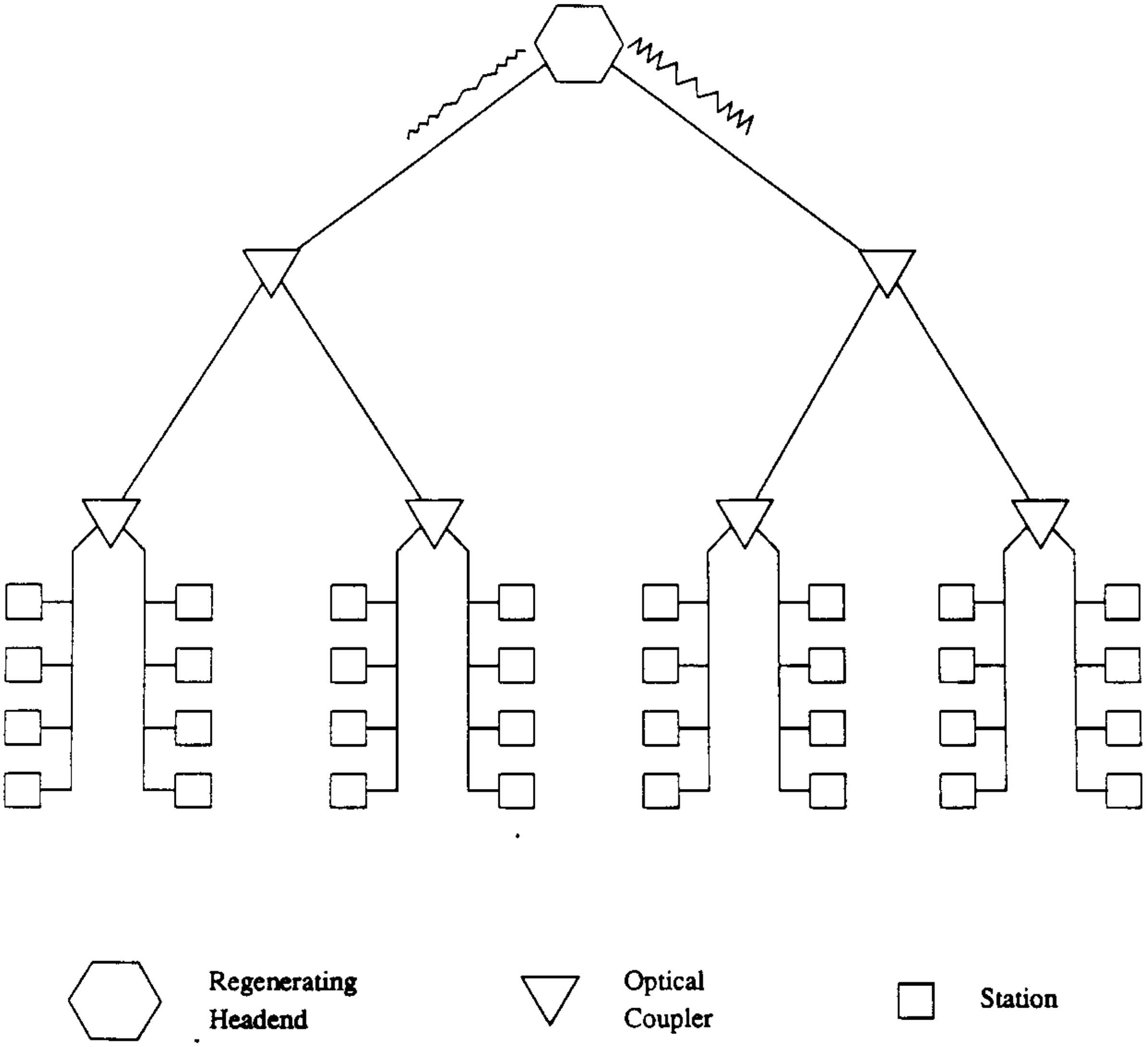


Figure 2.6: The Tree-Net Physical Topology.

budget P, which is defined as the worst-case difference between the light source's power output (in dBm) and the minimum power detectable at the light detector (in dBm), is on the order of 40 dB. We use the following conservative estimates for power reduction:

- Couplers: 3 dB for power splitting and 1 dB for connector loss
- Taps: 2 dB for power splitting and 1 dB for excess loss

If L and K are the number of levels in the tree and the number of stations in a bus-segment cluster, respectively, then the accumulated loss from a station to the headend must not exceed P:

$$4 \text{ dB/coupler} \times L \text{ couplers} + 3 \text{ dB/tap} \times K \text{ taps} \leq P \text{ dB}$$

Considering a balanced tree in which the number of stations N is equal to $2^L K$, we can rewrite this inequality as

$$N \leq 2^L(13.33 - 1.33L) \tag{2.1}$$

The maximum value of N to satisfy Equation (2.1) is 512, which can be achieved for seven-, eight-, and nine-level Tree-Nets ($L = 7, 8, 9$) with clusters of four, two, or one stations per bus segment ($K = 4, 2, 1$), respectively. By improving receiver sensitivity we could also increase the maximum number of stations in the Tree-Net.

The WON can be realized in several physical topologies, e.g., the star, tree, and bus. We show in Figure 2.7 an example of the WON with a star topology, which uses a multiport star coupler at its headend, and in Figure 2.8 an example of the WON with a tree topology. We have already presented in Figure 2.3 an example of the WON with a folded bus topology. We mention at this point that the bus topology has a major drawback in that it can only support a small population

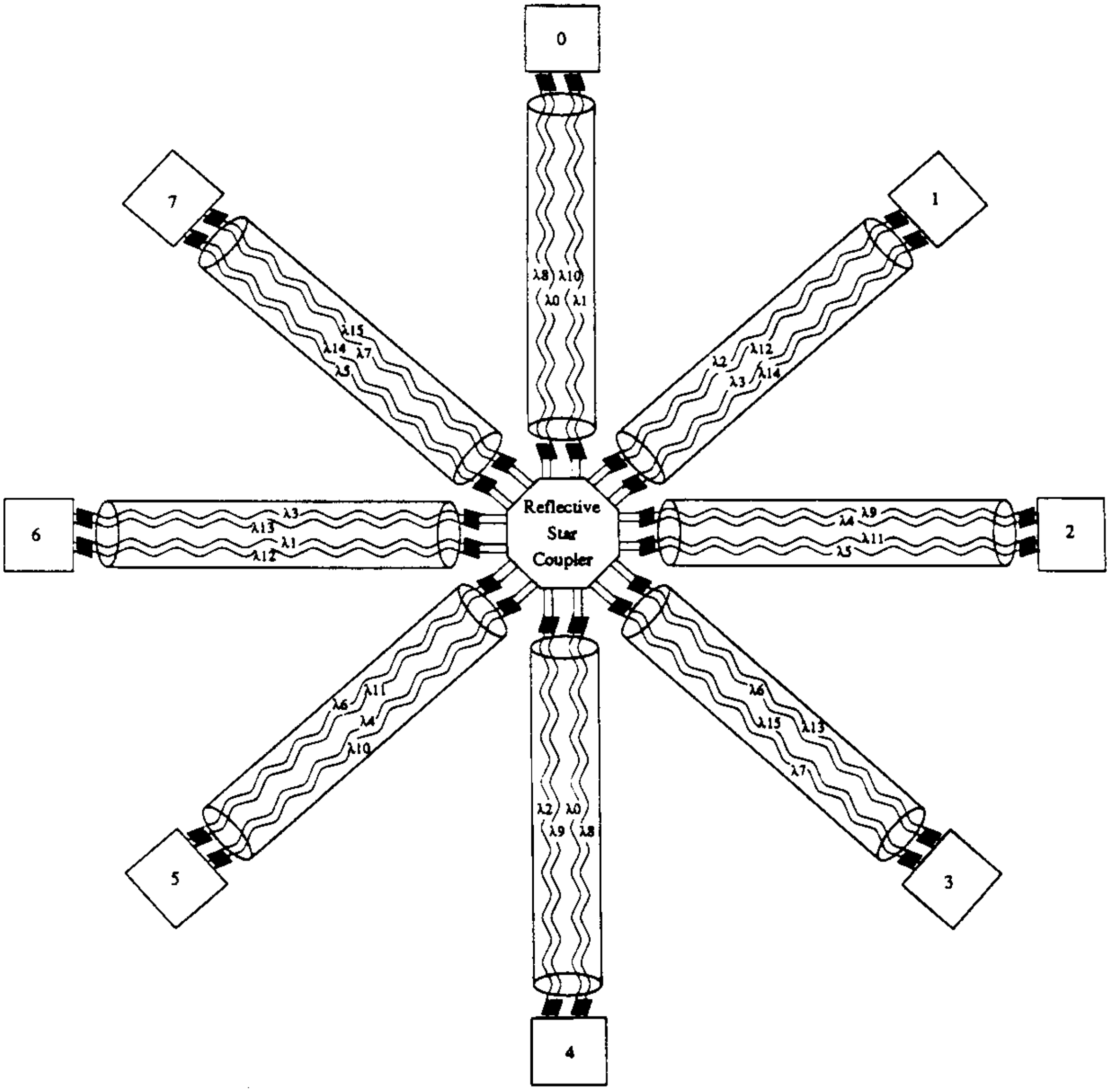


Figure 2.7: The Eight-Station ShuffleNet Realized as a Star.

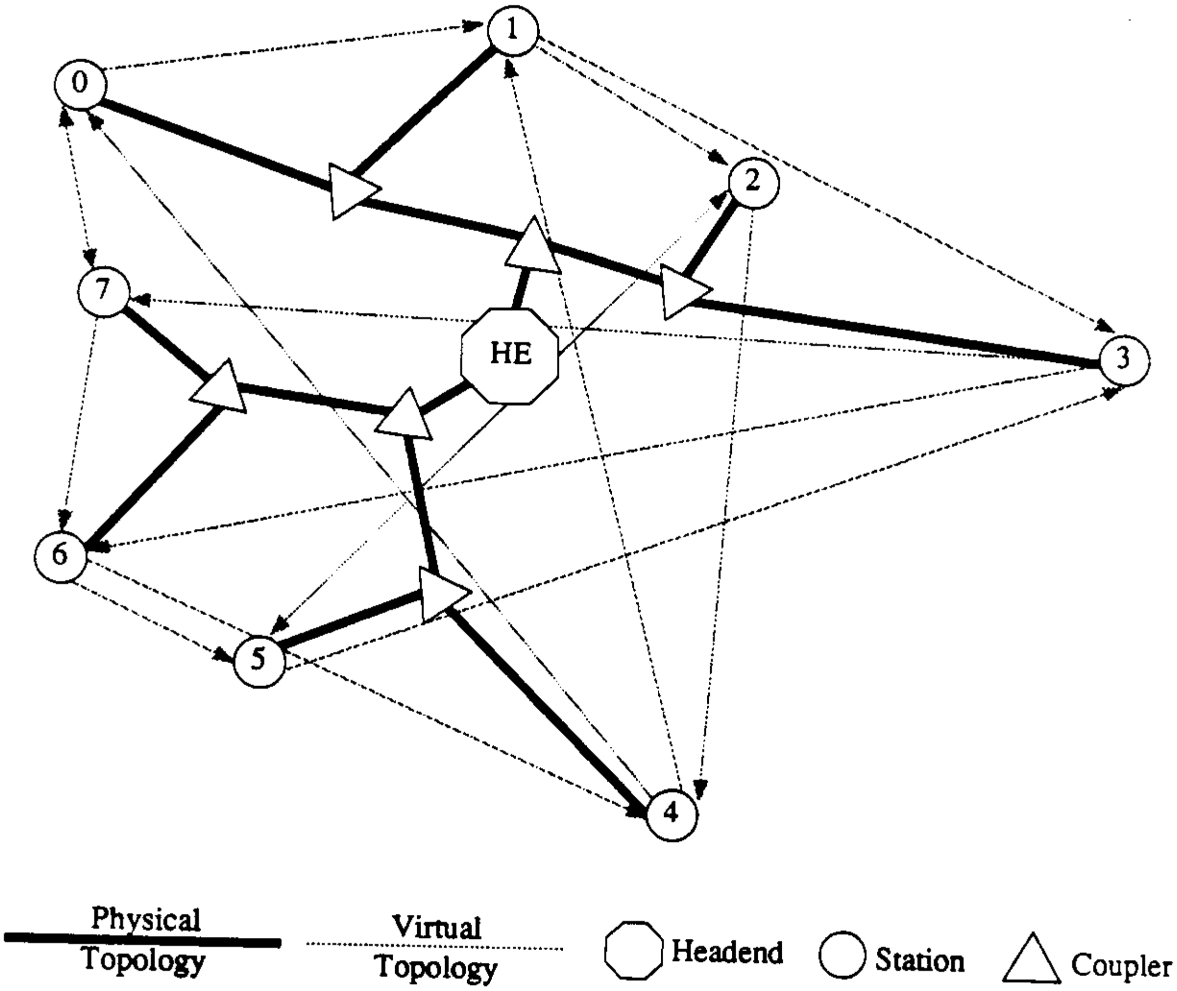


Figure 2.8: An Eight-Station WON Realized as a Tree.

of stations, because the loss of optical power grows linearly with the number of station taps on the bus [Pal88, pages 192–194].

2.1.3 Virtual Topology

An attribute that makes the WON unique among other communication networks is the flexibility that stems from its virtual topology. Numerous virtual topologies can be mapped onto a given physical topology. Figure 2.8 clearly illustrates this mapping principle by showing one specific virtual topology (indicated by shaded lines) overlaid on top of the fixed physical topology (indicated by heavy lines). Clearly, the shaded lines can be redrawn to effect a new virtual topology.

The virtual topology of the WON is naturally modeled as a directed graph (*digraph*) in which the existence of an arc from one node to another implies the cotuning of the corresponding stations' transmitter and receiver. We therefore review some definitions and notation related to digraphs. We assume that a digraph consists of a set of N nodes, usually labeled 0 to $N-1$, and has no parallel arcs, i.e., there is at most one arc from any given node to any other node, and that all the digraph's arcs are labeled with a nonnegative *distance* (or *length*). The *degree* (or *outdegree*) of a node is the number of arcs emanating from that node, and the degree of the digraph is the largest degree of any node. Recall that a digraph is p-regular if all its nodes have degree p; in such cases we will speak of the different arcs as emanating from or entering into (input or output) *ports* of the node, which we usually number from 0 to $p-1$. The distance of any directed path in the digraph is the arithmetic sum of all labels of the path's arcs, and the *shortest distance* from node i to node j is the minimum distance over all paths from node i to node j, denoted δ_{ij}. The *diameter* D of the digraph is defined as $D \triangleq \max_{i \neq j} \delta_{ij}$

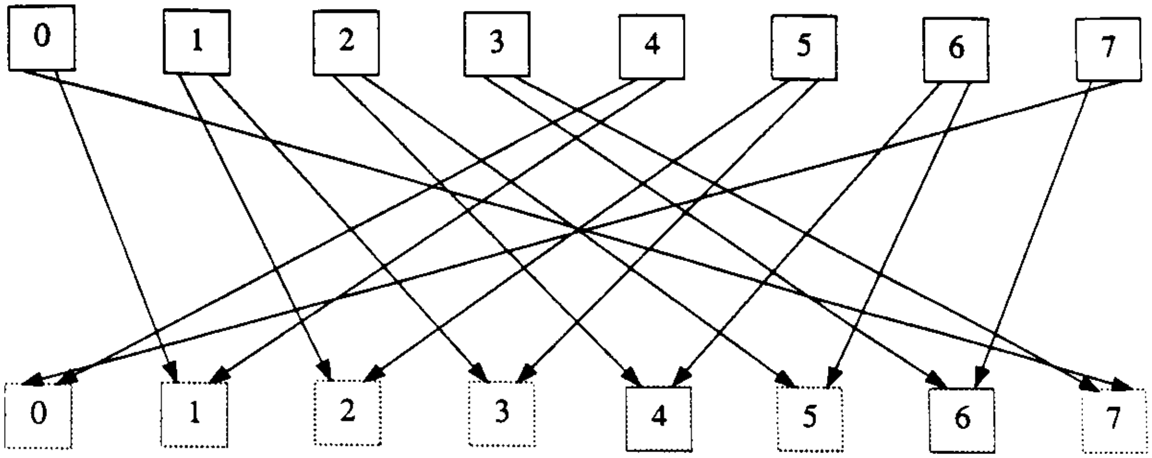


Figure 2.9: A Virtual Topology Based on the Modified de Bruijn Graph.

and the *average internode distance* as

$$\overline{D} \triangleq \sum_{i=0}^{N-1} \sum_{\substack{j=0 \\ j \neq i}}^{N-1} \frac{\delta_{ij}}{N(N-1)}$$

Using the convention that δ_{ii} is the shortest nonzero distance from node i back to itself, we define the *girth* of the digraph to be $\max_i \delta_{ii}$. The *density* of the digraph relates to the difficult problem of how to pack a large number of nodes into a small diameter and is defined as N/D.

The tuning of the WON in Figure 2.8 corresponds to the degree-two de Bruijn digraph on eight nodes which has been slightly modified by redirecting self loops; this virtual topology is shown in Figure 2.9. Note that this digraph is fully connected in the sense that any node is reachable via a directed path from any other node. The p-ary de Bruijn digraph on $N = p^n$ nodes (numbered from 0 to $N-1$) is constructed by connecting port j of node i to port l of node k, where $k = (pi + j) \bmod N$ and

$$l = \left\lfloor \frac{pi + j}{N} \right\rfloor$$

If each node's number is represented in the form of n p-ary digits, then these operations correspond to a left shift of the node's number followed by the concatenation of a single p-ary digit onto its right end—the original number is the source of an arc and the new number is its sink. This interconnection pattern is now slightly modified by eliminating the self loops from each node $i = p^n$ back to itself: we redirect port m of node $m(1 - p^n)/(1 - p)$ to port $(m + 1) \bmod p$ of node $[(m + 1) \bmod p](1 - p^n)/(1 - p)$. The de Bruijn digraph is well known for its high density—eliminating the self loops increases that density even more. The diameter D of the de Bruijn digraph is easily seen to be n since more than n shifts and concatenations will cause the original node's number to repeat. Although we know of no expression for its average internode distance, we derive simple lower and upper bounds on this quantity in Appendices A and B:

$$\frac{np^n - p^{n+1} + np + p - n}{p^n - 1} \leq \overline{D} \leq \frac{np^n + p^n - 1}{p^n - 1} - \frac{p}{p - 1}$$

We can likewise describe the ShuffleNet's virtual topology, viz., the recirculating p-ary perfect shuffle on $N = np^n$ nodes. We can express the adjacency relation of the digraph by specifying that node i's jth outgoing arc is the lth incoming arc of node k, where

$$l = \left\lfloor \frac{i \bmod p^n}{p^{n-1}} \right\rfloor$$

$$k = \left\{ p^n \left\lceil \frac{i}{p^n} \right\rceil + p \left[(i \bmod p^n) \bmod p^{n-1} \right] + j \right\} \bmod N$$

This interconnection produces an n-stage digraph with adjacent stages related by the perfect shuffle. Figure 2.4 shows this virtual topology for $n = 2$ and $p = 2$. One of the most attractive properties of ShuffleNet is its small diameter relative to its size: if we assume that all arcs are labeled with a distance of one, then no two nodes are separated by a distance of more that $2n - 1$, and the average internode distance is [AKH87]

$$\overline{D} = \frac{np^n(p-1)(3n-1) - 2n(p^n-1)}{2(p-1)(np^n-1)}$$

Unfortunately, the simple equations above do not apply if we wish to model a WON with nonuniform traffic, or if the different arcs of the WON's virtual topology have different lengths. These shortcomings highlight the need for more sophisticated and realistic models of the WON, one of which we shall present and develop later.

2.1.4 Channel Sharing in the WON

We might find it advantageous to design shared-channel virtual topologies, in which more than one station can be assigned to a WDM channel. There are at least four reasons to use such shared-channel virtual topologies in the WON:

- Channel sharing makes it possible to use fewer WDM channels than are used by the DCWON, which might be advantageous and/or necessary when there is a large number of stations and/or a shortage of WDM channels.

- Channel sharing makes it possible to implement any virtual topology with only one transceiver per station, which could significantly reduce the cost of a station.

- With shared channels we can implement more connected virtual topologies than are possible with dedicated channels, and these virtual topologies, if they are capable of carrying the network traffic, can provide better performance than the DCWON .

- The sharing of channels affords additional flexibility in routing, which can be exploited to more evenly balance traffic over all channels, thus reducing the chances of overutilizing any given channel. For example, Hluchyj and Karol [HK88] have given a routing procedure for the shared-channel ShuffleNet that routes uniform traffic along shortest paths in such a way that the traffic load on all channels is perfectly balanced (but no such procedure has yet been found for the dedicated-channel ShuffleNet).

Of course, there are added costs that accompany the use of shared channels in the WON, most notably the need for a multiaccess protocol, which can increase the complexity of the station and introduce more overhead into message transmission. Because of the high ratio of packet propagation time to packet transmission time in the WON, the time-division multiaccess protocol would be well suited for use on shared channels.

Figure 2.10 shows a SCWON implemented using four WDM channels and only one transceiver per station. The virtual topology of this SCWON is equivalent to that of ShuffleNet. This could significantly reduce the cost of the station, and the need for only four WDM channels permits a simpler design for the transceiver.

Figure 2.11 shows an asymmetric SCWON implemented using eight WDM channels and one transmitter and two receivers per station. Note that this SCWON's virtual topology, which is equivalent to ShuffleNet, has been defined in such a way that no two transmitters share a common WDM channel. Therefore, there is no need for a multiaccess protocol in this type of SCWON. The lack of a multiaccess protocol controller and the use of one fewer transmitter than in the DCWON imply that the station will have a lower cost than in the DCWON. Moreover, the asymmetric SCWON, with its eight WDM channels, requires only

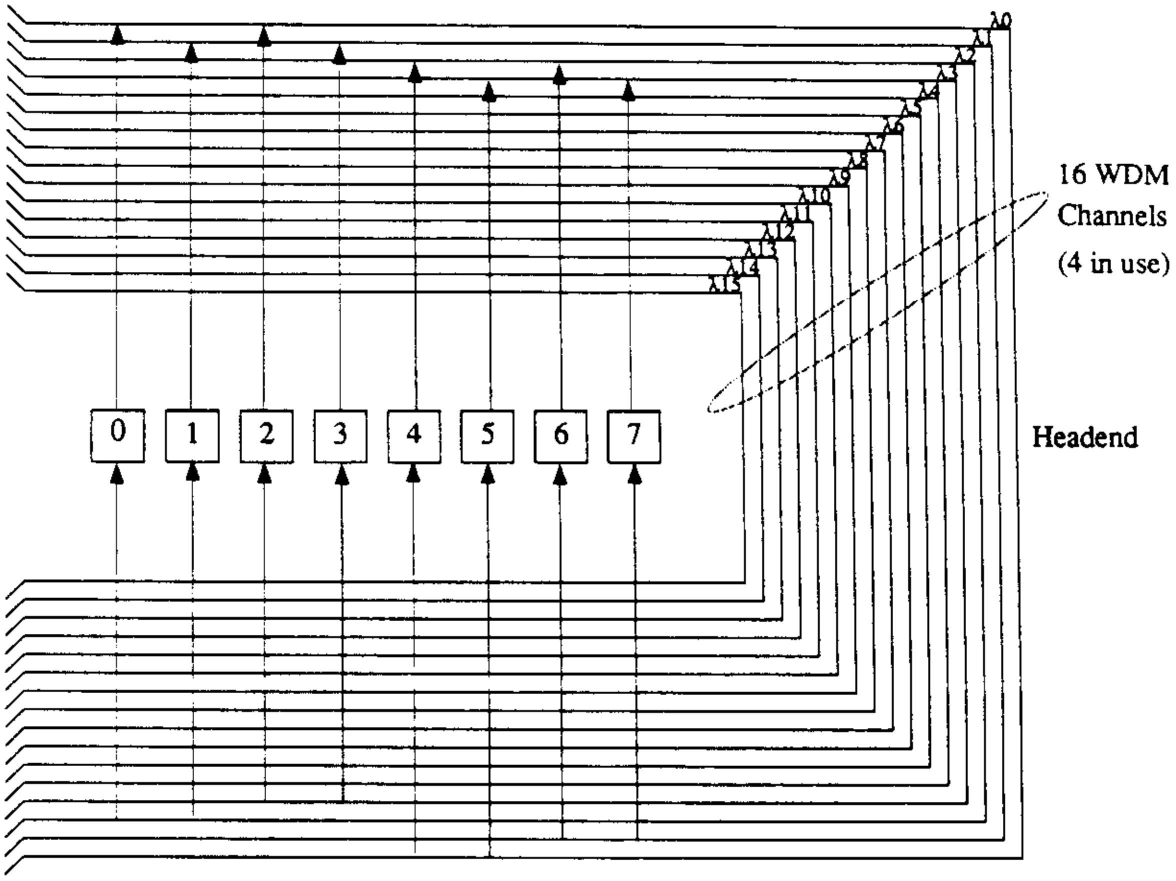


Figure 2.10: An SCWON with One Transceiver per Station.

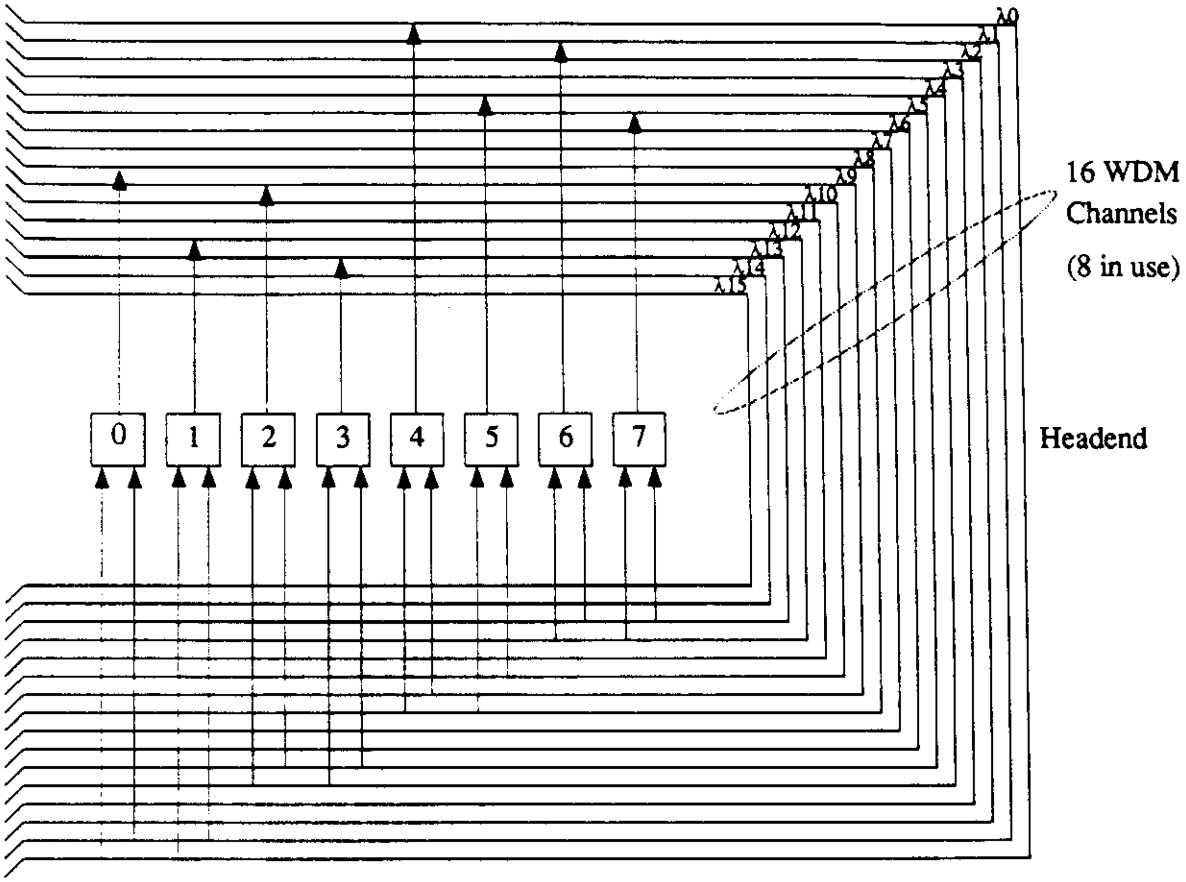


Figure 2.11: An Asymmetric SCWON.

half as many channels as the DCWON.

2.2 Design Goals for the WON

Each WON is intended to meet a number of specified requirements or goals. Three of the most important design goals for the WON are performance, cost, and dependability, each of which we discuss below.

2.2.1 Performance Issues in the WON

The performance of the WON can be measured in several ways. One of the most critical measures of performance is the maximum throughput that the WON can achieve. The maximum throughput is measured as the highest rate that the WON can deliver transmitted information to its users for a sustained period of time. The maximum throughput of the WON relates directly to the level of service that it delivers; more throughput corresponds to a more productive use of the WON. An enterprise operating the WON would generally wish to achieve as high a throughput as possible, as this would imply that more customers could use more network services, thus maximizing the enterprise's return on investment. The flip side of throughput is delay, which is usually defined as the amount of time that a packet takes to progress from its source to its destination and which is critical to many applications. For instance, time-critical data, such as voice traffic, can not tolerate an excessive delay in delivery. Other types of data communication, such as traffic associated with distributed or parallel computation, must have very low latency. For example, the remote procedure call, which forms the basis for numerous distributed computing systems implemented according to the client/server model.

will have to complete with minimal delay.

We generally consider only means or averages when we discuss network performance, but percentiles and distributions can be critical parameters in some situations. As an example, time-critical data may on average meet a specified limit, but even if the limit is only infrequently violated, this will have the same effect as having lost the data. Thus, we might wish to design a WON in which the 99th percentile of packet delay is below a given limit.

Several factors influence the performance of the WON. As in any network, the amount and pattern of traffic offered to the WON strongly affects the performance that it can achieve. One of the strongest influences on performance comes from the virtual topology of the WON, and the WON will exhibit a wide range of performance levels, depending on the virtual topology that it adopts. Other factors, such as the geographical layout of stations and the physical topology of the WON, have an impact on performance, especially delay. Delay is composed of propagation delay, which is the time that an optical signal takes to travel from point to point, and queueing delay, which is the time that a packet spends waiting for resources such as transmitters. The queueing delay of a packet is determined by the WON's policies for accessing resources and the demands being placed upon those resources. Thus, mechanisms such as multiaccess protocols for shared WDM channels and queueing disciplines can have a profound effect upon performance.

Having identified performance metrics such as throughput and delay, we need a way to easily evaluate these metrics if we are to design WONs with optimal performance.

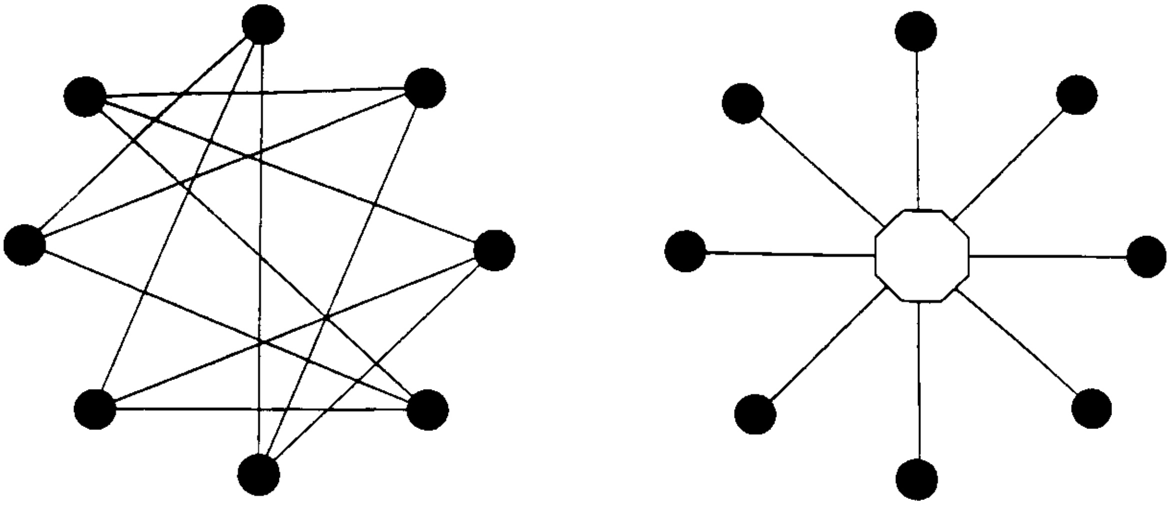


Figure 2.12: Cabling of a Point-to-Point Network Compared to Cabling of a WON.

2.2.2 Cost Issues in the WON

The successful design of a cost-effective WON would be an essential precondition to its acceptance by the market place. Since cost is always a prime concern in large systems, it is necessary to establish firm cost goals in the design of the WON.

Previous work has demonstrated that WONs such as ShuffleNet can provide nearly 1000 stations with an aggregate throughput of over 100 Gbps [Aca87, AKH87, BG89]. The WON, like the conventional point-to-point store-and-forward network, achieves high throughput by using its numerous channels (which correspond to links in the point-to-point store-and-forward network) for the concurrent transmission of packets. We believe, however, that the WON offers several advantages over the point-to-point store-and-forward network in a MAN setting, including:

- Reduced cabling cost. The simple example of Figure 2.12 compares the cabling cost of a point-to-point store-and-forward network configured in a

ShuffleNet interconnection with that of a WON with a physical star topology. The amount of cable in this point-to-point store-and-forward network is nearly 2.6 times the amount in the WON, and it is important to remember that the star is the least cost-effective physical topology for the WON!

- Greater adaptability and ease of configuration. To accommodate evolving traffic requirements or additional stations, the point-to-point store-and-forward network may have to reconfigure existing links, whereas the virtual topology of the WON can be redefined easily and without physical reconfiguration.

The development of a realistic cost model for the WON is not a simple task. The overall cost of the WON is reflected in areas such as cable length, station complexity, and optical component counts. A major contribution to the cost of the WON comes from the expense associated with constructing and maintaining its cable plant. Another source of cost in the WON is the station, whose transceivers, switch, and software all carry a nonnegligible price tag. As we expect the WON to use a large quantity of optical-fiber cable, our cost model will emphasize the dominant contribution that cable length makes to the overall cost. This is because the cost of procuring, laying, and maintaining a length of optical-fiber cable dwarfs the other costs in the WON, especially when the geographical region to be served is large. Moreover, total cable length is good *comparative* measure of cost, since we will usually be comparing the costs of WONs with the same number of stations and approximately equal numbers of optical components.

2.2.3 Dependability Issues in the WON

The term *dependability* refers to that attribute of a system that "allows *reliance to be justifiably placed on the service it delivers* [original emphasis]" [AL86]. A principal aspect of dependability in networks is fault tolerance. Fault tolerance should be designed into the WON from the outset, as the WON would provide a vital service, the loss of which—even in a temporary outage—could have a very detrimental impact on customers. It is important to remember that the WON, by virtue of its size and complexity, is susceptible to failure at several points. In the 2048-node ShuffleNet, for example, the probability that the entire network is fully operational, given that each station is available 99.9 percent of the time and the optical components never fail, is only about 0.13. Thus, it is clear that we can not expect a particular virtual topology to even be in force at a specific instant. This implies that the WON's virtual topology (and its physical topology, as well) should have a measure of robustness designed in.

It is crucial to understand the classes of faults that are expected to beset the WON, so that fault-tolerance mechanisms can be employed to detect, isolate, and recover from these faults. Faults that cause the failure of a station, an optical component, or a segment of optical-fiber are definitely to be expected and guarded against. Both the physical and virtual topologies must be designed to take the above-mentioned faults into account. In addition, protocols at the link, network, transport, and application layers should incorporate mechanisms to deal with noise-induced and intermittent errors.

Security is another aspect of dependability that needs to be addressed in the WON. A foremost concern of users of public or semipublic WONs would be privacy. The WDM-based channel structure of the WON allows each user—at least in principle—to tune to any channel. Unless measures are taken to prevent eavesdropping, an unauthorized user could have access to anyone's data. End-to-end encryption of the user data in each packet would be sufficient to protect data confidentiality, but these mechanisms can be expensive because of the requirement to manage and distribute keys and to install encryption and decryption devices at each station. A less expensive alternative is to separate customers into communities of interest, allow complete sharing of data within a community, and control the exchange of information between different communities. Partitioning the users into communities of interest could also thwart some denial-of-service attacks from outside the community. Such an approach is discussed in Chapter 4 and relies upon both physical- and virtual-topology design.

2.3 Pragmatics and Implementation

2.3.1 Lightwave Technology for the WON

Although we have not focused specifically on the lightwave technology upon which the WON is based, it is important to understand how such a network architecture would be implemented. The current and projected capabilities and limitations of lightwave technology are, therefore, very relevant to the construction of the WON.

There are four basic areas in which lightwave technology is critical to the WON architecture:

- Dense WDM of lightwave signals on a single-mode optical fiber
- Broadband amplification of lightwave signals
- Wavelength-agility of light sources and detectors

- Low-loss coupling of WDM channels

One of the first challenges to building the WON comes in the area of supporting multichannel operation on a single optical fiber. The use of WDM has been demonstrated in the laboratory and has already found limited application in the field [Kap85], but the number of channels used so far is small. To achieve truly dense WDM within a single optical fiber, we will almost certainly require coherent processing of optical signals, rather than direct detection, which is in wide use today. With the use of coherent techniques, we can encode information by modulating the amplitude, frequency, or phase of a beam of light. The encoded information can then be extracted from the coherent waveform by means of heterodyne or homodyne detection.

Coherent processing works in the optical domain much as in the electrical domain. A fixed-frequency (or wavelength) carrier is modulated by an information source, and the transmitted signal is demodulated by mixing the modulated signal with the output of a local oscillator to obtain an intermediate-frequency signal with a much lower frequency than the carrier. Coherent transmission systems are being demonstrated in the laboratory, and their performance is keeping pace with—if not approaching—that of direct-detection transmission systems [Kim87, VW89].

The use of coherent processing allows the formation of a large collection of WDM[2] channels on a single optical fiber. It has been estimated that a coherent lightwave system could support nearly 700 WDM 5-Gbps channels using 50-gigahertz spacing in the passbands centered at 1300 and 1550 nanometers [VW89]. Thus, we could reasonably expect to have several thousand 1-Gbps WDM channels available for use in the WON.

[2] The distinction between WDM, dense WDM, and optical FDM is sometimes made, but we use the term "WDM" to include all these variants.

Coherent systems are now approaching the level of noncoherent systems. There have been several demonstrations of coherent multichannel networks [Lin89], and up to eight 400-Mbps WDM channels have been operated simultaneously.

In the tree-like physical topologies, a broadband optical amplifier at the headend is highly desirable, if not absolutely necessary. Thus, the broadband amplifier is essential to the success of the WON. Such a broadband amplifier would be capable of simultaneously regenerating the optical signals on all WDM channels. As such, the broadband amplifier is the optical analog of the RF remodulating headend that is commonly found in CATV-based networks. The broadband optical amplifier can be an item of considerable expense in the WON and in the worst case could consist of a bank of receiver–transceiver pairs that convert optical signals to electronic signals and back again to optical signals. Furthermore, this device should be highly reliable. Thus, a cost-effective, reliable broadband optical amplifier would do much to promote the acceptance of the WON.

Clearly, the ability to tune the WON stations' transmitters and receivers to specific WDM channels is required in order to define an arbitrary virtual topology. Even more desirable are wavelength-agile transceivers, which can be tuned at any time, not only at the time of manufacture. The tuning and modulation of a light source generally dictates that the source be a laser device. Devices such as the distributed-feedback and the distributed Bragg reflector laser are capable of tuning over a wide range of optical frequencies [Lin89]. The stability and accuracy of a laser light source are also important, and devices that can be precisely tuned and that remain tuned to their designated wavelengths must be developed. Current lasers do not tolerate fluctuations in the ambient temperature very well. Similarly, the problem of setting the receiver to a specific WDM channel must be addressed.

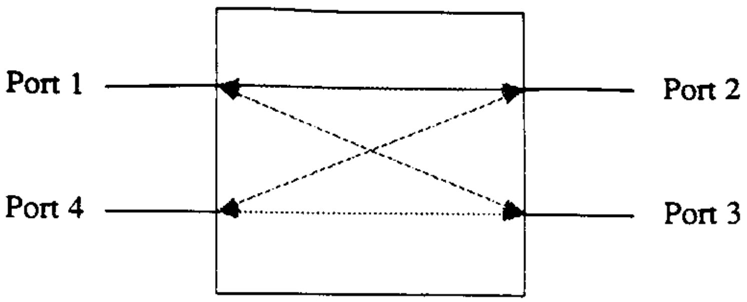


Figure 2.13: The Four-Port Optical Coupler.

In the DCWON, it is sufficient to have receivers that remain tuned to fixed WDM channels; it is then necessary to tune only transmitters. When receivers must be tunable, however, we must solve the problem of accurate and stable frequency synthesis in order to generate the required local oscillator.

The implementation of the WON's physical topology requires coupling devices to distribute optical power throughout the network. The canonical four-port coupler is shown in Figure 2.13. This device is used in the WON to form a junction between four segments of optical fiber. The devices allow the flow of optical power from port 1 to ports 2, 3, and 4. The coupler is often specified in terms of its loss characteristics; if we call port 1 the input port, port 2 the favored port, port 3 the tap port, and port 4 the isolated port, and let P_i denote the optical power incident on port i, then the following definitions apply with respect to the input port (port 1):

- ***Throughput loss***

$$L_{\mathrm{THP}} \triangleq -10 \log \frac{P_2}{P_1}$$

- *Tap loss*

$$L_{\mathrm{TAP}} \triangleq -10 \log \frac{P_3}{P_1}$$

- *Directionality*

$$L_{\mathrm{DIR}} \triangleq -10\log\frac{P_4}{P_1}$$

- *Excess loss*

$$L_{\mathrm{XS}} \triangleq -10\log\frac{P_2+P_3}{P_1}$$

Couplers for the WON would be designed to have throughput and tap losses of about 3 dB for even power splitting, low excess loss, and high directionality (which effectively isolates port 4). Light travels through the optical fiber in a noninterfering fashion, so that light can enter through any of the coupler's ports, as shown by the arrows in Figure 2.13.

The WON makes use of two different types of couplers: the three-port coupler and the star coupler, both of which are shown being used in Figures 2.5–2.8. The ideal three-port coupler, which is illustrated in Figure 2.14, allows no power to reach the isolated port and is simply a four-port coupler with high directionality, i.e., $L_{\mathrm{DIR}} = \infty$. Its function is to merge streams of photons in the upstream (i.e., from station to headend) direction and split streams of photons in the downstream (i.e., from headend to station) direction. It should be designed so that the merging and splitting of optical power are more or less invariant over the wavelengths of interest.

The star coupler is a device with N input and N output ports. If the input and output ports are identical, the coupler is called *reflective*, and if they are distinct, the coupler is called *transmissive*. The function of the star coupler is to distribute power evenly from any input port to all outputs ports. Thus, the ideal star coupler splits incoming optical signals N ways and has an *insertion loss* of $-10\log(1/N)$ dB. The star coupler can be modularly constructed from elemental

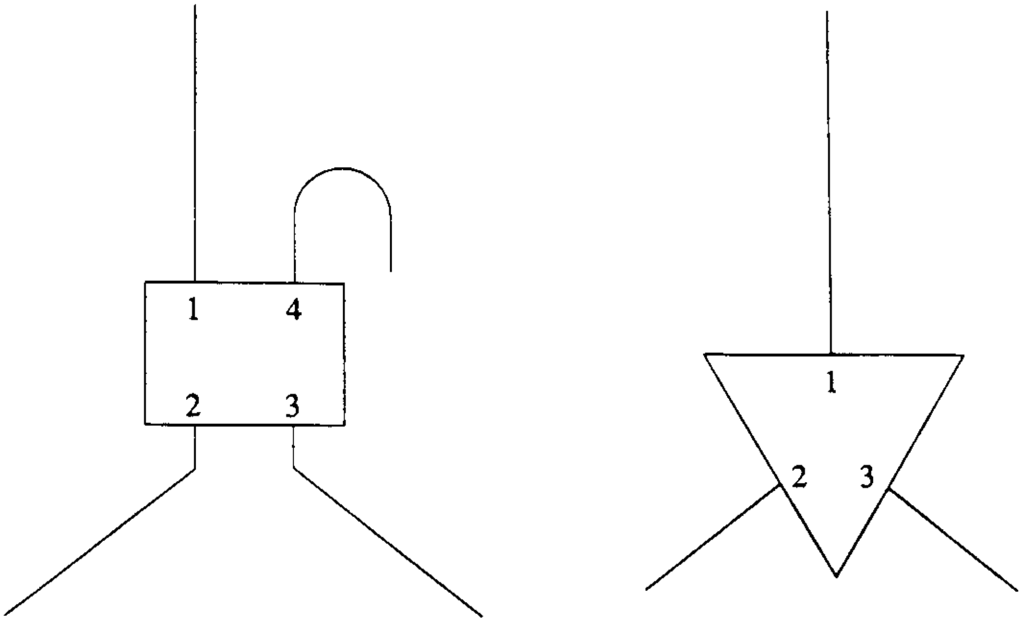
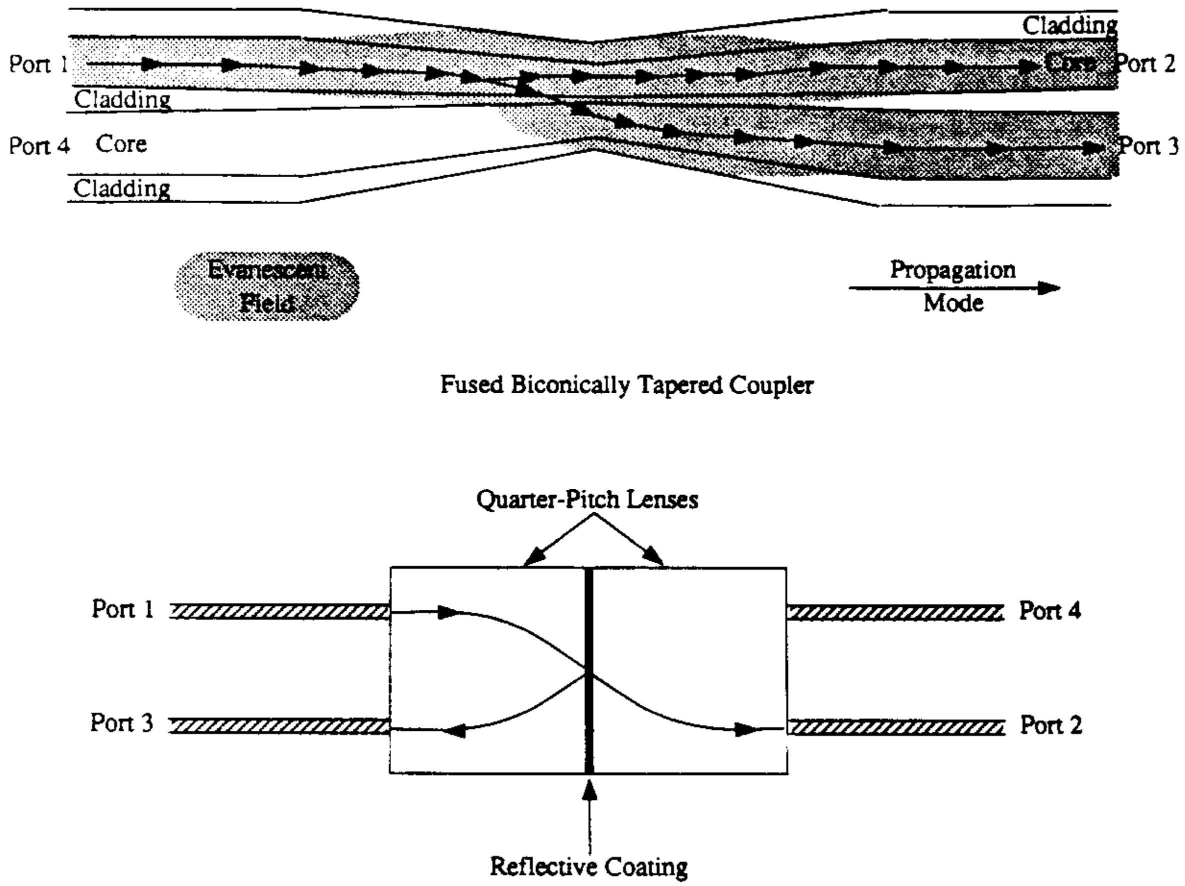


Figure 2.14: The Ideal Three-Port Coupler.

four-port canonical couplers, it can be constructed by fusing together a collection of optical fibers, or it can be constructed by using a planar waveguide [Dra89].

There are several ways to implement the coupler, two of which are illustrated in Figure 2.15. The fused biconically tapered coupler, shown at the top of Figure 2.15, is manufactured by twisting together two optical fibers and then applying heat and tension to fuse and taper the fibers. The optical properties of this construction cause power from the input port to couple into the favored and tap ports. The graded-index rod-lens coupler, also shown in Figure 2.15, uses quarter-pitch lenses to focus light from the input port onto a point of a half-silvered surface, from which it is then focused by the lenses onto the favored and tap ports.



Fused Biconically Tapered Coupler

Graded-Index Rod-Lens Coupler

Figure 2.15: Examples of Optical Coupling Devices.

2.3.2 Multiaccess Protocols for the SCWON

Multiaccess protocols are required by the SCWON in configurations that assign more than one transmitter per channel. With its high-speed WDM channels, the SCWON must use a multiaccess protocol that operates efficiently when the average transmission time is small compared to the end-to-end propagation delay on a channel. Another goal is simplicity, as an unduly complex multiaccess protocol controller could drive up the cost of the station.

Candidates for the multiaccess protocol include protocols as diverse as the Aloha and time-division multiple-access (TDMA) protocols. Recalling that the WON's principal use is as a MAN, we must keep in mind the fact that a WDM channel can have stations that are separated by large distances. Also, the high-speed character of the WDM channels implies that the time to transmit a packet—even a large packet—is short in comparison to the time that it takes the packet to propagate from the transmitting to the receiving station. Another special consideration in the SCWON that distinguishes it from other shared-channel systems is that its WDM channels are *unidirectional*: even though a station transmits on a specific WDM channel, it does not necessarily receive on that channel. Thus, a broadcasting station will not, in general, be able to hear its own broadcast, which contrasts with the typical shared-channel system. This is because the virtual topology of the WON is not constrained by any requirement to cotune transmitters and receivers of a station to the same WDM channel. In fact, such a constraint could severely limit the performance of the SCWON by eliminating certain virtual topologies from consideration or could increase the cost of the station by requiring an additional receiver for every transmitter.

Two schemes that will enable unidirectional channels to be shared are Aloha and TDMA. Aloha would use no feedback from the channel and would merely broadcast without checking the status of the transmission. If a collision occurs, the indication to retransmit would have to come from higher-level protocols. This solution is simple but inefficient. TDMA offers a more attractive solution that would preassign to each station on the channel a periodically repeating time slot with a position that is fixed relative to other stations' time slots. Thus, there must be some way to synchronize the actions of stations, so that they will not usurp each others' time slots. One method would be to provide a beaconing channel that carries a reference signal that is repeatedly broadcast by a designated station or even the headend. Each station would then have an inexpensive sensing port that would enable it to synchronize with the designated beacon station's signal. If each station knows how long it takes to beacon from the designated station to itself, then it can easily figure the time slot in which it has permission to transmit. This scheme would require an accurate clock in each station and a protocol to support the bidding for and assignment of time slots for access to the channel. A drawback of this fixed-assignment TDMA is the inflexibility owing to its inability to accommodate truly bursty traffic sources. Nonetheless, since bandwidth is relatively abundant, we can justify the waste of time slots that usually accompanies TDMA.

2.3.3 Routing in the WON

Because the WON is targeted for high-speed applications, its routing procedures must be streamlined. Routing must be done in real time and at link speeds. Generally speaking, routing can not be done in the WON the way it is done in conventional networks—there is not enough time to look information up in a routing table. Therefore, techniques such as source routing must be employed. In source routing the source, which somehow knows a route to reach the destination, specifies a detailed route from the source to the destination and inserts it into the header of the packet. Each intermediate station uses the information in the header to switch the packet correctly.

Source routing must be supported by a route-discovery protocol that informs source stations about routes to other destination stations. The route-discovery protocol can be invoked periodically or on demand, and routes can be cached at the source station for later use. Usually, shortest-path routing is used in order to minimize packet delay. Sessions between two stations are built on top of a virtual circuit that uses the same route for each packet exchanged in the circuit. This also ensures that the sequence of delivery is the same as the sequence of transmission, if no packets are lost. Error detection and recovery procedures at the virtual-circuit level can guarantee a reliable, sequenced data stream. However, if a link or station fails, routes that include that link or station can no longer be used, and a new route must be discovered and specified by the source station. Thus, the old virtual circuit must be torn down, and a new one must be established.

Also important is the issue of alternate routing when the primary route is experiencing congestion. Unlike the conventional network in which alternate routes are reflected in the routing table as a result of frequent broadcasts of routing and performance information, the WON requires a quick way to choose and specify an alternate route when congestion becomes a problem. A simple method consists of inserting a temporary segment into the the packet's source-specified route, so that the packet can follow the segment (away from congestion). This presumes that alternate paths can be prestored at a station

2.3.4 Network Management in the WON

Given the inherent flexibility of the WON in defining new virtual topologies as the need arises, it is important to have a facility to harness this flexibility. We call such a facility the integrated network management system.

The integrated network management system serves three primary functions: *monitoring* the state of the network, *analyzing* network statistics, and *controlling* network resources. As we shall see in Chapter 5, the problem of designing the WON's virtual topology can actually be viewed as a problem in network monitoring, analysis, and control. In this process a special network management center, which is a combination of computer(s) and administrator(s), is tasked with monitoring long-term trends in traffic patterns. When sufficient traffic statistics have been collected and processed, the network management center is responsible for determining—most likely via analytical modeling—the optimal virtual topology for the WON given the prevailing traffic conditions. Having determined the optimal virtual topology, the network management center must effect a changeover from the current to the new virtual topology. For this we must implement a protocol to control the virtual topology by coordinating the tuning of stations in an orderly fashion. Atomicity of the protocol would be important, as the prospect of a partially defined virtual topology could be disastrous from the point of view of connectivity. Obviously, the redefinition of the virtual topology of the WON is not something that is performed frequently. The monitoring step can require days or weeks, the analysis step can require minutes or hours, and the controlling step can require seconds.

The network management center would also participate in other aspects of WON operation, such as redundancy management in support of fault tolerance, as well as security monitoring and access control. The center could also be involved in the support of the SCWON's multiaccess protocol. For example, in the fixed-assignment TDMA scheme, the assignment of stations to time slots could be mediated by the network management center.

Of course, network management would also be used to provide the services typically required in any network. These services include accounting and configuration management, which are of concern to the network's operating company.

2.4 A Taxonomy of the WON

As we have seen in this chapter, there is a myriad of options that apply to the design of the WON. These different options exist in the areas of physical topology, virtual topology, and routing. Some of the many types of WONs are shown taxonomically arranged in Figure 2.16. This taxonomy describes a roadmap for our research, since it identifies the various design decisions that must be made in implementing a WON, and the goal of our research is to address and solve the specific problems that arise in making the aforementioned design decisions.

The tree in Figure 2.16 has three main branches: physical topology, virtual topology, and routing. Each branch of the tree represents a problem area in its own right. Although other physical topologies might emerge for use in the WON, the four physical topologies shown in Figure 2.16 (viz., star, tree, bus, and ring) have been specifically proposed for the WON and offer vastly different capabilities. We shall explore these physical topologies further in Chapter 4. The greatest number of options seems to arise in connection with the WON's virtual topology and includes wavelength agility, channel sharing, and symmetry. These aspects

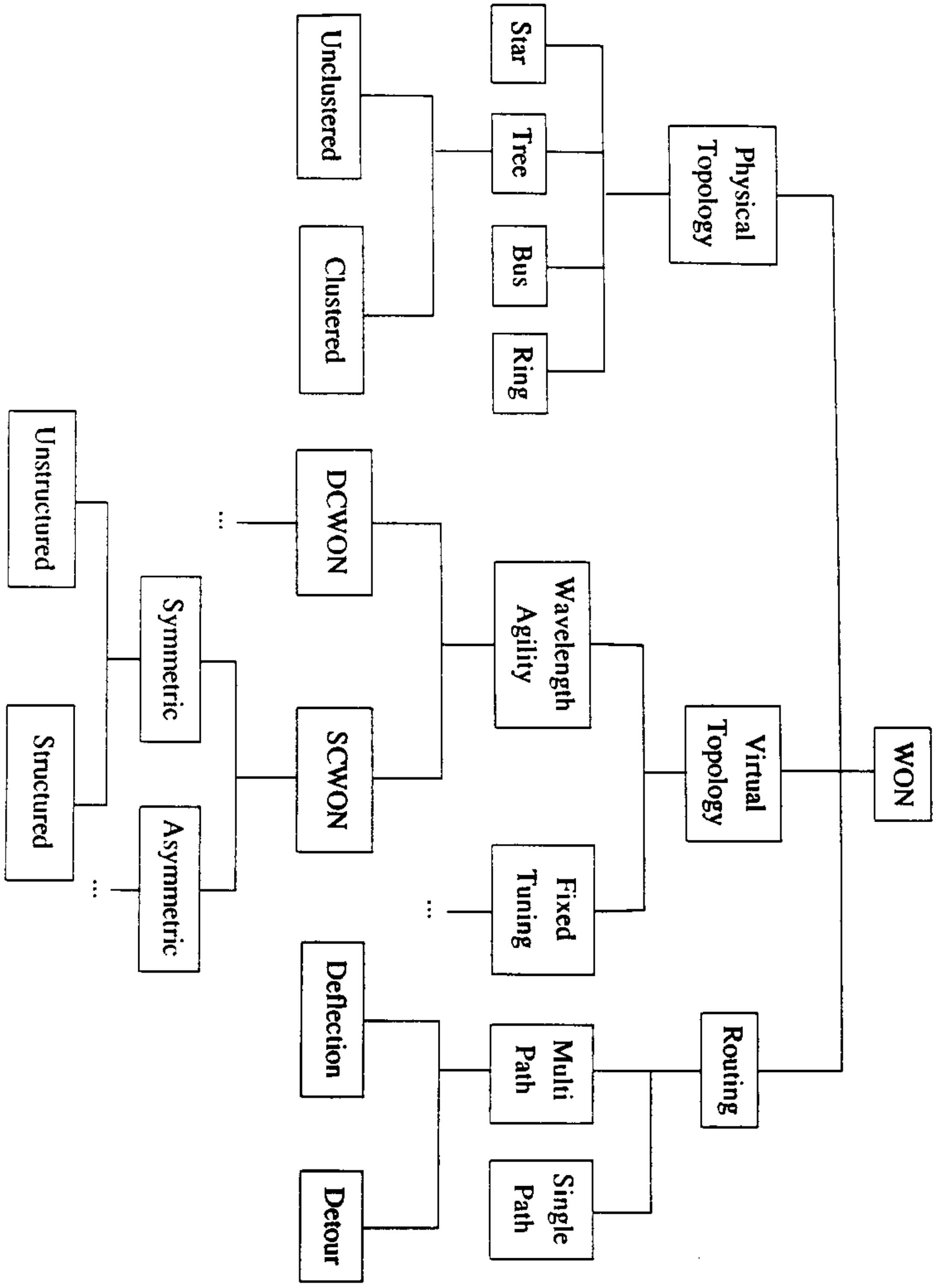


Figure 2.16: A Taxonomy of the WON.

of the WON's virtual topology will be comprehensively covered in Chapter 5. The problem of efficiently transporting packets through the WON—or routing—admits some interesting schemes such as deflection and detour routing, which we will examine in Chapter 6.

2.5 Summary

This chapter describes the architecture of the WON, concentrating on the many different design choices to be made in constructing the WON. The concept of a virtual topology, which can be defined independently of the WON's physical topology, is developed and illustrated by examples. We also describe the kinds of physical topologies that are applicable to the WON. We discuss the cost, performance, and dependability design goals for the WON, and outline approaches to achieving these goals. We mention practical issues in the implementation of the WON, including needed advances in lightwave technology, novel techniques for sharing and controlling WDM channels, and the procedures for fast routing and packet switching. We conclude with a taxonomy of the WON that succinctly summarizes the different design decisions to be made in building the WON.

Chapter 3

A Framework for the Design of the Wavelength-Division Optical Network

Having presented the WON architecture, which is the target of our research, we now formulate a group of problems that arise in the design and operation of this class of networks.

3.1 The Paradigm Shift Away from Traditional Network Design

Major advances in network design coincided with the early development of the Arpanet and other packet-switching wide area computer communication networks characterized by low-speed electronic transmission subsystems. The classical paradigm used in this design process is based on the decomposition of the process into the problems of *capacity assignment*, *flow assignment (routing problem)*, *capacity and flow assignment*, and *topology design*; the interested reader may consult [Ger73, GK77] for a comprehensive discussion of these problems. It is not our intent to give an in-depth presentation of these problems but rather to indicate why the classical paradigm is not adequate for design of the WON.

There are several fundamental reasons behind the paradigm shift in network design, and most of them can be traced back to basic changes occurring in the tradeoffs between lightwave and electronic technologies. With lightwave signal transmission, bandwidth is relatively abundant compared to electronic signal transmission, and lightwave channels can operate at far higher speeds than electronic channels. With WDM channels we can realize economies of scale impossible with copper cabling, i.e., the marginal cost of additional WDM channels is relatively low. Moreover, since electromagnetic signals—whether electrical or optical—propagate at roughly the same speed, the ratio of propagation time to transmission time is much greater on high-speed WDM channels than on low-speed electronic channels. The upshot of this is that a large proportion of a packet's delay in the WON is attributed to propagation delay rather than queueing delay, which is a reversal of the situation in traditional store-and-forward networks. Another consideration is that fast packet switching in the WON demands minimal processing at packet switches, if such switches are to keep up with high-speed WDM channels, and this limits the sophistication of the routing scheme.

Thus, when we consider Arpanet-style topological design, we see that its construction of point-to-point links between stations is not really applicable to the WON, which does not necessarily communicate between adjacent stations and incurs no additional cost when separate signals travel over a common optical fiber.

Similarly, the concept of capacity assignment has no clear analog in the WON, because its architecture assumes the use of wavelength-agile transceivers that can interoperate with each other, so that the wisdom of selecting different capacities for these transceivers is questionable. The advantages of solving for the optimal flow assignment in the WON are expected to be small, because propagation delay dominates queueing delay, so that bifurcated routing reduces delay very little. Furthermore, it is not clear that bifurcateed routing can be readily implemented, since stations must switch packets very rapidly, which suggests that the time spent consulting a routing table could prove problematic.

The WON's use of high-speed lightwave communication has encouraged the development of novel techniques such as the bufferless routing scheme described in Chapter 2. These techniques contrast sharply with those developed for use in conventional store-and-forward networks. Consequently, the analytical models developed to evaluate these conventional networks are not easily adapted to the evaluation of the WON. Therefore, we are often forced to develop new models of the WON, and from the application of these new models arise design problems fundamentally different from traditional network design problems.

Another factor that can be cited in the shift toward a new network design paradigm is that more powerful computers (especially parallel processors) are available today than in the heyday of the Arpanet, and these computing resources can enable the use of formerly cost-ineffective algorithms.

Given the inappropriateness of the classical paradigm for network design, we proceed to establish a new paradigm that has been tailored to the design of the WON.

3.2 Performance Models of the WON

In designing the WON to meet a specific performance goal, we require a straightforward method of evaluating the performance metric under study. In evaluating a specific performance metric such as mean packet delay, we could resort to a very accurate method of evaluation, e.g., simulation; but as simulation entails the use of significant computing time, this alternative can not be seriously considered for designing WONs, which could involve the evaluation of a large number of networks. By making certain simplifying assumptions about the WON, we might be able to construct models of the WON that are easily analyzed, preferably by basic mathematical techniques.

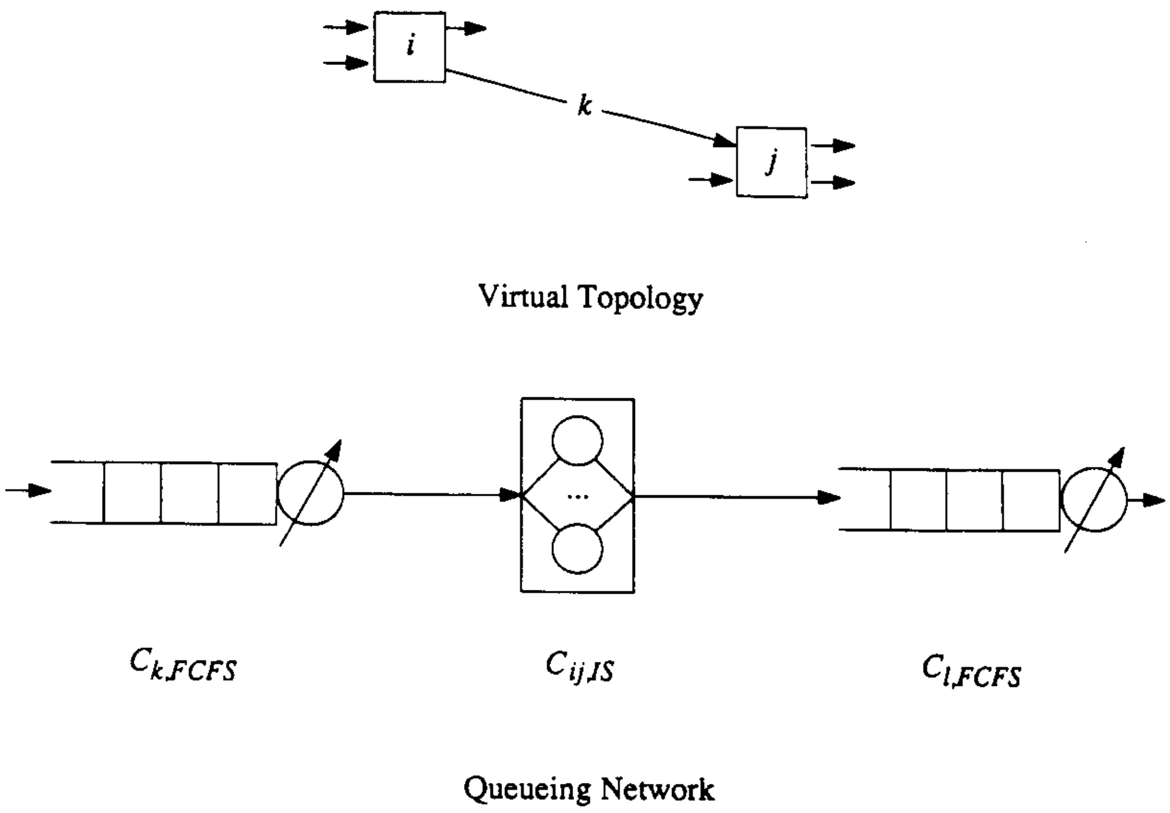
Before developing the mathematical model of the WON, it is convenient to define the following important terms. An $N \times N$ *distance matrix* (δ_{ij}) specifies the distance (in km) between stations i and j. This "glass" distance is measured as the length of optical fiber from station i to station j. The transmission signal propagates along this optical-fiber path at the speed of light through glass, which we take to be $\tilde{c} \approx 2 \times 10^5$ km per second. Thus, the time for a bit to propagate from station i to station j is $\delta_{ij}/\tilde{c}$ seconds. The $N \times N$ *traffic matrix* (γ_{ij}), specifies the rate (in packets per second) at which traffic flows from source station i to destination station j. The total amount of traffic offered to the network is defined to be $\gamma \triangleq \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_{ij}$. When all entries of the traffic or distance matrix are equal, we say that the traffic or distance matrix is uniform.

A simple measure of performance in the WON is the expected number of hops that a packet makes from its entry into the network until its exit from the network. If we define λ to be the sum of the rates of traffic flow on all WDM channels of the WON, then we have the following well-known relationship between the expected hop count $\mathbb{E}[\text{hops}]$, offered traffic γ, and carried traffic λ [Kle76, page 327]: $\gamma\mathbb{E}[\text{hops}] = \lambda$. If two networks have expected hop counts $\mathbb{E}[\text{hops}_1]$ and $\mathbb{E}[\text{hops}_2]$, with $\mathbb{E}[\text{hops}_1] > \mathbb{E}[\text{hops}_2]$, then $\lambda_1 > \lambda_2$ for a given γ. If the routing algorithm balances the traffic on all WDM channels, then the first network will saturate at a lower offered traffic load than the second. Thus, $\mathbb{E}[\text{hops}]$ can be considered to be a rough indicator of the maximum sustainable throughput in the WON.

A central attraction of the multistage recirculating p-ary shuffle and the p-ary de Bruijn graphs is that they both have a very low expected hop count under the assumption of uniform traffic.

Hop-count analysis, although important, has fundamental drawbacks. First, the case of nonuniform traffic has proven difficult to analyze. Second, since hop-count analysis does not take distances into account, hop count does not in general accurately correlate with packet delay, except in a few instances. For these reasons we demand a more authentic model of WON performance.

We model the N-station, K-WDM-channel WON as a Baskett-Chandy-Muntz-Palacios (BCMP) open queueing network [BCMP75]. The service centers of the queueing network are either infinite-server (IS) centers, which model propagation delay, or first-come–first-serve (FCFS) centers, which model queueing delay. Potentially, there may be as many as N^2 IS centers in the network (one for each station-to-station hop), denoted by $C_{ij,IS}$. Service center $C_{ij,IS}$ models the propagation delay experienced by a packet as it hops directly from station i to station j and thus has deterministic service time $\delta_{ij}/\bar{c}$. There are K FCFS centers in the network (one for each WDM channel of the WON). The FCFS service center $C_{k,FCFS}$ models the queueing delay experienced by a packet from the time that



Virtual Topology

Queueing Network

Figure 3.1: Fragment of the Queueing-Network Model of the WON.

it queues for transmission at channel k until the time it is successfully output to the channel. In general, $C_{k,FCFS}$ has a state-dependent service rate that depends upon the number n of packets queued for transmission at channel k. If we denote by $\mu_k(n)$ the service-rate function of center $C_{k,FCFS}$, then the time to transmit a packet of length $1/\mu$ bits when there are n packets in the queue is $1/\mu B\mu_k(n)$ seconds, where B is the channel speed in bits per second. We assume that for each state-dependent service center there exists some limiting value g_k for which $\mu_k(n) = \mu_k(g_k)$ for all $n \geq g_k$. Figure 3.1 illustrates the interrelationship of the FCFS and IS service centers in the queueing network and how they correspond to stations of the WON.

All packets offered to the WON are assumed to have their lengths chosen from an exponential distribution with mean $1/\mu$ bits. We make, for the sake of mathematical tractability, an independence assumption that as a packet completes a hop, its length is independently chosen anew from an exponential distribution with mean $1/\mu$. Packets destined for station j arrive to the user input port of station i as a Poisson process of intensity γ_{ij} packets per second.

With each pair of stations we associate a single route consisting of the sequence of hops used to send packets from the source station to the destination station. Given the routes of all source–destination pairs, we can determine the throughput at each service center of the queueing-network model. The quantities $\lambda_{ij,IS}$ and $\lambda_{k,FCFS}$ give the throughputs (in packets per second) at centers $\mathcal{C}_{ij,IS}$ and $\mathcal{C}_{k,FCFS}$, respectively.

In a BCMP queueing network in steady state, the mean queue length at any service center is identical to the queue length that would result if the center were isolated and driven by the same arrival rate. If we let the random variables $L_{ij,IS}$ and $L_{k,FCFS}$ represent the number of customers at centers $\mathcal{C}_{ij,IS}$ and $\mathcal{C}_{k,FCFS}$, respectively, then

$$\mathbb{E}[L_{ij,IS}] = \lambda_{ij,IS}\, \delta_{ij}/\tilde{c} \tag{3.1}$$

and [LS83, pages 153–155]

$$\begin{aligned} \boldsymbol{\mathbb{E}}[L_{k,FCFS}] &= p_{k0}\, \tilde{\rho}_k \Bigg\{ \frac{\mu_k(g_k)}{\mu_k(g_k-1)} \frac{1}{(1-\rho_k)^2} + \\ &\quad \sum_{i=0}^{g_k-2} i\, \tilde{\rho}_k^i \left[\frac{1}{\prod_{j=2}^{i+1} \mu_k(j)} - \frac{1}{\mu_k(g_k-1)\, [\mu_k(g_k)]^{i-1}} \right] \Bigg\} \end{aligned} \tag{3.2}$$

where $\rho_k \triangleq \lambda_{k,FCFS}/\mu B \mu_k(g_k)$, $\tilde{\rho}_k \triangleq \lambda_{k,FCFS}/\mu B$, and

$$p_{k0} = \left\{ \frac{\mu_k(g_k)}{\mu_k(g_k-1)} \frac{1}{1-\rho_k} + \sum_{i=0}^{g_k-2} \tilde{\rho}_k^i \left[\frac{1}{\prod_{j=1}^{i} \mu_k(j)} - \frac{1}{\mu_k(g_k-1)\, [\mu_k(g_k)]^{i-1}} \right] \right\}^{-1}$$

which is the probability that center $C_{k,FCFS}$ is idle.

Using Little's Result we may express the mean packet delay in terms of the offered load and the average number of packets in the network as follows

$$\mathbb{E}[T] = \frac{1}{\gamma}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[L_{ij,IS}] + \sum_{k=1}^{K}\mathbb{E}[L_{k,FCFS}]\right\} \tag{3.3}$$

By substituting Equations (3.1) and (3.2) into Equation (3.3) we obtain the basic formula for mean packet delay:

$$\begin{aligned}\mathbb{E}[T] \;=\; & \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\lambda_{ij,IS}\,\delta_{ij}}{\tilde{c}\,\gamma} \;+ \\ & \sum_{k=1}^{K}\frac{p_{k0}\,\tilde{\rho}_k}{\gamma}\left\{\frac{\mu_k(g_k)}{\mu_k(g_k-1)}\frac{1}{(1-\rho_k)^2} \;+\right. \\ & \left.\sum_{i=0}^{g_k-2} i\tilde{\rho}_k^i\left[\frac{1}{\prod_{j=2}^{i+1}\mu_k(j)} - \frac{1}{\mu_k(g_k-1)\,[\mu_k(g_k)]^{i-1}}\right]\right\}\end{aligned} \tag{3.4}$$

By choosing the appropriate form of the service-rate function $\mu_k(n)$ we can approximately model several multiaccess protocol schemes, and thus we can model the delay that originates from channel access in the shared-channel WON. The service-rate function can reflect the dependence of the protocol's performance upon the number of stations using the channel and the load that they are generating. The results obtained are in general only approximations. This is because multiaccess protocols do not usually provide service in FCFS order (or, for that matter, according to any of the other BCMP service disciplines). Moreover, channel delay will depend not only on the total backlog of packets awaiting transmission, but also on the way in which this backlog is distributed among the stations sharing the channel. The approximation is adequate for our main purpose, however, which is to compare the relative performance of different designs.

The problem of packet-buffer overflow is common to both the DCWON and SCWON and manifests itself when the transmitter can not keep up with the influx

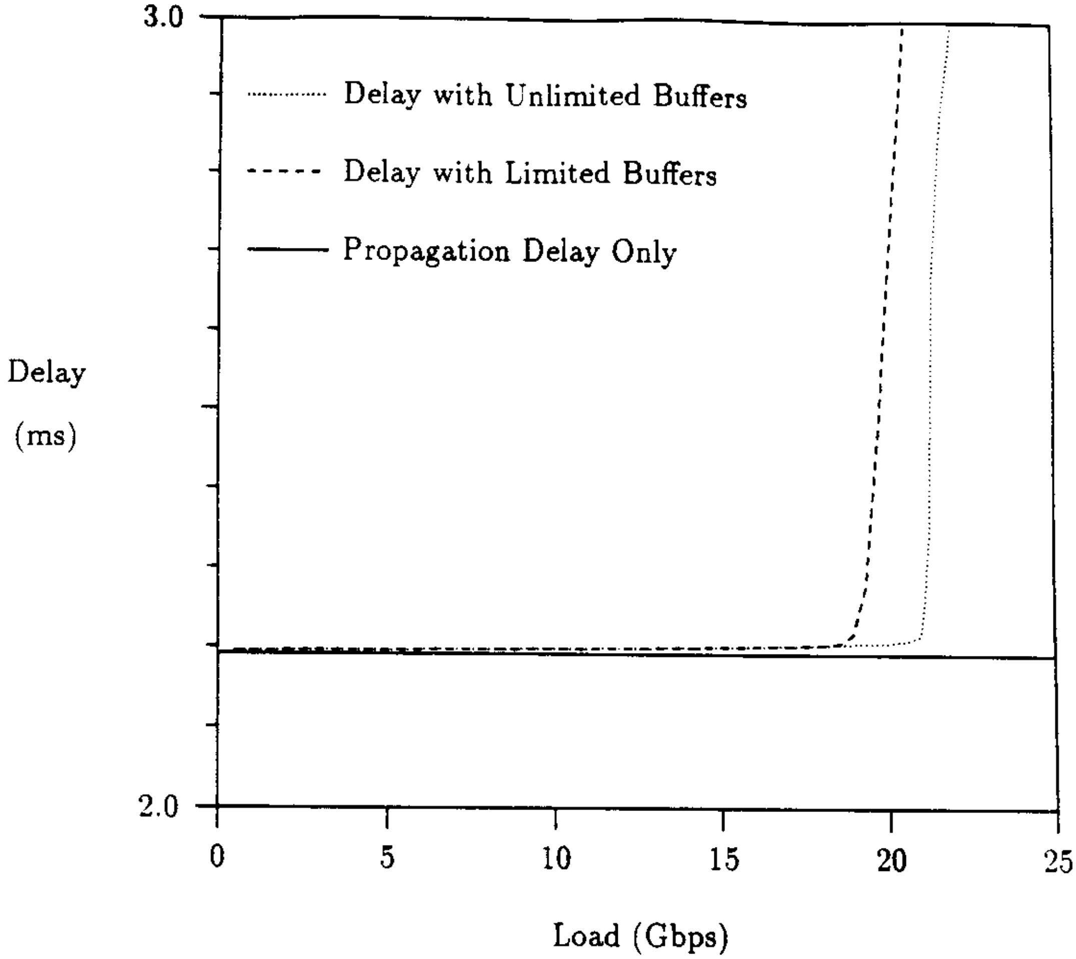


Figure 3.2: Mean Packet Delay in a 64-Station WON.

of packets. The severe impact upon performance of losing (and possibly retransmitting) packets can be modeled by using a lowered service rate when there are several packets awaiting transmission.

We illustrate in Figure 3.2 the behavior of mean packet delay as the traffic loading is scaled upward. The plot, which is for a 64-station two-transceiver DC-WON, assumes a mean packet size of 1000 bits, channels operating at 1 Gbps, uniform traffic, and stations that are all located 50 km from the headend. The dotted curve shows the mean packet delay in the case of unlimited packet buffers; it was obtained by using a constant service-rate function. The dashed curve shows the mean packet delay in the case of limited packet buffers; it was obtained by using a service-rate function that is reduced whenever the number of packets on the channel exceeds a threshold. The solid curve shows only the mean propagation delay, which is invariant as the traffic is scaled upward. We can see that the delay is essentially flat until we reach saturation, at which point the delay grows rapidly and without bound. This phenomenon, which resembles the behavior of the D/D/1 queueing system, is basically a result of two effects:

- There is a large discrepancy between packet transmission time and packet propagation time, which in the example considered above differ by more than two orders of magnitude.

- As we scale the traffic upward, the delay in the network remains relatively flat until a bottleneck channel reaches saturation, which quickly drives the overall delay to infinity.

The sharp knees in the delay curves underscore the need to operate the WON in such a way that all channels are utilized below their critical thresholds.

The plot in Figure 3.3 shows the reasonably close agreement between values of mean packet delay predicted by the analytical model and values of delay measured in actual simulations of the WON. The simulation model uses fixed-size packets of 1000 bits and assumes that channels operate at 1 Gbps, traffic is uniform, and all stations lie 50 km from the headend. The queueing-network model makes identical assumptions, with the exception that packet size is exponentially distributed. Also, in the simulation model traffic is generated by specifying a geometric probability

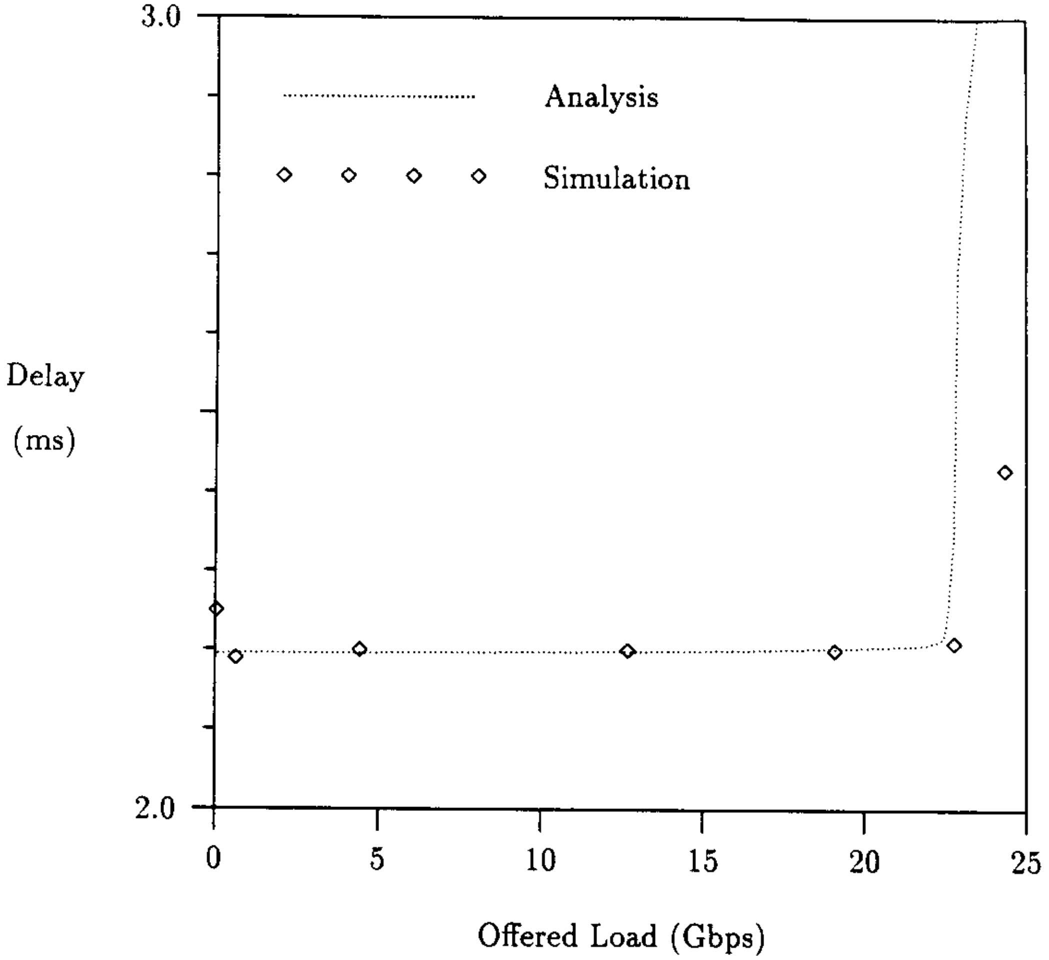


Figure 3.3: Predicted and Simulated Mean Packet Delay in a 64-Station WON.

distribution for packet interarrival times, whereas in the queueing-network model traffic is generated by specifying an exponential probability distribution for packet interarrival times. Both distributions are memoryless and produce comparable arrival statistics, however.

3.3 Cost Model of the WON

Several recurring and nonrecurring costs would be associated with the implementation of a WON. Components, including couplers, connectors, optical fibers, and other devices, would be a significant source of costs. Perhaps the greatest costs, however, would come from the actual construction of the WON. An operating enterprise would have to acquire or lease space for the great length of optical-fiber cables that make up the WON. Besides the capital cost of the optical-fiber cables, associated devices, and attached stations, there is a substantial cost to install this equipment. Given this great length of optical fiber in the WON, it is most natural to view its physical plant as essentially runs of cable. The overall cost of the WON could then be modeled as a function of the length of optical-fiber cable in the WON. We shall therefore assume that the cost of the WON is a linear function of the total length of optical fiber in the network.

Perhaps more restrictive than the linear cost model is the assumption that the length of optical fiber from one point to another is modeled by the Euclidean distance metric. Since we are not designing MANs in a desert, the assumption that optical fiber may be installed in straight lines is often invalidated by the fact that the existing infrastructure can cause optical fiber to be laid in zig-zags. One alternative, should this assumption prove untenable, would be to use other distance metrics, such as Manhattan or rectilinear distance, which might prove to be a more realistic model of how optical fiber is installed in an urban environment; the adaptation of our techniques to other well-defined distance metrics is straightforward.

It is worthwhile noting that the fidelity of the cost model is not necessarily the most important issue here. When we are comparing costs of extremely different topological classes, there is the danger that the model will yield costs that are not really comparable. However, the most promising topological classes for the WON, viz., the star and tree topologies, are very similar in their physical characteristics, so the use of the linear cost model will accurately *rank* these different physical topologies, assuming that they have been similarly implemented. We also note that *within* a given topological class, the specific costs of two different configurations of the same network will be used primarily for the purposes of comparing which configuration is more cost effective.

3.4 A Framework for WON Design

Our main objective in this section is to describe the WON design process as a decomposition of key problems in the design of the WON, with emphasis on how these problems interrelate.

In designing the WON to serve a specific metropolitan area, the network architect is immediately faced with a number of important design decisions to be made. First, the architect must choose a cable-plant design that permits reliable signaling between every transmitter–receiver pair in the WON. Engineering of the cable plant requires the selection of optical components to be used, e.g., optical fiber, couplers, connectors, transmitters, receivers, optical switches, etc. These components all possess characteristic losses that attenuate the lightwave signal, and therefore such losses must be kept at a minimum if adequate signal quality is to be maintained. Second, given the stations' locations and the general cable-plant design, where shall elements of the WON be physically located? Given the expense of installing and maintaining cable in an urban environment, it is reasonable to assume that the physical placement of the optical fiber and components of the WON must be chosen in order to achieve maximum cost savings. The third major design decision facing the network architect is how to assign stations to channels, given a specific physical topology and traffic matrix. A fourth decision requires the determination of a routing scheme that promotes a high level of performance and a low probability of congestion.

Ideally, it should be possible to solve the entire design problem in an integrated fashion, perhaps by jointly choosing the cable plant, physical topology, and virtual topology to minimize the cost of the WON subject to the constraint that packet delay shall not exceed a maximum acceptable threshold.[1] Since a combinatorial optimization of this scale appears impractical, we choose instead to decompose the problem into the three following subproblems, which are naturally derived from the overall design problem:

- *The Cable-Plant Design Problem (CPDP).* Given the station locations and available components, select the cable plant that provides an acceptable power margin for all transmitter–receiver pairs.

- *The Physical-Topology Design Problem (PTDP).* Given the constraints imposed by the cable-plant design, minimize the cost of the WON (as reflected

[1] This delay constraint encompasses the requirement that adequate signal quality be maintained, since lost or erroneous packets must be retransmitted, thereby driving up packet delay.

by the total length of optical fiber) by selecting the optimal physical deployment of optical fiber and components.

- *The Virtual-Topology Design Problem (VTDP).* Given the traffic matrix and the interstation separation distances dictated by the physical topology, minimize the mean packet delay by selecting the optimal virtual topology.
- *The Routing and Congestion-Control Problem (RCCP).* Determine a routing procedure that reduces the likelihood of congestion for specific physical and virtual topologies.

This solution approach, although it results in suboptimal designs, is quite reasonable, since the network architect's primary concerns are to make possible reliable signaling, to produce a cost-effective design, to construct a network that meets a given performance goal, and to develop a routing protocol that best adapts to dynamic fluctuations in the network, roughly in that order of precedence.

3.4.1 Cable-Plant Design

Since many aspects of the CPDP have been treated in detail by other authors [Bak86, Pal88], we will not say very much about it here. We do mention that cable-plant engineering is typically performed in a "cookbook" fashion in which a design is proposed, and the intrinsic losses of all components are tallied and checked against the transmitter-to-receiver power budget. There are numerous possibilities for the design of a WON's cable plant, and the design alternatives are limited solely by the inventiveness and experience of the cable-plant designer. Our approach to cable-plant design is somewhat less sophisticated—instead we suppose that the cable plant could be chosen cookbook-style from one of the common topological classes shown in Figure 4.2. These topological classes have been extensively discussed in the literature as to their suitability for use in optical fiber MANs [NTM85, GF88, BCF89]. That the cable plant will in fact accommodate the given transmitter-to-receiver power budget can be checked by totaling the worst-case power margin that results from attenuation in the optical fiber, coupling loss, connector loss, splice loss, transmitter power, and then by accounting for receiver sensitivity and dynamic range. If the power margin is inadequate, another cable-plant design must be considered.

3.4.2 Physical-Topology Design

The physical topology of the WON includes stationary equipment that can not be moved and nonstationary equipment that could be located in a number of different locations. Seeing that there are many different potential locations for a piece of equipment, the network designer would prefer to place the equipment in that spot that yields the highest benefit, if such a spot exists. The nonstationary category includes devices such as couplers and the headend, which can in principle be placed anywhere, while the stationary category is comprised principally of stations, which must be collocated near their attached users. We generally assume that once a location has been chosen for a piece of nonstationary equipment, that piece will remain there for the duration of its lifetime.

The PTDP involves determining the optimum placement of the WON's nonstationary equipment, given the locations of its stationary equipment. The criterion of optimality is cost, which we consider to be a function of the length of optical fiber in the WON. The statement of the PTDP presupposes the selection of the WON's basic cable plant, i.e., its basic topology. This basic topology is usually a star or tree, but it can be any topology that meets the WON's requirements, such as optical power or reliability. Once the network designer has committed to the use of a specific topology (or a small set of candidate topologies), the next logical step is the determination of the optimum geographical placement of devices.

Finding the optimum geographical placement of the WON's nonstationary equipment is an optimization problem related to other well-known mathematical programming problems. The PDTP corresponding to the star topology demands the placement of the central coupler in such a way that the radial arms end up as short as possible. Applied to the tree topology with $N = 2^n$ stations at its leaves, the PTDP consists of jointly locating the WON's $2^n - 1$ couplers so that the resulting tree is of minimum length. The star variant of the PTDP has a long tradition, having been previously studied by the mathematicians Fermat, Torricelli, and Steiner in earlier centuries. In this century the tree variant, which is a generalization of the original problem, has been studied in the context of locating economic facilities among centers of demand. We can informally state the PTDP in a quasimathematical format. We assume that the network designer has identified a promising topological class, so that the connectivity between couplers and stations is defined (though not the location of the couplers). The problem then becomes one of finding locations for the couplers that minimize the WON's cost:

Given:	Station–Coupler Connectivity
	Station Locations
Objective:	Minimize Total Length of Optical Fiber
Design Variables:	Coupler Locations
Constraints:	None

This problem is a continuous optimization problem that is amenable to solution by appropriate optimization techniques. Our goal is to develop and demonstrate implementable algorithms to solve the above problem in an acceptable amount of time.

In Chapter 4 we shall study the PTDP in greater detail.

3.4.3 Virtual-Topology Design

Given the network designer's powerful control over the definition of the WON's virtual topology, it is natural to attempt to exercise that control to full advantage. We have already noted the effect that the choice of virtual topology has on the performance of the WON. Certainly, it is obvious that the choice of an ill-conceived virtual topology could lead to disaster, e.g., were the virtual topology not to provide adequate connectivity by partitioning stations into disjoint groups. The structured virtual topologies, such as ShuffleNet or the de Bruijn digraph, appear to facilitate good performance in certain circumstances (e.g., uniform traffic), but it may be possible to improve upon these virtual topologies, especially when the network environment is modified (e.g., nonuniform traffic matrix).

The VTDP is the problem of identifying a virtual topology that optimizes a specific performance measure. Obviously, the VTDP depends on input parameters such as the traffic matrix and the distance matrix. The problem of finding small-diameter graphs, which is related to the VTDP, has been pursued by graph theorists and mathematicians, but in a much more abstract sense than we are advocating—the search for such graphs is one of analysis and elegant proofs, whereas ours relies much more on brute-force optimization. After all, the goal is to design high-performance networks.

Integrating what has been said, we can state the VTDP in a quasimathematical format. Basically, we seek to minimize the packet delay by choosing the right virtual topology for the given network conditions:

Given:	Distance Matrix
	Traffic Matrix
Objective:	Minimize Mean Packet Delay
Design Variables:	Virtual Topology
	Traffic Flows
Constraints:	Flow Conservation
	Station Connectivity

The VTDP, as we shall see in Chapter 5, is a challenging combinatorial optimization problem which is unlikely to have an efficient solution procedure. We therefore focus our efforts on finding algorithms that yield reasonably good solutions in an acceptable amount of time.

3.4.4 Routing and Congestion Control

To control the cost of the WON, we might wish to reduce the number of packet buffers in the station. It is thus possible that arriving packets might overflow the station's buffers during peak traffic intervals. Furthermore, the routes of packets can—if not intelligently chosen—cause congestion at various points in the WON. The effects of congestion, whether manifested as packet loss or excessive delay, are very deleterious to the performance of the WON. The throughput of the WON in periods of congestion could drop to levels well below the maximum achievable throughput, and delay could exceed the acceptable limit.

Although congestion usually results from excessive demands on a system resource, e.g., a high arrival rate, the failure of a resource can also cause congestion. Link or station failure in the WON can either cause a high probability of packet loss or an increase in delay. Because the wavelength-agile WON has the capability of altering its virtual topology, the loss of a resource will not be such a critical concern. We therefore concentrate on congestion caused by unexpected fluctuations in the workload.

Since there are several ways to route packets in the WON, the choice of a specific routing protocol should be aimed at achieving well-defined goals. To be sure, one of these goals is realizability in the WON, which is a nontrivial challenge when one considers the need to switch packets at hypothetical link speeds of 1 Gbps—this immediately rules out many adaptive routing techniques based on extensive routing tables, since updating and lookup are typically performed in software. To guarantee fast packet switching, one would most likely use a routing algorithm that can be implemented in hardware. Another goal is performance, i.e., to achieve low delay and high maximum throughput with the chosen routing algorithm. Such routing algorithms often rely on the avoidance of congestion by forcing a packet to take an alternate path when the original path is congested. Deflection and detour routing, which were mentioned in Chapter 2, are simple multipath-routing algorithms that implement congestion avoidance. These will be analyzed in Chapter 6, where we shall also study the problem of finding the optimal virtual topology to use with such a routing procedure.

3.5 Interplay Among the Design Problems

As we remarked earlier, the cable-plant design, physical-topology design, virtual-topology design, and routing problems could in principle all be solved as one huge monolithic problem were we not constrained by limits on our computational resources. Therefore, we are forced to decompose the monolithic problem into individual subproblems, which results in suboptimal designs. This is because decisions made at one point can affect those made at a later point, e.g., the choice a particular physical topology has an impact on the virtual topology, because its design uses the distance matrix as an input.

Also, the criteria for optimization differ from problem to problem. The PTDP's objective of minimizing cost can clash with the VTDP's objective of minimizing delay. For example, the star topology, with its direct cabling from station to headend, will produce smaller propagation delays than the tree topology, with its oblique cabling; but the PTDP will—if based completely on the cost criterion—favor the selection of the tree, much to the detriment of the WON's performance.

Another area in which a previous design decision could have a profound impact on subsequent decisions is in routing and congestion control. When using routing techniques such as detour routing, which diverts packets away from a congested port via a fixed alternate path, the choice of virtual topology determines whether or not alternate paths will be good, i.e., short. Therefore, a virtual topology that was chosen to minimize mean packet delay might not facilitate good flow control, because it was not optimized to offer good alternate paths.

An underlying difficulty is the problem of optimizing multiple objective functions. Cost and performance, for instance, are not orthogonal, and we can not necessarily expect to optimize one parameter without affecting the other. It is also difficult to perform mixed or multiple-objective optimization, because there is little basis for comparing different metrics. For example, it is not obvious how much performance one is willing to trade off against a unit of cost.

For these reasons it is necessary to be aware of how the solutions of the individual problems affect each other. We will attempt to gauge the effects of problem interplay at various points in the sequel. Specifically, in Chapter 5 we will report the results of a case study of physical- and virtual-topology design, with the intent of characterizing the degree of coupling between these two forms of topological design.

3.6 Summary

This chapter presents a queueing-network–based model of the WON that can be used to predict the mean packet delay in the WON. It also presents a simple cost-model that estimates the costs of installing and operating a WON, based on the total length of fiber in the WON. We apply these models to formulate specific problems in the design of the WON, including the cable-plant design problem, the physical-topology design problem, the virtual-topology design problem, and the routing and congestion-control problem.

Chapter 4

The Physical-Topology Design Problem

In this chapter we address the problem of selecting a physical topology for the WON. Since the cost of constructing and maintaining the physical topology is probably the single item of highest cost in the WON, it is extremely important that its design be executed with care and with the aim to keep its cost as low as possible.

We review the classes of physical topologies to be considered, mathematically formulate the PTDP, propose a solution procedure for the PTDP, and apply this procedure to solve specific instances of the PTDP.

4.1 Physical Topologies for the WON

As we have mentioned in Chapter 3, there are several different physical topologies that could serve as the infrastructure for signal distribution in the WON. Besides the star, tree, and bus topologies originally advocated by Acampora and his colleagues [Aca87, AKH87, AK89] for use in the WON, a novel type of ring topology has also been studied for the WON [Kar88]. Furthermore, we are interested in employing Tree-Net [GF88], which differs in key areas from the tree topology proposed by Acampora, as a physical topology for the WON.

Some physical topologies are better suited than others for use in the WON. The bus, in particular, can be excluded from serious consideration for all but the smallest of WONs. This is because the worst-case power loss in a bus additively includes the insertion losses of *all* bus taps, and thus the power budget must increase linearly with the number of taps, which in practice limits the maximum number of attached stations to less than 100. The star topology based upon the reflective or transmissive passive star coupler has the capacity for a much larger station population. The reflective or transmissive star coupler, upon which the star is based, is a passive optical device with N inputs and N outputs. The loss in the star coupler ascribable to power splitting is $-10\log(1/N)$, which implies that N can attain reasonably high values without violating the power budget. Tree-Net, too, can accommodate a reasonably large station population. We have seen in Chapter 2 that several configurations of Tree-Net (with varying degrees of clustering at the leaf nodes) maximize the station population relative to a given power budget. The ring topology [Kar88], which addressed the special concern of using multiple optical fibers when only a few WDM channels can be multiplexed onto a single optical fiber, will not be considered here. In summary, the physical topology can radically limit the station population of the WON, and the classes of topologies to be considered as design candidates should be selected with the ultimate station population and transceiver capabilities in mind.

The physical topology chosen for the WON impinges upon other design issues as well. Besides cost, which we soon discuss in detail, issues such as reliability and security surface in the choice of a physical topology. If tolerance of the failure of an optical fiber is a primary objective, then the star permits continuous operation with a minimal disruption of service, with the tree and bus faring somewhat worse. If other failure modes, such as coupler failure or headend failure, are anticipated, then all physical topologies are susceptible to degraded or interrupted service, and special measures should be taken to enable these devices to recover from such failures. One facet of security is the capability for groups of users to operate their own private subnetworks. A properly configured virtual topology in essence allows the establishment of such private subnetworks, but true physical separation of distinct subnetworks is more difficult to achieve. A desirable goal would be to partition the WON into several private subnetworks so that physical (not just virtual) access can be denied to members outside the prescribed group. Such physical separation can be achieved with all the topologies discussed—e.g., by providing separate cable plants for each subnetwork, perhaps with secure gateways between them—but the costs of this solution differ depending on the basic topology. Figure 4.1 gives a rough idea of how a two-group WON would be constructed using the three basic physical topologies. It is the star that appears to offer the most flexibility, as subnetworks can be redefined by merely altering connections at the headend, whereas a redefinition in the bus or the tree might require installing substantial runs of new optical fiber.

A reasonable approach to topological design in the WON would be to first identify promising topological classes, e.g., the tree or star, given the overall requirements for the cable plant, e.g., power budget, station population, and reliability, and then to optimize the candidate physical topologies from the viewpoint of cost.

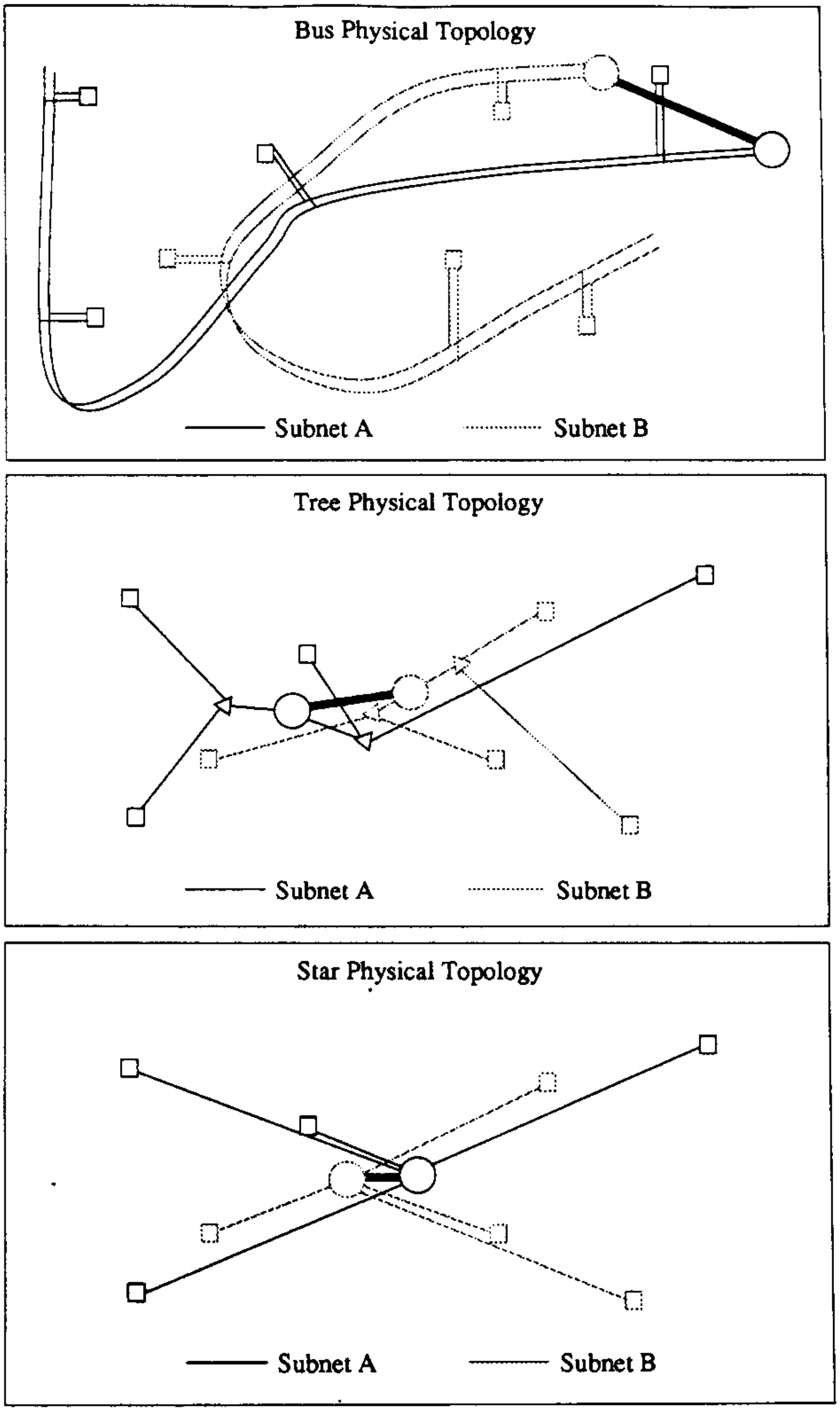


Figure 4.1: Providing Security in WONs with Different Physical Topologies.

In the sequel we restrict our consideration to star and tree topologies. The four topological classes that we examine are shown in Figure 4.2 and include the basic star and variants of Tree-Net with three degrees of clustering.

4.2 Problem Description and Formulation

Strictly speaking, the design of the WON's physical topology should embrace every aspect of its cable plant. This would include the exact specification of optical fibers to be used and their precise geographical layout; the determination of all coupler, tap, and connecter characteristics; and the type of regenerating headend to be used. We could go one step further, expecting to tabulate completely all component and installation costs, even to the extent of identifying the manufacturer of each piece of equipment. By necessity, we must go to a level of abstraction that simplifies the problem by eliminating nonessential details. Our version of the PTDP seeks to minimize the *relative* (rather than the absolute) cost of the WON by identifying the lowest-cost locations for major pieces of equipment used to implement a specific topological class.

To illustrate, given that the chosen "shape" of the WON's physical topology matches the tree shown in Figure 2.8 of Chapter 2, what positioning of its couplers would result in the shortest length of optical fiber? The answer is not obvious and could require a considerable amount of computation to obtain.

In formulating the PTDP we restrict ourselves to WONs that employ one of the topologies shown in Figure 4.2. This assumption is justified primarily on the basis of experience with optical topologies: of the bus, star, and variants of the tree topologies, only the star and trees have proven feasible from the point of view

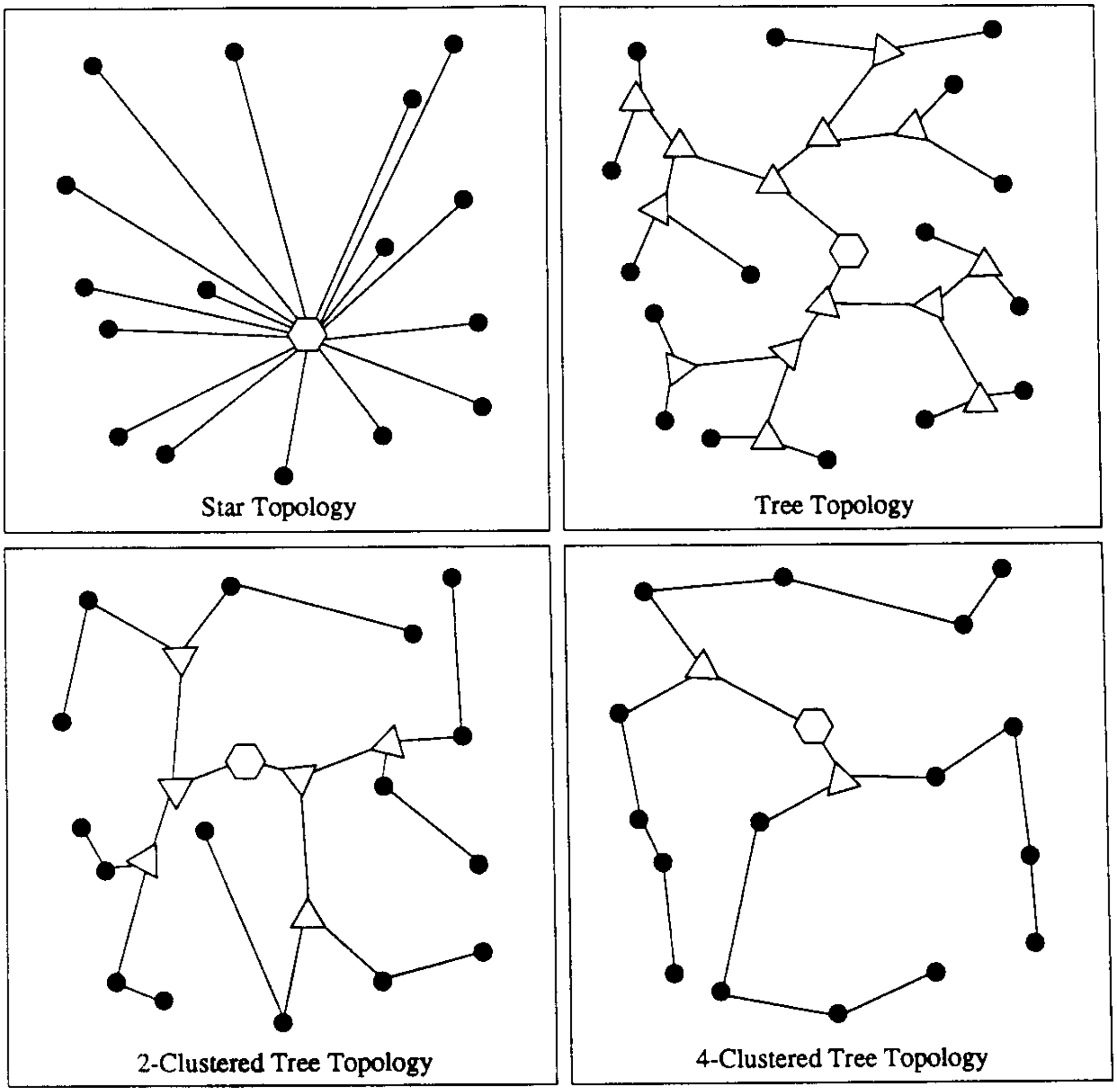


Figure 4.2: Four Topological Classes for the WON.

of satisfactory signal quality in networks of medium-to-large size [GF88]. The problem of which of the four topological classes of Figure 4.2 to choose as the basis for solving the PTDP is more problematic. Such a choice may be based upon additional criteria such as reliability requirements (e.g., the failure of a cable segment in the tree can isolate a large collection of stations from the remainder of the network). For the purposes of comparison we might evaluate *all* of the four topological classes. Thus the PTDP becomes a problem of finding locations for the couplers of the tree (the triangles of Figures 2.8 and 4.2), with the objective of minimizing the length of optical fiber (the solid lines of Figure 2.8 and 4.2) used in the WON.

The difficulty of finding a faithful cost metric is common to almost all cost-minimization problems, and the PTDP is no exception. In the formulation of the PTDP, we are assuming that the cost of the WON equates to the total length of optical fiber in the network, which is an acceptable assumption, but can be misleading in some cases, since it ignores other factors contributing to cost, such as component costs. Perhaps a more restrictive assumption is that the length of optical fiber from one point to another is modeled by the Euclidean distance metric, as the assumption that optical fiber may be installed in straight lines is often invalidated by the fact that optical fiber must be laid to cope with the existing infrastructure, which can result in nonlinear runs of optical-fiber cable. One alternative, should this assumption prove untenable, would be to use other distance metrics, such as Manhattan (rectilinear) distance, which might prove to be a more realistic model of how optical fiber is installed in an urban environment; the adaptation of our techniques to other well-defined distance metrics is straightforward. Despite these concerns, we will nonetheless retain the Euclidean distance metric to model the length of optical fiber between two points.

In its purest mathematical form the PTDP is an unconstrained optimization problem in which, given the locations of a set of points (stations) in the plane, the objective is to choose another set of points (couplers) so that the (not necessarily binary) spanning tree for these sets of points has minimal length. This problem, which is known as the geometric Steiner tree problem [Win87], is NP-hard in the strong sense [GJ79, page 209] and therefore essentially intractable. The geometric Steiner tree problem makes no restrictions on the degree of the tree's nodes, whereas couplers always correspond to nodes of degree three (input, favored, and tap ports). Furthermore, in the design of an actual WON, we must consider the (power, reliability, and security) constraints on the cable plant, as previously discussed, and we thus restrict the statement of the PTDP so that the objective is to find the optimal positioning of couplers, given one of the topological classes shown in Figure 4.2.

Considering the difficultly of solving the geometric Steiner tree problem, we find it necessary to first establish the basic physical topology and then to optimize the locations of its nonstationary devices. In the star topology this is trivial, as the topology is established simply by running optical fibers from all stations to the star coupler, which is then positioned to minimize the total length of optical fiber according to a procedure that we soon discuss. In the tree topologies, on the other hand, to establish a topology requires that we group stations together in order to run their optical fibers to a common coupler. This grouping should be done with care, because the final cost of the network strongly depends on the specific grouping, even after optimization. For instance, if we were to group stations at opposite ends of the plane, then the topology—although technically a tree—would appear more like a star. Thus we should attempt to group stations in close proximity to each other so that we can then run a small length of separate optical fibers to their common coupler; from the coupler to the headend they would then share a common optical fiber. This aspect of the problem is called the *station-coupler clustering problem* and relies on heuristic clustering algorithms. Obviously, if we are to use a tree with higher degrees of clustering at the leaves, then the clustering algorithm must group together stations that will reside at the leaves. The clustering proceeds hierarchically, since couplers also must be grouped together at shallower levels of the tree.

After we choose the desired topological class and heuristically group stations into clusters, we attempt to position the couplers so that the total length of optical fiber is minimized. This aspect of the problem, which we refer to as the *coupler location problem*, can be seen to be a special case of the multifacility location problem [Mie58, Rad88]. As in Figure 2.8, we can view the physical topology of the WON as a tree whose M nodes are partitioned into coupler and station nodes. If we let the vectors $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_M$ be the planar coordinates of the tree's nonstationary nodes (viz., couplers), and $\boldsymbol{V}_{M+1}, \boldsymbol{V}_{M+2}, \ldots, \boldsymbol{V}_{M+N}$ be the planar coordinates of the tree's stationary nodes (viz., stations), the PTDP then becomes a problem of selecting the locations of the coupler nodes so that the total length of the tree is minimized, i.e.,

$$\min f(\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_M) = \sum_{i=1}^{M} \sum_{j \in C(i)} \|\boldsymbol{V}_i - \boldsymbol{V}_j\| \qquad (4.1)$$

where $\|(x, y)\| \triangleq \sqrt{x^2 + y^2}$ is the Euclidean norm, and $C(i)$ is the set of nodes that are children of node i.

Clearly, the only inputs to the PTDP are the locations of the stations to be connected and the indication of what class of topology is being considered (i.e., the connectivity between stations and couplers). The principal outputs of the PTDP are the locations of the couplers, from which can be calculated the WON's distance matrix, whose entries are defined to be the "glass" distances that the lightwave signal must travel to get from one station to another. The distance matrix is then used as an input to the VTDP, which we will discuss in Chapter 5.

4.3 Solution Approach

When we start with a star or tree topology, there are two tasks involved with completely specifying its physical topology: connecting the stations and couplers,[1] and locating the couplers in the plane. As mentioned earlier, the first task is called the station–coupler location problem, and the second task is called the coupler location problem. The first task is performed by means of a clustering algorithm and the second by means of a location algorithm. Also, as in most optimization algorithms, the initialization of the algorithm is an important factor in the algorithm's subsequent performance. We next discuss the clustering and location algorithms, and the initialization of the location algorithm.

4.3.1 The Clustering Algorithm

The clustering algorithm is responsible for creating the tree that makes up the WON's physical topology. If we define an abstract distance metric on a test space, then cluster **analysis** attempts to aggregate data points of the test space into a collection of **clusters** so that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point in another cluster [JD88]. In the station–coupler clustering problem, the test space is the set of all station locations, and the distance metric corresponds to the Euclidean

[1] This procedure is trivial in the star since all its stations are connected to a single coupler.

distance between pairs of stations. The clustering proceeds hierarchically, reflecting the tree-like structure of the WON's physical topology.

Suppose that we wish to construct a k-clustered tree physical topology for N stations. We start off with N disjoint sets (or clusters) $\mathcal{K}_1$, $\mathcal{K}_2$, $\ldots$, $\mathcal{K}_N$, each consisting of a single station location. At the next level of the hierarchy, the clustering algorithm iteratively refines the initial clustering into sets $\mathcal{K}_{11}$, $\mathcal{K}_{12}$, $\ldots$, $\mathcal{K}_{1N_1}$, each containing no more than k station locations; this clustering corresponds to the clustering of stations at the leaf nodes of the physical topology. At the next level the refinement results in the partition $\mathcal{K}_{21}$, $\mathcal{K}_{22}$, $\ldots$, $\mathcal{K}_{2N_2}$, where each $\mathcal{K}_{2i}$ is the disjoint union of no more than two of the $\mathcal{K}_{1j}$. The refinement continues until a single cluster remains: $\mathcal{K}_{L1}$, where L is the number of levels in the tree. The end product of this process—often termed a *dendrogram* in the literature—is depicted in Figure 4.3. Notice that after the first two passes of clustering, each cluster has a unique coupler associated with it, e.g., the top-level coupler with $\mathcal{K}_{31}$ and the next-level coupler with $\mathcal{K}_{21}$.

A simple algorithm that comes immediately to mind is a greedy algorithm that scans the proximity matrix (i.e., the matrix of straight-line distances between all pairs of stations) and successively chooses the minimum entry, clusters the points corresponding to the entry, and deletes the row and column of the entry. Although this algorithm, which we show in Figure 4.4, effectively clusters the closest stations, it has a tendency to leave isolated "islands" that must be clustered at the end of the algorithm's execution. Unfortunately, these "islands" can be separated by substantial distances; therefore, the algorithm will often have a few clusters that consist of widely separated stations. The existence of widely separated station clusters is highly undesirable from the point of view of designing the physical

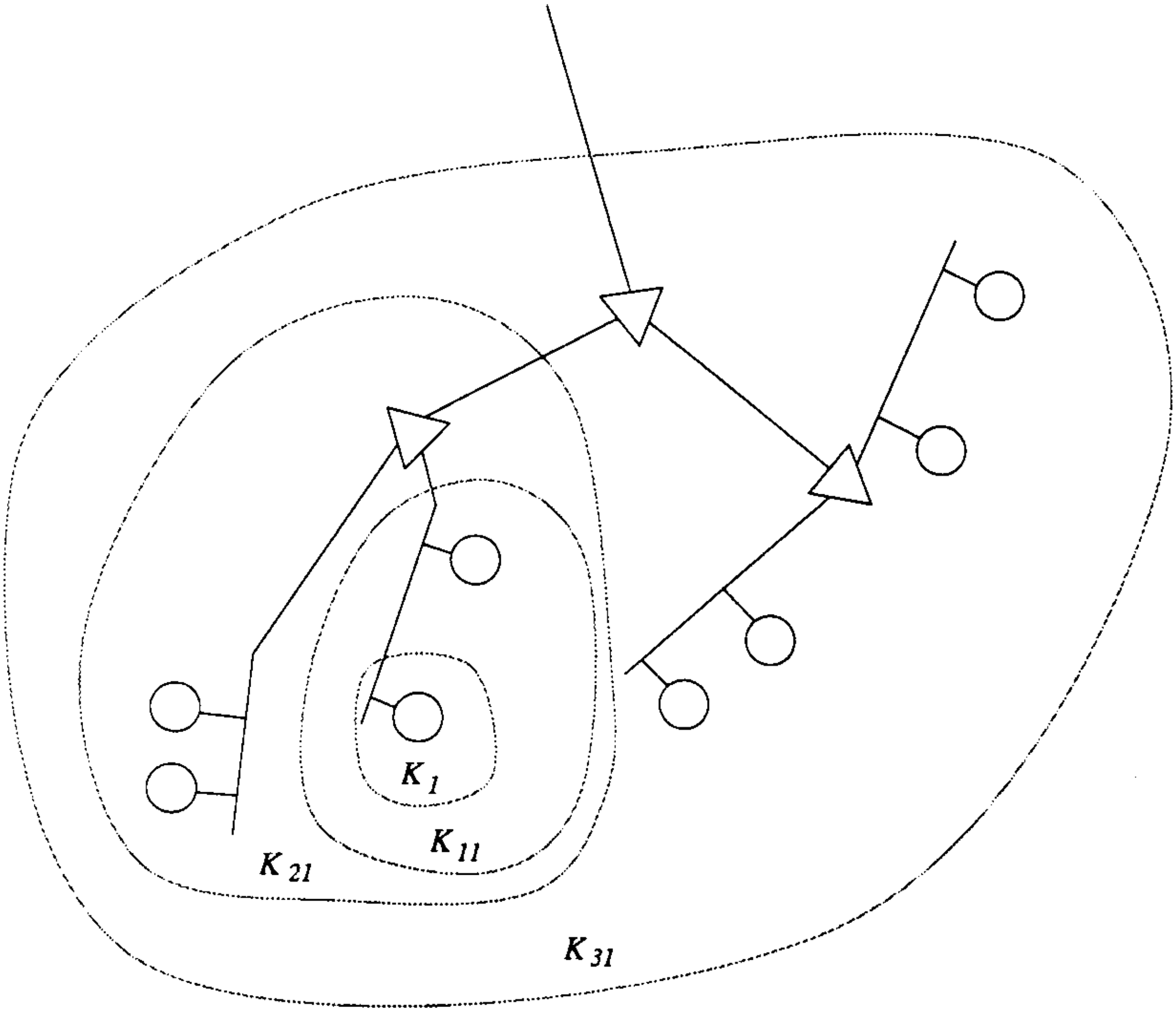


Figure 4.3: The Hierarchical Clustering of Stations and Couplers.

1. *Initialization.* Assign $S = \{1,\ 2,\ \ldots,\ N\}$.
2. *Minimum Distance Search.* Assign $T = \{i,\ j \mid \forall k{\in}S\ \forall l{\in}S$ distance$(i,\ j) \leq$ distance$(k,\ l)\}$.
3. *Cluster Formation.* Assign $\mathcal{K}(i) = T$.
4. *Termination Test.* Assign $S = S - T$. If set S is empty then halt, else go to 2.

Figure 4.4: The Greedy Clustering Algorithm.

topology. We therefore choose another approach.

The improved algorithm, which is shown in Figure 4.5, begins at the origin and scans the plane in a north-easterly direction, which is the basis for using the lexicographic ordering defined on $\mathbb{R}^2$ by $(x,\ y) \prec (w,\ z)$ if and only if either $x < w$ or $x = w$ and $y < z$. When it encounters a station, the algorithm searches for its nearest neighbor. The station and its neighbor are then clustered together, and the algorithm continues with the scan until no stations are left. Since the algorithm only clusters pairs of objects, it is necessary to repeat the algorithm if we wish to create clusters of three or more stations; in this case the algorithm should be applied to existing clusters as well as stations.

We emphasize the importance of the clustering algorithm, noting that the design of cost-effective physical topologies depends on our ability to cluster neighboring stations together so that they can share common runs of optical fiber. Since stations are not clustered in the star topology, the quality of the clustering algorithm's solutions is mainly an issue in the tree topologies. A poor clustering used

1. *Initialization.* Order all points lexicographically and place them in set S.
2. *Planar Scan.* Find point $(x,\ y)$, the lexicographic minimum of set S, and place it in cluster $\mathcal{K}(x,\ y)$.
3. *Cluster Formation.* Find the nearest neighbor of point $(x,\ y)$ and place it in cluster $\mathcal{K}(x,\ y)$.
4. *Termination Test.* Assign $S = S - \mathcal{K}(x,\ y)$. If set S is empty then halt, else go to 2.

Figure 4.5: The Scan-Driven Clustering Algorithm.

with a tree topology can produce a physical topology whose cost approaches that of the star. For instance, consider a set of stations located on the circumference of a circle, and suppose that each station is clustered with the station diametrically opposite it. In the tree topology every pair of stations would be joined by a coupler at the center of the circle, which results in a layout of optical fiber that strongly resembles the star. Obviously, then, a better clustering would be to couple pairs of adjacent stations.

4.3.2 The Location Algorithm

We first note that in Appendix C we prove that the function $f(\cdot)$ of Equation (4.1) is convex.[2] Therefore, any local minimum of $f(\cdot)$ must also be its global minimum

[2] A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex if it satisfies Jensen's inequality, i.e., if $\boldsymbol{W}_i \in \mathbb{R}^n$ for $i = 1,\ 2,\ \ldots,\ m$, then

$$f\left(\sum_{i=1}^{m} \alpha_i \boldsymbol{W}_i\right) \leq \sum_{i=1}^{m} \alpha_i f(\boldsymbol{W}_i)$$

[RV73, page 123]. So, once we identify a local minimum of $f(\cdot)$, we will have solved the problem of finding the minimum-cost solution to the PTDP. To find a local minimum we attempt to solve the system of nonlinear equations that results from setting the gradient of the objective function equal to the zero vector.

For the function $f(\cdot)$ in Equation (4.1), define

$$\nabla_i f(\boldsymbol{V}_1,\ \boldsymbol{V}_2,\ \ldots,\ \boldsymbol{V}_M) \triangleq \begin{bmatrix} \frac{\partial f}{\partial v_{ix}}(\boldsymbol{V}_1,\ \boldsymbol{V}_2,\ \ldots,\ \boldsymbol{V}_M) \\ \frac{\partial f}{\partial v_{iy}}(\boldsymbol{V}_1,\ \boldsymbol{V}_2,\ \ldots,\ \boldsymbol{V}_M) \end{bmatrix}^T$$

where $\boldsymbol{V}_i \triangleq (v_{ix},\ v_{iy})$. Computing the gradient of $f(\cdot)$ with respect to the coordinates of each coupler node i, we get

$$\nabla_i f(\boldsymbol{V}_1,\ \boldsymbol{V}_2,\ \ldots,\ \boldsymbol{V}_M) = \frac{\boldsymbol{V}_{p(i)} - \boldsymbol{V}_i}{\|\boldsymbol{V}_{p(i)} - \boldsymbol{V}_i\|} + \sum_{j \in C(i)} \frac{\boldsymbol{V}_i - \boldsymbol{V}_j}{\|\boldsymbol{V}_i - \boldsymbol{V}_j\|} \tag{4.2}$$

where node $p(i)$ is the parent of node i. Setting Equation (4.2) equal to 0 to minimize the function $f(\cdot)$, we obtain

$$\boldsymbol{V}_i = \frac{\frac{\boldsymbol{V}_{p(i)}}{\|\boldsymbol{V}_{p(i)} - \boldsymbol{V}_i\|} - \sum_{j \in C(i)} \frac{\boldsymbol{V}_j}{\|\boldsymbol{V}_i - \boldsymbol{V}_j\|}}{\frac{1}{\|\boldsymbol{V}_{p(i)} - \boldsymbol{V}_i\|} - \sum_{j \in C(i)} \frac{1}{\|\boldsymbol{V}_i - \boldsymbol{V}_j\|}} \tag{4.3}$$

If we define the vector function $F : I\!R^{2M} \rightarrow I\!R^{2M}$ by $F(\boldsymbol{V}_0,\ \boldsymbol{V}_1,\ \ldots,\ \boldsymbol{V}_M) = (\boldsymbol{W}_0,\ \boldsymbol{W}_1,\ \ldots,\ \boldsymbol{W}_M)$ where

$$\boldsymbol{W}_i = \frac{\frac{\boldsymbol{V}_{p(i)}}{\|\boldsymbol{V}_{p(i)} - \boldsymbol{V}_i\|} - \sum_{j \in C(i)} \frac{\boldsymbol{V}_j}{\|\boldsymbol{V}_i - \boldsymbol{V}_j\|}}{\frac{1}{\|\boldsymbol{V}_{p(i)} - \boldsymbol{V}_i\|} - \sum_{j \in C(i)} \frac{1}{\|\boldsymbol{V}_i - \boldsymbol{V}_j\|}}$$

then the system of $2M$ nonlinear equations in Equation (4.3) is equivalent to the fixed-point equation

$$F(\boldsymbol{V}_0,\ \boldsymbol{V}_1,\ \ldots,\ \boldsymbol{V}_M) = (\boldsymbol{V}_0,\ \boldsymbol{V}_1,\ \ldots,\ \boldsymbol{V}_M) \tag{4.4}$$

where $0 \leq \alpha_i \leq 1$ for $i = 1,\ 2,\ \ldots,\ m$, and $\sum_{i=1}^{m} \alpha_i = 1$.

1. *Initialization.* For $i = 1$ to M choose an initial value for the vector $\boldsymbol{V}_i$.
2. *Iteration.* For $i = 1$ to M assign
$$\boldsymbol{V}'_i = \frac{\frac{\boldsymbol{V}_{p(i)}}{\|\boldsymbol{V}_{p(i)}-\boldsymbol{V}_i\|} - \sum_{j\in C(i)} \frac{\boldsymbol{V}_j}{\|\boldsymbol{V}_i-\boldsymbol{V}_j\|}}{\frac{1}{\|\boldsymbol{V}_{p(i)}-\boldsymbol{V}_i\|} - \sum_{j\in C(i)} \frac{1}{\|\boldsymbol{V}_i-\boldsymbol{V}_j\|}}$$
3. *Convergence Test.* If $\sum_{i=1}^{M} \sum_{j\in C(i)} \|\boldsymbol{V}_i - \boldsymbol{V}_j\| - \|\boldsymbol{V}'_i - \boldsymbol{V}'_j\|$ meets the stopping criterion then stop, else for $i = 1$ to M assign $\boldsymbol{V}_i = \boldsymbol{V}'_i$ and go to 2.

Figure 4.6: The Location Algorithm.

To solve Equation (4.4), we can use the iterative form $\boldsymbol{X}^{(k+1)} = F(\boldsymbol{X}^{(k)})$ [Ger78], which yields the following formulas:

$$\boldsymbol{V}_i^{(k+1)} = \frac{\frac{\boldsymbol{V}_{p(i)}^{(k)}}{\left\|\boldsymbol{V}_{p(i)}^{(k)}-\boldsymbol{V}_i^{(k)}\right\|} - \sum_{j\in C(i)} \frac{\boldsymbol{V}_j^{(k)}}{\left\|\boldsymbol{V}_i^{(k)}-\boldsymbol{V}_j^{(k)}\right\|}}{\frac{1}{\left\|\boldsymbol{V}_{p(i)}^{(k)}-\boldsymbol{V}_i^{(k)}\right\|} - \sum_{j\in C(i)} \frac{1}{\left\|\boldsymbol{V}_i^{(k)}-\boldsymbol{V}_j^{(k)}\right\|}} \tag{4.5}$$

These iterated equations determine a sequence of locations $(\boldsymbol{V}_1^{(0)}, \boldsymbol{V}_2^{(0)}, \ldots, \boldsymbol{V}_M^{(0)})$, $(\boldsymbol{V}_1^{(1)}, \boldsymbol{V}_2^{(1)}, \ldots, \boldsymbol{V}_M^{(1)})$, ..., with the descent property [Ost77]:

$$f(\boldsymbol{V}_0^{(k+1)}, \boldsymbol{V}_1^{(k+1)}, \ldots, \boldsymbol{V}_M^{(k+1)}) \leq f(\boldsymbol{V}_0^{(k)}, \boldsymbol{V}_1^{(k)}, \ldots, \boldsymbol{V}_M^{(k)})$$

The limiting **point** $\boldsymbol{V}_i^*$ specifies the optimal location for the coupler node i.

This iterative procedure, which we depict in Figure 4.6, is essentially the one originally proposed by Weiszfeld [Wei36] and modified by Miehle [Mie58], who implemented an analog technique to solve the location problem using devices consisting of weights, strings, and pulleys arranged on a flat surface; and soap films between glass plates. It is not known whether this particular descent method converges to $(\boldsymbol{V}_1^*, \boldsymbol{V}_2^*, \ldots, \boldsymbol{V}_M^*)$, the minimizer of Equation (4.1); however, it appears to perform very well in practice.

Recently, Radó [Rad88] has provided a proof of convergence for a descent method that replaces the simple iterative scheme in Equation (4.5) with a step that requires the solution of a system of linear equations. We have opted to use the original procedure, because it is easy to implement and gives good results in practice.

A motivation for using Radó's algorithm is that it effectively deals with pathological situations that arise when the algorithm must contend with coincident or colinear points. When three or more stations are colinearly situated, there is not always a unique minimizer, since many points on their connecting line segment can be connected to the stations with the same amount of optical fiber, so we must check for this condition before proceeding with the coupler location algorithm. Moreover, if two stations share the same location, then it is clear that their coupler should be collocated with them, but this would cause the expression in Equation (4.5) to be undefined because of zeros in the denominators. We pragmatically eliminate such situations by using a modified Euclidean distance metric that fixes the distance between two points to be a small positive number whenever the points lie within a small common neighborhood of given radius. In this way the distance between two points is never equal to 0 (even if the points are collocated), and the expression in Equation (4.5) never takes on undefined values.

4.3.3 Good Initial Solutions for the Location Algorithm

It is necessary to begin the coupler location algorithm with some initial solution $(\boldsymbol{V}_0^{(0)}, \boldsymbol{V}_1^{(0)}, \ldots, \boldsymbol{V}_M^{(0)})$, the choice of which can significantly accelerate or retard the ultimate discovery of the optimum solution. Drawing from the physics of many-particle systems, we can define the *center of mass* of a set of points $\mathcal{W} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_n\}$, to be the point located in the midst of these points:

$$\mathrm{COM}(\mathcal{W}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_i$$

Placing a point at the center of mass of collection of points does not necessarily minimize the distance covered by the radial arms from the center of mass to all the points of the collection, but it is a reasonable approximation of the optimum [VR67].

To start the location algorithm, we must specify initial locations for all couplers in the WON. This is best performed in conjunction with the clustering algorithm, since it associates couplers with clusters of stations. The hierarchical clustering results in a tree whose leaf nodes are stations and whose nonleaf nodes are couplers. When the hierarchy of clusters has been formed, initial locations are assigned to the couplers from the headend (root) toward the leaves. Assume that the couplers are numbered from 1 to M and the stations are numbered from $M+1$ to $M+N$. In correspondence with the numbering of stations and couplers, the stations have locations $\boldsymbol{V}_{M+1}, \boldsymbol{V}_{M+2}, \ldots, \boldsymbol{V}_{M+N}$ and the couplers have yet-to-be-determined locations $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_M$. Let $\mathcal{D}_i$ be the set of locations of all stations in the subtree rooted at coupler i. As before, we let $p(i)$ represent the parent of node i.

We begin by assigning the headend coupler a location $\boldsymbol{V}_1$ at the center of mass of all stations:

$$\boldsymbol{V}_1 = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{V}_{M+i}$$

The procedure continues by assigning locations to the couplers of the tree in a breadth-first manner. At succeeding levels the coupler $\boldsymbol{V}_i$ is assigned a location at the center of mass of its parent coupler and all stations for which it is the root:

$$\boldsymbol{V}_i = \mathrm{COM}\left[\mathcal{D}_i \cup \{V_{p(i)}\}\right]$$

This tends to bias the location of couplers so that they fall more or less in the geometric centers of their subtrees. This also ensures that a common run of optical fiber feeds as deeply into the subtree as possible, which in turn conserves the length of optical fiber. The procedure is complete when all couplers have been assigned initial locations. The locations $\boldsymbol{V}_1$, $\boldsymbol{V}_2$, ..., $\boldsymbol{V}_M$ are then passed to the coupler location algorithm. Initial locations chosen in this way have proven to be good estimates of the optimized cost and are sometimes within a few percent of the optimum.

4.4 Empirical Studies

Having presented the PTDP and algorithms for its solution, we now turn to demonstrating the techniques on several problem instances. We first present a method for synthetically generating network geographies, i.e., an embedding of stations in the plane. We then perform and discuss a number of experiments to compare the minimized costs of the different topological classes for several different geographies.

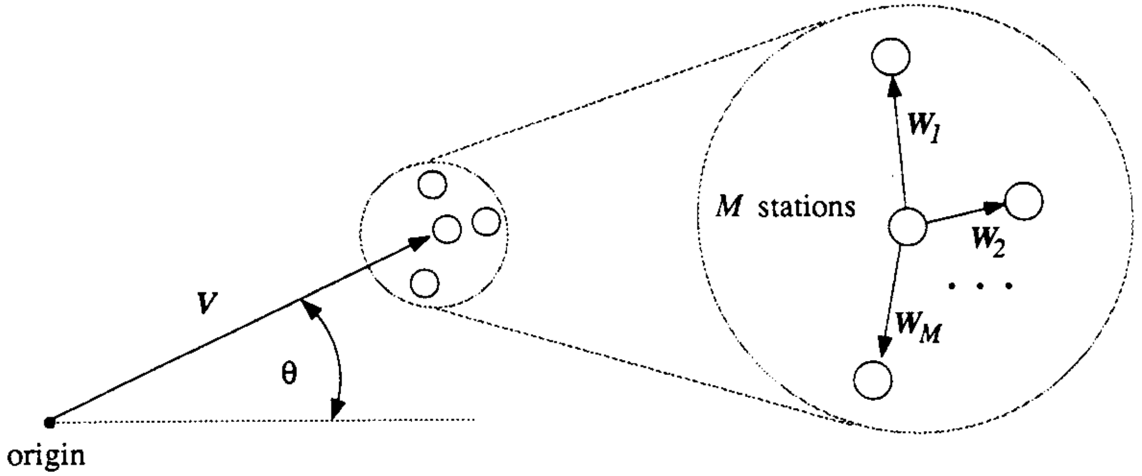


Figure 4.7: A Model for Generating WON Geographies.

4.4.1 The Geography Model

Since we do not have real data describing the geographical distribution of stations, we first require a method for artificially generating station locations in the plane. We would like to be able to specify many kinds of network geographies, including those with homogeneous densities, clumps of high-density clusters, and arbitrary distributions of interstation distances. To accomplish this we adopt the simple geography model illustrated in Figure 4.7, which probabilistically assigns locations to stations in the plane. In Figure 4.7 $\boldsymbol{V}$ and $\boldsymbol{W}_i$ are random vectors in $\mathbb{R}^2$, while M is a discrete random variable. We specify a random vector by choosing from a known probability distribution the vector's direction θ and magnitude $|\boldsymbol{V}|$. The procedure used is to repeat the following actions until the requisite number of station locations have been generated:

1. Generate the random vector $\boldsymbol{V}$.
2. Generate the random number M.

3. Generate M random vectors $\boldsymbol{W}_1,\ \boldsymbol{W}_2,\ \ldots,\ \boldsymbol{W}_M$.
4. Assign to the next M stations the locations $\boldsymbol{V}+\boldsymbol{W}_1,\ \boldsymbol{V}+\boldsymbol{W}_2,\ \ldots,\ \boldsymbol{V}+\boldsymbol{W}_M$.

The random vectors $\boldsymbol{W}_m$ are usually small displacements compared to $\boldsymbol{V}$ and serve only to create a cluster of locations centered about $\boldsymbol{V}$.

To generate network geographies for our study, we employ a gamut of probability distributions for use in the geography model. For vector magnitudes we use the following primitive random variables:

1. Constant (deterministic)
2. Uniform
3. Exponential
4. Normal

The vector's direction θ (in radians) is always chosen uniformly on the interval $[0,\ 2\pi)$. The random variable M is used to control the degree to which stations are geographically clustered, i.e., if $M \equiv 1$ then the stations' locations are statistically independent of each other. Using a constant value of 1 for M means that station locations are unclustered, whereas allowing M to assume values greater than 1 means that stations can have correlated locations.

Although the geography model is capable of generating a wide variety of network geographies, it is not obvious that they are actually representative of naturally occurring patterns of user locations. The synthetically generated network geographies are probably more homogeneous than naturally occurring ones. This is because the geography model treats the plane as a uniform disk, within which each random vector $\boldsymbol{V}$ is embedded according to a given probability distribution. On the other hand, a natural geography, with its variation of terrain and inhabitation, would not—except in rare cases—permit stations to fall just anywhere. In other words, real cities do not have a uniform pattern of inhabitation—bodies of water and mountains, around which cities often spring up, break up the patterns of settlement into a number of blob-like regions.

Our geography model takes much more of a "dart board" approach and thus simplifies reality somewhat. The model does, however, capture at least one important characteristic of urbanization, which is the propensity for cities to have a dense core surrounded by suburbs that are less and less densely populated the farther out one travels from the center.

It would be possible to generate synthetically network geographies that are more realistic than those we are now using. For instance, we could use demographic data for a specific city and randomly locate a station in a given neighborhood according to its relative population density. Therefore, if we use the population map of a city as a template for generating potential station locations, then unpopulated areas would host no stations, and the resulting station locations would be in proportion to the population density and thus fill in the contours of the city. Similar effect might be achieved with the geography model by later thinning out regions of the plane or superimposing two or more disks in an overlapping arrangement.

4.4.2 Experiments in Physical-Topology Design

We now discuss a series of experiments aimed at a comparative evaluation of different optimized physical topologies for the WON. The experiments involve solving the PTDP for a diverse set of problem instances. We then analyze the results of the experiments and draw conclusions.

Unless we state otherwise, the experiments are performed on a 64-station WON. The physical topology of the WON can be a star, tree, two-clustered tree, or four-clustered tree, where a k-clustered tree is one in which the leaves are clusters of k stations, as explained in Section 4.1. We generate six network geographies by allowing the magnitude of the random vector $\boldsymbol{V}$ to be chosen from either the uniform, exponential, or Gaussian probability distribution; the random variable M is then chosen either (deterministically) equal to 1 or (nondeterministically) from a geometric distribution with mean equal to 2. The geography parameter ranges from 1 to 6 and is interpreted as follows:

- geography $= 1$: uniform $\|\boldsymbol{V}\|$ and constant M
- geography $= 2$: exponential $\|\boldsymbol{V}\|$ and constant M
- geography $= 3$: Gaussian $\|\boldsymbol{V}\|$ and constant M
- geography $= 4$: uniform $\|\boldsymbol{V}\|$ and geometric M
- geography $= 5$: exponential $\|\boldsymbol{V}\|$ and geometric M
- geography $= 6$: Gaussian $\|\boldsymbol{V}\|$ and geometric M

The geographies numbered 1–3 are unclustered and those numbered 4–6 are clustered. The mean of $\|\boldsymbol{V}\|$ is fixed at 50 km; thus, the average distance from a station to the center of the plane is 50 km. No station is allowed to lie more than 100 km from the center of the plane. This combination of random variables results in a network geography similar to that indicated by the dots in Figure 4.8, which is produced by choosing $\|\boldsymbol{V}\|$ to be uniformly distributed and $M \equiv 1$. The four topo-

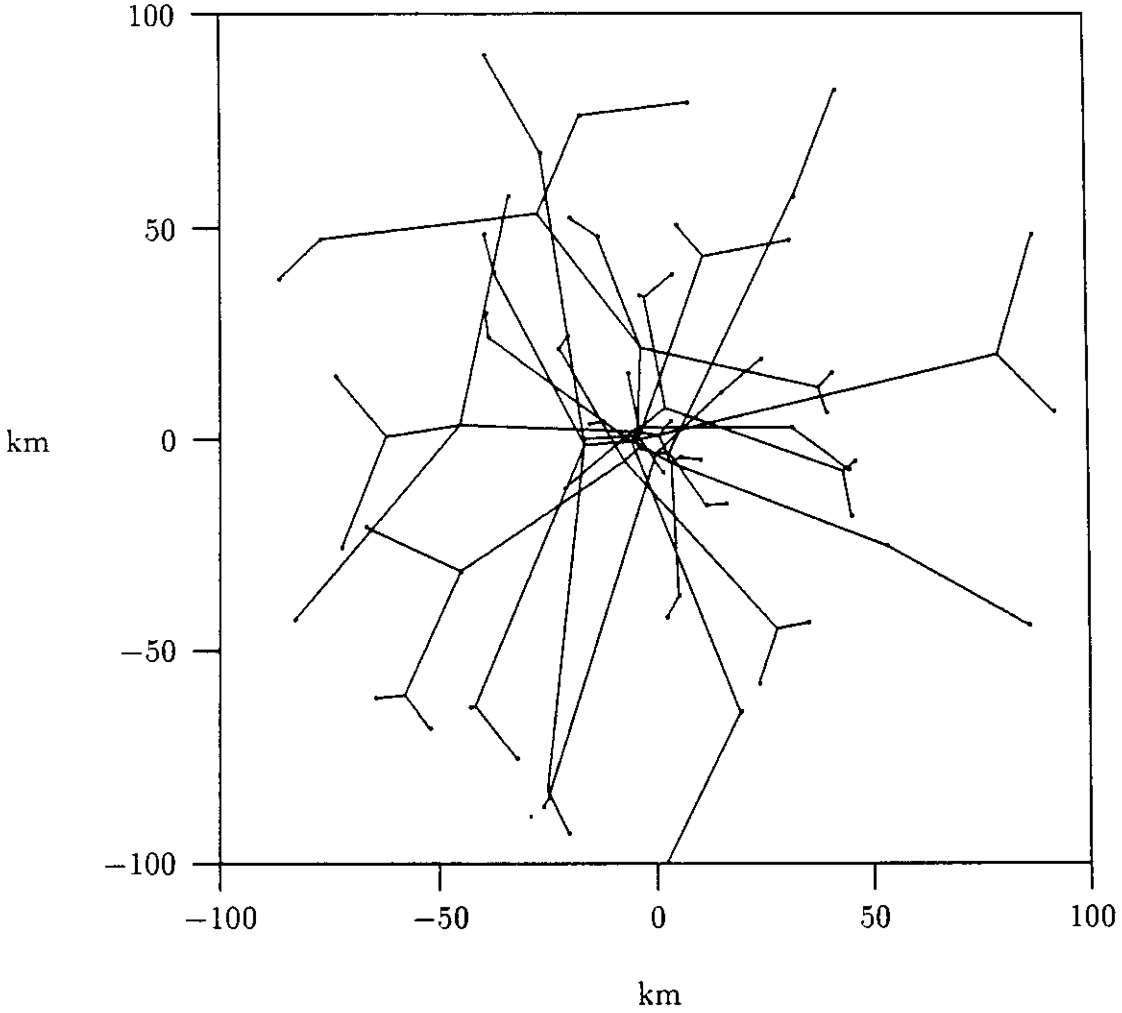


Figure 4.8: The Physical Topology for a 64-Station Tree.

logical classes and six network geographies correspond to 24 distinct experiments in physical-topology design.

Before we study the topological classes of Figure 4.2, we must first check that we can design a suitable cable plant for each of the topological classes. We assume that signal loss is 0.2 dB per km of optical fiber, 3 dB per optical tap, 4 dB per optical coupler, and 18 dB for the 64-port reflective star-coupler. If we amplify the signal at the headend of the trees, then each class of topology in our experiment can be shown to have a worst-case transmitter-to-receiver loss of 45 dB or less. Therefore, these classes are all feasible from the viewpoint of signal quality, and we next attempt to achieve optimum cost and performance for each.

We must keep in mind that comparing the costs of different topological classes is not free of pitfalls. Besides the cost associated with optical fiber, each physical topology uses a different number of couplers. In fact, different physical topologies may use radically different coupler designs. To illustrate, we see that the four topological classes in Figure 4.2 use, in addition to differing lengths of optical fiber, different types and quantities of components in their cable plants. Whereas the star uses only an $N \times N$ star coupler, the trees use three-port (or modified four-port) couplers, bus taps, and possibly a regenerating headend. Furthermore, each tree in Figure 4.2 uses a different number of these latter components; i.e., the ordinary tree uses 14 couplers and one headend; the two-clustered tree uses 16 taps, six couplers, and one headend; and the four-clustered tree uses 16 taps, two couplers, and one headend. Obviously, then, the final costs will include not only the cable-specific costs, but also the costs of each component as well. Therefore, these costs would have to be considered in the final analysis; their incorporation into the cost model would be straightforward, given that each component's cost is

Table 4.1: Cost Comparisons of Four Optimized Physical Topologies for the 64-Station WON.

Network Geography	Cost (km)			
	Star	Tree	2-Tree	4-Tree
1	3373	1966	1526	1227
2	3237	2107	1653	1678
3	3208	1826	1501	1281
4	3026	1367	1157	962
5	3256	1702	1457	1218
6	3169	1649	1205	946

known in advance. If, as we expect, the cost of cable installation and maintenance dominates the cost of components, then comparing different topological classes on the basis of cable length is acceptable.

We now solve the PTDP for 24 instances of the WON, corresponding to the four topological classes and six network geographies. The solution is obtained by applying the clustering and location algorithms of Figures 4.5 and 4.6, which yield the optimized costs of the WON for each problem instance, as shown in Table 4.1. The cost of the WON tends to decrease in the following order: star, tree, two-clustered tree, four-clustered tree. Taking the star as a point of comparison, we can reduce the cost of the WON by an average of 39, 52, and 57 percent by using the tree, two-clustered tree, and four-clustered tree, respectively, with the unclustered network geographies. Similarly, the average cost savings are 50, 60, and 70 percent with the clustered network geographies. The more ramified topologies obviously benefit from placing adjacent stations on common runs of optical fiber. The costs of the tree topologies are somewhat higher when the network geography consists of unclustered rather than clustered stations. It appears that tree topologies are able to take advantage of geographical clustering of stations to realize cost savings by running a shared fiber to a cluster, whereas a run of fiber to unclustered stations must split and then branch out to the dispersed stations. The cost savings from choosing the four-clustered tree over the star are notable—up to 70 percent in our example—but there remains the question of whether the oblique "glass" paths characteristic of the four-clustered tree will degrade performance substantially when compared to the direct "glass" paths of the star. We turn to answering this question in the next chapter.

The cost of the WON's physical topology is clearly affected by the specific network geography. We see in Table 4.1 that, given a clustered geography, we can achieve noticeably lower cost than when we are given an unclustered geography, at least where the tree topologies are concerned. The geographical clustering of stations in the plane facilitates the formation of compact clusters by the clustering algorithm. Thus, the closely situated stations in a cluster are able to share a run of optical fiber over greater distance than if they were widely separated, which brings down the cost of cabling. We note that the particular probability distribution chosen for $\|\boldsymbol{V}\|$ does not appear to affect significantly the cost of the WON.

An example of an optimized tree topology for the network geography generated from a uniformly distributed $\|\boldsymbol{V}\|$ and $M \equiv 1$, i.e., no clustering, is illustrated in Figure 4.8

We also optimize the physical topologies for WONs of various sizes and compare their costs. The network geography is generated from a uniformly distributed $\|\boldsymbol{V}\|$ without clustering (i.e., $M \equiv 1$), and the four topological classes are compared after optimization. The results, shown in Figure 4.9, for WONs ranging in size from 16 to 128 stations, further support the cost ranking of topological classes seen in the

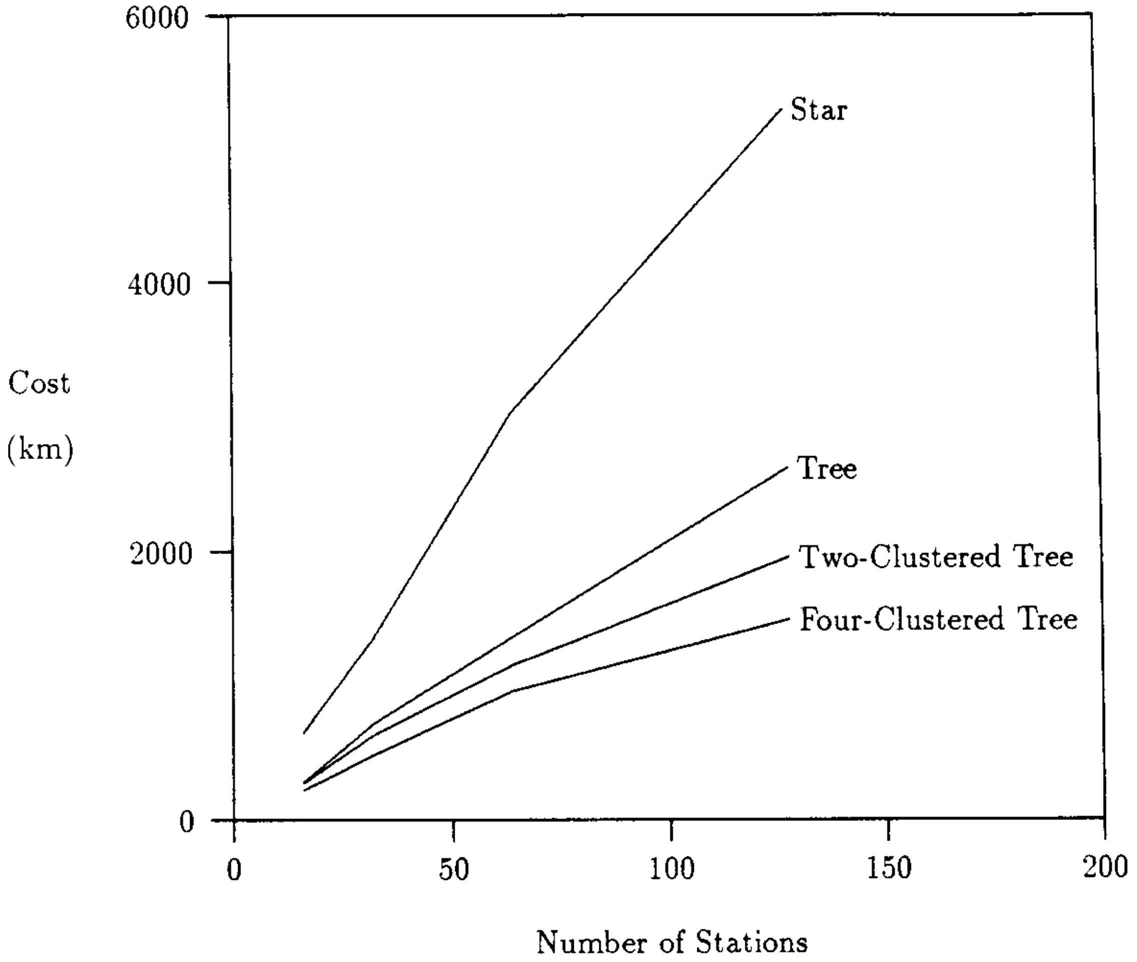


Figure 4.9: A Comparison of Costs of Four Physical Topologies.

previously presented data.

These experiments give an indication of the comparative cost of each of the four different topological classes after optimization. The experiments strongly support the use of the clustered tree as a physical topology for the WON, if cost is the primary concern. But, as we have pointed out earlier, the pursuit of maximum cost savings must be tempered with considerations of reliability and security. Moreover, the physical topology influences the level of performance that can be achieved after the virtual topology is designed, since every physical topology has a different distance matrix. The former considerations are not easily amenable to quantitative analysis, but we shall quantitatively address the latter consideration in the next chapter.

4.5 Summary

This chapter presents a mathematical formulation of the physical-topology design problem, which seeks to find a minimum-cost topology and placement of non-stationary equipment in the WON. The physical-topology design problem is decomposed into the station–coupler clustering problem and the coupler location problem. We propose a suboptimal algorithm to solve the station–coupler clustering problem and an optimal algorithm to solve the coupler location problem, and we apply the algorithms to the topological design of several WONs. Of the four basic topological classes that we consider for the WON, i.e., the star, tree, two-clustered tree, and four-clustered tree, the costs of the optimized tree topologies are roughly comparable to each other, but significantly less than the cost of the star topology. The cost of clustered topologies is seen to be lower than the cost of unclustered topologies, especially when the network geography is such that stations occur in localized clusters.

Chapter 5

The Virtual-Topology Design Problem

Without doubt, the most unique characteristic of the WON is its protean ability to assume—or even alter—its virtual topology independent of its physical attributes. The rich assortment of virtual topologies available to the network designer opens up the possibility of choosing that virtual topology that best satisfies some set of specific design criteria. In particular, this chapter addresses the problem of optimizing system performance by selecting a virtual topology that performs best under the conditions prevailing upon the WON.

We discuss the importance of the WON's virtual topology in determining system performance, explore the mathematical foundations of virtual-topology design, present a mathematical formulation of the VTDP, propose approaches for the solution of the VTDP, and validate our approaches by obtaining solutions for several problem instances.

5.1 Background

5.1.1 Dense Regular Digraphs of Low Degree

The wide interest in designing dense digraphs of low degree stems from the use of such digraphs in several fields of computer design. The problem can be traced back to the $(p,\ D)$-digraph problem, which requires one to find the largest (in terms of the number of nodes) digraph in which all nodes have outdegree p or less and no shortest path from one node to another consists of more than D hops.

It is well known [HS60] that the number of nodes N in any digraph of maximum degree p and diameter D satisfies the so-called Moore bound:

$$N \leq \frac{p^{D+1} - 1}{p - 1} \tag{5.1}$$

The inequality in Equation (5.1) is, in fact, strict when $p > 1$ and $D > 1$ [BT80].

In virtual-topology design we are more interested in the dual of the (p, D)-digraph problem, which is to find the N-node digraph with degree p that has the smallest diameter D. In this case the analog of Equation (5.1) is [II83]

$$D \geq \left\lceil \log_p[(p-1)N + p] \right\rceil - 1$$

In fact, the diameter of a digraph is only a crude indicator of its intrinsic value as a virtual topology for a WON. The diameter merely places a bound on the maximum number of hops required to get from one node to any other. The mean internode distance of a digraph would be a more realistic metric for evaluating the digraph. Under the assumption that traffic is uniform, the following lower bound on mean internode distance $\overline{D}$ holds [CCMS73]:

$$\overline{D} \geq \frac{L(N-1) + p(L+1) - p^{L+1}}{N - 1} \tag{5.2}$$

where

$$L \triangleq \left\lceil \log_p[(p-1)N+1]-1 \right\rceil$$

The search for dense digraphs has inspired an informal but wide-ranging competition among graph theorists, mathematicians, and computer scientists, who seek to outdo each other by constructing graphs that come closer to the Moore bound than the constructions of the previous record holders. Bermond *et al.*, who are prominent players in the competition, have periodically published tables of the known record holders for several values of p and D [BDQ82, BDQ86]. Record-breaking graphs are constructed by sophisticated mathematical techniques that rely on such diverse fields as combinatorics, group theory, and projective geometry. These graphs are constructible by following a simple recipe that specifies the way in which nodes are to be interconnected. As intriguing as the game is, it has only peripheral applicability to virtual-topology design. Although a small-diameter digraph, by virtue of its denseness, is also likely to have a small internode distance, there is no guarantee of this. Also, the graph constructions do not usually take into account the effects of traffic and distance matrices, i.e., these matrices are almost always assumed to be uniform. For reasons such as these we choose to take a more direct approach and explicitly optimize digraphs with respect to specific parameters and performance measures.

5.1.2 Virtual Topologies for the SCWON

The search for dense regular digraphs centers on digraphs in which all nodes have the same indegree and outdegree. While such digraphs are adequate to model the virtual topology of the DCWON, they do not model the virtual topology of the SCWON very well. In the SCWON an output port of a station can transmit to an input port of several stations. Therefore, a proper representation of the virtual topology of the SCWON consists of a mapping of nodes to channels. One mathematical structure that can model this is a hypergraph, which is defined as a collection of hyperedges, each of which is a set of nodes. The hyperedges correspond to WDM channels, and the nodes on the hyperedge are identified with stations tuned to the corresponding channel.

Next we characterize the limits of performance possible in the SCWON. We assume that the WON is subjected to a uniform traffic load of γ. Furthermore, we assume that channels are all shared by the same number of transmitters and that load can be perfectly balanced over all the channels. For channel performance we use the simple threshold model, in which the access delay on a channel is assumed to be negligible until the load on the channel reaches a threshold value of $\rho_{\max}$, past which point the delay is unbounded. Given that K channels are to be used in an SCWON of N stations and that each station has p transceivers, it is shown in Appendix A that the least possible value for the mean number of hops, H, in the WON is

$$H = \frac{LN}{N-1} - \frac{L-(L+1)pd+(pd)^{L+1}}{(N-1)(1-pd)^2} \tag{5.3}$$

where

$$L \triangleq \left\lceil \log_{pd}[(pd-1)N+1]-1 \right\rceil$$

and

$$d \triangleq \left\lceil \frac{N-1}{K} \right\rceil$$

This theoretical limit on performance applies when channel overload is avoided, and so if all channels were perfectly balanced, the following constraint on channel utilization must be satisfied: $\gamma H/K < \rho_{\max}$. In Figure 5.1 we show the best possible performance that could be achieved using shared channels by plotting

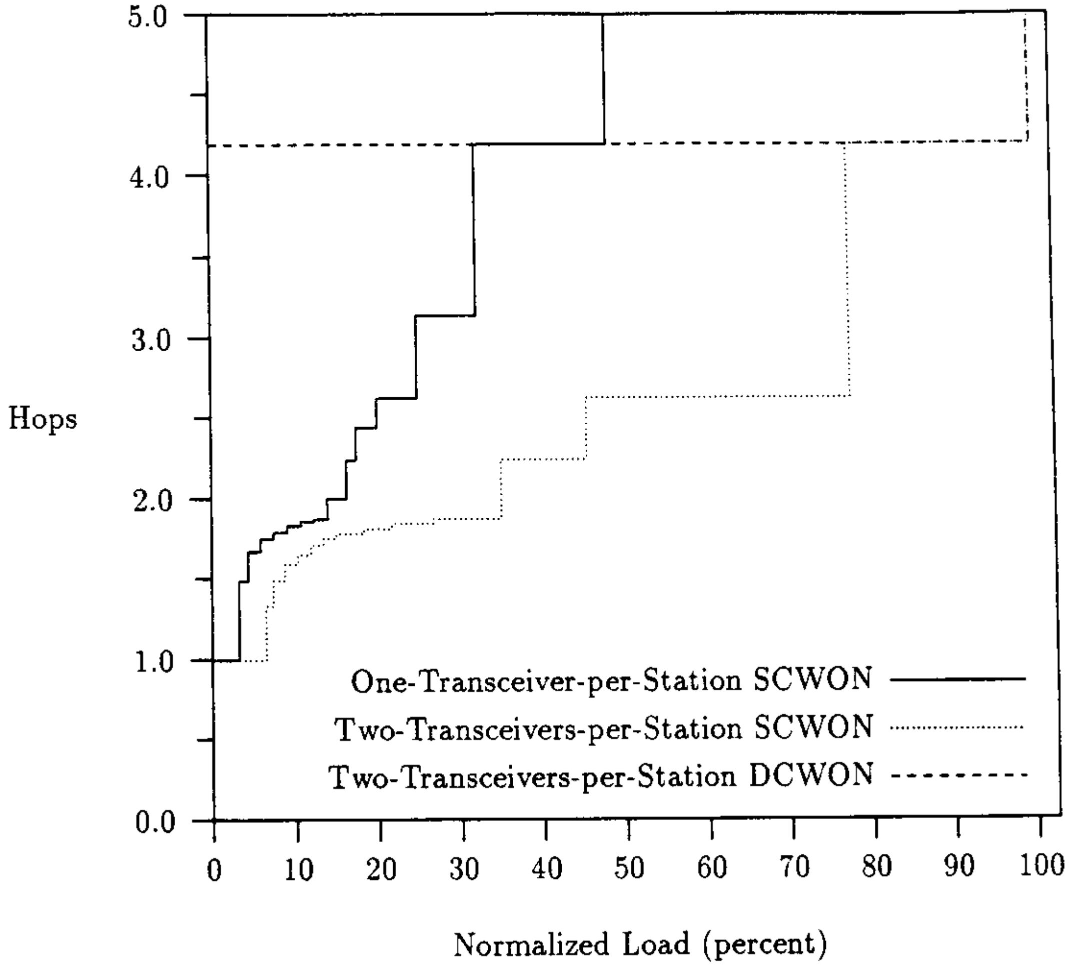


Figure 5.1: Lower Bounds on Mean Internode Distance in the SCWON.

the least value of Equation (5.3) that can be achieved in the 64-station WON for a specific value of γ. The value of $\rho_{\max}$ is held at 80 percent of a channel's capacity. The flat dashed curve corresponds to a two-transceivers-per-station WON in which only dedicated channels are used. The dotted curve represents a two-transceivers-per-station WON in which shared channels are permitted; a point on this curve corresponds to the minimum hop-count that might be theoretically achieved without overloading any of the perfectly balanced channels. The solid curve is similar to the dotted curve, but it applies to a one-transceiver-per-station SCWON. It can be seen that the SCWON offers performance that is superior to the DCWON for up to 78 percent of the highest throughput achievable by the DCWON.

5.2 Problem Description and Formulation

The VTDP is a combinatorial optimization problem in which the objective is to assign the WDM channels to stations in order to optimize the virtual topology with respect to the performance of the WON. Taking mean packet delay, as expressed by Equation (3.4), to be our primary performance metric, we note that under normal traffic conditions propagation delay forms the dominant component of the total delay in a typical WON, and thus the VTDP can be broadly viewed as an attempt to minimize the average distance traveled by a packet. It is important to note that this delay is also a function of the WON's traffic matrix, which specifies the rate of the traffic exchanged between any pair of stations.

Assuming that mean packet delay is our principal performance metric, we can formulate the VTDP as a nonlinear zero–one mathematical program. The decision variables are comprised of the tuning matrix, (κ_{ijk}), an entry of which specifies whether station i transmits to station j over channel k, and the routing matrix $(\pi_{ij}^{(lm)})$, an entry of which specifies the probability that a packet will be routed from station i to station j, given that the source of the packet is station l and its destination is station m. Both sets of decision variables (κ_{ijk}) and $(\pi_{ij}^{(lm)})$ assume discrete values from the set $\{0, 1\}$. The tuning matrix (κ_{ijk}) defines how stations are assigned to WDM channels and can be interpreted as the WON's virtual topology. By making the routing matrix $(\pi_{ij}^{(lm)})$ take discrete values, we are stipulating that the network use fixed single-path routing, in which there is only one path from any source to its destination. Since the routing matrix for given l and m must represent a valid path in the WON, we require that there exist positive integers $i_1, i_2, \ldots, i_n$ such that $\pi_{li_1}^{(lm)} = \pi_{i_1 i_2}^{(lm)} = \cdots \pi_{i_{n-1} i_n}^{(lm)} = \pi_{i_n m}^{(lm)} = 1$ and all other $\pi_{ij}^{(lm)} = 0$. The stipulation of single-path routing is not necessary—we could allow alternate routing and it would not affect the mean propagation delay because queueing effects are in essence ignored. The basic problem is to arrange the virtual topology of the network in a such a way that the routing procedures can deliver packets in the least amount of time. The problem may thus be expressed as a minimization of the mean packet delay, subject to connectivity and flow-conservation constraints. Let us define $\gamma_i \triangleq \sum_{j=1}^{N} \gamma_{ij}$ as the rate at which network traffic is generated by station i and $\eta_i \triangleq \sum_{j=1}^{N} \gamma_{ji}$ as the rate at which network traffic is consumed by station i. Recall that Equation (3.4) gives an approximate closed-form expression for the mean packet delay in the dedicated- or shared-channel WON, assuming that the routing procedure, distance matrix, and traffic matrix are known; this expression is used as the objective function of the VTDP. Formally, we propose to minimize the mean packet delay:

$$\mathbb{E}[T] \;=\; \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\lambda_{ij}\, \delta_{ij}}{\tilde{c}\, \gamma} \;+$$

$$\sum_{k=1}^{K} \frac{p_{k0}\,\lambda_k}{\gamma\,\mu B} \left\{ \frac{\mu_k(g_k)}{\mu_k(g_k-1)} \frac{1}{(1-\lambda_k/\mu B \mu_k(g_k))^2} + \right.$$
$$\left. \sum_{i=0}^{g_k-2} i \left(\frac{\lambda_k}{\mu B}\right)^i \left[\frac{1}{\prod_{j=2}^{i+1} \mu_k(j)} - \frac{1}{\mu_k(g_k-1)\,[\mu_k(g_k)]^{i-1}} \right] \right\} \tag{5.4}$$

subject to

$$\eta_i + \sum_{j=1}^{N} \lambda_{ij} = \gamma_i + \sum_{l=1}^{N} \lambda_{li} \quad \forall i \tag{5.5}$$

$$\lambda_k = \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_{ij} \kappa_{ijk} \quad \forall k \tag{5.6}$$

$$\lambda_{ij} = \sum_{l=1}^{N} \sum_{m=1}^{N} \pi_{ij}^{(lm)} \gamma_{lm} \quad \forall i \, \forall j \tag{5.7}$$

$$\pi_{ij}^{(lm)} \leq \sum_{k=1}^{K} \kappa_{ijk} \quad \forall i \, \forall j \, \forall l \, \forall m \tag{5.8}$$

$$\sum_{j=1}^{N} \pi_{ij}^{(lm)} \leq 1 \quad \forall i \, \forall l \, \forall m \tag{5.9}$$

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \kappa_{ijk} = 1 \quad \forall k \tag{5.10}$$

$$\sum_{k=1}^{K} \max_{j=1}^{N} \kappa_{ijk} = p \quad \forall i \tag{5.11}$$

$$\sum_{k=1}^{K} \max_{i=1}^{N} \kappa_{ijk} = p \quad \forall j \tag{5.12}$$

$$\lambda_{ij} \geq 0 \quad \forall i \, \forall j \tag{5.13}$$

$$\lambda_k \geq 0 \quad \forall k \tag{5.14}$$

$$\kappa_{ijk} \in \{0,1\} \quad \forall i \, \forall j \, \forall k \tag{5.15}$$

$$\pi_{ij}^{(lm)} \in \{0,1\} \quad \forall i \, \forall j \, \forall l \, \forall m \tag{5.16}$$

In Chapter 3 we have noted that queueing delay in the typical DCWON is negligible in comparison to propagation delay. By the time queueing delay approaches the level of propagation delay, there is a high likelihood that one or more stations of the network will have experienced serious problems with buffer overflow (and, consequently, packet loss). However, in the formulation of the VTDP for the SCWON, the objective function in Equation (5.4) must incorporate the appropriate state-dependent service rate to account for queueing delay caused by the station waiting to access its WDM channels. This is because propagation delay in the SCWON would be minimized by simply placing all stations on one common channel, which is clearly undesirable. Furthermore, the multiaccess scheme may be sensitive to factors such as traffic loading or station population on the channel. We must therefore penalize against this situation by using queueing delay to drive up the value of the objective function whenever a channel is unfavorably loaded.

There is a large collection of constraints in the VTDP, and we now explain the significance of each constraint in detail. First note that, in addition to the primary decision variables κ_{ijk} and $\pi_{ij}^{(lm)}$, we have introduced secondary decision variables λ_{ij} and λ_k, which represent the traffic flows through stations and channels, respectively. These traffic flows are related to the routing variables in terms of the system of linear equations shown in Equations (5.7) and (5.6) and can be efficiently computed from (γ_{ij}), (κ_{ijk}), and $(\pi_{ij}^{(lm)})$. Equations (5.5) and (5.6) are conservation-of-flow conditions, which specify that the traffic flowing into a station or a channel balances with the traffic flowing out of the station or channel. Equation (5.7) states that a routing chain contributes flow only to those hops over which the chain passes, Equation (5.8) permits flow only between stations tuned to the same WDM channel, and Equation (5.9) guarantees that the routing probabilities are valid probabilities. Equations (5.11) and (5.12) say that every station is assigned to precisely p receive and p transmit channels, while Equation (5.10), which applies only to the DCWON, says that each channel has exactly one transmitting and one receiving station. Finally, Equations (5.13)–(5.16) specify the possible values over which the primary and secondary decision variables may range.

The zero–one decision variables $(\pi_{ij}^{(lm)})$ specify a single-path route from each possible source station i to each possible destination station j. The use of bifurcated flows, wherein a packet from a given source to a given destination can take alternate paths, can reduce the delay below that which can be achieved by single-path routing. This reduction is realized by diverting some traffic from more-utilized links to less-utilized links, thereby alleviating some of the queueing delay. Such a routing scheme, however, is not easily implementable in the WON, where the need for fast routing favors single-path routing. A preferred solution would use source routing, in which the source station places the route to be taken by each packet into its header. Although source routing does not preclude bifurcation of flows, the bifurcation would certainly not be done on a packet-by-packet basis, as is usually practiced in multipath routing. Moreover, in contrast to the traditional networks in which transmission and propagation times are comparable, the expected improvement in delay from using bifurcated flows will generally be small, because propagation delay dominates queueing delay by a wide margin. Therefore, reducing the queueing delay by means of bifurcating flows will usually yield only marginal improvement.

In the next section we address the issue of how to solve the VTDP defined by Equations (5.4)–(5.16).

5.3 Solution Approaches

The VTDP, as specified in Equations (5.4)–(5.16), is intuitively a difficult problem to solve. Although problems related to the VTDP have been investigated, such as the task of finding the minimum-diameter regular graph of a given degree and size [TS79], the approaches used to attack these problems have been based upon heuristics or *ad hoc* techniques. For example, the approach used in [TS79], which was based on local (downhill) search, is not guaranteed to yield a global minimum. Moreover, these techniques often require extensive computation, which underscores the need to balance the quality of the solution against the amount of computation used to obtain this solution. In view of the long history of the $(p,\ D)$-digraph problem (see Section 5.1), which is a special case of the VTDP, it is reasonable to ask how hard it is to obtain solutions to the VTDP. We note that the VTDP is not intractable in the technical sense of the word, since we can provide an algorithm that solves it in time bounded by a polynomial in the problem size; but, as we shall soon see, it is intractable in practice.

If we pose a restricted version of the VTDP as the problem of finding the minimum mean internode distance among all N-node, p-regular digraphs, then the following is a polynomial-time algorithm for its solution:

1. Generate a new N-node digraph G with maximum outdegree p.
2. Evaluate the mean internode distance of G, and if it is the minimum seen so far then assign $G_{\min} = G$.
3. If all N-node digraphs of maximum outdegree p have been examined then **halt**, else go to 1.

If we represent a digraph in terms of its $N \times N$ adjacency matrix, then it is easy to see that there are $\binom{N}{p}$ distinct combinations of p 1s and $N - p$ 0s in a row of the matrix for an N-node digraph with maximum outdegree p. There are thus $N\binom{N}{p}$

possible adjacency matrices;[1] in other words, there are $O(N^{p+1})$ such digraphs. Assuming that the algorithm generates new adjacency matrices (i.e., digraphs) on a row-by-row basis, we see that the total time to generate the matrices is $O(N^{p+2})$. We could use the Floyd–Warshall all-pairs shortest paths algorithm [Flo62, War62] to evaluate the mean internode distance, which requires running time $O(N^3)$. The total running time for the algorithm, therefore, is $O(N^{p+5})$. Even for a WON of modest size, say with 100 two-transceiver stations, the algorithm would require on the order of 10^{14} basic operations, so that, given a computer that performs ten million basic operations per second, the running time would be about 116 days! Therefore, we must consider approximation algorithms to solve realistic instances of the VTDP.

The shared-channel VTDP, on the other hand, is intractable in the sense of complexity theory. We prove in Appendix D that the shared-channel VTDP is NP-complete, which suggests strongly that it probably can not be solved by a polynomial-time algorithm.

The solution method that we apply to the VTDP depends on whether or not we permit the use of shared channels in the WON being designed. We apply in the case of the DCWON the simulated annealing algorithm and in the case of the SCWON the genetic algorithm. Both algorithms are nondeterministic and find optima only in the ergodic sense: given enough running time, both algorithms yield the optimum solution of a well-behaved problem with probability 1. These algorithms can be computationally intensive but generally require far less running time than exhaustive search. To speed up the execution of the algorithms, they have been implemented as parallel programs suitable for execution on the Sequent

[1] These adjacency matrices represent digraphs with maximum outdegree p, but there is no limit on the indegree of a node. If the indegree is bounded, then the enumeration step must be modified to check that the generated digraph satisfies the constraint on maximum indegree.

Symmetry, a shared-memory multiprocessor.

The use of different algorithms for the dedicated- and shared-channel VTDPs is necessary because of their radically different problem structures—the formulation of the dedicated-channel VTDP is in terms of p-regular digraphs, while the formulation of the shared-channel VTDP is in terms of hypergraphs.[2]

Moreover, the cost functions, although both possessing the form of Equation (3.4), behave in fundamentally dissimilar ways, because the tendency to reduce hop count by assigning all stations to one common channel must be counterbalanced against the possibility of overloading such a channel in the SCWON.

We first discuss the dedicated-channel VTDP and then the shared-channel VTDP.

5.3.1 The Dedicated-Channel VTDP

We now address the problem of finding the virtual topology for a DCWON that provides the minimum average packet delay from source to destination, given the network's distance and traffic matrices. To solve the dedicated-channel VTDP, we apply the simulated annealing algorithm [KGV83], which is shown in Figure 5.2.

The simulated annealing algorithm draws heavily from the concepts of statistical mechanics and seeks to cast combinatorial optimization problems in terms of the physical process of cooling a substance from its liquid to its solid form (viz., annealing). To continue the analogy, combinatorial optimization is seen as being similar to physical annealing by virtue of the fact that both processes evolve toward a ground—or minimum-cost—state. In the simulated annealing algorithm for the

[2] A ***hypergraph*** is a collection hyperedges, each of which consists of a set of nodes. This generalizes the notion of a graph, in which an edge is a set of exactly two nodes.

1. *Initialization.* Select an initial temperature T and an initial state S.
2. *Epoch Initiation.* Start a new "epoch".
3. *Perturbation.* Randomly choose a neighbor state S' of S. Assign $\Delta = \text{cost}(S') - \text{cost}(S)$.
4. *Acceptance/Rejection.* Assign $S = S'$ with probability $\min(e^{-\Delta/T}, 1)$.
5. *Temperature Reduction.* If the "epoch" has expired then assign $T = rT$, else go to 3.
6. *Convergence Test.* If the state is "frozen" then halt, else go to 2.

Figure 5.2: The Simulated Annealing Algorithm.

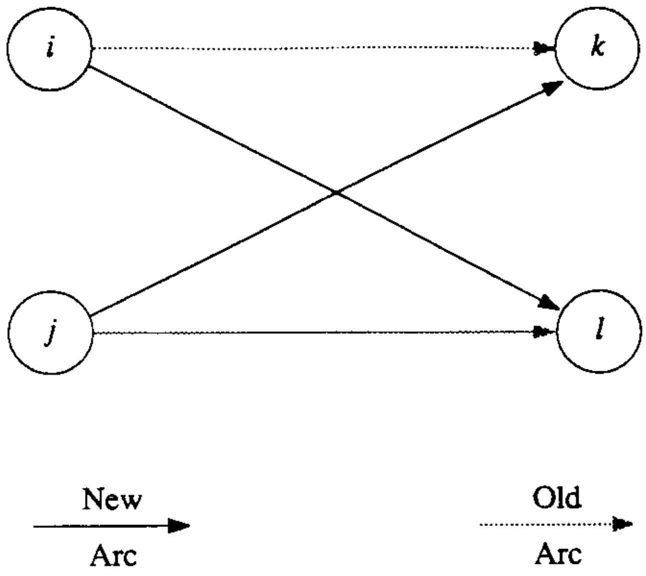


Figure 5.3: The Branch-Exchange Operation on Digraphs.

VTDP, a state corresponds a digraph, so that finding a "frozen" (or minimum-energy) state corresponds to finding a least-cost interconnection graph.

Given that a state corresponds to a p-regular digraph, we choose to perturb it by means of the branch-exchange operation in which a new graph is produced by swapping the targets of two arcs, as shown in Figure 5.3. The branch-exchange operation is attractive because a branch exchange applied to a regular graph produces another regular graph. The program loop in steps 2–5 of Figure 5.2 constitute what is known as the Metropolis algorithm [MRR$^+$53]. Steps 3 and 4 of the Metropolis algorithm can be viewed as a Monte Carlo simulation of the Markov process formed by memoryless transitions[3] from one state to the next. The "epoch" continues until the Markov process reaches a stage of equilibrium; thus each "epoch" must continue for a sufficient amount of time for equilibrium to occur.

[3] A memoryless transition is influenced only by the current state of the process and is independent of the process's history.

The stopping criterion requires that the optimization not continue for more than a specified number of "epochs" without an improvement in cost; when this criterion is satisfied, it is deemed that further improvement is unlikely, and hence the state is declared "frozen". By the time the algorithm has reached this point, few moves are being accepted, and the Metropolis algorithm is essentially performing local or downhill search.

The calculation of the cost function, since it involves finding all shortest paths in the graph being evaluated, is computationally expensive. Given that simulated annealing attempts to find a global minimum by traversing promising regions of the entire search space, we decided to try to speed up the optimization as much as possible. The simulated annealing algorithm was therefore implemented on a Sequent Symmetry parallel processor. The parallelization was accomplished using the Parallel Programming Library provided with the Symmetry's DYNIX operating system. The Parallel Programming Library is a collection of procedures that implement basic parallel programming primitives such as shared locks, synchronizing barriers, and process management, and that can be invoked from a C-language program [Bab88]. Instead of choosing to parallelize the basic annealing loop, as in [CRSV86, DKN87], we parallelized the computation of all shortest paths in the cost function by assigning to different processors the computation of all shortest paths from different source nodes. In other words, the only point at which parallelism is exploited in the simulated annealing algorithm of Figure 5.2 is in computing the function cost(S') in step 3.

Traffic between a particular source–destination pair is routed on a single shortest path without regard for balancing traffic flow along channels. This is known to be nonoptimal, but the need for fast packet switching and routing in the WON suggests that a simple, single-path routing procedure (possibly source routing) be used. Also, recalling that queueing delay contributes only nominally to packet delay, we expect that the extent to which mean packet delay can be reduced by diverting traffic flow is very small, except when traffic load is extremely high.

5.3.2 The Shared-Channel VTDP

In Chapter 2 we alluded to the use of channel sharing for improving the performance of the WON. On a shared channel a given station can transmit to more stations than it could on a dedicated channel, and this increased fanout opens up the possibility of defining denser interconnection graphs. As the extreme case we point out that assigning *all* stations to *one* channel would allow any packet to get to its destination in one hop. These considerations motivate us to study how much improvement can be expected when channel sharing is used in the WON.

The shared-channel VTDP, which seeks to find the virtual topology that permits minimum packet delay in the SCWON, differs from the dedicated-channel VTDP in both its formulation and solution. As discussed earlier, we can not ignore the queueing component of packet delay in the SCWON unless the WON is lightly loaded. Indeed, ignoring the queueing component would imply that the solution of the VTDP in the SCWON is best accomplished by collecting all stations on one channel, a solution that is in general unacceptable because of its tendency to saturate the single channel. We must therefore account for queueing delay by means of penalty functions or constraints in the formulation of the VTDP. For example, we could introduce the constraint that no channel should host more than a fixed, small number of stations. We have chosen to use the penalty-function approach: we explicitly use channel-access delay as our penalty function.

1. *Initialization.* Randomly select a population of M graphs and evaluate $\text{cost}(G_1), \ldots, \text{cost}(G_M)$.
2. *Selection.* Randomly select a subpopulation of K low-cost graphs.
3. *Crossover/Mutation.* Randomly "mate" pairs of the low-cost subpopulation to produce $K/2$ new offspring graphs; allow mutation. Evaluate the cost of each offspring.
4. *Ranking.* Include the offspring graphs into the population and "kill off" $K/2$ highest-cost graphs.
5. *Convergence Test.* If the stopping condition is satisfied then halt, else go to 2.

Figure 5.4: The Genetic Algorithm.

Our experience in solving the dedicated-channel VTDP with the simulated annealing algorithm motivated us consider alternative algorithms for solving the shared-channel VTDP. A major drawback of simulated annealing is the amount of running time that it requires, and this situation is further exacerbated by the fact that the evaluation of our chosen cost function also requires considerable computation time. Furthermore, using a new algorithm allows us the opportunity to compare the solution quality and execution time of different algorithms, which is especially important since the simulated annealing algorithm requires long running times on problems of large size. The genetic algorithm is an alternative to the simulated annealing algorithm that has achieved good results in solving combinatorial optimization problems. To solve the VTDP in the SCWON, we employed the genetic algorithm shown in Figure 5.4. We have also adapted the simulated annealing algorithm, so that it could be applied to solving the shared-channel VTDP. Since the branch-exchange operation produces only regular digraphs, it is not appropriate for generating digraphs that correspond to SCWONs, and our adaptation consisted of generating new shared-channel virtual topologies by the simple redirection of a single arc. Simulated annealing with arc redirection performs well and yields significant improvement in WON performance, but its running time is considerably greater than that of the genetic algorithm. We therefore decided to use the genetic algorithm exclusively for the solution of the shared-channel VTDP.

Given that the simulated annealing and genetic algorithms yield comparable results when applied to the shared-channel VTDP, why do we use the simulated annealing algorithm at all, as we do on the dedicated-channel VTDP? The crossover mechanism in step 3 of Figure 5.4 requires that we generate new digraphs from existing ones. As we shall soon see, there are very natural ways to effect crossover in digraphs that correspond to SCWONs. With DCWONs, i.e., regular digraphs, the situation is different; despite our best efforts, we are unable to discover a crossover mechanism that yields regular digraphs. All the crossover mechanisms that we have considered have a fair likelihood of generating digraphs in which at least one node has either too many or too few incoming or outgoing arcs. Thus, we are forced to use the simulated annealing algorithm on the dedicated-channel VTDP until a crossover and mutation mechanism that preserves digraph regularity can be identified.

In applying the genetic algorithm to the shared-channel VTDP, we have used two different crossover and mutation mechanisms. We began by using a "graph-splicing" mechanism that "mates" two graphs by taking a subset of nodes from one parent graph and the complementary subset from the other parent graph and placing both sets of nodes and their arcs together to construct a new "pseudograph". The pseudograph, which could contain dangling arcs, is then transformed into an offspring graph by a heuristic that connects dangling arcs to nodes without incoming arcs; random mutation consisting of the merging or splitting of channels could also be applied to the offspring graph. This graph-splicing crossover and mutation mechanism, while performing well on the lightly loaded SCWON, is ineffective when higher traffic loads are encountered. We therefore applied a second "graph-overlaying" crossover and mutation mechanism that produces offspring graphs by taking the first parent graph and adding to it all arcs of the second parent graph, as shown in Figure 5.5. Since this overlaying produces a graph in which nodes appear to have twice as many "transmitters" as possible, half the outgoing arcs at each node must be pruned. The pruning is based on a breadth-first search of the graph, which builds a tree of all shortest paths from a randomly chosen root node to every other node in the graph. Arcs are chosen to produce the "bushiest" tree possible so that mean path length would be kept short. We allow random mutation in a manner identical to the graph splicing case. The results for higher loads gotten by the graph-overlaying mechanism are substantially better than those gotten by the graph-splicing mechanism.

The genetic algorithm was parallelized—again, using the DYNIX Parallel Programming Library—in a very natural way by allowing different processors independently to manage the crossover and mutation of pairs of graphs. The running times of the genetic algorithm are considerably shorter than those of the simulated annealing algorithm. The genetic algorithm also seems to find low-cost graphs easily and consistently, though the quality of the solutions is typically not as good as those found by simulated annealing.

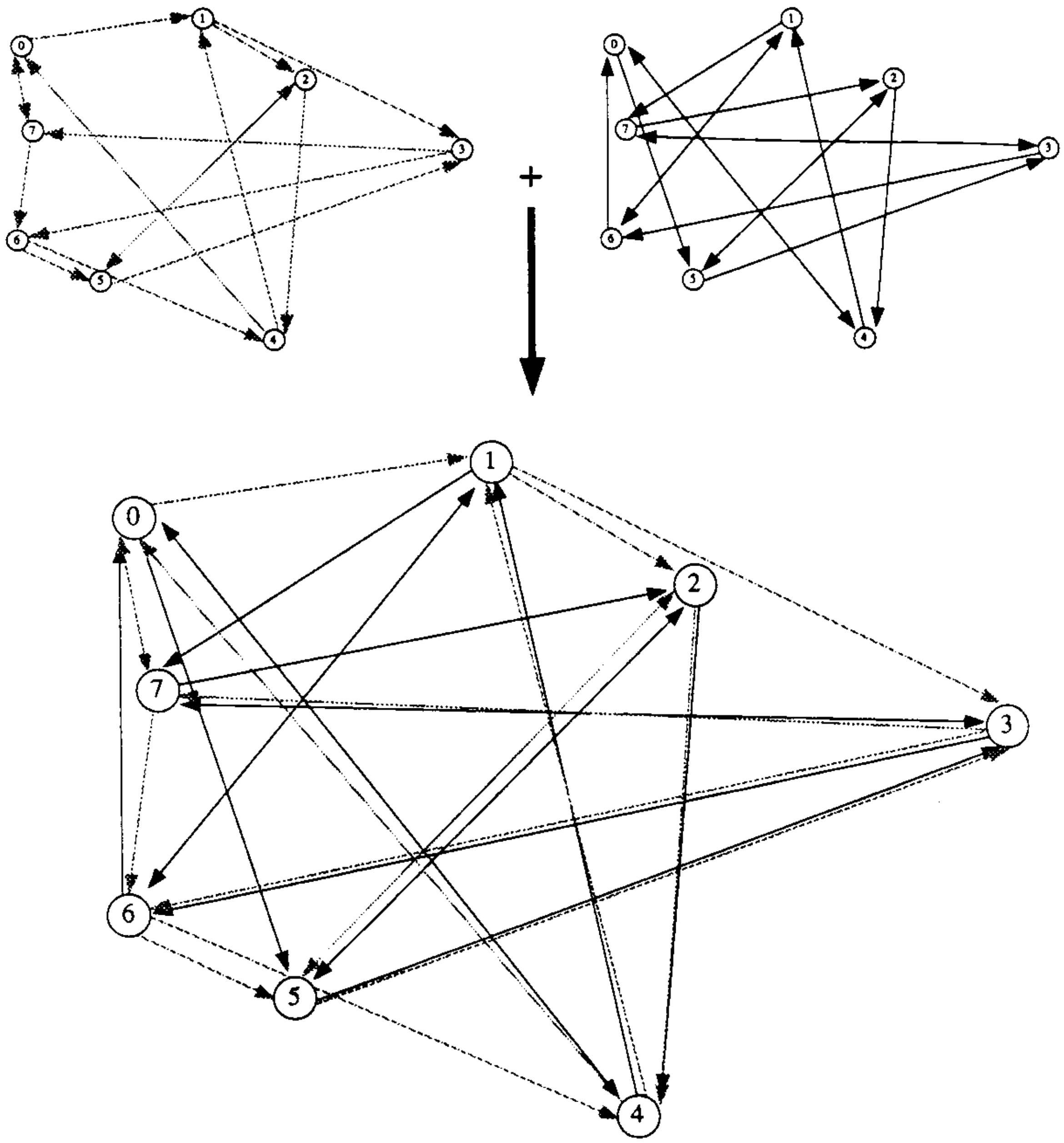


Figure 5.5: The Graph-Overlaying Operation Applied to Two Digraphs.

5.4 Empirical Studies

Next we explore the VTDP for both the DCWON and the SCWON. First we describe the technique used to generate instances of the VTDP. Having done that, we report on specific experiments in the design of the virtual topology of the DCWON and SCWON.

Throughout our experiments we compare the performance of the optimized WON against that of ShuffleNet. Our selection of ShuffleNet's virtual topology as the baseline for comparison is based not upon any desire to discredit ShuffleNet—indeed, ShuffleNet performs very well in the case of uniform traffic—but rather upon the fact that it is well established and is to date probably the best-studied virtual topology. Furthermore, we point out that the regularly structured virtual topologies, such as ShuffleNet and the de Bruijn graph, are often used as starting solutions for our optimization algorithms, which almost certainly ensures that the optimization produces an improvement.

Recalling the formulation of the VTDP in Equations (5.4)–(5.16), we see that the solution of the VTDP is influenced by the parameters of the problem, specifically the traffic and distance matrices (γ_{ij}) and (δ_{ij}). In what follows we treat these parameters as random variables that can change from network to network. For any set of experiments we choose different values for the entries of a matrix by selecting the values from a specified probability distribution. To illustrate, when we choose the interstation distances δ_{ij} from a deterministic distribution with mean 100 km, we are using a model in which all stations are equidistant from the headend. Likewise, we can independently choose each value of δ_{ij} from the uniform distribution with mean 100 km, which results in stations that are scattered over a disk of diameter 200 km, and the typical station is located 50 km from the headend. We assume that stations are scattered over the plane and that the length of optical fiber between station i and station j is given by δ_{ij}. In the following experiments the values of δ_{ij} are chosen by first scattering the stations of the WON over the plane so that their mean distance from the headend is 50 km; thus the mean distance between any pair of stations is 100 km, since a lightwave signal must travel from the first station to the headend and from there to the second station. We call the amount of variability in the distribution of interstation distance *scatter*; a scatter of 0 corresponds to a WON with equidistant stations, and higher scatter values imply that the stations are scattered more randomly over the plane.

Generating a distance matrix from a specified scatter is equivalent to generating the physical topology of the WON synthetically. Instead of going through the process of generating a network geography and designing an optimal physical topology, we go directly to the distance matrix, which represents a hypothetical physical topology.

We treat the traffic matrix similarly to the distance matrix: we fix the overall mean amount of traffic exchanged between pairs of stations, but vary the amount exchanged between specific pairs of stations according to a given probability distribution. The variability of traffic is referred to as *skew*, and a skew value of 0 corresponds to a uniform traffic matrix in which all stations exchange the same amount of traffic. The higher the value of skew, the less uniform the traffic pattern is.

We perform the experiments to be described in the sequel assuming a mean packet length of 1000 bits and 1-Gbps channel speeds. For the sake of tractability we assume that packet interarrival times and packet sizes are exponentially distributed; this assumption allows us to apply the formula for mean packet delay given in Equation (3.4).

5.4.1 An Example of Virtual-Topology Design in the DCWON

The experiments to be described in this subsection are aimed at determining the degree to which we can optimize the performance of the DCWON using the technique of simulated annealing. We are also concerned with comparing how a regularly structured interconnection graph, such as ShuffleNet, performs against an interconnection graph explicitly optimized for the given traffic matrix and station geography.

In executing the simulated annealing algorithm, several parameters must be specified, including the initial temperature, annealing schedule, etc. In these experiments we use the following parameter settings, which have been conservatively chosen to produce a very careful annealing that has good chance of locating the global optimum. We start the annealing at a temperature of 100 "degrees", and, since the cost function takes on values in the millisecond range, we normalize the cost function so that the uphill acceptance probability (i.e., $e^{-\Delta/T}$ in step 4 of Figure 5.2) starts off close to 1. In optimizing a DCWON of N stations, each epoch is continued until $8N$ moves have been examined or $5N$ moves have been accepted, whichever occurs first. The temperature is reduced by five percent from one epoch to the next. If there is no improvement in the cost function for five consecutive epochs, then the simulated annealing algorithm is terminated. Even after relaxing these parameter settings, we can still obtain good results from the simulated annealing algorithm. Rapid quenching, in which the temperature is reduced rapidly, can sometimes yield good results in much less time than careful annealing.

The cost function represents one of the problem-specific components of the simulated annealing algorithm. We equate cost with the mean packet delay of a specific network. The mean packet delay is computed for each virtual topology, traffic matrix, and distance matrix by means of Equation (3.4), which requires the calculation of the shortest-distance path from each source to each destination. The traffic matrix is chosen to provide a modest loading, so queueing delay is always negligible compared to propagation delay. During computation of the shortest-distance paths, we make no attempt to find alternate shortest-distance paths. However, the choice of alternate paths would not improve the cost by much, because the only savings that they could produce is in the area of queueing delay, which is already low. Furthermore, the branch-exchange operation used to generate new virtual topologies causes the routing procedure to examine new paths, because it gradually replaces old links with new ones. Therefore, old shortest-distance paths may no longer be valid in the new virtual topology, so that the algorithm has to find new ones.

For each value of N we allow five values for the skew ranging from 0 (uniform traffic) to 10 (highly asymmetric traffic). Specifically, the different values of skew correspond to the following methods for generating the traffic matrix:

- skew = 0: uniform traffic matrix
- skew = 1: traffic matrix entries chosen from a uniform probability distribution
- skew = 2: traffic matrix entries chosen from an exponential probability distribution

- skew = 3: traffic matrix entries chosen from a Bernoulli probability distribution with low variance

- skew = 10: traffic matrix entries chosen from a Bernoulli probability distribution with high variance

We also allow four values for the scatter parameter ranging from 0 (all stations equidistant from the headend) to 3 (two-thirds of the stations clustered near the headend). Specifically, the different values of scatter correspond to the following methods for generating the distance matrix:

- scatter = 0: uniform distance matrix

- scatter = 1: distance matrix entries chosen from a uniform probability distribution

- scatter = 2: distance matrix entries chosen from an exponential probability distribution

- scatter = 3: distance matrix entries chosen from a Bernoulli probability distribution with low variance

Throughout the set of experiments the traffic and distance matrices are randomly chosen in such a way as to keep the mean headend-to-station radius and the offered load fixed; the mean headend-to-station radius is held at 50 km, and the offered load ranges from about 50 Mbps for the eight-station DCWON to about 8 Gbps for the 64-station DCWON. The results are shown for DCWONs of 8, 24, and 64 stations in Tables 5.1, 5.2, and 5.3. This data shows that by selecting a new virtual topology for the WON we are able to improve its performance in all cases tested. The improvement over the ShuffleNet interconnection graph, which is used

Table 5.1: Comparison of Delays in the 8-Station DCWON.

			Delay (ms)		
N	skew	scatter	intial	best	gain
8	0	0	0.996	0.943	5 %
8	0	1	0.910	0.834	8 %
8	0	2	0.758	0.687	9 %
8	0	3	0.858	0.596	31 %
8	1	0	1.007	0.863	14 %
8	1	1	0.560	0.475	15 %
8	1	2	1.224	1.035	15 %
8	1	3	1.328	0.816	39 %
8	2	0	0.974	0.778	20 %
8	2	1	0.587	0.508	14 %
8	2	2	1.015	0.856	16 %
8	2	3	1.004	0.657	35 %
8	3	0	0.840	0.685	19 %
8	3	1	0.583	0.488	16 %
8	3	2	0.862	0.647	25 %
8	3	3	0.771	0.490	36 %
8	10	0	1.159	0.497	57 %
8	10	1	0.743	0.526	29 %
8	10	2	0.886	0.368	59 %
8	10	3	0.781	0.265	66 %

Table 5.2: Comparison of Delays in the 24-Station DCWON.

			Delay (ms)		
N	skew	scatter	intial	best	gain
24	0	0	1.630	1.552	5 %
24	0	1	1.471	1.235	16 %
24	0	2	1.344	1.013	25 %
24	0	3	1.179	0.571	52 %
24	1	0	1.614	1.484	8 %
24	1	1	1.532	1.318	14 %
24	1	2	1.054	0.807	23 %
24	1	3	0.558	0.436	22 %
24	2	0	1.665	1.396	16 %
24	2	1	1.576	1.223	22 %
24	2	2	1.067	0.778	27 %
24	2	3	0.576	0.443	23 %
24	3	0	1.718	1.279	26 %
24	3	1	1.625	1.126	31 %
24	3	2	1.086	0.745	31 %
24	3	3	0.589	0.446	24 %
24	10	0	1.741	0.845	51 %
24	10	1	1.602	0.837	48 %
24	10	2	1.112	0.404	47 %
24	10	3	0.584	0.407	30 %

Table 5.3: Comparison of Delays in the 64-Station DCWON.

			Delay (ms)		
N	skew	scatter	intial	best	gain
64	0	0	2.317	2.191	5 %
64	0	1	2.087	1.671	20 %
64	0	2	1.551	1.075	31 %
64	0	3	0.869	0.550	37 %
64	1	0	2.318	2.169	6 %
64	1	1	2.214	1.579	29 %
64	1	2	1.456	0.912	37 %
64	1	3	0.814	0.531	35 %
64	2	0	2.324	2.083	10 %
64	2	1	2.228	1.563	30 %
64	2	2	1.435	0.882	39 %
64	2	3	0.770	0.501	35 %
64	3	0	2.318	2.031	12 %
64	3	1	2.216	1.553	30 %
64	3	2	1.439	0.862	40 %
64	3	3	0.771	0.499	35 %
64	10	0	2.364	1.701	28 %
64	10	1	2.234	1.327	41 %
64	10	2	1.465	0.762	48 %
64	10	3	0.773	0.485	37 %

to initiate the optimization, ranges from as little as 5 percent to as much as 77 percent, depending on the specific problem parameters, and on average there is a 27-percent improvement in propagation delay. We can observe that the simulated annealing algorithm does not produce much of a gain in performance when skew and scatter are low. The small 5-percent improvement over ShuffleNet in the case of the DCWON with uniform traffic (skew = 0) and stations that are equidistant from the headend (scatter = 0) suggests that ShuffleNet performs quite well in such a case—this is, in fact, the case for which ShuffleNet was explicitly designed. As we increase the skew and scatter, however, ShuffleNet performance becomes less attractive, and the effectiveness of optimization, as evidenced in the dramatic improvement afforded by the simulated annealing algorithm, becomes significant.

The explanation for the dramatic improvement in propagation delay when the scatter is increased lies in the fact that a greater proportion of the WON's stations are situated close to the headend. The virtual topology chosen by the simulated annealing algorithm provides shortest-distance paths that are not necessarily minimum-hop paths: a packet traveling from its source to its destination will take shortcuts created by the presence of intermediate stations near the headend. The packet may use a number of the centrally located stations, but the overall distance (and therefore time) covered will be small. Thus, the WON effectively takes advantage of shortcuts created by stations near the headend, instead of using a smaller number of intermediate hops that could take the packet over a longer distance. In fact, this behavior is confirmed by examining the traffic flows produced during the calculation of the cost function in the simulated annealing algorithm, which reveals a heavier-than-normal loading of the shorter links between centrally located stations. This creates a kind of distributed switching center that consists of centrally located stations and their abbreviated hop distances. This form of *distributed cut-through* permits us to achieve better delay characteristics when a higher proportion of stations is situated close to the headend. It is even possible to add "dummy" stations near the headend for the purpose of promoting the incidence of distributed cut-through.

The next set of experiments examines the effects of optimization on both delay and throughput. Recalling that the ratio of carried load to offered load is equal to the mean number of hops in a network [Kle76], we hypothesize that by minimizing the weighted mean number of hops in the WON, we will also tend to increase the maximum throughput achievable by the WON. As in the previous set of experiments, we use the simulated annealing algorithm to minimize the number of hops traveled by a typical packet, without regard for the distance covered in a single hop. Such a procedure will result in decreased propagation delay, but the improvement is not as impressive as in the previous set of experiments, where minimizing the actual time delay was the prime objective. Thus, we do not vary the scatter parameter, since interstation distance would not affect the mean number of hops in the WON. The experiments are performed on WONs with 8, 24, 64, and 160 stations, and we examine five skew values ranging from 0 to 10, as in the previous experiments. To determine the maximum throughput that a particular network would support, we "pump up" the initial traffic load by a scaling factor and observe when the network reaches saturation, which is defined to be when the average buffer occupancy in some station exceeds a specified threshold—at this point the WON would be dropping too many packets. From each optimization run we are able to collect the following four items of information:

1. Mean packet delay for the initial ShuffleNet interconnection graph
2. Mean packet delay for the resulting optimized interconnection graph

Table 5.4: Joint Improvement of Delay and Throughput in the DCWON.

		Delay (ms)			Throughput (Gbps)		
N	skew	intial	best	gain	intial	best	gain
8	0	0.996	0.943	5 %	5.10	7.00	37 %
8	1	1.007	0.863	14 %	3.81	6.44	69 %
8	2	0.974	0.778	20 %	3.53	6.38	81 %
8	3	0.840	0.685	19 %	4.65	6.22	34 %
8	10	1.159	0.497	57 %	1.90	5.60	195 %
24	0	1.630	1.552	5 %	8.24	12.55	52 %
24	1	1.614	1.484	8 %	7.42	12.81	73 %
24	2	1.665	1.396	16 %	7.09	12.32	74 %
24	3	1.718	1.279	26 %	6.13	10.82	77 %
24	10	1.741	0.845	51 %	6.13	11.04	80 %
64	0	2.317	2.191	5 %	12.10	21.37	77 %
64	1	2.318	2.169	6 %	10.89	22.18	104 %
64	2	2.324	2.083	10 %	11.69	23.79	104 %
64	3	2.318	2.031	12 %	11.69	21.37	83 %
64	10	2.364	1.701	28 %	10.08	16.93	68 %
160	0	3.035	2.852	6 %	17.81	45.79	157 %
160	1	3.037	2.832	7 %	17.81	43.25	143 %
160	2	3.032	2.803	7 %	17.81	43.25	143 %
160	3	3.031	2.757	9 %	17.81	45.79	157 %
160	10	3.022	2.553	15 %	15.26	40.70	166 %

3. Maximum throughput for the initial ShuffleNet interconnection graph

4. Maximum throughput for the resulting optimized interconnection graph

The results of the experiment are shown in Table 5.4. Besides the improvement in delay that we saw in the previous set of experiments, we note that we can achieve an average improvement in throughput of 99 percent. We can also observe an interesting phenomenon in the data of Table 5.4: a small improvement in propagation delay yields a large improvement in the maximum sustainable throughput of the DCWON. This observation reinforces the need for optimization, since it not only gives better delay performance but makes for a WON that will carry heavier traffic loads.

It is interesting to note the different trends in Table 5.4. Most notably, we observe that the delay tends to increase as the skew increases in the initial (ShuffleNet) virtual topology. In the optimized virtual topologies, on the other hand, the delay decreases as the skew increases. Since higher skew implies greater variability in the traffic matrix's entries, the optimized virtual topology can achieve low delay by establishing a direct connection between pairs of stations with high traffic requirements. As an extreme example, a DCWON in which only two stations exchange traffic would achieve minimum delay by providing a direct connection between those two stations. We also observe that the trend in both ShuffleNet and optimized virtual topologies is for the maximum throughput to decrease as the skew increases. This tendency comes from the imbalance in traffic on WDM channels when skewed traffic is offered to the DCWON. Since the amount of traffic on a WDM channel is the sum of some subset of the traffic matrix's entries, greater variance in in each entry (each of which is chosen independently of the others) manifests itself as greater variance in the amount of traffic flow on each WDM channel. The amount of traffic flow on the maximally loaded channel will be highest when the variance in flow is greatest. Thus, saturation occurs earliest when traffic is unbalanced.

In Figures 5.6–5.10 we display the information of Table 5.4 more graphically. These five plots display for different skew values (i.e., 0–3, 10) the moderate-traffic delay versus the number of stations in the network. One curve is for the optimized DCWON and the other for ShuffleNet. The plots illustrate the extent to which

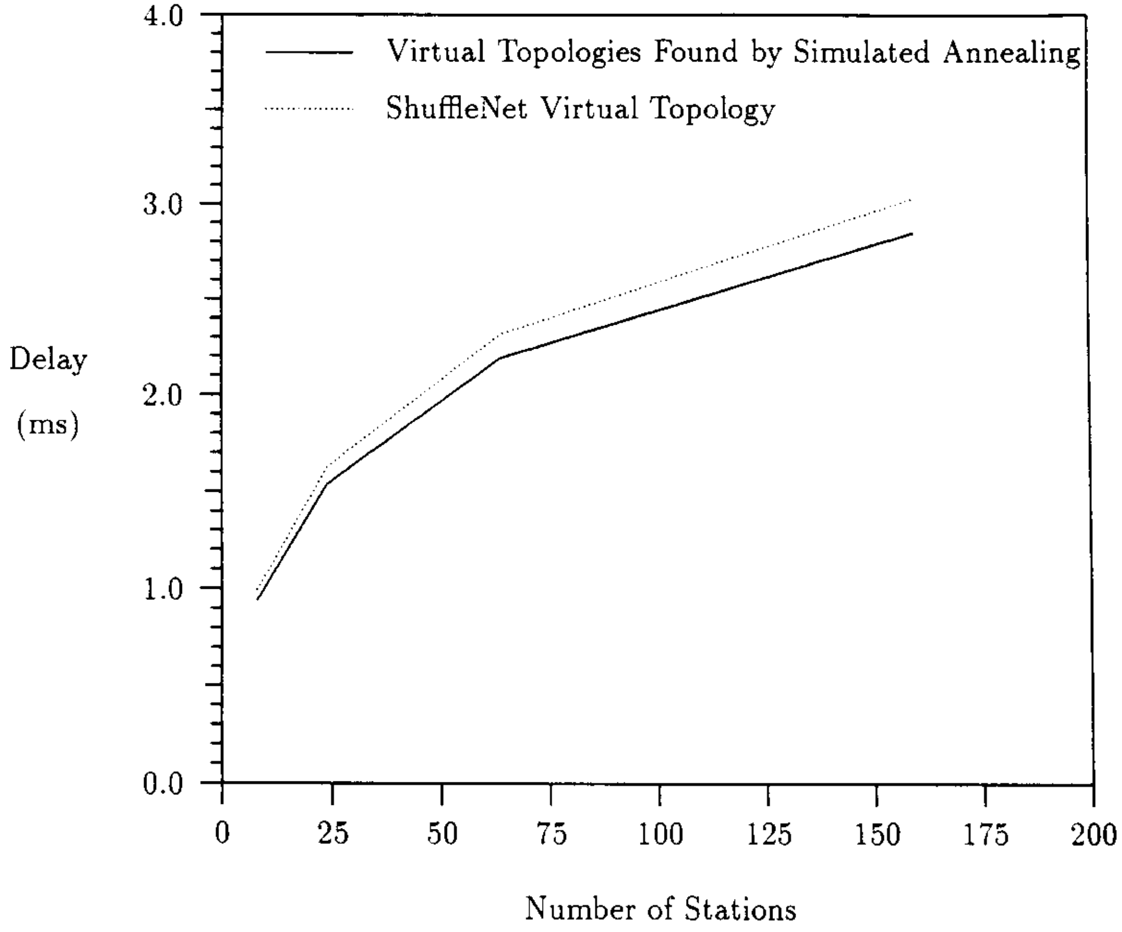


Figure 5.6: Comparison of Delay in Optimized and Unoptimized DCWONs with Uniform Distance and Traffic Matrices.

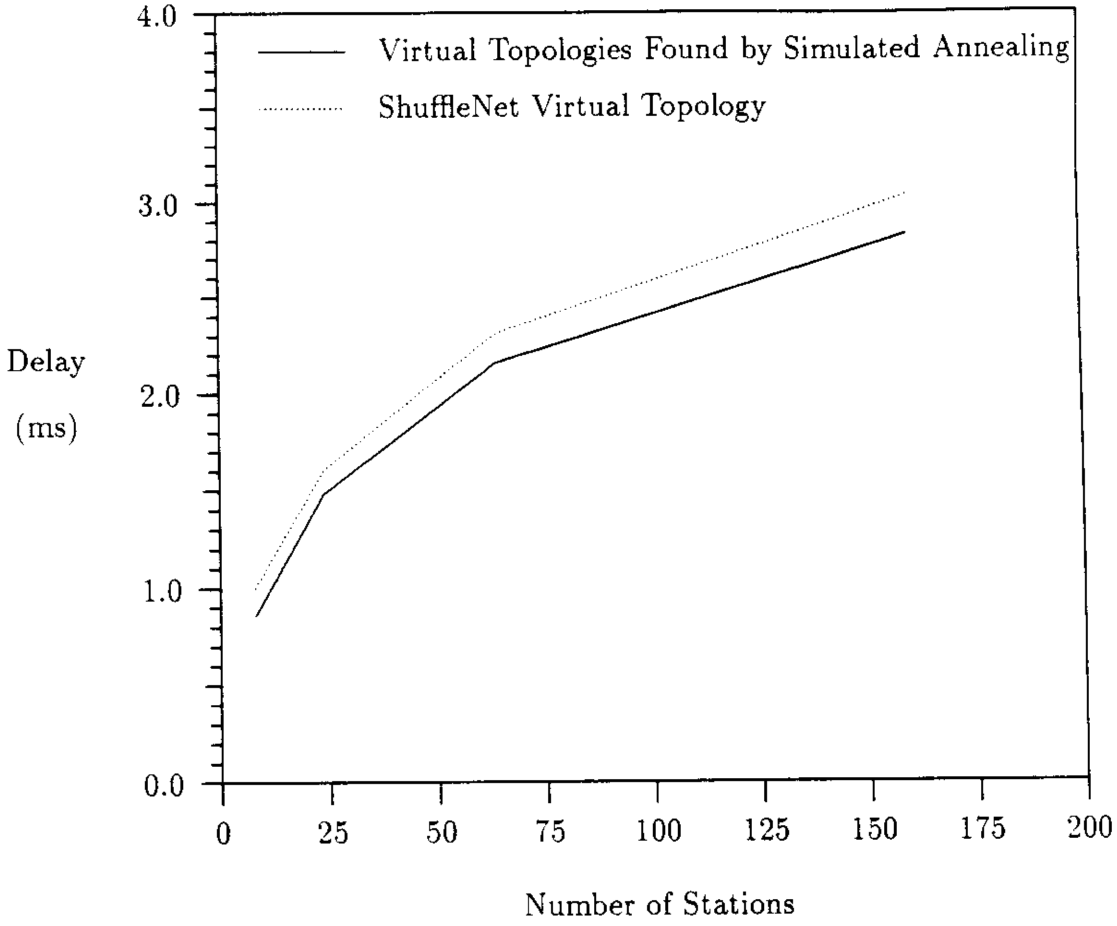


Figure 5.7: Comparison of Delay in Optimized and Unoptimized DCWONs with Uniform Distance Matrix (scatter = 0) and Nonuniform Traffic Matrix (skew = 1).

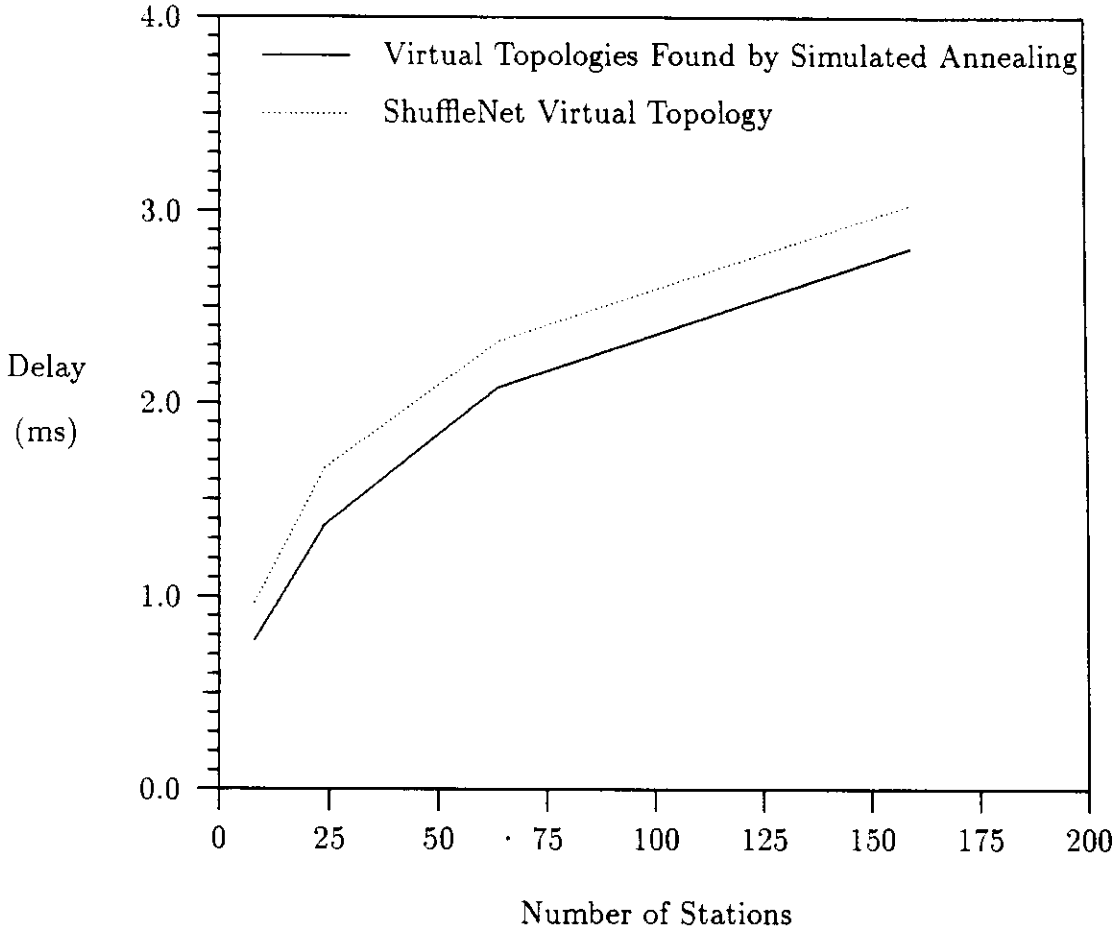


Figure 5.8: Comparison of Delay in Optimized and Unoptimized DCWONs with Uniform Distance Matrix (scatter = 0) and Nonuniform Traffic Matrix (skew = 2).

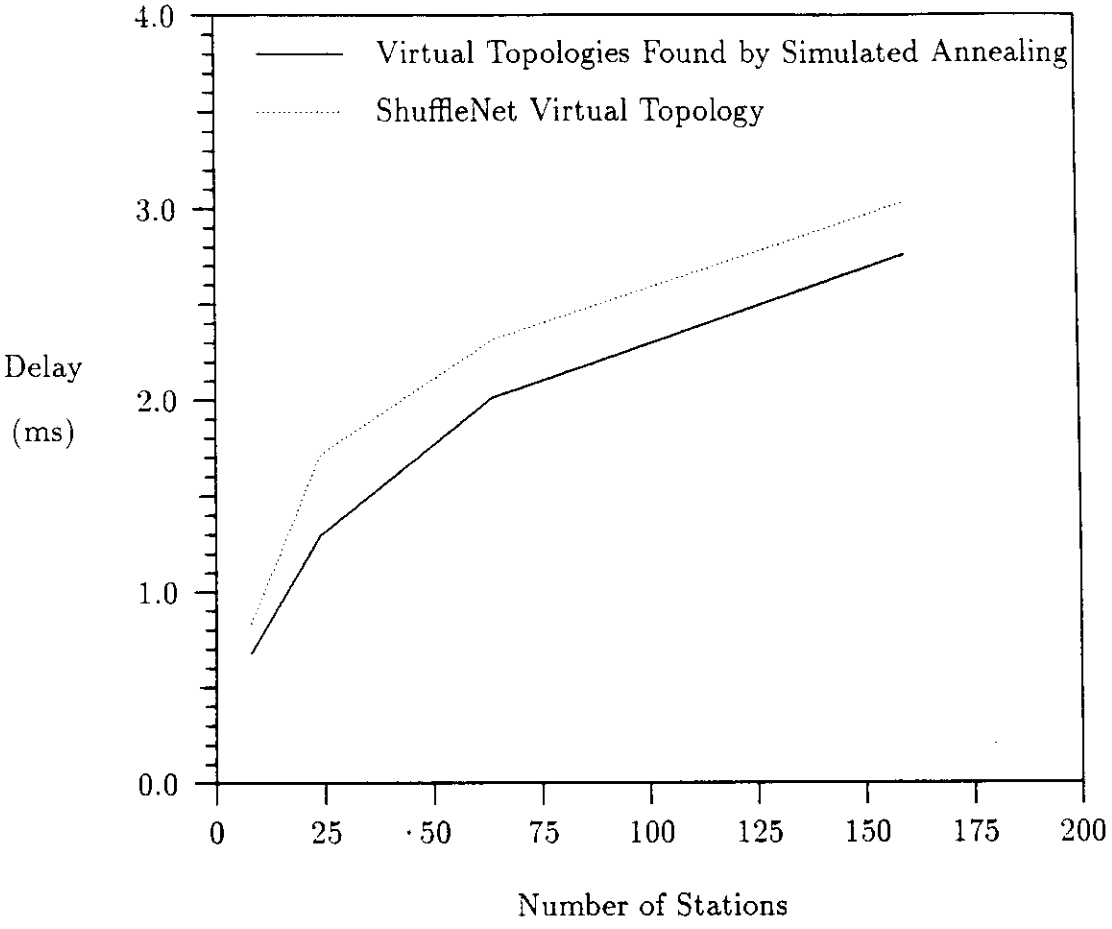


Figure 5.9: Comparison of Delay in Optimized and Unoptimized DCWONs with Uniform Distance Matrix (scatter = 0) and Nonuniform Traffic Matrix (skew = 3).

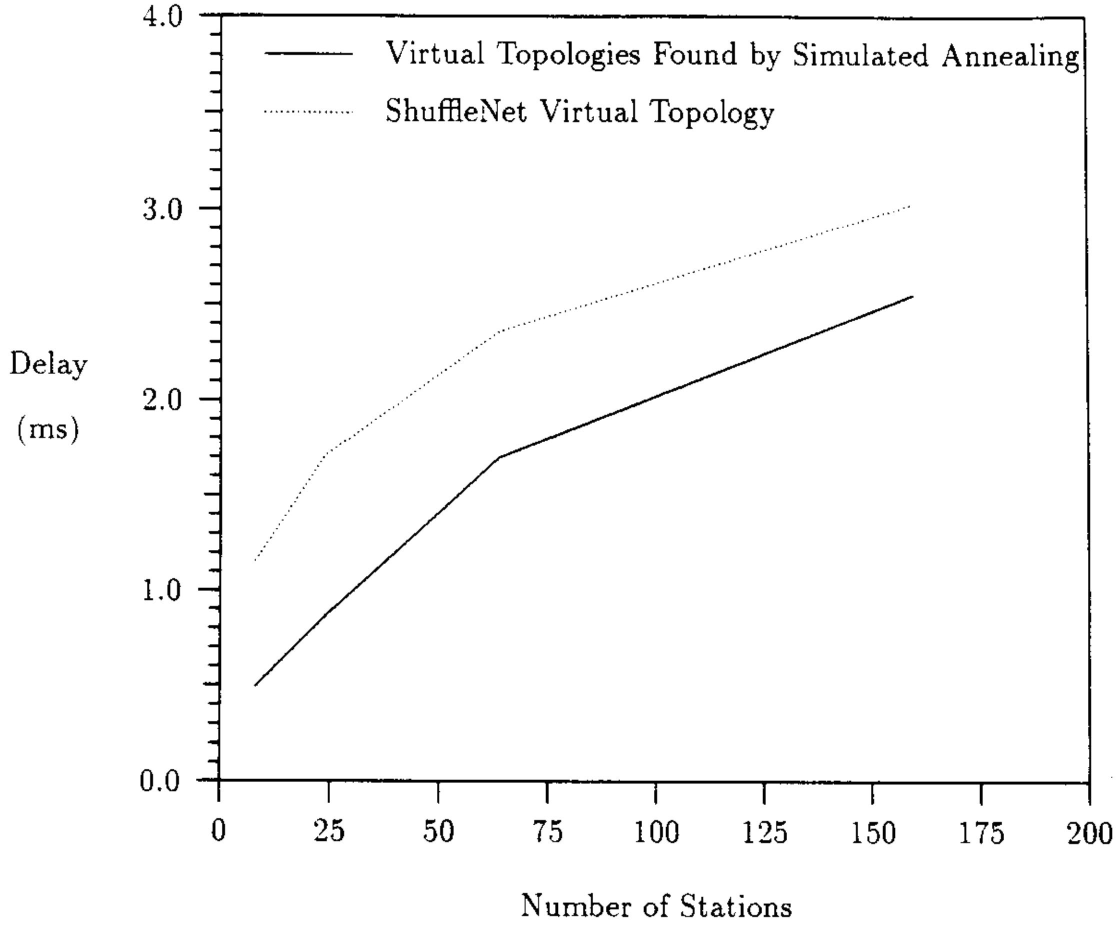


Figure 5.10: Comparison of Delay in Optimized and Unoptimized DCWONs with Uniform Distance Matrix (scatter = 0) and Nonuniform Traffic Matrix (skew = 10).

optimization can improve performance in the DCWON. The delay curve of the optimized DCWON lies well below that of ShuffleNet in every plot. In addition, we can see that the degree of improvement produced by optimization increases steadily as the traffic becomes more skewed.

If we study Tables 5.1–5.4 we observe that, for a fixed scatter, increasing the skew does not radically degrade performance in ShuffleNet. This finding corroborates the analysis of [EM88] which, using several mathematical models of traffic skew, concluded that the throughput of ShuffleNet does not degrade by more than about 50 percent when the assumption of uniform traffic is relaxed.

Finally, we show in Figure 5.11 a comparison of the mean packet delays in ShuffleNet and an optimized DCWON as the traffic load is scaled up. This plot more vividly illustrates the message of Table 5.4 by showing the mean packet delay as a function of traffic load. We see the usual domination of propagation delay until a critical load is reached, at which point the network saturates rapidly. The delay of the optimized DCWON is everywhere less than that of ShuffleNet, and its maximum throughput is nearly twice as great.

It is difficult to determine whether the solutions found by simulated annealing are optimal. The quality of solutions, however, appears to be high, in that we always see measurable improvements over and above starting solutions which ultimately appear to be good, and in cases where there is a known lower bound on the optimal solution, we come within a few percent of the bound. For example, in Figure 5.12 we compare the delays of DCWONs optimized by simulated annealing against a lower bound on the propagation delay for the DCWON with uniform distance and traffic matrices. The plot is for networks ranging in size from eight to 196 stations. This lower bound is computed from Equation (A.6) of Appendix A.

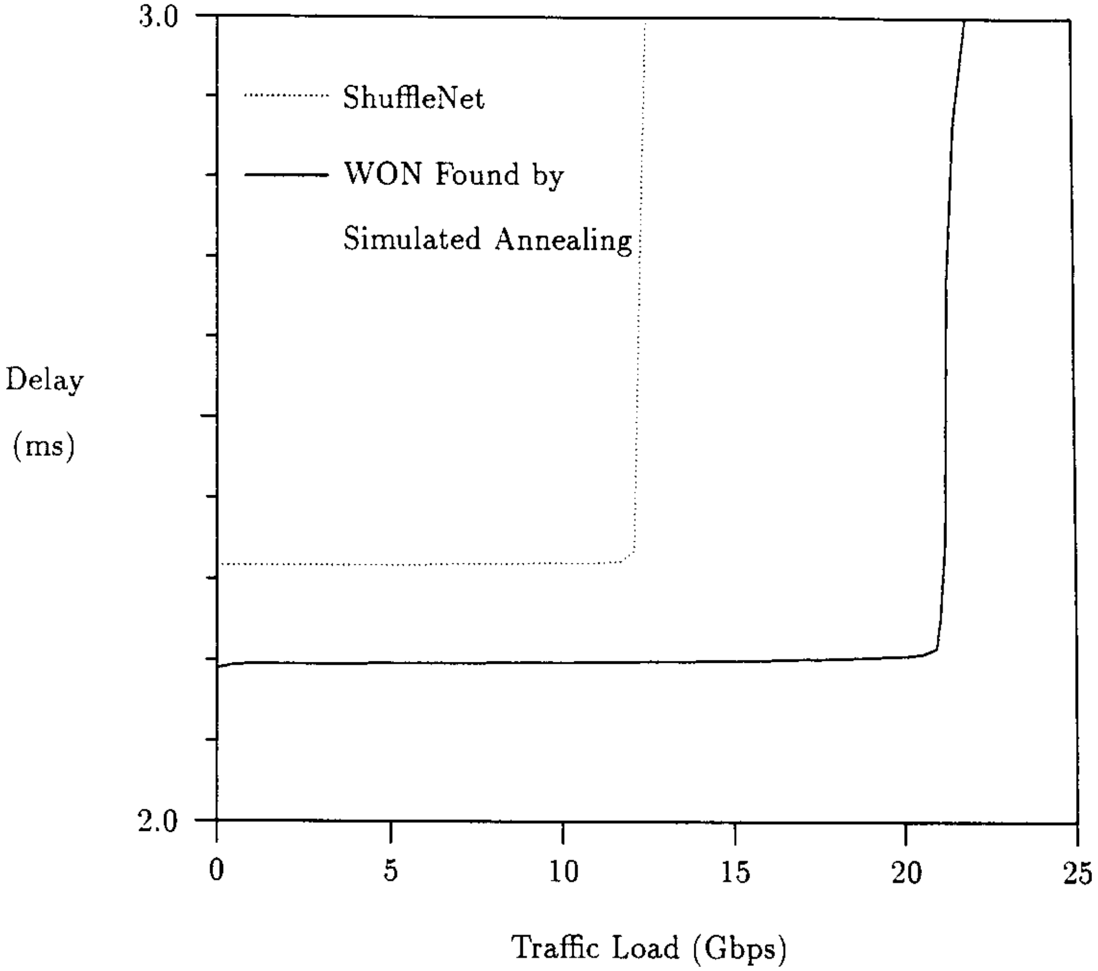


Figure 5.11: **Comparison of Delay** in ShuffleNet and an Optimized DCWON with Uniform **Distance and** Traffic Matrices.

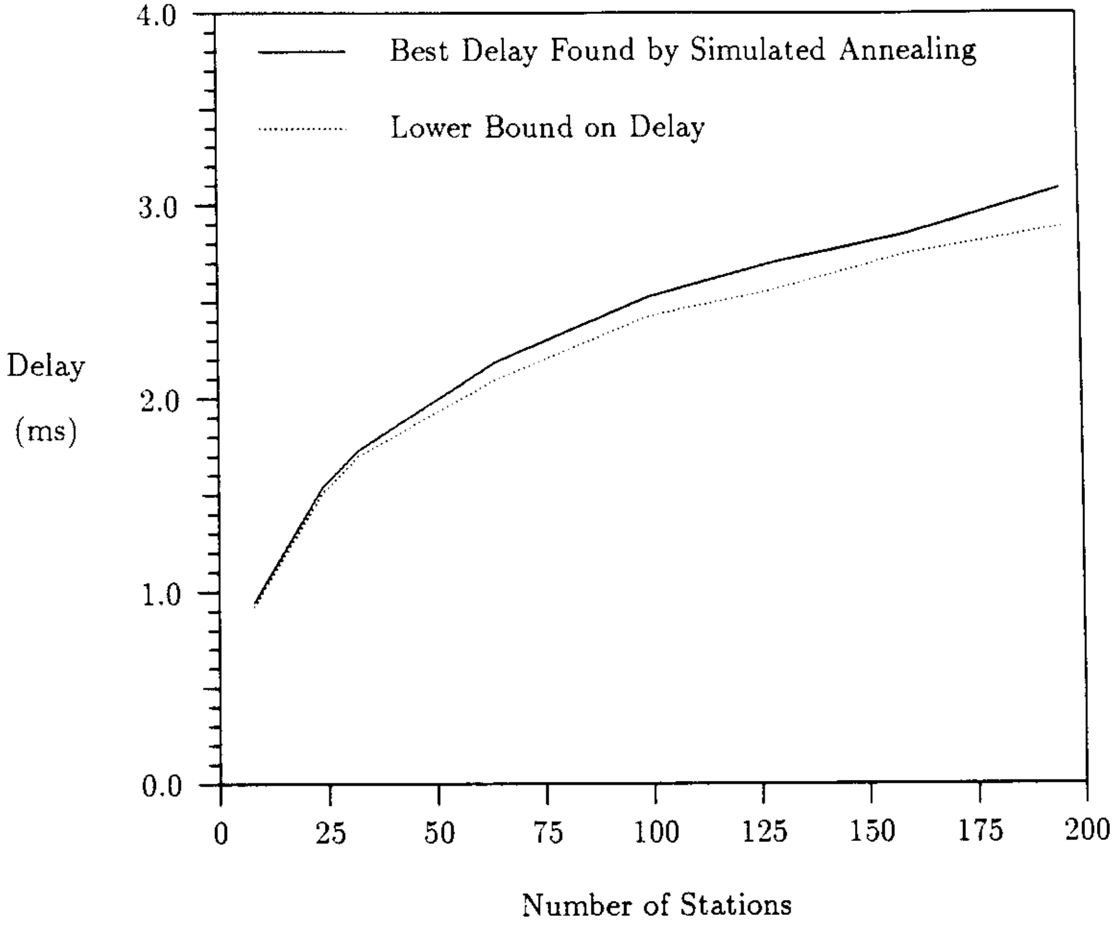


Figure 5.12: Delay of Optimized DCWONs Compared to a Theoretical Lower Bound (Uniform Distance and Traffic Matrices).

where it is formally derived, and it represents the best possible value to which we could ever aspire (although it is not necessarily realizable). It is encouraging to note that the optimized value never exceeds the lower bound by more than six percent.

5.4.2 An Example of Virtual-Topology Design in the SCWON

In Chapter 2 we alluded to the use of channel sharing for improving the performance of the WON. On a shared channel a given station can transmit to more stations than it could on a dedicated channel, and this increased fanout admits the possibility of defining denser interconnection graphs. As the extreme case, we point out that assigning *all* stations to *one* channel would allow any packet to get to its destination in one hop. However, there would be other sources of delay, including the time spent waiting to access the shared channel. These considerations motivate us to study how much improvement can be expected when channel sharing is used in the WON.

The shared-channel VTDP, which seeks to find the virtual topology that permits minimum packet delay in the SCWON, differs from the dedicated-channel VTDP in both its formulation and solution. As already discussed, we can not ignore the queueing component of packet delay in the SCWON unless the network is lightly loaded. Indeed, ignoring the queueing component would imply that the solution of the VTDP in the SCWON is best accomplished by collecting all stations on one channel. However, this solution is almost always unacceptable, because it can result in saturation of the channel or excessive access times for each station. We must therefore account for queueing delay by means of penalty functions or constraints in the formulation of the VTDP. For example, we could introduce the constraint that no WDM channel should host more than a fixed, small number of stations. We have chosen to use the penalty-function approach: we explicitly use channel-access delay as our penalty function.

The objective function we use is the formula for mean packet delay given in Equation (3.4). For this study we assume that shared WDM channels operate as pure Aloha channels [Abr70], which is not a very efficient access scheme but is fairly simple to implement. The principal drawback of Aloha is that work is wasted whenever packet collisions occur, and these can happen frequently if the channel is even moderately loaded. We would, therefore, prefer to operate the Aloha channel at a traffic loading that results in a fairly low rate of collisions. A rate of packet loss of more than, say, 3 percent could adversely affect packet delay by increasing the rate of retransmission. Using the classical formula $S = Ge^{-2G}$ [Abr70, Kle76], which relates channel throughput S to offered traffic load G in pure Aloha, we find that to keep retransmissions below 3 percent we must have $S > 0.97G$. Thus, we must have $Ge^{-2G} > 0.97G$, or $G < -(\ln 0.97)/2 \approx 0.02$. For our shared-channel model we therefore use a service-rate function that drops to 0 once a channel shared by two or more stations is more than 2 percent utilized; if the channel is less than 2 percent utilized, then the service rate is 1. Of course, when only one station is assigned to the channel, we use the simpler fixed-rate FCFS model, which, in principle, allows the channel to be perfectly utilized without any packet loss owing to collisions.

The genetic algorithm must be supplied with a number of parameters, including population size, mutation rate, etc. In these experiments we maintain a population of 50 SCWONs and choose parents of the next generation from the top 20 SCWONs by randomly selecting sixteen of them to mate and produce eight new offspring SCWONs. These eight offspring SCWONs are added to the population, which is then purged of the eight highest-cost SCWONs to bring the population size back down to 50. The purging can include the newly generated offspring SCWONs, if their cost is high enough to make them vulnerable. This process is repeated until we see no improvement in the lowest-cost SCWON for 50 consecutive generations. The stopping criterion reflects the judgement that further search would be unlikely to yield a better solution than has already been found. Mutations, consisting of the random reassignment of a transmitter or receiver to a channel, are introduced into about 20 percent of all offspring SCWONs.

To start the genetic algorithm we require an initial population of graphs with a fair amount of "genetic diversity". These graphs are mated and pass their characteristics on to their offspring so that characteristics with high survival value proliferate throughout the population. The initial population is constructed by taking a small set of regularly structured graphs, such as ShuffleNet, the MSN, the de Bruijn graph, the Moore graph, etc., and applying random mutation to them. This produces graphs with a good degree of variation and a fairly high overall survival value (survival value equates to low cost).

The running times of the genetic algorithm are considerably shorter than those of the simulated annealing algorithm for problems of equal size. The genetic algorithm also seems to find low-cost graphs easily and consistently, though the quality of the solutions is typically not as good as those found by simulated annealing.

We show the data from our experiments in Tables 5.5–5.7. The tables, given for SCWONs of size 24, 64, and 160, show the mean packet delays for a range of skew and scatter parameters. Skew and scatter values are chosen exactly as in

Table 5.5: Comparison of Delays in the 24-Station SCWON.

			Mean Packet Delay (ms)					
			Light Load			Moderate Load		
N	skew	scatter	initial	best	gain	initial	best	gain
24	0	0	1.630	0.501	69 %	1.618	1.576	3 %
24	0	1	1.471	0.538	63 %	1.535	1.456	5 %
24	0	2	1.344	0.452	66 %	1.019	1.002	2 %
24	0	3	1.179	0.453	62 %	0.632	0.514	19 %
24	1	0	1.614	0.522	68 %	1.611	1.559	3 %
24	1	1	1.532	0.554	64 %	1.549	1.429	8 %
24	1	2	1.054	0.516	51 %	1.143	0.900	21 %
24	1	3	0.558	0.500	11 %	0.979	0.567	42 %
24	2	0	1.665	0.501	68 %	1.618	1.510	7 %
24	2	1	1.576	0.543	64 %	1.537	1.352	12 %
24	2	2	1.067	0.539	50 %	1.209	0.895	26 %
24	2	3	0.576	0.528	9 %	1.016	0.652	36 %
24	3	0	1.718	0.515	70 %	1.640	1.452	11 %
24	3	1	1.625	0.532	67 %	1.534	1.170	24 %
24	3	2	1.086	0.561	48 %	1.239	0.839	32 %
24	3	3	0.589	0.556	6 %	1.025	0.613	40 %
24	10	0	1.741	0.521	69 %	1.504	1.169	22 %
24	10	1	1.602	0.583	64 %	1.601	1.157	28 %
24	10	2	1.112	0.432	61 %	1.083	0.675	38 %
24	10	3	0.584	0.427	27 %	0.878	0.474	46 %

Table 5.6: Comparison of Delays in the 64-Station SCWON.

			Mean Packet Delay (ms)					
			Light Load			Moderate Load		
N	skew	scatter	initial	best	gain	initial	best	gain
64	0	0	2.317	0.602	74 %	2.388	2.236	6 %
64	0	1	2.087	0.527	75 %	3.380	3.014	11 %
64	0	2	1.551	0.486	69 %	3.267	2.922	11 %
64	0	3	0.869	0.490	44 %	2.680	2.330	13 %
64	1	0	2.318	0.575	75 %	2.453	2.449	0 %
64	1	1	2.214	0.542	76 %	3.481	3.181	7 %
64	1	2	1.456	0.494	66 %	3.624	3.173	12 %
64	1	3	0.814	0.525	36 %	3.212	2.263	30 %
64	2	0	2.324	0.650	72 %	2.470	2.224	10 %
64	2	1	2.228	0.534	70 %	3.283	3.053	7 %
64	2	2	1.435	0.572	60 %	3.533	3.097	12 %
64	2	3	0.770	0.557	28 %	3.280	2.486	24 %
64	3	0	2.318	0.585	75 %	2.353	2.265	4 %
64	3	1	2.216	0.528	76 %	3.225	2.854	12 %
64	3	2	1.439	0.516	64 %	3.598	2.769	23 %
64	3	3	0.771	0.559	28 %	3.302	2.566	22 %
64	10	0	2.364	0.578	76 %	2.913	2.655	9 %
64	10	1	2.234	0.519	77 %	3.460	3.013	13 %
64	10	2	1.465	0.531	64 %	3.577	3.084	14 %
64	10	3	0.773	0.588	24 %	3.425	2.185	36 %

Table 5.7: Comparison of Delays in the 160-Station SCWON.

			Mean Packet Delay (ms)					
			Light Load			Moderate Load		
N	skew	scatter	initial	best	gain	initial	best	gain
160	0	0	2.904	1.533	47 %	2.905	2.873	1 %
160	0	1	2.628	0.745	72 %	2.630	2.261	14 %
160	0	2	2.065	0.884	57 %	2.112	1.999	5 %
160	0	3	1.001	0.510	49 %	41.254	0.676	98 %
160	1	0	2.906	1.047	64 %	2.907	2.904	0 %
160	1	1	2.330	0.683	71 %	2.331	2.188	6 %
160	1	2	2.640	0.673	75 %	2.686	2.535	6 %
160	1	3	1.612	0.742	54 %	2.666	1.810	32 %
160	2	0	2.900	1.717	41 %	2.901	2.865	1 %
160	2	1	2.328	1.955	16 %	2.329	2.288	2 %
160	2	2	2.633	0.659	75 %	2.683	2.631	2 %
160	2	3	1.616	0.612	62 %	15.643	1.923	88 %
160	3	0	2.985	1.466	51 %	2.896	2.853	1 %
160	3	1	2.325	0.806	65 %	2.327	2.308	1 %
160	3	2	2.640	1.477	44 %	2.687	2.510	7 %
160	3	3	1.618	0.617	62 %	15.582	1.777	89 %
160	10	0	2.888	1.909	34 %	2.889	2.877	0 %
160	10	1	2.333	0.860	63 %	2.335	2.212	5 %
160	10	2	2.631	0.957	64 %	2.684	2.222	17 %
160	10	3	1.617	0.627	61 %	28.194	2.404	91 %

the DCWON experiments of Section 5.4.1. In each table we show mean packet delays for the lightly loaded and the moderately loaded SCWON. In the light-load scenario there is very little traffic offered to the SCWON, and in the moderate-load scenario there is just enough traffic to produce queueing delays, but not so much that congestion is prevalent. Given each network size, we offer a fixed amount of traffic to the SCWON, regardless of the skew or scatter parameter; for example, we perform all the experiments with the moderately loaded 160-station SCWON assuming that approximately 7.6 Gbps are offered to the network. This means that sometimes a channel of the SCWON will be saturated, depending on the parameters of the experiment; for example, four of the entries in Table 5.7 show initial networks with comparatively high mean packet delays, which essentially means that the network is congested in those configurations. It can also be seen that we are able in all cases to eliminate the congestion by optimizing the virtual topology of the SCWON. We notice also that in all cases the congestion occurs when the scatter is at its peak value (viz., scatter = 3), which we explain by hypothesizing that the routing procedure is exploiting distributed cut-through to discover minimum-distance shortcuts provided by stations located near the headend, and this overutilizes a few specific channels in the SCWON.

We see in the data of Tables 5.5–5.7 the familiar drop-off in delay as scatter is increased. As in the DCWON, this is because the larger proportion of stations located near the headend can be used to provide shortcuts for traffic. Thus, we see that distributed cut-through can also be used to improve performance in the SCWON as well as in the DCWON. The genetic algorithm is able to improve upon the initial interconnection graphs—which either are or are directly derived from regularly structured graphs such as ShuffleNet and the de Bruijn digraph—but the results are different depending on the traffic load scenario. The average improvement in the lightly loaded SCWON is about 61 percent, compared with a 20-percent average improvement in the moderately loaded SCWON. It would appear easier to find the optimal virtual topology in the lightly loaded SCWON, perhaps because the goal of conglomerating as many communicating stations as possible on a single channel without overloading that channel is comparatively simple. It is interesting to note that in the moderately loaded SCWON channel sharing is only sparingly used: as traffic load increases, it becomes infeasible to share pure Aloha channels because of their sensitivity to load, and thus the optimization favors designs with a single transmitter per channel. Thus, the use of another multiaccess protocol, e.g., a protocol based on time-division multiplexing, would provide better performance at higher traffic loads. In fact, the experiments of the next section confirm that we can expect much better optimized performance when the SCWON uses fixed-assignment TDMA on its shared WDM channels.

5.5 The Impact of Physical-Topology Design on Virtual-Topology Design

As we have stated earlier, different physical topologies yield different distance matrices. We have seen in the previous section that the distance matrix exerts a strong influence on the propagation delay of the WON. Therefore, we expect to see some differences in WON performance when different physical topologies are used. We attempt in the next set of experiments to gauge the degree to which the choice of the physical topology affects the performance of the WON.

The experiments of the previous section used synthetically generated distance matrices, because this was the only way to make sure that the average "glass" distance between pairs of stations stayed constant from experiment to experiment. This was necessary to guarantee a fair comparison. Of course, this masks out the differences introduced by the physical topology. Also, the distance matrices did not correspond to actual physical topologies, since they were synthetically generated. Now, however, we wish to use distance matrices that result from actual physical topologies.

The following experiments postulate a network geography consisting of station clusters whose distances from the center of the plane are uniformly distributed with a mean of 50 km. We assume a WON with 64 stations, 1-Gbps channels, exponentially distributed packet interarrival times, and exponentially distributed packet sizes with a mean of 1000 bits. Shared WDM channels are assumed to use a fixed-assignment TDMA protocol that effectively allows stations to utilize up to 80 percent of the available channel capacity. The traffic matrices are generated according to the same procedure as in the experiments of the previous section, and the total offered traffic load was kept at a low-to-moderate level of 4 Gbps.

We optimize the physical topologies of the four topological classes (viz., the star, tree, two-clustered tree, and four-clustered tree). We then optimize the virtual topology for each of the four optimized physical topologies and the five skew values (i.e., 0–3, 10). The results for the DCWON are shown in Table 5.8. The optimized delay for the star physical topology, averaged over all five skew settings, is 1.351 ms, compared to 1.719 ms for the two-clustered tree (which has the highest overall delay). Thus, we can expect a decrease in delay of no more than about 21 percent when the star is used as the physical topology. The results for the SCWON, which are given in Table 5.9, are similar: the overall delay, after optimization of the WON's virtual topology, is 1.420 ms for the two-clustered tree, compared to 1.115

Table 5.8: Performance Comparison of Four DCWONs with Different Physical Topologies.

	Star		Tree		2-Tree		4-Tree	
Traffic Skew	Delay (ms)	Thruput (Gbps)	Delay (ms)	Thruput (Gbps)	Delay (ms)	Thruput (Gbps)	Delay (ms)	Thruput (Gbps)
0	1.435	24.7	1.798	23.6	1.830	20.1	1.759	25.8
1	1.425	24.3	1.774	23.4	1.816	19.5	1.733	25.1
2	1.372	24.1	1.737	22.7	1.771	19.8	1.684	25.8
3	1.357	26.3	1.674	23.6	1.726	21.3	1.669	28.0
10	1.168	19.2	1.458	19.2	1.452	16.8	1.467	22.4

Table 5.9: Performance Comparison of Four SCWONs with Different Physical Topologies.

	Star		Tree		2-Tree		4-Tree	
Traffic Skew	Delay (ms)	Thruput (Gbps)	Delay (ms)	Thruput (Gbps)	Delay (ms)	Thruput (Gbps)	Delay (ms)	Thruput (Gbps)
0	1.242	5.3	1.234	5.8	1.467	6.3	1.294	4.9
1	1.244	7.0	1.317	5.6	1.675	4.8	1.069	6.2
2	1.048	5.5	1.271	5.0	1.105	5.5	1.271	6.0
3	1.160	5.4	1.149	6.3	1.422	5.1	1.169	5.8
10	0.880	5.9	1.189	5.2	1.431	6.5	1.242	5.7

ms for the star, which represents a reduction of about 21 percent.

We also observe an overall decrease in mean packet delay of about 25 percent, when shared channels are used instead of dedicated channels. However, the maximum throughput for a given virtual topology is much less when channel sharing is used. Thus, if the traffic load were to increase in the SCWON, it might be necessary to reoptimize its virtual topology.

5.6 Summary

This chapter defines the dedicated- and shared-channel virtual-topology design problems. We derive lower bounds on the performance that can be achieved for a given traffic loading. We propose the simulated annealing algorithm and the genetic algorithm to solve the dedicated- and shared-channel virtual-topology design problems, respectively. We apply the algorithms to a large number of different problem instances and make the following observations:

- We can always find virtual topologies that surpass the performance (both in delay and throughput) of well-known WONs such as ShuffleNet.

- In network geographies with a fair proportion of stations located near the headend, the optimized virtual topology can achieve lower delays than an unoptimized virtual topology can achieve. The optimized virtual topology—taking advantage of the phenomenon of distributed cut-through—is able to use the centrally located stations as shortcuts that reduce propagation delay.

- There is somewhat of a difference between the delay that can be achieved by optimizing the virtual topology, given a star physical topology versus a tree physical topology, but the difference is about 20 percent, which is a small penalty compared to the cost savings that can be realized by using a tree rather than a star physical topology.

The optimization algorithms are shown to provide nearly optimal answers.

Chapter 6

The Routing and Congestion-Control Problem

Up to this point we have been concerned with problems that arise early in the design process, viz., physical- and virtual-topology design. We now turn our attention to problems that arise during the actual operation of the WON (but can often be anticipated early in the design process). We describe in this chapter a specific routing protocol for the WON and analyze its performance.

We note that this chapter is concerned only with dedicated-channel networks, unless otherwise indicated.

6.1 The Routing and Congestion-Control Problem

The Routing and Congestion-Control Problem (RCCP) is to find a routing procedure that delivers packets with a minimum delay and reduces the likelihood of congestion in the WON. The function of the routing protocol is to select the best path to the destination, and the function of the congestion-control protocol is to determine when to discard traffic—either at the entrance to or within the network—in order to relieve congestion [Ger81]. If we agree on a definition of what is meant by "best" (e.g., minimum delay) then the routing protocol can be seen as providing the best collection of paths between source–destination pairs. Occasionally, traffic demands exceed the capacity of the network, which necessitates the invocation of congestion-control procedures to reduce the traffic load to an acceptable level. Generally, the reduction of traffic is best applied at the entrance to the network, but sometimes it might be necessary to discard traffic after it has been admitted to the network.

Congestion occurs whenever the demand on a resource exceeds its capacity. These situations arise in basically two ways:

- fluctuations in demand
- fluctuations in capacity

While demand fluctuates because users make unpredictable requests for service, capacity fluctuates because resources fail. In either case we are concerned with increases in demand or reductions in service. An increase in demand is reflected in an abrupt change in the system's traffic requirements, e.g., the occurrence of traffic hotspots or sudden growth in the amount of traffic offered to the network. A reduction in service generally can be traced to the disappearance of a system resource, e.g., the failure of a switch or link. Although the problem of congestion after the depletion of resources is indeed an important one, we will consider only congestion caused by fluctuations in demand. This is because the depletion of resources in the WON, is, after all, an exceptional event, and—we believe—can almost always be handled by the reconfiguration of the virtual topology after the failure has been detected. At any rate, the study of strategies for rapid reconfiguration in the face of failures is a topic for future consideration (see Chapter 7).

If the routes taken by packets do not change with time, the routing algorithm is said to be *static* (or *quasistatic*, if the time between changes is long). If network conditions are not expected to change rapidly, then a static (or quasistatic) algorithm, in which the fixed set of routes is periodically recomputed, may be sufficient. If, on the other hand, network conditions are expected to behave more dynamically, then an *adaptive* routing procedure, in which the set of routes changes continually, might provide better performance. Static routing algorithms that use only one path between a source and destination are called *single-path*, while those that use several paths simultaneously are called *multipath.*

We have generally assumed that each station of the WON has ample packet buffers for a range of traffic loadings, e.g., the queueing-network model is based on unlimited waiting room for customers. In the infinite-buffer WON congestion control is not such a critical issue, since the consequences of congestion are usually limited to a temporary increase in delay. We note, though, that end-to-end flow control is still quite important because the source and destination could be speed-mismatched. However, the cost-effective station employs as few packet buffers as are necessary to achieve the desired level of performance. Therefore, there is a possibility that an arriving packet could find all buffers at its intended output port occupied. In this case congestion-control mechanisms should be invoked, i.e., the packet should be dropped or sent out through an alternate port. If congestion occurs frequently, then dropping packets is not a very palatable alternative.

There are strong interactions between routing and congestion-control protocols in the WON. There is also a strong need to keep the protocols relatively simple and streamlined in order to dovetail with the high-speed character of transmissions in the WON. In the WON mechanisms for routing and congestion control should ideally be implemented in hardware so that these functions are performed in real time at rates comparable to the link speed. The use of source routing, in which the source specifies in the header of each packet the complete list of station addresses via which the packet will travel to its destination, is attractive from the viewpoint of performance. Upon arrival at a station the packet's header is quickly analyzed and switched to the output port that goes to the packet's next hop. We note that each hop of the route can be efficiently coded by means of a few bits, because all that is necessary is to specify the output port at each station (at least when dedicated channels are used). This places the burden of route finding on the source station but obviates the need for a complex routing scheme in which the outgoing port is determined by lookup in a routing table. Implementation of table lookup in software would be extremely expensive in terms of performance. Although table lookup could be implemented in hardware, this might still not afford the low-latency switching that source routing does. Another, more important, reason for source routing, however, stems from its role in congestion control. When a packet arrives to a station and finds its preferred output port congested (which could mean that the number of packets vying for the port is above a predetermined threshold), then the packet can be given an alternate route by the station. One way to accomplish such alternate routing is for the station physically to insert a new sublist of station addresses into the packet header and place the packet on an alternate output port. If the alternate output port is less congested (i.e., fewer packets are queued there), then this procedure would reduce network congestion.

In this way routing and congestion control are naturally integrated in the WON.

In the scheme described above a packet is diverted away from the point of congestion by inserting an alternate route in the packet's header. It is not desirable to insert the entire rerouting information into the packet's header, as this would place an unreasonable processing burden on the station. The station would have to retrieve from storage or compute on the fly the alternate path to any other station. As we shall see in the next section, the station can reroute the packet to what would have been its next station via the alternate path. Since there is only one shortest path on which the packet can travel to get back to its former next station, it is easy to prestore and retrieve this alternate route in the station. Therefore, alternate-route selection can be done on the fly and with only a small amount of overhead to store the few alternate paths needed to get to the neighboring stations.

As the WON is intended to support the transport of both datagram (connectionless) and virtual-circuit (connection-oriented) traffic, it is necessary for the routing and congestion-control protocols to accommodate both. Recall that an important property of the virtual circuit is to preserve the order of transmission. If packets on the virtual circuit are adaptively rerouted, then there is no guarantee that they will arrive in order. Of course, we could rely on upper-layer protocols to sequence the packets properly, but this is not a preferred solution (nor is it, in most cases, feasible) for real-time virtual-circuit traffic such as streams of voice or video. It might, therefore, be desirable to award virtual-circuit packets higher priority than datagram packets, so that the high-priority packet is granted the primary route and the low-priority packet is sent on the alternate route in the case of contention for a congested output port. This solution, of course, works only if datagram traffic is a significant percentage of the total traffic volume, since

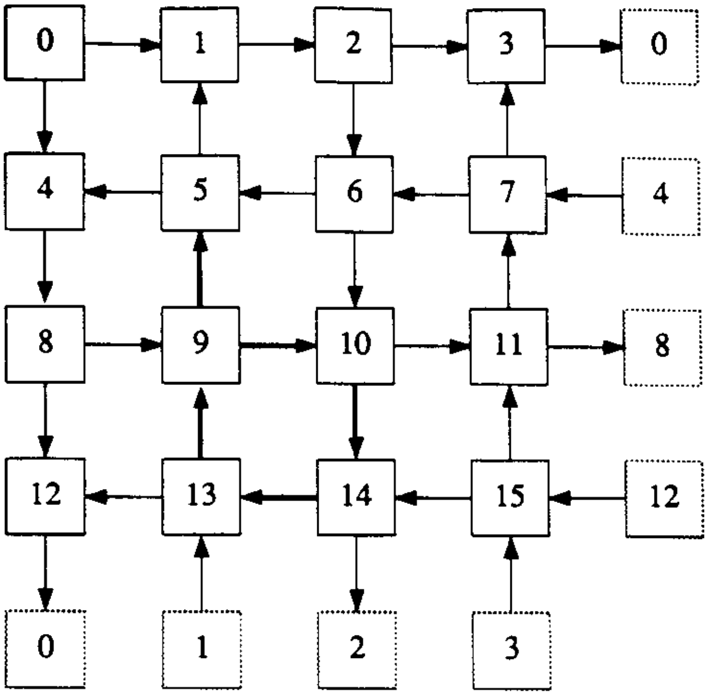


Figure 6.1: The Two-Dimensional Toroidal Virtual Topology.

frequent conflicts between datagram and high-priority packets could be resolved without misordering the high-priority packets. In addition, if high-priority packets conflict with each other, then one of them would have to be rerouted, which could result in a misordering.

The problem of routing and congestion control in specific instances of the WON has been studied by other researchers. The earliest work that is relevant to the RCCP in the WON is that of Maxemchuk, who focused on routing in the MSN [Max85, Max87, Max89]. The MSN, with its two-dimensional toroidal interconnect (shown in Figure 6.1), has the advantage that it can easily implement alternate-path routing by "deflecting" a packet so that it travels "around the block". Note that this never requires more than four additional hops; the original route can then be resumed after the deflection. This type of routing, which is called *deflection* *routing*, is attractive in the MSN because it makes it possible to use as few as one buffer per transmitter without suffering packet loss inside the network. Routing in ShuffleNet has also been examined. Acampora, Karol, and Hluchyj [AKH88] have demonstrated that there is a routing procedure that, when used in the shared-channel ShuffleNet, produces perfectly balanced loading on all WDM channels when traffic is uniform. Karol and Shaikh [KS88] have proposed a routing and congestion-control procedure that uses specific properties of ShuffleNet's recirculating perfect shuffle virtual topology. The procedure, which routes certain packets along nonoptimum routes when the preferred output port is busy, was shown to produce a more evenly balanced link loading and to eliminate bottlenecks caused by traffic hotspots when ordinary shortest-path routing is used.

Next, we examine detour routing, which is a simple generalization of deflection routing.

6.2 Detour and Deflection Routing

Deflection routing is used in the MSN to reduce buffering requirements. Under the assumption that time is slotted and that fixed-size packets[1] require one time slot for transmission, deflection routing has been proposed for dealing with stations that have as few as a single buffer per transmitter. When two packets simultaneously arrive at the two input ports of a station, there are two different cases that can arise: both packets are intended to exit via different ports or via the same port. The former case is free of conflict, but the latter case causes a problem if only one buffer is available at the output port, since one of the packets can not be

[1] Fixed-size packets in a time-slotted system are often referred to as "cells", but we will continue to use the term "packet".

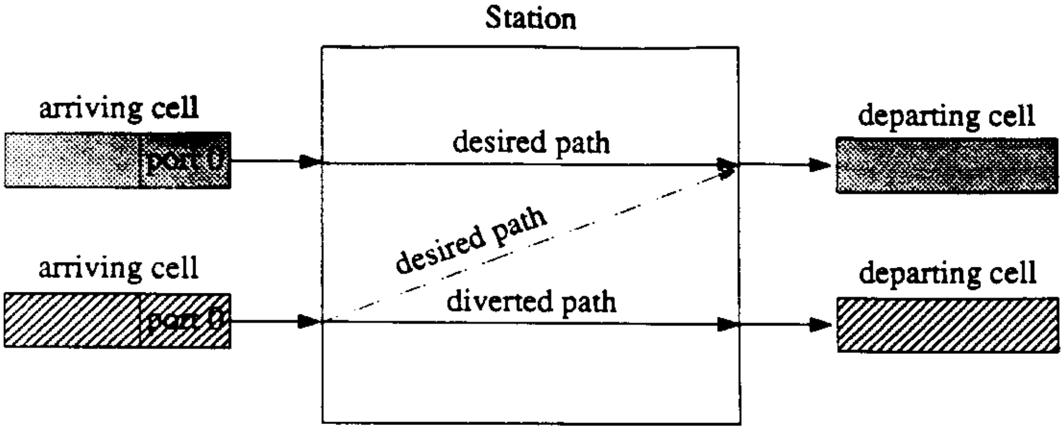


Figure 6.2: Bufferless Routing in the WON.

buffered and transmitted. The problem is resolved by deflection routing, which arbitrarily takes one of the packets and places it on the other output port, as shown in Figure 6.2. Because time is slotted and packets are transmitted during the slot, there will always be a free buffer available at the other output port. Because this routing protocol can guarantee—using only one transmission buffer per output port—the delivery of packets without any loss and without queueing, it is called *bufferless.*

Whenever a packet is diverted in the MSN, it merely takes the shortest path back to the current station, i.e., it takes a trip "around the block". For example, in Figure 6.1, a packet that originally was traveling from station 9 to station 5 would, if diverted, go along the path 10 → 14 → 13 → 9, and upon returning to station 9, it would reattempt the hop from 9 to 5. When source routing is used, deflection routing can be very easily implemented by keeping the precomputed loopback routes stored in the station, so that when a packet is diverted, the loopback route (e.g., a list of station addresses) can be automatically inserted into the part of the header containing the routing information: furthermore, this can be done by

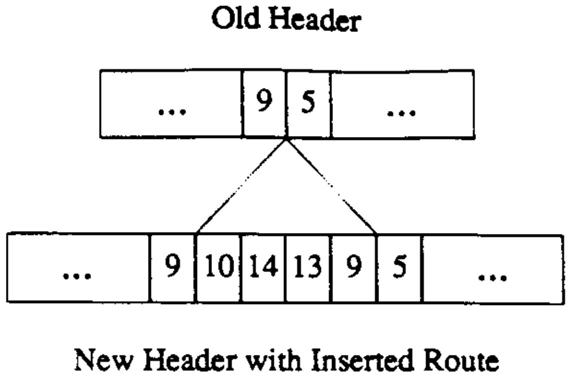


Figure 6.3: On-The-Fly Alternate Routing by Insertion of Routing Information.

hardware in real time. We illustrate the use of this technique in Figure 6.3.

Deflection routing can be used in the WON with any virtual topology. All that is required is that each station stores the loopback routes for each output port. We point out that deflection routing is appropriate only when all WDM channels are dedicated, since each transmitter must be able send out a queued packet during any arbitrary time slot in order to make room for a possible arrival in the next time slot. If a channel is shared by several transmitters with queued packets, only one transmitter can gain access to the channel during a time slot. Thus, the unsuccessful stations will still have their buffers occupied, and a packet arriving to any of them in the next time slot might have to be dropped for lack of buffer space. It is also clear that deflection routing applies equally well if there are more than one buffer per transmitter. In that case deflection will occur only when there are more packets in need of service than there are buffers available.

Deflection routing in the MSN deflects packets so that they return to the point at which they were deflected, but given a virtual topology different from the two-dimensional torus of the MSN, it might be possible to divert a packet to the

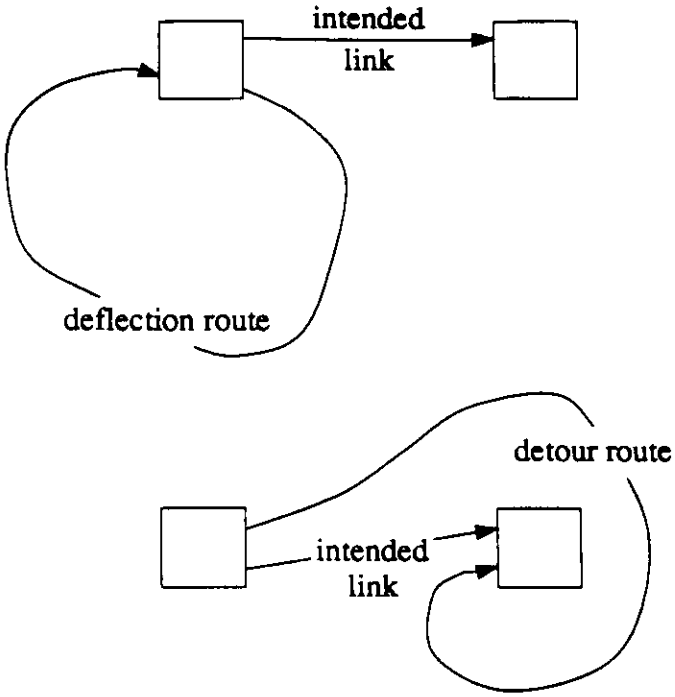


Figure 6.4: Deflection and Detour Routing.

intended next hop without returning to the point of deflection. By removing the restriction of returning to the point of deflection, we obtain a generalization of deflection routing that we call *detour routing.* We use the analogy of making an automobile detour around a congested segment of road—it makes more sense to take an alternate path to an intermediate goal rather than to loop back (around the block) to the current point. The distinction between deflection and detour routing is illustrated in Figure 6.4. A *detour* (with respect to a station's output port) is defined as the shortest path from the station's other port[2] to the station to which the original port was linked. Since a detour is a shortest path, it is never longer than a loop that returns to the original station and then proceeds to the intended

[2] Although we describe detour routing for WONs with two-transceiver stations, it generalizes easily to stations with more transceivers.

station. Intuitively, then, detour routing should perform better than deflection routing (which is, after all, just a special case of detour routing in which detours are not necessarily shortest paths).

It is also true that both detour and deflection routing need to prestore only a small number of detours. In a WON with the MSN virtual topology, for example, only three loopback paths are stored per station, i.e., the detours to be taken when one of the two output ports or the user output port is blocked. Likewise, in a two-transceiver WON exactly three detours need to be stored per station.

We point out that in both detour and deflection routing, deflections may be recursively nested, in the sense that a packet, having just been deflected, can be deflected once again from its detour. For instance, in Figure 6.1, within the detour shown in bold lines, there could be another detour, say at node 14, taking the packet on the loop 2 → 3 → 15 → 14. Eventually, however, all detours complete and the packet proceeds on its original route. There is, of course, a limit to how many detours a packet can take, since endless insertion of routing information into the packet's header could cause the packet to exceed its maximum length.

It should also be noted that detour routing and deflection routing require facilities for synchronizing the WON, so that time is divided into discrete slots, and this introduces additional complexity into the network.

A price we pay for detour (and deflection) routing is that packets may be discarded at the user input port before they are ever admitted into the network. Thus, the problem of admission control also plays an important role in congestion control. Several admission policies can be identified [BT89]. As a point in favor of detour (or deflection) routing, we note that it is, generally speaking, wiser to block packets at the entrance to the network than within the network.

6.3 Small-Girth Dense Digraphs

The toroidal interconnection of the MSN, shown in Figure 6.1, has the property that no deflection comprises more than four hops, i.e., its girth[3] is equal to four. Small-girth digraphs are attractive if we expect to invoke congestion-control mechanisms that divert traffic along alternate paths, be they based on deflections or detours. Our design criteria for the WON's virtual topology should thus encompass the need for small girth as well as small diameter. Girth and diameter are related, of course, since the girth of a digraph can never exceed its diameter by more than one. To keep the size of an alternate route small relative to the primary route's original size, i.e., to ensure that the length of an alternate route does not substantially exceed that of the primary route, we normally select the girth to be small in comparison to the diameter (or mean internode distance).

Simultaneously minimizing diameter *and* girth in a digraph creates a situation where these activities tend to militate against each other. For instance, known digraphs with small girth, e.g., the MSN virtual topology, have comparatively large diameter; and known digraphs with small diameter, e.g., the ShuffleNet virtual topology, have relatively large girth. The reason behind this diameter–girth tradeoff is that providing a short loopback path to a node means that an entire subtree of the node's reachability tree is eliminated from consideration. In Appendix E we prove an analog of the well-known Moore bound (see Chapter 5 or [TS79, FYdM84]) for digraphs with a constraint on the size of their girth. The bound on the number of nodes, N, in a digraph of degree p and diameter D when

[3] Recall that the girth of a digraph was defined in Chapter 2 to be the maximum over all nodes of the length of the shortest nontrivial path from each node back to itself.

the girth is constrained to be no more than C (for $C < D$) is given by

$$N(p,\ D,\ C) \leq \frac{p^D - 2p^{D-C} + 1}{p-1} + D - C \tag{6.1}$$

From this last equation it can be seen that constraining the digraph's girth sharply reduces the number of nodes possible in the digraph. For example, suppose we were to stipulate that no detour in a two-regular digraph should have length more than half that of the original route. One way to accomplish this would be to constrain the girth to be no more than half the diameter: $2C \leq D$. If we take D to be an even integer, $D = 2k$, then the bound on the number of nodes in a digraph with maximum girth, $C = k$, is given by Equation (6.1): $4^k - 2^{k+1} + k + 1$. Without a similar constraint on the digraph's girth, the Moore bound would be more than double this last quantity: $2^{2k+1} - 1$. Of course, these bounds are not necessarily achievable by actual digraphs, but they give an indication of how much we can expect to sacrifice by constraining the girth.

Given the unique structural properties of the MSN virtual topology and the MSN's favored position in the study of deflection routing, we can ask ourselves whether it is possible to construct better digraphs than the two-dimensional torus. Finding a digraph with minimum mean internode distance and girth no greater than a specified threshold is a combinatorial optimization problem with a key role in the design of a WON that uses detour routing. By using such an optimized digraph as the virtual topology for a DCWON, we also design a network that achieves optimum performance when detour routing is employed. We emphasize the difference between this aspect of design, which is a form of virtual topology design, and the VTDP studied in Chapter 5: the VTDP is directed at finding WONs with optimum performance when there are unlimited packet buffers, while the problem we are now considering deals with WONs in which the scarcity of packet buffers forces us to use detour routing.

We also point out that we are hampered by the lack of analytical results for expressing delay in networks that use detour routing. In actuality, we are not strictly concerned with diameter, mean internode distance, and girth, but rather with the performance that accompanies the use of a specific virtual topology in the WON. Although diameter, mean internode distance, and girth are rough gauges of a virtual topology's performance, we need to analyze explicitly the performance achievable with a given virtual topology in a WON using detour routing. We present an approach to estimating this performance in the next section.

6.4 Choosing the Right Virtual Topology ... Again

When we presented the VTDP in Chapter 5, our objective was to minimize the approximate expression for mean packet delay given in Equation (3.4). That formula was derived from a queueing-network model in which we assumed that there was a fixed number (which in the VTDP we always set to one) of routing chains between pairs of service centers. Detour routing could, in principle, use an infinite number of paths to get from a source to a destination. This is because a detour can occur at any point, even within another detour, and this could result in a recursive detour scenario that never terminates. Of course, such a scenario of boundless detouring would be highly unlikely. The upshot is that the VTDP, as stated, is not adequate for optimizing performance when detour routing is used. We therefore consider a modified formulation that can be applied to the optimization of virtual topologies when detour routing is used.

In the formulation of the routing problem, it will be convenient to present the problem for the special case when deflection routing is used. Therefore, we assume that our goal is to optimize the performance of a DCWON in which there is exactly one buffer per transmitter and deflection routing is used when transmission conflicts arise. The formulation can be generalized in a straightforward way to the case in which detour routing is used, but the notation is a bit more cumbersome.

We attempt to optimize the mean packet delay—or equivalently, the propagation delay, since each station has only a small number of packet buffers—of the WON by minimizing propagation delay subject to the constraint that the virtual topology has a weighted girth no greater than a specified value. The use of weighted girth, which we shall momentarily formalize, is prompted by the observation that the virtual topology should provide the shortest loopback paths at nodes that carry the most traffic and are thus most likely to deflect packets. A virtual topology in which the nodes processing the most traffic are not on the shortest loopback paths will not be very effective, because the detoured traffic must take longer detours, thus contributing to network congestion. Therefore, the object is not merely to constrain the length of *all* loopback paths to be less than a specified value, but rather to weight each loopback path according to its likelihood of being used as a detour.

We now present a model that can be used as a first-order approximation of performance when deflection routing is used in the WON with single-buffer stations. For the sake of simplicity, we assume that deflection (rather than detour) routing is used—the model for deflection routing is essentially the same as for detour routing, but easier to present. It is this model that is used for optimizing the virtual topology. A simple formulation involves assuming that a packet attempts to queue at a station's output port independently of other events occurring at the station. We define $\lambda_{jk}^{(i)}$ to be the fraction of time slots in which there is a packet at input port j of station i ready to be switched to output port k; restricting ourselves to two-transceiver stations, we see that $0 \leq j \leq 1$ and $0 \leq k \leq 2$, where output port number 2 is for the user. Now we can express the probability $\eta_k^{(i)}$ that two packets will simultaneously attempt to queue at output port k of station i, where $0 \leq k \leq 2$:

$$\eta_k^{(i)} = \lambda_{0k}^{(i)}\lambda_{1k}^{(i)} \tag{6.2}$$

Obviously, $\eta_k^{(i)}$ represents the fraction of traffic that must be diverted from output port k; if k equals 0 or 1, then the traffic will be diverted to output port 1 or 0, respectively, but if k equals 2 (i.e., the user port), then the traffic will be evenly distributed to output ports 0 and 1. Hence, the amount of traffic, $\Lambda_0^{(i)}$, diverted to output port 0 is given by

$$\Lambda_0^{(i)} = \eta_1^{(i)} + \frac{\eta_2^{(i)}}{2}$$

and, similarly, the amount of traffic, $\Lambda_1^{(i)}$, diverted to output port 1 is given by

$$\Lambda_1^{(i)} = \eta_0^{(i)} + \frac{\eta_2^{(i)}}{2}$$

Notice that the expression in Equation (6.2) is in terms of λ, and, hence, Λ is ultimately expressed in terms of λ. The calculation of λ is, therefore, necessary for the evaluation of the model. For each station the values of λ are essentially derived from the rate of traffic flow through the station. If we know the (primary) routes used by all source–destination pairs, then we can easily compute the rate of traffic flow from each input port to each output port of a station, viz., $\lambda_{jk}^{(i)}$.

Returning to the problem of finding the best virtual topology for a WON using deflection routing, we use the Λ as weights for the lengths of the loopback paths, and, instead of requiring a simple constraint on the virtual topology's girth, we require that the following expression not exceed a specified threshold:

$$\sum_{i=1}^{N}\sum_{k=0}^{1}\Lambda_k^{(i)}\delta_k^{(i)} \tag{6.3}$$

where $\delta_k^{(i)}$ is the length (in hops) of the shortest loopback path from output port k of station i back to itself. Equation (6.3), which is only an approximation, allows us to account for the fact that certain detours in a WON, by virtue of being more heavily used, are to be given greater weight than others.

To find optimal virtual topologies we again use the simulated annealing algorithm. The algorithm is applied in exactly the same way as in the VTDP, but the cost function is modified to reflect Equation (6.3). The cost function is propagation delay—or hop count, assuming a uniform distance matrix—with a penalty term that forces the function to infinity whenever the expression in Equation (6.3) exceeds a specified threshold. In this way the simulated annealing algorithm finds virtual topologies with minimum propagation delay—again, when there are no buffers, propagation delay is all that matters—and accounts for the degradation in performance that occurs when excessive detouring prevails.

6.5 Empirical Results

This section describes experiments conducted for the purpose of evaluating the comparative performance of detour and deflection routing with various virtual topologies. In these experiments we compare the performance of WONs with optimized virtual topologies against that of the MSN. In the optimized WON we assume that detour routing is used, and in the MSN we assume that deflection routing—deflection and detour routing are essentially equivalent in the MSN, since

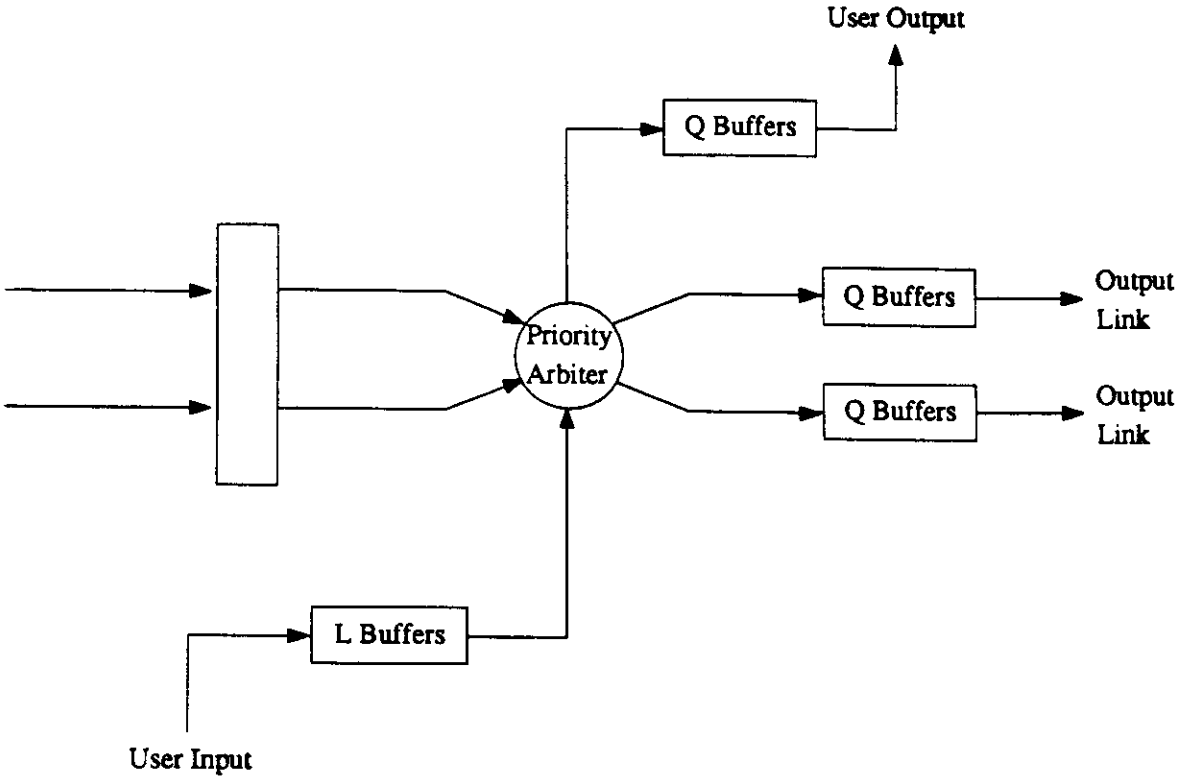


Figure 6.5: Station Model Used in the Simulations.

the shortest detour must always return to the detouring station—is used in such a way that the shortest path between any pair of stations makes a minimum number of right-angle turns.

We use a discrete-event simulation of the WON (and MSN) to evaluate its performance accurately. Events in the WON simulator are synchronized on time slots whose duration is adequate to transmit one fixed-size packet. The structural attributes of the simulated stations are shown in Figure 6.5. All simulations in the following experiments use two-port stations, the number of transmission buffers Q is either one or two, and the number of user input buffers L is fixed at eight. Since the user output buffer is essentially a transmission buffer, it too is set equal to Q. We run each simulation for 5000–8000 time slots, during which time we collect statistics such as the mean transport time that a packet spends in the network (starting with its admission into a transmission buffer), the end-to-end packet delay (starting from the packet's birth), the percentage of packets that are blocked because there are no user buffers, and the mean number of detours per packet.

We assume that the distance matrix is uniform, i.e., all stations are the same "glass" distance from each other. This distance is fixed at 50 km.

The WON is optimized by simulated annealing using a traffic matrix that exerts a reasonably heavy traffic load on the network. To assess the performance of the WON accurately, we run several simulations corresponding to different traffic loads and collect statistics from which delay and throughput figures are derived. The MSN's toroidal interconnection is used as the initial virtual topology in the optimizations. As described earlier, the cost function includes a penalty function that forces the cost to infinity whenever the weighted detour length typified by Equation (6.3) exceeds the threshold of 2.7 ms (which equates to a little more than five hops).

The first set of experiments is for the 64-station WON and measures the mean transport delay for the optimized WON and the MSN. We show in Figures 6.6 and 6.7 a comparison of mean transport delay for the 64-station optimized WON and MSN, with traffic skews of 0 and 1, respectively. Recall from Chapter 5 that a traffic skew of 0 corresponds to a uniform traffic matrix and a traffic skew of 1 corresponds to a traffic matrix whose entries are chosen from a uniform probability distribution. We also show in these graphs two separate curves for the MSN; one curve corresponds to the transport delay when the shortest-path routes

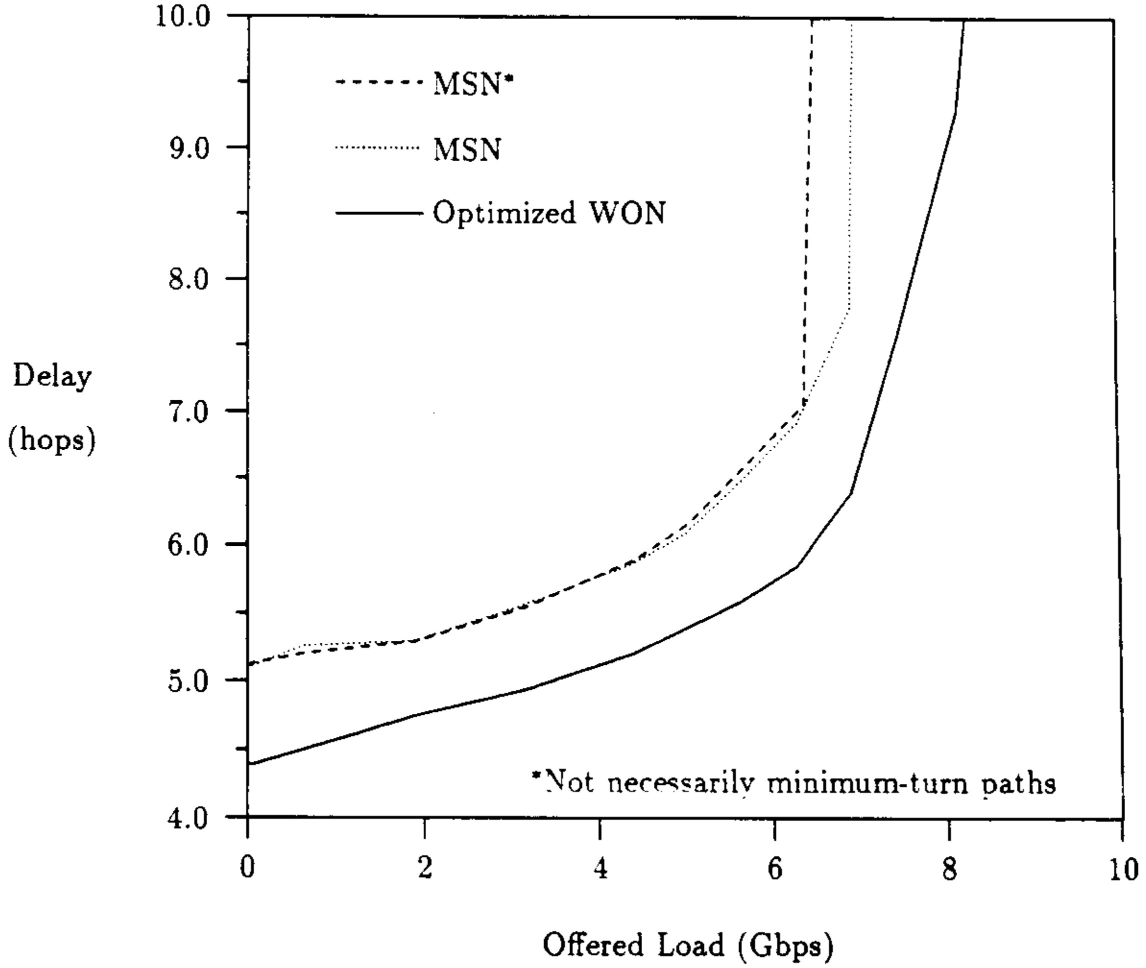


Figure 6.6: A **Comparison** of Delays in the 64-Station Optimized WON Using Detour Routing **and the** MSN Using Deflection Routing with Uniform Traffic (skew = 0) and **Single-Buffer** Stations ($Q = 1$).

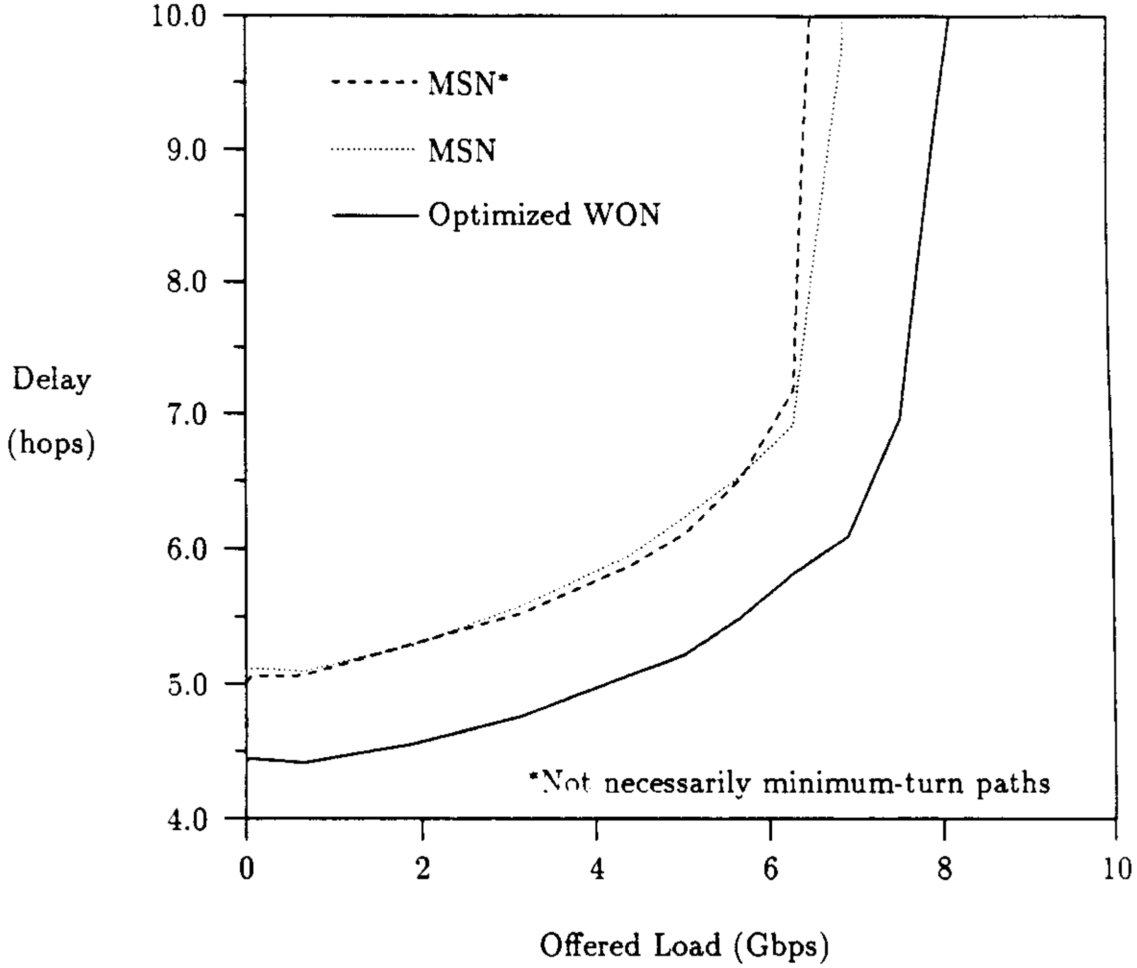


Figure 6.7: A Comparison of Delays in the 64-Station Optimized WON Using Detour Routing and the MSN Using Deflection Routing with Nonuniform Traffic (skew = 1) and Single-Buffer Stations ($Q = 1$).

take a minimum number of right-angle turns, and the other curve corresponds to transport delay when shortest-path routes are computed without any concern for minimizing right-angle turns. As predicted by other researchers [Max89], the use of minimum-turn routes allows us to achieve higher maximum throughput, since the probability of conflicting transmission-buffer access is reduced. For the rest of the experiments, only minimum-turn routes are used in the MSN.

The frequency with which packets are deflected from their primary routes decreases if we allocate more than one buffer per transmitter, and the next experiments attempt to quantify the improvement by examining WONs that use two packet buffers per transmitter. Figures 6.8 and 6.9 show comparisons of transport delay in the 64-station optimized WON and MSN for skew values of 0 and 1, respectively, when stations have two buffers per transmitter. The optimized WON used here is identical to that used in the previously described experiments with single-buffer WONs. In Figures 6.6–6.9 we see that it is possible to optimize the WON so that its transport delay and maximum throughput are consistently superior to the MSN, the latter being arguably a very good competitor, because of its special structural characteristics, i.e., its small girth. We also see that the maximum throughput can be increased substantially by adding just one more packet buffer per transmitter.

It is interesting to compare the performance of the limited-buffer WON with that of the unlimited-buffer WON. Referring back to Table 5.4 in Chapter 5, we see that the 64-station optimized DCWON could achieve a maximum throughput of approximately 21 Gbps. With only two buffers per transmitter, however, the WON can attain a throughput of nearly 16 Gbps before it saturates. Thus, with only two buffers per transmitter, we can realize about 75 percent of the throughput of the

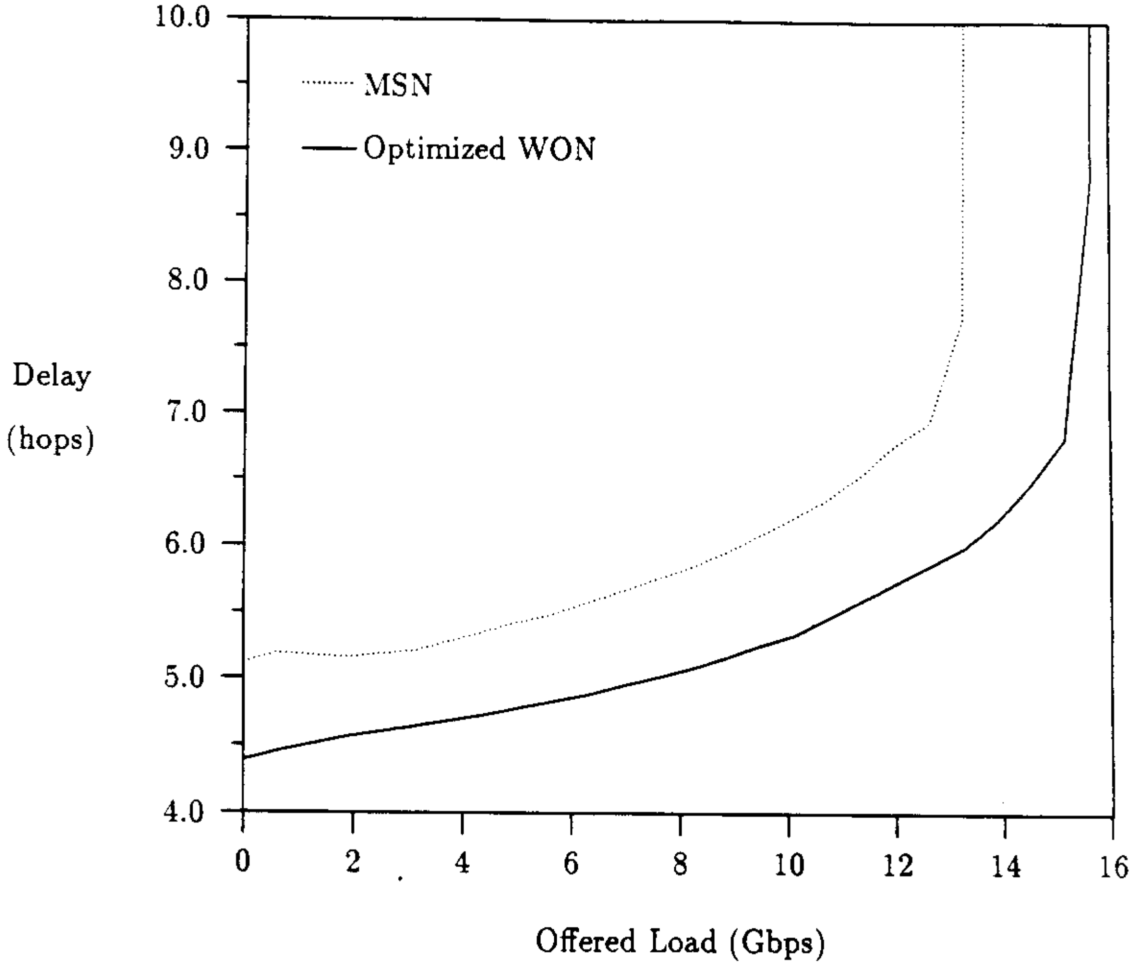


Figure 6.8: A Comparison of Delays in the 64-Station Optimized WON Using Detour Routing and the MSN Using Deflection Routing with Uniform Traffic (skew = 0) and Double-Buffer Stations ($Q = 2$).

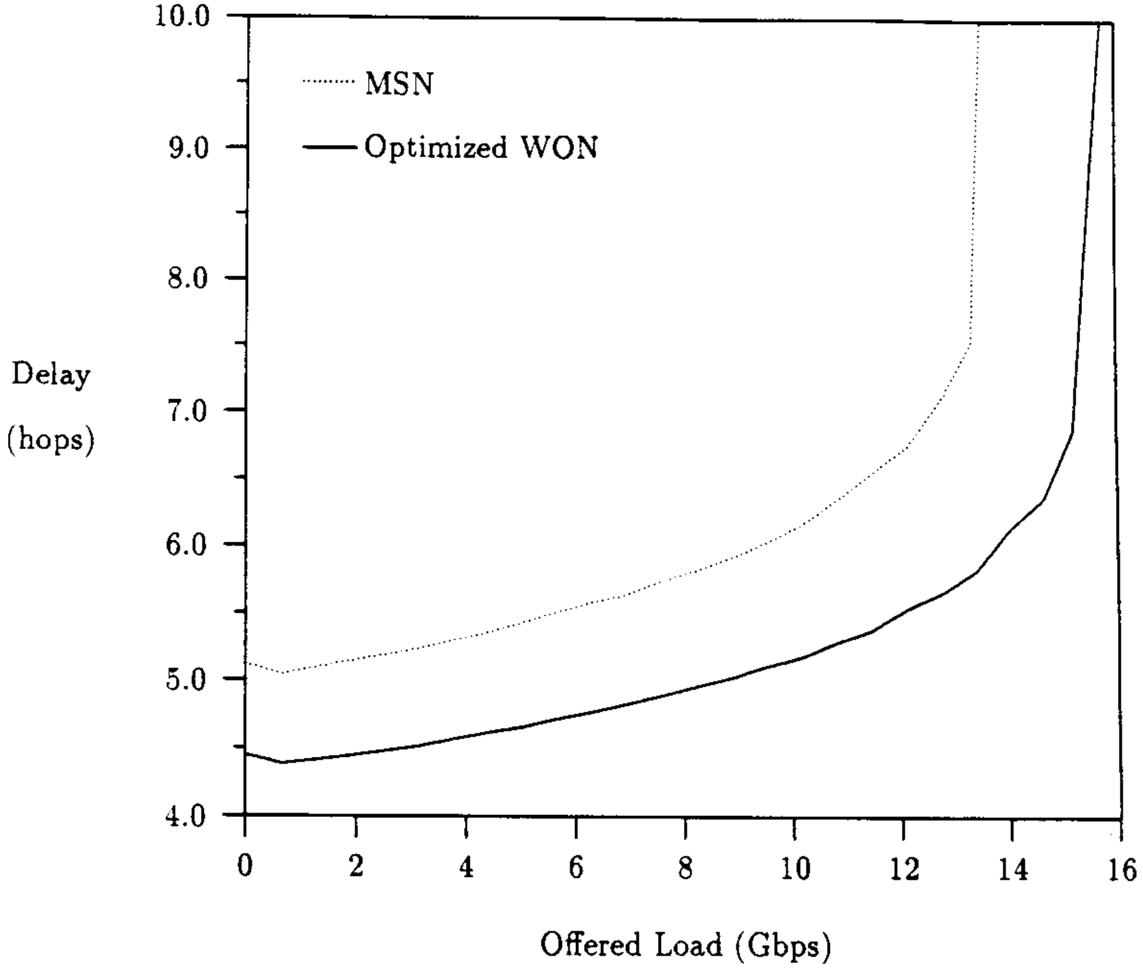


Figure 6.9: A Comparison of Delays in the 64-Station Optimized WON Using Detour Routing and the MSN Using Deflection Routing with Nonuniform Traffic (skew = 1) and Double-Buffer Stations ($Q = 2$).

unlimited-buffer WON. Clearly, the addition of packet buffers can reach a point where the performance improvement is only marginal. Comparing the curves of Figures 6.6–6.9 with the curve of Figure 5.11 in Chapter 5, we observe that whereas propagation delay is flat when single-path routes are used, propagation delay increases with traffic loading when detour or deflection routing is permitted. As traffic loading increases, the frequency of transmission-buffer contention increases, and this results in more packets taking detours. The net effect is for the typical packet to take more hops to get to its destination. The increase in hop count corresponds to a greater propagation delay. Since the route taken by a packet from a given source to a given destination never varies when stations have unlimited buffers, its propagation delay remains constant. Thus, the delay curves obtained when bufferless routing is used resemble delay in an M/M/1 queueing system (e.g., Figures 6.6–6.9), and the delay curves obtained when fixed-path routing is used resemble delay in a D/D/1 queueing system.

It is desirable to compare the performance of the optimized WON against the MSN for larger networks. The second set of experiments is for the 196-station WON and measures the mean transport delay for the optimized WON and the MSN. In Figures 6.10 and 6.11 we have plots of transport delay for the 196-station optimized WON and MSN for skew values of 0 and 1, respectively. Each station has one buffer per transmitter. We see that the transport delay and maximum throughput of the MSN can be consistently improved upon by the optimized WON.

6.6 Summary

This chapter addresses the problems of routing and flow control in WONs with limited packet buffers. We propose an efficient routing procedure, called detour

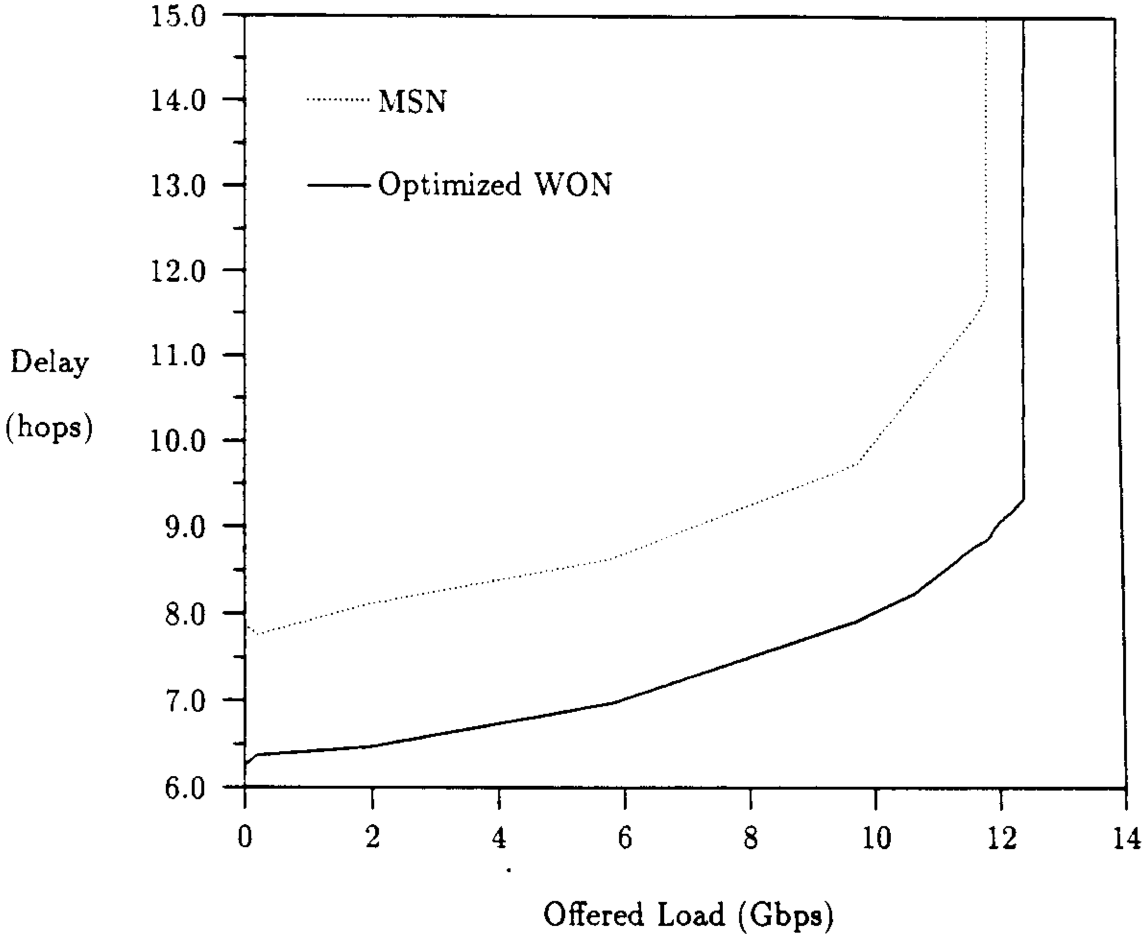


Figure 6.10: A Comparison of Delays in the 196-Station Optimized WON Using Detour Routing and the MSN Using Deflection Routing with Uniform Traffic (skew = 0) and Single-Buffer Stations ($Q = 1$).

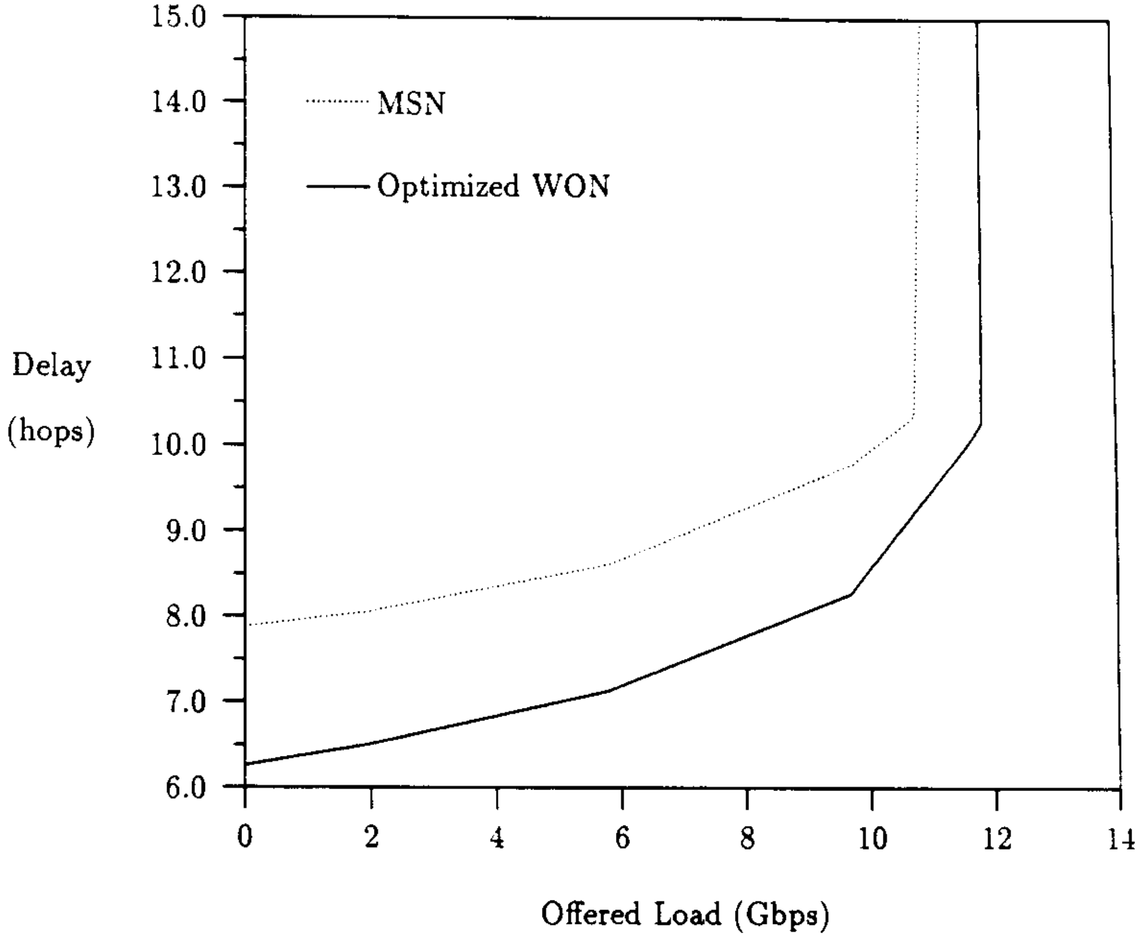


Figure 6.11: A Comparison of Delays in the 196-Station Optimized WON Using Detour **Routing and** the MSN Using Deflection Routing with Nonuniform Traffic (skew = 1) **and** Single-Buffer Stations ($Q = 1$).

routing, that prevents packet loss owing to buffer overflow by routing blocked packets along alternate paths. The virtual topology of the WON can be optimized with respect to the detour routing protocol to deliver superior performance. We show by simulation that detour routing in a properly optimized virtual topology achieves better performance than schemes such as deflection routing in the Manhattan Street Network.

Chapter 7

Conclusions and Recommendations for Future Research

7.1 Assessment of the Contributions of this Research

Although the objective assessment of this research is probably best left to the reader, we now point out what we feel are the most important contributions represented herein. In so doing, we wish to emphasize the novelty of our ideas and the fact that our methods are directed at producing tools that *work*.

The first major contribution of our work is the introduction of a new network architecture which is based on the concepts originally pioneered by Acampora [Aca87] and which holds great promise in meeting future networking requirements. The WON extends these concepts by fully developing the notion of wavelength agility, which can be harnessed to provide a malleable *virtual topology* independent of the chosen ***physical topology***. The virtual topology of the WON is entirely specified by the tuning of its wavelength-agile transceivers and vests the network administrator with considerable capabilities to manage the network and tune its performance.

The second major contribution of our work is in the area of WON design. We pose a number of uncontrived, realistic design problems that would have to be solved during the implementation of an actual WON. We identify the design of the physical topology as a principal cost driver in the WON and formulate the physical-topology design problem as a cost-minimization problem. We propose an algorithm that produces a minimum-cost physical topology with respect to a given network geography and topological class. Given an established physical topology for the WON, the choice of virtual topology exerts a profound influence on the performance (in terms of both delay and throughput) achieved by the WON. We formulate the virtual-topology design problem as a performance-optimization problem with respect to the network's traffic requirements. Depending on the type of WON being designed, we use either the simulated annealing or the genetic algorithm to optimize the virtual topology. The algorithms produce near-optimum solutions which perform dramatically better than structured digraphs that have been previously considered for fixed virtual topologies. We mention that we have implemented all algorithms and have performed hundreds of experiments to validate the algorithms and study their behavior.

Our third major contribution is the development of a new class of routing protocol especially well suited for use in high-speed lightwave networks such as the WON. To overcome the effects of congestion and reduce the probability of packet loss in buffer-limited stations, we propose a routing algorithm known as detour routing, a generalization of the well-known deflection routing protocol. Detour routing provides a limited amount of adaptability without the overhead that comes with a full-blown adaptive routing procedure, which is an important consideration in a high-speed transmission system such as the WON. We demonstrate via simulation the superiority of detour routing over deflection routing and develop a technique for optimizing the performance of the WON by tuning its virtual topology.

During the course of our research several challenging problems have presented themselves. As we consider them to be beyond the scope of our original plan of research, we note them as topics for future exploration. Next follow descriptions of some of these problems and thoughts on promising methods of attack.

7.2 Future Research in the Design and Analysis of the WON

At this point honesty compels us to point out what has *not* been accomplished. Although we have performed a large number of experiments, the maximum size of the networks we experimented with is less than 200 stations. Given the goal of building WONs to serve thousands of users, it is important to validate the proposed algorithms on larger problems. Should the algorithms not prove equal to the task, new heuristics would have to be developed. As suggested in Chapter 5, the simulated annealing algorithm can be adapted for fast optimization, e.g., rapid quenching. Such approaches should be tested on very large problems. It is also possible to adopt a hierarchical approach in which the WON is partitioned into communities of interest based on the traffic matrix. These communities of interest could first be optimized followed by an optimization of the interconnection of the communities of interest. Research and experience with TimberWolfSC, a simulated-annealing–based package for cell placement in integrated circuits, has shown that one can make significant improvements in computation time without sacrificing the quality of solutions by introducing better methods for controlling the simulated annealing algorithm, e.g., by the use of better annealing schedules [LDS88, Lee90]. The simulated annealing algorithms used in this research are basically hand-crafted and could benefit from some of the techniques used in the newer versions of TimberWolfSC.

The need for robust WON designs was mentioned in Chapter 2. Since stations can fail at any time, the virtual topology of the WON is always in jeopardy of being disrupted, and such a disruption can have a deleterious effect upon the performance of the WON. Can we design virtual topologies that tolerate the loss of a station without having to redefine completely the virtual topology of the WON? A virtual topology that could achieve this level of robustness could ease some of the burden of network management in the WON. Najjar and Gaudiot [NG90] have attempted to characterize the robustness or resilience of the hypercube virtual topology; some of their methods may be applicable to arbitrary virtual topologies as well. Also into the realm of robustness falls the problem of designing a virtual topology that maintains good performance when the traffic matrix varies. It would be interesting to gauge the degree to which a given virtual topology optimized for a given traffic matrix performs when the traffic matrix changes. This has been done for the ShuffleNet virtual topology [EM88] but not for arbitrary virtual topologies.

It has been pointed out in Chapter 5 that the problem of finding a minimum-diameter digraph with fixed outdegree is solvable in polynomial time. The obvious algorithm runs in time that is bounded by a polynomial of high degree, so it is little consolation that the problem is solvable in polynomial time. It is not known, however, whether we can find a polynomial-time algorithm that runs in a reasonable amount of time, i.e., in time bounded by a polynomial of low degree. If such an algorithm could be found, it would allow us to design optimal virtual topologies in certain classes of WONs. It could also have profound implications for the $(p,\ D)$-digraph problem.

We have noted in Chapter 4 that the design of the optimal physical topology of the WON requires solving the Steiner tree problem. Our approach avoids this challenge by decomposing the problem into clustering and location steps and is suboptimal. A more thorough study of the problem of physical-topology design would investigate techniques for solving the Steiner tree problem, perhaps adapting known heuristics to the problem-specific aspects of physical-topology design in the WON.

Another issue to be addressed is the problem of providing integrated services in the WON. Our traffic models are primarily oriented toward capturing the behavior of bursty traffic, so voice and video services in the WON must be examined more carefully. The large bandwidth made available by lightwave technology can be effectively used to provide broadcast and other services.

In our study of the routing and congestion-control problem, the amount of time consumed running simulations has convinced us of the desirability of a more efficient technique for evaluating WON performance when detour (or deflection) routing is used. It would therefore be advantageous to develop a closed-form model of WON performance under detour routing. Such a model would be of use in optimizing the virtual topology of the WON operating under detour routing (our analytical model works only for single-path routing with unlimited buffers). Although we do not believe that a truly closed-form solution exists, our preliminary analysis suggests that there is a simple iterative algorithm for computing flows when detour routing is used. Such a procedure could then replace the heuristic objective function that we employ in the optimization of the virtual topology.

Detour routing has an advantage over deflection routing in that it can tolerate link failures. By detouring around a failed link (which includes the failure of a transceiver serving the link), the packet could eventually reach its destination, which can not be said for deflection routing, because the packet would always return to the deflecting station and reattempt to use the failed link. Unless the link recovers, this process is futile, resulting in continuous looping. When a station fails, however, detour routing can not get around the failed station. This raises the possibility of a modification of detour routing in which it is possible for a packet to make a detour around a failed station as well as a congested link. This scheme would require a means of notifying a station when its neighbor has failed, since notification is not automatic with unidirectional links. It would also be necessary to store more alternate paths than in the original form of detour routing, i.e., detours for congested or failed output ports and detours for failed stations. The refinement and analysis of this form of detour routing is a topic for future study.

One issue that is more economic than technical is that of compatibility with existing networks and the plan for transition to the new system, were the WON to be adopted for use as a MAN.

Appendix A

Derivation of a Lower Bound on Mean Internode Distance in the SCWON

In this appendix we derive a lower bound on the mean number of hops in an N-station SCWONi. We assume that the WONs under consideration all have p transceivers. Also, we restrict our attention to SCWONs that are stable when offered uniform traffic totaling γ packets per second, where stability is taken to mean that no WDM channel carries more traffic than the threshold of $\rho_{\mathbf{max}}$ packets per second.

We define $\mathcal{W}_d$ to be the set of all N-station p-transceiver SCWONs that have K WDM channels, have no more than d stations assigned to any channel, and are stable when the WON is offered a uniform traffic load totaling γ packets per second. We also define

$$\mathcal{W} \triangleq \bigcup_{d=1}^{N} \mathcal{W}_d$$

We further define two functions on $\mathcal{W}$: $h(G)$ is equal to the mean number of hops in $G \in \mathcal{W}$, and $\lambda(G)$ is equal to the total traffic load generated on all channels of $G \in \mathcal{W}$. Finally, we define the expression

$$H_d \triangleq \frac{LN}{N-1} - \frac{L-(L+1)pd+(pd)^{L+1}}{(N-1)(1-pd)^2} \tag{A.1}$$

where

$$L \triangleq \left\lceil \log_{pd}[(pd-1)N+1]-1 \right\rceil \tag{A.2}$$

Claim A.1 $\forall G \in \mathcal{W}_d \quad h(G) \geq H_d$.

We can construct a *pd*-ary tree of N nodes in which only one nonleaf node can have fewer than *pd* children. Such a tree is constructed by starting with the root and assigning *pd* children to each node in a breadth-first manner (the last node may be assigned fewer than *pd* children before the limit of N nodes has been reached).

If we number the levels in the tree from 0 to L, then it is easy to show that

$$\frac{1-(pd)^L}{1-pd} < N \leq \frac{1-(pd)^{L+1}}{1-pd}$$

since the lefthand and righthand sides of the inequality are the number of nodes in a full *pd*-ary tree of L and $L+1$ levels, respectively. From this inequality we can derive

$$\log_{pd}[(pd-1)N+1]-1 \leq L < \log_{pd}[(pd-1)N+1] \tag{A.3}$$

Since L takes on integral values and Equation (A.3) has the form $x-1 \leq L < x$, it must be true that $L = \lceil x-1 \rceil$. Thus,

$$L = \left\lceil \log_{pd}[(pd-1)N+1]-1 \right\rceil$$

which agrees with the definition of L in Equation (A.2).

The number of nodes, M, at level L is equal to N minus the number at levels 0 through $L-1$, i.e.,

$$\begin{aligned} M &= N - \sum_{k=0}^{L-1} (pd)^k \\ &= N - \frac{1-(pd)^L}{1-pd} \end{aligned}$$

If we denote by S the sum of lengths of the (shortest) paths from the root to all other nodes in the tree, then

$$\begin{aligned} S &= ML + \sum_{k=0}^{L-1} k(pd)^k \\ &= LN - \frac{L-(L+1)pd+(pd)^{L+1}}{(1-pd)^2} \end{aligned}$$

Recalling the definition of H_d in Equation (A.1), we notice that

$$H_d = \frac{S}{N-1} \tag{A.4}$$

If we let δ_{ij} represent the number of hops in the shortest path from station i to station j in G, then the sum of shortest-path lengths from all stations to station j is $\sum_{i=1,i\neq j}^{N} \delta_{ij}$. Since each station has p receivers, none of which shares its channel with more than d transmitters, each station receives from pd or fewer other stations. It is obvious that $S \leq \sum_{i=1,i\neq j}^{N} \delta_{ij}$ because the tree represents the best possible routing. Hence

$$NS \leq \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \delta_{ij}$$

Since

$$h(G) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \delta_{ij}$$

we have, substituting NS for the double summation and applying Equation (A.4)

$$H_d = \frac{S}{N-1} \leq h(G)$$

□

Claim A.2 $\forall G \in \mathcal{W}_d \quad \lambda(G) = \gamma h(G) \geq \gamma H_d$.

The relationship $h(G) = \lambda(G)/\gamma$ is proved in [Kle76, page 327]. We then use Claim A.1 to establish that $\gamma h(G) \geq \gamma H_d$.

□

Claim A.3 $\forall G \in \mathcal{W} \quad \exists d$ such that $H_d \leq \lambda(G)/\gamma \leq K\rho_{\mathbf{max}}/\gamma$

The first inequality follows immediately from Claim A.2.

Let λ_k denote the load carried on WDM channel k. Then $\lambda(G) = \lambda_1 + \lambda_2 + \cdots + \lambda_K$. Since G is stable, we have that $\lambda_k \leq \rho_{\mathbf{max}}$ for each channel k. Thus $\lambda(G) \leq K\rho_{\mathbf{max}}$, from which follows the second inequality of the Claim.

□

From Claims A.1–A.3 it follows that

$$h(G) \geq \min\{H_d \mid H_d \leq K\rho_{\mathbf{max}}/\gamma, 1 \leq d \leq N\} \tag{A.5}$$

In particular, if we restrict the number of WDM channels to be pN and allow no more than one station per channel, as in the DCWON, then a lower bound on the expected number of hops under uniform traffic, $\mathbb{E}$[hops], is given by

$$\mathbb{E}[\text{hops}] \geq \frac{L(N-1) + p(L+1) - p^{L+1}}{N-1} \tag{A.6}$$

where

$$L \triangleq \lceil \log_p[(p-1)N + 1] - 1 \rceil$$

The lower bound in Equation (A.6) is well known and has been come to be known as the **Moore** bound [HS60].

Appendix B

Derivation of an Upper Bound on Mean Internode Distance in the Modified de Bruijn Digraph

In this appendix we derive an upper bound on the mean number of hops $\overline{D}$ in an N-station DCWON with a virtual topology based upon the modified de Bruijn digraph of degree p described in Chapter 2:

$$\overline{D} \leq \frac{np^n + p^n - 1}{p^n - 1} - \frac{p}{p-1} \tag{B.1}$$

where $n = \log_p N$. This upper bound, in conjunction with the lower bound derived in Appendix A, is useful in approximating the mean internode distance in the modified de Bruijn digraph. Thus, it can also be used to approximate the mean packet delay in a WON which uses the de Bruijn virtual topology and has uniform traffic and distance matrices. We first establish a mathematical result (Claim B.3) and then show how to apply this result to the de Bruijn digraph.

First define an *$(L,\ p,\ q)$-vector*, where L, p, and q are positive integers, to be a vector of $L+1$ positive integers $\langle n_0,\ n_1, \ldots,\ n_L\rangle$ such that $n_0 = 1$, $n_1 = q$, $n_k \leq pn_{k-1}$ for $k = 2,\ 3, \ldots,\ L$, and $\sum_{k=0}^{L} n_k = p^L$.

Claim B.1 *Let L and p be positive integers and $p > 1$. There is only one $(L,\ p,\ p-1)$-vector, and it is given by $\langle n_0,\ n_1, \ldots,\ n_L\rangle$ where*

$$n_i = \begin{cases} 1 & \textit{if } i = 0 \\ (p-1)p^{i-1} & \textit{if } 1 \leq i \leq L \end{cases}$$

By the definition of the $(L,\ p,\ p-1)$-vector we must have $n_k \leq pn_{k-1}$, $n_{k-1} \leq pn_{k-2}, \ldots, n_2 \leq pn_1$. Solving this recursion, we obtain $n_k \leq (p-1)p^{k-1}$. Therefore,

$$\sum_{k=0}^{L} n_k \leq 1 + \sum_{k=1}^{L} (p-1)p^{k-1} = p^L$$

Since equality must hold in this last equation, it must be true that $n_k = (p-1)p^{k-1}$ for $k = 1,\ 2, \ldots,\ L$.

□

Claim B.2 *Let L and p be positive integers, $p > 1$. If $\langle n_0,\ n_1, \ldots,\ n_L\rangle$ is an $(L,\ p,\ p-1)$-vector and $\langle m_0,\ m_1, \ldots,\ m_L\rangle$ is an $(L,\ p,\ p)$-vector, then there exists a positive integer k such that $0 \leq k \leq L$, $n_i \leq m_i$ for $i < k$, and $n_j > m_j$ for $j \geq k$.*

First we note that the $(L,\ p,\ p-1)$-vector $\langle n_0,\ n_1, \ldots,\ n_L\rangle$ is unique and satisfies the conditions of Claim B.1.

Let k be the smallest integer for which $n_k > m_k$. Such a k must exist, for if it did not then we would have $m_i \geq n_i$ for $0 \leq i \leq L$, and hence

$$\sum_{i=0}^{L} m_i \geq \sum_{i=0}^{L} n_i \tag{B.2}$$

But since we know that, in particular, $m_1 = p > p - 1 = n_1$, the inequality in Equation (B.2) must be strict, i.e.,

$$\sum_{i=0}^{L} m_i > \sum_{i=0}^{L} n_i$$

This latter inequality is a contradiction, since both summations must equal p^L, by the definition of $(L,\ p,\ p-1)$- and $(L,\ p,\ p)$-vectors.

We know that $n_k = (p-1)p^{k-1}$ by the definition of $\langle n_0,\ n_1, \ldots,\ n_L \rangle$. So

$$m_k < (p-1)p^{k-1} \tag{B.3}$$

It must be true that $n_j > m_j$ for $j \geq k$. If this were not so, then there would exist an integer $l \geq k$ such that $m_l \geq n_l$. Thus,

$$m_l \geq (p-1)p^{l-1} \tag{B.4}$$

again by the definition of $\langle n_0,\ n_1, \ldots,\ n_L \rangle$. Since $m_l \leq pm_{l-1}$, $m_{l-1} \leq pm_{l-2}, \ldots,$ $m_{k+1} \leq pm_k$, we can write

$$m_l \leq p^{l-k} m_k \tag{B.5}$$

Equations (B.3) and (B.5) imply that $m_l < p^{l-k}(p-1)p^{k-1}$, which contradicts Equation (B.4). Thus, k satisfies the conditions of the Claim.

□

Claim B.3 *Let L and p be positive integers, $p > 1$. If $\langle n_0,\ n_1, \ldots,\ n_L \rangle$ is an $(L,\ p,\ p-1)$-vector and $\langle m_0,\ m_1, \ldots,\ m_L \rangle$ is an $(L,\ p,\ p)$-vector, then*

$$\sum_{i=0}^{L} in_i > \sum_{i=0}^{L} im_i$$

First we note that the $(L,\ p,\ p-1)$-vector $\langle n_0,\ n_1, \ldots,\ n_L\rangle$ is unique and satisfies the conditions of Claim B.1. Let k satisfy the conditions of Claim B.2.

Define $q_i \triangleq n_i - m_i$.

$$\sum_{i=0}^{L} q_i = \sum_{i=0}^{L} n_i - \sum_{i=0}^{L} m_i = 0$$

By Claim B.2, $q_i \leq 0$ for $i < k$ and $q_i > 0$ for $i \geq k$. Therefore

$$-\sum_{i=0}^{k-1} q_i = \sum_{j=k}^{L} q_j$$

The quantity on both sides of this equation is positive, so that multiplying the left side of this equation by $(k-1)$ and the right side by k, gives us the following strict inequality:

$$(k-1)\sum_{i=0}^{k-1}(m_i - n_i) < k\sum_{j=k}^{L}(n_j - m_j)$$

From this it immediately follows that

$$\sum_{i=0}^{k-1} i(m_i - n_i) < \sum_{j=k}^{L} j(n_j - m_j)$$

Hence

$$\sum_{i=0}^{L} in_i > \sum_{i=0}^{L} im_i$$

□

We now use Claim B.3 to establish the bound in Equation (B.1). Consider the (p-ary) reachability tree of the N-node modified de Bruijn digraph (of degree p), rooted at an arbitrary node of the digraph. The reachability tree consists of $\log_p N + 1$ levels,[1] and if m_k is the number of nodes at level k, then there will be no more than pm_k nodes at level $k+1$. We also note that level 1 of the reachability tree always contains p nodes—this was ensured in the modified de Bruijn digraph by

[1] We number the levels from 0 to $\log_p N + 1$.

eliminating self loops. If we let $L = \log_p N$ and denote by m_k the number of nodes at level k of the tree, then it is clear that $\langle m_0,\ n_1, \ldots,\ m_L \rangle$ is an $(L,\ p,\ p)$-vector.

Now define

$$n_i = \begin{cases} 1 & \text{if } i = 0 \\ (p-1)p^{i-1} & \text{if } 1 \le i \le L \end{cases}$$

By Claim B.1 $\langle n_0,\ n_1, \ldots,\ n_L \rangle$ is an $(L,\ p,\ p-1)$-vector. Since $\langle m_0,\ m_1, \ldots,\ m_L \rangle$ and $\langle n_0,\ n_1, \ldots,\ n_L \rangle$ satisfy the hypothesis of Claim B.3, we can state that

$$\sum_{i=0}^{L} i m_i < \sum_{i=0}^{L} i n_i \tag{B.6}$$

The righthand side of Equation (B.6) can be recomputed as follows:

$$\begin{aligned} \sum_{i=0}^{L} i n_i &= \sum_{i=1}^{L} i(p-1)p^{i-1} \\ &= (p-1)\sum_{i=1}^{L} i p^{i-1} \\ &= (p-1)\frac{d}{dp}\sum_{i=1}^{L} p^i \\ &= (p-1)\frac{d}{dp}\left(\frac{p-p^{L+1}}{1-p}\right) \\ &= (L+1)N - 1 - \frac{p(N-1)}{p-1} \end{aligned} \tag{B.7}$$

We next choose a reachability tree of the modified de Bruijn digraph that has the largest average height, i.e., the tree such that

$$\frac{1}{N-1}\sum_{i=0}^{L} i m_i \tag{B.8}$$

is as large as possible. It is then clear that the mean internode distance $\overline{D}$ in the modified de Bruijn digraph is less than the expression in Equation (B.8). If we divide the expression in Equation (B.7) by $N-1$, then this quantity is greater than the expression in Equation (B.8). Hence

$$\overline{D} \le \frac{(L+1)N - 1}{N-1} - \frac{p}{p-1}$$

which, after the substitution $L = n$, is identical to Equation (B.1).

Appendix C

Proof of the Convexity of the PTDP

In this appendix we show that the objective function of the PTDP is convex. We demonstrate the convexity of the objective function by showing that its Hessian is positive definite.

The minimization of the objective function

$$f(\boldsymbol{V}_1,\ \boldsymbol{V}_2,\ \ldots,\ \boldsymbol{V}_M) = \sum_{i=1}^{M} \sum_{j \in C(i)} \|\boldsymbol{V}_i - \boldsymbol{V}_j\|$$

given in Equation (4.1) is greatly simplified if we can show that the function is convex,[1] as this implies that any local minimum is also a global minimum [RV73]. The function $f(\cdot)$ is a linear combination of the functions

$$g_{ij}(\boldsymbol{V}_i,\ \boldsymbol{V}_j) = \|\boldsymbol{V}_i - \boldsymbol{V}_j\|$$

[1] A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if it satisfies Jensen's inequality, i.e., if $\boldsymbol{W}_i \in \mathbb{R}^n$ for $i = 1,\ 2,\ \ldots,\ m$, then

$$f\left(\sum_{i=1}^{m} \alpha_i \boldsymbol{W}_i\right) \leq \sum_{i=1}^{m} \alpha_i f(\boldsymbol{W}_i)$$

where $0 \leq \alpha_i \leq 1$ for $i = 1,\ 2,\ \ldots,\ m$, and $\sum_{i=1}^{m} \alpha_i = 1$.

which we can rewrite as

$$g_{ij}\left[(x_1,y_1),(x_2,y_2)\right] = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$$

It is convenient to define

$$X = \frac{(x_1-x_2)}{\left[(x_1-x_2)^2+(y_1-y_2)^2\right]^{3/4}}$$

and

$$Y = \frac{(y_1-y_2)}{\left[(x_1-x_2)^2+(y_1-y_2)^2\right]^{3/4}}$$

Computing the Hessian of g, we find that

$$\begin{aligned}\nabla^2 g\left[(x_1,y_1),(x_2,y_2)\right] &= \begin{bmatrix} \frac{\partial^2 g}{\partial x_1\partial x_1} & \frac{\partial^2 g}{\partial x_1\partial y_1} & \frac{\partial^2 g}{\partial x_1\partial x_2} & \frac{\partial^2 g}{\partial x_1\partial y_2} \\ \frac{\partial^2 g}{\partial y_1\partial x_1} & \frac{\partial^2 g}{\partial y_1\partial y_1} & \frac{\partial^2 g}{\partial y_1\partial x_2} & \frac{\partial^2 g}{\partial y_1\partial y_2} \\ \frac{\partial^2 g}{\partial x_2\partial x_1} & \frac{\partial^2 g}{\partial x_2\partial y_1} & \frac{\partial^2 g}{\partial x_2\partial x_2} & \frac{\partial^2 g}{\partial x_2\partial y_2} \\ \frac{\partial^2 g}{\partial y_2\partial x_1} & \frac{\partial^2 g}{\partial y_2\partial y_1} & \frac{\partial^2 g}{\partial y_2\partial x_2} & \frac{\partial^2 g}{\partial y_2\partial y_2} \end{bmatrix} \\ &= \begin{bmatrix} YY & -XY & -YY & XY \\ -XY & XX & XY & -XX \\ -YY & XY & YY & -XY \\ XY & -XX & -XY & XX \end{bmatrix}\end{aligned}$$

We can show that the Hessian is positive semidefinite[2] as follows. Let z_1, z_2, z_3, and z_4 be real numbers. It is possible, though tedious, to show that

$$\sum_{i=1}^{4}\sum_{j=1}^{4}(\nabla^2 g)_{ij} z_i z_j = \left[(z_1 Y - z_2 X) - (z_3 Y - z_4 X)\right]^2 \geq 0$$

[2] A symmetric $n \times n$ matrix (a_{ij}) is *positive semidefinite* if for any z_1, z_2, $\ldots$, z_n

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} z_i z_j \geq 0$$

Thus, the functional form $g\left[(x_1, y_1), (x_2, y_2)\right]$ is convex. The objective function $f(\cdot)$ is a linear combination of the convex functional forms $g_{ij}\left[(x_1, y_1), (x_2, y_2)\right]$. Thus, Equation (4.1) is convex. This fact implies that any local minimum of Equation (4.1) is also the global minimum being sought after. Solution of the problem then hinges on efficient search procedures for finding a local minimum, and we may apply any of several well-known techniques for the optimization of a convex function, e.g., the Frank–Wolfe, Hooke–Jeeves, or gradient method.

Appendix D

The Shared-Channel VTDP is NP-Complete

We demonstrate in this appendix that the shared-channel VTDP is NP-complete. The proof is by polynomial transformation from a well-known NP-complete problem, the partition problem [GJ79].

First we state the partition problem in its yes/no decision form:

Instance A set $A = \{a_1,\ a_2,\ \ldots,\ a_N\}$ of positive integers.

Question Is there a subset $A' \subseteq A$ such that $\sum_{a \in A'} a = \sum_{a \in A - A'} a$?

Next we state the following yes/no decision problem, which can be seen to be a simple subproblem of the shared-channel VTDP:

Instance A set of N single-transceiver stations and an $N \times N$ traffic matrix (γ_{ij}).

Question Is there an assignment of stations to two channels such that the mean number of hops is 1 and the utilization of each channel is less than or equal to 1?

Clearly, if we could solve the shared-channel VTDP, then we could solve this decision problem, because its is a subproblem of the shared-channel VTDP. Thus, if the decision problem could be shown to be NP-complete, then the shared-channel VTDP would also be NP-complete. For convenience we will refer to the decision problem as the shared-channel VTDP in the remainder of this appendix, even though it is only a subproblem of the general problem.

We next define a polynomial transformation $\mathcal{T}(\cdot)$ that converts any instance x of the partition problem into an instance $\mathcal{T}(x)$ of the shared-channel VTDP in polynomial time. We then show that the answer to x is yes if and only if the answer to $\mathcal{T}(x)$ is yes. This is sufficient to demonstrate that the shared-channel VTDP is NP-hard; giving a nondeterministic polynomial-time algorithm for the shared-channel VTDP completes the proof of NP-completeness.

The transformation is specified by explaining how the numbers a_i are converted to the traffic matrix (γ_{ij}). We define (γ_{ij}) as follows:

$$\gamma_{ij} = \begin{cases} 2a_i/\sum_{j=1}^{N} a_j & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Notice that $\sum_{i=1}^{N} \gamma_{ii} = 2$. Thus, the transformation is defined by $\mathcal{T}[(a_i)_{i=1}^{N}] = (\gamma_{ij})_{i=1,\ldots,N}^{j=1,\ldots,N}$. It is obvious that such a transformation can be computed in polynomial time.

Claim D.1 *The answer to the partition problem with instance* $\{a_1, a_2, \ldots, a_N\}$ *is yes if and only if the answer to the shared-channel VTDP with instance* $(\gamma_{ij})_{i=1,\ldots,N}^{j=1,\ldots,N}$ *is yes.*

Suppose that the answer to the partition problem with instance $(a_i)_{i=1}^{N}$ is yes. Then there is a set $A' \subseteq A$ such that $\sum_{a\in A'} a = \sum_{a\in A-A'} a$. If we define $I \triangleq$

$\{i \mid a_i \in A'\}$ and $J \triangleq \{j \mid a_j \in A - A'\}$, then it easy to check that tuning the transmitters and receivers of stations from set I to channel 1 and those of stations from set J to channel 2 provides a positive answer to the transformed problem. Obviously, the mean number of hops is equal to 1 with this tuning since all stations are connected directly to themselves. The total load on channel 1 is equal to 1, as can be seen from the following:

$$\begin{aligned}
\sum_{i\in I}\sum_{j=1}^{N}\gamma_{ij} &= \sum_{i\in I}\gamma_{ii} \\
&= \sum_{i\in I}\frac{2a_i}{\sum_{j=1}^{N}a_j} \\
&= 2\cdot\frac{\sum_{i\in I}a_i}{\sum_{j=1}^{N}a_j} \\
&= 2\cdot\frac{1}{2} \\
&= 1
\end{aligned}$$

The load on channel 2 is similarly shown to be equal to 1.

Conversely, suppose that the transformed problem instance $(\gamma_{ij})_{i=1,\ldots,N}^{j=1,\ldots,N}$ has a yes answer. Let the set I denote the stations tuned to channel 1. It is obvious that each station's transmitter must be cotuned to the same channel as its receiver since every station must be able to reach itself in exactly one hop. Since $\sum_{i=1}^{N}\gamma_{ii} = 2$, each channel must carry a load of exactly 1, i.e., $\sum_{i\in I}\gamma_{ii} = 1$. Therefore, we let $A' \triangleq \{a_i \mid i \in I\}$. Hence,

$$\begin{aligned}
\sum_{a\in A'}a &= \sum_{i\in I}a_i \\
&= \sum_{i\in I}\left(\sum_{j=1}^{N}a_j\right)\gamma_{ii}/2 \\
&= \frac{1}{2}\cdot\left(\sum_{j=1}^{N}a_j\right)
\end{aligned}$$

which implies that A' induces a balanced partition of A.

□

Therefore, we have proven that the answer to an instance x of the partition problem is yes if and only if the answer to instance $\mathcal{T}(x)$ of the shared-channel VTDP is yes. Noting that the shared-channel VTDP has a nondeterministic polynomial-time algorithm, which consists of guessing a solution and checking whether is holds, we conclude that the shared-channel VTDP is NP-complete.

Appendix E

Derivation of a Bound for the (p, D, C)-Digraph Problem

In this appendix we derive an upper bound on the number of nodes, N, in a digraph of degree p, diameter D, and girth C.

First we remark that if $C \geq D$ then an upper bound on N is given by the well-known Moore bound for directed graphs [TS79, FYdM84]:

$$N \leq \frac{p^{D+1} - 1}{p - 1} \tag{E.1}$$

Equation (E.1) is true because a digraph in which all shortest circuits are as long as possible will have a best-case reachability tree in which no node is repeated until level D, as shown in Figure E.1

For the case in which $C < D$, we consider the reachability tree rooted at a distinguished starting node, such as is depicted in Figure E.2. In the best case, i.e., the case in which N is maximized, there will be one node at level 0, p nodes at level 1, ..., p^{C-1} nodes at level $C - 1$, and these nodes will all be distinct. Thus, we see that in the best case no node has a path back to itself containing

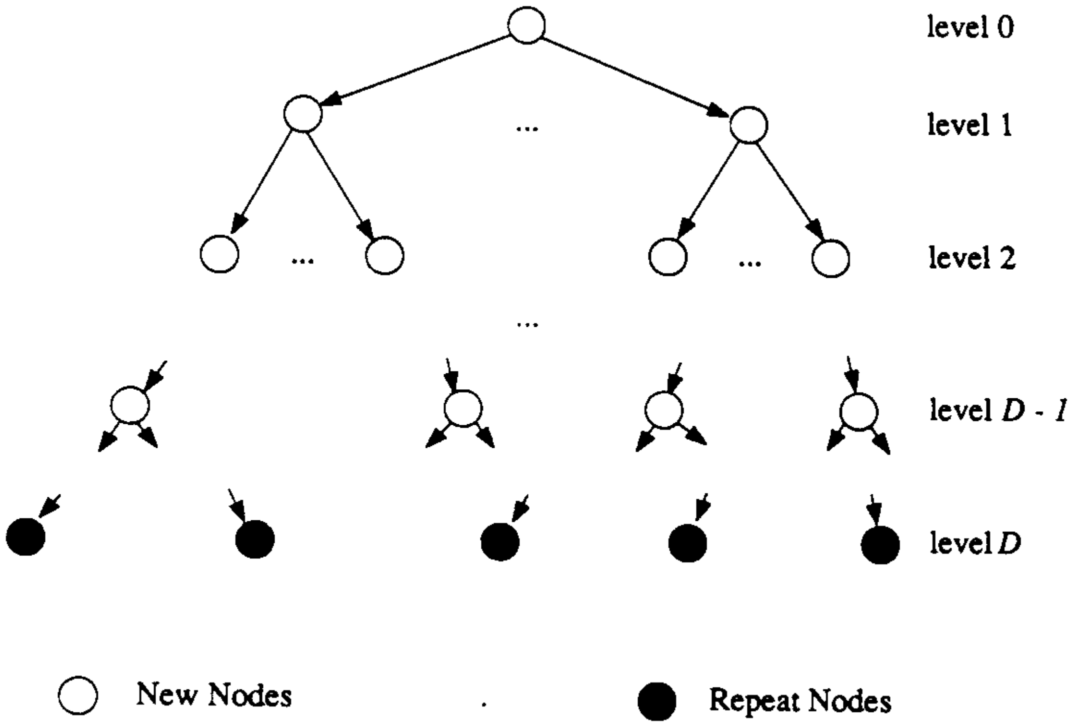


Figure E.1: Reachability Tree for a p-Regular Digraph of Maximum Diameter D.

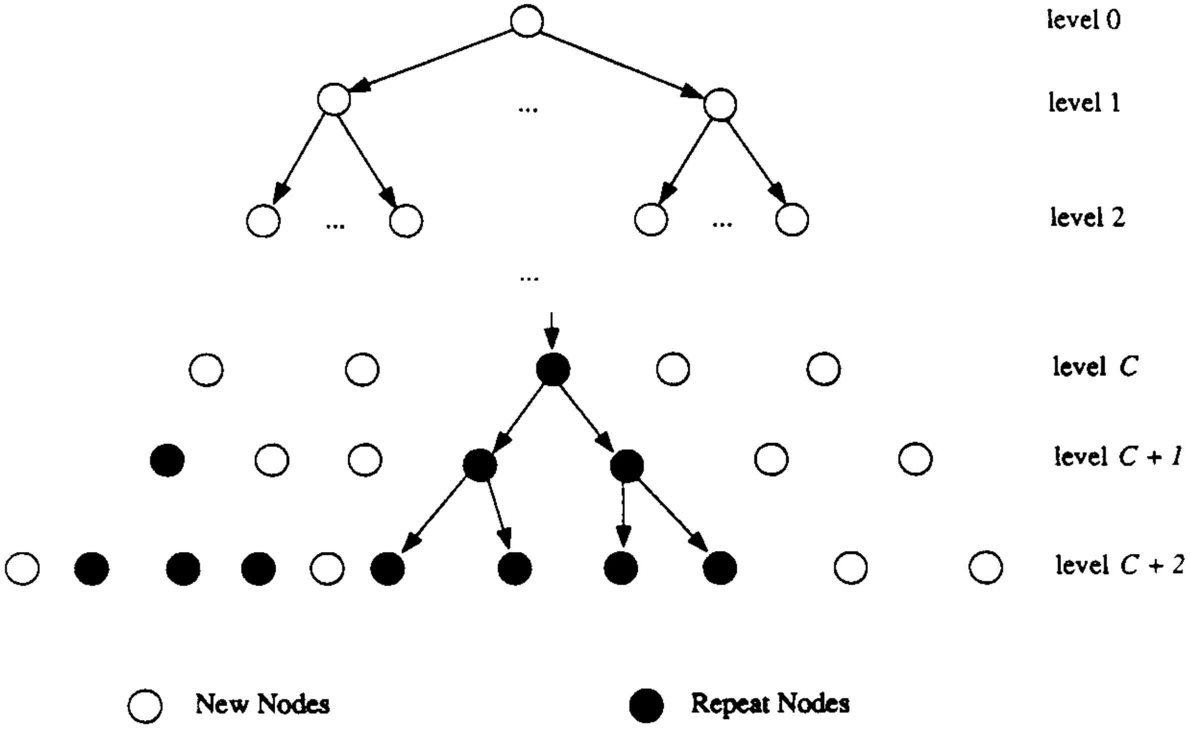


Figure E.2: Reachability Tree for a p-Regular Digraph of Maximum Diameter D and Girth C.

fewer than C hops. Suppose that the shortest nontrivial path from any node back to itself contains C hops. Then, although there are p^C nodes at level C of the reachability graph, one of these nodes will be—by virtue of the C-hop circuit from the root node back to itself—a repeat of the distinguished root node, which leaves $p^C - 1$ new nodes at that level. Call the subtree rooted at the repeated node of level C the *distinguished subtree*. Of p^{C+k} possible new nodes at level $C + k$, we can immediately rule out as repeats the p^k nodes of the distinguished subtree. Furthermore, $p^k - 1$ of the remaining nodes at this level are repeats since they are within C hops of the nodes at level k, one of whose repeats was counted in the distinguished subtree. Thus, $p^{C+k} - 2p^k + 1$ nodes at level $C + k$ *might* not be repeated nodes. We can continue this counting process until we reach level D, at which point all nodes must be repeats. Summing up the number of possible nodes in the reachability tree, we get

$$\begin{aligned} N &\leq \sum_{k=0}^{C-1} p^k + \sum_{k=0}^{D-C-1} \left(p^{C+k} - 2p^k + 1\right) \\ &= \frac{1-p^C}{1-p} + \frac{p^C - p^D}{1-p} + 2\frac{1-p^{D-C}}{1-p} + D - C \\ &= \frac{p^D - 2p^{D-C} + 1}{p-1} + D - C \end{aligned} \tag{E.2}$$

To recapitulate, we have, combining Equations (E.1) and (E.2), the following bound on the number of nodes, N, in a digraph of degree p with diameter D and girth C: $N \leq \boldsymbol{N(p,\ D,\ C)}$ where

$$N(p,\ D,\ C) \triangleq \begin{cases} \frac{p^{D+1}-1}{p-1} & \text{if } C \geq D \\ \frac{p^D - 2p^{D-C}+1}{p-1} + D - C & \text{if } C < D \end{cases} \tag{E.3}$$

The $(p,\ D,\ C)$-*digraph problem* is to find a digraph with the specified parameters whose number of nodes approaches the quantity given in Equation (E.3).

Bibliography

[Abr70] N. Abramson. The ALOHA system—Another alternative for computer communications. In *Proceedings of the Fall Joint Computer Conference, AFIPS Conference 37*, pages 281–284, 1970.

[Aca87] A. S. Acampora. A multichannel multihop local lightwave network. In *Proceedings of GLOBECOM '87*, pages 37.5.1–37.5.9, Tokyo, Japan, November 1987.

[AK89] Anthony S. Acampora and Mark J. Karol. An overview of lightwave packet networks. *IEEE Network*, 3(1):29–41, January 1989.

[AKH87] Anthony S. Acampora, Mark J. Karol, and Michael G. Hluchyj. Terabit lightwave networks: The multihop approach. *AT&T Technical Journal*, 66(6):21–34, November/December 1987.

[AKH88] A. S. Acampora, M. J. Karol, and M. G. Hluchyj. Multihop lightwave networks: A new approach to achieve terabit capabilities. In *Proceedings of ICC '88*, volume 1, pages 1478–1484, 1988.

[AL86] Algirdas Avižienis and Jean-Claude Laprie. Dependable computing: From concepts to design diversity. *Proceedings of the IEEE*, 74(5):629–638, May 1986.

[Bab88] Robert G. Babb II, editor. *Programming Parallel Processors*. Addison-Wesley, Reading, Massachusetts, 1988.

[Bak86] Donald G. Baker. *Local-Area Networks with Fiber-Optic Applications*. Prentice-Hall, Englewood Cliffs, New Jersey, 1986.

[BCF89] Flaminio Borgonovo, Enrico Cadorin, and Luigi Fratta. Performance evaluation of tree topology local area networks. In *Proceedings of the*

Ninth Conference on Measurement in Communication Systems, October 1989.

[BCMP75] Forest Baskett, K. Mani Chandy, Richard R. Muntz, and Fernando G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22(2):248–260, April 1975.

[BDQ82] J.-C. Bermond, C. Delorme, and J.-J. Quisquater. Tables of large graphs with given degree and diameter. *Information Processing Letters*, 15(1):10–13, August 1982.

[BDQ86] J.-C. Bermond, C. Delorme, and J.-J. Quisquater. Strategies for interconnection networks: Some methods from graph theory. *Journal of Parallel and Distributed Computing*, 3(4):433–449, December 1986.

[BG89] Joseph A. Bannister and Mario Gerla. Design of the wavelength-division optical network. Technical Report CSD-890022, UCLA Computer Science Department, Los Angeles, California, May 1989.

[Bib81] K. J. Biba. LocalNet: A digital communications network for broadband coaxial cable. In *Proceedings of COMPCON '81*, pages 59–63, Spring 1981.

[BT80] W. G. Bridges and S. Toueg. On the impossibility of directed Moore graphs. *Journal of Combinatorial Theory B*, 29:339–341, 1980.

[BT89] Allan M. Bignell and Terence D. Todd. SIGnet: A new ultra-high-speed lightwave network architecture. In *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 40–43, June 1989.

[CCMS73] V. G. Cerf, D. D. Cowan, R. C. Mullin, and R. G. Stanton. A lower bound on the average shortest path length in regular graphs. *Networks*, 4(4):335–342, December 1973.

[CG87] Imrich Chlamtac and Aura Ganz. Toward alternative high speed networks: The SWIFT architecture. In *Proceedings of IEEE INFOCOM '87*, pages 1102–1108, San Francisco, California, March 1987.

[CGK88] I. Chlamtac, A. Ganz, and G. Karmi. Circuit switching in multi-hop lightwave networks. In *Proceedings of the ACM SIGCOMM '88 Symposium*, pages 188–199, Stanford, California, August 1988.

[CGK89] I. Chlamtac, A. Ganz, and G. Karmi. Purely optical networks for terabit communication. In *Proceedings of IEEE INFOCOM '89*, volume 3, pages 887–896, Ottawa, Canada, April 1989.

[CRSV86] Andrea Casotto, Fabio Romeo, and Alberto Sangiovanni-Vincentelli. A parallel simulated annealing algorithm for the placement of macro-cells. In *Proceedings of the 1986 IEEE International Conference on Computer-Aided Design*, pages 30–33, Santa Clara, California, November 1986.

[DKN87] F. Darema, S. Kirkpatrick, and V. A. Norton. Parallel techniques for chip placement by simulated annealing on shared memory systems. In *Proceedings of the 1987 IEEE International Conference on Computer Design*, pages 87–90, Rye Brook, New York, October 1987.

[DQD88] Distributed queue dual bus (DQDB) metropolitan area network (MAN). Proposed Standard IEEE P802.6/D6-88/105, November 1988. Institute of Electrical and Electronics Engineers.

[Dra89] C. Dragone. Efficient $N \times N$ star couplers using Fourier optics. *Journal of Lightwave Technology*, LT-7(3):479–489, March 1989.

[EF83] Gregory Ennis and Peter Filice. Overview of a broad-band local area network protocol architecture. *IEEE Journal on Selected Areas in Communications*, SAC-1(5):832–841, November 1983.

[EM88] Martin Eisenberg and Nader Mehravari. Performance of the multichannel multihop lightwave network under nonuniform traffic. *IEEE Journal on Selected Areas in Communications*, 6(7):1063–1078, August 1988.

[FDD87] Fiber-distributed data interface (FDDI)—Token ring media access control (MAC). American National Standard for Information Systems ANSI X3.139-1987, July 1987. American National Standards Institute.

[Flo62] Robert W. Floyd. Algorithm 97: Shortest path. *Communications of the ACM*, 5(6):345, June 1962.

[FYdM84] Miguel A. Fiol, Luis Andres Yerba, and Ignacio Alegre de Miquel. Line digraph iterations and the $(d,\ k)$ digraph problem. *IEEE Transactions on Computers*, C-33(5):400–403, May 1984.

[Ger73] Mario Gerla. *The Design of Store-and-Forward (S/F) Networks for Computer Communications.* PhD thesis, Computer Science Department, University of California, Los Angeles, California, 1973. Technical Report UCLA-ENG-7319.

[Ger78] Curtis F. Gerald. *Applied Numerical Analysis.* Addison-Wesley, Reading, Massachusetts, second edition, 1978.

[Ger81] Mario Gerla. Routing and flow control. In Franklin F. Kuo, editor, *Protocols and Techniques for Data Communication Networks*, chapter 4, pages 122–174. Prentice-Hall, Englewood Cliffs, New Jersey, 1981.

[GF88] Mario Gerla and Luigi Fratta. Tree structured fiber optic MAN's. *IEEE Journal on Selected Areas in Communications*, SAC-6(6):934–943, July 1988.

[GJ79] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman and Company, New York, New York, 1979.

[GK77] Mario Gerla and Leonard Kleinrock. On the topological design of distributed computer networks. *IEEE Transactions on Communications*, COM-25(1):48–60, January 1977.

[HC] Zygmunt Haas and David R. Cheriton. *Blazenet*: A photonically implementable wide-area network. *IEEE Transactions on Communications.* to appear.

[HC87] Zygmunt Haas and David R. Cheriton. A case for packet switching in high-performance wide-area networks. In *Proceedings of the ACM SIGCOMM '87 Workshop*, pages 402–409, Stowe, Vermont, August 1987.

[HK88] Michael G. Hluchyj and Mark J. Karol. ShuffleNet: An application of generalized perfect shuffles to multihop lightwave networks. In *Proceedings of IEEE INFOCOM '88*, pages 4B.4.1–4B.4.12, New Orleans, Louisiana, March 1988.

[HS60] A. J. Hoffman and R. R. Singleton. On Moore graphs with diameters 2 and 3. *IBM Journal of Research and Development*, 4(5):497–504, November 1960.

[II83] Makoto Imase and Masaki Itoh. A design for directed graphs with minimum diameter. *IEEE Transactions on Computers*, C-32(8):782–784, August 1983.

[JD88] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[Jos88] Mark K. Joseph. *Architectural Issues in Fault-Tolerant, Secure Computing Systems.* PhD thesis, Computer Science Department, University of California, Los Angeles, California, June 1988. Technical Report CSD-880047.

[Kap85] F. P. Kapron. Fiber-optic system tradeoffs. *IEEE Spectrum*, 22(3):69–75, March 1985.

[Kar88] Mark J. Karol. Optical interconnection using ShuffleNet multihop networks in multi-connected ring topologies. In *Proceedings of the ACM SIGCOMM '88 Symposium*, pages 25–34, Stanford, California, August 1988.

[KGV83] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[Kim87] Tatsuya Kimura. Coherent optical fiber transmission. *Journal of Lightwave Technology*, LT-5(4):414–428, April 1987.

[Kle76] Leonard Kleinrock. *Queueing Systems, Volume II: Computer Applications.* John Wiley and Sons, New York, New York, 1976.

[KS88] Mark J. Karol and Salman Shaikh. A simple adaptive routing scheme for ShuffleNet multihop lightwave networks. In *Proceedings of GLOBECOM '88*, pages 1640–1647, Miami, Florida, November 1988.

[LDS88] Jimmy Lam, Jean-Marc Delosme, and Carl Sechen. An efficient simulated annealing schedule for row-based placement. In *MCNC International Workshop on Placement and Routing*, Research Triangle Park, North Carolina, May 1988.

[Lee90] Kai-Win Lee. *Global Routing of Row-Based Integrated Circuits.* PhD thesis, Department of Electrical Engineering, Yale University, New Haven, Connecticut, May 1990.

[Lin89] Richard A. Linke. Frequency division multiplexed optical networks using heterodyne detection. *IEEE Network*, 3(3):13–20, March 1989.

[LS83] Steven S. Lavenberg and Charles H. Sauer. Analytical results for queueing models. In Steven S. Lavenberg, editor, *Computer Performance Modeling Handbook*, chapter 3. Academic Press, New York, New York, 1983.

[Max85] N. F. Maxemchuk. Regular mesh topologies in local and metropolitan area networks. *AT&T Technical Journal*, 64(7):1659–1685, September 1985.

[Max87] Nicholas F. Maxemchuk. Routing in the Manhattan street network. *IEEE Transactions on Communications*, COM-35(5):503–512, September 1987.

[Max89] N. F. Maxemchuk. Comparison of deflection and store-and-forward techniques in the Manhattan street and shuffle-exchange networks. In *Proceedings of IEEE INFOCOM '89*, volume 3, pages 800–809, Ottawa, Canada, April 1989.

[Mie58] W. Miehle. Link length minimization in networks. *Operations Research*, 6:232–243, 1958.

[MRR+53] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[NG90] Walid Najjar and Jean-Luc Gaudiot. Network resilience: A measure of network fault tolerance. *IEEE Transactions on Computers*, 31(2):174–181, February 1990.

[NTM85] M. Mehdi Nassehi, Fouad A. Tobagi, and Michel E. Marhic. Fiber optic configurations for local area networks. *IEEE Journal on Selected Areas in Communications*, SAC-3(6):941–949, November 1985.

[Ost77] Lawrence M. Ostresh, Jr. The multifacility location problem: Applications and descent theorems. *Journal of Regional Science*, 17:409–419, 1977.

[Pal88] Joseph C. Palais. *Fiber Optic Communications*. Prentice Hall, Englewood Cliffs, New Jersey, second edition, 1988.

[Rad88] Francisc Radó. The Euclidean multifacility location problem. *Operations Research*, 36(3):485–492, May-June 1988.

[RV73] A. Wayne Roberts and Dale E. Varberg. *Convex Functions*. Academic Press, New York, New York, 1973.

[SON88a] Synchronous digital hierarchy bit rates. CCITT Recommendation G.707, ITU, Geneva, Switzerland, November 1988.

[SON88b] Network node interface for the synchronous digital hierarchy. CCITT Recommendation G.708, ITU, Geneva, Switzerland, November 1988.

[SON88c] Synchronous multiplexing structure. CCITT Recommendation G.709, ITU, Geneva, Switzerland, November 1988.

[TS79] Sam Toueg and Kenneth Steiglitz. The design of small-diameter networks by local search. *IEEE Transactions on Computers*, C-28(7):537–542, July 1979.

[VR67] Roger C. Vergin and Jack D. Rogers. An algorithm and computational procedure for locating economic facilities. *Management Science*, 13(6):B–240–B–254, February 1967.

[VW89] R. S. Vodhanel and R. E. Wagner. Multi-gigabit/sec coherent lightwave systems. In *Proceedings of ICC '89*, pages 14.4.1–14.4.6, Boston, Massachusetts, June 1989.

[War62] Stephen Warshall. A theorem on Boolean matrices. *Journal of the ACM*, 9(1):11–12, January 1962.

[Wei36] E. Weiszfeld. Sur le point pour lequel la somme de distances de n points donnés est minimum. *Tohoku Mathematical Journal*, 43:355–386, 1936.

[Win87] Pawel Winter. Steiner problem in networks: A survey. *Networks*, 17:129–167, 1987.