Network Reliability and Fault Tolerance

Muriel Médard

medard@mit.edu

Laboratory for Information and Decision Systems

Room 35-212

Massachusetts Institute of Technology

77 Massachusetts Avenue, Cambridge, MA 02139

Steven S. Lumetta

lumetta@uiuc.edu

Coordinated Science Laboratory

University of Illinois Urbana-Champaign

1308 W. Main Street, Urbana, IL 61801

1 Introduction

The majority of communications applications, from cellular telephone conversations to credit card transactions, assume the availability of a reliable network. At this level, data are expected to traverse the network and to arrive intact at their destination. The physical systems that compose a network, on the other hand, are subjected to a wide range of problems, ranging from signal distortion to component failures. Similarly, the software that supports the high-level semantic interface

often contains unknown bugs and other latent reliability problems. Redundancy underlies all approaches to fault tolerance. Definitive definitions for all concepts and terms related to reliability, and, more broadly, dependability, can be found in [AAC+92].

Designing any system to tolerate faults first requires the selection of a fault model, a set of possible failure scenarios along with an understanding of the frequency, duration, and impact of each scenario. A simple fault model merely lists the set of faults to be considered; inclusion in the set is decided based on a combination of expected frequency, impact on the system, and feasibility or cost of providing protection. Most reliable network designs address the failure of any single component, and some designs tolerate multiple failures. In contrast, few attempt to handle the adversarial conditions that might occur in a terrorist attack, and cataclysmic events are almost never addressed at any scale larger than a city.

The temporal characteristics of faults vary widely, but can be roughly categorized as permanent, intermittent, or transient. Failures that prevent a component from functioning until repaired or replaced, such as the destruction of a network fiber by a backhoe, are considered permanent. Failures that allow a component to function properly some of the time are called intermittent. Damaged connectors and electrical components sometimes produce intermittent faults, operating correctly until mechanical vibrations or thermal variations cause a failure, and recovering when conditions change again. The last category, transient faults, is usually the easiest to handle. Transient faults range from changes in the contents of computer memory due to cosmic rays, to bit errors due to thermal noise in a demodulator, and are typically infrequent and unpredictable. The difference between an intermittent fault and a transient fault is sometimes solely one of frequency; for transient faults, a combination of error-correcting codes and data retransmission usually provides adequate protection.

Redundancy takes two forms, spatial and temporal. Spatial redundancy replicates the components or data in a system. Transmission over multiple paths through a network and the use of error-correction codes are examples of spatial redundancy. Temporal redundancy underlies automatic repeat request (ARQ) algorithms, such as the sliding window abstraction used to support reliable transmission in the Internet's Transmission Control Protocol (TCP). A reliable network typically provides both spatial and temporal redundancy to tolerate faults with differing temporal persistence. Spatial redundancy is necessary to overcome permanent failures in physical compo-

nents, while temporal redundancy requires fewer resources and is thus preferable when dealing with transient errors.

Beyond the selection of a fault model, several additional problems must be considered in the design of a fault-tolerant system. A system must be capable of detecting each fault in the model, and must be able to isolate each fault from the functioning portion of the system in a manner that prevents faulty behavior from spreading. As a fault detection mechanism may detect more than one possible fault, a system must also address the process of fault diagnosis (or localization), which narrows the set of possible faults and allows more efficient fault isolation techniques to be employed. An error identified by a system need not necessarily be narrowed down to a single possible fault, but a smaller set of possibilities usually allows a more efficient strategy for recovery.

Fault isolation boundaries are usually designed to provide fail-stop behavior for the desired fault model. The term fail-stop implies that incorrect behavior does not propagate across the fault isolation boundary; instead, failed components cease to produce any signals. Fail-stop does not imply self-diagnosis; components adjacent to a failed component may diagnose the failure and deliberately ignore any signals from the failed component, but the physical system design must allow such a decision. In a router, for example, the interconnect between cards controlling individual links must provide electrical isolation to support fail-stop behavior for failed cards. A bus-based computer interconnect does not allow for fail-stop, as nothing can prevent a failed card from driving the bus lines inappropriately. In modern, high-end servers, such buses have been replaced by switched networks with broadcast capability in order to enable such isolation. The eradication of similar phenomena in the move from shared to switched Ethernets in the mid-1990's was one of the main administrative advantages of the change, as failed hosts are much less likely to render a switched network unusable by flooding it with continuous traffic.

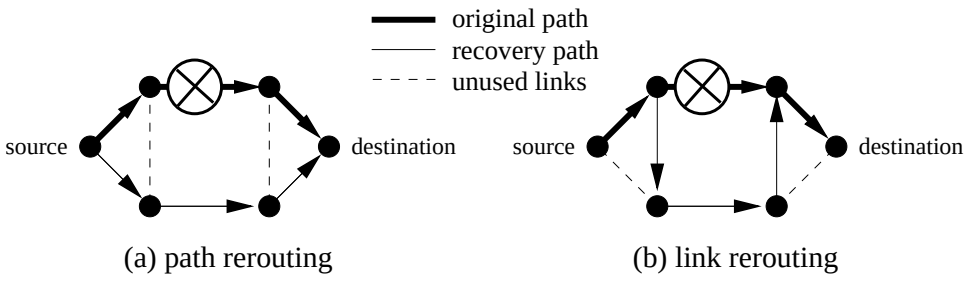
Two models of network service have dominated research and commercial networking. The first is the telephony network, or more generally a network in which quasi-permanent routes called circuits deliver fixed data capacity from one point to another. In digital telephony, a voice circuit requires 64 kbps; a single lightpath in a wavelength division multiplexed (WDM) optical network may deliver up to 40 Gbps, but is conceptually similar to the circuit used to carry a phone call. The second network service model is the packet-switched data network, which evolved from the early ARPANET and NSFNET projects into the modern Internet. Packet-switched networks do

not usually provide strong guarantees on delivered data rate or maximum delay, but are typically more efficient than circuit-oriented designs, which must base guaranteed agreements on worst-case traffic load scenarios.

For the purposes of our discussion, the key difference between these two models lies in the fact that applications using packet-switched networks can generally tolerate more serious service disruptions than can those based on circuit-switched networks. The latter class of applications may assume that data rate, delay, and jitter guarantees provided by the network will be honored even when failures occur, whereas minor disruptions may occur even in normal circumstances on packet-switched networks due to fluctuations in traffic patterns and loads. Fault tolerance issues are thus addressed in markedly different ways in the two types of networks. In packet-switched networks like the Internet, users currently tolerate restoration times of minutes [LABJ00, LAWV01], whereas fault tolerance for circuit-switched networks can be considered a component of quality of service (QoS) [MK96, PO97], and is typically achieved in milliseconds, or, at worst, seconds.

The majority of this article focuses on fault tolerance issues in high-speed backbone networks, such as wide-area networks (WAN's) and metropolitan-area networks (MAN's). Such networks are predominantly circuit-based and carry heavy traffic loads. As even a short down time may cause substantial data loss, rapid recovery from failure is important, and these networks require high levels of reliability. Backbone networks are generally are implemented using optical transmission and, conversely, fault tolerance in optical networks is typically considered in the context of backbone networks [GR00, ZS00]. In these networks, a failure may arise because a communications link is disconnected or a network node becomes incapacitated. Failures may occur in military networks under attack [Gre95], as well as in public networks, in which failures, albeit rare, can be extremely disruptive [SB00, Chap. 8].

The next section provides an overview of fault detection mechanisms and the basic strategies available for recovery from network component failures. Sections 3 and 4 build on these basics to illustrate recovery schemes for high-speed backbone networks. Sections 5 and 6 examine simple and more complex topologies and discuss the relationship between topology and recovery. Section 5 highlights ring topologies, as they are a key architectural component of high-speed networks. Section 6 extends the concepts developed for rings by overlaying logical ring topologies over physical mesh topologies. We also discuss some link and node-based reliability schemes



(a) path rerouting (b) link rerouting

Figure 1: Path and link rerouting. The failure is marked with an $\otimes$.

that are specifically tailored to mesh networks. Although the text focuses on approaches to fault tolerance in high-speed backbone networks, many of the principles also apply to other types of networks. In Section 7, we move away from circuit-switched networks and examine fault tolerance for packet-switched networks, and in particular the Internet. Finally, Section 8 discusses reliability issues for local area networks (LAN's).

2 Failure Detection and Recovery

A wide variety of approaches have been employed for detection of network failures. In electronic networks with binary voltage encodings (*e.g.*, RS-232), two non-zero voltages are chosen for signaling. A voltage of zero thus implies a dead line or terminal. Similarly, electronic networks based on carrier modulation infer failures from the absence of a carrier. Shared segments such as Ethernet have been more problematic, as individual nodes cannot be expected to drive the segment continuously. In such networks, many failures must be detected by higher levels in the protocol stack, as discussed later in this section.

The capacity of optical links makes physical monitoring a particularly important problem, and many techniques have been explored and used in practice. Optical encoding schemes generally rely on on-off keying, *i.e.*, the presence of light provides one signal, and its absence provides a second. With single-wavelength optics, information must be incorporated into the channel itself. One approach is to monitor time-averaged signal power, using an encoding scheme that results in a predictable distribution of on and off frequencies. A second approach utilizes overhead bits in the channel, allowing bit error rate (BER) sampling at the expense of restricting the data format used by higher levels of the protocol stack. A third approach employs a sideband to carry a pilot tone.

These approaches are complementary, and can be used in tandem.

A WDM system typically applies the single-wavelength techniques just mentioned to each wavelength, but the possibility of exploiting the multiplexing to reduce the cost of failure detection has given rise to new techniques. A single wavelength, for example, can be allocated to provide accurate estimates of BER along a link. Unfortunately, this approach may fail to detect frequency-dependent signal degradation. Pairing of monitoring wavelengths with data wavelengths reduces the likelihood of missing a frequency-dependent failure, but is too inefficient for most networks.

The approaches discussed so far have dealt with failure detection at the link level. With circuit-switched networks, the receiver on any given path can directly monitor accumulated effects along the entire path. The techniques discussed for a single wavelength can also be employed for a full path with optically transparent networks. With networks that perform optoelectronic conversion at each node, only in-band information is retained along the length of the path, and overhead in the data format is typically necessary for failure detection. Path-based approaches are advantageous in the sense that they may cover a broader set of possible failures. They get to the root of the problem: something went wrong getting from the sender to the receiver. Link-based approaches, however, make fault localization simpler, an important benefit in finding and repairing problems in the network. In practice, most backbone networks use a combination of link and path detection techniques to obtain both benefits.

Additional fault tolerance is often included in higher levels of a network protocol stack. Most protocols used for data networking (as opposed to telephony), for example, include some redundancy coding for the purposes of error detection. Typically, feedback from these layers is not provided to the physical layer, although some exceptions do exist in LAN's, such as the use of periodic packet transmissions and inference of failures when no packet arrives (see Section 8 for more detail). Instead, the error detection schemes allow the network to tolerate transient errors through temporal redundancy, *i.e.*, retransmission. Voice channels and other redundant forms of data also utilize error correction or other error tolerance techniques in some cases. A telephone circuit crossing an Asynchronous Transfer Mode (ATM) network may lose an occasional cell to a cyclic redundancy check (CRC) failure. In such a case, the cell is discarded, and the voice signal regenerated by interpolation from adjacent cells. This interpolation suffices to make a single cell loss undetectable to humans, thus as long as the transient errors occur infrequently, no loss is

noticed by the people using the circuit.

The choice of failure detection methods used in a backbone network is intertwined with the choice of strategies for restoring circuits that pass through a failed element of the network. Path monitoring, for example, does not readily provide information for failure localization. Correlated failures between paths may help to localize failures, but typically a more careful investigation must be initiated to find the problem. Path monitoring also requires that failure information propagate to the endpoints of the path, delaying detection. Link monitoring allows more rapid and local response to failures, but does not require such an approach. Instead, failure information can be propagated to the ends of each path crossing a link, while the localized failure information is retained for initiating repairs and for dynamic construction of future paths. At the algorithmic level, circuit rerouting schemes can be broadly split into path-based and link- or node-based approaches.

Prompted by the increasing reliance on high-speed communications and the requirement that these communications be robust to failures, backbone networks have generally adopted self-healing strategies to automatically restore functionality. The study of self-healing networks is often classified according to the following three criteria (see [BPF94a, Dov91], for instance) the use of link (line) rerouting versus path (or end-to-end) rerouting, the use of centralized computation versus distributed computation, and the use of precomputed versus dynamically computed routes. A succinct comparison of the different options can be found in [Wu92, pp. 291–294] and in [JJB+94]. For path recovery, when a failure leaves a node disconnected from the primary route, a back-up route, which may or may not share nodes and links with the primary route, is used. Link rerouting usually refers to the replacement of a link by links connecting the two end nodes of the failed link. When the rerouting is precomputed, the method is generally termed protection. Thus, path protection refers to precomputed recovery applied to connections following a particular path across a network. Link or node protection refers to precomputed recovery of all the traffic across a failed link or node, respectively. Figure 1 illustrates path and link rerouting. Protection routes are precomputed at a single location, and are thus centralized, although some distributed reconfiguration of optical switches may be necessary before traffic is restored. Restoration techniques, on the other hand, can rely on distributed signaling between nodes, or upon allocation of a new path by a central manager.

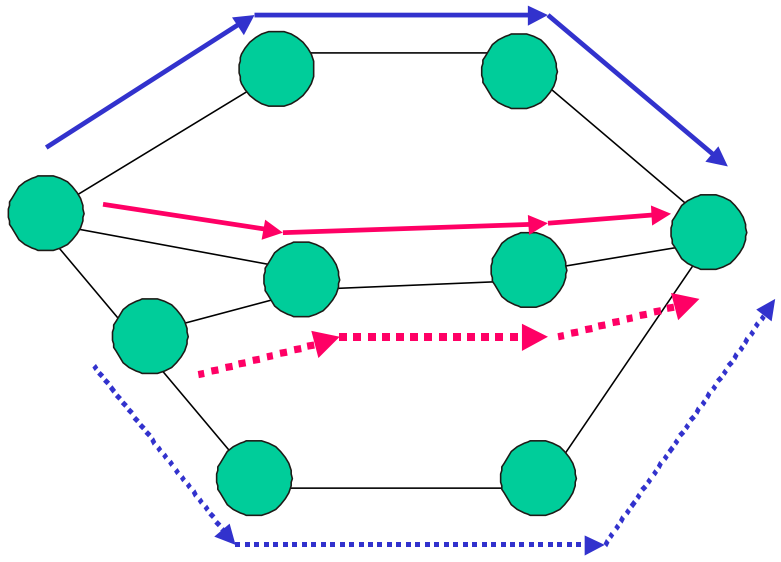


Figure 2: Path protection and associated bandwidth sharing on the back-up.

3 Path-based Schemes

Protection schemes, in which recovery routes are preplanned, generally offer better recovery speeds than restoration approaches, which search for new routes dynamically in response to a failure and generally involve software processing [NON94, VVS94]. The Synchronous Optical NETwork (SONET) specification, for example, requires that recovery time with protection approaches be under 60 milliseconds. Recovery can be achieved in tens of milliseconds using opto-mechanical add-drop multiplexers [TYKK94, Sos94], and in a few microseconds using acousto-optical switches [EDP90, WW92]. In contrast, dynamic distributed restoration using digital cross-connect systems (DCS) for ATM or SONET [Gro87, YH88, FY94, SOOH93] typically targets a two second recovery time goal [Wu94, Sos94, KA93]. Dynamic centralized path restoration for SONET [GKS96] may even take minutes [Wu94, BPG92]. The performance of several algorithms is given in [CBMS93, BCS93]. Restoration does typically require less protection capacity, however.

In this section, we focus on path protection, as the majority of current backbone networks utilize such techniques. Path protection trades longer recovery times for reduced capacity requirements relative to the link-based approaches discussed in the next section. These tradeoffs are discussed in more depth in [DW94, Fri97, BVCD97]. Path protection involves finding, for each

circuit, a back-up route (or path). Figure 2 shows two primary routes and their corresponding back-up routes. For each circuit, the two routes do not overlap on any links, implying that no single link failure can affect both a primary route and its back-up.

Path protection can itself be divided into several categories: one-plus-one (written 1+1), one-for-one (written 1:1), and one-for-N (1:N). In the first case, all data are sent along two routes simultaneously, a primary and a back-up. The two routes are link-disjoint for tolerance to link failures, or node-disjoint (excluding source and destination) for tolerance to node failures. The receiver monitors incoming traffic on the primary route; when a component along the primary route fails, the receiver switches to the back-up signal. The back-up route is typically the longer of the two, ensuring that no data are lost due to a single failure. Because both primary and back-up routes carry live traffic, the 1+1 approach is sometimes referred to as live back-up. Recovery using live back-up is extremely fast, as it is based on a local decision by a single node. Protection capacity requirements are high, however, as the back-up channel cannot be shared among connections.

The other two approaches, together known as event-triggered back-up, require less network capacity devoted to protection than does live back-up. The penalty is loss of some data when a failure occurs as well as slower restoration times relative to live back-up. With event-triggered back-up, the back-up path is only activated after a failure is detected. As with live back-up, the receiver monitors the primary path, but rather than acting locally when it detects a failure, the receiver notifies the sender that a failure has occurred on the primary path, at which point the sender begins sending traffic on the back-up path. All data transmitted or in flight between the time of the failure and the sender switching over to the back-up route are lost. With 1:1 protection, optical crossconnects are preconfigured for a particular route. Sharing and re-use of the back-up route is therefore somewhat restricted. With 1:N protection, back-up resources can be shared between any set of circuits for which the primary routes do not have resources in common, as illustrated by the two circuits in Figure 2. Two primary routes passing through a single link, for example, cannot share back-up resources; if that link should fail, only one of the two routes could be recovered. The need to configure optical crossconnects along the back-up route adds further delay to restoration time and increases data loss with 1:N protection, but sharing of back-up resources reduces protection capacity requirements by roughly 15 to 30% relative to 1+1 protection in an all-optical network. Traffic grooming, in which traffic can be assigned to wavelengths in

granularities smaller than whole wavelengths, can allow for much more effective sharing.

All forms of protection require adequate spatial redundancy in the network topology to allow advance selection of two disjoint routes for each circuit. For link (node) protection, this requirement translates to a need for two-edge (-vertex) redundant graphs; in other words, a graph must remain completely connected after removal of any single edge (vertex, along with all adjacent edges). For path protection, the same condition suffices, as shown by Menger's theorem [Men27, Sto92], which states that, given a two-edge (-vertex) redundant graph, two edge- (vertex-) disjoint routes can be found between any two vertices in the graph. A variety of schemes based on Menger's theorem have been proposed, such as sub-network connection protection (SNCP) and variations thereof [Bha94, Sha95, Suu74, ZA91, WK90, MD76], which establish two paths between every pair of nodes. However, Menger's theorem is only the starting point for designing a recovery algorithm, which must also consider rules for routing and reserving spare capacity. Path protection over arbitrary redundant networks can also be performed with trees, for example, which are more bandwidth efficient for multicast traffic [IR88, MFGB98].

With ATM, path rerouting performed by the private network node interface (PNNI) tears down virtual circuit (VC) connections after a failure, forcing the source node to establish a new end-to-end route. Back-up virtual paths (VP's) can be predetermined [KHT95] or selected jointly by the end nodes [LZL94]. Source routing, which is used by ATM PNNI, can be preplanned [HCGK97] or partially preplanned [MK96].

4 Link and Node-Based Schemes

As with path rerouting, methods commonly employed for link and node rerouting in high-speed networks can be divided into protection and restoration, although some hybrids schemes do exist [SOOH93]. The two types offer a tradeoff between adaptive use of back-up (or "spare") capacity and speed of restoration [CBJ+94, KA93]. Dynamic restoration typically involves a search for a free path using back-up capacity [HKSM94, GK93, Bak91] through broadcasting of help messages [CBMS93, FY94, Gro87, JBB+94, KA93, Wu94]. The performance of several algorithms is given in [BCS93, CBMS93]. Overheads due to message passing and software processing render dynamic processing slow. For dynamic link restoration using digital cross-connect systems, a two

second restoration time is a common goal for SONET [FY94, Gro87, KA93, Sos94, Wu94, YH88]. Preplanned methods, or link protection, depend mostly on look-up tables and switches or add-drop multiplexers. For all-optical networks, switches may operate in a matter of microseconds or nanoseconds and propagation delay dominates switching time.

Link and node protection can be viewed as a compromise between live and event-triggered path protection. Although not as capacity-efficient as 1:N path protection [BVCD97, LZL94], link protection is more efficient than live path back-up, as back-up capacity is shared between links. All traffic carried by a failed link or node is recovered independent of the circuits or end-to-end routes associated with the traffic. In particular, the two nodes adjacent to the failure initiate recovery, and only nodes local to the failure typically take part in the process. Back-up is not live, but triggered by a failure. Overviews of the different types of protection and restoration methods and comparison of the trade-offs among them can be found in [DW94, RM99, VVS94, ADDH94, XM99].

The fact that link and node protection are performed independently of the particular traffic being carried does provide an additional benefit. In particular, these approaches are independent of traffic patterns, and can be preplanned once to support arbitrary dynamic traffic loads. Path protection does not provide this feature; new protection capacity may be necessary to support additional circuits, and routes chosen without knowledge of the entire traffic load, as is necessary when allocating routes online, are often suboptimal. This benefit makes link and node restoration particularly attractive at lower layers, at which network management at any given point in the network may not be aware of the origination and destination, or of the format [Wu94] of all the traffic being carried at that location.

Link rerouting in ATM usually involves a choice of new routes by nodes adjacent to the failure [AFS+95, KST92].

5 Rings

Rings have emerged as one of the most important architectural building blocks for backbone networks in the MAN and WAN arenas. While ring networks can support both path-based and link or node-based schemes for reliability, rings merit a separate discussion because of the practical importance of the ring architecture and of its special properties.

Rings are the most common means of implementing both path and link protection in SONET, which is the dominant protocol in backbone networks. The building blocks of SONET networks are generally self-healing rings (SHR's) and diversity protection (DP) [THW91, Was91, WB90, SWC93, SGM93, SF96, STW95, GHS+94, WW92, TYKK94, Wu94, Wu92, WKC89, HT92]. SHR's are unidirectional path-switched rings (UPSR's) or bidirectional line-switched rings (BLSR's), while DP refers to physical redundancy in which a spare link (node) is assigned to one or several links (nodes) [Wu92, pp. 315-22].

In SONET, path protection is usually performed by a UPSR, as illustrated by Figure 3(a), in which a route in the clockwise direction is replaced by a route in the counterclockwise direction in response to a failure. Link rerouting is performed in a SONET BLSR using a technique known as loop-back in which traffic is sent back in the direction from which it came. Figure 3(b) illustrates this operation, in which a route in the clockwise direction is looped back (redirected) onto a counterclockwise route at the node adjacent to the failure. After traveling around the entire ring to the other end of the failed link, the route is looped back again away from the failed link, rejoining the original route. Note that, with the exception of the failed link, the final back-up route includes all of the original route. The waste of bandwidth due to traversing both directions on a single link, known as back-hauling, can be eliminated by looping back at points other than the failure location [Mag97, KTK94]. In case of failure of a node on a BLSR, the failure is handled in a manner similar to a link failure. The failure of a node is equivalent to the failure of a meta-link consisting of the node and the two links adjacent to it. The only difference is that network management must be able to purge any traffic directed to the failed node from the network.

Loop-back operations can be performed on entire fibers or on individual wavelengths. With fiber-based loop-back, all traffic carried by a fiber is backed by another fiber, regardless of how many wavelengths are involved. If traffic is allowed in both directions in a network, fiber-based loop-back relies on four fibers, as illustrated in Figure 4. In WDM-based loop-back, restoration is performed in a wavelength-by-wavelength basis.

WDM-based loop-back requires at least two fibers. Figure 5 illustrates WDM-based loop-back. A two-fiber counter-propagating WDM system can be used for WDM-based loop-back, even if traffic is allowed in both directions. Note that WDM loop-back as shown on Figure 5 does not require any change of wavelength: traffic initially carried by λ_1 is backed up by the same

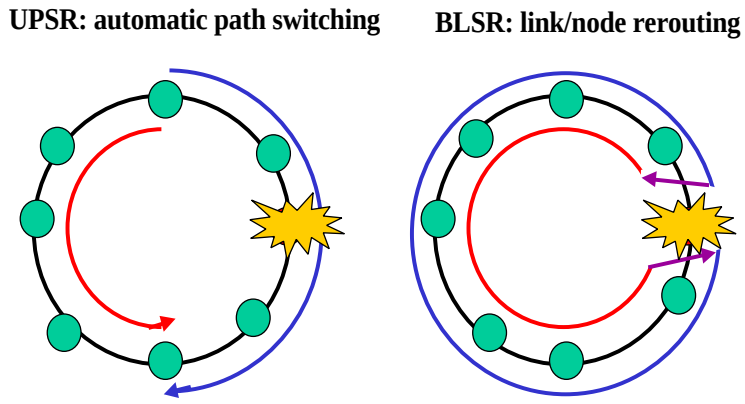


Figure 3: Path protection and node protection in a ring.

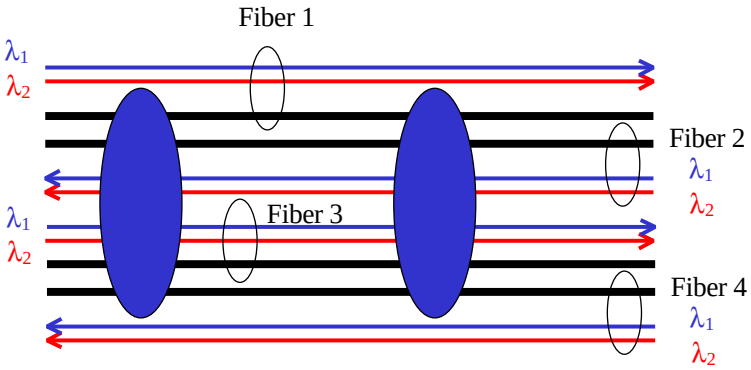


Figure 4: Four-fiber system with fiber-based loop-back. Primary traffic is carried by fiber 1 and by fiber 2. Back-up is provided by fiber 3 for fiber 1 and by fiber 4 for fiber 2.

wavelength. Obviating the need for wavelength changing is economical and efficient in WDM networks. One could, of course, back up traffic from λ_1 on fiber 1 onto λ_2 on fiber 2, if there were advantages to such wavelength changing, for instance in terms of wavelength assignment for certain traffic patterns. We can easily extend the model to a system with more fibers, as long as the back-up for a certain wavelength on a certain fiber is provided by some wavelength on another fiber. Moreover, we may change the fiber and/or wavelengths from one fiber section to another. For instance, the back-up to λ_1 on fiber 1 may be λ_1 on fiber 2 on a two-fiber section and λ_2 on fiber 3 on another section with four fibers. Note, also, that we could elect not to back up λ_1 on fiber 1 and instead use λ_1 on fiber 1 for primary traffic. The extension to systems with more fibers,

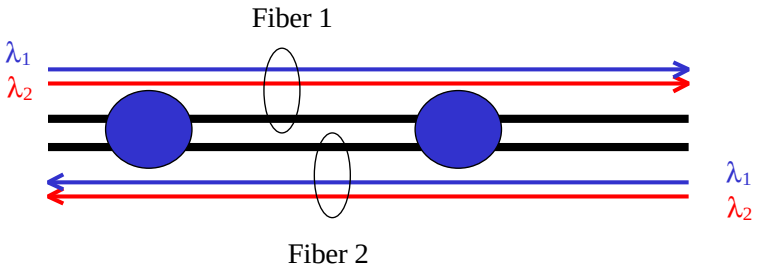


Figure 5: Two-fiber WDM-based loop-back. Primary traffic is carried by fiber 1 on λ_1 and by fiber 2 on λ_2. Back-up is provided by on λ_1 on fiber 2 for λ_1 on fiber 1. λ_2 on fiber 2 is backed up by λ_2 on fiber 1.

inter-wavelength back-ups and back-ups among fiber sections can be readily done.

The finer granularity of WDM-based recovery systems provides several advantages over fiber-based systems. First, if fibers carry at most half of their total capacity, only two fibers rather than four are necessary to provide recovery. Thus, a user need only lease two fibers, rather than paying for unused bandwidth over four fibers. On existing four-fiber systems, fibers could be leased by pairs rather than fours, allowing two leases of two fibers each for a single four-fiber system. The second advantage is that, in fiber based-systems, certain wavelengths may be selectively given restoration capability. For instance, half the wavelengths on a fiber may be assigned protection, while the rest may have no protection. Different wavelengths may thus afford different levels of restoration QoS, which can be reflected in pricing. In fiber-based restoration, all the traffic carried by a fiber is restored via another fiber. If each fiber is less than half full, WDM-based loop-back can help avoid the use of counterpropagating wavelengths on the same fiber. Counterpropagating wavelengths on the same fiber are intended to enable duplex operation in situations that require a full fiber's worth of capacity in each direction and which have scarce fiber resources. However, counterpropagation on the same fiber is onerous and reduces the number of wavelengths that a fiber can carry with respect to unidirectional propagation. WDM-based loop-back may make using two unidirectional fibers preferable to using two counterpropagating fibers, for which one fiber is a back-up for the other.

When more than one ring is required, rings must be interconnected. In SONET, the usual method to handle nodes shared between rings is called matched nodes. Figure 6 shows matched

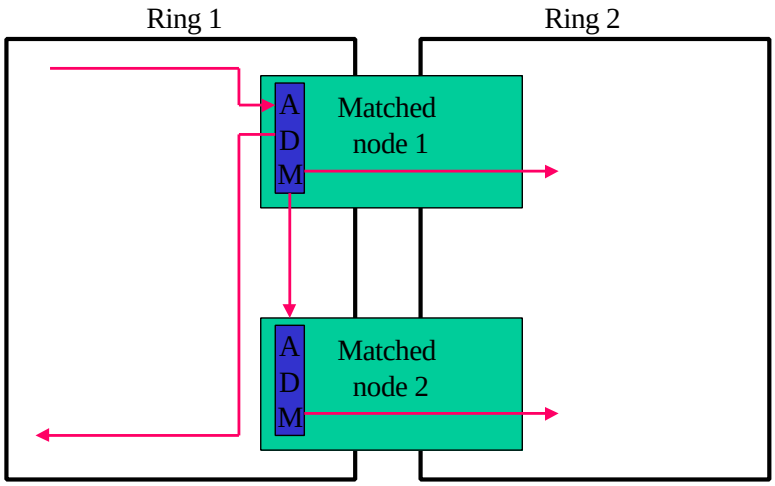


Figure 6: Matched nodes.

nodes under normal operating conditions. Consider traffic moving from ring 1 to ring 2; traffic in the reverse direction is handled similarly. Under normal operation, matched node 1 is responsible for all inter-ring communications. Matched node 1 houses an add-drop multiplexer (ADM) that performs a drop and continue operation. The drop and continue operation consists of duplicating all traffic through matched node 1 and transmitting it to matched node 2. Thus, matched node 2 has a live back-up of all the traffic arriving to matched node 1, and mirrors the operation of matched node 1. However, under normal operating conditions, ring 2 disregards the output from matched node 2. Failure of any node other than the primary matched node is handled by a single ring in a stand-alone manner. Failure of the secondary matched node treats intra-ring and inter-ring traffic differently. Note that, depending on the failure mode of the primary matched node, the failure may be seen by both rings or by a single ring. Indeed, failures may occur only on access cards interfacing with one or the other ring, or a wholesale failure may be detected by both rings. Loop-back is performed on all the rings that see the failure. Intra-ring traffic is recovered within the ring wherein it lies and inter-ring traffic is handled by the second node. How to extend matched nodes to cases other than simple extensions of the topology shown in Figure 6 is generally not known. For instance, the case in which a node is shared by more than one ring is difficult. Similarly, the case in which two adjacent rings share links without duplication of resources, as shown in Figure 8, is complicated in the case of shared nodes.

Many schemes besides SONET exist or have been proposed to enable ring-based networks, usually using optical fiber as the medium. Fiber Distributed Data Interface (FDDI) [Ros89, Ros90, LaM91] is such a scheme. Both FDDI and IEEE 802.5 control access to the ring by passing an electronic token from node to node. Only the node with the token is allowed to transmit. Multiple ring topologies may be interconnected through a hub [JL98], or rings may coexist in a logically interconnected fashion over a single physical ring [MBL+97b, MBL+99, MBL+97a], or rings may be arranged hierarchically [BDF+, JL98, LG97].

The IEEE 802.17 Resilient Packet Ring (RPR) Working Group has recently been set up to investigate the use, mainly at the MAN, of an optical ring architecture coupled with a packet-based MAC. The purpose of this project is to combine the robustness of rings with a flexible MAC that is well-suited to current optical access applications [Gro].

6 Mesh Networks

In this section, we expand our exploration of topologies to redundant meshes. Restricting a network to use only DP and SHR's is a constraint that has cost implications for building and expanding networks [WKC89]; see [Sto92] for an overview of the design of topologies under certain reliability constraints. Ring-based architectures may be more expensive than meshes [BPF94b, WKC89], and as nodes are added, or networks are interconnected, ring-based structures may be difficult to preserve, thus limiting their scalability [WKC89, Wu92, WKC88]. However, rings are not necessary to construct fault tolerant networks [NV91, WH91]. Mesh-based topologies can also provide redundancy [Sto92, JHC93, WKC88]. Even if we constrain ourselves to always use ring-based architectures, such architectures may not easily bear changes and additions as the network grows . For instance, adding a new node, connected to its two nearest node neighbors, will preserve mesh structure, but may not preserve ring structure. Our arguments indicate that, for reasons of cost and extensibility, mesh-based architectures are more promising than interconnected rings.

Algorithmic approaches to general mesh restoration are often difficult, however, and implementations can be substantially more complex. To address this problem, many techniques attempt to find rings within the meshes. Overlays using rings are obtained by placing cycles atop existing mesh networks. Each such cycle creates a ring. Service protection or restoration is then generally

obtained on each ring as though it were a physical ring. Covering mesh topologies with rings is a means of providing both mesh topologies and distributed, ring-based restoration. Numerous approaches ensure link restorability by finding covers of rings for networks. Many of these techniques have been proposed in the context of backbone networks in order to enable recovery over mesh topologies.

One such approach is to cover nodes in the network by rings [Was91]. In this manner, a portion of links are covered by rings. If primary routings are restricted to the covered links, link restoration can be effected on each ring in the same manner as in a traditional SHR, by routing back-up traffic around the ring in the opposite direction to the primary traffic. Using such an approach, the uncovered links can be used to carry unprotected traffic, *i.e.*, traffic that may not be restored if the link that carries it fails. However, under some conditions it may not be possible to cover all nodes with a single ring, or the length of the resulting may be undesirable. A large ring forces long routes for many connections. Such long routes have several drawbacks, both from the point of view of routing (reduced wavelength-assignment efficiency) and from the point of view of communications (excessive jitter).

To allow every link to carry protected traffic, other ring-based approaches ensure that every link is covered by a ring. One approach to selecting such covers is to cover a network with rings so that every link is part of at least one ring [Gro92]. Several issues arise concerning the overlap and interconnection of such rings. Many of these issues are similar to issues encountered in SONET network deployment. The two main issues are management of links logically included in two rings and node management for ring interconnection.

The first issue concerns the case in which a single link is located on two rings. If that link bears a sufficient number of fibers or wavelengths, the two rings can be operated independently over that link, as shown in Figure 7. However, the resources available to the overlay network may require sharing the resources over that link. Figure 8 shows such a network, in which only a single wavelength is available to the overlay network. In such a case, the logical fibers must be physically routed through available physical fibers, with network management acting to ensure that conflicts are avoided on the shared span. Such operations incur significant overhead.

The second issue relates to node interconnection among rings. Minimizing the amount of fiber required to obtain redundancy using ring covers is equivalent to finding the minimum cycle cover

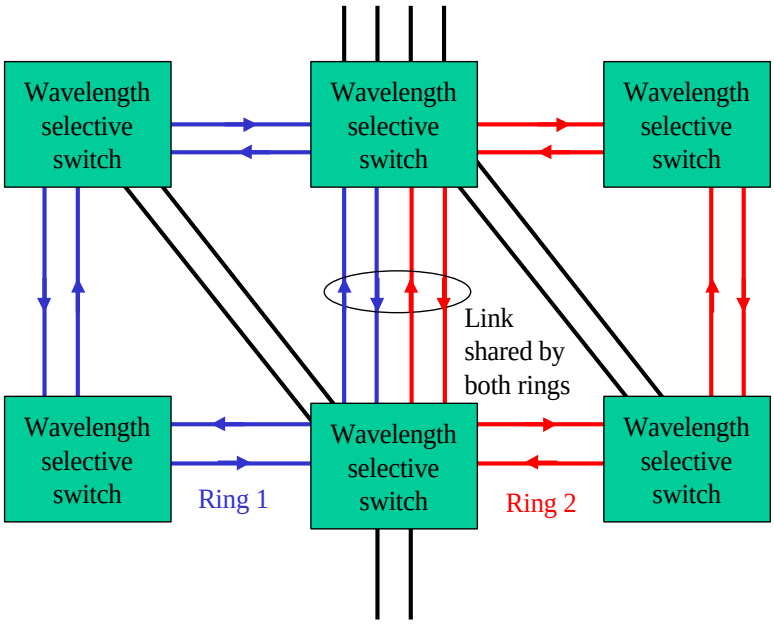


Figure 7: Two rings traversing separate resources over the same link.

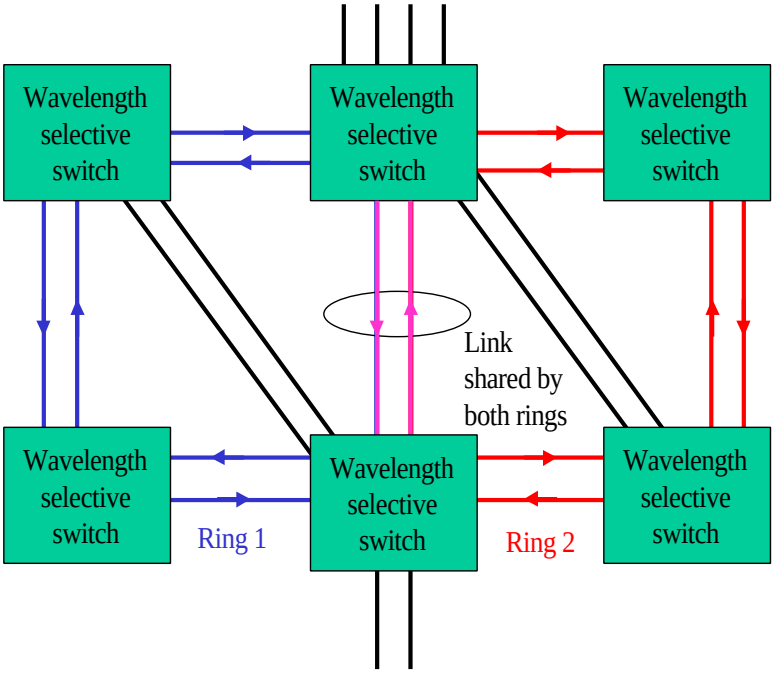


Figure 8: Two rings traversing shared resources over the same link.

of a graph, an NP-complete problem [Tho97, ILPR81], although bounds on the total length of the cycle cover may be found [Fan92].

A more recent approach to ring covers, intended to overcome the difficulties of previous approaches, is to cover every link with exactly two rings, each with two fibers. The ability to perform loop-back style restoration over any link in mesh topologies was first introduced in [ESH97, ES96], using certain types of ring covers. In particular, [ESH97] considers link failure restoration in optical networks with arbitrary two-link redundant arbitrary mesh topologies and bidirectional links. The approach is an application of the double-cycle ring cover [Jae85, Sey79, Sze73], which selects cycles in such a way that each edge is covered by two cycles. For planar graphs, the problem can be solved in polynomial-time; for non-planar graphs, it is conjectured that double cycle covers exist, and a counterexample must have certain properties unlikely to occur in network graphs [God85]. Cycles can be used as rings to perform restoration. Each cycle corresponds either to a primary or a secondary two-fiber ring. Let us consider a link covered by two rings, rings 1 and 2. If we assign a direction to ring 1 and the opposite direction to ring 2, ring-based recovery using the double cycle cover uses ring 2 to back up ring 1. This recovery technique is similar to recovery in conventional SHR's, except that the two rings that form four-fiber SHR's are no longer co-located over their entire length. In the case of four fiber systems, with two fibers in the same direction per ring, we have fiber-based recovery, because fibers are backed up by fibers. Extending this notion to WDM-based loop-back, each ring is both primary for certain wavelengths and secondary for the remaining wavelengths. For simplicity, let us again consider just two wavelengths. Figure 9 shows that we cannot assign primary and secondary wavelengths in such a way that a wavelength is secondary or primary over a whole ring.

The use of double cycle covers can also lead to asymmetric restoration times for a bidirectional connection. In particular, the links and nodes used to recover traffic crossing a link often depends on the direction of the traffic, with each direction being recovered by a separate cycle. The two directions on a link thus have different restoration times and timing jitter, which can lead to problems for bidirectional connections. In contrast, both SHR's and generalized loop-back (discussed later in the section) avoid these problems by protecting bidirectional traffic with unique, bidirectional restoration paths.

Cycle covers work well for link failures, but have drawback for recovery from node failures,

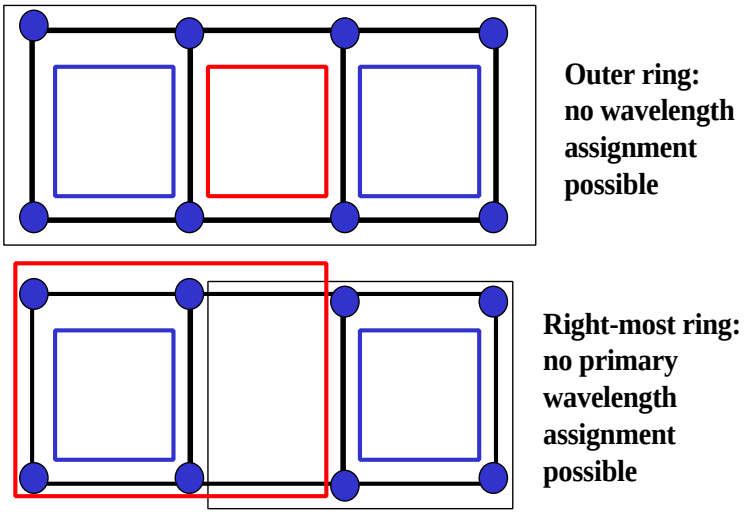


Figure 9: Example showing the problems of applying double cycle covers to wavelength recovery.

particularly when failures occur at nodes that are shared by one than one link. While node recovery can be effected with double cycle ring covers, such restoration requires cumbersome hopping among rings. Moreover, if a link or node is added to a network, the cover of cycles can change significantly, limiting the scalability of double cycle covers. These drawbacks are a general property of ring embeddings, and are already found in SONET networks.

In order to avoid the limitations of ring covers, an approach using preconfigured protection cycles, or p-cycles, is given in [GS98]. A p-cycle is a cycle on a redundant mesh network. Links on the p-cycle are recovered by using the p-cycle as a conventional BLSR. Links not on the p-cycle are recovered by selecting, along the p-cycle, a path connecting the nodes at either end of the failed link. Some difficulty arises from the fact that several p-cycles may be required to cover a network, making management among p-cycles necessary. A single p-cycle may be insufficient because a Hamiltonian circuit might not exist, even in a two-connected graph. Even finding p-cycles that cover a large number of nodes may be difficult. Some results [Fou85, Jac80, ZLY85] and conjectures [HJ85, Woo75] exist concerning the length of maximal cycles in two-connected graphs. The p-cycle approach is in effect a hybrid ring approach, which mixes link protection (for links not on the p-cycle) with ring recovery (for links on the p-cycle).

Another approach to link restoration on mesh networks, which we term generalized loop-back,

was first presented in [MFB99]. The principle behind generalized loop-back is to select a directed graph, called the primary graph, such that another directed graph, called the secondary, can be used to carry back-up traffic for any link failure in the primary. Construction of a primary involves selection of a single direction for each link in the network. Loop-back then occurs along the secondary graph in a manner akin to SONET BLSR. Figure 10 demonstrates generalized loop-back for a simple network. In the figure, only two fibers per link are shown—one primary fiber and its corresponding secondary fiber. When the link $[Y, X]$ fails, traffic from the primary digraph floods onto the secondary digraph starting at Y. The secondary digraph carries this back-up traffic from one endpoint of the failed node to the other endpoint, possibly along multiple paths. When traffic reaches X (along the first successful path), it is again placed on the primary fiber, as though no failure had occurred. Unnecessary back-up paths are subsequently torn down. The fact that multiple paths may exist for restoration allows us to reclaim some arcs (fibers) from secondary digraphs to carry additional traffic. The capacity efficiency obtained in this manner is, for typical networks, in the order of 20% over methods, such as double cycle cover, that require half of the network capacity to be devoted to recovery.

7 Packet-Based Approaches

This section looks more closely at techniques particular to packet-based networks, giving more details for the failure detection techniques mentioned in Section 2 and the recovery schemes used when failures are detected. Some packet-based networks, such as FDDI, are based on ring topologies and use failure detection and recovery techniques nearly identical to those described in previous sections. For packet networks built with redundant mesh topologies, the approach is substantially different.

One protocol of particular interest and importance was developed in the mid-80's [Per85] to allow redundant interconnection of LAN's for reliability while avoiding problems of infinite routing loops. The model, known as an extended LAN, uses bridges to connect LAN's and to forward traffic between LAN's as necessary. The bridges cooperate in a distributed fashion to select a minimum spanning tree (MST) from the full connectivity graph. All traffic is then routed along the MST, preventing cycles. Periodic configuration messages are sent from the root of the tree and

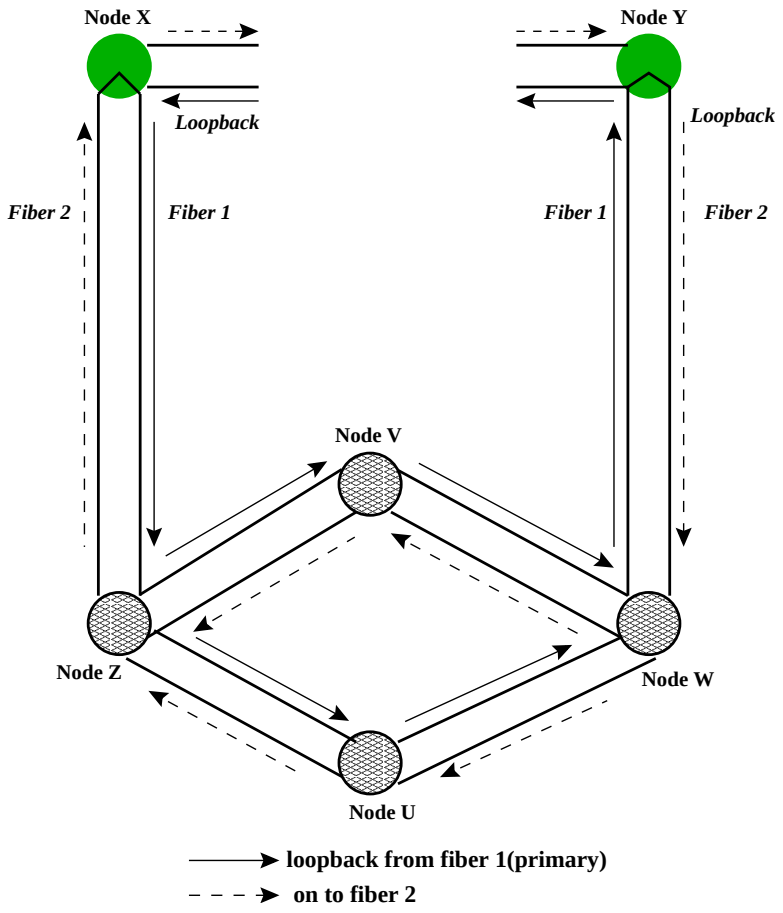


Figure 10: Generalized loop-back.

forwarded over all LAN's. Failure detection then relies on timeout mechanisms: if a bridge fails to hear the configuration message on any LAN to which it is attached, it acts to find a new route from the spanning tree to that LAN. The simplest method is to restart the search entirely, but some optimizations are possible.

The MST approach serves its purpose, allowing redundant topologies in packet-switched networks without allowing for routing loops. However, restricting traffic to flow along a tree can severely limit the capacity of the extended LAN, and can force traffic between two LAN's that are close in the original mesh to follow a long path through the tree, consuming some of the capacity of many other LAN's in the process.

Autonet, developed in the early 90's, solved this problem to some extent by allowing the use of all links in the network [SBB+91]. Routing cycles and deadlocks in Autonet were prevented

through the use of up*/down* (read “up star, down star,” and alluding to regular expressions) routes. With this approach, all routers are assigned a unique number, and packets between two hosts in the network must follow a route that moves in a monotone sequence upwards followed by a monotone sequence downwards before exiting the network. Any route obeying this constraint can be used. Consider a cycle or self-cycle in routes, and label each link in the cycle as either up or down, depending on the identifiers assigned to the routers at the end of the link. Obviously, a cycle must contain both up and down links, and in particular must contain a two-link section with the first link down and the second up. No route can legally follow both links in this section, however, implying that deadlocks are impossible, *i.e.*, all up*/down* routes are mutually deadlock-free. Autonet relied on timeouts built into the switch hardware for detection of failed links and node, but was otherwise quite similar to the extended LAN approach in terms of reliability.

Within the last decade, many vendors of packet-based networks have recognized the importance of redundancy, and have introduced hardware and software support for combining physical channels into single logical channels between high-end switches. While this approach may seem fairly natural in a packet-based network, in which utilization is already based on statistical multiplexing of the links, some complexities must be addressed. These complexities arise from an assumption by higher-level protocols, in particular the Transmission Control Protocol (TCP), that packets sent through a network arrive in order of transmission. Use of multiple physical routes to carry packets from a single TCP connection often violates this assumption, causing TCP’s congestion control mechanisms to drastically cut bandwidth. To address this issue, link aggregation schemes try to restrict individual TCP connections to specific links, and rely on the availability of many connections to provide good load balancing and capacity benefits from aggregation. Failure detection, as with Autonet, is generally handled in hardware, and results in routing reorganization. Unlike many backbone networks, the capacity of most packet-switched networks degrades in the presence of failures, encouraging network architects and managers to operate in somewhat risky modes in which inadequate capacity remains available after certain failures. This phenomenon can be observed even in high-end packet-based systems, including some SAN’s backing bank operations.

In the wide area, fault tolerance in packet-switched networks relies on a combination of physical layer notification and rerouting by higher-level routing protocols. The Border Gateway Pro-

tocol (BGP) [RL95], a peer protocol to the Internet Protocol (IP), defines the rules for advertising and selecting routes to networks in the Internet. More specifically, it defines a homogeneous set of rules for interactions between Autonomous Systems (AS's), networks controlled by a single administrative entity. With each AS, administrators are free to select whatever routing protocol suits their fancy, but AS's must interact with each other in a standard way, as defined by BGP. BGP explicitly propagates failure information in the form of withdrawn routes, which cancel previously advertised routes to specific networks.

From the point of view of recovery in the context of optical backbone networks, the overhead required for packet-based systems depends critically on what functionalities are implemented in the optical domain. Restoration, in which, after a failure, excess bandwidth is claimed for the purpose of providing alternate routes to traffic around a failure, is challenging in the optical domain, since it requires operating on the whole data stream and possibly separating packets from a stream. In order to avoid packet-level operations, flow switching on a stream-by-stream basis, for instance using multi-protocol label switching (MPLS), is a promising alternative. Such stream-based operations are more amenable to optical processing. For recovery, stream-based processing reduces roughly to circuit-based recovery.

Packet-switched approaches for optical access seek to perform some subset of the functionalities of traditional opto-electronic packet-based networks optically [Mod99]. These functionalities may be header recognition, buffering, packet insertion, packet reading, packet retrieval, rate conversion. Performing such operations in the optical domain is challenging and no consensus has emerged regarding implementation. However, certain general statements can be made. Operations, such as buffering a stream, that involve significant timing issues or that introduce loss and distortion on the data stream, tend to be challenging. Replicating a stream, for instance, can be done using passive optical splitters and is therefore relatively straightforward. Merging streams, on the other hand, is challenging because of timing issues. The most challenging operations are the ones performed at the packet level. Again, different levels of difficulty arise. Reading signals from an optical data stream is possible by removing a fraction of the signal power and operating on that fraction. Retrieving a packet (reading the packet and removing it from the stream) is difficult because it involves performing an operation on the whole stream, as well as timing, phase and polarization issues. Thus, operations such as packet-switching are also challenging because of issues

of timing and speed of optical switches. Thus, fully optical packet-switched systems replicating the entire operations of electronic systems are still distant.

8 High-Speed LAN's

The vast majority of the proposed architectures for LAN's consist of star topologies or of networks built from combinations of star topologies, in which some type of switch, router, or other type of hub, is placed in the center of a topology and each node is directly connected to the hub [HRS93]. The emergent 10 Gb/s standard (IEEE 802.3ae) for LAN's and MAN's also allows for optical stars and trees. From the point of reliability, stars present many weaknesses. In particular, a failure at the hub may entail failure of the whole network. However, other failures may occur even without outright failure of the hub. If the hub passively broadcasts, total failure of the hub is unlikely. However, many partial failure scenarios exist: amplifier failures; port connection failures, at the access nodes or at the hub; transmitter or receiver failures at access nodes, for instance because of laser failures; cabling failures in the fiber itself. Such failures entail the failure of one or more arms of the star.

The center of a star topology is inherently a single point of failure, making complete replication of the system necessary to support recovery. Operation of a fully redundant system is difficult, however, as illustrated by existing reliable networks based on star topologies. Many enterprise networks and storage area networks (SAN's), for example, are built as stars.

Such systems typically use single-wavelength optical connections rather than WDM, and rely on electronic switching. Enterprise networks are usually based on Gigabit Ethernet (GigE), while SAN's are based on Fibre Channel (FC). Network interface cards (NIC's), housing both a receiver and a transmitter, are optically connected to an electronic switch. The switch is closer to a traditional router than to the passive broadcast hubs or wavelength-selective switches discussed in the context of star-based WDM LAN's. For such networks, redundancy is obtained by full duplication of all resources, as shown in Figure 11.

In addition to replication, the two switches must be connected. Consider the case of failure of the primary NIC in server 1. Server 1 communicates via the secondary switch. Requiring other servers also to communicate via the secondary switch is undesirable. Indeed, although we show

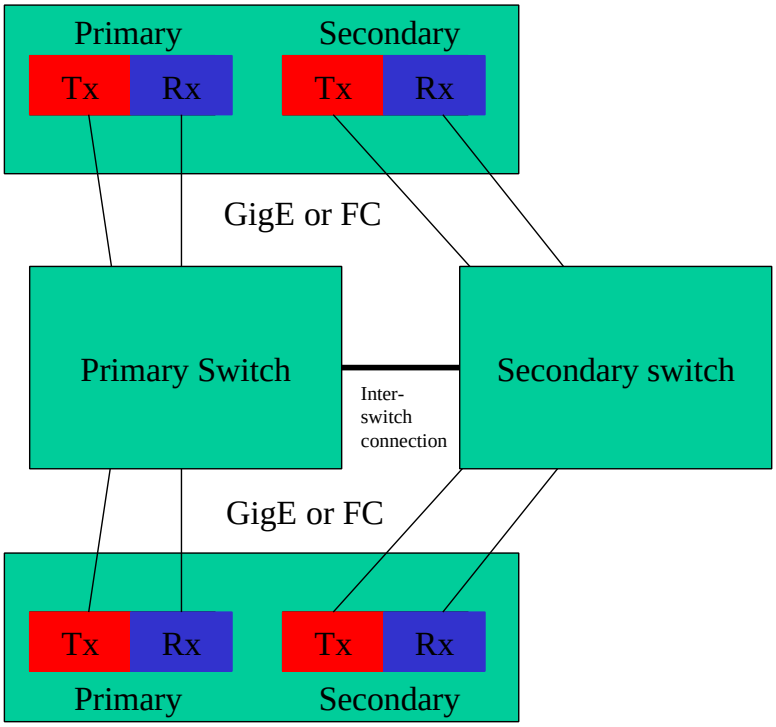


Figure 11: Redundant architecture in an enterprise network or LAN using traditional star topologies.

just two servers, such networks typically have many servers connected to them and reconfiguring so many connections simultaneously is difficult. Moreover, there is some delay involved in creating new connections through switch 2 owing to initialization overheads. To avoid reconfiguration at all servers, all servers other than server 1 continue to communicate with the primary switch and the two switches communicate with each other via the inter-switch connection.

In the context of optical networks, an inter-switch connection translates into connection between two hubs. In order to manage such an inter-hub connection, the hub needs to be equipped with far greater capabilities than simple optical broadcasting. Thus, it would appear that optical star dedicated networks will be difficult to deploy and that the means of providing robustness available in traditional star topologies cannot be easily extended to optical access networks.

The star topology for LAN's connected to a backbone is not limited to optical applications. Such an architecture has been proposed, for instance, for the Integrated Services LAN (ISLAN) defined by IEEE 802.9 using unshielded twisted pair.

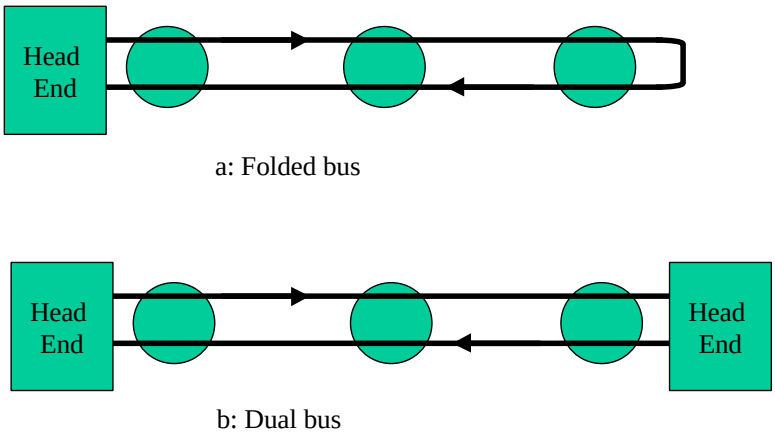


Figure 12: Dual and folded buses.

While stars and topologies built from stars dominate in the LAN's, LAN's are also built using bus schemes. Bus schemes allow nodes to place and retrieve traffic using a shared medium. Figure 12 shows a folded bus and a dual bus. In a folded bus, a single bus, originating at a head end, serves all nodes. Typically, nodes use the bus first as a collection bus, onto which they place traffic (in the left to right direction in Figure 12(a). The last node folds back the bus to make it travel in the right to left direction. In the right to left direction, nodes collect traffic placed onto the bus. The traffic may be read only or read and removed. In the dual bus architecture, two buses are used, each with its own head end. Folded and dual buses are simple options for LAN's and certain types of MAN's. In particular, they offer an effective way of sharing bandwidth among several users and are therefore attractive to allow nodes to access optical bandwidth, whether it be a full fiber, a few wavelengths, or a single wavelength.

Folded and dual buses suffer from reliability drawbacks. Figure 13 shows a folded bus and a dual bus after a failure. Partial recovery can be effected by creating a bus on either side of the failure. For a dual bus architecture, the node immediately upstream of the failure needs to be able to fold the bus. In order to re-establish full connectivity after a failure, the end nodes of the original buses must be able to connect outside of the original buses to transmit traffic that was destined to traverse the cut.

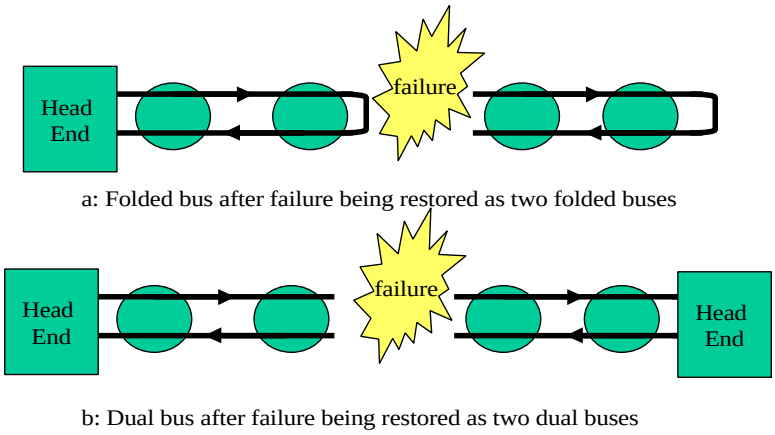


Figure 13: Dual and folded buses after a failure.

References

[AAC+92] T. Anderson, A. Avizienis, W. C. Carter, A. Costes, F. Cristian, Y. Koga, H. Kopetz, J. H. Lala, J. C. Laprie, J. F. Meyer, B. Randell, A. S. Robinson, L. Simoncini, and U. Voges. *Dependability: Basic Concepts and Terminology*. Springer-Verlag, Wien, Austria, 1992.

[ADDH94] J. Anderson, B.T. Doshi, S. Dravida, and P. Harshavardhana. Fast restoration of ATM networks. *IEEE Journal on Selected Areas in Communications*, 12(1):128–136, January 1994.

[AFS+95] M. Azuma, Y. Fujii, Y. Sato, T. Chujo, and K. Murakami. Network restoration algorithm for multimedia communication services and its performance characteristics. *IEICE Transactions on Communications*, E78-B(7):987–994, July 1995.

[Bak91] J.E. Baker. A distributed link restoration algorithm with robust preplanning. In *Proceedings IEEE GLOBECOM*, volume 1, pages 10.4.1–10.4.6, 1991.

[BCS93] J. Bicknell, C. E. Chow, and S. Syed. Performance analysis of fast distributed network restoration algorithms. In *Proceedings IEEE GLOBECOM*, volume 3, pages 1596–1600, 1993.

[BDF+] A. Bianco, V. Distefano, A. Fumagalli, E. Leonardi, and F. Neri. A-posteriori access strategies in all-optical slotted WDM rings. In *Global Telecommunications Conference*.

[Bha94] R. Bhandari. Optimal diverse routing in telecommunication fiber networks. In *Proceedings IEEE INFOCOM*, volume 3, pages 11c.3.1–11.c.3.11, May 1994.

[BPF94a] A. Banerjea, C.J. Parris, and D. Ferrari. Recovering guaranteed performance service connections from single and multiple faults. In *globecom*, volume 1, pages 162–166, 1994.

[BPF94b] A. Banerjea, C.J. Parris, and D. Ferrari. Recovering guaranteed performance service connections from single and multiple faults. In *globecom*, volume 1, pages 162–166, 1994.

[BPG92] M. Barezzani, E. Pedrinelli, and M. Gerla. Protection planning in transmission networks. In *icc*, pages 316.4.1–316.4.5, 1992.

[BVCD97] N. Wauters B. Van Caenegem and P. Demeester. Spare capacity assignment for different restoration strategies in mesh survivable networks. In *icc*, volume 1, pages 288–292, 1997.

[CBJ+94] R.S.K. Chng, C.P. Botham, D. Johnson, G.N. Brown, M.C. Sinclair, M.J. O'Mahony, and I.Hawker. A multi-layer restoration strategy for reconfigurable networks. In *globecom*, volume 3, pages 1872–1878, 1994.

[CBMS93] C.E. Chow, J. Bicknell, S. McCaughey, and S. Syed. A fast distributed network restoration algorithm. In *Proceedings of the 12^{th} International Phoenix Conference on Computers and Communications*, volume 1, pages 261–267, March 1993.

[Dov91] R. Doverspike. A multi-layered model for survivability in intra-LATA transport networks. In *globecom*, pages 2025–2031, 1991.

[DW94] R. Doverspike and B. Wilson. Comparison of capacity efficiency of dcs network restoration routing techniques. volume 2, 1994.

[EDP90] D. Edinger, P. Duthie, and G.R. Prabhakara. A new answer to fiber protection. pages 53–55, April 9, 1990.

[ES96] G. Ellinas and T. E. Stern. Automatic protection switching for link failures in optical networks with bi-directional links. In *globecom*, 1996.

[ESH97] G. Ellinas, T. E. Stern, and A. Hailemarian. Link failure restoration in optical networks with arbitrary mesh topologies and bi-directional links. 1997.

[Fan92] G. Fan. Covering graphs by cycles. In *SIAM*, volume 5, pages 491–496, November 1992.

[Fou85] I. Fournier. Longest cycles in 2-connected graphs of independence number α. In North-Holland, editor, *Cycles in Graphs*, pages 201–204. Annals of Discrete Mathematics 115, 1985.

[Fri97] T. Frisanco. Optimal spare capacity design for various protection switching methods in atm networks. In *icc*, volume 1, pages 293–298, 1997.

[FY94] H. Fujii and N. Yoshikai. Double search self-healing algorithm and its characteristics. volume 77, pages 975–995, 1994.

[GHS+94] L. M. Gardner, M. Heydri, J. Shah, I. H. Sudborough, I. G. Trollis, and C. Xia. Techniques for finding ring covers in survivable networks. In *Proceedings of GLOBECOM*, volume 3, 1994.

[GK93] A. Gersht and S. Kheradpir. Real-time bandwidth allocation and path restorations in SONET-based self-healing mesh networks. In *Proceedings of the IEEE International Conference on Communications*, volume 1, pages 250–255, 1993.

[GKS96] A. Gersht, S. Kheradpir, and A. Shulman. Dynamic bandwidth-allocation and path-restoration in SONET self-healing networks. In *IEEE Transactions on Reliability*, volume 45, pages 321–331, June 1996.

[God85] L. Goddyn. A girth requirement for the double cycle cover conjecture. In North-Holland, editor, *Cycles in Graphs*, pages 13–26. Annals of Discrete Mathematics 115, 1985.

[GR00] O. Gerstel and R. Ramaswami. Optical layer survivability - an implementation perspective. In *IEEE Journal on Selected Areas in Communications*, pages 1885–1899, 2000.

[Gre95] C.J. Green. Protocols for a self-healing network. *Proceedings of the Military Communications Conference (MILCOM)*, 1:252–256, 1995.

[Gro] Resilient Packet Ring Working Group.

[Gro87] W.D. Grover. The selfhealing^{TM} network. In *globecom*, pages 1090–1095, 1987.

[Gro92] W.D. Grover. Case studies of survivable ring, mesh and mesh-arc hybrid networks. In *globecom*, pages 633–638, 1992.

[GS98] W.D. Grover and D. Stamatelakis. Cycle-oriented distributed preconfiguration: Ring-like speed with mesh-like capacity for self-planning network reconfiguration. In *Proceedings of the IEEE International Conference on Communications*, volume 2, pages 537–543, 1998.

[HCGK97] D.K. Hsing, B.-C. Cheng, G. Goncu, and L. Kant. A restoration methodology based on pre-planned source routing in ATM networks. *icc*, 1, 277-282 1997.

[HJ85] R. H*ä*ggkvist and B. Jackson. A note on maximal cycles in 2-connected graphs. In North-Holland, editor, *Cycles in Graphs*, pages 205–208. Annals of Discrete Mathematics 115, 1985.

[HKSM94] S. Hasegawa, A. Kanemasa, H. Sakaguchi, and R. Maruta. Dynamic reconfiguration of digital cross-connect systems with network control and management. In *globecom*, pages 28.3.2–28.3.5, 1994.

[HRS93] P.A. Humblet, R. Ramaswami, and K.N. Sivarajan. An efficient communication protocol for high-speed packet-switched multichannel networks. In *IEEE Journal on Selected Areas in Communications*, volume 11, pages 568 –578, May 1993.

[HT92] E.L. Hahne and T.D. Todd. Fault-tolerant multimesh networks. In *globecom*, pages 627–632, 1992.

[ILPR81] A. Itai, R.J. Lipton, C.H. Papadimitriou, and M. Rodeh. Covering graphs with simple circuits. *SIAM Journal of Computing*, 10:746–750, 1981.

[IR88] A. Itai and M. Rodeh. The multi-tree approach to reliability in distributed networks. Number 79, 1988.

[Jac80] B. Jackson. Hamilton cycles in regular 2-connected graphs. *Journal Comb. Theory Ser. B*, 29:27–46, 1980.

[Jae85] F. Jaeger. A survey of the double cycle cover conjecture. In North-Holland, editor, *Cycles in Graphs*. Annals of Discrete Mathematics 115, 1985.

[JBB+94] D. Johnson, G.N. Brown, S.L. Beggs, C.P. Botham, I. Hawker. R.S.K. Chng, M.C. Sinclair, and M.J. O'Mahony. Distributed rstoration strategies in telecommunications networks. In *Proceedings of the IEEE International Conference on Communications*, volume 1, pages 483–488, 1994.

[JHC93] R-H Jan, F-J Hwang, and S-T Cheng. Topologocial optimization of a communication network subject to a reliability constraint. *IEEE Transactions on Reliability*, 42(1), March 1993.

[JJB+94] D. Johnson, G.N. Johnson, S.L. Beggs, C. Botahm, I. Hawker, R.S.K. Chng, M.C. Sinclair, and M.J. O'Mahony. Distributed restoration strategies in telecommunications networks. In *icc*, volume 1, pages 483–488, 1994.

[JL98] T.S. Jones and A. Louri. Media access protocols for a scalable optical interconnection network, May 1998.

[KA93] H. Kobrinski and M. Azuma. Distributed control algorithms for dynamic restoration in dcs mesh networks: Performance evaluation. In *globecom*, volume 3, pages 1584–1588, 1993.

[KHT95] R. Kawamura, H. Hadama, and I. Tokizawa. Implementation of self-healing function in ATM networks. *Journal of Network and Systems Management*, 3(3), 243–264 1995.

[KST92] R. Kawamura, K. Sato, and I. Tokizawa. High-speed self-healing techniques utilizing virtual paths. 5^{th} *International Network Planning Symposium*, May 1992.

[KTK94] Y. Kajiyama, N. Tokura, and K. Kikuchi. An atm vp-based self-healing ring. volume 12, January 1994.

[LABJ00] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. Delayed Internet Routing Convergence. In *Proceedings of ACM SIGCOMM Conference*, pages 175–87, 2000.

[LaM91] R.O. LaMaire. FDDI performance at 1 Gbit/s. In *IEEE International Conference on Communications*, pages 174 –183, 1991.

[LAWV01] C. Labovitz, A. Ahuja, R. Watterhofer, and S. Venkatachary. The Impact of Internet Policy and Topology on Delayed Routing Convergence. In *Proceedings of INFOCOM*, 2001.

[LG97] A. Louri and R. Gupta. Hierarchical optical interconnection network HORN: scalable interconnection network for multiprocessors and multicomputers, January 1997.

[LZL94] N.D. Lin, A. Zolfaghari, and B. Lusignan. ATM virtual path self-healing based on a new path restoration protocol. *globecom*, 2:794–798, 1994.

[Mag97] R.B. Magill. A bandwidth efficient self-healing ring for b-isdn. In *Proceedings of ICC*, 1997.

[MBL+97a] M.A. Marsan, A. Bianco, E. Leonardi, A. A. Morabito, and F. Neri. SR/sup 3/:a bandwidth-reservation MACprotocol for multimedia applications over all-optical WDM multi-rings. In *INFOCOM '97*, volume 2, 1997.

[MBL+97b] M.A. Marsan, A. Bianco, E. Leonardi, F. Neri, and S. Toniolo. An almost optimal MACprotocol for all-optical WDM multi-rings with tunable transmitters and fixed receivers. In *IEEE International Conference on Communications*, volume 1, pages 437 – 442, May 1997.

[MBL+99] M.A. Marsan, A. Bianco, E. Leonardi, A. Morabito, and F. Neri. All-optical WDMmulti-rings with differentiated qos. In *IEEE Communications Magazine*, volume 37, pages 58 –66, Feb. 1999.

[MD76] P. Mateti and N. Deo. On algorithms for enumerating all circuits of a graph. volume 5, March 1976.

[Men27] K. Menger. *Zur allgemeinen Kurventheorie*. Fundamenta Mathematicae, 1927.

[MFB99] M. Médard, S. G. Finn, and R. A. Barry. Wdm loop-back recovery in mesh networks. In *Proceedings IEEE INFOCOM*, 1999.

[MFGB98] M. Médard, S. G. Finn, R. G. Gallager, and R. A. Barry. Redundant trees for automatic protection switching in arbitrary node-redundant or edge-redundant graphs. In *Proceedings of ICC*, 1998.

[MK96] K. Murakami and H.S. Kim. Virtual path routing for survivable ATM networks. In *IEEE/ACM Transactions on Networking*, volume 4, 1996.

[Mod99] E. Modiano. Wdm-based packet networks. In *IEEE Communications Magazine*, volume 37, pages 130–135, March 1999.

[NON94] R. Nakamura, H. Ono, and K. Nishikawara. Reliable switching services. In *globecom*, volume 3, pages 1596–1600, 1994.

[NV91] K.T. Newport and P.K. Varshney. Design of survivable communication networks under performance constraints. *IEEE Transactions on Reliability*, 40:433–440, October, 1991.

[Per85] R. Perlman. An Algorithm for Distributed Computation of a Spanning Tree in an Extended LAN. In *Proceedings of the 9^{th} Symposium on Data Communications*, pages 44–53, Whistler Moutain, British Columbia, Canada, 1985. SIGCOMM'85.

[PO97] D.J. Pai and H.L. Owen. An algorithm for bandwidth management with survivability constraints in ATM networks. In *icc*, volume 1, pages 261–266, 1997.

[RL95] Y. Rekhter and T. Li. A Border Gateway Protocol 4 (BGP-4). Internet Engineering Task Force RFC 1771, March 1995.

[RM99] S. Ramamurthy and B. Mukherjee. Survivable WDM mesh networks, part I - protection. In *Proceedings IEEE INFOCOM*, pages 744–751, 1999.

[Ros89] F.E. Ross. An overview of FDDI: the fiber distributed data interface. In *IEEE Journal on Selected Areas in Communications*, volume 7, pages 1043 –1051, Sept. 1989.

[Ros90] F.E. Ross. Fiber distributed data interface: an overview. In *15th Conference on Local Computer Networks*, pages 6–11, 1990.

[SB00] T. E. Stern and K. Bala. *Multiwavelength Optical Networks: A Layered Approach*. Prentice Hall, Upper Saddle River, New Jersey, 2000.

[SBB+91] M. D. Schroeder, A. D. Birrell, M. Burrows, H. Murray, R. M. Needham, T. L. Rodeheffer, E. H. Satterthwaite, and C. P. Thacker. Autonet: A High-Speed, Self-Configuring Local Area Network using Point-to-Point Links. *jsac*, 9:1318–35, October 1991.

[Sey79] P.D. Seymour. Sums of circuits. In U.S.R. Murty J.A. Bondy, editor, *Graph Theory and Related Topics*, pages 341–355. Academic Press, NY, 1979.

[SF96] J. Shi and J. Fonseka. Interconnection of self-healing rings. In *icc*, volume 1, 1996.

[SGM93] J. B. Slevinsky, W. D. Grover, and M. H. MacGregor. An algorithm for survivable network design employing multiple self-healing rings. In *Proceedings of GLOBECOM*, volume 3, pages 1568–1573, 1993.

[Sha95] S.Z. Shaikh. Span-disjoint paths for physical diversity in networks. In *Proceedings of the IEEE Symposium on Computers and Communications*, pages 127–133, 1995.

[SOOH93] H. Sakauchi, Y. Okanoue, H. Okazaki, and S. Hasegawa. Distributed self-healing control in SONET. *Journal of Network and Systems Management*, 1(2):123–141, 1993.

[Sos94] J. Sosnosky. Service application for sonet dcs distributed restoration. volume 12, pages 59–68, 1994.

[Sto92] M. Stoer. *Design of Survivable Networks*. Springer-Verlag, 1992.

[STW95] C.-C. Shyur, S.-H Tsao, and Y.-M. Wu. Survivable network planning methods and tools in taiwan. September 1995.

[Suu74] J. W. Suurballe. Disjoint paths in a network. pages 125–145, 1974.

[SWC93] C.-C. Shyur, Y.-M. Wu, and C.-H. Chen. A capacity comparison for sonet self-healing ring networks. In *Proceedings of GLOBECOM*, pages 1574–1578, 1993.

[Sze73] G. Szekeres. Polyhedral decomposition of cubic graphs. *J. Australian Math. Soc.*, 8:367–387, 1973.

[Tho97] C. Thomassen. On the complexity of finding a minimum cycle cover of a graph. In *SIAM*, volume 26, pages 675–677, June 1997.

[THW91] M. Boyden T.-H. Wu, R.H. Caldwell. A multi-period design model for survivable network architecture selection for sdh/sonet interoffice networks. volume 40, pages 417–432, October 1991.

[TYKK94] M. Tomizawa, Y. Yamabayashi, N. Kawase, and Y. Kobayashi. Self-healing algorithm for logical mesh connection on ring networks. volume 30, pages 1615–1616, September 15 1994.

[VVS94] J. Veerasamy, S. Vemkatesan, and J.C. Shah. Effect of traffic splitting on link and path restoration planning. In *globecom*, volume 3, pages 1867–1871, 1994.

[Was91] O.J. Wasem. An algorithm for designing rings for survivable fiber networks. volume 40, 1991.

[WB90] T.-H. Wu and M. E. Burrows. Feasibility study of a high-speed sonet self-healing ring architecture in future interoffice networks. pages 33–51, November 1990.

[WH91] T.H. Wu and S. Fouad Habiby. Strategies and technologies for planning a cost-effective survivable network architecture using optical switches. *IEEE Transactions on Reliability*, 8(2):152–159, February 1991.

[WK90] J. S. Whalen and J. Kenney. Finding maximal link disjoint paths in a multigraph. In *Proceedings of GLOBECOM*, pages 403.6.1–403.6.5, 1990.

[WKC88] T.H. Wu, D.J. Kolar, and R.H. Cardwell. Survivable network architectures for broad-band fiber optic networks: Model and performance comparison. *IEEE Journal of Lightwave Communications*, 6(11), November 1988.

[WKC89] T.-H. Wu, D. J. Kolar, and R. H. Cardwell. High-speed self-healing ring architectures for future interoffice networks. In *Proccedings of GLOBECOM*, volume 2, pages 23.1.–23.1.7, 1989.

[Woo75] D.R. Woodall. Maximal circuits of graphs II. *Studia Sci. Math. Hungar.*, 10:103–109, 1975.

[Wu92] T.-H. Wu. *Fiber Network Service Survivability*. Artech House, 1992.

[Wu94] T.-H. Wu. A passive protected self-healing mesh network architecture and applications. volume IEEE/ACM Transactions on Networking, pages 40–52, February 1994.

[WW92] T.-H. Wu and W.I. Way. A novel passive protected sonet bidirectional self-healing ring architecture. In *jlt*, volume 10, September 1992.

[XM99] Y. Xiong and L.G. Mason. Restoration strategies and spare capacity requirements in self-healing ATM networks. *IEEE Journal of Lightwave Communications*, 7(1):98–110, February 1999.

[YH88] C.H. Yang and S. Hasegawa. Fitness: Failure immunization technology for network service survivability. In *globecom*, volume 3, pages 47.3.1–47.3.6, 1988.

[ZA91] W. T. Zaumen and J. J. Garcia-Luna Aceves. Dynamics of distributed shortest-path routing algorithms. In *Proceedings of the 21st SIGCOMM Conference*, volume 21, pages 31–43. ACM Press, September 3-6 1991.

[ZLY85] Y. Zhu, Z. Liu, and Z. Yu. An improvement of Jackson's result on Hamilton cycles in 2-connected graphs. In North-Holland, editor, *Cycles in Graphs*, pages 237–247. Annals of Discrete Mathematics 115, 1985.

[ZS00] D. Zhou and S. Subramanian. Survivability in optical networks. In *IEEE Network*, pages 16–23, November/December 2000.